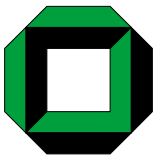

Universität Karlsruhe
Fakultät für Informatik
Institut für Telematik
76128 Karlsruhe

Architektur vernetzter Systeme - Drahtlose Netze

Seminar Wintersemester 2000/2001

Herausgeber:
Verena Kahmann
Rainer Ruggaber
Dr. Jochen Seitz
Prof. L. Wolf

Universität Karlsruhe
Institut für Telematik

Interner Bericht 2001-02
ISSN 1432-7864

Inhaltsverzeichnis

Vorwort

Drahtlose Netze gewinnen eine immer größere Bedeutung für die Anbindung von vorwiegend mobilen Nutzern. Dieser Seminarband, der in der Reihe "Architektur vernetzter Systeme" erscheint, hat zum Ziel verschiedene Aspekte drahtloser Netze zu untersuchen und vorzustellen, und trägt deshalb den Untertitel "Drahtlose Netze". Dabei kommen neben der Technik verschiedener drahtloser Netzwerktechnologien, auch deren Einsatzbereiche zur Sprache. Aufbauend auf die Netzwerktechnologie ist es notwendig, geeignete Protokolle der Vermittlungsschicht und der Transportschicht zu entwickeln. Im Bereich der Anwendungen werden sowohl neue Dienste vorgestellt, wie auch Mechanismen, um bestehende Dienste in einer Umgebung betreiben zu können, in der Verbindungsabbrüche häufig auftreten. Konkret werden die folgenden Themenstellungen behandelt:

Drahtlose lokale Netze nach IEEE 802.11
Drahtlose lokale Netze nach IEEE 802.11 werden bei vielen Firmen vermehrt eingesetzt, um Computer einfach und schnell miteinander zuverbinden. Die technischen Grundlagen für Bitübertragungschicht und Medienzugriffssteuerung werden kurz vorgestellt, wie auch die Probleme und rechtlichen Aspekte.

QoS in IEEE 802.11-Netzen
Der IEEE 802.11-Standard ist ein robuster und weitverbreiteter Standard für drahtlose lokale Netzwerke. Drei verschiedene Übertragungsvarianten der physikalischen Schicht, die wiederum beliebig mit zwei verschiedenen Funktionsweisen der MAC-Schicht kombiniert werden können, sind vorgesehen, die im vorliegenden Semesterbeitrag vorgestellt werden. Dabei liegt das Augenmerk auf den damit einhergehenden Dienstgütezusicherungen. Die Möglichkeit, die Bandbreite eines drahtlosen Netzes in Abhängigkeit verschiedener Anwendungen zu verteilen, wird ergänzend angesprochen, ist im aktuellen 802.11-Standard jedoch noch nicht verankert.

Drahtlose Campus-Netze
In den letzten Jahren wurden von verschiedenen Universitäten WLAN-Projekte nach IEEE 802.11 initiiert. Diese Netze unterscheiden sich in der Motivation, im Aufbau und im Betrieb. Diese Arbeit hat das Ziel, diese Unterschiede beispielhaft an drei Universitäten aufzuzeigen. Dabei spielen auch Authentifizierung und Verwaltung eine Rolle. Der letzte Punkt hängt unmittelbar mit der eingestetzten Hardware zusammen, die daher auch im Beitrag Berücksichtigung findet.

HiperLAN/2
HiperLAN/2 ist ein Wireless LAN Standard der 2. Generation. Er wurde aufgrund des gestiegenen Bedarfs an WLANs geschaffen. Durch verschiedene Modulationsverfahren sind Bandbreiten von bis zu 54 Mbit/s (auf Schicht 1) möglich. HiperLAN/2 integriert Funktionen zur Verschlüsselung und Authentififzierung, aber auch Dienstgüte-

Parameter (QoS). HiperLAN/2 definiert allerdings nur die ersten zwei Layer und ist somit auf die Layer anderer Netzwerk-Protokolle angewiesen. Um diese optimal umsetzen zu können, wurde der Convergence Layer mit speziellen Sublayern geschaffen. Sublayer existieren beispielsweise für ATM und Ethernet, aber auch UMTS.

GPRS (General Packet Radio Service)
Diese Seminararbeit beschäftigt sich mit dem Thema GPRS (General Packet Radio Service). Sie entstand aus den gewonnenen Erkenntnissen eines viermonatigen Praktikums bei der Firma 12snap in München. Ziel war es, zu untersuchen, inwieweit das Geschäftsmodell, auf das später noch näher eingegangen werden soll, durch GPRS beeinflusst wird. Daher gliedert sich die Arbeit in zwei Teile. Der erste Abschnitt wird GPRS unter dem Gesichtspunkt der technischen Neuerungen betrachten, sozusagen die Perspektive der Netzwerkbetreiber. Der zweite, etwas kürzere Teil wird GPRS aus den Augen eines Serviceanbieters beleuchten. Hierbei wird zuvor auf das Beispiel von i-mode eingegangen, einem System, das sich in Japan schon großer Beliebtheit erfreut und GPRS sehr ähnlich ist. Darauf aufbauend wird auf die Kombination WAP und GPRS ausführlicher Bezug genommen. Mit den konkreten Implikationen der Firma 12snap wird dieser zweite Hauptteil abgerundet.

Optimierungen von Mobile IP
Mobile IP wurde von der Internet Engineering Task Force (IETF) als Erweiterung des IP-Protokolls entwickelt. Damit gab es einen Ansatz, der eine Datenübertragung aus dem Internet bei gleichzeitiger Mobilität unterstützte. Doch ließ dieser Ansatz noch Wünsche offen. In dieser Arbeit werden einige Verfahren vorgestellt, die eine Verbesserung von Mobile IP zum Ziel haben. Erster Ansatz ist die Verbesserung der micro-Mobilität, die sich HAWAII und Cellular IP zur Aufgabe gemacht haben. Bei einem weiteren Vorschlag geht es um die Optimierung beim Routing vom Correspondent Node zum Mobile Node. Die wichtigsten Ansätze sind die zur Verbesserung der Sicherheit sowohl der Netzwerke als auch der Verbindungen zum Mobile Node, die in FATIMA und mit AAA-Servern gemacht werden.

Alternative Transportprotokolle für drahtlose Netze
TCP hat sich als ungeeignet für den Einsatz in drahtlosen Netzen herausgestellt.Dies liegt daran, daß die in Funknetzwerken häufigen Paketverluste von TCP als Stauanzeichen interpretiert werden. TCP führt alle Segmentverluste auf einen Stau im Netzinneren zurück und drosselt somit seine Senderate. Dies verursacht eine deutliche Verschlechterung des Durchsatzes auf der drahtlosen Strecke. In dieser Arbeit sollen zunächst Gründe für die schlechte Performance von TCP im drahtlosen Bereich untersucht werden. Danach werden die einzelnen Lösungsansätze klassifiziert und vorgestellt. Zum Schluss wird noch auf den Sonderfall der Ad-hoc Netzwerke eingegangen.

Positionsbestimmung in drahtlosen Netzen
Die Bestimmung der Position eines mobilen Endgerätes in einem drahtlosen Netz ist

ein in den letzten fünf Jahren vielseitig diskutiertes Thema. In dieser Seminararbeit werden einige der wichtigsten Techniken zur Positonsbestimmung vorgestellt. Dazu gehören TOA, TDOA, AOA, Messung des Phasenwinkels und Auswertung der Signalstärke. Durch die Zielsetzung der E–911 Bestimmungen der FCC gilt insbesondere der Positionsbestimmung in GSM-Netzen besondere Aufmerksamkeit. Auch in Gebäuden gibt es vielfältige Anwendungsmöglichkeiten für Lokalisationssysteme. Da dort GSM und GPS nicht funktionieren, wird in diesem Zusammenhang auf Ansätze zur Positionsbestimmung mittels drahtloser lokaler Netze und Infrarottechnik eingegangen.

Kontextabhängige Dienste
Der Begriff Kontext umfasst alle Eigenschaften der Umgebung sowie die Situation, in der ein Ereignis stattfindet. Kontextabhängige Dienste sind Dienste bzw. Anwendungen die Kenntnisse haben über den aktuellen Kontext, und die diese Kenntnisse nutzen, um sich auf verschiedenste Art und Weise an die aktuelle Situation und Umgebung anzupassen. Die vorliegende Arbeit beschreibt zunächst den Begriff "kontextabhängige Dienste" genauer und zeigt, wie Kontextinformation gewonnen werden kann. Es werden vielfältige Möglichkeiten aufgezeigt, Kontextinformation in zukünftigen Anwendungen und Systemen umzusetzen. Gleichzeitig wird gezeigt, dass es bereits heute zahlreiche Systeme gibt, die Kontextinformation erfolgreich nutzen. Dabei wird insbesondere der von der Universität Lancaster entwickelte interaktive Reiseführer GUIDE vorgestellt.

Middlewaresysteme in partitionierbaren Netzen
In diesem Aufsatz wird der "Jgroup"-Ansatz der Universität Bologna vorgestellt, der auf eine differenzierte Transparenz für die Teilnehmer in Netzen mit verteilten Anwendungen abzielt. Hervorzuheben ist, dass Jgroup partitionierbare Netze betrachtet und spezielle Tools entwickelt hat, die durch Partitionen entstehenden Komplikationen zu lösen. Es werden drei Komponenten beschrieben: Der Gruppenkommunikationsdienst, die Methode der Anfragestellungen und der Dienst zum Verschmelzen von Partitionen. Durch die Verwendung replizierter und über dem Netz verteilter Dienste wird eine höhere Leistungsfähigkeit (bezüglich Zuverlässigkeit und Verfügbarkeit) erreicht, als in den bisher eingesetzten Middleware-Systemen wie z.B. CORBA. Im Anschluss an die Vorstellung des Jgroup Konzepts werden Vergleiche zu anderen verteilten Systemen gezogen und die Anstrengungen der Object Management Group für ein drahtloses CORBA beschrieben, das sich mit einer in Teilaspekten vergleichbaren Aufgabenstellung beschäftigt.

Mensch-Maschine-Interaktion im Bereich Mobiler Endgeräte
Der Einsatz mobiler Geräte lässt die Verwendung verschiedener Anwendungen an jedem Ort und zu jeder Zeit zu. Voraussetzungen sind jedoch eine hohe Leistungsfähigkeit und kleine Ausmaße der Geräte. Da einige Ein-/Ausgabemechanismen nicht beliebig verkleinerbar sind, ist der Einsatz neuer und angepasster Mechanismen notwendig. Dieser Beitrag beschreibt zunächst einige dieser Mechanismen. Er stellt Verfahren zur Beurteilung vor und vergleicht ausgewählte Eingabemechanismen.

Bevor nun die einzelnen Ausarbeitungen der Seminarbeiträge präsentiert werden, möchten wir allen beteiligten Studenten für ihre engagierte Mitarbeit danken, ohne die weder der Erfolg des Seminars noch die Anfertigung des vorliegenden Berichts möglich gewesen wäre. Hierzu haben auch die nach den einzelnen Vorträgen stattfindenden Diskussionen maßgeblich beigetragen.

Karlsruhe, im Februar 2001

Verena Kahmann Rainer Ruggaber Jochen Seitz Prof. L. Wolf

Drahtlose lokale Netze nach IEEE 802.11

Christian Gladigow

Kurzfassung

Drahtlose lokale Netze nach IEEE 802.11 werden bei vielen Firmen vermehrt eingesetzt , um Computer einfach und schnell miteinander zuverbinden. Die technischen Grundlagen für Bitübertragungschicht und Medienzugriffssteuerung werden kurz vorgestellt,wie auch die Probleme und rechtlichen Aspekte.

1 Einleitung

Zwei der bekanntesten Vertreter der IEEE 802.x Norm sind die Ethernet-Definition (802.3) und die Token Ring-Definition (802.5) Eines der Ziele von IEEE 802.11 war die Sicherstellung der herstellerübergreifenden Kommunikation bei drahtlosen lokalen Netzen. Der 802.11 Standard definiert die beiden untersten Schichten des OSI-Modells, die Bitübertragungsschicht (Physical Layer,PHY) und die Sicherungsschicht (Data Link Layer),letztere wird noch in die Medienzugriffskontrollschicht (Media Access Control Layer, MAC) und die Logical Link Control-Schicht(LLC) unterteilt. Die LLC stellt die Schnittstelle zur Netzwerkschicht dar, ab der dann sich nichts mehr von den anderer Definitionen unterscheidet.

2 Charakterisierung von drahtlosen lokalen Netzen

Drahtlose lokale Netze können bezüglich der Systemarchitektur in zwei unterschiedliche Varianten eingeteilten werden, in die Infrastrukturnetze und in Ad-hoc-Netzwerk . Bei Infrastrukturnetzen sind ein oder mehrere Endgeräte mit einem Access Point über eine Funkstrecke verbunden .Alle Stationen und der zugehörige Access Point bilden ein sogenannten Basic Service Set(BSS). Durch ein Distribution System (DS) werden mehrere BSS miteinander verbunden, dies kann über Funkstrecke oder über fest verlegte Kabel geschehen. Unter anderem verbindet das DS die drahtlosen Netze mit einem Portal, das wiederum beliebige andere Netze mit einander verbindet.

Bei Ad-hoc-Netzen bilden mehrere Stationen ,die mit dem selben Verfahren und auf den selben Frequenzen kommunizieren ein BSS.Mehrere Ad-hoc-Netze können entweder durch genügend großen räumlichen Abstand oder durch Verwendung von unterschiedlichen Trägerfrequenzen gebildet werden. Dann können sich auch mehrere Netze überlappen.

3 Physikalische Schicht

Drahtlose Funknetze werden auf der physikalischen Ebene nach zwei unterschiedlichen, nicht zu einander kompatiblen Verfahren aufgebaut. Die 3. Variante der drahtlosen Netze sind die Infrarotnetze, die ebenfalls von der 802.11- Definition erfasst werden. Diese unterscheiden sich aber grundsätzlich von dem oben genannten, da sie als Trägermedium Licht verwenden.

Probleme bei der Funkübertragung ergeben sich besonders durch die Mehrwegeausbreitung (siehe Abb1) die durch Reflexion,Beugung der Funkwellen zum Beispiel an Wänden, Decken, Möbeln ,usw. entstehen können [Schi00]. Durch die starke Interferenzen der auf unterschiedlichen Wegen zum Empfänger eintreffenden Signalen kann es Zeitweise zu erheblichen Einbrüchen des empfangene Signals kommen .

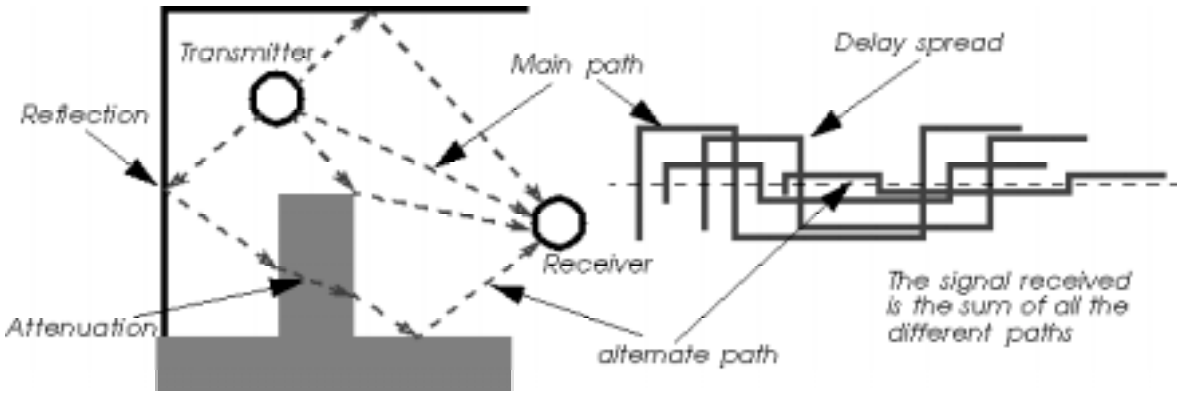

Abbildung 1: Entstehung und Folge der Mehrwegausbreitung

Bei der Spreizbandtechnik besteht der Trick darin, das zu übertragende Signal auf einen breiteren Frequenzspektrum zu übertragen wie es eigentlich nach dem Shannon-Hartley-Gesetz notwendig wäre. Man erreicht damit, dass sich in jedem Teil des Frequenzspektrums die übertragene Signalleistung (spektrale Leistungsdichte) wesentlich verringert. Dies macht die Signalübertragung unter anderem gegen schmalbandigere Störungen, sei es durch natürliche Störquellen oder gezielte Störversuche anderer Sender in die Frequenzbereich unempfindlicher. Darüber hinaus sorgt die Bandspreizung für eine deutlich gleichmäßiger Ausleuchtung des Raums.

3.1 Direct Sequence Spread Sprectrum (DSSS) bei 802.11

Aus dem Shannon-Hartley-Gesetz ergibt sich, dass bei einer konstanten Störabstand und einer Vergrößerung der Übertragungsbandbreite die übertragbare Datenrate proportional zunimmt. Das DSSS-Verfahren nutzt dies dazu aus, die zu sendenden Nutzdaten zunächst mit einem bestimmten Bit-Muster, der sogenannten Speizfrequenz, zu kodieren: es verknüpft Nutzdaten und Code-Sequenz binär mit einem Exklusiv-Oder.

Vor der Übertragung muss der Sender das kodierte Signal mit einem Modulationsverfahren in den gewünschten Frequenzbereich transformieren. Als Modulationsverfahren kommt entweder Differential Binary Phase Shift Keying (BPSK) oder Differential Quadrature Phase Shiftkeyinging(QPSK) zum Einsatz. BPSK gilt als sicherer gegen

Abhören, bietet allerdings nur eine Datenrate von 1 Mbit. QPSK verdoppelt die Datenrate durch ein anderes Kodierverfahren. Innerhalb des ISM-Bandes stehen hierfür in Europa neun Frequenzbänder mit je 22 MHz Bandbreite zur Verfühgung, wobei sich benachbarte Bänder jeweils überlappen. Die Spreizsequenz besteht aus so genannten 'Chips' wobei ein Bit 11 Chips entspricht . Der zu sendende Datenstrom hat jetzt eine um den Faktor elf höhere Datenrate als das ursprüngliche zu sendende Signal: somit nimmt auch die erforderliche Übertragungsbandbreite um diese Spreizfaktor zu.

Der Empfänger demoduliert zunächst das Signal und multipliziert das Ergebnis mit der selben Spreizfrequenz wie beim Senden, indem er beide mit einem binären XOR verknüpft, und rekonstruiert wieder das Orginalsignal. Diese Verfahren ist weitgehend unempfindlich gegenüber Übertragungsstörungen, wie schmalbandige Störquellen. Die Orginaldatenbits lassen sich über jeweils eine Chiplänge ' T ' relativ einfach wieder restaurieren.

Der Clou an diesem aufwendigen Verfahren ist die Auswahl einer geigneten Spreizsfrequenz. Diese berechnet das sendende Gerät mit Hilfe eines Pseudo-Zufallsgenerators so, das das modulierte Signal einen bandbegrenzten ('rosa') Rauschen gleicht. Damit erreicht man auch einen wirkungsvollen Schutz gegen unberechtigten Abhören des Signals. Wenn nun der Sender die Amplitude des gesendeten Signals so niedrig hält, dass sie nur niedrig über oder sogar unterhalb der allgemeinen Rauschleistung liegt, können Dritte zunächst nicht erkennen, dass überhaupt ein Nutzsignal vorliegt. Selbst wenn ein unbefugter Lauscher ein Signal empfängt, wird ihm ohne genaue Kenntnis der beim Senden verwendeten Spreizsequenz die Entschlüsselung des kodierten Datenstroms nicht gelingen.

3.2 Frequency Hopping Spread Spectrum(FHSS)

Beim Frequenzsprungverfahren teilt man den im ISB-Band zur Verfügung stehenden Frequenzbereich in 80 Kanäle mit je 1 Mhz Bandbreite auf. Die Bandspreizung realisiert der Sender, indem er mit einer bestimmten Sprungfrequenz in festen Zeitabständen von einem Kanal zum nächsten wechselt. Wenn diese Zeitabstände kürzer sind als die zum Senden eines einzelnen Nutzdatenbits (bei 2 Mbit/s sind dies 500 ns) spricht man von Fast Frequencyhopping(FFH) , sonst von Slow Frequency Hopping (SFH). IEEE 802.11 benutzt für SFH nur 2,5 Sprünge pro Sekunde.

Für die Wahl der Sprungsequenz gelten ähnliche Kriterien wie für die Spreizfrequenz beim DSSS: Der Sender muss das Signal mit der Zeit so über die zur Verführung stehenden Kanäle verteilen, dass das Frequenzspektrum einem rosa Rauschen gleicht. Ein Empfänger kann das Signal nur vom Rauschen unterscheiden und es dekodieren , wenn er die Sprungsequenz des Senders kennt.

4 Medienzugriffssteuerung

Der MAC-Layer besteht im Funk-LAN aus mehreren funktionellen Blöcken. Sämtliche Blöcke arbeiten unabhängig von Datenraten oder physischen Eigenschaften des

eingesetzten Physical Layer. Es stehen auf unterster Ebene drei verschiedene Übertragungsverfahren zur Verfügung, die aus Sicht einer auf das Netz zugreifenden Applikation völlig transparent erscheinen. Schließlich wird über eine einheitliche MAC-Ebene zugegriffen.

4.1 CSMA/CA

Als Zugriffsverfahren kommt CSMA/CA zum Einsatz (Carrier Sense Multiple Access with Collision Avoidance). Das Verfahren ähnelt dem im Ethernet (CSMA/CD - Collision Detection), allerdings ist im Funknetz das Erkennen einer kollidierenden Übertragung (zwei Stationen senden parallel) mit physikalischen Mechanismen nicht möglich. Daher muß eine Station, die auf dem Funk-LAN Daten senden will, zunächst überprüfen, ob sich bereits Datenpakete im Äther befinden. Wenn das Medium (also die Funkstrecke) für eine definierte Zeit frei ist, darf die Station einen Frame (Datenblock) senden.

Ist das Medium besetzt, wählt die Station nachdem, das Medium wieder frei ist, ein zufälliges Zeitintervall (Backoff Interval) und verzögert ihre Aussendung um dieses Intervall,nachdem sie einen Timer herunterzählt. Ist das Medium dann immer noch frei, kann die Station senden. Falls das Medium in der Zwischenzeit wieder besetzt ist, hält der Timer an. Irgendwannist der Timer gleich Null , so daß die Station auf jeden Fall eine Sendeberechtigung erhält.

4.2 RTS/CTS

Request to send (RTS) und Clear to send(CTS) werden beim sogenannten 'Hidden Node Problem' benötigt. Dieses Situation entsteht dadurch, wenn sich eine Station im Empfangsbereich von zwei weiteren Stationen befindet, diese sich aber nicht gegenseitig empfangen können. Es wäre dann möglich, dass die beiden Stationen ein freies Medium feststellen und beide anfangen zu senden, was wiederum eine Kollision zur Folge hätte. Diese Situation kann umschifft werden, indem das in der IEEE 802.11 Definition festgelegte RTS Signal von Sender zum Empfänger gesendet wird , falls das Medium frei ist. Empfängt der Empfänger das RTS-Signal, antwortet er mit einen CTS-Signal, worauf alle anderen Stationen ihre Übertragung verschieben. Danach kann die eigentliche Signalübertragung beginnen.

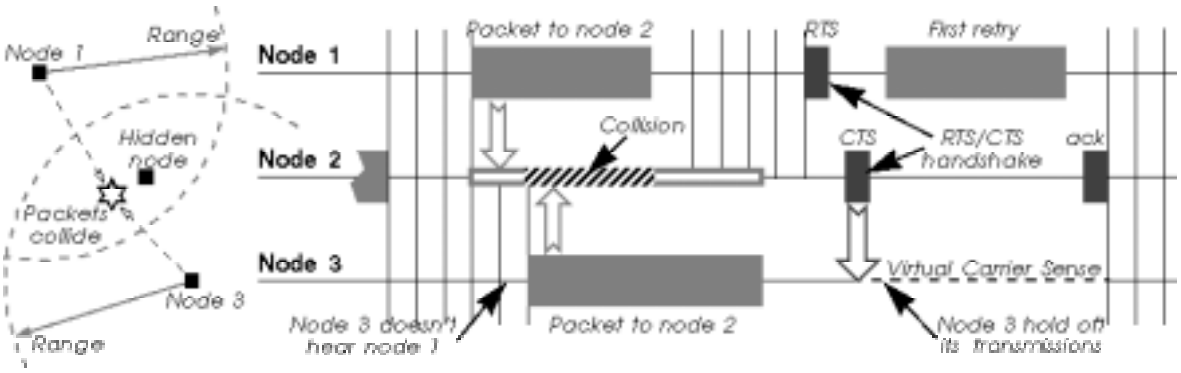

Abbildung 2: Zeitlicher Ablauf von RTS/CTS Packeten

5 Leistungsmerkmale

Sowohl DSSS als auch FHSS kommen auf Übertragungsraten von 1 oder 2 Mbit pro Sekunde. Die Datenraten im Funk-LAN sind also noch weit niedriger als im Ethernet. Allerdings gibt es schon einem Standard, der Datenraten von über 10 Mbit/s ermöglicht. Damit sind dann beispielsweise erste drahtlose Video-Conferencing-Anwendungen möglich. Die Reichweite der Funk-LAN-Komponenten unterscheidet sich je nach Typ, Bauweise und Sendeleistung. Kleingeräte besitzen manchmal nur eine Reichweite von weniger als einem Meter, größere Systeme können bis zu 20 Kilometer überbrücken. Dabei hängt die Reichweite einer Funk-LAN-Verbindung sehr stark von den örtlichen Gegebenheiten ab. So wird das Funksignal beispielsweise durch Häuserecken und Wände gedämpft oder reflektiert, wie bereits in Abschnitt 3 beschrieben.

6 Probleme

Bedenken bestehen hinsichtlich der Sicherheit von Daten im Funk-LAN. Über Verschlüsselung läßt sich vermeiden, dass Unberechtigte das Funknetz einfach abhören. Das gewährleistet zwar Schutz, dennoch ist das Mithören im Funk-LAN wesentlich unkomplizierter möglich als beispielsweise in einem einem normalen Netz. Spezielle Verschlüsselungsfunktionen für Funk-LANs gibt es nicht, da diese auf höherer OSI-Ebene ansetzen.

Probleme mit der Interoperabilität einzelner Funk-LANs, beispielsweise beim Betrieb zweier getrennter Funk-LANs in einem Gebäude (wenn mehrere Unternehmen sich ein Bürogebäude teilen und mindestens zwei ein Funk-LAN einsetzen) sind mittlerweile gelöst. So trennt eine Kennung (ähnlich den MAC-Adressen bei Netzwerkkarten) den Datenverkehr zwischen den beiden Netzen, womit sämtliche Komponenten und die Funknetze voneinander getrennt bleiben.

Weitere Bedenken kommen aus einer anderen Ecke: Wie schon bei Mobiltelefonen befürchtet man durch Funk-LANs weiteren Elektrosmog mit gesundheitsschädlichen Auswirkungen. Die einzigen belegten biologischen Nebenwirkungen elektromagnetischer Wellen sind thermischer Art. So führen elektromagnetische Wellen, die biologisches Material durchdringen, zu einer schnelleren Bewegung der Molekülatome im Körper, was zu einer geringfügigen Erwärmung führt. Auf diesem Prinzip basiert der Mikrowellenherd. Allerdings ist die Erwärmung, die ein Funk-LAN in Lebewesen erzeugt, so gering, daß man sie nicht spürt. Andere biologische Nebenwirkungen des Elektrosmogs sind bislang nicht belegt.

7 Rechtlich Aspekte

Basis für den Einsatz des Funk-LANs als Campus-Netz ist die Verfügung 122/1997 des Bundesministeriums für Post und Telekommunikation (BMPT) vom 21. Mai 1997. Danach sind Funkverbindungen, die über ein Privatgelände hinaus in öffentlichen Raum reichen, an sich genehmigungspflichtig. Die besagte Verordnung erlaubt jedoch generell

die grundstücksübergreifende Anwendung von Funk-LANs für den internen Unternehmensbedarf. Somit steht dem unbürokratischen Einsatz von Funk-LANs gemäß IEEE 802.11 zumindest in Deutschland nichts im Wege.

8 Zusammenfassung

Der IEEE 802.11 Standart beschreibt eine flexsible drahtlose Erweiterung für lokale Netze, welche beispielsweise in Firmen und Universitäten vermehrt eingesetzt wird. In dieser Ausarbeitung wurden die technischen Grundlagen dises Standarts beschrieben. Mit der Spreizspektrumstechnik wurde ein wichtiges Übertragungsverfahren erläutert welches zur Reduktion von Multi-Path-Effekten benutzt wird. Auf der Ebene der Medienzugriffsschicht definiert der Standard mit CSMA/CA ein dem Ethernet ähnliches Verfahren, welches jedoch einen Algorithmus zur Kollionsvermeidung nutzt. Weiterentwicklungen des Standards bieten einerseits höhere Übertragungsraten, andererseits befinden sich verbesserte Mediezugriffverfahren in der Entwicklung durch die 802.11-Arbeitsgruppe der IEEE.

Literatur

[Baer00] Klaus-Dieter Baer. Angefunkt. *Linux Magazin* (8), 2000, S. 98–103.

[Baer01] Klaus-Dieter Baer. Lan-Funker. *Linux Magazin* (2), 2001, S. 106–109.

[GlHK01] Garry Glendown, Heinz W. Huber und Olaf Krause. Äthernetz. *iX* (1), 2001, S. 56–57.

[Schi00] J. Schiller. *Mobilkommunikation*. Net.com. Addison–Wesley. 2000.

[Wunn99] Lukas Wunner. Funkmodems im 2.4GHz ISM-Band, Frequently Asked Questions. http://www.wunner.de/people/lukas/funk/, Juli 1999.

Abbildungsverzeichnis

QoS in IEEE 802.11-Netzen

Matthias Unholzer

Kurzfassung

Der IEEE 802.11-Standard ist ein robuster und weitverbreiteter Standard für drahtlose lokale Netzwerke. Drei verschiedene Übertragungsvarianten der physikalischen Schicht, die wiederum beliebig mit zwei verschiedenen Funktionsweisen der MAC-Schicht kombiniert werden können, sind vorgesehen. Die Distributed Coordination Function der MAC-Schicht arbeitet im reinen Wettbewerbsverfahren und bietet somit keinerlei harte Garantien für QoS. Wenn man zur Unterstützung eines zentralen Zugangspunktes (Infrastrukturnetz) die Point Coordination Function der MAC-Schicht einsetzt, wird die Koordination des Medienzugriffes zuverlässiger, wobei auch hier insgesamt nur „bestmögliche" Dienstgüte erwartet werden kann, weil der Overhead zunimmt und der zentrale Zugangspunkt ebenfalls dem Wettbewerb ausgesetzt ist, bevor er die Medienkontrolle übernehmen kann. Deshalb kommt es bei beiden MAC-Versionen darauf an, verschiedene variable Parameter zu analysieren und so gut wie möglich zu verbessern. Es geht hierbei in erster Linie um Zeitintervalle, welche die Wartezeiten bis zum nächsten Zugriffsversuch eines Teilnehmers vorgeben. Auch Strategien, mit welcher Wahrscheinlichkeit innerhalb der Wettbewerbsfenster auf das Medium zugegriffen wird, sind von Bedeutung. Energiesparmaßnahmen, die der Dienstgüte entgegenwirken, müssen optimiert werden. Die Möglichkeit, die Bandbreite eines drahtlosen Netzes in Abhängigkeit verschiedener Anwendungen zu verteilen, wird ergänzend angesprochen, ist im aktuellen 802.11-Standard jedoch noch nicht verankert.

1 Einleitung

Drahtlose lokale Netze (Wireless Local Networks, WLANs) bieten gegenüber drahtgebundenen lokalen Netzen (Local Area Networks, LANs) einige Vorteile. Im wesentlichen sind es die Bewegungsfreiheit innerhalb des Empfangsbereiches, die Möglichkeit von Direktverbindungen zwischen mobilen Endgeräten (Stations, STAs) in Ad-hoc-Netzen sowie der Entfall aufwendiger und unpraktischer Verkabelungen.

In dieser Ausarbeitung wird die Dienstgüte eines WLAN nach dem IEEE 802.11-Standard von 1997 behandelt. Es werden verschiedene Verfahren untersucht, die es erlauben, die Dienstgüte zu verändern und bestimmten Anforderungen anzupassen. Diese Anforderungen entsprechen im wesentlichen denen, die auch an LANs gestellt werden wie z.B. möglichst hohe Übertragungsrate, niedrige Bitfehlerrate und geringe Zugriffsverzögerung. Hinzu kommen dann spezifische Anforderungen an drahtlose Systeme wie z.B. fehlerfreier Wechsel von einer Funkzelle (Basic Service Set, BSS) zur

anderen innerhalb eines infrastrukturbasierten Netzes (Roaming) oder der Umgang mit versteckten Endgeräten bei Ad-hoc-Netzen. Bei den zuerst genannten Anforderungen (Übertragungsrate, Bitfehlerrate, Zugriffsverzögerung), die sich auf den eigentlichen Übertragungsvorgang beziehen, können Veränderungen nur innerhalb des Rahmens vorgenommen werden, den ein Funksystem grundsätzlich erlaubt.

Die Funkeinheiten des 802.11-Standards lassen maximal 2 Mbit/s zu im Gegensatz zu beispielsweise 100 Mbit/s eines Fast-Ethernet. Vorschläge und Weiterentwicklungen des Standards, oft hervorgehend aus proprietären Lösungen, existieren bereits. Als Beispiele sind hierfür die IEEE Working Groups 802.11a, 802.11b und 802.11e zu nennen. 802.11a arbeitet mit einer höheren Trägerfrequenz (5 Ghz anstatt 2,4 GHz), die eine Übertragungsrate von bis zu 54 Mbit/s ermöglicht. Diese Weiterentwicklung läßt sich mit dem 802.11-Standard von 1997 kombinieren, da die Medienzugriffssteuerung kompatibel ist. Es werden jedoch andere Funkadapter benötigt. Die Weiterentwicklung 802.11b von 1999 kann über das bisherige 2,4 GHz-Band höhere Datenraten übertragen. Dies wird durch geänderte Einträge in den Steuerinformationen der physikalischen Übertragungsrahmen möglich. Auch hier besteht Kompatibilität zum 802.11-Standard von 1997. Die Arbeitsgruppe IEEE 802.11e beschäftigt sich mit der Erweiterung des Standards im Hinblick auf adaptive Anpassung von QoS (siehe Abschnitt 2.5).

Der Schwerpunkt der Dienstgüteanforderungen an ein 802.11-Netz (von 1997) muß auf die möglichst effektive Nutzung der theoretisch möglichen Übertragungsrate von maximal 2 Mbit/s ausgerichtet werden. Genauso gilt das für die höheren Übertragungsraten von 802.11a und 802.11b, wobei in dieser Ausarbeitung vom Original-Standard (1997) ausgegangen wird, da er die technische Grundlage aller 802.11-Netze darstellt. Grundsätzliche Problematik ist hierbei das geteilte Medium, d.h. alle mobilen Endgeräte und ein evtl. vorhandener zentraler Zugangspunkt greifen auf ein Übertragungsmedium zu. Die Funktionsweise der Zugriffsverfahren (Im wesentlichen sind das Wettbewerbsverfahren im 802.11-Standard.) und die Modulationsvarianten der Funkeinheiten entscheiden dabei über die zu erwartende Dienstgüte. Es liegt keine gesicherte Reservierung von Ressourcen vor, weshalb der 802.11-Standard in die Klasse der „bestmöglichen Dienstgüte" einzuordnen ist. Bevor jedoch in Abschnitt 2 auf konkrete Verfahren zur Beeinflussung der Dienstgüte eingegangen wird, müssen zunächst noch Fragen zur technischen Überprüfbarkeit von QoS gestellt werden: Wie lauten die Dienstgüteparameter bzw. -eigenschaften und mit welchen Methoden können sie ermittelt werden? Zum Verständnis der Lösungsvorschläge muß zudem ein Überblick über den 802.11-Standard geschaffen werden.

Der 802.11-Standard erstreckt sich über Schicht 1 (physikalische Schicht, PHY) und Schicht 2a (Medienzugriffssteuerung, MAC) der Sicherungsschicht des OSI-Referenz-Modells. Oberhalb dieser beiden Schichten bietet 802.11 die gleiche Schnittstelle wie die anderen 802.x-Familienmitglieder (z.B. Ethernet, 802.3). Eine Anwendung, die auf einer STA betrieben wird, erkennt also nicht, über welches Medium (Funk oder Kabel) sie ihre Daten übermittelt. Nur für den Anwender werden die oben geschilderten Einschränkungen spürbar. Diese gilt es, möglichst gering zu halten.

Die physikalische Schicht von 802.11 sieht drei Übertragungstechniken vor, zwei Funk- und eine Infrarotvariante:

- FHSS (Frequency Hopping Spread Spectrum):

  - Bandspreizung durch Wechsel von 79 verschiedenen 1 MHz-Bändern (Kanäle) im lizenzfreien 2,4 GHz-Band (In Japan sind es momentan nur 23 Kanäle.)
  - 2-stufige GFSK-Modulation (1 Bit wird auf 2 verschiedene Frequenzen codiert): ermöglicht Bandbreite von 1 Mbit/s
  - 4-stufige GFSK-Modulation (2 Bits werden auf 4 verschiedene Frequenzen codiert): ermöglicht Bandbreite von 2 Mbit/s

- DSSS (Direct Sequence Spread Spectrum):
  - arbeitet ebenfalls im lizenzfreien 2,4 GHz-Band
  - XOR-Verknüpfung auf die Nutzdaten mit einem Barker-Code (11 Chips)
  - DBPSK-Modulation: ermöglicht 1 Mbit/s Bandbreite
  - DQPSK-Modulation: ermöglicht 2 Mbit/s Bandbreite
  - OFDM-Modulation: ermöglicht bis zu 54 Mbit/s Bandbreite (nur 802.11a)

- Infrarot (Diffuse Infrared, DFIR):
  - arbeitet im Basisband
  - diffuses Licht mit 850-950nm, etwa 10m Reichweite
  - Bandbreite von 1 bzw. 2 Mbit/s möglich

Die bereits angesprochene Weiterentwicklung 802.11b bietet in FHSS-Systemen 3 Mbit/s an und in DSSS-Systemen immerhin 10 Mbit/s.

Der Dienstzugangspunkt (Service Access Point, SAP) der PHY-Schicht bietet, wie oben geschildert, der MAC-Schicht je nach Übertragungstechnik und Modulationsverfahren verschieden Bandbreiten an. Hierfür ist, genau betrachtet, die Physical Layer Convergence Protocol (PLCP-)Unterschicht zuständig. Desweiteren wird ein Clear Channel Assessement Signal (CCA) übertragen, welches der MAC-Schicht signalisiert, ob das Medium frei oder belegt ist. Die Physical Medium Independent (PMD-)Unterschicht moduliert die Daten und kodiert/dekodiert die Signale.

Die MAC-Schicht hat im wesentlichen die Aufgabe, den Zugriff auf das von allen STAs (und evtl. einem AP) gemeinsam genutzte Medium zu koordinieren. Im 802.11-Standard werden zwei Datendienste unterstützt, der asynchrone und der zeitbeschränkte. Der zeitbeschränkte Datendienst kann nur in einem infrastrukturbasierten 802.11-Netz realisiert werden, da in diesem Fall ein zentraler Zugangspunkt die Steuerung des Medienzugriffs durch die PCF (Point Coordination Function)übernehmen muß.

Der zentrale Zugangspunkt (Access Point, AP) ist am Distribution System (DS) angeschlossen, das die Verbindung zum gewünschten Kommunikationspartner herstellt. Dieser kann sich im gleichen BSS des rufenden Teilnehmers befinden oder auch in einem beliebigen anderen. Verbindungen zu anderen (evtl. drahtgebundenen) Netzen werden vom DS über das Portal hergestellt (siehe Abbildung 1). Wird nur der asynchrone Datendienst gefordert, so kann ein 802.11-Netz auch im Ad-hoc-Modus betrieben werden, d.h. es werden Direktverbindungen zwischen den STAs aufgebaut (siehe Abbildung 2). Hier wird die Steuerung von der Distributed Coordination Function (DCF)

übernommen, die natürlich auch ein Infrastrukturnetz mit asynchronem Datendienst unterstützt. Die Abbildungen 1 und 2 orientieren sich an [Schi00].

Abbildung 3 zeigt zusammenfassend den Zusammenhang und die Kombinationsmöglichkeiten der verschiedenen Ausführungen der beiden relevanten Schichten.

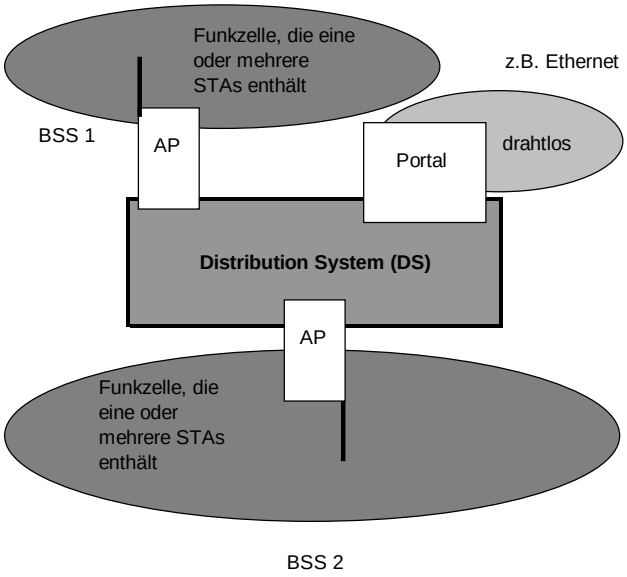

Abbildung 1: Systemarchitektur eines infrastrukturbasierten 802.11-Netzes

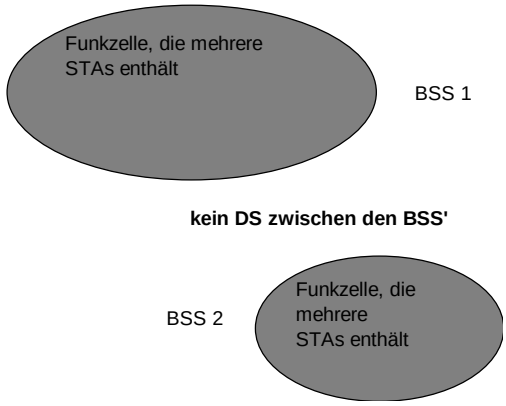

Abbildung 2: Systemarchitektur eines Ad-hoc-802.11-Netzes

Zur Beurteilung der Übertragungseigenschaften ist der sogenannte Overhead von Bedeutung. Hierauf haben die Übertragungsrahmen beider Schichten Einfluß. Die Übertragungsrahmen der beiden Varianten der PHY-Schicht bestehen im wesentlichen aus drei Feldern, nämlich der PLCP-Präambel, dem PLCP-Paketkopf und dem von der MAC-Schicht übergebenen MAC-Rahmen. Dieser wiederum besteht aus einem Paketkopf, den eigentlichen Nutzdaten und einem Prüfsummenfeld. Die Paketköpfe, die PLCP-Präambel und das Prüfsummenfeld addiert und verglichen mit der Länge der eigentlichen Nutzdaten ergeben den Overhead. Dieser sollte möglichst klein sein, um die Übertragungskapazität der Funkeinheiten hauptsächlich den Nutzdaten zukommen zu lassen. Später werden wir sehen, daß weitere Steuerinformationen der MAC-Schicht den Overhead noch weiter vergrößern können.

Benny Bing [Bing99] prüft die Leistung des Funkübertragungsweges mit zwei verschiedenen 802.11-APs mit Ethernet-Schnittstelle, dem WavePOINT für DSSS und dem

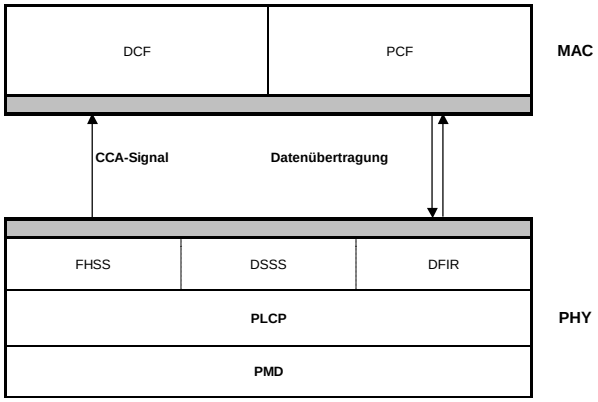

Abbildung 3: 802.11-Protokollarchitektur

Spectrum24 für FHSS. Hierüber werden Ethernet-Rahmen unterschiedlicher Länge übertragen. Da bei oben genanntem Versuchsaufbau nur zwei Funkeinheiten miteinander kommunizieren und deshalb keine realistische Wettbewerbssituation zwischen drei und mehreren STAs aufkommt, wird hier die Leistungsfähigkeit der PHY-Schicht zunächst isoliert betrachtet. Der Analyse der MAC-Schicht wird dann mit objektorientierten programmierbaren Simulationswerkzeugen Rechnung getragen. Entscheidende Parameter in einem realistischen 802.11-Netz wie die Anzahl der am Netz beteiligten STAs, die Größe der Funkzellen, das Auftreten versteckter Endgeräte bei Ad-hoc-Netzen, die Belastung des Netzwerks je nach Anzahl und Größe der Datenrahmen sowie der Wechsel einer STA von einem BSS zum anderen (Roaming) können dann zur Beurteilung der QoS-Parameter herangezogen werden. Ein solches Simulationswerkzeug (CSIM18)wird beispielsweise in [Eber00] beschrieben. Es handelt sich hierbei um eine Programmbibliothek für $C/C++$.

2 Verfahren zur Beeinflussung der Dienstgüte in einem 802.11-Netz

2.1 Die physikalische Übertragung

Die folgenden Ausführungen orientieren sich am 802.11-Standard von 1997, sind jedoch auf die höheren Datenraten der Weiterentwicklungen übertragbar. Der im vorigen Abschnitt angesprochene Overhead spielt die entscheidende Rolle bei der theoretisch möglichen Ausnutzung der von der PHY-Schicht zur Verfügung gestellten Übertragungsrate. Die MAC-Schicht übergibt MAC-Datenrahmen nach Abwarten des DIFS (DCF Interframe Space) an die PHY-Schicht. Dieser wird in den PHY-Rahmen eingebettet und an die Empfangseinheit übermittelt. Handelt es sich um eine Unicast-Übertragung, so antwortet der Empfänger nach Abwarten des SIFS (Short IFS) mit einem Bestätigungs-MAC-Rahmen (ACK), der wiederum in einen PHY-Rahmen eingebettet und zurückgesendet wird. Die bei diesem Vorgang auftretenden Rahmenformate und -längen sind in Abbildung 4 gezeigt, wobei die PLCP-Präambel und der PLCP-Paketkopf grundsätzlich nur mit 1 Mbit/s übertragen werden.

Die Länge der verschiedenen IFS, deren Notwendigkeit später in Abschnitt 2.2 erklärt wird, sind im 802.11-Standard wie folgt definiert:

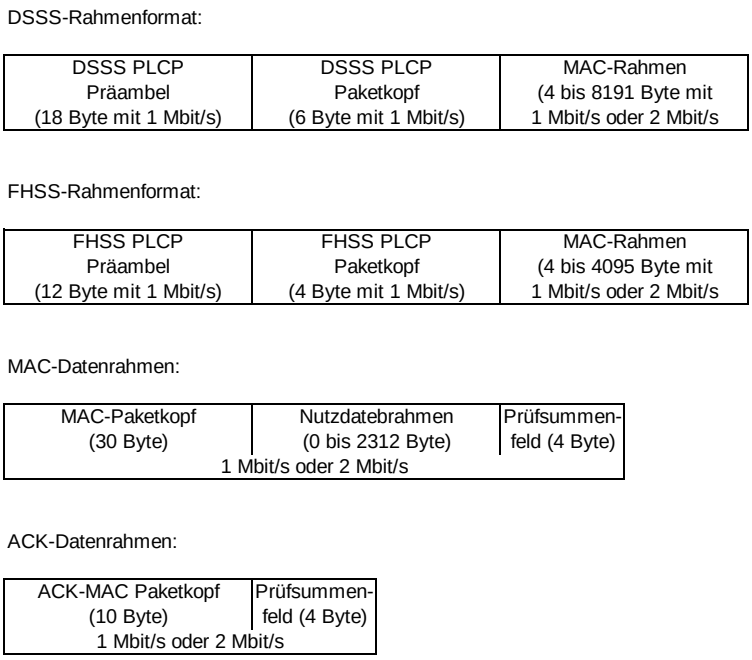

Abbildung 4: 802.11-Rahmenformate

- DSSS:
  - $SIFS = 10\mu s$
  - $DIFS = 50\mu s$
- FHSS:
  - $SIFS = 28\mu s$
  - $DIFS = 128\mu s$

Den Übertragungsvorgang und die dabei entstehenden Overheads zeigen die Abbildungen 5 und 6. Die Übertragungszeiten der PLCP-Präambel und des PLCP-Paketkopfes müssen doppelt angerechnet werden, da zwei PHY-Rahmen gesendet werden, nämlich für die Datenübertragung und die darauf folgende Bestätigung. In Abschnitt 2.2.1 wird ein Wettbewerbsverfahren beschrieben, das hier bereits in Erscheinung tritt, aber wegen des fehlenden Wettbewerbs keine Verbesserung erzielt. Es wird in diesem Kapitel deshalb noch vernachlässigt.

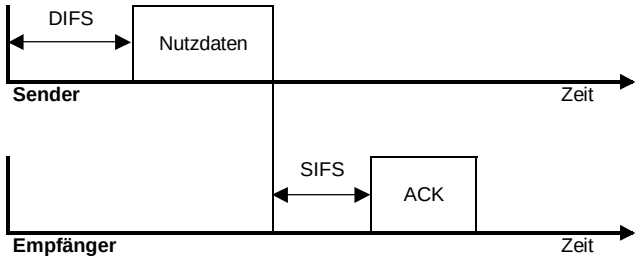

Abbildung 5: Ablauf einer Unicast-Übertragung mit ACK

Überträgt man nun mit den beiden Übertragungsverfahren jeweils einen Ethernet-Rahmen mit maximaler Länge (1518 Byte), so ergibt sich durch die Overheads folgende

Overheads	DSSS mit WavePOINT und optionalen 2 Mbit/s	FHSS mit Spectrum24 1 Mbit/s
DIFS	50 µs	128 µs
SIFS	10 µs	28 µs
PLCP Präambel und Paketkopf	24 * 8 Bits / 1 Mbit/s = 192 µs (* 2)	16 * 8 Bits / 1 Mbit/s = 128 µs (* 2)
MAC Paketkopf und Prüfsummenfeld	34 * 8 Bits / **2** Mbit/s = 136 µs	34 * 8 Bits / 1 Mbit/s = 272 µs
ACK-MAC Paketkopf und Prüfsummenfeld	14 * 8 Bits / **2** Mbit/s = 56 µs	14 * 8 Bits / 1 Mbit/s = 112 µs
Overhead gesamt	636 µs	796 µs

Abbildung 6: Overheads

effektive Nutzung der Übertragungsrate: 90.5% bei DSSS, da $\frac{1.810Mbit/s}{2Mbit/s} = 0.905$ mit $1.810Mbit/s = \frac{1518 \cdot 8Bits}{6072\mu s + 636\mu s}$ wobei $6072\mu s$ die theoretisch mögliche Übertragungszeit der eigentlichen Nutzdaten bei 2 Mbit/s angibt. Die hinzukommenden $636\mu s$ stehen für den auftretenden Overhead bei DSSS. Bei FHSS ergibt sich bei gleicher Berechnung (mit den entsprechenden Werten eingesetzt) eine effektive Nutzung der Übertragungskapazität von 94.1%.

Dieses Ergebnis zeigt, daß das FHSS-Verfahren trotz größerem Overhead eine bessere effektive Nutzung der Übertragungsrate aufweist als das DSSS-Verfahren. Der Grund hierfür ist, daß trotz der höheren Übertragungsrate der MAC-Rahmen von 2 Mbit/s (DSSS) die PLCP Paketköpfe und Präambeln nur mit 1 Mbit/s übertragen werden. Das Verhältnis zwischen der Ausnutzung der Gesamtübertragungsrate, mit denen auch die Nutzdaten übertragen werden, und den Steuerdaten wird dadurch noch ungünstiger. Bei beiden Übertragungsverfahren gilt jedoch erwartungsgemäß, daß der negative Einfluß der Overheads um so geringer wird, je größer die Rahmen der zu übertragenden Nutzdaten sind.

Die in Abschnitt 1 kurz angesprochene von Benny Bing [Bing99] verwendete Versuchsanordnung für einen drahtlosen Übertragungsweg zwischen zwei APs bestätigt den theoretisch ermittelten negativen Einfluß der Overheads. Bei beiden Übertragungsverfahren steigt die Übertragungsrate des drahtlosen Übertragungsweges mit zunehmender Größe der Nutzdatenrahmem, die vom Ethernet übergeben werden. Beim DSSS-Verfahren (realisiert mit WavePOINT) werden maximal 1.67 Mbit/s bei der maximalen Größe eines Ethernet-Rahmens von 1512 Byte erreicht. Das FHSS-Verfahren zeigt eine Besonderheit in Verbindung mit Spectrum24. Auch hier steigt die Übertragungsrate mit zunehmender Ethernet-Rahmengröße und erreicht ihr Maximum bei 0.58 Mbit/s (512 Byte) und bleibt dann konstant bei diesem Wert. Spectrum24 fragmentiert größere Ethernet-Datenrahmen fix auf 512 Byte. Dies scheint zunächst ein Nachteil zu sein, da hierdurch die Leistungsfähigkeit von FHSS nur noch mit 58% ausgenutzt wird. Die Fragmentierung bietet jedoch einen anderen Vorteil, nämlich eine kleinere Rahmenfehlerrate. Bei kleineren Rahmen gehen beim Rahmenverlust weniger Bits auf einmal verloren als bei größeren Rahmen.

2.2 Der Medienzugriff

Die MAC-Schicht ist für Roaming, Authentifizierung der STAs, Energiesparmaßnahmen und vorrangig für den Medienzugriff zuständig. Das Grundprinzip der 802.11-

Medienzugriffssteuerung ist ein Wettbewerbsverfahren, das nach bestimmten Wartezeiten (Prioritäten), den Interframe Spaces (IFS), anläuft. Der Sender (ein AP oder eine STA), der den Wettbewerb für sich entscheidet, kann dann einen Rahmen (Daten- oder Steuerrahmen) über das Medium senden. Die IFS treten in unterschiedlichen Längen auf:

- Distributed Coordination Function IFS (DIFS):
  - längste Wartezeit vor dem Wettbewerbsverfahren und damit niedrigste Priorität
  - Verwendung beim asynchronen Datendienst, der auch als Distributed Coordination Function (DCF) bezeichnet wird
- Point Coordination Function IFS (PIFS):
  - Verwendung nur im infrastrukturbasierten Netz mit AP vorgesehen
  - liegt zwischen den Wartezeiten von DIFS und SIFS
  - unterstützt zeitbeschränkten Datendienst, bei dem der AP nur den PIFS abwarten muß, bevor es auf das Medium zugreifen kann (bezeichnet als Point Coordination Function)
- Short IFS (SIFS):
  - kürzeste Wartezeit und damit höchste Priorität
  - Zeit, die Steuernachrichten abwarten müssen

Das sich an den DIFS anschließende Wettbewerbsverfahren wird dann entscheiden, wer den Zugriff letztlich erhält. Graphisch sieht das dann so aus:

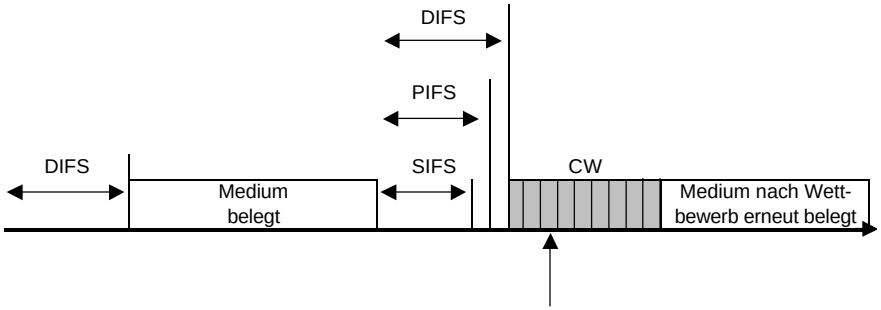

Abbildung 7: Medienzugriffsverfahren des 802.11-Standard

In den nächsten Abschnitten wird nun auf diese Wettbewerbsverfahren eingegangen, die einen entscheidenden Einfluß auf wesentliche QoS-Parameter haben.

2.2.1 Asynchroner Datendienst mit CSMA/CA-Verfahren und Backoff-Algorithmus

Das CSMA/CA-Verfahren (Carrier Sense Multiple Access with Collision Avoidance) mit Backoff-Algorithmus stellt den Standard-Zugriffsmechanismus in einem 802.11-Netz dar. Eine STA hört das Medium ab und beginnt, wenn das Medium durch das

CCA-Signal der PHY-Schicht als frei erkannt wird, nach Abwarten des DIFS mit dem Senden. In diesem Fall werden dann die in Abschnitt 2.1 beschriebenen Übertragungseigenschaften angeboten. Treten jedoch mehrere STAs in Erscheinung, kommt es zur Wettbewerbssituation. Jede STA wählt nach Abwarten des DIFS eine zufällige Backoff-Zeit (Anzahl von Zeitschlitzen des Wettbewerbsfensters) und greift nach dieser Zeit auf das Medium zu, wenn es zu diesem Zeitpunkt als frei erkannt wird. Ist es nicht frei, dann hat bereits eine andere STA erfolgreich auf das Medium zugegriffen, und der Wettbewerb geht nach dem nächsten DIFS erneut von vorne los. Dieses Verfahren (auch DCF genannt) unterstützt sowohl infrastrukturbasierte Netze als auch Ad-hoc-Netze.

Die DCF besitzt keinerlei QoS-Garantien und scheint ziemlich unfair zu sein, weil es keine Rolle spielt, wie lange eine STA schon Zugriffsversuche unternimmt. Deshalb kommt nun der Backoff-Algorithmus zum Tragen, der einer schon lange wartenden STA eine höhere statistische Wahrscheinlichkeit für erfolgreichen Zugriff bieten soll. Die Anzahl der zufällig gewählten Zeitschlitze werden in der Wartephase pro vergangenem Zeitschlitz heruntergezählt. Erreicht dieser Zähler den Wert Null, so greift die STA auf das Medium zu. Wird das Medium schon vorher von einer anderen STA belegt, so hält der Zähler vorher an und beginnt mit dem bereits heruntergezählten Wert die nächste Wettbewerbsphase. Dies soll eine gewisse Gerechtigkeit schaffen, nach langer Wartezeit früher erneut zugreifen zu dürfen.

Bei wachsender Anzahl konkurrierender STAs steigt jedoch die Wahrscheinlichkeit dafür, daß die gewählten Zeitschlitze sehr eng beieinander liegen oder sich sogar überdecken. Das Ergebnis sind häufige Kollisionen. Aus diesem Grund wird bei jeder Kollision, die auf ein stark belastetes Medium hindeutet, die Größe des Wettbewerbsfensters verdoppelt, damit die Wahrscheinlichkeit, eng beieinander liegender Zeitschlitze für den Zugriff zu wählen, abnimmt. So werden auch Kollisionen unwahrscheinlicher. Allerdings werden jetzt die Zugriffsverzögerungen vergrößert, weil ja nun durch das größere Wettbewerbsfenster die Wartezeiten im Durchschnitt ansteigen. Dies führt zur Notwendigkeit, noch andere Methoden heranzuziehen, die im Standard zusätzlich verankert werden müßten.

Hierzu muß man die Wahrscheinlichkeiten betrachten, mit denen eine STA die Zeitschlitze für den Zugriff zufällig auswählt. Im Initialisierungszustand der Wettbewerbsphase ist für jeden Zeitschlitz innerhalb eines Wettbewerbsfensters (Contention Window, CW) die Wahrscheinlichkeit ausgewählt zu werden, gleich groß. Der ausgewählte Zeitschlitz sei mit CW_1 bezeichnet ($0 \leq CW_1 \leq CW$), wobei CW für die Anzahl aller Zeitschlitze eines Wettbewerbsfensters steht. Dies zeigt folgende Abbildung:

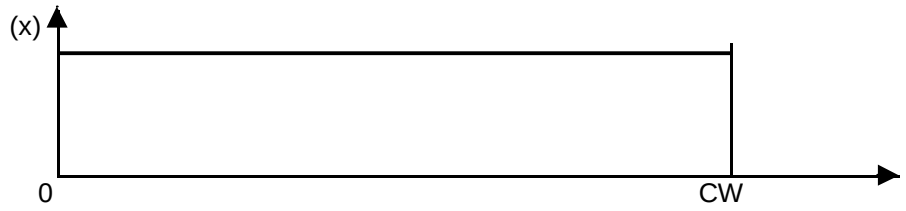

Abbildung 8: Gleichverteilung der Zugriffswahrscheinlichkeit

Eine STA, die keinen Zugriff bekommt, kehrt nun, wie oben beschrieben, mit einer reduzierten Wartezeit in die Wettbewerbsphase zurück. Der neue Zeitschlitz (CW_2) für den Zugriff liegt nun im Bereich von $0 \leq CW_2 < CW_1$. Schlägt der Versuch erneut

fehl, so bleibt der Zähler für den nächsten Zugriffsversuch wieder bei einem neuen Zeitschlitz (CW_3) stehen mit der Eigenschaft $0 \leq CW_3 < CW_2$. Dieser Vorgang wiederholt sich immer wieder bis zum erfolgreichen Zugriff. STAs, die innerhalb dieser Zeit neu in die Wettbewerbsphase eintreten, wählen nun wieder aus der gesamten Länge des CW ihren Zeitschlitz für den Zugriff zufällig aus. Auch für diese STAs beginnt dann der eben beschriebene Prozeß: Die Zugriffszeiten verschieben sich in Richtung Zeitschlitz 0. Sind viele STAs im Wettbewerb und kommen ständig neue hinzu, so vergrößert sich nach [WSFW97] die Wahrscheinlichkeit für alle STAs, einen Zeitschlitz im Anfangsbereich des CW zu wählen mehr und mehr. Also wird es hier, da dann wieder viele ausgewählte Zeitschlitze für den Zugriff dicht beieinander liegen, vermehrt zu Kollisionen kommen. Aus der ursprünglichen Gleichverteilung entwickelt sich eine rechtsschiefe stufenförmige Verteilung:

Abbildung 9: Zugriffswahrscheinlichkeiten nach längerer Wettbewerbsphase

Nun muß eine Methode gefunden werden, die Gleichverteilung des Initialisierungszustandes annähernd aufrecht zu erhalten. Hier sollen zwei Lösungsmöglichkeiten gezeigt werden:

Als erste Möglichkeit kann die im Initialisierungszustand anzutreffende Gleichverteilung $f(x) = \frac{1}{CW}$ derart modifiziert werden, daß bei großer Anzahl von Stationen die Wahrscheinlichkeit dafür, daß eine neu hinzutretende Station einen Zeitschlitz im oberen Bereich des CW wählt, zunimmt. Die Konzentration im unteren Bereich soll dadurch abgeschwächt werden. Nach [WSFW97] könnte die Verteilung dann so aussehen: $f(x) = \frac{a+1}{CW^{a+1}} \cdot x^a$ $(a \geq 0; 1 \leq x \leq CW)$ mit $x = CW_x$ und $a = \textit{Anzahl der Stationen}$ ($a = 0$ ergibt wieder die alte Gleichverteilung). Ist a groß ($a = 10$ wurde bei [WSFW97] als günstig ermittelt), so steigt die Wahrscheinlichkeit, einen Zeitschlitz im oberen Bereich des CW zu wählen, exponentiell an und sieht dann in etwa so aus:

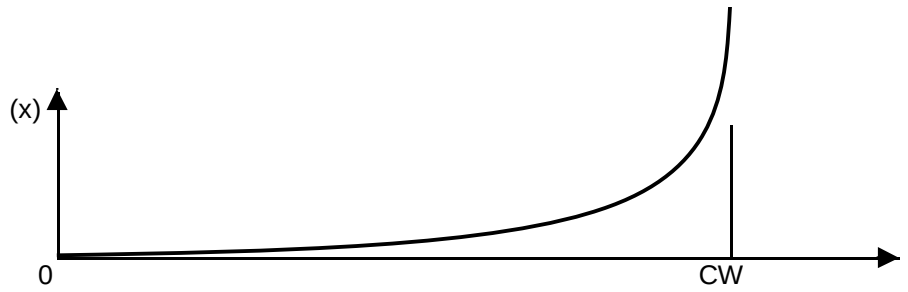

Abbildung 10: Modifizierte Verteilung der Zugriffswahrscheinlichkeiten

Eine weitere Möglichkeit, die Auswahl der Zeitschlitze neu hinzukommender STAs in den oberen Bereich des CW zu verschieben, ist die Information aller STAs hierüber, welcher Zeitschlitz CW^* den letzten Zugriff erhielt. Anschließend müssen sich die erneut

in den Wettbewerb tretenden STAs im Bereich von 0 bis $CW - CW^*$ konkurrieren. Neu hinzutretenden Stationen wird der Bereich $CW - CW^*$ zugeordnet (also im oberen Bereich des CW). Befindet sich CW^* nun im unteren Bereich des CW, so könnte man annehmen, daß bis zu diesem Zeitpunkt der Wettbewerb schon lange andauert, also eine große Last vorliegt. Um so weiter nach rechts werden dann die Zeitschlitze der neu hinzutretende STAs verschoben. Man kann dieses Verfahren demzufolge auch als lastabhängig bezeichnen.

In [WSFW97] wird gezeigt, daß beide Verfahren die ursprüngliche Gleichverteilung annähernd aufrecht erhalten. Die anschließende Netzwerksimulation mit acht Stationen bestätigt den Erfolg: Bei möglichst hohem Durchsatz und geringen Zugriffsverzögerungen in Abhängigkeit von der Netzwerklast übertreffen beide Verfahren den ursprünglichen Backoff-Algorithmus, wobei das lastorientierte Verfahren das beste Ergebnis erzielt.

2.2.2 Asynchroner Datendienst im Ad-hoc-Modus mit RTS/CTS-Mechanismus

Versteckte Endgeräte stellen in WLANs (im Ad-hoc-Modus) ein nicht zu umgehendes Problem dar. Es seien drei STAs (A, B und C), wie in folgender Abbildung gezeigt, gegeben:

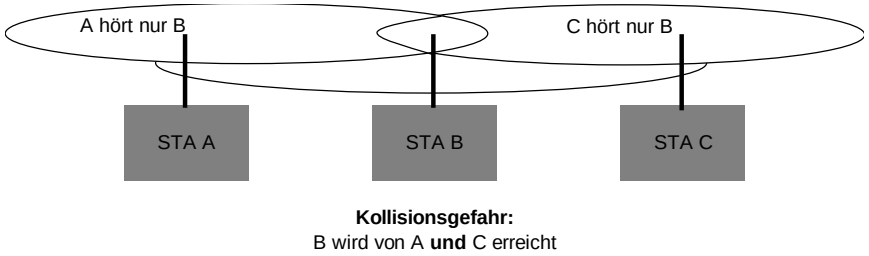

Abbildung 11: Versteckte Endgeräte

STA A will an STA B senden und erkennt nicht, daß STA C auch an STA B senden will, da deren Reichweite nur bis STA B reicht. STA C ist also für STA A versteckt. Da aus diesem Grund STA A trotzdem zu senden beginnt, kollidieren die Signale von STA A und STA C bei STA B. Der Sendevorgang wird gestört.

Dieser Problematik soll mit der Aktivierung des RTS/CTS-Mechanismus begegnet werden. Ist es einer STA gelungen nach Ablauf der im vorigen Abschnitt beschriebenen Zugriffsverfahren auf das Medium zuzugreifen, sendet sie zunächst einen (Request-To-Send) RTS-Steuerrahmen. Hierin wird die Adresse des gewünschten Teilnehmers und die Dauer der geplanten Datenübertragung übermittelt. Alle Stationen, die den RTS-Steuerrahmen empfangen, setzen ihren Net Allocation Vector (NAV) auf die Länge der zu erwartenden Datenübertragung. Der Wert der NAV stellt nun für alle STAs, die nicht direkt angesprochen waren, den frühestmöglichen Zeitpunkt dar, ab dem erneut auf das Medium zugegriffen werden darf. Der adressierte Empfänger des RTS-Steuerrahmens sendet dann nach Abwarten des SIFS ein (Clear-To-Send) CTS-Steuerrahmen an den Absender zurück. Dabei werden die NAV der anderen STAs (auch versteckter!) auf den verbleibenden Wert der nun folgenden Datenübertragung gesetzt. Nach Abwarten des

SIFS kann der Sender nun den Datenrahmen senden, wobei dieser dann vom Empfänger nach Abwarten des SIFS mit einem ACK-Rahmen (Bestätigung) beantwortet wird. Nun ist die Zeit des NAV abgelaufen, und der freie Wettbewerb beginnt wieder. Man bezeichnet den RTS/CTS-Mechanismus als virtuelle Reservierung des Mediums. Graphisch sei das in Abbildung 12 veranschaulicht.

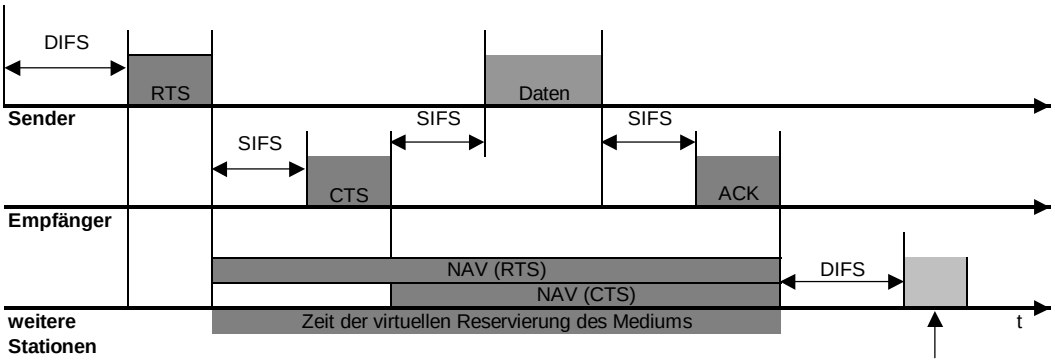

Abbildung 12: Medienzugriff mit RTS/CTS-Mechanismus

Es muß dabei aber beachtet werden, daß der RTS/CTS-Mechanismus das Netz mit der Übermittlung der RTS/CTS-Steuerrahmen zusätzlich belastet und dadurch die Zugriffsverzögerungen erhöht werden. Man muß also herausfinden, in welcher Situation es sich überhaupt lohnt, den RTS/CTS-Mechanismus einzusetzen. Außerdem ist der RTS/CTS-Mechanismus nicht in der Lage, das Problem versteckter Endgeräte vollständig zu lösen. Dies sei an folgendem Beispiel aus [WSFW97] gezeigt (Anordnung der STAs siehe Abbildung 13):

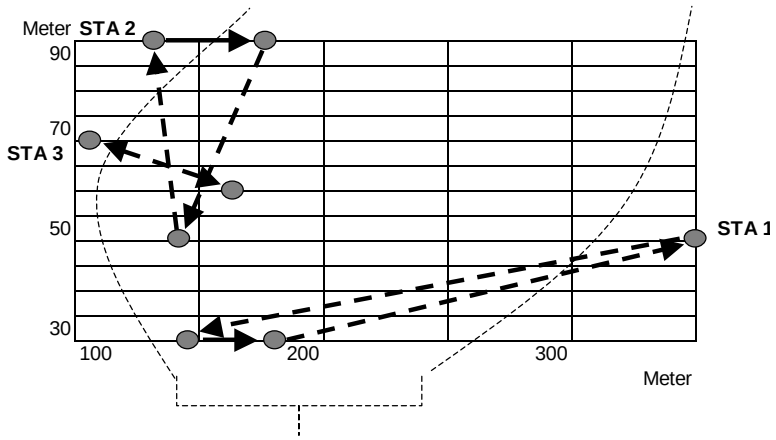

Abbildung 13: Beispiel für realistische Anordnung von STAs

Es handelt sich hierbei um ein Ad-hoc-Netz mit acht STAs, wobei die STA 1 vor STA 2 und STA 3 versteckt ist. Eine deutliche Verbesserung des Durchsatzes in Abhängigkeit von der Systemlast wird in dieser praxisnahen Konfiguration des Netzes nicht für alle acht Stationen erreicht, wenn der RTS/CTS-Mechanismus eingeschaltet wird. Dies hängt damit zusammen, daß die voreinander versteckten Stationen nach einer vergangenen Reservierung des Mediums ihre Backoff-Zähler für einen erneuten Zugriffsversuch gemeinsam neu starten, jedoch nicht in der Lage sind den Beginn eines Sendevorganges der jeweiligen anderen versteckten Station zu erkennen.

Die Belastung des Netzes mit den RTS/CTS-Steuerrahmen läßt das in Abschnitt 2.1 beschriebene Problem der Overheads noch größer werden. Eine weitere realistische Netzkonfiguration [WSFW97] (acht STAs, jedoch ohne versteckte Endgeräte) zeigt, daß bei einer relativ kleinen Nutzdatenrahmenlänge von beispielsweise 64 Byte die Werte des Durchsatzes und der Zugriffsverzögerung unter steigender Netzwerklast bei eingeschaltetem RTS/CTS-Mechanismus durchweg schlechter sind als ohne RTS/CTS-Mechanismus. Das Verhältnis der Steuerrahmen zu den eigentlichen Nutzdaten ist jetzt noch ungünstiger als in Abschnitt 2.1 beschrieben. Wird die Größe der Nutzdatenrahmen auf 682 Byte angehoben, wirkt sich der RTS/CTS-Mechanismus vorteilhaft auf Durchsatz und Zugriffsverzögerung aus, der Overhead wird kleiner. Aus diesem Grund wird ein RTS-Schwellenwert festgelegt, der bestimmt, ab welcher Nutzdatenrahmengröße der RTS/CTS-Mechanismus eingeschaltet wird. Dieser liegt nach den Simulationsergebnissen von [WSFW97] bei etwa 200 Byte, wobei die üblichen hinzukommenden Overheads der PHY-Schicht angenommen werden (siehe Abschnitt 2.1). Unter diesen Voraussetzungen kann der RTS/CTS-Mechanismus auch ohne versteckte Endgeräte Vorteile bringen.

2.2.3 Optionaler zeitbeschränkter Datendienst im Infrastrukturnetz mit PCF und Polling

Dienstgüte kann mit DCF nur „so gut wie möglich" realisiert werden. Aus diesem Grund sieht der 802.11-Standard eine zweite Implementierung der MAC-Schicht vor, nämlich die bereits erwähnte PCF. In Abschnitt 1 wurde die zugehörige Systemkonfiguration erklärt. Ein Point Co-Ordinator übernimmt im AP eine zentrale Steuerungsfunktion, weshalb die PCF nur im Infrastrukturnetz realisiert werden kann. Der AP hört (wie die STAs) in das Medium hinein und kann dann bereits nach Abwarten des PIFS auf das Medium zugreifen. Ab diesem Zeitpunkt beginnt die sogenannte wettbewerbsfreie Periode, in der vom Point Co-Ordinator ein Polling-Verfahren eingesetzt wird. Der Point-Co-Ordinator beginnt, den einzelnen Stationen in Abwärtsrichtung (D) der Reihe nach einen Datenrahmen zu senden. Nach jedem gesendeten Datenrahmen wartet der Point Co-Ordinator, ob eine Station nach Abwarten eines SIFS mit einem Datenrahmen in Aufwärtsrichtung (U) antwortet. Hat der Point-Co-Ordinator alle Stationen abgefragt, läßt er abschließend mit einem CF_{end}-Steuerrahmen (Contention Free end) wieder den Wettbewerb zu. Die Verfahren der vorhergehenden Abschnitte greifen dann wieder. Die gesamte Zeitspanne, in der der Point Co-Ordinator plant, die Medienzugriffssteuerung zu übernehmen, wird als Superrahmen bezeichnet. Die Länge des geplanten Superrahmens wird bei Beginn des ersten D-Rahmens in den NAV der STAs gespeichert. So werden in der wettbewerbsfreien Periode unberechtigte Zugriffsversuche vermieden. Den Ablauf dieses Verfahrens zeigt Abbildung 14.

Während der wettbewerbsfreien Periode (Superrahmen) wird nun deterministische Dienstgüte (maximale Zugriffsverzögerung) geboten, weil das Zugriffsschema des Polling-Verfahrens die Übertragungseigenschaften des Systems bestimmt und bekannt ist. Da das Wettbewerbsverfahren jedoch auch den Zugriff des Point Co-Ordinators hinausschieben kann, muß die Übertragungsdauer eines maximalen Datenrahmens, der dann gerade noch übertragen werden darf, zur maximalen Zugriffsverzögerung hinzuaddiert werden. Die Übertragungsrate des Netzes wird jedoch gegenüber der DCF nicht deutlich verbessert, weil die PCF einen erheblichen zusätzlichen Overhead erzeugt. Einerseits wird versucht, die verfügbare Bandbreite gerecht aufzuteilen, indem

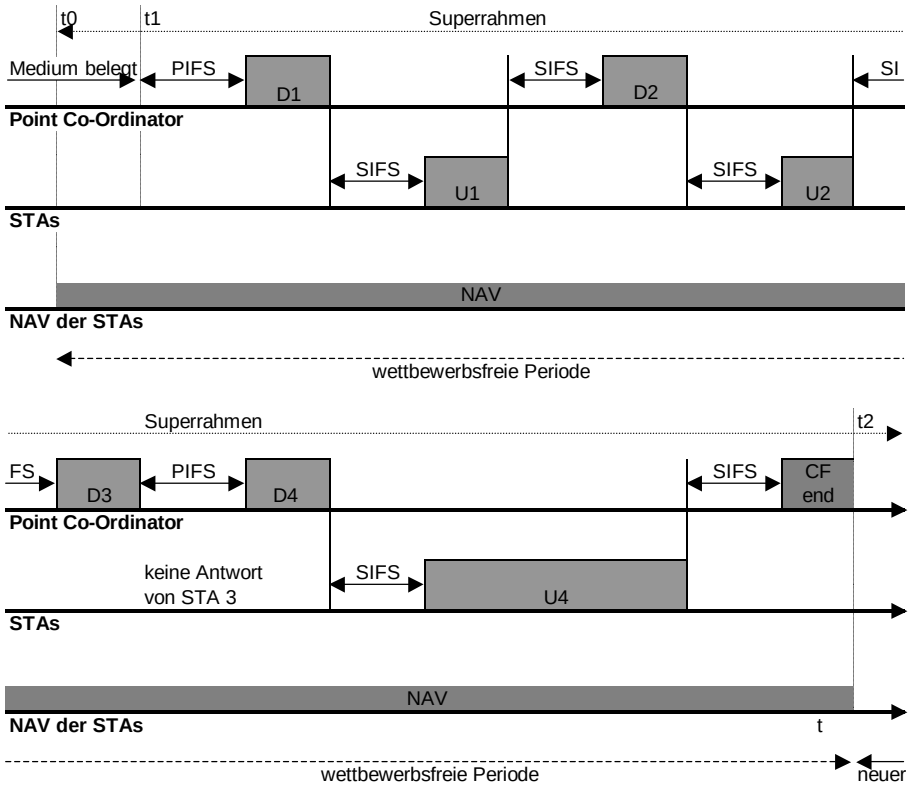

Abbildung 14: Medienzugriffssteuerung mit PCF

alle STAs sukzessive abgefragt werden, andererseits werden auch STAs abgefragt, die nicht sendewillig sind, und so wird Bandbreite verschenkt.

In [WSFW97] wird ein BSS mit acht STAs simuliert. Beide Betriebsarten (DCF und PCF) werden hier auf ihren Durchsatz in Abhängigkeit von der Netzwerklast untersucht. Der Vorteil, den PCF hier für sich verbuchen kann, fällt eher gering aus. Bei DCF werden 66% der zu verarbeitenden Netzwerklast durchgesetzt, bei PCF sind es dagen 70%. Ob dieser Unterschied von 4% die Investition in wesentlich aufwendigere Funkadapter mit PCF lohnt, muß in spezifischen Anwendungsfällen simuliert und dann entschieden werden.

2.3 Leistungssteuerung und ihre Auswirkungen auf QoS

Die Stromversorgung mobiler Endgeräte erfolgt in der Regel durch Batterien oder Akkus, deren Kapazität beschränkt ist. Deshalb sieht der 802.11-Standard eine Leistungssteuerung vor. Dabei soll unbeschäftigten (auf Datenrahmen wartenden) STAs ein Standby-Modus ermöglicht werden, der den Stromverbrauch herabsetzt. Die grundlegende Funktionsweise dieser unabdingbaren Leistungssteuerung soll in diesem Abschnitt erklärt werden. Dabei werden Auswirkungen auf QoS ermittelt. Auch auf die Sendestärke der Funkeinheiten, die den Stromverbrauch beeinflußt, wird eingegangen. Die Leistungssteuerung kann in Ad-hoc- und infrastrukturbasierten Netzen realisiert werden.

Hierzu werden jetzt zwei Zustände der STAs eingeführt: schlafend (sleep) und wach (awake). Jede 802.11-Funkeinheit besitzt eine interne Uhr mit einer besonderen Funk-

tion, der Timing Synchronisation Function (TSF). Die TSF veranlaßt die STAs dazu, in regelmäßigen Abständen aufzuwachen. Dann können die STAs ihre (evtl. zwischengespeicherten) Daten an die gewünschten Ziele schicken. Idealerweise sollten deshalb alle Stationen gleichzeitig aufwachen. Sind diese Zeitpunkte wegen nicht mehr richtig synchronisierter Uhren verschoben, werden die Daten in den STAs zwischengespeichert. Die Synchronisierung der Uhren erfolgt durch das Aussenden von Beacon-Rahmen (Leuchtfeuer) durch den AP (oder der STAs untereinander im Ad-hoc-Modus). Um den aufgewachten STAs bevorstehende Datenübertragungen anzukündigen, werden zusammen mit dem Beacon-Rahmen Informationen übertragen, welche STAs mit Datenrahmen zu rechnen haben und deshalb bis zur erfolgreichen Datenübertragung wach bleiben müssen. Für Unicast-Nachrichten sind das eine Traffic-Indication-Map (TIM) und für Broadcast-Nachrichten eine Delivery-Traffic-Indication-Map (DTIM). Im Ad-hoc-Modus versuchen die STAs zum gleichen Zweck Ad-hoc-TIMs (ATIMs) gegenseitig auszutauschen. Nun können die STAs nach Ende der Datenübertragung mit Hilfe interner Parameter entscheiden, ob es sich lohnt bis zum nächsten (von der TSF geplantem) Aufwachen erneut in den Schlafzustand zu gehen oder gleich wach zu bleiben. Die hier konkurrierenden Parameter sind bei möglichst vielen Schlafphasen die hohe Energieeinsparung und andererseits bei möglichst vielen Wachphasen die geringere Zugriffsverzögerung. Die unterschiedlichen TIMs sowie die Beacon-Rahmen werden in bestimmten Intervallen versendet. Jetzt stellt sich die Aufgabe, die Länge dieser Intervalle so zu bestimmen, daß ein angemessenes Energiesparverhalten bei gleichzeitig geringen Zugriffsverzögerungen bzw. hohem Durchsatz erreicht wird.

Hierzu wird in [RöWW98] ein Ad-hoc-802.11-Netz mit acht Stationen und DSSS untersucht. Versteckte Endgeräte seien hier nicht betrachtet. Der anfallende Datenverkehr wird aus den Aufzeichnungen eines realistischen Ethernet-Datenverkehrs simuliert. Unter verschiedenen Netzwerkauslastungen wird der Durchsatz des Netzes (Byte/s) in Abhängigkeit der Längen der ATIM-Fenster und Beacon-Intervalle gemessen. Bei der Auslastung des Netzwerkes von etwa 61% ergibt sich folgendes Ergebnis:

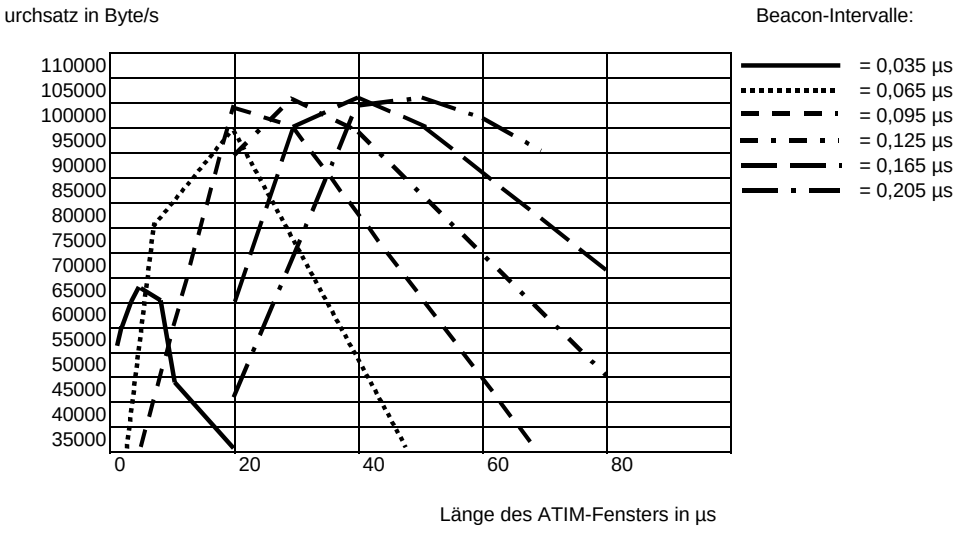

Abbildung 15: Durchsatz in Abh. verschiedener Beacon-Intervalle und ATIM-Fenster

Man kann erkennen, daß es mehrere optimale Kombinationsmöglichkeiten der Längen von Beacon-Intervallen und ATIM-Fenstern gibt. Das Verhältnis sollte etwa 1/4 zu 1/3 sein.

Eine weitere Größe, die zu konträren Ergebnissen führt, ist die Sendeleistung der Funkeinheiten. Umso geringer die Sendeleistung ist, umso höher ist der Energiespareffekt. Geringe Sendeleistung führt jedoch zu höheren Bitfehlerraten.

In [EbWo99] wird zur Analyse dieser Problematik ein 802.11-Netz herangezogen, das als DSSS-System mit 2 Mbit/s realisiert ist. Trägt man in diesem Versuchsaufbau die Sendeleistung (TX Power in dBm), die gerade erforderlich ist, um den Empfänger zu erreichen, an der Bitfehlerrate (Bit Error Rate, BER) ab, so erhält man folgendes Schaubild:

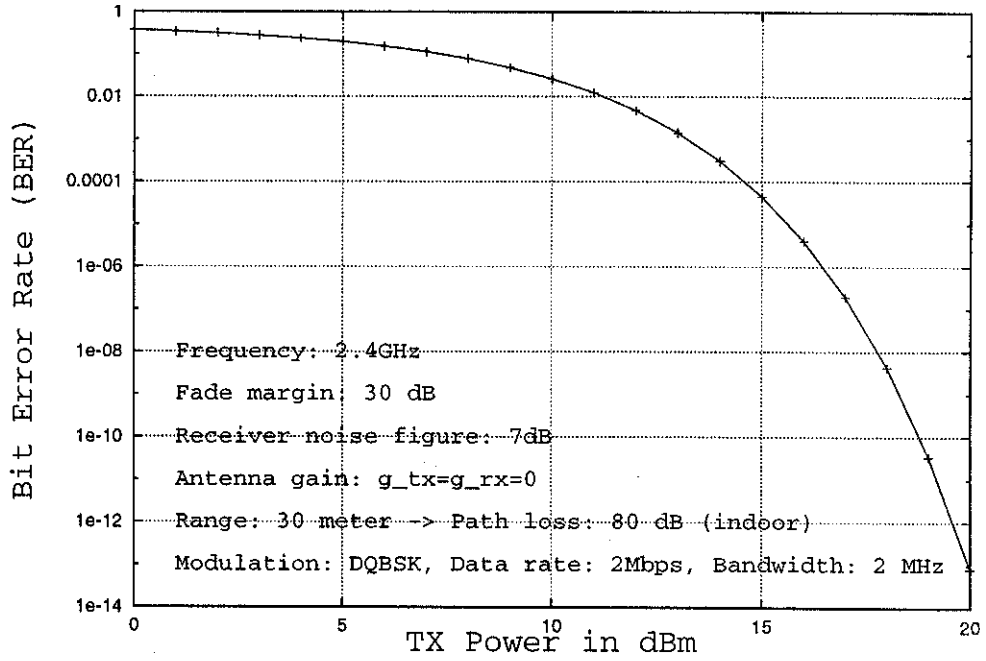

Abbildung 16: Bitfehlerrate in Abh. von der Sendeleistung

Dieses Schaubild zeigt, daß mit zunehmender Sendeleistung die Bitfehlerrate deutlich abnimmt. In der Praxis reicht diese Betrachtung alleine jedoch nicht aus. Es kommen immer zusätzliche Faktoren hinzu, die die Übertragung stören können. Beispielsweise werden STAs durch ihren Benutzer im Raum bewegt und so die Entfernung zum AP oder einer anderen STA variiert. Dies hat zur Folge, daß auch die tatsächlich beim Empfänger ankommende Sendeleistung und damit die Bitfehlerrate variiert. Diese Veränderungen können natürlich nicht exakt meßtechnisch erfaßt werden. Hierzu bedient man sich eines Wahrscheinlichkeitsmodells, dem Gilbert-Elliot-Kanal-Modell. Es kennt zwei Zustände: den Zustand besonders hoher Bitfehlerrate (Bad) und den Zustand besonders niedriger Bitfehlerrate (Good). Die Übergänge von Good nach Bad oder die Möglichkeit, daß sich der Zustand nicht ändert, sind mit bestimmten Übergangswahrscheinlichkeiten beschrieben. Dieses praxisnahe Modell kann den Idealverlauf aus Abbildung 16 deutlich verändern und muß deshalb berücksichtigt werden.

In [EbWo99] wird nun im weiteren wieder die oben erwähnte Netzkonfiguration betrachtet. Hierbei soll der Begriff der Protokolleffizienz genauer betrachtet werden. Protokolleffizienz gibt das Verhältnis der erfolgreich übertragenen Bits zu der Anzahl der insgesamt zu übertragenden Nutzdaten-Bits an. Der Wert liegt zwischen 0 und 1, wobei 1 (alle Nutzdaten-Bits erfolgreich übertragen) durch die PHY und MAC Overheads nie erreicht werden kann. Ein weiter Koeffizient, die effektive Übertragungsleistung,

soll noch betrachtet werden. Er gibt das Verhältnis der tatsächlichen Sendeleistung zur Protokolleffizienz an. Dieser Wert bedeutet also, wieviel Sendeleistung tatsächlich auch erfolgreich genutzt wird. In [EbWo99] werden diese Werte ins Verhältnis zur Größe der übertragenen Nutzdatenrahmen und zur Anzahl der im Netz agierenden STAs gesetzt. Das Ergebnis dieser Simulationen soll hier kurz verbal dargestellt werden. Es zeigt sich, daß für jede Kombination aus Sendeleistung, Paketgröße und Anzahl der Station eine bestmögliche Protokolleffizienz erreichen läßt. Dies läßt folgende Schluß zur Optimierung der Übertragungsqualität zu: Die Arbeitsweise der PHY-Schicht und der MAC-Schicht müssen aneinander angepaßt werden, d.h. die MAC-Schicht muß der PHY-Schicht mitteilen, welche Sendeleistung für den aktuellen Netz-Zustand erforderlich ist (Protokollharmonisierung). Die 802.11-Funkeinheiten lassen eine Variation der Sendeleistung stufenweise zu.

2.4 Roaming zur Gewährleistung wirklicher drahtloser Flexibiltät

Eine einzelne Funkzelle eines 802.11-Netzes (bisher auch BSS genannt), hat innerhalb eines Gebäudes die Reichweite der Größe eines Zimmers oder eines Stockwerkes je nach Wandstärken. Deshalb sind innerhalb eines Gebäudes oder eines ganzen Firmengeländes mehrere Zugangspunkte (APs) notwendig. Ein drahtloses Netz macht jedoch nur wirklich Sinn, wenn sich der Anwender mit seinem mobilen Endgerät (STA) im gesamten Bereich bewegen kann. Dabei sollen Schwankungen der Übertragungsrate und der Zugriffsverzögerungen möglichst klein bleiben und ein Abbruch der Verbindung unbedingt verhindert werden. Es muß also ein Mechanismus existieren, der die STAs von einem zum anderen BSS weiterleitet (Roaming). Hierfür ist die beste Voraussetzung ein Infrastrukturnetz mit DS. Jede STA, die sich innerhalb eines BSS befindet, ist im DS registriert. Stellt eine STA fest, daß die Signalstärke des aktuellen AP zu gering ist, wird die Suche (Scanning) nach einem neuen AP gestartet. Diese Suche kann passiv geschehen, indem die Stärke der empfangenen Beacon-Rahmen verglichen wird und ein AP mit der größten Signalstärke ausgewählt wird. Es gibt auch eine aktive Variante, bei der die STA ein Suchsignal sendet, das von den APs beantwortet wird. Auch hier wird dann die AP mit der größten Signalstärke ausgewählt. Die STA sendet dann eine Anfrage zur Aufnahme, die beantwortet wird. Ist die Antwort eines neuen AP positiv, wird dies dem DS zur Aktualisierung der Datenbank gemeldet und der Zellenwechsel ist vollzogen.

Im Ad-hoc-Netz ist im eigentlichen Sinn kein Roaming möglich, da alle STAs, die einem BSS angehören, in einem räumlichen Zusammenhang stehen müssen (Direktkontakt). Insofern wir hier nur die aktive Suche eingesetzt, um ein BSS aufzumachen. Das bedeutet, die STAs ermitteln dann, welche Kommunikationspartner sich in ihrer Reichweite befinden und ein BSS bilden könnten.

2.5 Anwendungsspezifische Anpassung von QoS

In diesem Abschnitt wird ein Verfahren angesprochen, daß im aktuellen 802.11-Standard nicht verankert ist, sich jedoch in allen WLANs realisieren läßt. Es geht hierbei um die adaptive Anpassung des QoS in Abhängigkeit der Anforderungen einer oder mehrerer verschiedener Anwendungen (z.B. Audio- oder Viedeoübertragungen).

Ein Verfahren dieser Art wird in [BiCL98] beschrieben. Zunächst muß ermittelt werden, welche Bandbreite eine bestimmte Anwendung benötigt. Hierzu wird eine Nutzenfunktion eingeführt. Sie beginnt ab einer bestimmten Bandbreite, die die Anwendung benötigt, um überhaupt betrieben werden zu können. Diese steigt dann (beispielsweise) abschnittsweise linear an. Die Schwellen geben Punkte an, ab denen sich weiterer Zuwachs der Bandbreite besonders positiv auf die Anwendung auswirkt oder anderenfalls keine so große Auswirkung zeigt. Diese Information wird nun einem Bandbreiten-Zuteilungs-Algorithmus zugänglich gemacht, der jedoch darauf angewiesen ist, daß ein zentraler Zugangspunkt (AP) die Verteilung der Bandbreite auf die STAs und damit auf die verschiedenen Anwendungen übernehmen kann. Die hierfür zur Verfügung stehende Gesamtbandbreite des Systems ist natürlich nicht konstant. Sie ist abhängig von der zu jedem Zeitpunkt bereits vergebenen Bandbreite und variiert durch Bewegung der STAs, die zu Schwankungen der Übertragungskapazität führt. Die Funktion des Bandbreiten-Algorithmus wird zusätzlich noch von der Einteilung der Anwendungen in bestimmte Serviceklassen bestimmt:

- Zugesicherte Dienstgüte (1): Eine minimale Bandbreite wird einer Anwendung zugesichert.
- Aktive Anpassung(2): Die Dienstgüte wird, soweit möglich, gemäß der Nutzenfunktion und weiterer additiver Werte zugeteilt.
- Passive Dienstgüte(3): Die Dienstgüte wird, soweit möglich, nur nach der Nutzenfunktion zugeteilt.

Der Bandbreiten-Zuteilungs-Algorithmus funktioniert dann folgendermaßen: Die erforderliche Bandbreite der zur Verfügung stehenden Gesamtbandbreite wird zunächst der Serviceklasse (1) zugeteilt. Die noch verbleibende Bandbreite wird dann der Serviceklasse (2) und anschließend der Serviceklasse (3) zugeordnet.

Im zentralen Zugangspunkt (AP) befindet sich ein Allocation Controller (AC), der die Bandbreiten gemäß des Bandbreiten-Zuteilungs-Algorithmus verteilt. Hierfür benötigt er Informationen über die Serviceklassen der einzelnen Anwendungen und über ihre Nutzenfunktionen. Diese erhält er durch die Adaptation-Handler (AH), die auf den einzelnen STAs, den jeweiligen Anwendungen entsprechend, individuell programmiert sind. Weiter AHs, die im AP installiert sind, liefern Informationen über die aktuell verfügbare Bandbreite im Netz. Die Unterstützung dieser Vorgehensweise erfordert eine erweiterte MAC-Schicht. Hiermit beschäftigt sich die 802.11e Working-Group.

3 Schlussfolgerungen und Aussicht

In der Einleitung wurde der 802.11-Standard in die Klasse der „bestmögliche Dienstgüte" eingeordnet. Diese Eigenschaft bedeutet, daß der 802.11-Standard keinerlei harte Garantien für die Dienstgüte übernehmen kann. Die Feinabstimmung verschiedener Parameter in der Medienzugriffssteuerung (siehe Abschnitt 2.2) lassen jedoch durchaus Variationsmöglichkeiten zu. Dabei ist vor allem zu berücksichtigen, daß ein 802.11-Netz mit relativ einfachen Funkadaptern zu realisieren ist, wenn die DCF eingesetzt wird. Der Einsatz adaptiver Methoden (siehe vorhergehender Abschnitt) erfordert zwar

einen hohen Programmieraufwand, sichert jedoch der Serviceklasse (1) deterministische Dienstgüte zu.

Die PCF, die ausschließlich im Infrastrukturnetz eingesetzt werden kann, bietet erweiterte Möglichkeiten zur QoS-Unterstützung (siehe Abschnitt 2.2.3). Sobald der Zugangspunkt die Medienkoordination übernommen hat, lassen sich bessere Aussagen zur Übertragungsrate und zum Durchsatz machen als bei DCF. Von zugesicherten Dienstgüteeigenschaften kann aber auch hier nicht gesprochen werden, weil der durch Steuerinformationen zusätzlich verursachte Overhead der Übertragungsrate entgegenwirkt. Dabei ist noch der wesentlich höhere Hardware-Aufwand für die PCF-tauglichen Funkadapter zu beachten.

Abschließend kann man sagen, daß ein 802.11-Netz am effizientesten ist, wenn es mit der DCF betrieben wird. Die beschriebenen Einschränkungen für QoS sind hinzunehmen, dafür arbeitet der Standard durch seine Einfachheit robust und ist kostengünstig. Für breitbandige und echtzeitabhängige Anwendungen stehen mittlerweile auch andere drahtlose Netze zur Verfügung, deren technischer und administrativer Aufwand jedoch weit über dem des 802.11-Standard liegt (z.B. Hiperlan). Als besonders kostengünstige Alternative steht vor allem für Ad-hoc-Verbindungen Bluetooth zur Verfügung.

Um die vom 802.11-Standard bekannten Vorteile der Einfachheit und Robustheit zukünftig für komplexere Anwendungen verwenden zu können, wird an seiner Weiterentwicklung gearbeitet. Beispiele hierfür wurden in Abschnitt 1 sowie in Abschnitt 2.5 genannt. Die Working-Group 802.11e, basierend auf 802.11b, geht dabei konkret auf QoS-Modifikationen ein.

Literatur

[BiCL98] Giuseppe Bianchi, Andrew T. Campbell und Raymond R.-F. Liao. On Utility-Fair Adaptive Services in Wireless Networks. Technischer Bericht, Politecnico di Milano, Dipartimento di Elettronica e Informazione and Center for Telecommunications Research, Columbia University, 1998.

[Bing99] Benny Bing. Measured Performance of the IEEE 802.11 Wireless LAN. Technischer Bericht, Department of Electrical and Computer Engineering, University of Maryland, College Park, 1999.

[Eber00] Jean-Pierre Ebert. An IEEE 802.11 WLAN Simulation Modell. Technischer Bericht, Telecommunication Networks Group, Technical University Berlin, 2000.

[EbWo99] Jean-Pierre Ebert und Adam Wolisz. Power Saving in Wireless LANs: Analyzing the RF Transmission Power and MAC Retransmission Trade-Off. Technischer Bericht, Telecommunication Networks Group, Technical University Berlin, 1999.

[EbWo00] Jean Pierre Ebert und Adam Wolisz. Combined Tuning of RF Power and Medium Access Control for WLANs. *Mobile Networks and Applications*, 2000.

[ESWW00] Jean-Pierre Ebert, Björn Stremmel, Eckhardt Wiederhold und Adam Wolisz. An Energy-efficient Power Control Approach for WLANs. *Journal of Communications and Networks*, 2000.

[KrRe00] Gerhard Krüger und Dietrich Reschke. *Telematik*. Fachbuchverlag Leipzig im Carl Hanser Verlag München Wien. 2000.

[RöWW98] Christian Röhl, Hagen Woesner und Adam Wolisz. A Short Look on Power Saving Mechanisms in the Wireless LAN Standard Draft IEEE 802.11. Technischer Bericht, Telecommunication Networks Group, Technical University Berlin, 1998.

[Schi00] Jochen Schiller. *Mobilkommunikation*. Addison-Wesley. 2000.

[WSFW97] Jost Weinmiller, Morten Schläger, Andreas Festag und Adam Wolisz. Performance Study of Access Control in Wireless LANs - IEEE 802.11 DWFMAC and ETSI RES 10 Hiperlan. Technischer Bericht, Electrical Engineering Department, Technical University Berlin, 1997.

[WWEW96] Jost Weinmiller, Hagen Woesner, Jean-Pierre Ebert und Adam Wolisz. Analyzing and Tuning the Distributed Coordination Function in the IEEE 802.11 DWFMAC Draft Standard. Technischer Bericht, Electrical Engineering Department, Technical University Berlin, 1996.

Abbildungsverzeichnis

Drahtlose Campus-Netze

Christian Biedermann

Kurzfassung

In den letzten Jahren wurden von verschiedenen Universitäten WLAN-Projekte nach IEEE 802.11 initiiert. Diese Netze unterscheiden sich in der Motivation, im Aufbau und im Betrieb. Diese Arbeit hat das Ziel, diese Unterschiede beispielhaft an drei Universitäten aufzuzeigen. Es handelt sich dabei um die Carnegie-Mellon-University in Pittsburgh, PA, die Universität Rostock sowie die Universität Karlsruhe. Neben den oben genannten Punkten Motivation, Aufbau und Betrieb im Allgemeinen werde ich außerdem auf die Authentifizierungsmechanismen sowie das Managementkonzept der WLAN eingehen. Der letzte Punkt hängt unmittelbar mit der eingestetzten Hardware zusammen. Daher werde ich diese ebenfalls vorstellen. Zum Schluß werde ich einen kurzen Ausblick auf die Zukunft dieser Netze geben.

1 Einleitung

1.1 *Wireless Local Area Network* (WLAN)

1.1.1 diverse Standards

Die folgende Auflistung stellt eine Auswahl der Standards dar, welche von verschiedenen Universitäten diskutiert wurden.[Burc00]

- Systeme, wie Bluetooth, IRDA, DECT haben aufgrund spezieller Gegebenheiten, wie z.B. Übertragung per Infrarot, eine zu geringe Reichweite. Sie sind daher für ein Campus-Netzwerk nicht geeignet.
- GSM bietet als ein weltweiter Standard für Mobilkommunikation zwar die nötige Reichweite, jedoch sind Nachteile wie eine niedrige Bandbreite und kostenpflichtige Frequenzbänder nicht zu unterschätzen. Erweiterungen wie HSCSD, GPRS oder EDGE wären nötig um akzeptable Übertragungsraten zu erzielen. Diese Zusätze sind größtenteils jedoch noch nicht in der Praxis erprobt.
- UMTS ist, nur die Leistungsmerkmale betrachtet, eine überlegenswerte Alternative. Jedoch sind die Kosten für die Frequenzbereiche in manchen Ländern nicht sehr billig (siehe Deutschland). Zudem wird dieser Standard erst 2003 in der Praxis verfügbar sein. Eine ausgiebige Testphase und somit Erfahrung im Betrieb liegen bisher nicht vor.

- IEEE 802.11
  Dies ist der Standard, den die meisten Universitäten für ihre drahtlosen Campusnetze einsetzten. Ich werde diesen Standard daher im weiteren Verlauf ausführlich behandeln.

1.1.2 Vor- und Nachteile eines WLAN?

Vorteile

- Flexibilität[Schi00]
  Der Benutzer ist mit seinem Endgerät nicht mehr an einen bestimmten Ort gebunden. Er kann sich also frei bewegen, sofern er dabei nicht das abgedeckte Gebiet verläßt.

- Vereinfachte Planung[Schi00]
  Es werden für die Clients keine weiteren Anschlußbuchsen und -kabel mehr benötigt. Selbst die Basisstationen könnten über Richtfunkstrecken verbunden werden. Allerdings wird dies meistens über ein kabelgebundenes Netzwerk realisiert. Auch ist es mit einem WLAN möglich, spontane Kleinstnetze, sogenannte Adhoc-Netze, aufzubauen. Hierfür ist meist nur die Präsenz einer Basisstation notwendig.

- Erweiterte Entwurfsfreiheit[Schi00]
  Die Entwicklung kleinerer tragbarer Geräte ist nur dann möglich, wenn das störende Kabel durch eine Funkverbindung abgelöst wird.

- Robustheit[Schi00]
  Solange die Stromversorgung sichergestellt ist, ist der Betrieb eines WLAN möglich. Eventuell durch Brand oder andere Einwirkungen auftretende Kabelbrüche sind ausgeschloßen.

- Kosten[Schi00]
  Als Beispiel reicht ein *Access Point* (AP) aus, um einen Hörsaal zu versorgen. Die Anzahl der Studenten spielt dabei (fast) keine Rolle. Bei einer verdrahteten Variante müssten alle Plätze mit einem Anschluß ausgestattet werden.

Nachteile

- Dienstgüte[Schi00]
  Fast alle Dienstgüteparameter eines WLAN haben gegenüber einem verdrahteten Netzwerk schlechtere Werte. Die Bandbreiten sind niedriger (10 Mbit/s statt 100-1000 Mbit/s), Übertragungsfehlerraten sind wesentlich höher (ca. 10^{-4} statt 10^{-12}) und die Verzögerungen/ Verzögerungsschwankungen aufgrund der Fehlerkorrekturmaßnahmen größer.

- Sicherheit[Schi00]
  Die Verwendung eines lizenzfreien Frequenzbandes kann unter Umständen zu Problemen mit anderen Geräten führen. Als Beispiel sei hier die handelsübliche Mikrowelle genannt, weswegen einige Hersteller in ihren Treiberkonfigurationen

den Menüpunkt „Microwave Oven Robustness“ implementiert haben [Erns00]
Außerdem ist die Kommunikation durch das Medium Luft sehr leicht abhörbar. Daher müssen Produkte Möglichkeiten zur automatischen Datenverschlüsselung anbieten.

- Kosten[Schi00]
  Zur Zeit sind die Kosten für die WLAN-Komponenten (AccessPoints, Netzkarten) noch sehr hoch. Mit zunehmender Massenfertigung könnten sich die Preise jedoch bald anpassen.

1.1.3 Sicherheit im WLAN

Einige Sicherheitsmechanismen sind bereist durch den Standard IEEE 802.11 vordefiniert und werden von fast allen Herstellern in die Endgeräte bzw. Basisstationen implemetiert. So z.B. die Sicherung des Netzzuganges durch Einsatz einer Service Set ID (Funknetzname) oder die Verwaltung der Nutzer durch Access-Control-Listen in den Basisstationen. Ein anderer vom Standard vorgesehener Sicherheitsmechanismus ist die Verschlüsselung der Kommunikation durch *Wired Equivalency Privacy* (WEP). Hier sieht IEEE 802.11 die Verwendung von WEP 40-bit vor. Ein verbesserte 128-bit Variante ist bereits proprietär von einigen Herstellern verfügbar. Ich werde auf diese Mechanismen nocheinmal in Kap.1.4.1 eingehen.
Außerdem ist es selbstverständlich möglich, die Sicherheitsstrukturen mit zusätzlichen Mechanismen, welche bereits aus dem drahtgebundenen LAN bekannt sind, zu verbessern. Hier sei nur an die zusätzliche Verschlüsslung auf Applikationsebene durch Kerberos oder das Zwischenschalten einer Firewall erinnert.[Malb00]

1.2 WLAN an Universitäten

1.2.1 Gründe für den Einsatz von IEEE 802.11 an Universitäten

Folgende Punkte bilden die entscheidenen Vorteile des IEEE 802.11 gegenüber den in Kap. 1.1.1 vorgestellten WLAN-Standards [Tava00]:

- weltweiter Standard verfügbar
  Da bereits ein weltweiter Standard nach IEEE 802.11 existiert, kann auf proprietäre Lösungen verzichtet werden. Eine spätere kostenintesive Umrüstung ist demzufolge nicht zu erwarten.

- Die Technik für IEEE 802.11 ist von zahlreichen Anbietern und Herstellern verfügbar.

- es ist eine hohe Bandbreite (derzeit 11 Mbit/s, in den nächsten Jahren bis zu 54 Mbit/s verfügbar) möglich. Bei den früheren Versionen des 802.11-Standards waren nur Bandbreiten im Bereich 1-2 Mbit/s möglich.

- Dank der Zugehörigkeit des 802.11 zur 802.xx-Familie ist eine leichte Integrierbarkeit in bestehende Netzwerktechniken gegeben. Es werden die selben Schichten im ISO/OSI-Referenzmodell implementiert.

- Durch eine einfache Installation der Komponenten ist eine Anwendung des WLAN auch ohne technische Detailkenntnisse möglich.

1.2.2 Anwendungen des WLAN an Universitäten

Neben den allgemeinen Vorteilen eines Netzzuganges wie z.B. die Durchführung von Recherche-Aufgaben im Internet an einem beliebigen Ort oder das Abrufen von Electronic- und Hypermedia-Büchern ohne den Einsatz von CD-ROM, bietet das drahtlose Netz einige weitere Anwendungsmöglichkeiten. Unter den Oberbegriffen „Virtuelle Universität“ und „Remote Lecturing“ verbergen sich Stichworte wie „Internet-basierte Vorlesungsunterstützung mit Studenteninteraktion“ oder „universitätsweites Lernen“. Es soll mit einem drahtlosen Netz möglich sein, Videokonserven von Vorlesungen abzuspielen, oder aber einfach Folien während der Vorlesung aus dem Netz zu laden. Sicherlich ist dies alles auch mit einem lokalen, drahtgebundenen Netzwerk möglich. Jedoch bietet ein WLAN die Möglichkeit, dies an jedem beliebigen Ort (in Karlsruhe z.B. im Schloßgarten) abzurufen. Theoretisch kann man also die Vorlesung im Schloßgarten verfolgen und per Interaktion dennoch an Übungen teilnehmen.
Nicht zu vergessen ist die Nutzung des WLAN als Forschungs- und Testnetz für die entsprechenden Arbeitsgruppen an der Universität.[Tava00]

1.3 IEEE 802.11

Der Standard, der das Fundament für ein WLAN bildet ist der IEEE 802.11. Dieser Standard wurde 1997 durch das *Institute of Electrical and Electronics Engineers* (IEEE) vorgestellt. In ihm sind die Bitübertragungsschicht sowie die Medienzugriffsverfahren spezifiziert. Diese erste Variante sah verpflichtend eine Datenrate von 1 Mbit/s und optional von 2 Mbit/s vor. Als Medienzugriffsverfahren wurden *Frequency Hopping Spread Spectrum* (FHSS) und *Direct Sequence Spread Spectrum* (DSSS) definiert. Ende 1999 wurde die erste Spezifikation auf IEEE 802.11b erweitert.

1.3.1 Spezifikationen IEEE 802.11b

- Es wird das lizenzfreie *Industrial, Science, Medical* (ISM)-Frequenzband bei 2,4 GHz benutzt.

- Datenraten von 1 Mbit/s bis zu 11 Mbit/s
  Wie bei allen Arten der Kommunikation, welche elektromagnetische Wellen als Informationsträger nutzen, kommt es auch bei WLAN zum Abfall der Datenrate, wenn die Entfernung zum Sender (Abb. 1) zunimmt oder das Signal durch andere Umstände (z.B. Abdeckung) schwächer wird. Sollten mehrere Nutzer denselben Zugangspunkt benutzen, wird hierbei die verfügbare Datenrate auf die Nutzer verteilt.

- es ist sowohl möglich, ein fest etablierte Funknetz als auch ein spontanes adhoc-Netz aufzubauen.

- es wird ausschließlich DSSS verwendet. Das früher eingesetzte FHSS wird nicht mehr verwendet.

Abbildung 1: Reichweite und Übertragungsgeschwindigkeit [Tava00]

- die Endgeräte können nun mit einem standardisierten Energiesparmodus ausgerüstet werden. Der Standard schreibt das Verhalten einer „schlafenden“ Netzwerkkomponente vor.
- Als Medienzugriffsverfahren wird CSMA/CA eingesetzt.

1.3.2 Strukturformen

- Ad-hoc-Netz [Schi00]
  Mehrere Clients bilden untereinander ohne die Nutzung eines AP ein spontanes Netzwerk. Diese Form nennt man Ad-hoc-Netz. Diese Struktur kann vor allem in Vorlesungen bzw. Übungen von Interesse sein, indem z.B. eine Programmieraufgabe vor Ort gelöst wird und das Ergebnis über das Ad-hoc-Netz an den Beamer weitergeleitet wird.
- autonome Netzform mit AP [Schi00]
  APs sind in der Regel festinstallierte Hardware. Ein AP und an diesem angemeldete Clients bilden zusammen ein so genanntes *Basic Service Set* (BSS). Sind die APs untereinander verbunden, so nennt man diesen Verbund *Distribution System* (DS). Mit *Extended Service Set* (ESS) bezeichnet man die Menge aller angeschloßenen BSS inklusive dem DS. In einem solchen ESS kann sich der Client frei bewegen, seine Netzanbindung wird durch Roaming sichergestellt. Der Begriff „Roaming“ ist an dieser Stelle im eigentliche Sinne falsch. Richtig wäre der Begriff „Handover“. In der Literatur wird jedoch meistens von Roaming gesprochen. Die Datenbestände der AP's werden über das DS aktualisiert, d.h. ein Client, welcher den AP wechselt, wird in die Liste des neuen AP eingetragen und aus der Datenbank des alten AP's automatisch entfernt.
- hybride Netzstruktur [Schi00]
  Das DS kann über ein Portal mit jedem anderen 802.xx-Netzwerk verbunden werden. Daraus ergibt sich die Möglichkeit das WLAN an das bestehende Campusnetz anzuschließen und damit die Anbindung ans Internet zu ermöglichen . Die

Zugehörigkeit der 802.11-Spezifikation zur 802.xx-Familie garantiert die Kompatibilät unter den Standards.

- Die Anbindung ans Internet wird in vielen Fällen durch ein *Virtual Local Area Network* (VLAN) realisiert.

Auf einem IEEE 802.11-Netz können beliebige Protokolle eingesetzt werden, da sich die Spezifikation nur auf die unteren beiden Schichten des ISO/OSI-Referenzmodells beziehen.

1.4 Hardware

1.4.1 Access Point

Zahlreiche Hersteller bieten inzwischen APs an. Die Geräte unterstützten größtenteils folgende Funktionen [Erns00]:

- *Access Control List* (ACL)
  In diesen Listen können die registrierten MAC-Adressen der Clientkarten eingetragen werden. Der AP erlaubt nur eingetragenen Adressen den Zugang zum WLAN.

- *Dynamic Host Configration Protocol* (DHCP)
  Den Clients wird bei der Anmeldung eine IP-Adresse zugewiesen. Sollte der AP einen übergeordneten DHCP-Server im Backbone feststellen, schaltet er seine eigene DHCP-Funktion aus.

- IP-Masquerading
  Der AP gibt die internen IP-Adressen der Clients nach außen als eine aus. Dadurch können mehrere Clients eine globale IP-Adresse gemeinsam nutzen.

- *Network Address Translation* (NAT)
  Hierdurch werden die internen IP-Adressen 1:1 in externe übersetzt. Die Struktur des WLAN bleibt so nach außenhin verborgen.

- *Wired Equivalent Privacy* (WEP)
  Mit WEP soll das Abhören der 802.11-Funkverbindungen unterbunden werden. Die meisten Hersteller bieten zur Zeit WEP 40 an. Dies ist ein RC4-Algorithmus mit einem 64-Bit-Schlüssel. Da hierbei pro Frame 24 Bit als so genannter Inital Vector vom Algorithmus selbst berechnet werden, kann der Anwender selbst nur 40 Bit des Schlüssels frei wählen.
  Einige Hersteller haben ein proprietäres WEP 128 implementiert.

- Protokoll-Filter
  Der Standard 802.11 ist, wie bereits erwähnt, protokolltransparent. Um die Netzlast niedrig zu halten, kann man nur bestimmte Protokolle auf dem AP zulassen.

Die Möglichkeiten zur Konfiguration und Administration sind von Hersteller zu Hersteller verschieden. Zum Einsatz kommen z.B. Telnet, SNMP, HTTP oder spezielle Windows-Tools.

1.4.2 Client

Fast jeder Hersteller von APs bietet die passenden Client-Karten an. Für Notebook im PCMCIA-Format und für PCs als Steckkarte. Die Authentifizierung eines Clients am Netzwerk ist nicht in jedem Falle vom Standard IEEE 802.11 festgelegt. In der Praxis werden verschiedenen Authentifizierungsmechanismen verwendet. Diese werden später erläutert. [Erns00]

2 Reale Umsetzung: Beispiele

Wie eingangs erwähnt, haben mehrere Universitäten in den letzten Jahren mit WLAN-Projekten begonnen. Folgende Universiäten möchte ich kurz näher beleuchten:

1. Universität Rostock
2. Universität Karlsruhe
3. Carnegie Mellon University Pittsburgh,PA

2.1 *Wireless Infrastructure for Students and Staff* (WISS) (Universität Rostock)

2.1.1 Motivation

Die Universität Rostock suchte eine praktikable, skalierbare und kostengünstige Lösung für folgende Probleme [Tava00]:

- es herrscht ein Mangel an Arbeitsräumen für Studenten
- der Bedarf an netzwerkgestützten Rechnerarbeitsplätzen wird immer größer
- Die Verkabelung von historischen Gebäuden wäre aufgrund von Denkmalschutzbestimmungen schwierig
- es stehen nur begrenzte Mittel für Neu- und Umbaumaßnahmen zur Verfügung

Als Lösung wurde in Rostock der Aufbau eines WLAN als Ausbau des bestehenden LAN beschloßen.

2.1.2 Geschichte und aktueller Zustand des Netzes

- seit November 1999 wurden Planung und Konzeption des WLANs vorangetrieben
- Im Februar 2000 begannen die eigentlichen Installationsarbeiten
- Anfang April 2000 wurde das WISS in Betrieb genommen

Da in Rostock die Planungen immer nur eine Ergänzung des bestehenden LAN vorsahen, werden auf absehbare Zeit nicht alle der Universität zugehörigen Gebäude durch das Funknetz abgedeckt werden. Die Dokumentationen der zuständigen Abteilung (Informatik Fakultät der Universität Rostock) weißen eindeutig auf dieses hin. So ist zum Beispiel die Funkvernetzung der Studentenwohnheime nicht geplant, da diese bereits durch ein bestehendes LAN abgedeckt werden. [Tava00]

2.1.3 Sicherheit und Management

Anmeldeverfahren

Nach folgendem Muster wird das Anmeldeverfahren für WISS an der Universität Rostock durchgeführt [Gern00]:

1. *Service Set ID*(SSID) = Netzwerkname
   Zuerst wird vom AP geprüft, ob der Client den selben Netzwerknamen (Wireless Domain Name) wie er selbst besitzt. Erst nach Übereinstimmung erfolgt die eigentliche Authentifizierung des Users.

2. ACL
   An der Universität Rostock wird, wie an der CMU, auf die in den APs implementierten ACLs zurückgegriffen. Dies bedeutet, daß jeder User die MAC-Adresse seiner Client-Karte, sei es für den PC oder das Notebook, zuerst registrieren lassen muß. Über die MAC-Adressen ist eine weltweit eindeutige Identifikation der Netzwerkkomponente möglich. Die MAC-Adresse wird nun in die ACL aller AP's eingetragen. *Network Interface Cards* (NIC), deren MAC-Adresse nicht in der ACL vorkommen, können über die AP's keine Pakete verschicken.

3. DHCP
   Über die MAC-Adresse wird dem Client im VLAN durch den DHCP-Server eine IP-Adresse zugeteilt.

4. WEP
   Es kommt zu guter Letzt zum Abgleich der Verschlüsselungscodes. Stimmen auch diese überein, wird dem Client der Zugang zum Netz gewährt.

Zusammenfassung

Insgesamt nutzen die Sicherheits- und Authentifizierungsprozeduren beim WISS ausschliesslich die, vom Standard gegebenen bzw. der Hardware implementierten Möglichkeiten. Daß es auch Erweiterungen gibt und wo die Schwächen der Standardprozeduren liegen, wird uns die Universität Karlsruhe anhand DUKATH näherlegen.

2.1.4 Zukünftige Entwicklung des Netzes

Da das WISS als Ergänzung zum bisherigen LAN konzipiert ist, werden schon verkabelte Gebäude vorerst nicht ans WLAN angebunden. Dies ist nicht zuletzt auch eine Frage der Kosten.

2.2 *Drahtlose Universität Karlsruhe (TH)*(DUKATH) (Universität Karlsruhe)

2.2.1 Motivation

DUKATH ist nicht nur, wie bei den bisher vorgestellten Universitäten, eine Erweiterung des bestehenden LAN, es verfolgt vielmehr die Zielsetzung, neue Wege in Lehre und Forschung anzubieten. Es sollen neue Möglichkeiten im Bereich der Vorlesungen, Seminare, Besprechungen, usw. zur Verfügung gestellt werden. Aber selbstverständlich ist auch die Ergänzung des wired LAN in den Bereichen, welche bisher nicht abgedeckt werden konnten, eine Aufgabe von DUKATH.[Wolf00b]

2.2.2 „Lebenslauf“ und Zukunft

Zu Beginn des Aufbaues wurden das Rechenzentrum, das Fakultätsgebäude der Informatik und das zentrale Verwaltungsgebäude am Ehrenhof mit Basisstationen versorgt. Inzwischen wurden weitere Accesspoints sowohl innerhalb von Gebäuden als auch zur Außenversorgung installiert.
Die weitere Planung sieht die Einrichtung der DUKATH-Komponenten in jedem neuen, sowie in etlichen bestehenden Hörsälen und auch in der Universitätsbibliothek vor. Je nach Inanspruchnahme des Netzes und sich daraus ergebender Notwendigkeit ist eine spätere Einführung von Mobile IP nicht ausgeschlossen.

2.2.3 Sicherheit und Management

Die Universität Karlsruhe entschied sich in ihrer WLAN-Architektur für eine über die Standardkomponenten weitergehende Implementierung von Sicherheitsmechanismen. Die von der Hardware angebotenen Möglichkeiten waren entweder zu schwach oder mit einem zu hohem Maß an adminstrativen Aufwand verbunden. So ist z.B. die Registrierung der MAC-Adressen für kleine Kartenanzahlen noch realisierbar. Steigt die Anzahl jedoch in den Bereich von mehreren Tausenden, wird die Administration schnell unübersichtlich. Ausserdem wäre mit einer gestohlenen Karte trotzdem der Zugang zum Netz möglich.
Die Nutzung des Funknetznamens (SSID) ist bei einer Netzgröße wie sie bei DUKATH angestrebt wird, ebenfalls nicht als Sicherungsverfahren geeignet. Wie geheim ist eine solche SSID bei mehreren tausend Nutzern?

Zusätzliche Maßnahmen bei DUKATH

Da bei DUKATH auf die Nutzung der ACL verzichtet wurde, ist das WLAN prinzipiell jedem Client, welcher die korrekte Zugangskonfiguration besitzt, zugänglich. Eine Kommunikation innerhalb DUKATHs ist demzufolge ohne Authentifizierung möglich, da dem Client vom im WLAN befindlichen DHCP-Server eine interne IP-Adresse zugewiesen wird. Wie in Abb.2 dargestellt, ist der Zugang zum KLICK (das kabelgebundene Campus-Netz) und damit zum Internet jedoch nur über einen dedizierten Übergangspunkt möglich. Um nun eine Kommunikation aus dem WLAN heraus aufzubauen, ist eine Authentifizierung mit *Benutzername und Passwort* (UID/PWD) an

einem der Gateways nötig. Die Verbindung zwischen Client und Gateway wird mittels PPTP (Abb.3) hergestellt.
Bei PPTP übernehmen der *PPTP Access Concentrator* (PAC) und der *PPTP Network Server* (PNS) die Aufgaben eines *Network Access Servers* (NAS). Der PAC übernimmt die Rolle des Einwahlservers, d.h. das Endgerät baut eine Verbindung zu ihm via PPP auf. Der Name Access Concentrator leitet sich aus der Aufgabe, mehrere PPP-Verbindungen zu bündeln und zu multiplexen, ab. Der PNS kapselt die multigeplexten PPP-Packete und schickt diese über den PPTP-Tunnel an den PNS. Der PNS entpackt das erhaltene Packet und deplext die verschiedenen PPP-PDUs. Der PNS schickt nun die eigentlichen Packete an den Empfänger. Das so getunnelete Protokoll muß nicht zwingend IP sein. Es ist auch möglich z.B. IPX über PPTP zu tunneln.
Durch den Einsatz von Techniken zum Aufbau eines *Virtuelles Privates Netzwerk* (VPN) ist es möglich, Client und PAC zusammenzulegen. Das Gateway übernimmt dann die Rolle des PNS. Über diesen Tunnel, und nur über diesen, ist es dem Client möglich mit Rechnern außerhalb des WLAN zu kommunizieren. Seine Pakete laufen alle über den Router zum Gateway und von dort zum eigentlichen Adressaten.

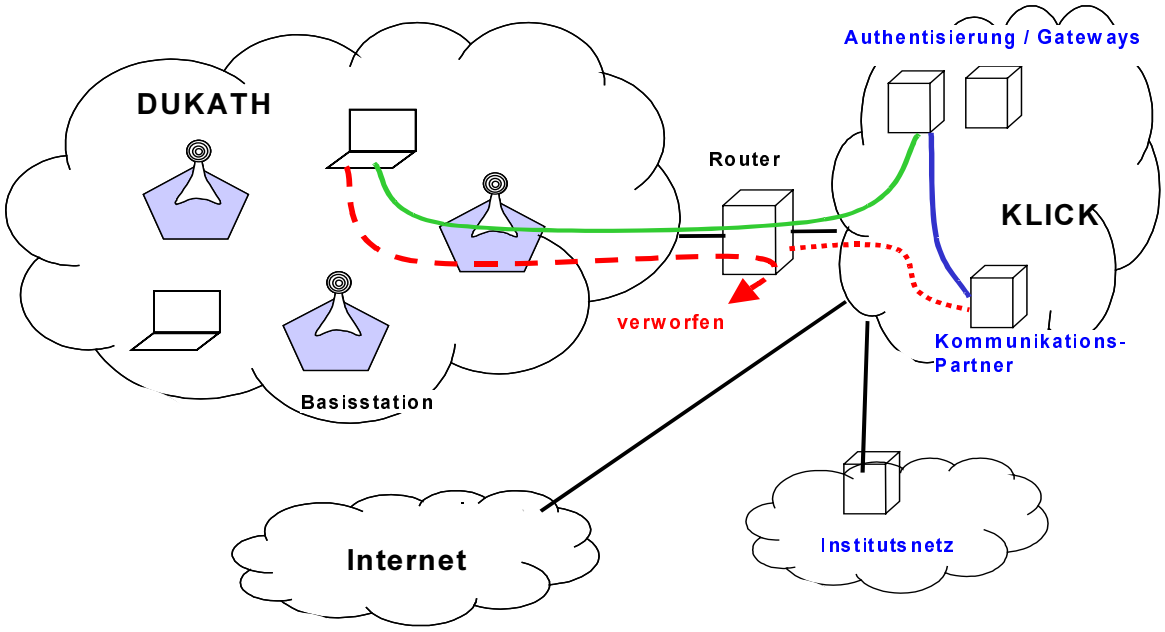

Abbildung 2: Erforderliche Anmeldung zur Nutzung von DUKATH [Wolf00a]

Vorteile des Mehraufwandes

- Durch die Authentifizerung am entsprechenden Gateway ist es dem Benutzer möglich, sich direkt in sein Institut einzuwählen, um so die Netzkomponenten (z.B. Netzdrucker) vor Ort zu nutzen.
- Die Registrierungsprozedur der WLAN-Karte entfällt. Bei Diebstahl oder Weitergabe ist ein Zugriff auf das KLICK bzw. Internet nicht möglich.
- Es gibt nur ein Passwort für DFÜ-Einwahl und Nutzung des DUKATH-Netzes. Da die Gateways im Passwortverteilmechanismus eingebunden sind, werden sie immer mit den aktuellen Dialin-Accounts und Passwörtern versorgt.

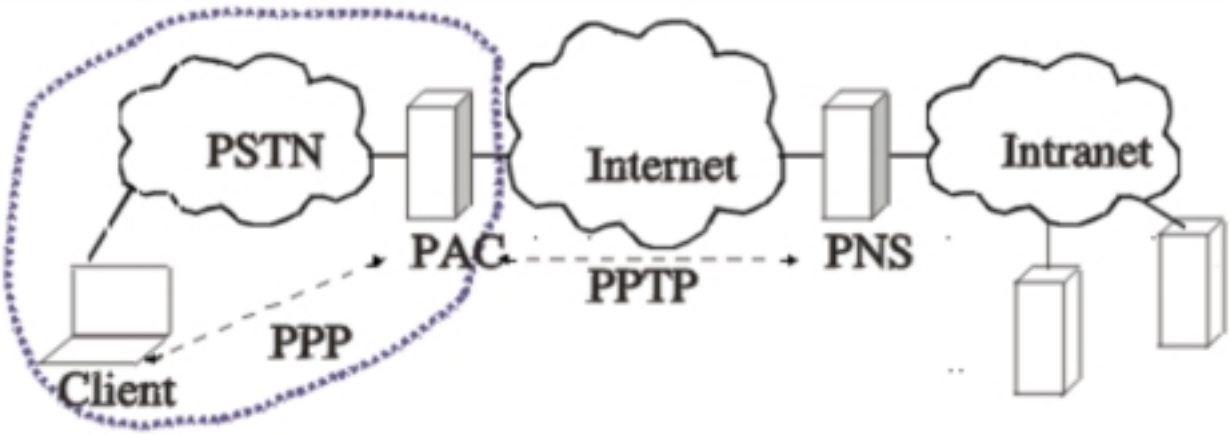

Abbildung 3: PPTP - schematische Darstellung [Wolf00b]

- Mittels PPTP ist eine Verschlüsselung möglich (Abhörsicherheit).

- Sollten unterschiedliche Nutzer das selbe Notebook verwenden, so ist sichergestellt, daß diese trotzdem in ihr jeweiliges Insitut eingebunden werden. Dies geschieht einzig über das UID/PWD, und damit die Authentifizierung am entsprechenden Gateway, ohne daß die Konfiguration des Clients dafür geändert werden müsste.

2.3 Wireless Andrew (Carnegie Mellon University Pittsburgh, PA)

2.3.1 Motivation, Geschichte und aktueller Zustand

1994 als Forschungsnetzwerk geplant, hatte Wireless Andrew (benannt nach dem Industriellen Andrew Carnegie und dem Finanzier Andrew Mellon) ursprünglich die Aufgabe WLAN-Projekte zu unterstützen. Heute ist Wireless Andrew eines der größten, nach Angaben der CMU sogar das größte bestehende WLAN. Es wird schon lange nicht mehr nur zu Forschungszwecken genutzt, sondern bietet inzwischen auch Studenten und Mitarbeitern der CMU die Möglichkeit die Vorteile eines WLAN zu nutzen. Wireless Andrew wurde bisher nach dem Standard IEEE 802.11 betrieben und besaß somit z.B. nur eine maximale Datenübertragungsrate von 2 MBit/s. Aber im Juli und August 2000 wurde Wireless Andrew auf den Standard 802.11b aktualisiert.
Die APs sind, wie in 1.3.2 beschrieben, über das *Distributed System* miteinander verbunden. Bei Wireless Andrew wird das DS durch ein eigens für diesen Zweck neuverlegtes IEEE 802.3-Backbone gebildet. Auch hier gibt es einen dedizierten Übergang zum restlichen Campus-Netz. Der Router ist angewiesen nur Pakete hindurchzulassen, dessen Adressat tatsächlich auch auf der anderen Seite liegt. Zum Beispiel können Broadcast-Pakete, die gesendet wurden, um einen im Roaming verlorenen Client wiederzufinden, so abgefangen werden, bevor diese das Campus-Netz belasten.

2.3.2 Sicherheit und Management

Außer der Registrierung der MAC-Adressen und der Überprüfung des Netzwerknamens (SSID) besitzt dieses WLAN keine weiteren Sicherheitsmechanismen. Das WEP-Konzept wurde bisher noch nicht implementiert, da man auf Seiten der Carnegie Mellon University auf die Einführung von Verschlüsselungsverfahren wie Kerberos auch im drahtlosen Bereich warten möchte.
Eine Unterteilung in Subnetze wäre inzwischen wünschenswert von Seiten der Administrative in Pittsburgh, jedoch würde mit den jetzigen Standards (IEEE 802.11b und IP) ein Roaming nicht mehr möglich sein.
Im Übrigen ist das Abhören der Verbindungen ein bekanntes Problem und es wurde auch schon die Einführung von WEP diskutiert. Allerdings würde dies nur die Abhöraktionen von „außen“ unterbinden. Da aber jeder authentifizierte Benutzer den selben WEP-Schlüssel verwenden würde, wäre das interne Abhören noch nicht unterbunden. So beantworten die Verantwortlichen alle Fragen zu diesem Thema mit dem Kommentar, daß Wireless Andrew eben ein offenes Netz sei.

2.3.3 Zukünftige Entwicklung des Netzes

Der Ausbau von Wireless Andrew zu Handheld Andrew ist bereits in Planung. Mit Handheld Andrew soll es dann auch möglich sein z.B. mit PDAs ins Netz zu gehen.
Die Entwicklung von Mobile IP könnte das oben angesprochene Subnet-Problem lösen. Auch hier wartet die Carnegie Mellon University auf eine entsprechende Standardisierung.

Literatur

[Burc00] Frank Burchert. Drahtlose Kommunikationstechniken. http://wiss.informatik.uni-rostock.de/veranstaltungen/vortraege/burchert.pdf, 7 2000.

[Erns00] Dusan Zivadinovic Ernst Ahlers. Datenkuriere. *c't* Band 22, 2000, S. 260–273.

[Gern00] Thomas Gerneth. WLAN-Systeme und Komponenten. http://wiss.informatik.uni-rostock.de/veranstaltungen/vortraege/gerneth.pdf, 7 2000.

[Malb00] Tobias Malbrich. WLAN-Systeme im universitären Einsatz. http://wiss.informatik.uni-rostock.de/veranstaltungen/vortraege/malbrich.pdf, 7 2000.

[Schi00] Jochen Schiller. *Mobilkommunikation*. Addison-Wesley. 2000.

[Tava00] Djamshid Tavangarian. Einsatz von WLAN - Projekt WISS an der Universität Rostock. http://wiss.informatik.uni-rostock.de/veranstaltungen/vortraege/tavangarian.pdf, 7 2000.

[Wolf00a] Lars Wolf. DUKATH - das drahtlose Netz der Universität Karlsruhe. http://www.uni-karlsruhe.de/ DUKATH/Folien/dukath4pik-kopfzeile.pdf.pdf, 11 2000.

[Wolf00b] Lars Wolf. DUKATH - Drahtlose Kommunikation an der Universität Karlsruhe. http://www.uni-karlsruhe.de/ DUKATH/Folien/wlan-bmbf-info-101100.pdf, 11 2000.

Abbildungsverzeichnis

HiperLAN/2

Jochen Dinger

Kurzfassung

HiperLAN/2 ist ein Wireless LAN Standard der 2. Generation. Er wurde aufgrund des gestiegenen Bedarfs an WLANs geschaffen. Durch verschiedene Modulationsverfahren sind Bandbreiten von bis zu 54 Mbit/s (auf Layer 1) möglich. HiperLAN/2 integriert Funktionen zur Verschlüsselung und Authentifizierung, aber auch Dienstgüte-Parameter (QoS). HiperLAN/2 definiert allerdings nur die ersten zwei Layer und ist somit auf die Layer anderer Netzwerk-Protokolle angewiesen. Um diese optimal umsetzen zu können, wurde der Convergence Layer mit speziellen Sublayern geschaffen. Sublayer existieren beispielsweise für ATM und Ethernet, aber auch UMTS. Der Standard zählt zu den besten Entwicklungen in diesem Marktsegment.

1 Einleitung

Die Nachfrage für drahtlose lokale Netzwerke stieg in den letzten Jahren stetig an. Dies hat mehrere Gründe. Zum einen sind dies die schnelle Installation ohne bauliche Veränderungen, wie z.B. Mauerdurchbrüche und Kabelgewirr. Zum anderen erlauben die Wireless LANs (WLAN) eine Vernetzung über Grundstücksgrenzen hinweg, was teure Standleitungen erspart. WLAN eröffnen auch neue Möglichkeiten. In Krankenhäusern können beispielsweise Notebooks zur Visite eingesetzt werden, um direkt auf den kompletten Datenbestand zugreifen zu können. Dies ist erst mit einem WLAN möglich, da die Endgeräte mobil sind und sich frei in einer Funkzelle bewegen können. Aufgrund der relativ geringen Sendeleistung, vergleichbar einem schnurlosen Telefon, treten auch keine Störungen medizinischer Geräte auf.

Inzwischen ist der Markt für WLANs so groß, dass verschiedene Standards entstanden sind, die eine Kommunikation zwischen Geräten unterschiedlicher Hersteller ermöglichen sollen, wie z. B. der IEEE 802.11 bzw. die schnellere Variante 802.11b. Sie ermöglichen unter sehr guten Funkbedingungen Bandbreiten von bis zu 11 Mbit/s. Aufgrund des Bedarf an höheren Bandbreiten sind zur Zeit die Standards der 2. Generation in Ausarbeitung. Die zwei wichtigsten sind hierbei 802.11a und HiperLAN/2.

1.1 BRAN-Projekt

Die ETSI (European Telecommunications Standards Institute) ist eine europäische nonprofit Organisation, die Telekommunikationsstandards erarbeitet. Sehr bekannte und

weltweit eingesetzte Standards der ETSI sind GSM und DECT. Um sich speziell mit den Möglichkeiten der funkbasierten Datenübertragung auseinanderzusetzen, wurde das Broadband Radio Access Network Projekt (BRAN) im Frühjahr 1997 ins Leben gerufen. Es soll Standards zur schnellen Funkübertragung, mit relativ geringen Gerätepreisen, schaffen. Zunächst wurde HiperLAN1 standardisiert, welches eine Alternative zu IEEE 802.11 darstellt. Dieser Standard fand jedoch keine Verbreitung, da die Hersteller schon funktionsfähige 802.11 Geräte hatten und keinen zweiten Standard unterstützen wollten. Momentan beschäftigt sich das BRAN Projekt mit drei verschiedenen Standards zur Datenübertragung per Funk.

HiperLAN/2 ist der Nachfolger von HiperLAN1. Dieser Standard ist gedacht für einen schnellen und einfachen Aufbau eines WLAN. Datenraten werden bis zu 25 Mbit/s (netto) unterstützt, was die Übertragung von Multimediainhalten (z.B.: Video) ermöglicht. Die Endgeräte können mobil sein und sich während der Übertragung bewegen.

Hiperaccess ist ein festes drahtloses Netz, d.h. ein Ersatz zum drahtgebundenen Netz. Eine Bewegung der Endgeräte, während sie im Netz eingebucht sind, ist hierbei nicht möglich. Die Reichweite liegt jedoch bei 2-10 km. Der Einsatzort solcher Systeme wird im outdoor-Bereich sein, wie z.B. der Vernetzung eines Campus.

Hiperlink ist ein Richtfunk-Standard und unterstützt sehr schnelle Punkt-zu-Punkt Verbindungen (bis zu 155 Mbit/s), wobei hierbei die Sender und Empfänger statisch sind. Es kann beispielsweise als Teilstrecke eines herkömmlichen Netzes oder zur Verbindung der Access Points eines WLANs eingesetzt werden.

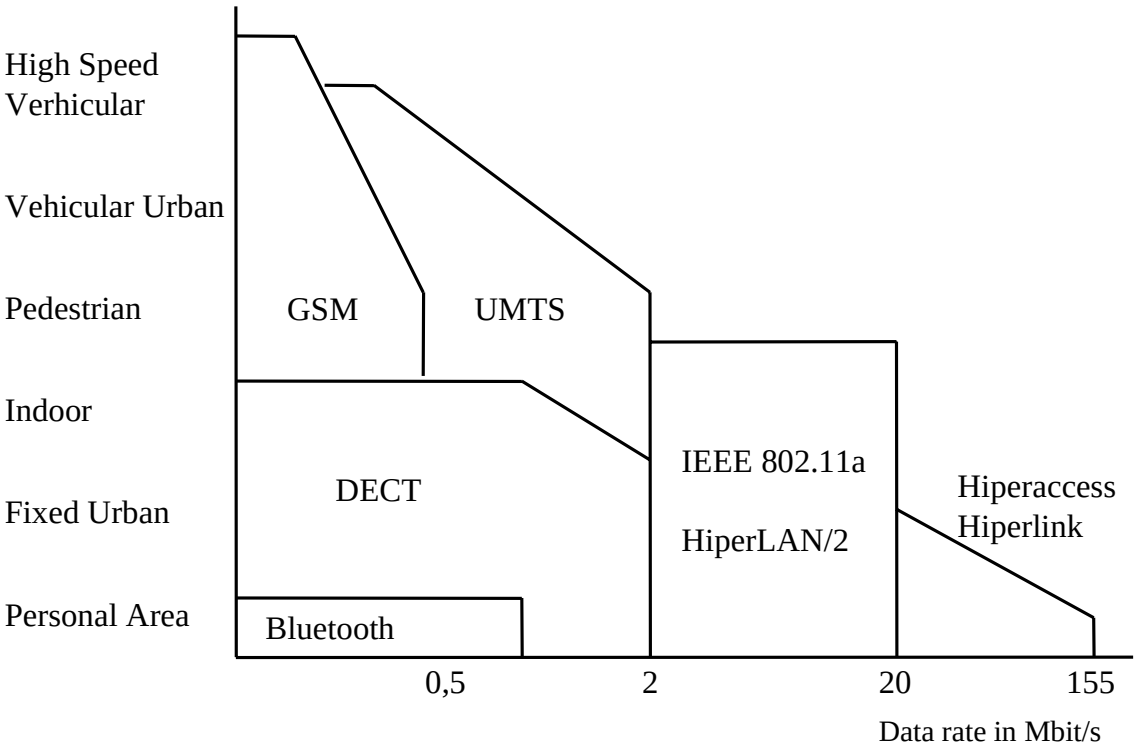

Abbildung 1: Mobilität vs. Geschwindigkeit ([AHKM00])

Die Standards des BRAN Projekt ordnen sich, wie in Abbildung 1 sichtbar, als Hochgeschwindigkeitslösung mit relativ geringer, bis zu keiner Mobilität in die Standards der Funkkommunikation ein. Aufgrund verschiedener Eigenschaften, wie Geschwindigkeiten, Reichweiten und Einsatzzwecke hat jeder Standard seine Daseinsberechtigung. Es

sind auch Kombinationen denkbar, die in nachfolgenden Abschnitten noch betrachtet werden. Nachfolgend werden die Eigenschaften von HiperLAN/2 genauer erläutert.

2 HiperLAN/2

2.1 Überblick

HiperLAN/2 ist eine Abkürzung für HIgh PErfomance Radio Local Area Network. HiperLAN/2 hat meistens eine Topologie, wie sie in Abbildung 2 dargestellt ist. Die mobilen Terminals (MT) kommunizieren per Funk mit einem Access Point (AP). Die APs vermitteln dabei zwischen den MTs und einem Festnetz. Eine Verbindung zwischen zwei MTs über den AP ist auch möglich.

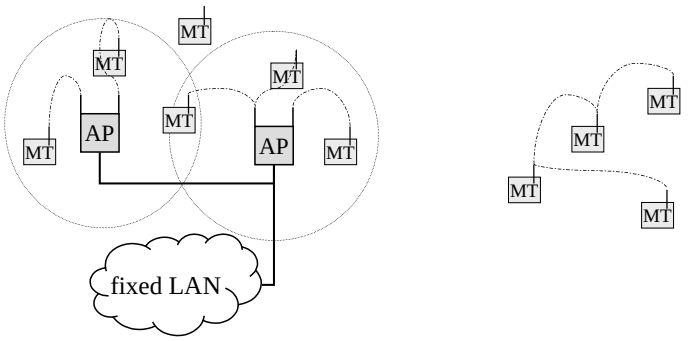

Abbildung 2: HiperLAN/2 Topologie (mit Access Point/Ad-Hoc)

Außerdem gibt es eine Ad-Hoc Betriebsart, welche die Verbindung zwischen mehreren MTs ohne AP ermöglicht.

Wenn ein MT mehrere APs empfängt, wird zunächst die Empfangsstärke gemessen, um dann den besten AP auszuwählen. Falls sich die Empfangssituation, aufgrund von Störeinflüssen oder durch die Bewegung des MT ändert, wird ein Handover durchgeführt. Bei einem Handover werden alle offenen Verbindungen des MTs zum AP an den anderen AP übergeben. Die Kommunikation zwischen den APs erfogt nicht über HiperLAN/2 sondern über das Festnetz (z.B.: Ethernet, Hiperaccess).

Die Verbindungen sind im Gegensatz zu anderen WLAN-Protokollen verbindungsorientiert, da HiperLAN/2 auch Quality of Service (QoS) unterstützt. QoS ist notwendig, um die Übertragung von Multimedia-Inhalten (z.B.: Video, Audio) sicherstellen zu können. Für solche Anwendungen bietet sich eine unidirektionale point-to-multipoint Verbindung an, welche vom AP in Richtung der MTs möglich ist. Außerdem existiert auch noch ein Broadcast-Kanal.

HiperLAN/2 bietet eine high-speed Übertragung, d.h. auf Schicht 1 (physikalische Schicht) ist eine Übertragung mit bis zu 54 Mbit/s möglich. Diese Leistung verringert sich aber aufgrund von Fehlerkorrektur und Protokoll-Overhead auf 25 Mbit/s. Solche Geschwindigkeiten sind nur mit Hilfe von modernsten Modulationsverfahren möglich, welche aber nicht unter allen Bedingungen einsatzfähig sind. HiperLAN/2 erlaubt es deshalb, das Modulationsverfahren während der Übertragung zu ändern. Dieses Verfahren wird im folgenden Abschnitt näher betrachtet.

Um Störungen auf der Funkstrecke zu vermeiden, ist es notwendig, dass APs in überlappenden Funkzellen verschiedene Frequenzen benutzen. Um sich eine aufwendige manuelle Planung, wie bei GSM, zu ersparen haben, die APs eine dynamische Frequenzzuordnung. Die APs passen sich so den vorherrschenden Gegebenheiten ständig an.

Die Verschlüsselung des Datenverkehrs ist bei einem WLAN natürlich notwendig, da jeder im Bereich einer Funkzelle den kompletten Datenverkehr mithören kann. HiperLAN/2 unterstützt dies durch die symmetrischen Verfahren DES und 3DES. Der Schlüsselaustausch findet nach dem Diffie-Hellmann Verfahren statt. Eine gegenseitige Authentifikation zwischen MT und AP ist entweder mit vorher ausgetauschten Schlüsseln oder mit Hilfe von public Keys möglich. Weitere Authentifizierungsverfahren, wie beispielsweise ein Verzeichnisdienst (X.509), sind in höheren Schichten möglich, jedoch nicht mehr im Standard spezifiziert.

Da ein WLAN für den Einsatz von mobilen Endgeräten konzipiert ist, ist eine Stromsparfunktion notwendig. Das MT kann zu jeder Zeit dem AP mitteilen, das es in den Schlafmodus wechselt. Der Schlafmodus besteht aus einzelnen Schlafperioden, nach deren Ablauf das MT auf eine Meldung des AP wartet, um danach wieder zu schlafen oder die vom AP zwischengespeicherten Daten in Empfang zu nehmen.

Der HiperLAN/2 Protokoll-Stack hat eine flexible Architektur, die eine Anbindung an verschiedene Netzprotokolle ermöglicht. HiperLAN/2 beschreibt die untersten zwei Schichten des ISO/OSI Basisreferenzmodells (physical-, DLC-Layer) sowie einen sogenannten Convergence Layer (CL). Der CL dient als Vermittlungsschicht zwischen dem WLAN und anderen Netzwerk-Protokollen.

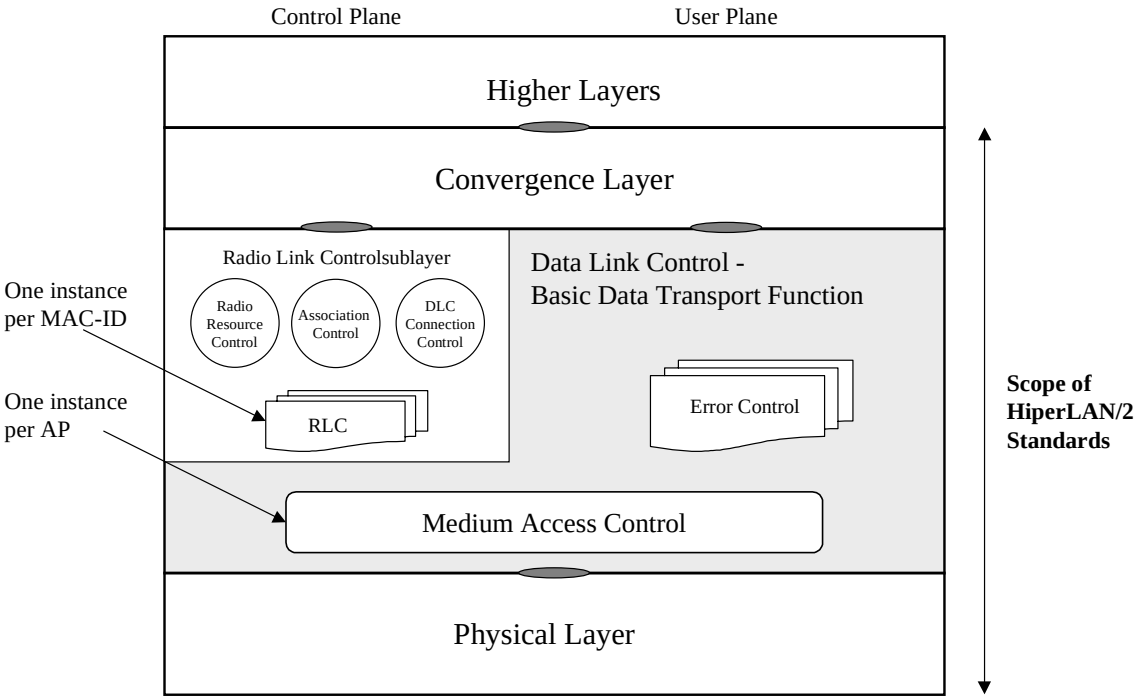

Abbildung 3: HiperLAN/2 Layer-Modell

Das Protokoll ist in Anlehnung an ISDN in ein Control- und ein User-Ebene unterteilt. Hierbei ist die User-Ebene für die Übertragung der Benutzerdaten zuständig (vgl. ISDN: B-Kanal), während die Control- Ebene (vgl. ISDN: D-Kanal) zur Kontrolle der

Verbindung eingeführt wurde. Bei HiperLAN/2 wird die Unterscheidung nur im Data Link Control Layer (DLC) gemacht.

Die drei Schichten werden nun näher beschrieben.

2.2 Physical Layer

Der physical Layer (PHY) sendet und empfängt die Daten und tauscht sie mit dem DLC aus. Er wurde in [BRAN00a] definiert. Zur Kommunikation sind HiperLAN/2 in Europa zwei Frequenzbänder zugeteilt. Das untere Band von 5150 bis 5350 MHz und das obere im Bereich 5470-5725 MHz, wobei die Sendeleistung im unteren Band nicht mehr als 200 mW betragen darf. Im oberen Band ist bis zu 1 W erlaubt. Das untere Frequenzband ist für den indoor-Einsatz gedacht, während das obere für den outdoor-Bereich bestimmt ist. Als Kanalabstand wurde 20 MHz gewählt. Dies ermöglicht einerseits eine schnelle Übertragung und andererseits 19 verschiedene Kanäle. Die Frequenzbänder und Sendestärken in den USA und Japan sind anders. In Amerika ist beispielsweise eine Sendeleistung mit bis zu 4 W erlaubt.

Die vom DLC erhaltenen Potokoll Data Units (PDU) werden durch eine Preample ergänzt und dann mit Hilfe des Orthogonal Frequency Division Multiplexing (OFDM) gesendet. Die grundlegende Idee bei OFDM ist eine breitbandige Übertragung durch Aufteilung des Signals in mehrere parallele Bit-Ströme, die jeweils einen Subcarrier modulieren. Dadurch wird das Kanalspektrum in nicht selektive Subkanäle aufgeteilt. Ein Vorteil von OFDM ist die Robustheit gegenüber Mehrwege-Ausbreitung mit Frequenzüberlagerung. Unter einer Mehrwege-Ausbreitung versteht man den Effekt, das ein Signal aufgrund von Reflexionen unterschiedliche Laufwege nehmen kann. Dies führt dann zu unterschiedlichen Laufzeiten und somit zur Überlagerung des Signals. Bei HiperLAN/2 wird ein Kanal in 52 Subcarrier aufgeteilt, von denen 48 zur Datenübertragung und 4 zur Dekodierung des Signals notwendig sind (z.B.: Bestimmung der Phasenlage). OFDM wurde erst durch den Einsatz von guten Signalen Prozessoren (DSP) und die Fast Fourier Transformation in Send- und Empfangsbausteinen effizient realisierbar. Dieses Verfahren wird inzwischen auch in einigen anderen Bereichen, wie beispielsweise bei ADSL und digitalen Radio (DAB), eingesetzt.

Mode	**Modulation**	**Code rate**	**PHY bit rate in Mbit/s**	**bytes/OFDM symbol**
1	BPSK	1/2	6	3,0
2	BPSK	3/4	9	4,5
3	QPSK	1/2	12	6,0
4	QPSK	3/4	18	9,0
5	16QAM	9/16	27	13,5
6	16QAM	3/4	36	18,0
7	64QAM	3/4	54	27,0

Tabelle 1: PHY-Modes

Ein Besonderheit von HiperLAN/2 ist die dynamische Link Adaption, die je nach Signalstärke und Fehlerfreiheit des Signals eines der 7 Modulationsverfahren auswählt.

2.3 Data Link Control Layer

Der DLC (siehe [ETSI00c]) baut eine logische Verbindung zwischen dem AP und dem MT auf und regelt den Zugriff auf das Medium. In dieser Schicht gibt es die oben erwähnte Trennung zwischen User- und Control-Ebene. Der DLC besteht aus mehreren Sublayern.

- Medium Access Control (MAC) Protokoll
- Error Control (EC) Protokoll
- Radio Link Control (RLC) Protokoll, mit den Unterpunkten DLC Connection Control (DCC), Radio Resource Control (RRC) und Association Control Function (ACF)

2.3.1 MAC-Protokoll

Das MAC-Protokoll wird benutzt, um auf das Funk-Medium zuzugreifen. Es basiert auf Time Division Multiple Access (TDMA) und Time Division Duplex (TDD). Dies ermöglicht eine simultane Benutzung des Zeitschlitzes, bei HiperLAN/2 als MAC-Frame bezeichnet, sowohl in Richtung des MT (downlink) als auch ein Richtung des AP (uplink). Ein MAC-Frame hat immer eine Länge von 2 ms.

Die komplette Kontrolle geht vom AP aus. Dadurch ist eine zentrale und eindeutige Vergabe der Kanäle in der Up- bzw. Downlink Phase gewährleistet. Dies ermöglicht auch eine dynamische Zuordnung.

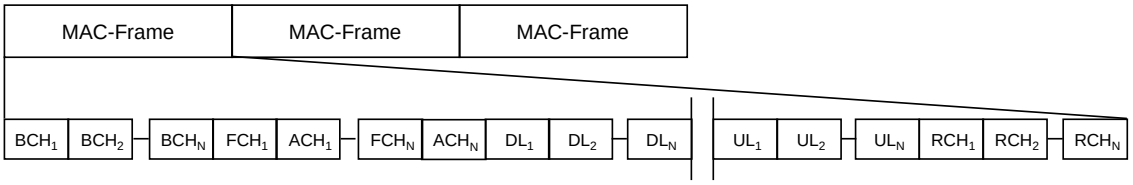

Abbildung 4: MAC-Frame

Zur besseren Strukturierung des MAC-Protokolls wurde es in Transportkanäle und logische Kanäle unterteilt.

Transport Kanal	Richtung	PHY mode	Länge (in octets)
BCH	Downlink	BPSK 1/2	15
FCH	Downlink	BPSK 1/2	n*27
SCH	Down-/Uplink	je nach FCCH	9
LCH	Down-/Uplink	je nach FCCH	54
ACH	Downlink	BPSK 1/2	9
RCH	Uplink	BPSK 1/2	9

Tabelle 2: Transportkanäle

Der Broadcast Channel (BCH) enthält Kontroll Informationen und wird in jedem MAC-Frame vom AP aus gesendet. Der Frame Control Channel (FCH) beschreibt die Länge

der Down- und Uplink Phasen. Über den Random Access Channel (RCH) teilen die MTs dem AP den Bedarf an Uplink Zeitschlitzen mit. Die Userdaten werden mit Hilfe des Short Transport Channel (SCH) und des Long Transport Channel (LCH) übertragen. Die einzelnen Transport Kanäle sind mit einem CRC versehen. Auf eine genaue Beschreibung der Blöcke wird hier verzichtet.

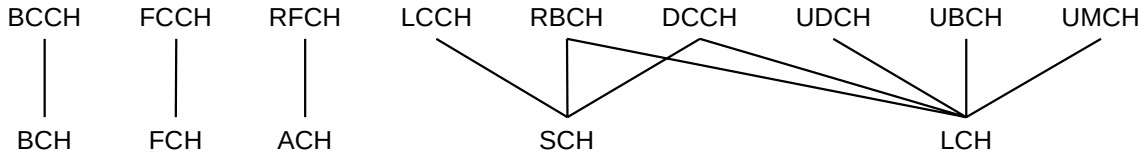

Abbildung 5: Mapping von logischen in Transport Kanäle - downlink

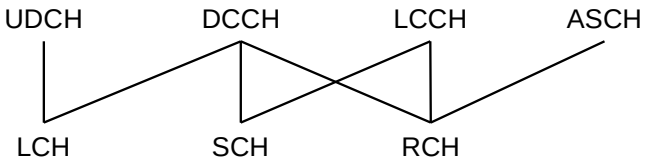

Abbildung 6: Mapping von logischen in Transport Kanäle - uplink

Der Zusammenhang zwischen logischen Kanäle und Transportkanälen wird in Abbildung 5 und 6 dargestellt.

logischer Kanal	**Richtung**
Broadcast Control Channel (BCCH)	downlink
Frame Control Channel (FCCH)	downlink
Random Access Feedback Channel (RFCH)	downlink
RLC Broadcast Channel (RBCH)	downlink
Dedicated Control Channel (DCCH)	up-/downlink
User Broadcast Channel (UBCH)	downlink
User Multicast Channel (UMCH)	downlink
User Data Channel (UDCH)	up-/downlink
Link Control Channel (LCCH)	up-/downlink
Association Control Channel (ASCH)	uplink

Tabelle 3: Logische Kanäle

Da HiperLAN/2 Broad-, Multi- und Unicast Übertragung unterstützt gibt es drei verschiedene Kanäle zur Übertragung von Nutzdaten. Dies sind der UBCH, UMCH und UDCH. Sie transportieren die eigentlichen Daten, wobei ein Broadcast oder Multicast nur in Richtung der MTs möglich ist. Für Multi- und Broadcast-Betrieb stehen, wie beim Ethernet, spezielle MAC-IDs zur Verfügung. Der BCCH enthält Kontrollinformationen. Der FCCH gibt Auskunft über die nachfolgende Pakete sowie ihre Modulationsart. Die RFCH und ASCH Kanäle dienen zur Anforderung von Übertragungskapazität bzw. zu deren Bestätigung. Die Fehlerkorrektur (EC) tauscht Informationen über den LCCH aus. Die DCCH und RBCH enthalten Kontrollinformationen über die Verbindung.

Für den MAC-Durchsatz ergibt sich nun folgende Formel

$$Throughput_{MAC} = \lfloor \frac{L_{LCH}}{\lceil \frac{54}{BpS_{LCH}} \rceil} \rfloor * \frac{48 * 8}{2ms}$$

L_{LCH} entspricht der Anzahl der möglichen OFDM-Symb. im MAC-Frame und BpS_{LCH} der Anzahl der übertragbaren Bytes pro OFDM-Symbol. Die Übertragungsrate hängt hierbei hauptsächlich von der Modulation ab.

2.3.2 Error Control

Zur Sicherung der Verbindung wurde das Error Control (EC) Protokoll spezifiziert. Das EC kennt folgende drei Funktionsweisen:

- *Acknowledged Mode:* Jedes LCH-Paket muss in diesem Modus durch ein ACK-Paket vom Empfänger bestätigt werden. Dieser Modus wird zur Sicherung von Unicast-Paketen (UDCH) eingesetzt, da bei einer Broadcast- oder Multicast-Übertragung zu viel Overhead entstehen würde. Wenn die Bestätigung ausbleibt, wird das Paket nach einem time-out erneut gesendet. Der Empfänger kann bei erkannten Fehlern auch eine Neusendung durch ein NACK-Paket veranlassen.
- *Repetition Mode:* Der Sender sendet hierbei das LCH-Paket mehrfach aus, um durch Redundanz eine geringere Verlustrate zu erreichen. Auf diese Weise ist auch eine Sicherung von Broadcast-Übertragungen möglich. Für Multicast ist dies nicht möglich, da mehrere Multicast-Übertragungen zu viel Last erzeugen würden und für eine Unicast-Verbindung nicht sinnvoll.
- *Unacknowledged Mode:* Dieser Modus bietet keinen Schutz gegen Verlust von Paketen. Die Netzlast ist hierbei allerdings auch geringer, da weniger Kontrollinformationen gesendet werden. Es kann bei allen drei Verbindungsarten eingesetzt werden.

Zur Sicherung von Multicast-Paketen wurde der sog. n*Unicast Modus definiert. Dieser ist einer Unicast-Übertragung gleich zusetzen und nutzt somit den Acknowledged Mode. Weitere Funktionen des EC sind die Flusskontrolle und Generierung bzw. Überprüfung der CRCs. Die Flusskontrolle stellt die sequentielle Weitergabe der Pakete an den Convergence Layer durch Nummerierung sicher. CRCs sind notwendig, um die Integrität der einzelnen Pakete zu sichern.

Die oben genannte Formel ist somit für den Durchsatz des DLC nur bedingt richtig. Es gilt nämlich:

$$Throughput_{DLC} = Throughput_{MAC} * (1 - PER_{PHYMode})$$

$PER_{PHYMode}$ stellt die Fehlerrate mit dem selektierten PHY-Modus dar. Dabei wird deutlich, dass immer eine Anpassung des PHY-Modus notwendig ist, da sonst die Fehlerrate steigen kann und trotz höherer physikalischer Übertragungsrate der Gesamtdurchsatz sinkt.

2.3.3 Radio Link Control

Das RLC Protokoll regelt mit Hilfe der Association Control Function (ACF), der Radio Resource Control Function (RRC) und der DLC User Connection Control Function (DCC) die Kommunikation zwischen MT und AP.

Der Verbindungsaufbau und Abbau ist in der ACF definiert und geht folgendermaßen von statten. Das MT startet den Aufbau, indem es die Signalstärken der einzelnen APs misst. Nun teilt das MT dem AP mit der besten Signalstärke den Verbindungswunsch, über den BCCH, mit. Der AP vergibt dann zunächst eine MAC-ID, welche für diese Funkzelle eindeutig sein muss. Danach tauschen AP und MT die unterstützten PHY-modes und Convergence Layers aus. Die notwendige Authentifizierung und Verschlüsselung wird auch vereinbart. Falls Verschlüsselung gewünscht ist, werden nun zunächst mit dem Diffie-Hellmann Verfahren die Schlüssel ausgetauscht, die eine mögliche folgende Authentifizierung auch sichern. Als Verschlüsselungsverfahren wird DES oder 3DES benutzt. Nun ist eine einseitige oder gegenseitige Authentifizierung mit Hilfe von vorher ausgetauschten Schlüsseln möglich. Eine Public Key Infrastruktur wird zwar unterstützt, setzt aber Funktionen höherer Schichten voraus. Somit ist auch eine Authentifizierung nach MD5, HMAC und RSA möglich. Das MT kann nun einen eigenen DCCH anfordern und DLC User Verbindungen anfordern, wobei für eine einzelnen DLC-Verbindungen immer nur eine Qualtity of Service Unterstützung (z.B.: min. 64kbit/s Bandbreite) gelten kann. Der Verbindungsabbau findet entweder explizit durch eine Nachricht des MT an den AP statt oder implizit durch Überschreiten eines time-out.

Die DCC regelt die Kommunikation innerhalb eines DCCH. Die RRC unterstützt Handover, dynamische Frequenzzuordnung und Power Save. Das Handover findet statt, wenn das MT einen anderen AP besser empfängt. Die vorhanden Verbindungen gehen dabei auf den neuen AP über. Die Informationen zwischen den APs müssen über eine feste Verbindung ausgetauscht werden. Die dynamische Frequenzzuordnung kann MTs beauftragen, Messungen über die von benachbarten APs verwendeten Frequenzen, durchzuführen. Darauf hin kann der AP den MTs einen Frequenzwechsel mitteilen.

2.4 Convergence Layer

Der Convergence Layer (CL) wurde in HiperLAN/2 definiert ([BRAN00b]), um eine Schnittstelle zwischen dem DLC-Layer und den höheren Layern eines anderen Netzwerkstandards zu schaffen. Auf die Definition höherer Schichten wurde in HiperLAN/2 verzichtet, da es nur den Last-Hop einer Verbindung bewerkstelligen soll. Somit sollte es für die Vermittlung verschiedenster Netzwerk-Protokolle offen sein, was wiederum zu einer hohen Komplexität in den höheren Schichten geführt hätte. Außerdem müßten die Daten durch mehr Schichten wandern, was zu einer Verzögerung der Übertragung geführt hätte. Der CL bildet somit die Eigenschaften verschiedener Netztypen auf die Merkmale von HiperLAN/2 ab.

Genau genommen gibt es zwei CL, zum einen ist dies ein Zell-basierter-CL und zum anderen ein Paket-basierter-CL. Der Zell-basierte-CL wurde zu Unterstützung zellenbasierter Netze, wie beispielsweise ATM, geschaffen. Diese haben sehr kurze Rahmen-Längen, um auch kleine Datenströme ohne große Verzögerungen übertragen zu können.

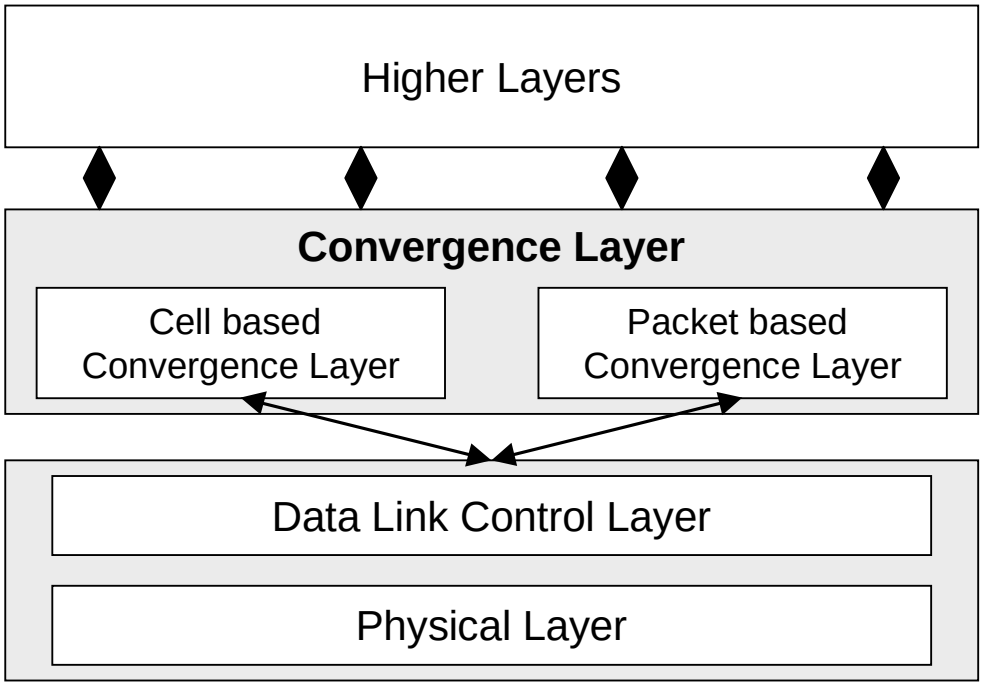

Abbildung 7: Convergence Layer ([ETSI00d])

Der Paket-basierte-CL eignet sich zur Umsetzung typischer Paket-orientierter Netze, wie Ethernet bzw. IP. Diese zwei CL allein bieten jedoch noch keine Schnittstelle zu der nächst höheren Schicht, sondern nur zum DLC. Hierfür gibt es noch sog. Service Specific Convergence Sublayer (SSCS), welche für jeden Netztyp verschieden sind.

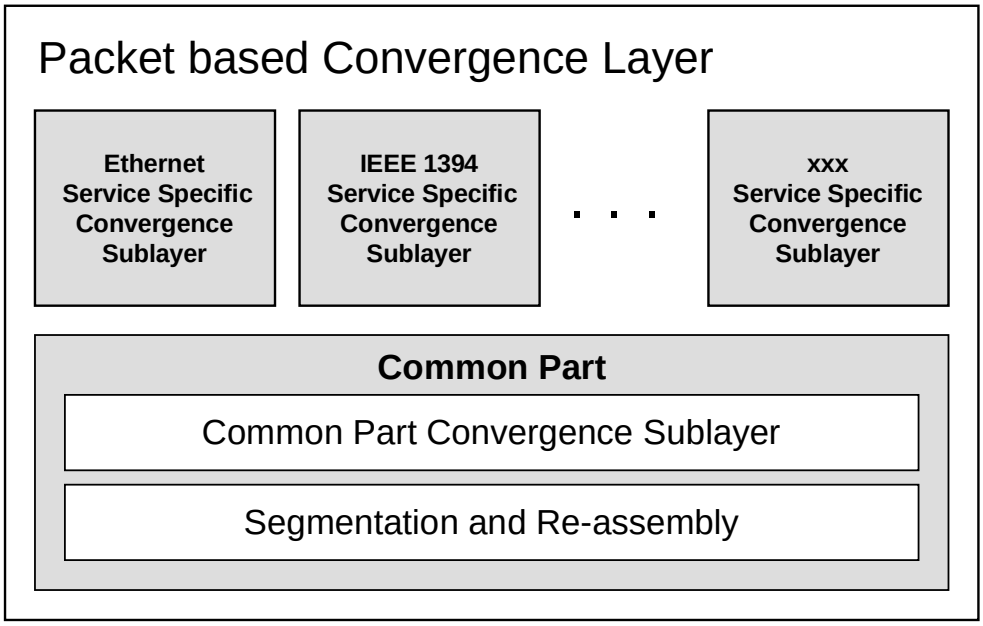

Abbildung 8: Paket basierter Convergence Layer ([ETSI00a])

Der allgemeine Teil des Paket-basierten-CL ist für die Teilung der vom SSCS gelieferten Pakete und die Zusammensetzung der vom DLC empfangenen Pakete zuständig. Die SSCS sorgen dann für die vollständige Transparenz des HiperLANs gegenüber dem festen Netz. SSCS gibt es für Ethernet, IP und IEEE 1394 (Firewire).

Der Zell-basierte-CL wurde geschaffen um die Merkmale von ATM und UMTS unterstützen zu können. HiperLAN/2 kann somit UMTS in sog. Hot Spots unterstützen. Hot Spots sind Bereiche, in denen eine sehr viel Datenverkehr stattfindet, wie beispielsweise bei Messen und in Hotels. Eine Versorgung durch HiperLAN/2 könnte aufgrund der über 10-mal größeren Bandbreite Entlastung schaffen und gleichzeitig neue Dienste verfügbar machen.

2.5 weitere Features

2.5.1 Ad Hoc

HiperLAN/2 kann in zwei Modi betrieben werden. In der bisher beschriebenen Betriebsart ging die Kontrolle immer von einem AP aus. Es ist aber auch ein Ad-Hoc Betrieb möglich (siehe [Peet00]). Bei dieser Betriebsweise gibt es keinen AP, sondern nur MTs, welche untereinander kommunizieren wollen. Der eigentliche Betrieb funktioniert genauso wie mit einem AP, wobei ein MT als Central Controller (CC) fungieren muß. Der CC wird durch ein dynamisches Selektionsverfahren festgelegt. Wenn ein MT CC-fähig ist, versucht es zunächst, sich einem bestehenden Subnetz anzuschließen, mißlingt dies aufgrund der Authentifizierung oder weil keine Geräte vorhanden sind, versucht es, selbst CC zu werden. Es sendet dann mehrere sog. Probing Frames über den BCCH. Danach scannt es die anderen Frequenzen nach weiteren MTs, die auch CC werden wollen. Wenn ein anderes MT entdeckt wird, tritt es für eine gewisse Zeit zurück und versucht dann, durch einen erneuten Scan sich diesem anzuschließen. Ein MT ist dann CC, wenn es während des kompletten Selektions-Prozesses, der aus 10 Probing-Phasen besteht, kein anderes MT registriert hat. Die Wahrscheinlichkeit, dass es mehr als ein CC gibt, ist relativ gering. Dies kann nur dann vorkommen, wenn die Scan-Phasen der MTs zur gleichen Zeit stattfinden. Die rechnerische Wahrscheinlichkeit hierfür ist kleiner $2 * 10^{-5}$.

2.5.2 Quality of Service

Unter Quality of Service (QoS) versteht man die Unterstützung von garantierten Bandbreiten und Verzögerungszeiten für einzelne Datenströme. Dies ist bei der Übertragung von Audio- (z. B.: Telefon) und Videodaten notwendig. Aufgrund der verbindungsorientierten und zentralisierten Struktur von HiperLAN/2 ist es möglich, QoS Parameter anderer Protokolle wie ATM zu garantieren. Dies gilt bei einem Funkmedium natürlich nur mit Einschränkungen, da es von vielen Dingen gestört werden kann. Jedoch kann der AP mit Hilfe eines guten Scheduling der Up- bzw. Downlink Kanäle die Einhaltung der Parameter erreichen.

2.6 HiperLAN/2 vs. IEEE 802.11a

Die IEEE spezifiziert momentan die Weiterentwicklung des 802.11 Standards. Der neue Standard 802.11a soll ähnliche Bandbreiten wie HiperLAN/2 bieten. Außerdem sendet er im gleichen Frequenzband. HiperLAN/2 hat aber einige Vorteile. Aus Sicherheitsaspekten sind hier die fehlende Authentifizierung und schwache Verschlüsselung durch

40-bit RC4-Schlüssel zu nennen. Die Verbindung mit anderen Netzen ist nur für Ethernet standardisiert und somit nicht modular aufgebaut. Außerdem fehlen wichtige Eigenschaften, wie QoS Unterstützung und dynamische Frequenzzuordnung.

2.7 Performance

Die Performance eines WLAN hängt stark von den Gegenheiten ab. Beispielsweise werden WLANs die im ISM-Band (2,4 GHz) senden, von Mikrowellen-Geräten und Bluetooth-Geräten gestört. Außerdem spielen funkspezifische Probleme wie Interferenz durch Multi-Pfad-Ausbreitung eine Rolle. Die Anzahl der beteiligten MTs ist dabei auch von entscheidender Bedeutung, da sie einen Protokoll-Overhead erzeugen, der netto Datenrate reduziert. Daher sind Meßwerte relativ ungenau und meist nicht für konkrete Szenarien berechenbar.

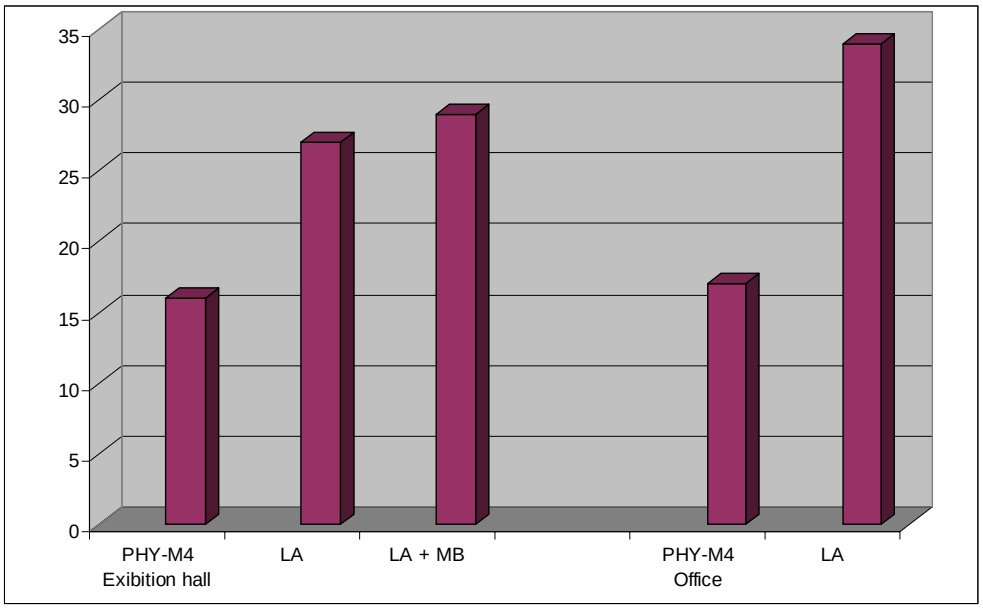

Abbildung 9: Performance in Mbit/s ([John99])

Trotzdem wurden natürlich Versuchsmessungen durchgeführt. Bei einem Versuch in einer Ausstellungshalle zeigte sich, wie auch beim zweiten Versuch in einem Büro Gebäude, die Leistungsfähigkeit der Link Adaption, welche im Vergleich zum festen PHY-Mode 4 einen Perfomance Zuwachs von ca. 15 Mbit/s auf 25 Mbit/s bringt. Außerdem war festzustellen, dass der Einsatz von Multibeam-Antennen den Zuwachs weiter vergrößert.

	BPSK 1/2	**BPSK 3/4**	**QPSK 1/2**	**QPSK 3/4**
1 RCH	30	39	48	60
3 RCH	27	39	45	57
5 RCH	27	36	45	57
7 RCH	27	36	42	54

Tabelle 4: max. Anzahl von MTs bei Sprach-Übertragungen in einer Zelle ([LeMi00])

Tabelle 4 zeigt mit wie vielen MTs eine Sprach-Kommunikation gleichzeitig möglich ist. Die einzelnen Voice-Streams wiesen dabei eine Bandbreite von 64 kbit/s auf. Die Abhängigkeit von den Random Access Channel (RCH) ist relativ gering, während der Zusammenhang zwischen Modulations-Verfahren und Anzahl der MTs erwartungsgemäß groß ist. Die max. Anzahl wird nicht durch die Bandbreite begrenzt, sondern durch die Verzögerungszeit bestimmt.

3 Ausblick

Die wichtigsten Teile des Standards wurden im Frühjahr 2000 verabschiedet und veröffentlicht. Allerdings fehlen immer noch Teile, wie beispielsweise die Spezifikation des IEEE 1394 Sublayers und die Network Management Spezifikation. Es sind auch weitere Konzepte, wie Forwarding ([EsVW00]), in Bearbeitung. Das Forwarding soll eine Erweiterung der Funkzelle ohne zusätzliche APs ermöglichen, indem ein forward-fähiges MT einem MT außerhalb der Reichweite des AP die Daten weiterleitet. Bei diesem Verfahren gibt es allerdings Probleme durch zusätzlichen Datenverkehr und größere Verzögerungszeiten, was die Einhaltung von QoS Parametern erschwert. HiperLAN/2 wird durch das HiperLAN/2 Global Forum (H2GF, [H2GF00]) unterstützt. Es wurde von Bosch, Dell, Ericsson, Nokia, Telia and Texas Instruments gegründet und hat sich zum Ziel gesetzt die Interoperabilität, Anwendungen und das Marketing von HiperLAN/2 voranzutreiben. Inzwischen sind dem H2GF ca. 40 Firmen beigetreten, darunter in der Mobilfunk-Branche führende Firmen, wie Siemens, Lucent, NTT. Teilweise sind diese Firmen auch an der endgültigen Spezifikation beteiligt.

Aufgrund des großen industriellen Engagement sind die Chancen für eine schnelle Markteinführung gut. Ericsson hat im Dezember 2000 einen Prototyp vorgeführt der schon beinahe serienreif war. Die Steckkarten für ein MT (Notebook) besaßen schon PCMCIA-Maße und waren einsatzfähig. Mit fertigen Produkten wird im vierten Quartal 2001 gerechnet.

Im November 2000 wurde zudem beschlossen die 5 GHz Globalisation Study Group zu gründen. Diese soll die Aktivitäten der verschiedenen WLAN Standards vereinen. Neben HiperLAN/2 und IEEE 802.11a gibt es in Japan einen dritten Standard, der MMAC. Unabhängig davon sind auch Bestrebungen im Gange, die Frequenzbänder der ETSI und der IEEE zu harmonisieren.

4 Resümee

HiperLAN/2 zählt, mit seinen zukunftsorientierten Funktionen, zu den besten WLAN-Standards. Es muß keine Kompatibilität mit alten Standards bieten und wurde daher ohne Kompromisse entwickelt, was sich auch in der Unterstützung durch die Industrie zeigt.

Literatur

[AHKM00] Kurt Aretz, Martin Haardt, Walter Konhäuser und Werner Mohr. The Future Of Wireless Communications. Dresden, Germany, 2000. Proceedings of European Wireless 2000.

[BRAN00a] BRAN (Hrsg.). ETSI TS 101 475 V1.1.1, PHY-Layer. technische spezifikation, The European Telecommunications Standards Institute, April 2000.

[BRAN00b] BRAN (Hrsg.). ETSI TS 101 683 V1.1.1, System Overview. technische spezifikation, The European Telecommunications Standards Institute, Februar 2000.

[EsVW00] Norbert Esseling, Harbinder S. Vandra und Bernhard Walke. A Forwarding Concept For HiperLAN/2. Dresden, Germany, 2000. RWTH Aachen, Proceedings of European Wireless 2000.

[ETSI00a] ETSI (Hrsg.). ETSI TS 101 493-1 V1.1.1, Packet Based Convergence Layer, Part 1. technische spezifikation, The European Telecommunications Standards Institute, April 2000.

[ETSI00b] ETSI (Hrsg.). ETSI TS 101 493-2 V1.1.1, Packet Based Convergence Layer, Part 2 Ehternet. technische spezifikation, The European Telecommunications Standards Institute, April 2000.

[ETSI00c] ETSI (Hrsg.). ETSI TS 101 761-1 V1.1.1, DLC-Layer. technische spezifikation, The European Telecommunications Standards Institute, April 2000.

[ETSI00d] ETSI (Hrsg.). ETSI TS 101 763-1 V1.1.1, Cell Based Convergence Layer, Part 1. technische spezifikation, The European Telecommunications Standards Institute, April 2000.

[H2GF00] H2GF. HiperLAN/2 Global Forum. www.hiperlan2.com, 2000.

[John99] Martin Johnsson (Hrsg.). HiperLAN/2 - The Broadband Radio Transmission Technology Operating In The 5 GHz Frequency Band. White paper, The HiperLAN/2 Global Forum, 1999.

[KaMa00] Arndt Kadelka und Arno Masella. Serving IP Quality Of Service With HiperLAN/2. Dresden, Germany, 2000. RWTH Aachen, Proceedings of European Wireless 2000.

[LeMi00] Prof. Luciano Lenzini und Dr. Eng. Enzo Mingozzi. Performamce Evaluation Of HiperLAN/2 Bandwith Allocation. Dresden, Germany, 2000. University of Pisa, Proceedings of European Wireless 2000.

[Peet00] Jörg Peetz. HiperLAN/2 Ad Hoc Network Configuration By CC Selection. Dresden, Germany, 2000. RWTH Aachen, Proceedings of European Wireless 2000.

Abbildungsverzeichnis

Tabellenverzeichnis

GPRS (General Packet Radio Service)

Robert Soukup

Kurzfassung

Diese Seminararbeit beschäftigt sich mit dem Thema GPRS (General Packet Radio Service). Sie entstand aus den gewonnenen Erkenntnissen eines viermonatigen Praktikums bei der Firma 12snap in München. Ziel war es, zu untersuchen, inwieweit das Geschäftsmodell, auf das später noch näher eingegangen werden soll, durch GPRS beeinflusst wird. Daher gliedert sich die Arbeit in zwei Teile. Der erste Abschnitt wird GPRS unter dem Gesichtspunkt der technischen Neuerungen betrachten, sozusagen die Perspektive der Netzwerkbetreiber. Der zweite, etwas kürzere Teil wird GPRS aus den Augen eines Serviceanbieters beleuchten. Hierbei wird zuvor auf das Beispiel von i-mode eingegangen, einem System, das sich in Japan schon großer Beliebtheit erfreut und GPRS sehr ähnlich ist. Darauf aufbauend wird auf die Kombination WAP und GPRS ausführlicher Bezug genommen. Mit den konkreten Implikationen der Firma 12snap wird dieser zweite Hauptteil abgerundet.

1 GPRS - Die Technik

1.1 Paketbasierte Datenübertragung

Die Wurzeln von GPRS gehen bis in die Mitte der 90er Jahre zurück. Schon damals erkannte man, dass Circuit-Switched-Verkehr, also die Datenübermittlung über eine fest geschaltete Verbindung, nicht für jede Informationsübertragung geeignet ist. Während es für Sprache prinzipiell keinen günstigeren Weg gibt, erwies sich eben diese Technik als nicht sonderlich elegant für die Übermittlung von Daten. Denn die Beanspruchung von Ressourcen ist unabhängig von der tatsächlichen Auslastung. Das heißt, die generell knappen Funkkanäle werden nicht wirklich effektiv ausgelastet.

Die Übertragung von Daten in Paketform erweist sich dagegen als überaus geeignet, kleine, zeitlich unregelmäßig auftretende Datenmengen durch Netzwerke zu versenden. Das Internet ist wohl das beste Beispiel dafür. GPRS sollte daher als ein Instrument gesehen werden, welches einerseits dem Netzwerkbetreiber eine effektivere Auslastung seiner physikalisch limitierten Infrastruktur ermöglicht und zweitens dem Endkunden ein seinen weiteren Nutzungsanforderungen besser entsprechendes Übertragungsmedium zur Verfügung stellt.

1.2 Paketübermittlung in anderen Mobilfunksystemen

GPRS setzt auf Mobilfunksystemen, die nach dem sogenannten Zeitscheibenverfahren (TDMA, Time Division Multiple Access) arbeiten, auf. Die beiden wohl bekanntesten, digital arbeitenden sind D-AMPS (Digital Advanced Mobile Phone System) und GSM (Global System for Mobile Communications). Beide Systeme decken fast zwei Drittel des Weltmarktes ab. Zu der Familie der TDMA-basierten System zählt noch das japanische PDC System (Personal Digital Cellular). Auch hier gibt es wieder eine zugehörige Paketübertragungstechnik. Das System heißt PDC-P, ist aber besser unter dem Markennamen "i-mode" bekannt. Auf i-mode soll später noch ausführlicher eingegangen werden, weil dieses System bereits seit Februar 1999 in Betrieb ist und bisher von über 16 Millionen, das heißt einem Drittel aller Mobilfunkkunden der NTT Tochter DoCoMo, genutzt wird. Ob dieser Erfolg GPRS beschieden sein kann, soll ebenfalls mit diskutiert werden.

AT&T in den USA betreibt bereits ein auf Paketübertragung basierendes System, welches CDPD heißt (Cellular Digital Packet Data). CDPD setzt ebenfalls auf der AMPS-Technik auf, ist aber in seinen technischen Möglichkeiten nicht wirklich mit GPRS vergleichbar. Zum Beispiel können Datenpakete nur versandt werden, wenn der Teilnehmer gerade kein Telefonat führt. Die theoretisch maximale Geschwindigkeiten ist auch nur 19.2 kbps. Ein Vorteil ist allerdings, dass das System bereits in über 60 Großstädten der USA angeboten wird und daher ein gewisser Erfahrungsschatz mit dieser Technik gesammelt werden konnte.

Aber auch grundlegend andere Mobilfunkstandards haben ihre zugehörige Paketübermittlungstechnik. CDMA (IS-95, Code Division Multiple Access), auch "Bandspreizverfahren" genannt, vergibt die Ressourcen, indem es über einen speziellen Zahlencode eine Funkverbindung für einen Teilnehmer exklusiv von anderen Teilnehmern unterscheidet. Dieses Verfahren erlaubt es, das gesamte Funkspektrum von allen Teilnehmern gleichzeitig zu nutzen. Auch CDMA hat bereits ein funktionsfähigen Paketaufsatz, der von Sprint PCS in den USA Endkunden angeboten wird. Das sogenannte CDMAone ist technisch nicht so ausgereift wie GPRS, konnte sich aber in der Praxis schon mehrere Jahre bewähren.

Insgesamt muss eingestanden werden, dass CDMA TDMA in vielen Belangen überlegen ist, obwohl ebenfalls Probleme in der Praxis auftreten. So zum Beispiel das Phänomen der "Zellatmung", was plastisch beschreibt, dass die Größe der Zelle abhängig von der Teilnehmerzahl ist. Grund dafür ist, dass die Leistung der Basisstation auf die Anzahl der Endgeräte aufgesplittet wird. Trotzdem wird diese Technologie die Basis des in Westeuropa einzuführenden UMTS-Standards (Universal Mobile Telephone Service), dem Wideband-CDMA, bilden. Dafür wurden bereits 1997 bei der ETSI (European Telecommunication Standard Institute) die Weichen gestellt.

1.3 Die Technik aus Sicht von GSM

Da in Deutschland seit dem Abschalten des C-Netzes am 31.12.2000 kein alternatives System zu GSM für den breiten Massenmarkt mehr existiert, werden sich die folgenden Diskussionen auf GSM, dem Global System for Mobile Communication, beschränken.

Die Abbildung 1 zeigt vereinfacht die Struktur eines um GPRS erweiterten GSM-Netzwerks. Das GSM-Netzwerk besteht im Grunde aus sieben Hauptkomponenten.

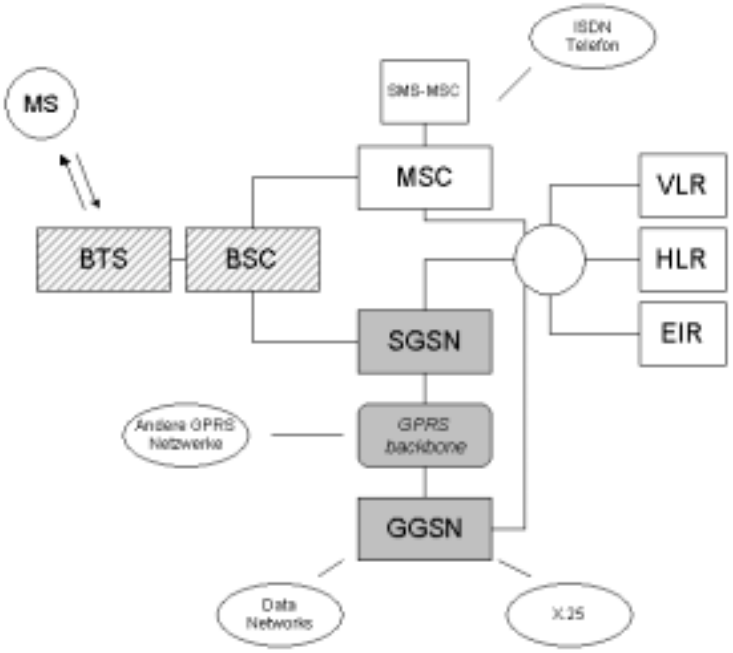

Abbildung 1: GSM-Netzwerk mit GPRS Erweiterung (grau).

Ausgehend von der BTS (der Base Transceiver Station), die im Wesentlichen nichts anderes ist als eine Antenne, wird das Signal zu einem BSC (Base Station Controller) weitergeleitet. Ein BSC ist in der Regel für mehrere BTS zuständig. Von da aus weitergehend wird das Signal im MSC, dem Mobile Switching Centre, verarbeitet. Ein MSC kann ungefähr eine Million Nutzer gleichzeitig betreuen, was die herausragende Stellung des MSC im Netzwerk verdeutlicht.

Das angeschlossene Home Location Register (HLR), bzw. das Visitor Location Register (VLR) enthalten Daten über netzzugehörige Nutzer bzw. roamende Nutzer aus anderen Netzwerken, die sich im Einzugsbereich des entsprechenden MSC befinden. Das EIR (Equipment Identity Register) ist eine zentrale Einheit, die Informationen über Nutzer und Mobiltelefone speichert. Die EIRs der Netzwerkbetrieber werden regelmäßig über einen Zentralrechner in Dublin synchronisiert. Ziel ist es unter anderem, gestohlene Mobiltelefone aufzuspüren. Der SMS-MSC ist für die Versendung und den Empfang der SMS (Short Message Service) zuständig.

Wie leicht aus der Abbildung zu entnehmen ist, bestehen weitere Verbindungen in entsprechende Netzwerke, auf die hier nicht weiter eingegangen werden soll.

Die bestehenden GSM-Infrastruktur wird nun um drei Hauptkomponenten erweitert.

- Der Gateway GPRS Support Node (GGSN) dient als Schnittstelle zu externen Netzen. Hier werden die Paketdatenprotokoll-Adressen ausgewertet und auf die IMSI (International Mobile Station Identity) der jeweiligen Mobilstation umgesetzt. Die Datenpakete werden hier entkapselt und entsprechend den Optionen des Netzprotokolls an die nächste Instanz der Netzschicht versendet.

- Der Serving GPRS Support Node (SGSN) dient zur funktionalen Unterstützung der Mobilstationen. Hier werden z.B. die Adressen der Teilnehmer eines Grup-

penrufes aus den GPRS-Registern (GR) abgefragt. Die Funktion des SGSN und des GGSN können auch in einer Einheit realisiert werden.

- Alle GPRS-bezogenen Daten werden im GPRS-Register (GR) gespeichert, welches man als Teilbereich des HLR ansehen kann [Walk00].

Wie die Abbildung 1 auch leicht erkennen lässt, wird der gesamte Datenverkehr im GPRS über den SGSN abgewickelt. Letztlich stehen sich hier MSC und SGSN auf einem Niveau ebenbürtig gegenüber, wobei der MSC nach wie vor für den Cicuit-Switched-Verkehr zuständig ist und jeglicher Paketverkehr über die SGSNs abgewickelt wird. Um beide Einheiten koordiniert zu verwalten, muss eine Signalleitung vom SGSN zum MSC oder zu mehreren MSCs, bzw. von mehreren SGSNs zu einem MSC gelegt werden.

1.4 Routing einer Datenübertragung initiiert von der Mobilstation (MS)

Initiiert eine MS eine Datenübertragung, wird diese über BTS und BSC (zusammen BSS, Base Station Subsystem) zu einem entsprechenden SGSN geleitet. Die Daten werden im SGSN entkapselt und gemäß ihrer Adressinformation zum entsprechenden GGSN geroutet. Dieser wiederum schickt die Daten in das entsprechende Packet Data Network (PDN) weiter. Ein GGSN ist also nichts weiter als ein Router.

Die Antwort aus dem PDN (z.B. Internet) wird nun von einem GGSN empfangen. Dieser prüft den Routingkontext, der dieser Zieladresse zugeordnet ist, und fragt den entsprechenden SGSN ab. Das Paket wird danach gekapselt und zum SGSN gesendet, der wiederum über das BSS die Information an die MS weitersendet.

Die Kapselung der Daten vom SGSN zum GGSN und umgekehrt sowie zwischen einzelnen GGSN ist notwendig, um alle Paketdatenprotokolle zuzulassen, selbst wenn sie vom SGSN nicht unterstützt werden. Untereinander kommunizieren die GSNs (GPRS Support Nodes) über das sogenannte GPRS Tunnel Protocol (GTP). Die Kapselung scheint aufwendig, macht GPRS jedoch offen für jegliche externe Paketstandards.

Abschließend mögen noch folgende Bemerkungen den technischen Rahmen einer GSM-Netzwerkerweiterung um GPRS erläutern:

Ein wichtiger Punkt wird nämlich zu häufig in der Literatur übersehen, weil er scheinbar selbstverständlich zu sein scheint - die Nutzung der alten BSS-Infrastruktur.

Für alle Betreiber gilt, dass mindestens ein Software-Update zu erfolgen hat. Dieser kann problemlos bei laufendem Betrieb erfolgen. Problematischer dagegen sind Hardwareveränderungen. Ein Unternehmen, das bundesweit seinen Dienst anbietet, verfügt über mehrere tausend BTS (D1 besitzt 39'000 BTS an 16'000 Standorten [gol00]). Dieser eventuelle Umbau würde ganz erhebliche Kosten nach sich ziehen. D1 und D2, die zu den Pionieren im GSM-Mobilfunk zählen, hatten hier unterschiedliche Voraussetzungen. Während D2 Vodafone mit Ericsson-Infrastruktur mit einem reinen Software-Update auskam, musste D1 einige seiner BTS komplett austauschen.

Obwohl mir keine exakten Angaben vorliegen, so interessiert vielleicht, daß D1 ca. 300 Millionen DM für den Ausbau von GPRS investieren musste. Das entspricht letztlich weniger als 25 DM pro potentiellem Kunde [gol00].

1.5 Mobiltelefone

Um die Möglichkeiten von GPRS, auf die ausführlicher noch später eingegangen werden soll, auch nutzen zu können, sind neue mobile Endgeräte erforderlich. Bis Ende November letzten Jahres hatten nur drei Hersteller Alcatel, Sagem und Ericsson GPRS-taugliche Geräte vorgestellt.

Grundsätzlich können die MS in drei Kategorien untergliedert werden.

- Funktelefonklasse A umfasst alle Mobiltelefone, die simultan Paket- und Circuit-Switched-Verkehr bearbeiten können. (Datenübertragung an einen Laptop wird nicht unterbrochen bei eingehendem Telefonat, sofern entsprechende Funkressourcen verfügbar sind).

- Zur Funktelefonklasse B gehören alle Telefone, die automatisch zwischen beiden Modi wechseln können. Eine Datenübertragung kann nach einem Telefonat wieder automatisch aufgenommen werden. Diese Kategorie wird vermutlich von den Herstellern am ehesten favorisiert, weil sie im Vergleich zur Klasse C einen guten Kompromiss von technischem Aufwand und Anwenderfreundlichkeit darstellt.

- Klasse C umfasst die Mobiltelefone, die manuell von einem Modus in den anderen umgeschaltet werden müssen. D.h. eingehende Daten werden abgewiesen, wenn der Benutzer gerade ein Gespräch führt.

Eine ganz entscheidende Fähigkeit von GPRS wurde bisher noch nicht angesprochen, nämlich die Möglichkeit, einem einzelnen Nutzer mehrere Funkkanäle zur Verfügung zu stellen. Ein kleiner GSM-Exkurs kann leichter begreiflich machen, wie das Grundprinzip der Funkkanalaufteilung funktioniert.

GSM arbeitet wie bereits angesprochen nach einem Zeitscheibenverfahren, indem es eine Frequenz für Bruchteile von Sekunden abwechselnd mehreren Nutzern zur Verfügung stellt. Dies ist auch korrekt innerhalb einer Zelle. Mehrere Zellen unterscheiden sich jedoch nach unterschiedlichen Frequenzen. Je nach Frequenz (D1, D2 arbeiten mit 900 MHz; E-Plus und Viag mit 1800 MHz.) teilen sich Uplink (MS zu BTS) und Downlink (BTS zu MS) kleine Frequenzbänder von 25 MHz bzw. 75 MHz. Diese werden wiederum in Carrier von je 200 KHz unterteilt. Das ergibt 124 Carrier für die D-Netze und 374 für die E-Netze. Zusätzlich wird jeder Carrier wieder in acht Zeitschlitze unterteilt, was das GSM-System im Grunde zu einer Kombination aus FDMA (Frequency Divison Multiple Access) und TDMA (Time Divison Multiple Access) macht.

Da es aber wie bereits erwähnt in Deutschland mehr Zellen gibt als mögliche Carrier, müssen Frequenzen wiederverwendet werden. In der Literatur wird häufig von einem Frequenzrecycling gesprochen. Die 124 Carrier unterschiedlicher Frequenzen können also systematisch wiedergenutzt werden. So kann eine Zelle in Karlsruhe mit den selben Frequenzen arbeiten wie auch in Berlin oder Hamburg. Dieser Recyclingfaktor liegt bei durchschnittlich neun, einer Zahl, die ihre Herkunft wohl eher der Empirie als wissenschaftlichen Überlegungen verdankt. Zumindest besagt diese, dass ich pro Zelle ungefähr $124 \,/\, 9 \approx 13$ verschiedene Frequenzen benutzen kann. Jede Frequenz lässt wiederum acht Zeitschlitze zu, was meine maximale Nutzerzahl auf $124 \,/\, 9 \cdot 8 \approx 110$ pro Zelle beschränkt (in den D-Netzen). Diesem limitierenden Faktor wurde schon

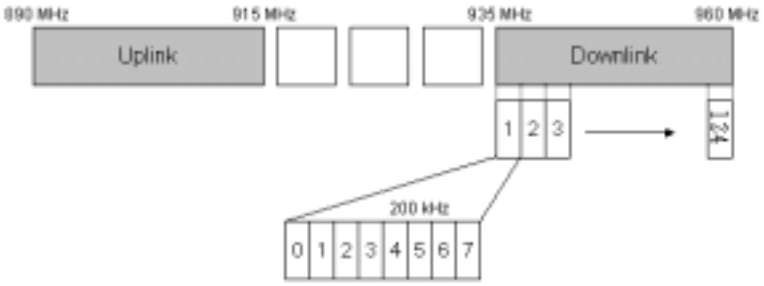

Abbildung 2: Spektrum des GSM 900MHz - Bandes und Zeitschlitzaufteilung

versucht, mit extrem kleinen Zellen (sogenannten Picozellen, Durchmesser < 100m) Abhilfe zu verschaffen. Dies ist allerdings teuer und wird daher nur in Ballungsgebieten (Fußgängerzonen, Stadien) angewandt.

Wenn es nach dem Willen der Netzwerkbetreiber geht, sollte der Endkunde bald mehr als nur die reine Sprachkommunikation über das mobile Netzwerk ausrichten. Von sprachunabhängigen Diensten wie z.B. eMail ist die Rede. Die logische Konsequenz wäre, dass noch mehr Verkehr im Netz anfällt. Damit würde das GSM-System sehr schnell an seine natürlichen Grenzen stoßen.

GPRS bietet hier eine finanziell tragbare Alternative für den Netzwerkbetreiber. Die im GSM-Exkurs erwähnten Zeitschlitze pro Frequenz können nun je nach Bedarf den Benutzern automatisch zugewiesen werden. Und damit nicht genug: ähnlich wie HSCSD (High Speed Circuit Switched Data) kann ein Nutzer auch mehrere Kanäle gleichzeitig benutzen. Der GPRS-Standard bietet technisch die Möglichkeit, einem Benutzer alle acht Zeitschlitze einer Frequenz auf einmal zuzuordnen. Da dies von der Netzwerkseite her relativ klar definiert ist, folgt diese Ausführung auch im Abschnitt über die Mobiltelefone, wo dies weit weniger zwingend gelöst wurde. Hier stoßen die Mobilfunkgerätehersteller schnell an die Grenzen des Machbaren. Batterie- und Prozessorleistungen sind hier die limitierenden Faktoren. Die bereits kurz erwähnten Modelle von Ericsson und Sagem sind sogenannte 4+1 bzw. 2+1 Modelle. Das heisst, sie unterstützen maximal vier Kanäle im Downlink (BTS zu MS) und einen im Uplink (MS zu BTS), bzw. zwei im Empfang und einen für Senden. Die letztlich für die Geschwindigkeit entscheidende Komponente (oder besser bottleneck) wird das Mobiltelefon sein.

Abschließend soll hier noch Klarheit bezüglich der möglichen Datenübertragungsgeschwindigkeiten geschaffen werden, da in vermeintlichen Fachzeitschriften über GPRS immer wieder die seltsamsten Zahlen veröffentlich werden.

Herkömmliches GSM verfügte über eine maximale Datenübertragungsgeschwindigkeit von 9,6 kbps. Das liegt daran, dass für die Kanalkodierung eines Zeitschlitzes ein sehr aufwändiger Algorithmus verwandt wird, der mit Rücksicht auf den ursprünglich ungenügenden Ausbau der Netzwerke und der damit verbundenen Notwendigkeit nach verstärkter Fehlerkorrektur entwickelt wurde.

Mit der Einführung von GPRS werden diese Beschränkungen teilweise aufgehoben. Channel Coding Scheme 2 wird mit 13,4 kbps pro Zeitschlitz Channel Coding Scheme 1 ablösen. Nun werden die künftigen, technisch möglichen Datenübertragungsraten eventuell leichter verständlich. 8 * 13,4 kbps sind 107,2 kbps. Zahlen jenseits dieser Grenze

gehen von anderen Channel Coding Schemes aus, die derzeit nicht aktuell sind, weil sie den praktischen Notwendigkeiten nach ausreichender Fehlerkorrektur nicht genügend Rechnung tragen. Im Prinzip sind Channel Coding Scheme 3 und 4 bereits fest definiert und standardisiert. Es ließen sich damit Übertragungsgeschwindigkeiten jenseits von 20 kbps pro Zeitschlitz erreichen. Allerdings ist der Unterschied zwischen Labor- und Praxisbedingungen ziemlich eklatant und der Grund für manche Kopfschmerzen bei den Netzwerkbetreibern.

1.6 Zukunft von GPRS

Das in der Einführung befindliche GPRS gründet sich auf den Standardisierungen von 1997. Es existieren bereits die verabschiedeten Spezifikationen für GPRS 1999 (Phase 2), die in etwa zwei Jahren eingesetzt werden könnten. Diese umfassen die folgenden Punkte, die hier nicht weiter ausgeführt werden sollen, obwohl sie durchaus praktische Relevanz aufweisen.

Unterstützung von Point-to-Multipoint (PTM):

- Der PTM-Multicast wird an die Benutzer innerhalb einer geographischen Zone geschickt. ähnlich dem heutigen Cell-Broadcast werden keine Empfangsbestätigungen versandt.
- Der PTM-Group Call kann ebenfalls auf ein gewisses Gebiet beschränkt werden, richtet sich im Grunde aber an eine vordefinierte Benutzergruppe. Eine Empfangsbestätigung ist hier optional.
- Der IP-Multicast entspricht dem IP-Multicast im Internet, welches derzeit allerdings nur einen geringen Anteil am Gesamtverkehr ausmacht. Der Vorteil liegt hierbei darin, dass die Pakete im Netzwerk vervielfältigt werden. Eine mögliche Anwendung sind Videodateien, die simultan von mehreren Benutzern empfangen werden können, ohne dass dies den Datenverkehr vervielfacht.

GPRS Echtzeitdienste:

- Die GPRS Phase 1 von 1997 enthält kein QoS (Quality of Service). QoS definiert die Mindestgüte eines Dienstes (z.B. Mindestrate bei der Datenübertragung). Es ist daher verständlich, dass ein solcher Schritt noch nicht vor der eigentlichen Einführung eines echten Netzwerkes definiert werden kann. Viele Probleme von GPRS werden sicher erst im laufenden Betrieb erkennbar.

Verbessertes GPRS Charging und Billing:

- GPRS Phase 1 definiert z.B. kein Prepaid. Es ist verwunderlich, dass angesichts des großen Erfolgs von im Voraus bezahlten Gesprächsgebühren, diese technische Hürde nicht eher genommen wurde. Allerdings befand sich Prepaid zu Zeiten der ersten GPRS-Standardisierungsverhandlungen selbst noch in der Aufbauphase.

Zugang zu Intranets, VPN und M2M:

- GPRS Phase 2 wird netzwerkseitig die Möglichkeiten von VPNs (Virtual Private Networks) bieten. Dazu werden ferngesteuerte Zugänge von z.B. Maschine-zu-Maschine definiert.

FIGS (Fraud Information Gathering System):

- Dies umfasst weiterführende Möglichkeiten, Missbrauch und Diebstahl von mobilen Endgeräten einzuschränken.

2 Paketübertragung in der Praxis

Wie bereits im ersten Teil erwähnt, existiert bereits ein dem GPRS sehr ähnliches System. Der japanische Telekommunikationsgigant NTT lässt durch seine Tochter DoCoMo ein PDC-Netzwerk betreiben, welches über 50 Millionen Kunden umfasst. Seit ungefähr zwei Jahren bietet NTT DoCoMo seinen Kunden ein auf Pakettransfer basierendes System an, welches den Namen i-mode trägt. Dies ist so nicht ganz korrekt, weil sich dadurch leicht Träger und Service vertauschen lassen. i-mode ist der inhaltübertragende Aufsatz auf einem GPRS-ähnlichen Netzwerk. Es könnte also auch i-mode auf GPRS geben. Diese Lösung ist übrigens nicht ganz absurd, weil NTT-DoCoMo indirekt zu fast 12% an E-Plus beteiligt ist.

i-mode konnte in den letzten 24 Monaten ein Drittel der NTT DoCoMo-Kunden gewinnen.

Folgende Gründe waren ausschlaggebend.

- günstige Preislegung
- umfangreiches Serviceangebot (Dienste, Information, Unterhaltung, Shopping, eMail, etc.)
- technische Wunderwerke als Mobilgeräte (zum Teil selbst entwickelt durch NTT), Farbbildschirme
- hohe Verfügbarkeit (zeitlich und räumlich)

Offensichtlich unwichtig ist die derzeitige Datenübertragungsgeschwindigkeit von 9,6 kbps, die der von GSM entspricht.

i-mode von NTT DoCoMo bietet seinen Kunden etwa 5.000 offizielle Seiten im cHTML Format an. Dazu kommen etwa doppelt so viele, inoffizielle Seiten von Drittanbietern. cHTML ist ein vereinfachtes HTML (Compact Hypertext Markup Language), das sich auf das Wesentlichste beschränkt (Text, GIF, Formatierungen, Tabellen).

An dieser Stelle muss wieder ein Exkurs eingefügt werden, der helfen soll, das europäische Pendant zu cHTML zu verstehen - WML (Wireless Markup Language). Das

Wireless Application Protocol (WAP) ist das Protokoll dieser komplexen, auf XML-basierenden Sprache, die speziell für die Bedürfnisse von mobilen Endgeräten entwickelt wurde.

WAP / WML haben folgende entscheidende Vorteile:

- Die Unterstützung von Active Server Pages, Java-Servlets und CGI-Skripten ist definiert.
- Ein eigner WAP-Gateway ist nicht notwendig, um Inhalte zur Verfügung zu stellen.
- WML wurde speziell für kleine Bildschirme entwickelt.
- WAP läuft über den Circuit-Switched Teil, den Paket-basierten Teil und über den Signalteil eines Netzwerks.
- Es ist netzwerkunabhängig.
- WML/WAP ermöglicht es, Cookies zu setzen.
- Dynamische WML-Pages sind ebenfalls möglich.
- Eine Layer des WAP-Stacks ist WTLS (Wireless Transport Layer Security). WTLS unterstützt Verschlüsselung und Authentifizierung zwischen Server und Client. Dies entspricht gewissermaßen der Sicherheit von SSL-Verschlüsselungen im Internet.
- WAP ist ein gemeinsamer Kompromiss der größten Spieler in der Industrie (Endgeräte- und Netzwerkhersteller sowie Netzwerkbetreiber).
- Ein User Agent Profile wird unterstützt, welches Endgeräte-spezifische Dienste ermöglicht (Der Gerätetyp kann damit identifiziert werden.).
- WAP unterstützt die Wireless Telephony Architecture, welche es erleichtert Telefonspezifika und Datenspezifika zu integrieren (z.B. Telefonbuch im Handy)

Die Nachteile von WAP:

- Es ist extrem schwierig, Verbindungen aufzubauen und zu halten. (Schuld ist allerdings der bisherige Cicuit-Switched-Verkehr.)
- WAP war bisher (v1.2 verabschiedet im Dez. 99) nicht stabil.
- WML-Seiten dürfen je nach Mobiltelefon 5 kb nicht überschreiten und sind in der Regel voll von Tags und Overhead (eine Ja/Nein-Antwort könnte theoretisch mit einem Bit (0 oder 1) gelöst werden; WAP muss etwa 100 Bit übertragen)
- Ein WML-Script muss in separater Datei gespeichert werden und kann nicht Inhalt der aufrufenden WML-Seite sein.

- Es ist in der Regel nicht einfach, den befindlichen Inhalt eines Dienstes (meist im HTML-Format) zu WML-Seiten zu konvertieren. Existierend Hilfsprogramme automatisieren wegen der Komplexität von WML meist nur einen kleinen Teil des technisch Denkbaren.

- WAP blieb bisher weit hinter den Erwartungen zurück, was die künftige Akzeptanz beim Endkunden in Frage stellt.

Zurückkehrend zu i-mode kann nun der Unterschied zwischen WML und cHTML besser verstanden werden. cHTML ist verhältnismäßig leicht zu erstellen und die Seiten werden stabil von den Endgeräten unterstützt. Damit wurden die Türen für die breite Massenanwendung geöffnet. Ein weiterer Punkt ist die günstige Preisstruktur, die NTT-DoCoMo eingeführt hat. Die Übermittlung von Daten in Paketform erlaubt nämlich, den Preis nicht nur nach der Nutzungszeit zu bestimmen, sondern gleichermaßen nach der Menge der übertragenen Daten. NTT-DoCoMo verlangt E 0.0024 je 128 bytes, die vom oder zum Mobilteil gesendet wurden. Dazu kommt eine generell erhöhte monatliche Abgabe von E 2.40, um Zugang zum paketbasierten Teil des Netzwerkes zu erhalten.

Wenn vom Serviceanbieter (Drittanbieter) gewünscht, arrangiert NTT-DoCoMo auch jeglichen Zahlungsverkehr, der über das Netzwerk abgeschlossen wird. Das heißt, dass – Dank einer eigenen Bezahlungsplattform – Lastschriften für genutzte Dienstleistungen oder bezogene Waren auf der Telefonrechnung des Kunden erscheinen können. Diesen Service lässt sich NTT-DoCoMo aber mit einer 9%-igen Marge versilbern, was bei Waren den Spielraum von Anbietern deutlich einschränkt. Informationsdienstleister wie CNN Japan, die es Kunden zum Beispiel gestatten, aktuelle Nachrichten zu abonnieren, erhalten damit allerdings eine technische Unterstützung, die es ihnen überhaupt erst einmal ermöglicht, sogenannte Microdienste im DM 2 bis DM 5 Segment anzubieten. Derartige Services sind bei i-mode sehr beliebt, machen aber am eigentlichen Datenverkehr nur einen geringen Prozentsatz aus. i-mode lebt zu fast 60% von Spielen und Zeitvertreiben, die nicht ohne weiteres nach Europa übertragen werden könnten.

Der große Erfolg von i-mode liegt auch zu einem erheblichem Teil in der hohen Verfügbarkeit der Technik. Endgeräte und Infrastruktur entsprechen den Anforderungen des Marktes. Der Kunde hat eine enorme Auswahl an verschiedenen Endgeräten. Seit kurzem bieten auch Ericsson und Nokia PDC-P-Geräte mit cHTML-Browser an. Zuvor beschränkte sich das Angebot auf die großen japanischen Hersteller und NTT selbst.

Interessant ist sicher auch, dass die i-mode Sparte von NTT-DoCoMo nur etwa 300 Mitarbeiter beschäftigt. Das zeigt, wie effektiv und erfolgreich ein solcher paketbasierter Standard eingeführt und betrieben werden kann.

Wie bereits angedeutet, kann der Erfolg von i-mode in Japan nicht 1:1 auf Europa übertragen werden. Kulturelle Unterschiede lassen nicht zu, Nutzerverhalten einfach zu kopieren. Trotzdem glaube ich, dass GPRS und WAP (die wahrscheinlichste Kombination) bald einen Durchbruch in Europa erleben werden. Dazu komme ich in einem folgenden Abschnitt unter "Always on - always connected".

3 Die 12snap AG München - Eine Fallstudie

Die 12snap AG wurde im September 1999 von sechs Personen gegründet. Heute beschäftigt das Unternehmen über 120 Mitarbeiter in vier Ländern. Die dritte Finanzierungsrunde wurde im Herbst mit über 72 Millionen DM abgeschlossen. Ein Börsengang ist eventuell für 2002 geplant.

Der Service von 12snap ist bisher nur im D2 Vodafone-Netz verfügbar. Cell-Broadcast (CBC) ermöglicht allen Nutzern, die im konkreten Fall den Kanal 123 auf ihrem Handy freigeschaltet haben, ständig Angebote zu Auktionen und Festpreisverkäufen sowie News zu erhalten. Der Empfang dieser Nachrichten ist selbstverständlich kostenlos und auf 99% aller heute verfügbaren Handys möglich. 12snap versteht sich als Infotainment Anbieter, der im besonderen junge Menschen zwischen 18 und 35 Jahren ansprechen möchte.

3.1 Cell-Broadcast (CBC)

Die Geschäftsidee von 12snap gründet sich zu einem großen Teil auf die Möglichkeiten einer stiefmütterlich behandelten Technik, dem CBC. CBC ist ein point-to-omnipoint-Übertragungsverfahren, welches wie auch SMS den Signalkanal des GSM-Netzwerks nutzt. Dabei werden alle Teilnehmer erreicht, die sich im Empfangsbereich einer bestimmten Antenne befinden. 12snap nimmt diesen Service bundesweit in Anspruch, das heißt, alle Antennen strahlen das gleiche Signal ab.

CBC ist keine proprietäre GSM-Erfindung. Technisch ist es in CDMA, AMPS und PDC Netzwerken ebenfalls möglich. Nur sind mir keine Anwendungsbeispiele bekannt.

3.2 12snap und GPRS

Ausgangspunkt dieses Dokuments war die Fragestellung, inwieweit GPRS das Geschäftsmodell von 12snap beeinflussen könnte.

Bevor hier allerdings auf diese ganz konkrete Fragestellung eingegangen wird, soll dem derzeitigen Entwicklungsstand von GPRS in Deutschland noch Aufmerksamkeit geschenkt werden. Generell sollte das potentiellen Dienstanbietern helfen, besser in die Zukunft planen zu können.

Das folgende gesammelte Wissen stammt vom November letzten Jahres, als ich mein Praktikum beendete, und ist die ausgewertete Information von unzähligen Newsletters, Pressemitteilungen und Internet News-Groups. GPRS ist heute schon in manchen Gebieten verfügbar. Zum Zeitpunkt des Seminars werden sich viele Dinge zusätzlich noch einmal geändert haben.

3.2.1 GPRS in Deutschland (Stand: Nov 00)

Wie bereits mehrmals erwähnt, ist das Angebot an GPRS-fähigen Endgeräten noch sehr dürftig. Nach den GPRS-Spezifikationen wird es vorerst auch keine Prepaid-Handys geben. Prepaid ist Teil der GPRS Phase 2, die erst in etwa zwei Jahren implementiert

wird. Da Prepaid den derzeitigen Boom des Mobilfunksektors entscheidend vorangetrieben hat, ist GPRS ohne Prepaid nur schwer vorstellbar. Vodafone in Großbritannien, das zu 60% Prepaid-Kunden zählt, hat daher schon angedeutet, dass ein proprietäres System Prepaid von Beginn an mit GPRS ermöglichen soll. Das wiederum legt die Vermutung nahe, dass D2 infolge der Konzernzugehörigkeit ebenfalls Prepaid anbieten wird, was andererseits die anderen Netzbetreiber in Zugzwang bringen könnte. Insgesamt fehlten im November aber auch noch klare Aussagen bezüglich der Preisfestsetzung von GPRS.

D2 Vodafone D2 kündigte einen GPRS-Versuch im Juli letzten Jahres an. Der Test sah vor, jede Stunde GPRS-Nutzung mit Eur 0.25 zu bepreisen. Dazu kam allerdings noch eine datenabhängige Gebühr von etwa Eur 0.20 je 10 Kb. Ein zweites Preismodell schlug eine monatliche Pauschale von Eur 5 bis 10 vor, die bereits alle Datentransfers enthalten sollte.

D1 Telekom T-Mobil unternahm einen ähnlichen Test Ende September, der eine monatliche Grundgebühr von ungefähr Eur 6 vorsah und für jedes 10 Kb Paket Eur 0,35 berechnen wollte. Daneben gab es noch einen Vorschlag B, der ähnlich wie D2 Vodafone eine Gebühr pro Stunde plus ein datenmengenabhängiges Entgelt plante.

Diese – im Vergleich zu i-mode – sehr hohen Preise würden eine weite Verbreitung von GPRS in Deutschland unterbinden. Bevor aber zu eilig pessimistische Schlüsse gezogen werden, sollte besser bis zur Einführung von GPRS in den nächsten Tagen und Wochen gewartet werden. Wenn GPRS einmal stabil läuft, wird mit Sicherheit auch der Preis gesenkt, um das Netzwerk weiteren Nutzern zu öffnen.

Zurück zur Problematik, die 12snap und andere Dienstanbieter beschäftigt.

Grundsätzlich kann vorausgesagt werden, dass es sich zu Beginn nicht um ein Massenprodukt handeln wird. Typische Anwendungen werden sich auf die Belange von High-End-Kunden beschränken, die Zugang zu Unternehmens-LANs nutzen und eMails empfangen bzw. versenden wollen. Daher wird GPRS auch nicht mehr mit höchster Priorität bei 12snap betrachtet.

Ein großer Vorteil, den paketbasierter Verkehr gegenüber Circuit-Switched-Verkehr bietet, wurde bisher nicht angesprochen: GPRS bietet die prinzipielle Möglichkeit, ähnlich einer Internetverbindung im Rechenzentrum, immer on-line zu sein. Das wird enorme Konsequenzen für das Verhalten der Endkunden haben. Deshalb soll hier noch einmal ausführlicher auf die bereits angekündigte

3.3 "Always on - always connected"

Phrase eingegangen werden, mit der jeder Artikel über GPRS eingeleitet wird.

Ein Mobiltelefon muss demnach erst einmal in einen Zustand (Ready) gebracht werden, der es dem Gerät erlaubt, Daten auszutauschen. Dazu muss sich die MS an GPRS "attachen" ("ankoppeln"). Hierbei wird dem Endgerät eine eindeutige, temporäre Verbindungskennung zugeordnet; ähnlich der IP-Nummer bei der Modemeinwahl in einen

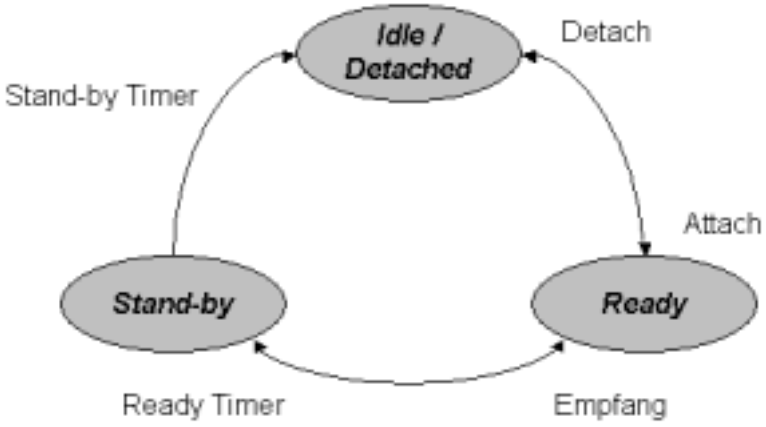

Abbildung 3: Zustandsmodell des GPRS-Endgeräts

ISP. Nach erfolgreichem Einbuchen wird mit dem SGSN ein Routingkontext ausgehandelt, der sinngemäß festlegt, welche Dienste erlaubt und welche Zugriffe auf das Public Data Network möglich sind. Wie schon im ersten Teil erwähnt, speichert das GPRS-Register (GR) die notwendige Information. Nach dem Festlegen der Dienste wird der Routingkontext im entsprechenden GGSN aktualisiert. Es ist zwingend, dass ein Netzwerk mit mobilen Kunden auf die laufende Aktualisierung dieser Daten angewiesen ist.

Der Zustand "Ready" informiert den SGSN über jeden Zellwechsel. Werden im "Ready"-Modus keine Aktionen wahrgenommen, wird nach einer gewissen Zeit automatisch der Zustand "Stand-by" eingenommen. Befindet sich die MS im Modus "Stand-by", wird nur eine Aktualisierung vorgenommen, wenn eine gewisse Gruppe von Zellen, die sogenannte Routing Area, verlassen wird. Sollte mit einer neuen Routing Area auch der Einzugsbereich eines neuen SGSN betreten werden oder gar der GGSN gewechselt werden, muss diese Information entsprechend im Netzwerk verarbeitet werden.

Ein Timer überwacht auch den "Stand-by"-Zustand. Werden keine Aktionen getätigt, wird auf "Idle" geschalten. Dies entspricht dem Ausgangszustand "Detached". Es können keine Daten empfangen werden.

Deutlich wird in jedem Fall, dass die Verwaltung eines Netzwerks mit mobilen Endkunden recht komplex werden kann. Das wird sich der Betreiber auch entsprechend vergüten lassen, doch darüber lässt sich nur spekulieren.

Der paketbasierte Verkehr bietet also dem Endkunden die Möglichkeit, immer empfangsbereit für Daten und Gespräche zu sein. Letztere werden wegen der Dienstgüte nach wie vor noch im Circuit-Switched-Teil des Netzwerks verarbeitet werden. Die GPRS-Spezifikationen der Phase 1 definieren aber grundsätzlich auch Sprache über IP.

Die eben erläuterte Möglichkeit des "Immer-verbunden-seins" und das technische Grundverständnis von GPRS erlauben nun einige interessante

3.4 Rückschlüsse und Implikationen für 12snap und andere Serviceanbieter.

- WAP wird auf GPRS seinen endgültigen Durchbruch erleben. Deshalb sollte das WAP-Angebot der Firma verbessert werden. WAP läuft mittlerweile in sich stabil. Der große Datenoverhead wird allerdings dem Kunden in Rechnung gestellt. Der Vorteil der Paket-Verbindung ist, dass sogenannte Dropped-Calls per Definition nicht mehr auftreten können. Sie werden nur noch als Verzögerung wahrnehmbar sein.

- Künftig werden SMS ebenfalls über den Paket-basierten Teil des Netzwerkes verschickt werden können. Die äquivalente Datenmenge einer SMS kostet einen i-mode Kunden derzeit Eur 0.003, das heißt ungefähr einen halben Pfennig. Dazu kommt noch, dass der Signalkanal, den der Netzbetreiber heute nutzt, um SMS zu versenden, zunehmend überlastet ist. D.h. SMS wird künftig durch direktes Zusenden von Information (abonnierte Dienste wie Nachrichten, Horoskope, etc.) an den Endkunden abgelöst. Das wirft allerdings weitere Fragen des unlauteren Datenversands auf, die hier nicht weiter diskutiert werden sollen.

- Ein Geschäftsmodell, das auf geteilten Gewinnen aus dem SMS-Versand aufbaut, wird über kurz oder lang begraben werden müssen, weil SMS einem rapiden Preisverfall unterliegen werden (160 Zeichen kosten in der i-mode-Welt weniger als ein Pfennig).

- CBC ist eine interessante Technik, die allerdings zum Tode verurteilt ist. Sie ermöglicht ein gießkannenartiges Verteilen von Information, das dem Bedürfnis des Endkunden nach individueller Behandlung nicht ausreichend entgegenkommt. Ein vergleichbarer, individuell zugeschnittener Service in i-mode würde etwa DM 8 pro Monat kosten. Diese Kalkulation geht von 1'800 Nettokontakten im Monat aus, wie sie derzeit über CBC erreicht wird.

- Es ist leider noch unklar, inwieweit WML-Seiten an den Endkunden zugesandt werden können. Hier ergeben sich eventuell vielversprechende Möglichkeiten für Direktmarketing.

- GPRS wird zumindest anfänglich teuer sein. Es ist fraglich, ob sich das Kundensegment der Firma 12snap mit dem der frühen GPRS-Kunden decken wird.

- QoS wird bald eine größere Rolle spielen. GPRS ermöglicht grundsätzlich die Einteilung der Kunden nach verschiedenen Güteklassen, wie in einem kleinen Exkurs noch gezeigt werden soll.

GPRS definierte verschiedene Dienstgüteprofile, die beim "Attachen" an GPRS "ausgehandelt" werden. Folgende fünf Eigenschaften spielen in ein Dienstgüte- oder QoS-Profile mit ein:

1. Dringlichkeit (Drei Zustände, die die Priorität des Verkehrs gegenüber der restlichen Last im Netz vergleichbar machen)

2. Verzögerung (Vier Klassen definieren, inwieweit ein Paket einer bestimmten Größe Verzögerungen im Netzwerk unterliegen darf. Dabei kann natürlich nicht geregelt werden, welche Verzögerungen in Public Data Networks außerhalb des Netzwerkbetreibers anfallen können).

3. Verlässlichkeit (Fünf Verlässlichkeitsklassen charakterisieren den Datenverkehr nach der Wahrscheinlichkeit für Datenverlust, Reihenfolgefehler, Übermittlungsfehler und Mehrfachauslieferung)

4. Durchsatzklassen (Diese klären, welche mittlere Datenrate oder welche Spitzenrate dem Kunden ermöglicht werden darf, oder welche ihm ermöglicht werden muss)

5. Parallele Nutzung von Sprach- und Datendiensten (siehe Klassen A,B und C von MS). Hierbei wird die Nutzung netzwerkseitig definiert.

GPRS kann QoS demnach individuell für jeden Kunden vereinbaren und vordefinieren. Es wird verschiedene Kundenklassen geben. Der Dienstanbieter muss nun überlegen, welches QoS-Profile sich am ehesten mit dem seines "typischen Kunden" deckt. Entsprechend sollte das Serviceangebot gestaltet werden.

- Mit diesem Teil des QoS teilt der Netzwerkbetreiber also seine Kunden direkt nach Priorität ein. Für die Firma wiegt dieser Punkt aber weniger schwer, weil der Dienst nicht wirklich von der Datenübertragungsgeschwindigkeit abhängt und nur selten wirklich zeitkritisch ist. i-mode bewies, dass Geschwindigkeit nicht allein entscheidet. Wichtig ist allerdings, zu erkennen, dass der "klassische Endkunde" möglicherweise selten Zugang zu mehr als einem Kanal haben wird, was vom Applikationsprogrammierer beachtet werden sollte.

4 Fazit

GPRS wird in wenigen Monaten zum Alltag gehören. WAP wird nun – verdient oder unverdient – seinen endgültigen Durchbruch auf GPRS erleben. Vor allem durch die Möglichkeiten des "Always on - always connected" wird sich das Nutzerverhalten zu herkömmlichem GSM tiefgreifend ändern. Voraussetzung dafüer ist allerdings eine angemessene Preislegung, die ganz entscheidend für den Erfolg oder Misserfolg von GPRS sein wird. Grundsätzlich wird es uns Endkunden aber möglich sein, viele, vor allem zeitkritische Dinge (eMail, Kartenbuchungen, Sonderangebote, etc.) von nun im Bus, Bahn oder vor dem Kino/Theater direkt zu erledigen. Zu den traditionellen Anwendungen werden aber auch neue hinzukommen: Spiele, Bezahlung über Handy, Preisvergleiche, etc. Natürlich ist vieles auch schon heute möglich, GPRS wird es für den Endkunden nur finanziell tragbar machen – hoffentlich. EDGE und UMTS werden irgendwann schließlich nur noch Bandbreite bringen; die Revolution hat dann bereits stattgefunden.

Literatur

[CDG01] CDMA Development Group, 2001. http://www.cdg.org.

[Eber99] Jörg Eberspächer. *GSM, Global System for Mobile Communication : Vermittlung, Dienste und Protokolle in digitalen Mobilfunknetzen, 2.Auflage.* Verlag Teubner, Stuttgart, Leipzig, Wiesbaden. 1999.

[Eric00] Ericsson. http://www.ericsson.com/gsm, 2000.

[ETS99] European Telecommunication Standards Insitute, 1999. http://www.etsi.org.

[gol00] T-D1 führt Hochgeschwindigkeits- Mobilfunktechnik GPRS ein, 2000. http://www.golem.de/0006/8288.html.

[gpr00] Mobile Applications Initiative, 2000. http://www.gprsworld.com.

[GSM01] http://www.gsmworld.com, 2001.

[Nort00] Northstream. http://www.northstream.se, 2000.

[NTT01] NTT DoCoMo Net, 2001. http://www.nttdocomo.co.jp.

[Walk00] Bernhard Walke. *Mobilfunknetze und ihre Protokolle, Band 1, 2.Auflage.* Verlag Teubner, Stuttgart, Leipzig, Wiesbaden. 2000.

Abbildungsverzeichnis

Optimierungen von Mobile IP

Olaf Kleine

Kurzfassung

Mobile IP wurde von der Internet Engineering Task Force (IETF) als Erweiterung des IP-Protokolls entwickelt. Damit gab es einen Ansatz, der eine Datenübertragung aus dem Internet bei gleichzeitiger Mobilität unterstützte. Doch ließ dieser Ansatz noch Wünsche offen. In dieser Arbeit werden einige Verfahren vorgestellt, die eine Verbesserung von Mobile IP zum Ziel haben. Erster Ansatz ist die Verbesserung der micro-Mobilität, die sich HAWAII und Cellular IP zur Aufgabe gemacht haben. Bei einem weiteren Vorschlag geht es um die Optimierung beim Routing vom Correspondent Node zum Mobile Node. Die wichtigsten Ansätze sind die zur Verbesserung der Sicherheit sowohl der Netzwerke als auch der Verbindungen zum Mobile Node, die in FATIMA und mit AAA-Servern gemacht werden.

1 Einleitung

Die ständige Erreichbarkeit und die Möglichkeit, jederzeit Informationen abrufen zu können wird für die Gesellschaft immer wichtiger. So finden Notebooks und Palm-Rechner eine immer größere Verbreitung. Doch nur auf Daten zugreifen zu können, die vorher zu Hause aufgespielt wurden befriedigt nicht immer das Bedürfnis nach Aktualität. Ziel ist es deshalb, von verschiedenen Orten aus Informationen aus dem Internet verfügbar zu machen. Wird im Auto oder im Zug gearbeitet, kann es auch vorkommen, daß während der Übertragung von Daten aus dem Internet der Zugriffspunkt des Computers wechselt. Dafür ist das Internet aber bisher nicht ausgelegt. Um den Wunsch nach Mobilität entgegenzukommen wurde deshalb Mobile IP entwickelt.

2 Mobile IP

Bei den üblichen Routingverfahren im Internet ist der topologische Zugriffspunkt eines Computers durch seine IP-Adresse zum größten Teil bestimmt. Um Mobilität zu ermöglichen, umgeht Mobile IP diese enge Bindung zwischen IP-Adresse und Zugriffspunkt des Computers.

Ein Computer, der seinen Zugriffspunkt ins Internet wechselt, wird Mobile Node genannt. Befindet er sich in seinem eigenem Netzwerk (Home Network), so kommuniziert er wie gewöhnlich mit anderen Computern (Correspondent Node). Das heißt, der Mobile

Node sendet bei seinen Anfragen an den Correspondent Node seine IP-Adresse als Absender mit. Die Antworten des Correspondent Node werden dann an diese IP-Adresse gesendet. Aufgrund dieser IP-Adresse wird die Antwort zum Home Network geleitet und schließlich zum Mobile Node.

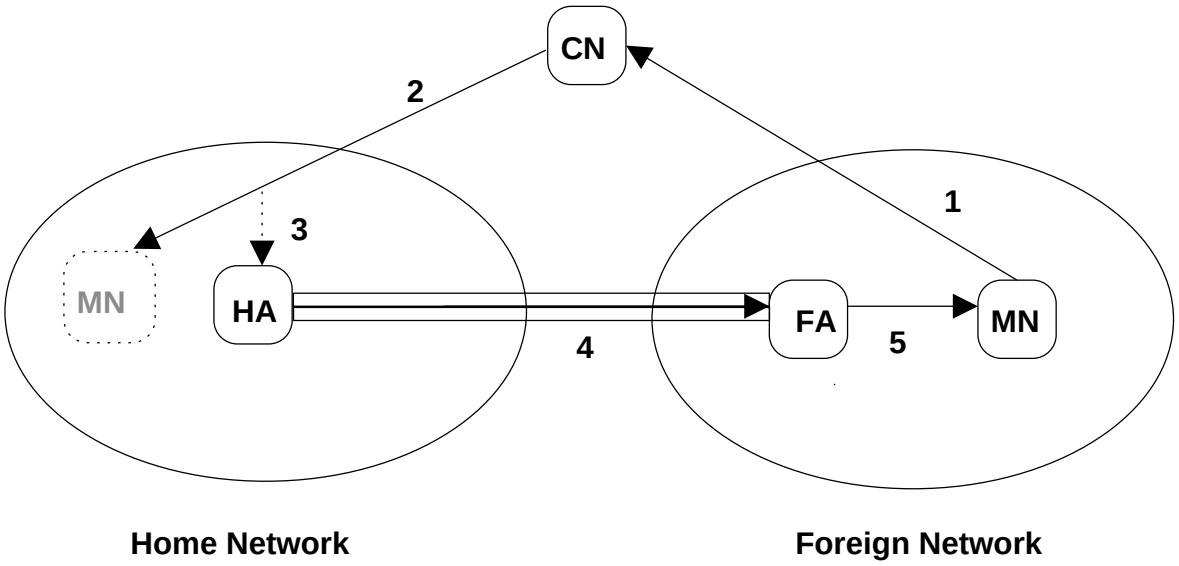

Abbildung 1: *Datenverkehr bei Mobile IP:* Die Nachricht vom Mobile Node geht an den Correspondent Node (1), dieser schickt die Antwort an den vermeintlichen Aufenthaltsort der Mobile Node in das Home Network (2). Diese Antwort wird von dem Home Agent abgefangen (3) und an den Foreign Agent getunnelt (4). Dieser sendet die Antwort dann an den Mobile Node (5).

Wechselt nun der Mobile Node in ein anderes Netzwerk (Foreign Network), so läuft die Kommunikation über Mobile IP. Voraussetzung dafür ist ein Home Agent im Home Network sowie ein Foreign Agent im Foreign Network. Aufgabe des Home Agent ist es, den aktuellen Foreign Agent zu kennen, der für den Mobile Node zuständig ist. Der Foreign Agent macht durch Advertisement-Nachrichten auf sich aufmerksam. Diese Advertisement-Nachrichten empfangen alle an das selbe Medium angeschlossene Geräte wie der Foreign Agent (Link Layer). Empfängt der Mobile Node eine solche Advertisement-Nachricht, so kann er daran erkennen, daß sein Zugriffspunkt ins Internet sich geändert hat. Des Weiteren kann anhand der Advertisement-Nachricht eine Care-of-Address (Stellvertreter-Adresse) festgestellt werden. Diese Care-of-Address ist die IP-Adresse des Foreign Agent. Um seinen neuen Aufenthaltsort dem Home Agent mitzuteilen, generiert der Mobile Node nun eine Registration Request -Nachricht. Diese enthält die neue Care-of-Address und wird über den Foreign Agent an den Home Agent gesendet. Als Bestätigung erhält der Mobile Node vom Home Agent eine Registration Reply -Nachricht. Die beim Home Agent registrierte Care-of-Address des Mobile Node hat jedoch nur eine zeitlich begrenzte Gültigkeit. Vor Ablauf dieser Frist muß sich der Mobile Node erneut beim Home Agent melden, um auch weiterhin Daten empfangen zu können.

Will nun der Mobile Node mit einem Correspondent Node kommunizieren, so schickt er als Absender seine normale IP-Adresse. Die Antwort des Correspondent Node wird somit zum Home Network geleitet. Hier wird sie vom Home Agent abgefangen und quittiert (siehe auch Abbildung 1). Der Home Agent tunnelt dann die Antwort an die Care-of-Address, die zur Zeit bei ihm registriert ist. Tunneln bedeutet hier, daß die empfangenen Daten mit Absender- und Zieladresse als eine neue Nachricht aufgefasst, und mit einer neuen Absenderadresse (die des Home Agent) und einer neuen Empfängeradresse (die des Foreign Agent) versehen werden. Dadurch gelangt die Antwort des Correspondent Node zum Foreign Agent. Dieser enttunnelt die Antwort und leitet sie an den Mobile Node weiter.

Als Variante gibt es noch die sogenannte co-located Care-of-Address. Hierbei wird dem Mobile Node im Foreign Network eine eigene neue (topologisch korrekte) IP-Adresse zugeteilt. Nach der Registrierung dieser co-located Care-of-Address beim Home Agent werden somit die Daten direkt zum Mobile Node getunnelt. Auf diese Art und Weise kann der Foreign Agent eingespart werden.

Zum Schutz gegen Abfangversuche der Nachrichten an den Mobile Node ist beim Registrieren der Care-of-Address beim Home Agent eine Authentifikation notwendig. Dies betrifft sowohl die Registration Request -Nachricht als auch die Registration Reply -Nachricht. Die Authentifikation setzt ein gemeinsames Geheimnis zwischen dem Home Agent und dem Mobile Node voraus.

3 Granularität der Mobilität

Für jeden Wechsel des Zugriffspunktes vom Mobile Node ins Internet ist eine Registrierung beim Home Agent über das Internet notwendig. Datenpakete an den Mobile Node, die der Home Agent vor dem Abschluss der neuen Registrierung noch an den vorhergehenden Foreign Agent schickt, gehen verloren. Aus diesem Grunde sollte für ein effizientes Arbeiten von Mobile IP ein Wechsel des Zugriffspunktes nicht zu häufig auftreten. Somit ist Mobile IP besonders für sogenannte macro-Mobilität gut geeignet.

Nun ist aber zu erwarten, daß Mobile IP hauptsächlich bei Funknetzen eingesetzt wird. Der Boom bei den Mobiltelefonen ist ein deutliches Zeichen dafür, daß mobiler Datenaustausch gewünscht wird. Eine Ausstattung von Palm-Rechnern oder Notebooks mit Funkschnittstellen hat schon begonnen. Bei einem Funknetz wird ein Gebiet in Funkzellen unterteilt. Der Datenverkehr in einer Funkzelle wird über eine Basisstation abgewickelt. Dies geschieht auf dem Link-Layer und ist damit der Ansatzpunkt für Mobile IP. Somit hat eine Basisstation die Aufgaben eines Foreign Agent, und für jeden Wechsel von einer Funkzelle in die nächste ist eine neue Registrierung beim Home Agent notwendig. Um eine höhere Übertragungsleistung für den Einzelnen bereitstellen zu können, sind die Funkzellen aber eher klein. Möchte man nun im fahrenden Auto oder Zug an seinem Notebook arbeiten, so erfolgt sehr häufig ein Wechsel der Funkzelle. Dieses häufige Wechseln des Zugriffspunktes ins Internet wird micro-Mobilität genannt. Für diese micro-Mobilität ist Mobile IP zu träge. Aus diesem Grunde ist dies der erste Ansatzpunkt für Optimierungen.

3.1 HAWAII

Beim Handoff-Aware Wireless Access Internet Infrastructure (HAWAII) [RPTV+99] übernimmt Mobile IP nur noch die grobe Verfolgung des Aufenthaltsortes des Mobile Node. So bekommt der Home Agent nur den Wechsel von einem Foreign Network in ein anderes mitgeteilt. Die micro-Mobilität, nämlich den Wechsel des Zugriffspunktes ins Internet innerhalb eines Foreign Networks, übernimmt das Foreign Network selber.

Ein HAWAII-Netzwerk hat als zentralen Zugangspunkt ins Internet den Domain Root Router. Alle Daten, die an dieses Subnetz gesendet werden, gehen über diesen Router. Der innere Aufbau des Subnetzes besteht aus Routern und Basisstationen, die mit dem Mobile Node in Kontakt treten können. Jeder dieser Router und Basisstationen hat einen Routingeintrag zum nächsten Router, um Daten an den Domain Root Router zu senden. Kommt ein Mobile Node in ein HAWAII-Subnetz, so wird ihm eine co-located Care-of-Address zugeteilt. Die Basisstation, mit der der Mobile Node kommuniziert, erzeugt eine Registration Request -Nachricht an den Home Agent. Erfolgt eine positive Antwort vom Home Agent, so wird als nächstes ein Pfad vom Domain Root Router zu dieser Basisstation angelegt. Dazu erzeugt die Basisstation eine Power-up-update-message und sendet sie an den nächsten Router in Richtung Domain Root Router. Jeder Router, der diese Power-up-update-message erhält, setzt einen entsprechenden Routingeintrag für die Mobile Node, sodaß der Pfad wieder zurückverfolgt werden kann. Diese Routingeinträge sind nur eine bestimmte Zeit lang gültig und müssen dann wieder erneuert werden. Wenn die Power-up-update-message den Domain Root Router erreicht, dann sendet dieser eine Bestätigung an die Basisstation zurück. Nach so erfolgreich angelegten Pfad vom Domain Root Router zum Mobile Node erzeugt die Basisstation ein Registration Reply -Nachricht und sendet sie an den Mobile Node.

Ändert nun der Mobile Node innerhalb eines HAWAII-Netzwerkes seine Basisstation, so behält er seine co-located Care-of-Address bei. Dadurch bedarf es keiner erneuten Registrierung beim Home Agent. Statt dessen sendet der Mobile Node eine Registration Request -Nachricht an die neue Basisstation. Diese sorgt für eine Änderung des Pfades vom Domain Root Router zur neuen Basisstation. Ist dieses erfolgreich abgeschlossen, so sendet die Basisstation eine Registration Reply -Nachricht an den Mobile Node.

Für die Änderung des Pfades existieren zwei Algorithmen. Der Forwarding Path Setup -Algorithmus ist optimiert für ein kabelloses Netz, bei dem der Mobile Node nur mit einer Basisstation auf einmal kommunizieren kann. Ist der Mobile Node hingegen in der Lage, mit mehreren Basisstationen auf einmal in Kontakt zu stehen, so ist der Non-Forwarding Path Setup -Algorithmus besser geeignet. Gemeinsame Voraussetzung ist, daß anhand der Registration Request -Nachricht die alte Basisstation festgestellt werden kann. Dies ist durch die Erweiterung Previous Foreign Agent Notification Extension (PFANE) gegeben. Diese Erweiterung wurde für die Routing Optimization eingeführt, die in Kapitel 4 vorgestellt wird.

Beim Forwarding Path Setup -Algorithmus erzeugt die neue Basisstation eine Path-setup-update-message mit seiner eigenen Adresse, und sendet diese an die alte Basisstation (siehe auch Abbildung 2). Die alte Basisstation setzt ihren Routingeintrag für den Mobile Node auf die Adresse des Routers, der der neuen Basisstation näher ist. An diesen Router wird die Path-setup-update-message ebenfalls geschickt. Bei einer Baumstruktur wird so die Botschaft solange nach oben weitergeleitet, bis der gemeinsame Vorfahre von der alten und der neuen Basisstation erreicht ist. Auf diese Art und

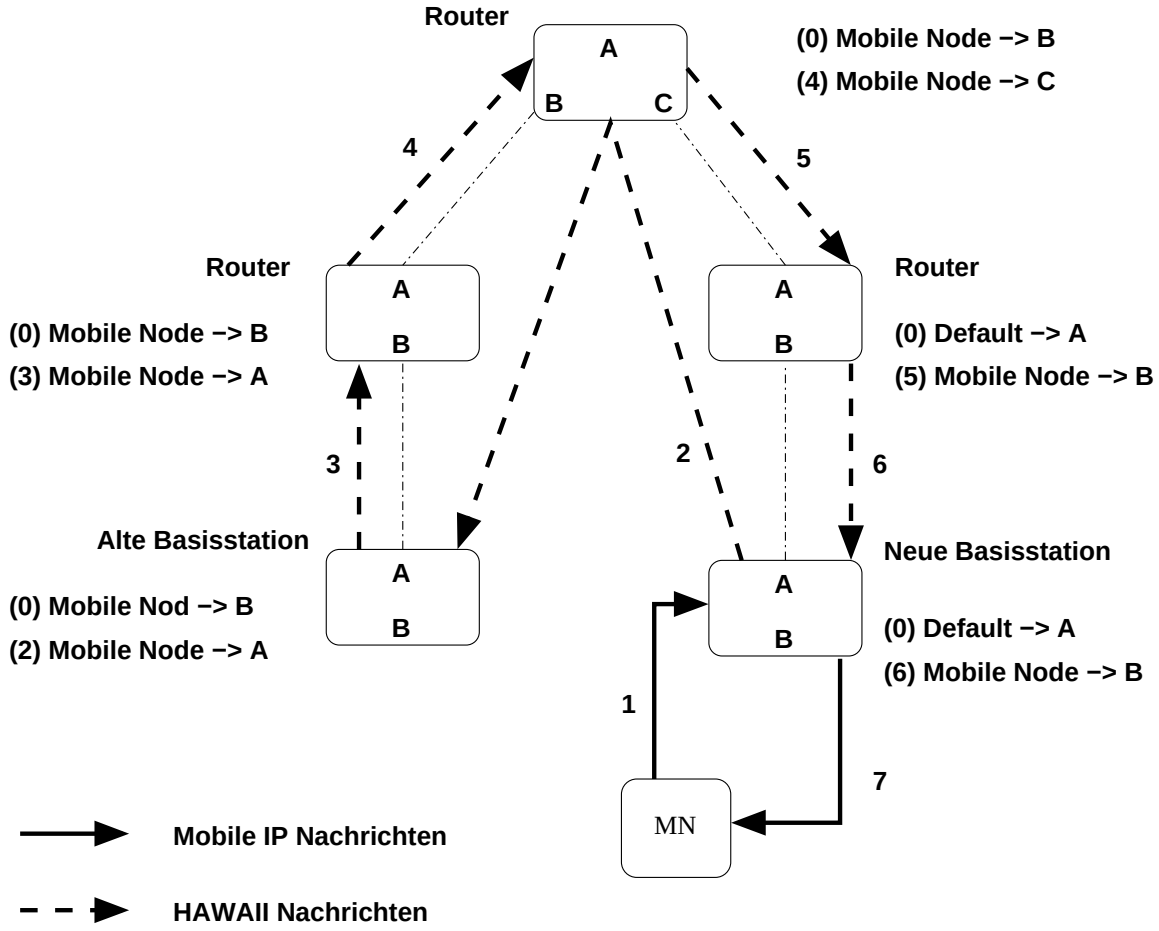

Abbildung 2: *Forwarding Path Setup -Algorithmus:* Die Routingeinträge für den Mobile Node sind im Bild angegeben. In Klammern steht die Nummer der Nachricht, die den Eintrag erzeugt hat. Der Anfangswert hat die Nummer 0.

Weise wird der Teil des Pfades aktualisiert, der nicht mehr für die Kommunikation mit dem Mobile Node gebraucht wird. Nun setzt sich das Verfahren fort, und erreicht, der Baumstruktur nach unten folgend, die neue Basisstation. Diese setzt auch ihren Eintrag für den Mobile Node und sendet ihm eine Registration Reply -Nachricht.

Beim Non-Forwarding Path Setup -Algorithmus werden die Erneuerungen der Routingeinträge in umgekehrter Reihenfolge vorgenommen (siehe auch Abbildung 3). So setzt zuerst die neue Basisstation seinen Routingeintrag für den Mobile Node. Die Path-setup-update-message enthält nun die Adresse der alten Basisstation. Sie wird von der neuen Basisstation an den voreingestellten Router auf dem Weg zum Domain Root Router geschickt. In jedem durchquerten Router wird ein Eintrag für das Weiterleiten der Daten an den Mobile Node erzeugt. Auf diese Weise kommt die Path-setup-update-message bis zu dem gemeinsamen Vorfahren von der neuen und der alten Basisstation. Die Path-setup-update-message wird auf dem Pfad zur alten Basisstation weitergeleitet. Auf diesem nun nicht mehr gebrauchten Pfad werden die Einträge für den Mobile Node auf den jeweils voreingestellten Router geändert. Somit weisen diese Einträge in Richtung des gemeinsamen Vorfahren. Erreicht die Path-setup-update-message die alte Basisstation, so sendet diese eine Quittierung an die neue Basisstation, welche daraufhin dem Mobile Node eine Registration Reply -Nachricht sendet.

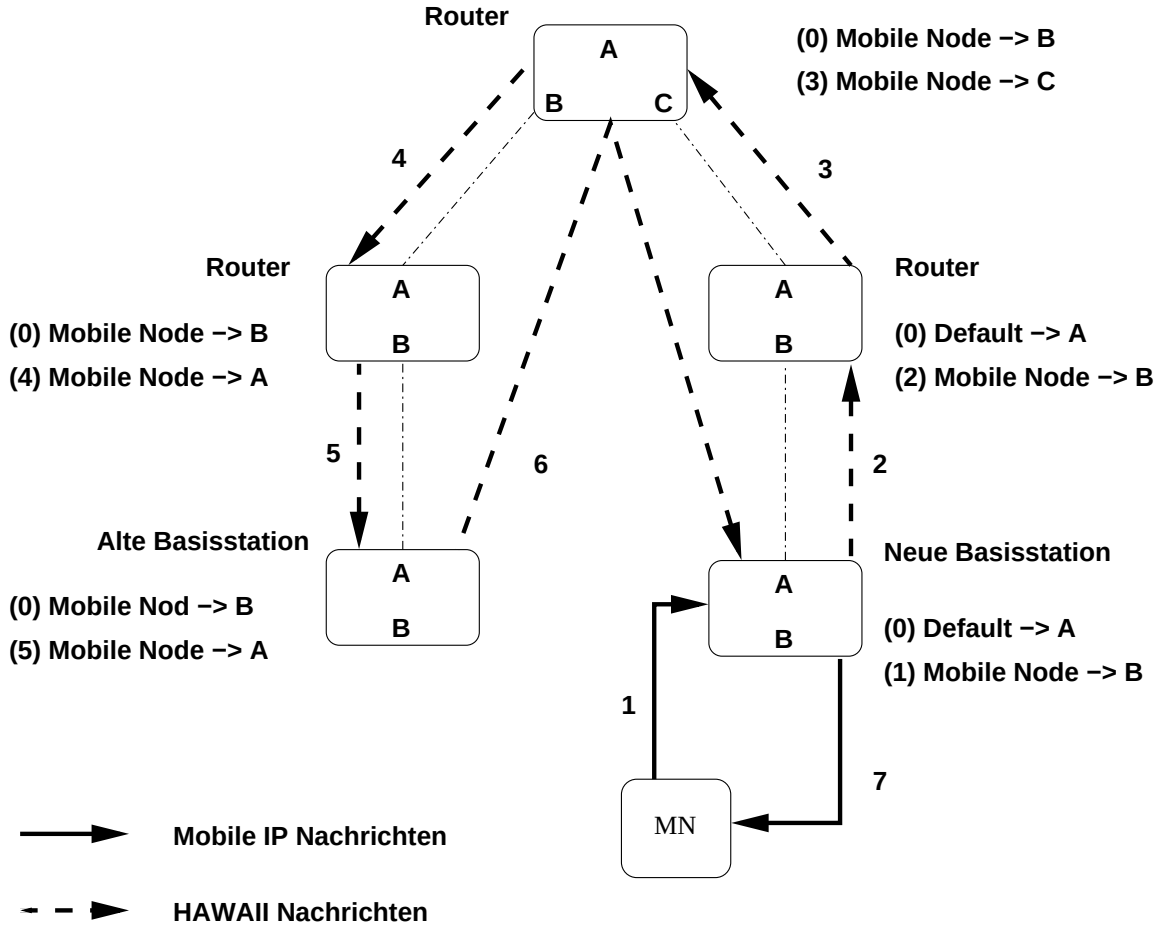

Abbildung 3: *Non Forwarding Path Setup -Algorithmus:* Die Routingeinträge für den Mobile Node sind im Bild angegeben. In Klammern steht die Nummer der Nachricht, die den Eintrag erzeugt hat. Der Anfangswert hat die Nummer 0.

Somit sind bei einer Änderung des Zugriffspunktes des Mobile Node nur die Router betroffen, die auf dem Pfad von der alten und der neuen Basisstation zum gemeinsamen Vorfahren liegen. Insbesondere ist der Home Agent nicht betroffen sondern nur das Foreign Network. Dies führt dazu, daß eine Änderung schneller durchgeführt werden kann.

Ziel von HAWAII war es, micro-Mobilität zu gewährleisten und gleichzeitig kompatibel zum ursprünglichen Mobile IP-Protokoll zu sein. Allerdings wurde der Registrierungsablauf dabei zerteilt. Der Mobile Node registriert sich nur noch bei der Basisstation. Zur Absicherung sind diese Nachrichten authentifiziert mit einem gemeinsamen Schlüssel zwischen dem Mobile Node und der Basisstation bzw. dem HAWAII-Netzwerk. Die Basisstation registriert dann den Mobile Node beim Home Agent. Diese Nachricht ist wiederum authentifiziert mit einem gemeinsamen Schlüssel zwischen der Basisstation und dem Home Agent. Der Mobile Node bekommt nur noch Registration Reply -Nachrichten, die nicht, wie vorgesehen, vom Home Agent authentifiziert sind, sondern von der Basisstation. Es bedarf also eines hohen Vertrauens vom Home Agent gegenüber dem Foreign Network.

3.2 Cellular IP

Ein zweiter Ansatz zur Verbesserung der micro-Mobilität ist Cellular IP [Valk99]. Ziel dieser Entwicklung war es vor allem, eine kostengünstige Lösung für den Aufbau des Netzes sowie eine einfache Administration und Erweiterbarkeit zu erreichen.

Wie bei HAWAII wird der Home Agent nur noch bei einem Wechsel des Mobile Node von einem Foreign Network zu einem anderen unterrichtet. Ändert der Mobile Node hingegen innerhalb des Foreign Network seinen Zugangspunkt, so wird dies intern im Foreign Network behandelt.

Der Aufbau des Foreign Network bei Cellular IP ist eine Baumstruktur. An der Wurzel sitzt der Wurzelrouter als Foreign Agent, der den gemeinsamen Zugang ins Internet für den gesamten Baum bereitstellt. Jeder innere Knoten in der Baumstruktur ist ein Router. Die Blätter sind Basisstationen, die den Funkkontakt mit dem Mobile Node herstellen.

Jede Basisstation und Router kennt seinen Vorfahren in dieser Baumstruktur. Tritt nun ein Mobile Node in dieses Foreign Network ein, so erhält es als Care-of-Address die Adresse des Foreign Agent, also die des Wurzelrouters. Die Registration Request -Nachricht des Mobile Node wird dabei von der Basisstation an den Vorfahren in der Baumstruktur hochgereicht. Die Nachricht wird solange hochgereicht, bis sie den Wurzelrouter, also den Foreign Agent, erreicht hat. Dieser kann dann, wie in Mobile IP vorgesehen, die Nachricht an den Home Agent weiterleiten. Bei jedem Hochreichen wird für den Mobile Node ein Registrierungseintrag im Cache des Routers erstellt, der angibt, von welchen Vorfahren die Nachricht kam. Auf diese Weise wird ein Pfad von der Wurzel bis zur Basisstation hergestellt, mit dem der Mobile Node verbunden ist. Diese Einträge in den Caches haben nur eine begrenzte Gültigkeit, und müssen nach Ablauf dieser Frist wieder aufgefrischt werden.

Sendet nun ein Correspondent Node eine Nachricht an den Mobile Node, gehen die Daten, wie bei Mobile IP vorgesehen, zuerst zum Home Agent. Dieser tunnelt die Daten zum Foreign Agent, in diesem Fall also zum Wurzelrouter. Diese Daten müssen nun durch die Baumstruktur zurück zum Mobile Node geleitet werden. Jeder Knoten verwendet die Einträge in seinem Cache, um zu ermitteln, an welchen Nachfahren er die Daten weiterleiten muß. So erreichen sie schließlich die Basisstation, in deren Funkzelle sich der Mobile Node befindet.

Wechselt nun der Mobile Node die Basisstation, so reicht es, eine Nachricht an den Wurzelrouter zu senden. Dadurch wird automatisch wieder ein neuer Pfad in der Baumstruktur erstellt. Der alte Pfad bleibt jedoch erhalten. Insbesondere existieren nun im gemeinsamen Vorfahren von der alten und der neuen Basisstation zwei Registrierungseinträge für ein und den selben Mobile Node. Ankommende Nachrichten an den Mobile Node werden hier dupliziert und auf beiden Wegen weitergeleitet. Befindet sich der Mobile Node gerade an der Grenze der beiden Funkzellen, so ist auf dieser Weise sichergestellt, daß zumindest eine Nachricht ankommt. Werden beide Nachrichten empfangen, können die tiefer liegenden Schichten des Protokolls dies erkennen und eine Nachricht entfernen. Der alte Pfad erlischt, wenn die Frist für die Einträge abgelaufen ist.

Die Wahl einer günstigen Frist, in der die Einträge in den Caches gültig sein sollen, hängt von dem Zustand des Mobile Node ab. Erwartet der Mobile Node keine Daten,

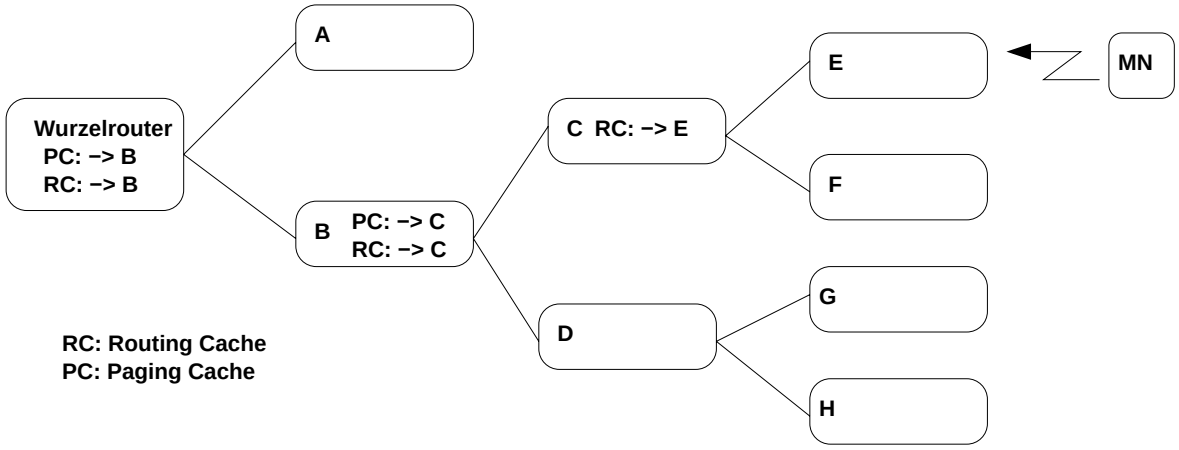

Abbildung 4: *Routing bei Cellular IP:* Die Nachricht an den Mobile Node werden mit Hilfe der Einträge in den Caches weitergeleitet. Paging Cache müssen dabei nicht in jedem Knoten vorhanden sein.

ist eine möglichst lange Frist sinnvoll. Dadurch ist es selten notwendig, daß der Mobile Node Daten sendet, um die Einträge in den Caches aufzufrischen. Dies spart Energie bei den Mobile Nodes, welche oft mit Batterien auskommen müssen, aber auch die Übertragungskapazität einer Funkzelle wird dadurch nicht unnötig belastet. Erwartet oder empfängt hingegen der Mobile Node Daten, so ist eine kurze Frist für das Auffrischen der Einträge in den Caches effektiver. Dadurch kommt es seltener vor, daß ein Mobile Node noch bei mehreren Basisstationen gleichzeitig eingetragen ist, und somit werden auch seltener Duplikate erzeugt. Diese würden nur die Sendestationen und die Verbindungsleitungen unnötig belasten.

Um keinen Spagat zwischen zwei widersprüchlichen Vorgaben machen zu müssen, werden diese zwei Zustände bei Cellular IP unterschieden (Abbildung 4). So werden die Caches der Knoten unterteilt in einen Routing Cache und einen Paging Cache. Im Routing Cache stehen die Einträge der Mobile Nodes, wenn Daten empfangen werden. Hier sind die Zeitspannen, die diese Einträge gültig sind, eher kurz. Somit muß der Mobile Node oft Daten senden, um seine Position möglichst aktuell anzugeben. Empfängt der Mobile Node hingegen keine Daten, so stehen die Einträge im Paging Cache. Diese müßen seltener aufgefrischt werden. Um einzelne Knoten möglichst einfach zu gestalten und somit kostengünstig zu machen, ist es nicht einmal notwendig, jeden Knoten mit einem Paging Cache zu versehen. Knoten ohne Paging Cache senden Nachrichten an alle Nachkommen. Empfängt der Mobile Node dann diese Daten, wechselt sein Zustand auf Empfang, und er meldet sich mit einer Nachricht an den Wurzelrouter, sodaß wieder ein aktueller Pfad zu ihm existiert.

4 Optimierungen beim Routing

Befindet sich ein Mobile Node in einem Foreign Network, ergibt sich bei der Kommunikation mit dem Correspondent Node ein Dreieck, da die Antworten von dem Correspondent Node erst über den Home Agent geleitet werden. Dies ist insbesondere dann unerwünscht, wenn sich der Correspondent Node im selben Subnetz befindet wie der Mobile Node. Auch wird der Home Agent stark belastet, wenn dieser für mehrere Mobi-

le Nodes zuständig ist. Um dieses Dreieck zu unterbinden wurde Mobile IP mit Routing Optimierung ausgebaut [PeJo00].

Bei der Routing Optimierung speichern die Correspondent Nodes die aktuelle Care-of-Address für die Mobile Nodes ab, mit denen sie in Verbindung stehen. Statt die Daten dann erst an den Home Agent zu schicken, tunneln die Correspondent Nodes die Daten direkt zum Foreign Agent des Mobile Node. Dadurch wird das Dreieck eingespart. Wurde die aktuelle Care-of-Address eine Zeitlang gespeichert, ohne erneut gesetzt worden zu sein, wird sie gelöscht und eventuell noch anfallende Daten an die normale IP-Adresse des Mobile Node geschickt. Ist der Mobile Node immer noch in einem Foreign Network, so werden diese Datenpakete vom Home Agent abgefangen und an die Care-of-Address des Mobile Node weitergeleitet. Gleichzeitig teilt der Home Agent dem Correspondent Node die aktuelle Care-of-Address des Mobile Node mit, sodaß der Correspondent Node die Daten wieder direkt an den Foreign Agent des Mobile Node schicken kann. Bei jeder Aktualisierung der Care-of-Address bei einem Correspondent Node muß sich der Home Agent authentifizieren.

Wechselt nun der Mobile Node seinen Foreign Agent, so müssen auch die Correspondent Nodes verständigt werden. Dies übernimmt der Home Agent. Bekommt er den neuen Aufenthaltsort des Mobile Node mitgeteilt, so sendet der Home Agent eine Aktualisierung der neuen Care-of-Address an die entsprechenden Correspondent Nodes. Dazu hält er eine Liste der Correspondent Nodes bereit, die die letzten Aktualisierungen der Care-of-Address quittiert haben. Doch nicht nur der Home Agent muß über einen Wechsel des Foreign Agent informiert werden. Auch der vorhergehende Foreign Agent muß über den neuen Aufenthaltsort des Mobile Node Bescheid wissen. Dafür erzeugt der Mobile Node schon für das Anmelden beim neuen Foreign Agent eine authentifizierte Botschaft an den vorhergehenden Foreign Agent. Der neue Foreign Agent leitet diese Botschaft direkt weiter an den vorhergehenden Foreign Agent. Für diesen Zweck wurde die Previous Foreign Agent Notification Extension (PFANE) definiert, die angibt, wie die Adresse des vorhergehenden Foreign Agent in der Registration Request -Nachricht angegeben wird. Der vorhergehende Foreign Agent setzt dann einen Eintrag, damit alle noch bei ihm ankommende Daten für den Mobile Node direkt zum neuen Foreign Agent getunnelt werden. Positiver Nebeneffekt ist dabei, daß der vorhergehende Foreign Agent gleich eventuelle Ressourcen, die für den Mobile Node reserviert waren, freigeben kann. Datenpakete, die beim normalen Mobile IP verloren gingen, weil der Home Agent sie nach dem Wechsel des Foreign Agent noch an den vorhergehenden Foreign Agent geschickt hatt, können nun auch ans Ziel geleitet werden.

Bekommt nun der vorhergehende Foreign Agent noch Daten für den Mobile Node, so tunnelt er diese weiter an den neuen Foreign Agent. Gleichzeitig meldet er dem Home Agent diesen Vorfall, der daraufhin diesem Correspondent Node die neue Care-of-Address mitteilt.

5 FATIMA

Das Internet hat in der heutigen Wirtschaft einen großen Stellenwert eingenommen. Über das Internet werden Einkäufe getätigt oder Konten bei Banken verwaltet. Für eine Firma kann es teuer werden, wenn ihr Internetanschluß ausfällt. Deshalb wird

immer mehr Wert auf die sichere Abschottung des eigenen Subnetzes gelegt. Diese Aufgabe übernehmen Firewalls.

Bei Mobile IP ist vorgesehen, daß der Mobile Node als Absender seine richtige Adresse von seinem Home Network angibt. Befindet sich nun der Mobile Node im Foreign Network, so ist diese Absenderadresse topologisch falsch. Übliche Firewalls erkennen dies und interpretieren es als Vortäuschung falscher Tatsachen (Spoof-Attack). Sicherheitshalber werden deshalb solche Nachrichten herausgefiltert. Möchte man nun ein Subnetz zu einem Foreign Network ausbauen, müßte diese Funktion bei der Firewall ausgeschaltet werden. Dies würde aber zu einem Verlust an Sicherheit führen. Um die Sicherheit zu wahren, wurde die Firewall-Aware Transparent Mobility Architecture (FATIMA) entwickelt [MPSS00].

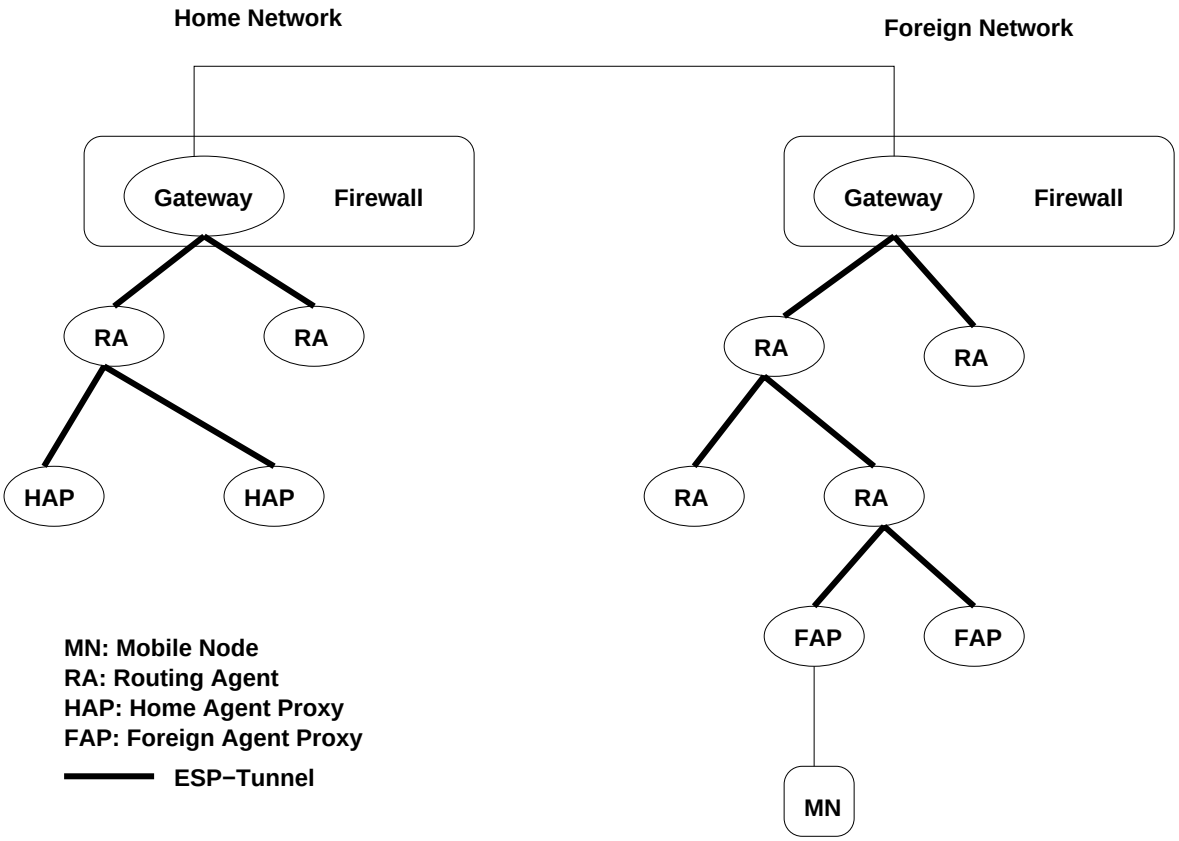

Abbildung 5: *Netzaufbau bei FATIMA:* Der Datenverkehr des Mobile Node wird, so lange es geht, durch ESP-Tunnel geleitet.

FATIMA betrifft sowohl das Foreign Network als auch das Home Network. Alle sicherheitsrelevanten Einstellungen sind in der jeweiligen Gateway konzentriert, um sie leicht administrieren zu können. Die Gateway tritt nach außen hin als Home Agent bzw. als Foreign Agent auf. Um Skalierbarkeit zu gewährleisten, ist eine Baumstruktur aus Routing Agents im jeweiligen Netzwerk möglich. An den Blättern sitzen dann Foreign Agent Proxys bzw. Home Agent Proxys. Um vorgetäuschte Daten ausschließen zu können, werden für alle Verbindungen in der Baumstruktur das Encapsulated Security Payload -Protokoll (ESP) benutzt. Hierbei findet eine Authentifizierung des jeweiligen Absenders statt. Optional kann auch eine Verschlüsselung der Daten verlangt werden. Die Foreign Agent Proxys stellen den Kontakt zu einem Mobile Node her. Als Care-of-Address wird die Adresse der Gateway bekannt gegeben. Die Registration Re-

quest -Nachricht wird an die Gateway des Foreign Networks über die Baumstruktur weitergeleitet. Jeder dabei passierte Foreign Routing Agent speichert in einem Eintrag, woher der Registrierungswunsch kam, um Daten auch wieder zurücksenden zu können. Doch ist dieser erste Eintrag zeitlich nur stark begrenzt haltbar. Die Gateway kann nun überprüfen, ob sie den Mobile Node zulassen will und sendet die Registration Request -Nachricht weiter an den Home Agent. Ist das Home Network ebenfalls ein FATIMA-Network, wird auch hier die Registration Request -Nachricht mit Hilfe der IP-Adresse des Mobile Node über die Baumstruktur an einen Home Agent Proxy weitergeleitet. Der Home Agent Proxy setzt einen Eintrag für den aktuellen Aufenthaltsort des Mobile Node und quittiert den Registrierungswunsch mit der Registration Reply -Nachricht. Diese wird an den Foreign Agent weitergeleitet, von ihm in der Baumstruktur wieder nach unten gereicht und vom Foreign Agent Proxy an den Mobile Node übermittelt. Beim Herunterreichen werden dann auch die bisher nur zeitlich stark begrenzten Einträge für das Routen auf eine längere Gültigkeitsdauer gesetzt. Die Kommunikation zwischen dem Home Agent und dem Foreign Agent muß dabei immer authentifiziert werden. Dafür muß ein gemeinsames Geheimnis zwischen jedem Foreign Agent und Home Agent existieren, wenn keine Public Key Infrastruktur besteht, jedoch nicht zwischen jedem Foreign Agent Proxy und Home Agent Proxy, was die Zahl der benötigten Geheimnisse stark reduziert.

Ändert nun ein Mobile Node seinen Zugriffspunkt innerhalb eines Foreign Networks, so muß sie sich erneut authentifizieren. Im Gegensatz zu HAWAII oder Cellular IP ist deshalb auch eine Kontaktierung des Home Agent notwendig. Datenverkehr wird nur in soweit eingespart, als daß ein eventuelles Verständigen eines Correspondent Node bei Routing Optimization nicht notwendig ist, da die Care-of-Address gleich bleibt. Um dennoch die Mobilität innerhalb des Foreign Network zu unterstützen, ist eine Erweiterung des Protokolls um die Fast Handoff Extension möglich [Mink99]. Stellt der Mobile Node fest, daß er den Anschluss nur innerhalb des Foreign Networks gewechselt hat, kann er sich bereit erklären, nur mit dem Foreign Agent die neue Registrierung durchzuführen. Dabei reicht es aus, wenn sich der Mobile Node und der Foreign Agent authentifizieren.

6 Authentifizierung Autorisierung und Abrechnung (AAA)

Mobile IP sieht in seiner Basisdefinition keine obligatorische Authentifizierung zwischen dem Foreign Agent und dem Mobile Node bzw. dem Home Agent vor. Nun sind aber solche Authentifizierungen gerade für die micro-Mobilität von Nöten, um große Sicherheitslücken zu vermeiden. Mit Hilfe einer AAA-Infrastruktur ist ein Erzeugen von Schlüsseln für die Authentifikation zwischen dem Foreign Agent und dem Mobile Node möglich.

Ein Mobile Node, das in ein fremdes Netz kommt, sendet sein Registration Request -Nachricht an den Foreign Agent. Dieser leitet die Nachricht an den AAA-Server in seinem Netz weiter. Der AAA-Server des Foreign Network macht den AAA-Server des Home Network ausfindig und sendet ihm die Registration Request -Nachricht. Der Nachrichtenaustausch zwischen den AAA-Servern ist dabei authentifiziert und verschlüsselt. Der Datenaustausch zwischen den AAA-Servern kann auch über einen Broker ablaufen,

falls die AAA-Server keinen direkten gemeinsamen Schlüssel besitzen. Der AAA-Server des Home Network entschlüsselt die Nachricht und bestimmt den zuständigen Home Agent. Zusätzlich generiert der AAA-Server Schlüssel für die Authentifikation zwischen dem Foreign Agent und dem Mobile Node (FA-MN-Key) und dem Foreign Agent und dem Home Agent (FA-HA-Key). Die Schlüssel und die Registration Request -Nachricht wird an den Home Agent gesendet. Der Home Agent nimmt die notwendigen Einträge für den Mobile Node vor und sendet eine Registration Reply -Nachricht mitsamt den Schlüsseln zurück an den AAA-Server im Home Network. Der AAA-Server verschlüsselt die Nachricht und leitet sie weiter an den AAA-Server im Foreign Network. Hier werden die Nachrichten entschlüsselt, und, für den Internen Datenverkehr wieder verschlüsselt, an den Foreign Agent geschickt. Dieser meldet dann die Registration Reply -Nachricht und die Schlüssel weiter an den Mobile Node. Damit ist ein authentifizierter Nachrichtenaustausch zwischen dem Foreign Agent und dem Mobile Node bzw. dem Home Agent möglich, ohne vorher für jedes mögliche Paar einen Schlüssel definieren zu müssen.

7 Zusammenfassung

Die zunehmende Verbreitung von mobilen Computer macht die Notwendigkeit von mobiler Kommunikation deutlich. Doch ist Mobile IP in erster Linie für die macro-Mobilität konzipiert worden. Für die micro-Mobilität wurde es nicht entwickelt. Da aber mobile Geräte hauptsächlich über Funkschnittstellen arbeiten, ist eine Unterstützung der micro-Mobilität gefordert. Die Verbesserungsansätze, um dies möglich zu machen, bieten aber nicht die notwendige Sicherheit bei der Authentifizierung der Nachrichten zwischen dem Foreign Agent und dem Mobile Node. Eine Authentifizierung war im Grundkonzept von Mobile IP auch nur zwischen dem Home Agent und dem Mobile Node vorgesehen. Um einen Austausch authentifizierter Nachrichten ohne den Umweg über den Home Agent zu ermöglichen, ist eine Infrastruktur mit z.B. AAA-Servern notwendig. Es wird sich zeigen, ob dieser Ansatz sich durchsetzt, oder ob es andere Verfahren gibt, die micro-Mobilität und Sicherheit von Anfang an berücksichtigen.

Literatur

[Mink99] S. Mink. Konzeption einer Firewall-Architektur für Mobile IP. *Diploma Thesis, Universität Karlsruhe (TH), Institut für Telematik*, Oktober 1999.

[MPSS00] Stefan Mink, Frank Pählke, Günter Schäfer und Jochen Schiller. FATIMA: A Firewall-Aware Transparent Internet Mobility Architecture. In *Fifth IEEE Symposium on Computers and Communications (ISCC 2000)*, Antibes, France, Juli 2000.

[PeJo00] Charles Perkins und David B. Johnson. Route Optimiziation in Mobile IP. *Internet Draft „draft-ietf-mobileip-optim-10.txt“*, November 2000.

[Perk] Charles E. Perkins. Mobile Networking Through Mobile IP. *Internet Page „HTTP://computer.org/internet/v2n1/perkins.htm“.*

[Perk96] C. Perkins. IP Mobility Support. *RFC 2002*, Oktober 1996.

[RPTV+99] R. Ramjee, T. La Porta, S. Thuel, K. Varadhan und L. Salgarelli. IP micro-mobility support using HAWAII. *Internet Draft „draft-ietf-mobileip-hawaii-00.txt“*, Juni 1999.

[Valk99] András G. Valkó. Cellular IP: A New Approach to Internet Host Mobility. *ACM Computer Communication Review* 29(1), Januar 1999, S. 50 – 65.

Abbildungsverzeichnis

Alternative Transportprotokolle für drahtlose Netze

Colin Schulz

Kurzfassung

TCP hat sich als ungeeignet für den Einsatz in drahtlosen Netzen herausgestellt.Dies liegt daran, daß die in Funknetzwerken häufigen Paketverluste von TCP als Stauanzeichen interpretiert werden. TCP führt alle Segmentverluste auf einen Stau im Netzinneren zurück und drosselt somit seine Senderate. Dies verursacht eine deutliche Verschlechterung des Durchsatzes auf der drahtlosen Strecke. In dieser Arbeit sollen zunächst Gründe für die schlechte Performance von TCP im drahtlosen Bereich untersucht werden. Danach werden die einzelnen Lösungsansätze klassifiziert und vorgestellt. Zum Schluss wird noch auf den Sonderfall der Ad-hoc Netzwerke eingegangen.

1 Einleitung

Der aufkommende Boom für mobile und tragbare Computer verlangt Netzwerktechnologien, die den Benutzer in die Lage versetzen, jederzeit und ohne die Notwendigkeit von Kabeln oder einer festen Infrastruktur auf Informationen zugreifen zu können. Hierfür müssen nicht nur Verbesserungen im Bereich der Hardware erfolgen, auch die bestehenden Protokolle müssen für das mobile Umfeld angepasst werden. Intensive Forschungen wurden hierfür im Bereich von Mobile IP betrieben, welches mittlerweile zu einem von der IETF vorgeschlagenem Standard ausgereift ist. Seit einigen Jahren konzentrieren sich die Bemühungen vermehrt auch auf die Transportschicht. Hier spielt TCP als verbindungsorientiertes und zuverlässiges Protokoll eine zentrale Rolle. Als eines der größten Probleme von TCP im mobilen Umfeld erweisen sich die in TCP vorhandenen Mechanismen zur Flusskontrolle. Diese interpretieren jeden Paketverlust als Anzeichen eines Staus und veranlassen somit eine Drosselung der Senderate. Da im mobilen Bereich Paketverluste keine Seltenheit sind, erfährt ein unverändertes TCP hier drastische Performanceeinbussen.

.

2 Probleme von TCP im mobilen Umfeld

Die grundsätzliche Aufgabe von TCP ist es, zuverlässige Ende-zu-Ende Verbindungen zwischen zwei Anwendungen anzubieten. TCP überträgt die Daten der Anwendungsschicht in Segmenten, die durchnummeriert sind. Der Empfänger schickt eine Quittung

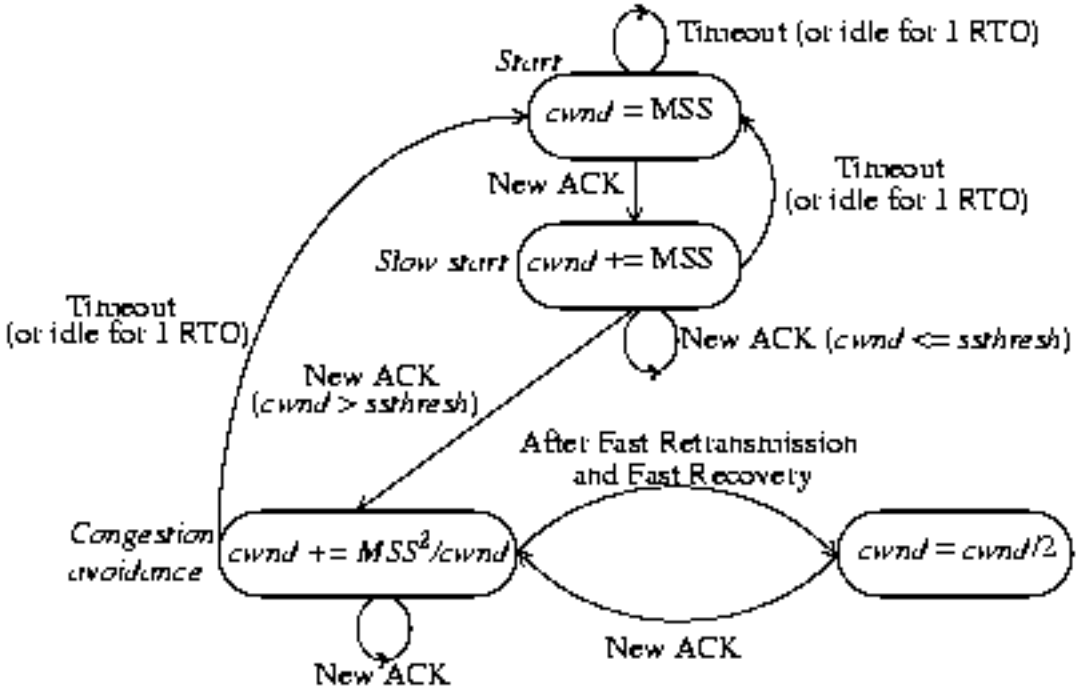

Abbildung 1: Zustandsautomat für das Sendefenster

(ACK) für jedes korrekt empfangene Segment. Um die korrekte Übertragung sicherzustellen, berechnet TCP eine Prüfsumme. TCP garantiert eine zuverlässige Auslieferung der Daten.

2.1 Flusskontrolle

Darüberhinaus hat TCP noch einen Mechanismus zur Flusskontrolle, der bewirken soll, daß sich Stausituationen im Netz schnell wieder auflösen. Bemerkt TCP anhand eines Paketverlustes einen Stau im Netz, so drosselt es seine Senderate. Hierfür unterhält jeder Sender fortlaufend mehrere Daten über eine TCP Verbindung:

1. ein Staufenster(Congestion Window) welches angibt, wie viele Segmente gesendet werden dürfen ohne eine Bestätigung zu erhalten
2. ein RTO (Retransmission Timeout) Wert. Ist nach Ablauf der RTO-Zeit für ein gesendetes Segment noch keine Quittung eingetroffen, so wird der Verlust des Segmentes angenommen und die Übertragung wiederholt. Der RTO Wert berechnet sich aus dem Mittelwert der gemessenen Paketumlaufzeit (RTT-Round Trip Time) plus vier mal deren Standardabweichung.

2.1.1 Der Slow-Start Algorithmus

Am Anfang einer Verbindung, aber auch nach jedem Timeout für den Empfang eines ACK startet der Sender im Slow-Start Modus. Der Slow-Start Algorithmus funktioniert folgendermaßen:

Der Sender startet mit einer Fenstergröße von 1, er darf also ein Segment senden. Erhält er für dieses Segment innerhalb der vorgesehenen Zeit eine Bestätigung, so darf er das Staufenster verdoppeln, also zwei Segmente senden. Erhält er auch für diese eine Bestätigung, so darf er danch 4 Segmente auf einmal schicken usw. Die Größe des Staufensters steigt also exponentiell, bis sie einen Schwellwert erreicht, danach geht das exponentielle Wachstum in ein lineares über. Die Phase, in der die Fenstergröße nur noch linear steigt, nennt man Congestion-Avoidance- Phase. Sie ist notwendig, da bei rein exponentiellem Wachstum irgendwann die Schritte der Vergrösserung sehr groß werden würden und auf einmal einen Stau auslösen könnten.

Es gibt jetzt zwei Situationen, in denen TCP Staukontrollmechanismen ergreift:

1. Eine fehlende Bestätigung für ein Segment, d.h. der Zeitgeber (RTO) für den Empfang der Bestätigung läuft aus. Hier leitet TCP die folgenden Massnahmen ein:

   - das Segment wird erneut übertragen
   - die Größe des Staufensters wird auf 1 gesetzt
   - der Stauschwellenwert wird auf die Hälfte der Größe, welches das Staufenster vor dem Timeout hatte, gesetzt.
   - der RTO Wert wird verdoppelt

2. Der Empfang von drei gleichen Bestätigungen (DUPACKS) für ein Segment. Die Bestätigungen in TCP sind kumulativ, d.h. sie geben an, bis zu welchem Segment der Empfänger *alle* Segmente korrekt empfangen hat. Für Segmente, die zwar korrekt, aber nicht in der richtigen Reihenfolge empfangen wurden (etwa weil ein vorgehendes Segment verloren ging oder im Netz länger verzögert wurde) kann der Empfänger somit nur die selbe Quittung schicken wie für das letzte in Reihenfolge empfangene Segment. Nun kann es immer wieder mal vorkommen, daß Pakete im Netz umgeordnet werden, z.Bsp. weil sie unterschiedliche Wege gehen. Deshalb wartet TCP auf den Empfang von drei DUPACKS, bevor es den Verlust des Segmentes annimmt. In den früheren TCP Versionen wurden nach dem Empfang von drei DUPACKS die selben Maßnahmen ergriffen wie nach einem Timeout, d.h. der Sender ging in den slow-start Modus. Die Tatsache, dass ein Empfänger DUPACKS sendet zeigt jedoch, daß er weiterhin Daten vom Sender erhält. Deshalb kann davon ausgegangen werden, dass keine schwere Stausituation vorliegt. Alles in allem werden nach dem Empfang von drei DUPACKS die folgenden Maßnahmen ergriffen:

   - das letzte unbestätigte Segment wird sofort (ohne auf den Ablauf des Timers zu warten) erneut übertragen. Dieses Verhalten nennt man schnelle Übertragungswiederholung (Fast Retransmit)
   - der Stauschwellenwert wird auf die Hälfte der Größe, welches das Staufenster vor dem Fast Retransmit hatte gesetzt.
   - das Staufenster wird auf den neuen Stauschwellenwert gesetzt und der Sender fährt im Congestion Avoidance Modus fort.

   Da der Sender nicht den Slow-Start Mechanismus aktiviert, nennt man dieses Verhalten Fast Recovery, da er sich schnell von dem Paketverlust erholt.

2.2 Auswirkungen der Mobilität auf TCP

Man sieht, daß diesen Mechanismen zur Staukontrolle eine Grundannahme vorausgeht, und zwar daß Paketverluste ausschließlich aufgrund von Stausituationen entstehen. TCP wurde für den Einsatz in Festnetzen entwickelt. Diese arbeiten sehr zuverlässig, weshalb Verluste durch Übertragungsfehler eher selten sind. Ein Paketverlust ist mit hoher Wahrscheinlichkeit auf eine Überlastung an irgendeinem Punkt auf dem Weg durch das Netz zurückzuführen. Einem überlasteten Router, dessen Pufferkapazität erschöpft ist, bleibt nichts anderes übrig als neu ankommende Pakete zu verwerfen.

In drahtlosen Netzen ist die Annahme, daß Paketverluste nur auf Grund einer Stausituation entstehen nicht mehr korrekt. Hier gibt es viele mögliche Ursachen für einen Paketverlust:

- eine sehr hohe Bitfehlerrate aufgrund von physikalischen Übertragungsfehlern durch Abschattung, Mehrwegeausbreitung, Interferenz, etc.
- Verbindungsunterbrechungen, wenn physikalische Hindernisse die Funksignale blockieren
- Unterbrechungen der Übertragung bzw Fehlweiterleitungen während eines Handover

In drahtlosen Netzen treten Paketverluste demnach viel häufiger auf als in Festnetzen. TCP interpretiert jedoch jeden Segmentverlust als Anzeichen für einen Stau im Netz und drosselt folglich seine Senderate. Dies führt dazu, daß die Performance des Transportprotokolls drastisch einbricht.

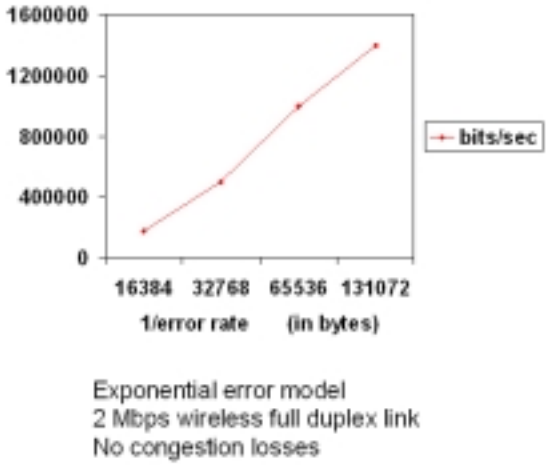

Abbildung 2: Auswirkung verschiedener Fehlerraten auf den Durchsatz von TCP

Da TCP ein fester Bestandteil des heutigen Internet ist und auch wesentlich zu dessen Stabilität beiträgt, kann man das Transportprotokoll nicht grundsätzlich verändern, indem man beispielsweise den Slow-Start Algorithmus verändert oder abschafft. Vielmehr sollte jede Verbesserung kompatibel mit den bestehenden Implementierungen sein und idealerweise nur Veränderungen beim mobilen Knoten erfordern.

In den letzten Jahren wurden eine Reihe von Erweiterungen und Ergänzungen von TCP vorgestellt, die das Problem TCP über drahtlosen Link beheben sollten. Diese

Verfahren unterscheiden sich in vielerlei Hinsicht so daß eine Klassifizierung nach den folgenden Gesichtspunkten möglich ist:

1. nach dem gewählten Problemschwerpunkt

   - die hohe Bitfehlerrate
   - kurze und lange Verbindungsunterbrechungen
   - Unterbrechungen/Verluste während eines Handover

   Anhand des Problemschwerpunktes lässt sich auch die Eignung der Verfahren für unterschiedliche Netze feststellen, da in jedem Szenario (drahtloses LAN, ad-hoc-Netzwerk, Mobilfunknetz) ein anderes Problem im Vordergrund steht.

2. hinsichtlich dem Ort ihrer Implementierung:

   - beim Sender (im Festnetz angenommen)
   - beim Empfänger
   - in einem Zwischensystem
   - einer Kombination dieser Punkte

3. nach der Art des Ansatzes:

   - Ende-zu-Ende.
   - Split connection
   - Link Layer

3 Mechanismen für das Problem der hohen Verlustrate

3.1 Performance Enhancing Proxys

Ein PEP ist ein Zwischensystem, welches dazu benutzt wird, die negativ auf die Performance wirkenden Charakteristiken eines Links durch spezielle Mechanismen auszugleichen.

3.1.1 Indirect TCP

Indirect TCP teilt die Verbindung in zwei Teilverbindungen: Eine für den Festnetzteil und die andere für die drahtlose Strecke. Der Trennungspunkt kann der Fremdagent oder die Basisstation im Mobilfunknetz sein. Pakete, die für den mobilen Knoten bestimmt sind werden nun vom Proxy quittiert und zwischengespeichert. Danach kümmert sich der Proxy um die zuverlässige Auslieferung an den mobilen Knoten. Dabei kann er ein spezielles, an die Eigenschaften der drahtlosen Strecke angepasstes TCP verwenden. Aber auch ein unverändertes TCP erzielt wegen der kürzeren RTT auf der Teilstrecke eine bessere Performance. Dieser Ansatz hat folgende Vorteile:

- Fehler auf der drahtlosen Strecke pflanzen sich nicht ins Festnetz fort
- gute Performance aufgrund der kurzen RTT auf der drahtlosen Teilstrecke und somit schnelle Erholung von Übertragungsfehlern

Ein großer Nachteil ist der Verlust der Ende-zu-Ende Semantik von TCP. Nach Erhalt einer Quittung denkt der Sender, daß sein Segment vom Empfänger richtig empfangen wurde. Die Quittung bedeutet jedoch lediglich, dass der Proxy das Segment empfangen hat und es ist nicht garantiert, daß er das Segment auch ausliefern können wird. Des weiteren sind grössere Puffer in den Trennungspunkten notwendig. Durch die Zwischenspeicherung der Pakete erhöht sich auch die für ein Handover benötigte Zeit, da alle vom Proxy schon quittierten, aber noch nicht ausgelieferten Segmente an den neuen Proxy weitergeleitet werden müssen.

3.1.2 Snooping TCP

Einer der größten Nachteile von I-TCP ist der Verlust der Ende-zu-Ende Semantik. Der Fremdagent bestätigt hier Segmente, die der mobile Knoten noch gar nicht erhalten hat. Dies verhindert Snooping TCP indem es einen sogenannten Schnüffelagenten im Zugangspunkt einführt. Dieser hört den Paketfluss in beide Richtungen mit. Des weiteren puffert er alle Segmente die an den mobilen Knoten gehen und kann somit im Fall eines Verlustes auf der drahtlosen Strecke eine lokale Übertragungswiederholung durchführen. Eine weitere Aufgabe des Schnüffelagenten ist das Herausfiltern von doppelten Bestätigungen (DUPACKs) die vom mobilen Knoten kommen. Dadurch werden unnötige Fast-Retransmits beim Sender verhindert. In der Literatur wird Snoop als Link-Layer Ansatz, der Kenntnis von TCP hat bezeichnet. Snoop führt lokale Übertragungswiederholungen durch, benutzt hierfür jedoch die Bestätigungen der Transportschicht.

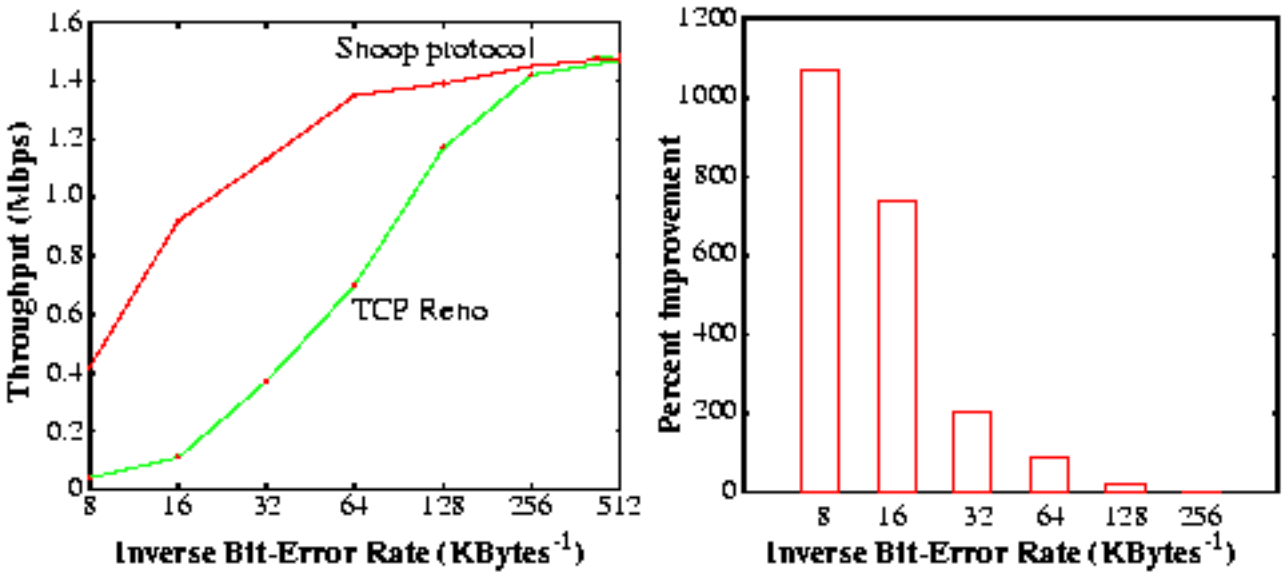

Abbildung 3: Vergleich der Performance von TCP Reno und dem Snoop Protokoll bei verschiedenen Fehlerraten.

3.2 Link-Layer-Ansätze

Die Grundidee von Link-Layer-Ansätzen ist es, die hohen Bitfehlerraten auf dem drahtlosen Link durch Mechanismen der Sicherungsschicht auszugleichen. Dies kann durch Übertragungswiederholungen geschehen, aber auch durch Techniken wie Vorwärtsfehlerkorrektur(FEC). Während man durch Vorwärtsfehlerkorrektur die Fehlerraten auf ein erträgliches Niveau abzusenken versucht, kann man durch Übertragungswiederholungen den Link-Layer zuverlässig machen. Der Vorteil von lokalen Übertragungswiederholungen ist, daß sie das Problem dort beheben wo es entsteht, nämlich auf dem drahtlosen Link. Es gibt hier verschiedene Implementierungen, die sich in den folgenden Punkten unterscheiden:

1. Wie oft wird die Übertragung wiederholt?
   - begrenzte Anzahl von Übertragungsversuchen
   - unbegrenzte Anzahl von Wiederholungen, also bis die Übertragung erfolgreich war (zuverlässiger LL)
2. Was löst die Übertragungswiederholung aus?
   - Link Layer Timeout
   - Link Layer ACKs(NACKS)
   - andere Mechanismen (wie bei Snoop, wo die ACKs der Transportschicht benutzt werden)

Trotz optimierter lokaler Übertragungswiederholungen kann TCP eine schlechte Performance aufweisen, und zwar aus den folgenden Gründen:

1. der Versuch, die Daten lokal zu übertragen dauert zu lange und löst somit trotzdem ein Timeout bei TCP aus
2. durch nicht reihenfolgetreue Auslieferung der Segmente wird unnötigerweise ein Fast Retransmit bei TCP ausgelöst

Die meisten LL-Protokolle streben keine Auslieferung der Pakete in der richtigen Reihenfolge an. Gehen bei einem TCP Empfänger jedoch Segmente in der falschen Reihenfolge ein, so führt dies zur Generierung von DUPACKs, diese wiederum veranlassen den Sender eine schnelle Übertragungswiederholung durchzuführen. Es hat sich gezeigt, daß das erste Problem (Übertragungswiederholung durch Timeout bei TCP) weitaus weniger gravierend ist als das zweite. Dies liegt daran, daß bei TCP die RTO-Werte grobkörniger berechnet werden (sie betragen meistens ein Vielfaches von 500 ms). Dies erlaubt eine gewisse Anzahl von Link-Layer Übertragungswiederholungen, bevor ein Timeout bei TCP auftritt. Das zweite Problem ist schwerwiegender, da bei einem Fast Retransmit immer gleichzeitig auch Staukontrollmaßnahmen ergriffen werden. Um ein unnötiges Auslösen eines Fast Retransmit zu vermeiden sollte das LL-Protokoll möglichst eine in-order Auslieferung der Pakete anstreben.

3.2.1 Delayed DUPACKS

[Univ97] stellten 1999 einen Ansatz vor, der das durch Paketumordnungen verursachte Problem beheben soll. Empfängt der mobile Knoten ein Segment out-of-order, so schickt er seine Bestätigung (ein DUPACK) hierfür nicht sofort, um der Link-Level Übertragungswiederholung Zeit zu geben, das fehlende Segment nachzuliefern. Dadurch soll vermieden werden, daß der Sender ein Fast Retransmit durchführt, obwohl das Segment nicht verloren ging, sondern nur länger verzögert wurde. Allerdings kann er seine DUPACKs nicht für unbegrenzte Zeit zurückhalten, da sonst ein viel schlimmerer Fall eintreten würde, nämlich ein Timeout beim TCP-Sender. Aus diesem Grund kann dieser Ansatz die unnötigen fast-retransmissions zwar reduzieren, aber nicht vollständig beseitigen.

3.2.2 Der Eifel Algorithmus

Auch bei einem hoch angesetzten RTO-Wert kann es vorkommen, daß es beim Sender zu einem Timeout kommt und er das Segment erneut überträgt, obwohl dieses (z.Bsp durch mehrfache LL-Übertragungsversuche) nur länger verzögert wurde. Kurz nachdem der Sender das Segment erneut übertragen hat, trifft nun doch die Quittung für das ursprünglich übertragene Segment ein. Der Sender kann jedoch aus diesem ACK nicht erkennen, ob es sich für die Quittung für das ursprünglich oder das erneut übertragene Segment handelt. Deshalb muss er annehmen, dass das ACK seine Übertragungswiederholung bestätigt und kann auch die nach dem Timeout eingeleiteten Staumaßnahmen nicht rückgängig machen. Die selbe Situation tritt auch nach einer aufgrund von 3 DUPACKs durchgeführten Übertragungswiederholung ein. Auch hier kann der Sender nicht unterscheiden, ob es sich bei der eintreffenden Quittung um die Bestätigung des Originals oder der Wiederholung handelt. Der Grundgedanke des Eifel-Algorithmus ist es, den Sender mit den notwendigen Informationen auszustatten, um die ACKs voneinander unterscheiden und somit wissen zu können, wann eine Übertragungswiederholung unnötig stattgefunden hat.

Der Eifel Algorithmus läßt den Sender in jedes abgehende Segment einen Zeitstempel schreiben. Der Empfänger gibt diese Zeitstempel in seinen ACKs zurück . Der Sender speichert bei sich den Zeitstempel jeder Übertragungswiederholung. Nun kann durch einen einfachen Vergleich mit dem Zeitstempel der ankommenden Bestätigung feststellen, welche Übertragung bestätigt wird. Merkt der Sender, dass eine Übertragungswiederholung unnötig stattgefunden hat, so kann er die eingeleiteten Staukontrollmaßnahmen rückgängig machen.

3.3 Ende-zu-Ende Mechanismen

3.3.1 Efficient fast retransmit (EFR)

Ein weiterer Grund für schlechte TCP Performance kann das häufige Warten auf das Auslaufen des Timers für die Übertragungswiederholung sein. Geht ein einzelnes Segment (z.B. aufgrund eines Übertragungsfehlers) verloren, so wiederholt der TCP-Sender die Übertragung ja erst nach dem Auslaufen der RTO-Zeit, es sei denn, er erhält vorher 3 DUPACKs und kann eine schnelle Übertragungswiederholung durchführen. Oftmals

werden vom Empfänger aber gar nicht genügend DUPACKs gesendet, um beim Sender eine schnelle Übertragungswiederholung auslösen zu können. Dies ist immer der Fall, wenn das Sendefenster eine Größe zwischen einem und vier Segmenten hat. In diesen Fällen wartet der Sender unnötig mit seiner Übertragungswiederholung. Außerdem muss er die nach einem Timeout in den Slow-Start Modus gehen.

Die von EFR vorgeschlagene Lösung ist relativ einfach: Der Sender berechnet fortlaufend die Anzahl der DUPACKs die bei ihm einen Fast Retransmit auslösen. Ist also das Sendefenster klein, so wird auch der Schwellwert für ein Fast Retransmit herabgesetzt.

Simulationen haben gezeigt, dass dieser simple Ansatz eine Leistungssteigerung von 10 -15 Prozent erreicht.

3.3.2 Limited Transmit

Ähnlich wie EFR will auch dieser Ansatz unnötige Timeouts bei kleinem Sendefenster vermeiden. Hierfür wird dem Sender erlaubt, für jedes der ersten zwei ankommenden DUPACKs noch ein weiteres *neues* Segment zu senden. Hat der Sender also beispielsweise bis Segment fünf gesendet(bei einer Fenstergröße von zwei), und erhält nun ein DUPACK das den korrekten Empfang bis Segment drei bestätigt, so sendet er nun Segment sechs, obwohl er eigentlich nur zwei ausstehende Segmente haben darf.

Die Übertragung dieser Segmente garantiert, daß auf jeden Fall genügend DUPACKs generiert werden, um ein Fast-Retransmit auszulösen.

3.3.3 Selektive Empfangsbestätigung

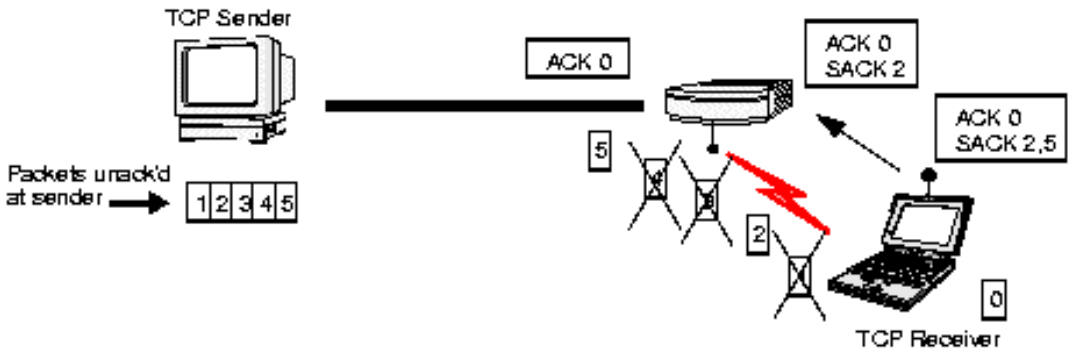

Abbildung 4: Selektive Empfangsbestätigung:

Die Bestätigungen in TCP sind kumulativ, d.h. sie bestätigen, bis zu welchem Segment alle korrekt empfangen wurden. Geht ein einzelnes Paket verloren, so muß der Sender unter Umständen alle Pakete ab dem verlorengegangenen erneut übertragen. Dies ist offensichtlich eine Verschwendung von Bandbreite. Durch SACKs (Selective Acknowledgements -Selektive Empfangsbestätigungen) kann der Empfänger auch einzelne Pakete bestätigen. Der Sender braucht somit nicht mehr alle Pakete erneut zu übertragen, sondern nur noch die fehlenden. Der SACK-Vorschlag sieht vor, daß in jedes ACK eine Bitmaske der bisher gesehenen Segmente trägt.

Eine Alternative zu diesem Vorschlag ist SMART (simple method to aid retransmissions). Hier trägt jede Bestätigung zwei Informationen:

Erstens, bis zu welchem Segment alle korrekt empfangen wurden. Zweitens, die Nummer des Segments, das die Bestätigung ausgelöst hat.

Hiermit kann sich der Empfänger selber eine Liste der vom Empfänger bis dahin erhaltenen Segmente zusammenstellen. Dadurch wird der ACK Header verkleinert ond somit der Overhead reduziert.

3.3.4 Explicit loss Notification

Unter ELN Ansätzen versteht man Mechanismen, die den Sender darüber informieren, daß ein Paketverlust aus einem anderen Grund als einer Stausituation stattgefunden hat. Die Nachricht kann vom Empfänger, einer Basisstation oder anderen Elementen im Netz generiert werden. Der Sender braucht daraufhin sein Sendefenster nicht zu verkleinern, da ja keine Stausituation vorliegt. Die Frage ist, wer diese Nachricht senden soll. Eine Möglichkeit ist der mobile Empfänger, der bei Eingang eines fehlerhaften Segments in der nächsten rausgehenden Bestätigung ein spezielles ELN-Bit setzen kann. Allerdings kann der mobile Knoten nicht immer feststellen, zu welcher Verbindung ein fehlerhaftes Segment gehört, da der TCP-Header selbst fehlerhaft sein kann. Die andere Alternative läßt den Fremdagenten die ELN-Nachricht senden. Dafür muß der Fremdagent den Datenstrom mitlesen und kann, falls er DUPACKs vom mobilen Knoten empfängt, das ELN-Bit in den Bestätigungen setzen und diese dann an den Sender weiterleiten.

4 Ansätze für die Auswirkungen der Mobilität

Eine hohe Bitfehlerrate ist nicht das einzige Problem im mobilen Umfeld. Durch die Mobilität des Knotens kann es immer wieder zu Verbindungsunterbrechungen kommen. Ein typisches Beispiel ist ein Benutzer, der durch einen Tunnel ohne Funkanbindung fährt.

Auch kann es während eines Handover zu Paketverlusten kommen, und zwar wenn sich immer noch Pakete auf dem Weg zum alten Zugangspunkt befinden, obwohl sich der mobile knoten schon bei einem neuen Zugangspunkt registriert hat.

TCP bemerkt den Abbruch der Verbindung zunächst durch das Ausbleiben von Bestätigungen des mobilen Knotens. Dies führt zu einem Timeout und einer Übertragungswiederholung. Der Sender versucht nun immer wieder, das Segment zu übertragen. Nach jedem verunglückten Übertragungsversuch wird der RTO-Timer, der die Übertragungswiederholungen steuert, verdoppelt bis er sein Maximum von einer Minute erreicht hat. Der Sender versucht nun also jede Minute, das Segment erneut zu übertragen. Dies probiert er 12 mal, danach bricht er die TCP Verbindung ab. Selbst wenn zwischenzeitlich die Verbindung zum mobilen Knoten wiederhergestellt werden kann, werden auf Grund des hohen RTO-wertes für eine längere Zeit (nach Herstellung der Verbindung) keine Daten gesendet. Die im folgenden vorgestellten Ansätze haben also drei grundsätzliche Ziele:

- einen Abbruch der TCP-Verbindung bei einer längeren Verbindungsunterbrechung vermeiden
- die vom Sender eingeleiteten Staukontrollmechanismen nach einem Verbindungsabbruch verhindern/mildern
- bei Wiederherstellung der Verbindung erreichen, dass sofort wieder mit dem Senden begonnen wird

4.1 M-TCP

Ein Ansatz, der wieder die Hilfe eines PEP benötigt, ist M-TCP. Ähnlich wie I-TCP, segmentiert auch M-TCP die Verbindung. Zwischen dem Partner im Festnetz und dem Überwachungsknoten (Supervisory Host,SH) wird ein unverändertes TCP eingesetzt, zwischen dem SH und dem mobilen Knoten kommt ein modifiziertes TCP zum Einsatz. Der SH ist verantwortlich für den Datenaustausch zwischen den zwei Partnern, jedoch unternimmt er keine Paketpufferung und führt auch keine Übertragungswiederholungen durch.

Der SH überwacht den Paketfluß vom und zum mobilen Knoten. Empfängt er über einen längeren Zeitraum keine Bestätigung vom mobilen Knoten, so nimmt er an, daß die Funkverbindung unterbrochen ist. Daraufhin setzt er beim Sender durch ein Zero Window Advertisement(ZWA) die Fenstergrösse auf null. Dadurch wird der Sender in den persistenten Modus gezwungen, d.h sein Zustand ändert sich nicht, unabhängig davon, wie lange er vom Empfänger getrennt ist.

Bemerkt der SH wieder eine Anbindung des mobilen Knotens, so setzt er das Sendefenster beim Sender wieder auf den alten Wert. Der Sender kann nun die Übertragung sofort und mit der alten Senderate fortsetzen. Durch diesen Ansatz wird verhindert, dass nach Verbindungsabbrüchen bzw. Handoffs die Staukontrollmechanismen ausgelöst werden.

Wie alle Mechanismen, die einen PEP verwenden, erfordert auch dieser Ansatz, dass der Proxy in der Lage sein muss, den TCP Verkehr mitzulesen.

4.2 Schnelle Übertragungswiederholung

Bei einem Handover kommt es häufig zu Paketverlusten, wenn die Pakete noch an den alten Fremdagenten geroutet werden, obwohl der mobile Knoten schon an eine andere Station angebunden ist. Da der TCP Sender nach diesem Paketverlust auf das Auslaufen des RTO-Timers warten muss,um die Übertragung zu wiederholen, kommt es nach jedem Zellenwechsel zu unnötigen Pausen in der Kommunikation, die um die 0.8 Sekunden betragen. Caceres [CaIf94] schlägt ein Verfahren vor, daß nach einem Handover eine schnelle Übertragungswiederholung forcieren soll: Sobald der mobile Knoten an seinem neuen Zugangspunkt angebunden ist, schickt er drei DUPACKs für das letzte von ihm empfangene Segment. Nach Empfang dieser DUPACKs führt der Sender sofort ein Fast Retransmit durch, wodurch die Pause in der Übertragung auf 50 ms reduziert wird. Dieses Verfahren erfordert nur minimale Änderungen in der Software des mobilen Knotens und bringt eine wesentliche Leistungssteigerung.

4.3 Freeze TCP

Vorübergehende Verbindungsunterbrechungen sind ebenfalls ein häufiges Phänomen im mobilen Umfeld. M-TCP versuchte dieses Problem mit Hilfe eines PEP zu lösen, der den Sender in den persistenten Modus versetzte, sobald der mobile Knoten nicht erreichbar war. Freeze TCP wählt einen ähnlichen Ansatz, jedoch wird hier auf den PEP verzichtet. Bei Freeze TCP beobachtet der mobile Knoten fortlaufend die Stärke des Signals, das er von der Basisstation empfängt. Dadurch soll er drohende Verbindungsabbrüche oder unmittelbar bevorstehende Handover bemerken können. Merkt der mobile Knoten, dass ein Verbindungsabbruch bevorsteht, so schickt er ein ZWA aus. dadurch wird der Sender in den persistenten Modus versetzt. Während der Sender im persistenten Modus ist, sind alle Verbindungsparameter (Sendefenster, Timer) eingefroren. Deshalb wird auch bei einer längeren Unterbrechung der Schicht-2 Verbindung die Transportverbindung nicht abgebrochen. Ausserdem kann der Sender bei Wiederaufnahme der Verbindung mit der alten Senderate fortfahren. Da der mobile Knoten als erster bemerkt, dass er wieder eine Verbindung hat, wird vorgeschlagen, ihn nach Wiederanbindung sofort 3 DUPACKs aussenden zu lassen, damit der Sender sofort seine Übertragung fortsetzt. Die grösste Schwierigkeit bei diesem Ansatz ist natürlich, den Verbindungsabbruch richtig vorherzusagen. Dies erfordert die Kooperation zwischen der MAC-Schicht und TCP.

Freeze TCP hat den großen Vorteil, daß nur geringe Änderungen auf Seiten des mobilen Knotens benötigt werden.

5 TCP in Ad-hoc Netzwerken

Die Zunehmende Verbreitung tragbarer Computer wird dazu führen, dass in Zukunft Ad-hoc Netwerke an Bedeutung gewinnen werden. Ad hoc Netzwerke werden immer dort benötigt, wo eine feste Infrastruktur nicht vorhanden ist oder zerstört wurde. Ein Beispiel wäre eine spontane Arbeitsgruppe von Studenten mit ihren Laptops. Auch in einem ad-hoc Netzwerk sollte ein zuverlässiger Datentransfer möglich sein, des weiteren müssen auch hier Stausituationen vermieden werden. In Festnetzen werden Flusskontrolle und zuverlässige Datenauslieferung von TCP erbracht. Die Frage ist, ob man TCP für die in ad-hoc Netzwerken auftretenden Probleme anpassen kann. Diese sind:

- Häufige Routenwechsel, diese führen zu
  - Paketumordnungen
  - Unterbrechungen, wenn die aktuelle Route zusammenbricht
- Übertragungsfehler und große Verzögerungszeiten, da bei jedem hop auf Zugang zum Medium gewartet werden muß.
- Unfairness durch Interaktion zwischen den TCP und den MAC Backoff-Timern

[t0499] kommen zu dem Ergebnis, daß TCP allein die Probleme in Ad-hoc Netzwerken nicht bewältigen kann. Sie schlagen eine sorgfältige Wahl des MAC-Protokolls und zusätzliche Optimierungen der Routing-Verfahren vor. Zur Verbesserung der Fairness schlagen sie vor, die bestehenden Exponentiellen Backoff-Mechanismen der MAC-Protokolle durch etwas mildere Strategien zu ersetzen.

5.1 Explicit Link Failure Notification (ELFN)

In Ad-hoc Netzwerken kommt es bei Mobilität der Benutzer häufig vor, daß die Route vom Sender zu Empfänger sich ändert oder kurzzeitig zusammenbricht. Um dies zu vermeiden, sind natürlich in erster Linie Verbesserungen der Routing-Algorithmen notwendig. Aber auch im Bereich von TCP sind Veränderungen notwendig, da die Unterbrechungen und Segmentverluste ja bekanntermassen sehr negative Auswirkungen auf das Transportprotokoll haben. Bei diesem Problem kann man jedoch eine Lösung wie die von Freeze TCP nicht anwenden, da die Engeräte ja keine Kenntnisse über den Zustand aller hops im Netz haben und somit eine Verbindungsunterbrechung nicht vorhersagen können.

Das Ziel des ELFN Ansatzes ist es, den Sender über Link und Routenzusammenbrüche zu informieren, damit dieser keine Staumassnahmen ergreift. Bemerkt das Routing-Protokoll das Versagen einer Route, so schickt es eine Nachricht an den Sender mit dem TCP/IP Paketkopf des Paketes, daß die Nachricht ausgelöst hat. Der TCP Sender kann daraufhin die entsprechende Verbindung in einen Stand-by-Modus versetzen, wobei in periodischen Intervallen ein Probepaket gesendet wird um zu testen, ob eine Route gefunden wurde. Falls das Probepaket bestätigt wird, so verläßt der Sender den Stand-by-Modus und setzt die Übertragung mit den vor der Unterbrechung gültigen Werten für Sendefenster und RTO fort.

6 Bewertung und Vergleich der Mechanismen

Wie man sieht, gibt es eine ganze Reihe von Ansätzen, die das Problem "TCP über drahtlosen Link" versuchen zu lösen. Keiner von diesen Mechanismen kann als vollkommen bezeichnet werden, da jeder Vorteil(beispielsweise die Einfachheit der Implementierung) auch wieder Nachteile mit sich bringt (in diesem Fall vielleicht eine schlechtere Leistung).

Ein ideales TCP müßte jederzeit unterscheiden können, ob ein Paketverlust aufgrund einer Stausituation oder wegen eines Übertragungsfehlers aufgetreten ist. Somit könnte es das Segment einfach erneut übertragen, ohne Staumaßnahmen zu ergreifen. Ein ideales darunterliegendes Netzwerk hingegen sollte Fehler und somit auch die Mobilität vor TCP verstecken, Fehler sollten effizient und transparent korrigiert werden.

Die vorgestellten Mechanismen versuchen eine dieser beiden Idealvorstellungen zu erreichen. Deshalb kann man sie grundsätzlich auch in zwei Gruppen aufteilen: Die einen versuchen die Mobilität vor TCP zu verstecken, während die anderen TCP explizit darauf aufmerksam machen.

[BPSK97] verglich die Performance von mehreren Ansätzen und kam zu dem Ergebnis, daß ein Link-Layer Ansatz mit Kenntnis von TCP und zusätzlich implementierten selektiven Empfangsbestätigungen(LL-SMART-TCP-Aware) die beste Leistung erzielt. Dieser Ansatz war in jedem Szenario besser als Ansätze, die eine Aufteilung der TCP Verbindung erfordern (Split Connection), wodurch gezeigt ist, daß es nicht erforderlich ist die TCP-Verbindung aufzuteilen, um gute Performance zu erzielen.

LL-SMART-TCP-Aware hat jedoch den Nachteil, daß die Hilfe eines PEP benötigt wird. Der Nachteil von Lösungen die einen PEP benötigen (Snoop, I-TCP, M-TCP)

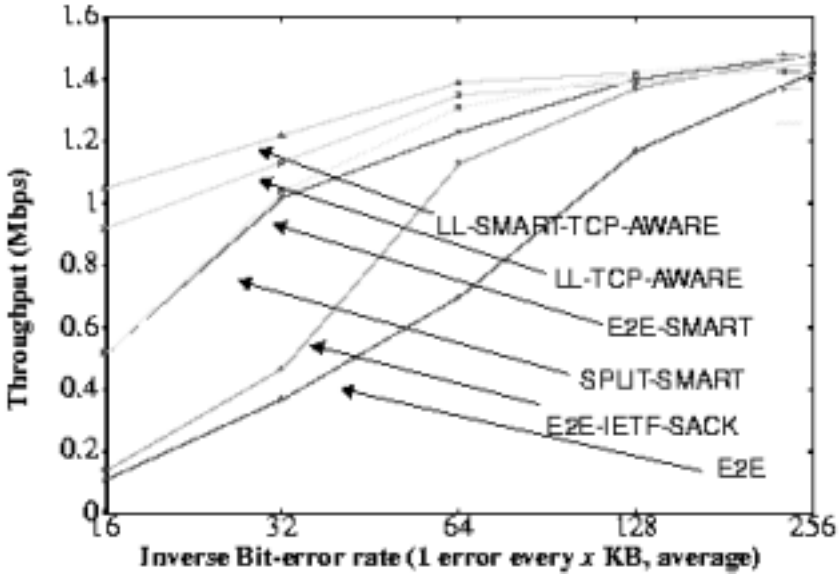

Abbildung 5: Leistungsvergleich verschiedener Mechanismen.(E2E: Standard-TCP)

ist, daß der PEP in der Lage sein muß, den TCP-Verkehr mitzulesen. Wird, wie bei IPv6, die gesamte IP-Nutzlast verschlüsselt übertragen, so muß der PEP in alle Sicherheitsmechanismen mit eingeweiht werden oder das Verfahren ist nutzlos.

Ende-zu-Ende Mechanismen sind in der Bewältigung der vielen Paketverluste nicht so effektiv wie lokale Übertragungswiederholungen. Trotzdem zeigen einige Ansätze, daß erhebliche Leistungssteigerungen auch ohne die Unterstützung von Zwischensystemen möglich sind. Desweiteren ist diese Art von Ansätzen, besonders wenn sie nur Änderungen am mobilen Knoten erfordert, am schnellsten und einfachsten zu Implementieren. Ein Beispiel für einen solchen Mechanismus ist Freeze-TCP, wo allerdings eine Kooperation zwischen der MAC-Schicht und TCP erforderlich ist. Informationsaustausch zwischen den Schichten scheint jedoch unumgänglich zu sein, will man TCP für den drahtlosen Bereich anpassen.

Das Beispiel der Ad-Hoc Netzwerke zeigt, wie gerade die Heterogenität drahtloser Netze die Entwicklung einer Lösung erschwert. Es ist schwierig, einen Ansatz zu finden, der für jede Art von drahtlosem Netzwerk geeignet ist, zu unterschiedlich sind die speziellen Charakteristiken der einzelnen Szenarien.

Ein weiterer wichtiger Aspekt bei der Bewertung, der bid jetzt noch wenig erforscht wurde, ist die Energie-Effizienz. Da mobile Endgeräte meistens batteriebetrieben sind, sollten die Mechanismen diesen Punkt auch berücksichtigen. So sollte die Anzahl der vom mobilen Endgerät doppelt empfangen oder übertragenen Segmente gering gehalten werden.

Zum Abschluß noch einmal eine übersichtliche Darstellung der wichtigsten vorgestellten Mechanismen:

Ansätze für die hohe Verlustrate			
Ansatz	Kategorie	Änderungen beim:	Mechanismus
I-TCP	Split-Connection	FA	Aufteilen der TCP-Verbindung
Link-Layer	Link-Layer	MK, FA	Lokale Übertragungswiederholung
SNOOP	Link-Layer	FA	Lokale Übertragungswiederholung und Filtern von DUPACKs
Delayed DUPACKS	Link-Layer	MK	Herauszögern von DUPACKS
Eifel	Link-Layer	FP	Schnelle Erholung von unnötigen Übertragungswiederholungen
ELN	Ende-zu-Ende	MK	Mobiler Knoten informiert Sender über einen Verlust
SMART/SACK	Ende-zu-Ende	MK,FP	Erneute Übertragung ausschließlich verloren gegangener Segmente
EFR	Ende-zu-Ende	FP	Dynamische Anpassung der Anzahl von DUPACKs für ein Fast Retransmit
Ansätze für die Auswirkungen der Mobilität			
M-TCP	Split-Connection	FA	Fremdagent versetzt Sender in persistenten Modus bei Unterbrechung der Verbindung
Freeze-TCP	Ende-zu-Ende	MK	Mobiler Knoten versetzt Sender in persistenten Modus vor einer Unterbrechung
Fast Retransmit	Ende-zu-Ende	MK	Mobiler Knoten sendet drei DUPACKs nach jedem Handover

Tabelle 1: MK: Mobiler Knoten, FA: Fremdagent, FP: Festnetz-Partner

Literatur

[AlBF00] M. Allman, H. Balakrishnan und S. Floyd. Enhancing TCP´s Loss Recovery Using Limited Transmit. *Internet Draft „draft-ietf-tsvwg-limited-xmit-00.txt“*, August 2000.

[BaBa94] A. Bakre und B.R. Badrinath. I-TCP:Indirect TCP for Mobile Hosts. 1994.

[Bala98] H. Balakrishnan. *Challenges to Reliable Data Transport over Heterogenous Wireless Networks*. Dissertation, University of California at Berkeley, 1998.

[BPSK97] H. Balakrishnan, V.N. Padmanabhan, S. Seshan und R. Katz. A Comparison of Mechanisms for Improving TCP Performance over Wireless Links. *IEEE/ACM Transactions on Networking* 5(6), Dezember 1997.

[BrSi97] K. Brown und S. Singh. M-TCP:TCP for Mobile Cellular Networks. 1997.

[CaIf94] R. Caceres und L. Iftode. Improving the Performance of Reliable Transport Protocols in Mobile Computing Environments. *IEEE Journal on Selected Areas in Communicatio*, August 1994.

[GJPG99] T. Goff, J.Moronski, D.S. Phatak und V. Gupta. Freeze-TCP: A true end-to-end TCP enhancement mechanism for mobile environments. 1999.

[GMKB99] J. Griner, G. Montenegro, M. Kojo und J. Border. Performance Enhancing Proxies. *Internet Draft „draft-ietf-pilc-pep-01.txt“*, Dezember 1999.

[Labs96] Bell Labs (Hrsg.). SMART Retransmission with Performance Overload and Random Losses. White Paper, Bell Laboratories, Mai 1996.

[Schi99] Jochen Schiller. *Mobilkommunikation*. Addison Wesley. 1999.

[t0499] *TCP Performance in Wireless Multi-Hop Networks*, Februar 1999.

[TaTT98] Y. Tamura, Y. Tobe und H. Tokuda. EFR: A Retransmi Scheme for TCP in Wireless LANs. 1998.

[Univ97] Texas A&M University (Hrsg.). Delayed Duplicate Acknowledgements: A proposal to improve performance of tcp in wireless links. White Paper, -, Dezember 1997.

[Univ99] Texas A&M University (Hrsg.). Impact of Routing and Link-Layers on TCP Performance in Mobile Ad Hoc Networks. White Paper, -, Mai 1999.

Abbildungsverzeichnis

Tabellenverzeichnis

Positionsbestimmung in drahtlosen Netzen

Matthias Beck

Kurzfassung

Die Bestimmung der Position eines mobilen Endgerätes in einem drahtlosen Netz ist ein in den letzten fünf Jahren vielseitig diskutiertes Thema. In dieser Seminararbeit stelle ich einige der wichtigsten Techniken zur Positonsbestimmung vor. Dazu gehören TOA, TDOA, AOA, Messung des Phasenwinkels und Auswertung der Signalstärke. Durch die Zielsetzung der E–911 Bestimmungen der FCC gilt insbesondere der Positionsbestimmung in GSM Netzen besondere Aufmerksamkeit. Neben diesem und anderen Zielen beschreibe ich auch kurz vorgeschlagene Lösungsansätze. Auch in Gebäuden gibt es vielfältige Anwendungsmöglichkeiten für Lokalisationssysteme. Da dort GSM und GPS nicht funktionieren, gehe ich auch auf Ansätze zur Positionsbestimmung mittels drahtloser lokaler Netze und Infrarottechnik ein.

1 Einleitung

Die Bestimmung des Aufenthaltsorts von Teilnehmern in mobilen Netzen hat vor allem in den letzten fünf Jahren an Bedeutung gewonnen. Der technologische Fortschritt, die weite Verbreitung von Mobiltelefonen und mobilen Datennetzen und auch gesetzliche Regelungen haben die Entwicklung in diesem Bereich beschleunigt. Ich möchte in dieser Seminararbeit Ziele und technische Möglichkeiten der Positionsbestimmung in drahtlosen Netzen vorstellen. Die Arbeit ist in vier Teile gegliedert. Im ersten Teil stelle ich die grundlegenden Ansätze zur Positionsbestimmung in drahtlosen Netzen vor. Der zweite Abschnitt befasst sich mit GSM Netzen, darauf folgt im dritten Teil die Positionsbestimmung in Gebäuden. Eine kurze abschließende Zusammenfassung bildet den letzten Teil der Arbeit.

2 Grundlegende Methoden zur Positionsbestimmung

Der grundlegende Ansatz zur Bestimmung der Position eines mobilen Endgerätes in einem drahtlosen Netz besteht darin, aus den Eigenschaften der übertragenen Signale Informationen über den möglichen Aufenthaltsort zu gewinnen. Bei nahezu allen Methoden werden mehrere Signale von unterschiedlichen Sendern beziehungsweise ein Signal durch verschiedene Empfänger ausgewertet. Dadurch ergeben sich zwei grundsätzliche Klassen von Positionsbestimmungssystemen [DrNS98]:

- *Self–Positioning*
  Hier wird die Positionsbestimmung vom Empfänger selbst durchgeführt. Dabei ist zu bedenken, dass bei mobilen Endgeräten die zur Verfügung stehende Leistung begrenzt ist. Ein großer Vorteil liegt in der guten Skalierbarkeit dieses Ansatzes, da jedes Gerät für sich selbst genügend Rechnerleistung mitbringen muss, die benötigten Signale jedoch von allen gemeinsam genutzt werden können.

- *Remote–Positioning*
  Dabei wird die Positionsbestimmung von einem entfernten System durchgeführt. Deshalb muss die Leistung des Systems für die Anzahl der möglichen Teilnehmer ausreichen. Nur dieser Ansatz ist möglich, wenn keine Veränderungen an vorhandenen mobilen Endgeräten erfolgen sollen.

Es gibt ein Vielzahl von Methoden, die aus Signalen Informationen über die Position ableiten. Diese können in unterschiedlichen drahtlosen Netzen angewandt werden. Die folgenden gehören zu den wichtigsten [DrNS98]:

- Messung der Signallaufzeit (Time of Arrival, TOA)
- Messung der Signallaufzeitdifferenz zwischen zwei Stationen
  (Time Difference of Arrival, TDOA)
- Bestimmung des Empfangswinkels (Angel of Arrival, AOA)
- Bestimmung der Trägerphase
- Bestimmung des stärksten Senders / Auswertung der Signalstärke [KrBR97]

Durch all diese Methoden werden geometrische Orte für die mögliche Position des mobilen Geräts bestimmt. Durch Messungen bezüglich verschiedener Sendestationen erhält man unterschiedliche Kurven, deren Schnittpunkt dann die gesuchte Position darstellt. In den meisten Netzen kann man davon ausgehen, dass sich alle Sender und Empfänger in einer Ebene befinden und das Problem der Positionsbestimmung damit nur in zwei Dimensionen gelöst werden muss. Die Methoden müssen nicht in *Reinkultur* angewandt werden. Verschiedene Kombinationen von Messungen ergeben neue Systeme und nutzen die Vorteile der verschiedenen Ansätze.

2.1 Time of Arrival (TOA)

Bei dieser Methode wird die Zeit die ein Signal für die Strecke zwischen mobilem Endgerät und Basisstation benötigen gemessen. Dazu müssen jedoch alle Sender und Empfänger genau synchronisierte Uhren besitzen [RaRW96]. Zusätzlich muss jeder Empfänger wissen, wann das Signal abgesendet wurde — was entweder genau festgelegte Sendezeitpunkte oder einen Zeitstempel erfordert.

Alternativ kann auch die Dauer eines Signallaufs vom Sender zum Empfänger und zurück gemessen werden. Bei bekannter Verzögerungszeit für die Erzeugung des Echos lässt sich daraus die gesuchte Signallaufzeit auch ohne Synchronisation ermitteln. Die Bestimmung der Positionen stellt sich bei bekannten Senderstandorten relativ einfach

dar. Es gilt: $R_i = c \cdot t_i$ mit R_i Entfernung zwischen Sender i und Empfänger, c Lichtgeschwindigkeit und t_i gemessene Signallaufzeit. Das mobile Endgerät befindet sich auf einem Kreis mit Radius R_i. Führt man drei Messungen durch, so ist der gesuchte Ort der genau bestimmte Schnittpunkt der drei Kreise (Abbildung 1).

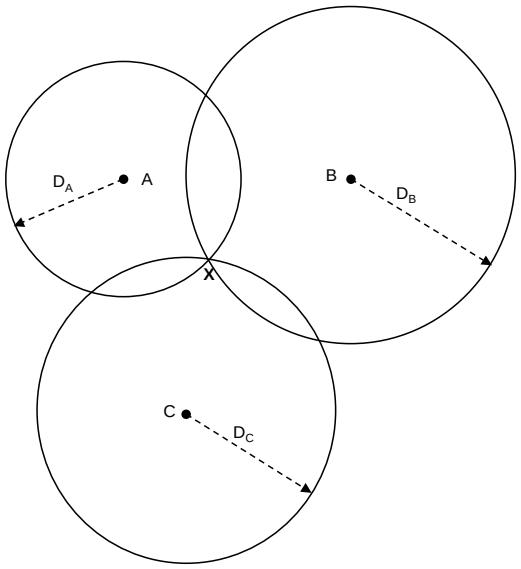

Abbildung 1: Positionsbestimmung durch TOA; A,B und C sind Basisstationen, X die gesuchte Position

2.2 Time Difference of Arrival (TDOA)

Das Problem der aufwendigen Synchronisation umgeht TDOA durch die Bestimmung des Laufzeitunterschieds den Signale verschiedener Sender aufweisen. Als Ergebnis erhält man einen Hyperboloid der den geometrischen Ort der Punkte mit konstanter Differenz der Abstände von den entsprechenden Stationen darstellt[RaRW96]. Der Hyperboloid ist durch folgende Gleichung bestimmt:

$$R_{A,B} = c * t_{A,B} = \sqrt{(X_A - x)^2 + (Y_A - y)^2} - \sqrt{(X_B - x)^2 + (Y_B - y)^2} \qquad (1)$$

Dabei ist $R_{A,B}$ die Differenz der Abstände zwischen mobilem Endgerät und festen Stationen A und B mit den Koordinaten (X_A, Y_A) und (X_B, Y_B). Diese berechnet sich aus TDOA $t_{A,B}$ und der Ausbreitungsgeschwindigkeit c. (x, y) ist der gesuchte Standort. Die gesuchte Position ergibt sich als gemeinsame Lösung mindestens zweier Gleichungen vom Typ (1) (siehe Abbildung 2).

Die TDOA Methode kann — wie auch TOA — sowohl als Self–Positioning System als auch als Remote–Positioning System realisiert werden. Beim ersteren „hört“ das mobile Endgerät auf mehrere Basisstationen, beim zweiteren empfangen mehrere Basisstationen ein Signal. Unabhängig davon benötigt man eine gewisse Synchronität der Uhren in den Basisstationen[DrNS98], da die Stationen entweder als Sender gleichzeitig oder mit bekanntem Zeitunterschied senden müssen und als Empfänger die ermittelten Ankunftszeiten vergleichbar sein müssen.

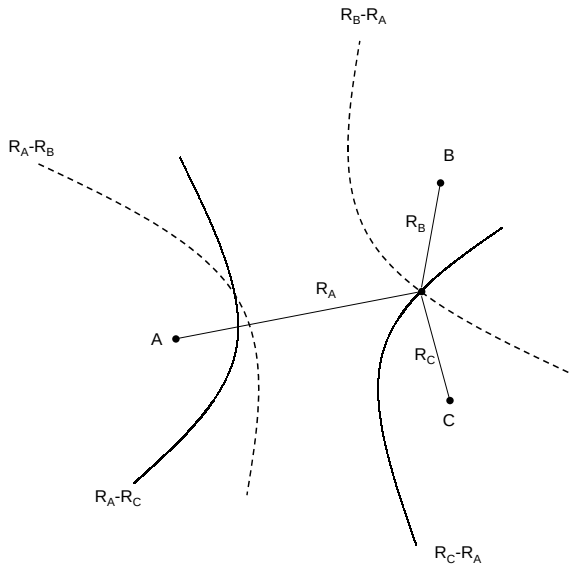

Abbildung 2: Bestimmung der Position als Schnittpunkt zweier Hyperboloide

2.2.1 Implementierungsmöglichkeiten

Bei TDOA Systemen gliedert sich die zu bewältigende Aufgabe in zwei Schritte. Erstens müssen aus verrauschten Signalen TDOA Werte geschätzt werden. Zweitens gilt es die daraus resultierenden quadratischen Gleichungssysteme zu lösen.

Der klassische Ansatz zur Bestimmung der TDOA Werte $R_{i,j}$ ist die Kreuz–Korrelations–Methode. Sei $s(t)$ das übertragene Signal, $x_i(t) = s(t - d_i) + n_i(t)$ das Signal, das bei Empfänger i um d_i Sekunden verzögert und durch $n_i(t)$ verrauscht ankommt. Ebenso beschreibt $x_j(t) = s(t - d_j) + n_j(t)$ das bei Empfänger j um d_j verzögert und durch $n_j(t)$ verrauscht ankommende Signal. Die Kreuz–Korrelation $\hat{R}_{x_i,x_j}(\tau)$ zwischen den beiden Signalen ergibt sich durch die Integration über das Produkt der beiden Signale, wobei eines um den Faktor τ verzögert ist.

$$\hat{R}_{x_i,x_j}(\tau) = \frac{1}{T} \int_0^T x_i(t) x_j(t - \tau) dt \tag{2}$$

Der Wert des Parameters τ, für den die Kreuz–Korrelationsfunktion ihren maximalen Wert annimmt, entspricht dann dem *maximum likelihood* Schätzer für den TDOA Wert. Das Auflösungsvermögen dieser Methode ist bei Mehrwegeausbreitung jedoch auf $\frac{1}{B}$ (B: Bandbreite des empfangenen Signals) begrenzt [KrBR97]. Die durch die Mehrwegeausbreitung verursachte Verbreiterung des Maximums der Kreuz–Korrelationsfunktion versuchen neuere Algorithmen in den Griff zu bekommen. Dabei konnten beispielsweise für BPSK und QPSK gute Ergebnisse trotz starker Interferenzen erzielt werden [ChGa92].

Die Lösung der Gleichungen des Typs (1) ist der zweite Schritt der Postionsbestimmung [RaRW96]. Da die Gleichungen nicht linear sind, ist dies nicht trivial. Es gibt verschiedene Ansätze, deren Lösungen unterschiedliche Eigenschaften aufweisen. Eine einfachere stellt die Linearisierung der Gleichungen durch eine Taylor Entwicklung dar. Diese Technik kann jedoch unter gewissen Voraussetzungen zu großen Fehlern in der

errechneten Position trotz relativ genauer Schätzung der TDOA Werte führen. Dies ist z.B. der Fall, wenn das mobile Endgerät von beiden Basisstationen weit entfernt steht.

Ist das Gleichungssystem nicht überdeterminiert, so kann man auch eine exakte Lösung berechnen [Fang90]. Das bedeutet für den zweidimensionalen Fall, dass zwei TDOA Messungen und somit Funkkontakt mit drei Basisstationen ausreichen um eine Position zu bestimmen. Sind jedoch zusätzliche Messungen vorhanden, so kann diese Methode die zusätzliche Information nicht nutzen [RaRW96].

2.3 Angel of Arrival (AOA)

Die Bestimmung des Winkels aus dem ein Signal empfangen wird stellt eine weitere wichtige Methode dar. Sie findet zum Beispiel beim Radar Anwendung. Kennt man den Einfallswinkel des Signals, so ergibt sich daraus eine Gerade auf der sich das betreffende mobile Endgerät befinden muss. Kennt man zwei oder mehr solcher Geraden von unterschiedlichen Basisstationen, so ist der Schnittpunkt der gesuchte Standort [DrNS98].

Für AOA sind theoretisch also Messungen bzgl. zweier Stationen ausreichend. Generell finden bei AOA Antennen mit ausgeprägter Richtwirkung Anwendung, was die Realisation auf mobilen Endgeräten schwierig macht [RaRW96]. Bei den Antennen handelt es sich meistens um eine Antennengruppe mit Abständen zwischen den Antennen im Bereich der halben Wellenlänge der Trägerfrequenz. Dies ermöglicht den Einfallswinkel durch die Phasensprünge, die an den einzelnen Antennen zu beobachten sind, zu modellieren. Die Genauigkeit dieser Modellierung hängt stark von der Qualität der Empfängerhardware ab. Die an den einzelnen Antennen ankommenden Signale müssen in möglichst identischer Art und Weise verarbeitet werden, was mit steigender Antennanzahl schwieriger wird. Darin liegt mit der größte Kostenanteil bei AOA Systemen, weshalb die Anzahl der Antennen so gering wie möglich gehalten werden sollte. Weiterer Aufwand entsteht durch die Notwendigkeit, die Gruppenantenne in regelmäßigen Abständen zu kalibrieren[KrBR97].

2.3.1 Implementierungsmöglichkeiten

Neben der Messung der Phasensprünge an den einzelnen Antennen gibt es weitere Algorithmen zur Bestimmung des Einfallwinkels. Ein einfacher Ansatz besteht darin, den Hauptstrahl der Richtungsantenne über das zu beobachtende Gebiet zu steuren und dabei die empfangene Leistung zu bestimmen[KrBR97]. Dadurch erhält man eine Verteilung der Empfangsleistung über das Gebiet der Einfallwinkel. Problematisch ist bei dieser Methode, dass die Auflösung auf die Breite des Strahls der Richtungsantenne begrenzt ist. In Netzen mit Mehrwegeausbreitung funktionieren jedoch beide Ansätze nur schlecht. In solch einer Umgebung können Algorithmen, die auf maximum likelihood (ML) Schätzern basieren, wesentlich bessere Ergebnisse erzielen [RaRW96], da diese gegenüber dabei auftretenden kohärenten Signalen robuster sind [ZiWa88].

2.4 Bestimmung der Trägerphase

Die Bestimmung der Trägerphase ermöglicht einen sehr exakte Positionsbestimmung. Diese Methode findet beispielsweise beim GPS Anwendung. Auch hier misst der GPS Empfänger die Laufzeit mehrerer eingehender Signale. Die Trägerfrequenz ist für jeden Satelliten gleich. Dies wird durch Modulation mit orthogonalen Codes im Spread Spectrum Verfahren ermöglicht. Durch Erzeugung des gleichen Codes im Empfänger und Autokorrelation kann dann die genaue Signallaufzeit gemessen werden. Da der Empfänger hierzu immer auf die Trägerfrequenz eingestellt bleiben muss und eine Regeneration des Trägers bei GMSK Signalen schwierig ist, ist diese Methode in vielen Mobiltelefonnetzen nicht praktikabel [DrNS98]. Eine gute Einführung in die Funktionsweise von GPS findet sich in [EnMi99] und [Gett93].

2.5 Bestimmung des stärksten Senders / Auswertung der Signalstärke

Eine einfache Methode zu Positionsbestimmung ist die Bestimmung des stärksten Senders. Dabei wird die Position des Mobilgeräts mit der Position der Sendestation mit dem stärksten Signal gleich gesetzt. Eine Verbesserung des Ergebnisses lässt sich dadurch erzielen, dass man eine Vielzahl von Sendern, die nur ihre eigene Position übermitteln, verwendet [RaRW96]. Die Auflösung ist dann durch die Feinheit des Sendernetzes bestimmt. Eine Verfeinerung kann durch die Bestimmung mehrerer stärkster Sender erfolgen.

Ähnlich geht man bei der Auswertung der Signalstärke vor. Kennt man die Dämpfung auf dem Übertragungsweg, so kann aus der Signalstärke die Entfernung zum Sender ermittelt werden. Dies resultiert praktisch in einem TOA Ansatz.

Beide Ansätze finden primär in Gebäuden und auf lokaler Ebene ihre Anwendung, da sich dort genügend Sender verteilen lassen und mittels Wellenausbreitungsmodellen die Dämpfung in Gebäuden relativ gut bestimmbar ist.

2.6 Probleme der TDOA und AOA Methoden im praktischen Einsatz

Neben den erwähnten Schwierigkeiten bei der Schätzung der TDOA und AOA Werte unter Mehrwegeausbreitung und Interferenz gibt es andere Probleme, die die Anwendung in der Praxis erschweren. Eine Annahme, die für die beiden Methoden gemacht wird, ist die Existenz einer Sichtverbindung zwischen Sender und Empfäger [KrBR97]. Befindet sich in dem durch Mehrwegeausbreitung erzeugten Bündel von Signalen keines, das ohne Reflexion und damit ohne Änderung der Richtung beziehungsweise Weglänge zum Empfänger gelangt, so sind realistische Zeit– und Winkelschätzungen nur schwer möglich.

Eine weitere Schwierigkeit in der Praxis ist, dass beide Techniken mindestens zwei oder drei Basisstationen benötigen um eine Positionsbestimmung durchzuführen. In Realität sind aber zum Beispiel die mobilen Telefonnetze gerade darauf ausgelegt, dass eine gute Funkverbindung mit niedrigem Signal–Rausch–Verhältnis nur zwischen Mobiltelefon

und einer Basisstation besteht. Diese Art von Raummultiplex ermöglicht eine bessere Ausnutzung der Kanalressourcen, da der vorhandene Frequenzbereich mehrmals an verschiedenen Orten genutzt werden kann [Schi00]. In Städten spielen bei der Bestimmung der Zellen primär Kapazitätsgesichtspunkte eine Rolle [RKWR98]. Dort sind deswegen genügend *hörbare* Basisstationen zu erwarten. Dies ist möglich, da benachbarte Zellen unterschiedliche Frequenzbereiche nutzen und die gegenseitige Störung zweier Zellen über eine andere Zelle hinweg aufgrund der starken Dämpfung sehr unwahrscheinlich ist. In ländlichen Gebieten kann jedoch nicht generell mit der Erreichbarkeit mehrerer Stationen gerechnet werden. In einer Untersuchung für das Advanced Mobile Phone System (AMPS) wurden entlang des Interstate–95 highways in den USA Messungen durchgeführt [RKWR98]. Dabei stellte sich heraus, dass die Wahrscheinlichkeit in ländlichen Gegenden drei Basisstationen ausreichend stark empfangen zu können bei 35% lag, während in Stadtgebieten ein Wert von 84% ermittelt wurde. Wie zu erwarten, war in ländlichen Gebieten oftmals eine Station sehr gut zu empfangen, weitere dann nur noch in seltenen Fällen.

Diese in der realen Welt auftretenden Probleme werden teilweise durch die erwähnten Implementationen gelöst. Robuster zeigen sich die Algorithmen, die — soweit vorhanden — mehr Messungen als für die Lösung des Positonsproblems nötig sind, berücksichtigen. Schwierig ist meiner Meinung nach eine Lösung für den Fall ohne Sichtverbindung zu finden, da man bei Empfang des Signals dessen konkreten Ausbreitungspfad rekonstruieren müsste.

3 GSM Netze

Mobilfunksysteme die auf dem *Global System for Mobile communications (GSM)* Standard basieren stellen mit einem Marktanteil von über 50% (Stand Oktober 1999) das weltweit erfolgreichste Mobilfunksystem dar [Schi00]. Damit ist GSM auch für Systeme zur Positionsbestimmung eines der wichtigsten Anwendungsgebiete.

3.1 Ziele und Anwendungen der Positionsbestimmung in GSM Netzen

3.1.1 Die E-911 Regelung der FCC

Einen entscheidenden Schub erhielt das Thema Positionsbestimmung in mobilen Telefonnetzen durch eine Entscheidung der US amerikanischen Federal Communications Commission (FCC) im Jahre 1996 [RKWR98]. Für die FCC steht in diesem Zusammenhang die Sicherheit im Vordergrund [RaRW96]. So wurde in dieser Zeit in der Öffentlichkeit bekannt, daß das vorhandene Notrufsystem für mobile telefonierende Anrufer inadequat sei. Etwa jeder fünfte Notruf wurde damals von einem Mobiltelefon abgesetzt und etwa 25 Prozent dieser Anrufer kannten ihren aktuellen Standort nicht. Nachdem 1997 erste Maßnahmen wie die Möglichkeit des Notrufs ohne Benutzervalidierung eingeführt wurden, trat Anfang 1998 die erste Phase des drahtlosen Enhanced 911 (E–911) Service in Kraft. Diese beinhaltete die Möglichkeit des Rückrufs mobiler Anrufer sowie die Identifikation der Funkzelle des Anrufers. Für die zweite Phase von E-911

— diese wird am 1. Oktober 2001 für Netzbetreiber in den USA Vorschrift — werden wesentlich umfangreichere Maßnahmen nötig. So verlangt dieser Service die Übermittlung der Anrufernummer, Ortsangabe der bentutzten Basisstation und eine Schätzung der Position des Anrufers in Länge und Breite. Dabei ist eine Genauigkeit von 125 m in 67% aller Fälle gefordert.

Diese Anforderungen stellen eine große technische und finanzielle Herausforderung für die Netzbetreiber dar. Andererseits sind durch die Einführung von E-911 auch andere Dienste die auf der Lokalisierung der Kunden beruhen möglich. Somit stellt die geforderte Positionsbestimmung auch eine Chance für die Betreiber dar, sich am Markt zu unterscheiden [RKWR98].

3.1.2 Mögliche Anwendungen der Positionsbestimmung

Eine Marktstudie [Grou99] ergab, daß der Markt für positionsbasierte Dienste im Jahr 2004 ein Volumen von bis zu $4 Miliarden erreichen kann. So zeigten beispielsweise 54% der Befragten Interesse an einem durch Positionsbestimmung unterstützten Pannendienst. Neben den sicherheitsrelevanten Anwendungen sind auch Dienste in den Bereichen lokalisierte Informationsdienste, ortsabhängige Abrechnung — vergleichbar mit dem Genion Dienst von ViagInterkom — und Verfolgung (Tracking) von Paketen, Fahrzeugen und Menschen vorstellbar [McCa99]. Bei der Verfolgung von Menschen mag einem zurecht der Gedanke an eine schöne neue Welt durch den Kopf gehen. Pilotanwendungen mit Straffälligen und Alzheimerkranken zeigen jedoch auch das Nutzenpotential dieses Ansatzes [ZPBM98].

Auch für die Netzbetreiber ergeben sich Nutzungsmöglichkeiten, die über die Erfüllung des E-911 Standards hinausgehen [DrNS98]. Mikroskopisch betrachtet ermöglicht eine exakte Positionsangabe eine bessere Entscheidung beim handover zwischen verschiedenen Funkzellen. Die langfristige Beobachtung von Positionen im Netz ermöglicht außerdem eine wesentlich genauere und bessere Ressourcenplanung für das Netzwerk.

Eine erhebliche Erweiterung der denkbaren Applikation wird mit der weiteren Verbreitung mobilen Internets einhergehen. Insbesondere die aufgrund kleiner Displays und Tastaturen unkomfortable Benutzung mobiler Endgeräte könnte durch Kontextsensitivität stark verbessert werden. Mobiles Internet ist als Thema zu umfangreich um in dieser Arbeit behandelt zu werden. Eine Einführung in einige der verwendeten Technologien findet sich in [Schi00].

3.2 Bestimmung der Position in GSM Netzen

Um die Anforderungen der E-911 Regelung zu erfüllen gibt es verschiedene technische Realisationsmöglichkeiten. So bestünde theoretisch auch die Möglichkeit Mobiltelefone mit der Fähigkeit des Global Positioning Systems (GPS) auszustatten. GPS bietet den Vorteil extrem hoher Genauigkeit und Skalierbarkeit, da jeder Empfänger seine Position autark mit Hilfe von Satellitensignalen bestimmt. Solche Empfänger kosten mittlerweile unter $100 [EnMi99]. Um diesen Preis weiter zu senken, wird ein Ansatz untersucht, bei dem der Empfänger die Signale der Satelliten für eine kurze Zeit aufzeichnet und per Datenübertragung an eine zentrale Stelle zur Auswertung geschickt wird [RaRW96]. Damit spart man sich aufwendige Signalverarbeitungshardware zur Bestimmung der

Signallaufzeiten und Rechenleistung zur Bestimmung der Position. Da aber die von der FCC geforderte Lösung auch mit allen sich bereits im Umlauf befindenden Telefonen funktionieren muss, kann GPS nicht ausschließlich zum Einsatz kommen [DrNS98].

Meiner Meinung nach wäre GPS eine interessante Möglichkeit für genauere Systeme. Problematisch erscheint mir der recht knappe Zeitrahmen. Da Mobiltelefone nur eine Lebenszeit von wenigen Jahren haben, wäre ein kompletter Austausch der Geräte innerhalb von fünf Jahren wohl möglich. Danach könnte man die Präzision und Ausgereiftheit von GPS nutzen. So investieren die teilweise hoch verschuldeten Netzanbieter jetzt in relativ aufwendige Hardware und werden deshalb auch nach einigen Modellzyklen nicht wieder auf ein neues System umsteigen. Damit kommen wohl andere, in GSM Netzen direkt anwendbare Technologien zum Einsatz.

Einige der in Abschnitt 2 vorgestellten Methoden finden eine interessante Anwendung in der Positionsbestimmung für GSM Netze. Diese bieten in ihrer ursprünglichen Spezifikation bereits Möglichkeiten zur groben Berechnung der benötigten Werte.

3.2.1 Der GSM Standard — für die Positionsbestimmung wichtige Details

Eine genaue Einführung in das GSM Netz würde den Rahmen dieser Arbeit natürlich sprengen. Dennoch möchte ich kurz auf einige technische Details eingehen, die für konkrete Versuche benutzt wurden. Eine umfassendere Beschreibung findet sich in [Schi00], ich orientiere mich hier an [DrNS98].

GSM nutzt für die Abwärts– und Aufwärtsrichtung jeweils einen eigenen 200Khz Kanal (Richtungsduplex). Auf jeden dieser Kanäle wird ein starres Zeitmultiplexverfahren angewandt, das wiederum jeden Kanal in acht Blöcke unterteilt. Weiterhin benutzt GSM verschiedene logische Kanäle. Über Verkehrskanäle (Traffic Channel, TCH) werden alle Nutzdaten übertragen, über mehrere Steuerkanäle (Control Channel, CCH) verschiedene Steuerdaten, unter anderem zur Synchronisation. Die Daten für die einzelnen Kanäle werden in einer relativ komplizierten Rahmen–Hierarchie übertragen. Dabei kommen verschiedene *Burst* Formate zur Übertragung in den einzelnen Zeitschlitzen zum Einsatz, so z.B. der *normal Burst*. Für die Positionsbestimmung sind dabei insbesondere die darin enthaltene feste 26–bit Trainingssequenz interessant. Sie befindet sich in der Mitte des Bursts und dient zur Anpassung des Empfängers an die Kanaleigenschaften. Dazu wird auch der oben beschriebene Korrelationsansatz verwendet und bietet damit also eine Möglichkeit zur Laufzeitbestimmung. Eine weitere so nutzbare Übertragung ist der 64–bit *synchronization burst*. Problematisch ist dabei die Breite der Spitze der Korrelationsfunktion. Sie ist für alle Sequenzen 4 bit und reduziert damit das zeitliche Auflösungsvermögen erheblich. Ein weiters Problem ergibt sich durch die Tatsache, dass ein GSM Empfänger sich an die vorhandene Mehrwegeausbreitung durch einen Entzerrer anpasst und die ankommenden Signale kombiniert auswertet. Wie in Abschnitt 2.6 erwähnt, benötigen Positionsbestimmungs–Systeme jedoch das Signal der direkten Sichtverbindung. Weitere Informationen können aus der Einteilung des Gebiets in Funkzellen gewonnen werden. Die Verwendung von sektorisierten Antennen in jeder Zelle ermöglicht eine zusätzliche Einschränkung der möglichen Position.

Wegen der technischen Eigenschaften sind im GSM–Netz direkte Zeitmessungen für einen korrekten Betrieb nötig. So müssen die mobilen Geräte ihren Sendezeitpunkt um die Signallaufzeit anpassen, damit der Zeitmultiplexrahmen nicht verletzt wird. Die

benötigte Versatzzeit (Timing Advance, TA) wird dem mobilen Endgerät durch die Basisstation mitgeteilt. Ein Vergleich der Versatzzeiten verschiedener hörbarer Basisstationen liefert die *observed time difference (OTD)*, welche zur Steuerung der Übergabe zwischen zwei Funkzellen genutzt werden kann.

3.2.2 Positionsbestimmung innerhalb des GSM Standards

Nutzt man nur die durch den Standard gegebenen Möglichkeiten, so ergeben sich im wesentlichen zwei Methoden [RKWR98], [DrNS98]. Die TA ist eine Maß für die Signallaufzeit und wird dem Mobiltelefon auch bei einem Wechsel der zuständigen Basisstation mitgeteilt. Erzwingt man künstlich zwei Wechsel der Basisstation — deren Hörbarkeit vorausgesetzt — so hat man insgesamt drei Messungen zur Verfügung, was eine Positionsbestimmung ermöglicht. Da die TA jedoch in Bit–Perioden gemessen wird ist die Genauigkeit der Messung auf 554m begrenzt. Außerdem können sich Wechsel zu ungünstigen Basisstationen negativ auf die Sprachqualität auswirken.

OTD Bestimmungen werden ohne Wechsel der Basisstation durchgeführt und könne als Schätzung der TDOA verwendet werden. Jedoch ist auch hier die Auflösung auf 554m begrenzt. Zusätzlich müssten die Basisstationen, um zumindest diese Genauigkeit zu erreichen, genau synchronisiert sein. Die so genannte Pseudosynchronisation ist wiederum nur in Bitgenauigkeit sichergestellt wodurch die Qualität der Berechnungen zusätzlich leidet. Ein in [SpMa98] und [Spir99] gezeigter Ansatz um die erzielte Genauigkeit zu erhöhen verwendet den aus der Satellitentechnik kommenden Kalman Filter. Dabei wird die Position nicht aus den durch OTD bestimmten TDOA Werten berechnet sondern mittels Kalman Filter geschätzt, wodurch eine Verbesserung der Genauigkeit erzielt wird.

3.2.3 Mögliche System–Architekturen und Erweiterungen

Um eine die E–911 Vorschriften erfüllende Positionsbestimmung durchführen zu können sind technische Erweiterungen der Netze nötig, die eine genauere Messung der Signallaufzeiten und die Synchronisation der Basisstationen ermöglichen [DrNS98]. Eine weitere grundsätzliche Frage ist, an welcher Stelle im System die Positionsbestimmung erfolgen soll. Grundsätzlich kann dies im mobilen Gerät oder an einem zentralen Ort im Netz erfolgen. Erstere Lösung hat den Vorteil guter Skalierbarkeit, zweitere kann auch ohne Modifikationen an existierenden Mobiltelefonen funktionieren. Solch eine Modifikation ist zum Beispiel nötig, wenn Trainingssequenzen für eine hochauflösende Bestimmnung der Laufzeit verwendet werden sollen.

Ein weiterer zu beachtender Punkt ist die anfallende Datenmenge. So ist einerseits bei einem Notruf nicht nur die Position zu bestimmen, sondern auch an eine lokale Rettungstation zu melden. Ein weitaus größeres Datenaufkommen ist zu bewältigen, falls von den Netzwerkbetreibern zusätzliche positonsbasierte Dienste angeboten werden. Müssen dazu viele Positionsdaten regelmäßig gespeichert, aktualisiert und abgefragt werden, so sind aufgrund der verteilten Architektur des GSM Netzes auch geeignete Datenbanken nötig. Die in [MaDo00] vorgestellte hierarchische Struktur ist in der Lage, trotz der nicht hierarchischen Nummernstruktur schnelle Zugriffe zu leisten.

4 Positionbestimmung in Gebäuden

Neben der vor allem durch die E–911 Bestimmungen verursachten Aktivität im Bereich der Positionsbestimmung im Freien wurden in den letzten Jahren einige Anwendungen für solche System in Gebäuden entwickelt. Die im Freien realisierten Systeme können jedoch nicht einfach in Gebäuden benutzt werden [BaPa00]. Auf zellularen Netzen basierende Systeme sind insbesondere aufgrund der starken Störungen in Gebäuden zu ungenau. GPS kann die nötige Genauigkeit liefern [Gett93], benötigt jedoch Sichtverbindung mit 3–4 Satelliten. Lösungsansätze wurden für mobile Endgeräte in drahtlosen lokalen Netzen (Wireless Local Area Network, WLAN) entwickelt (Abschnitt 4.2). Ein primär rein zur Positionsbestimmung entwickeltes und auf Infrarottechnik basierendes System ist ActiveBadge (Abschnitt 4.3). Zuerst möchte ich jedoch kurz auf die zukünftigen und teilweise bereits erprobten Anwendungen von Ortsinformationen in Gebäuden eingehen.

4.1 Mögliche Anwendungen positionsbestimmender Systeme in Gebäuden

Eine Anwendungsgebiet liegt im militärischen Bereich [KrPB98]. Ein Programm des DARPA hat ein System zum Ziel, in dem kleine Kampfgruppen mit Kommunikationshardware ausgestattet werden [Agen]. Diese soll auch in der Lage sein, in Gebäuden Positionsbestimmung durchzuführen. Zusätzlich sollen auch Funktionen zur taktischen Unterstützung integriert werden. Polizei und Feuerwehr sind an ähnlichen Anwendungen interessiert [LPLaY00b].

Auch kommerzielle Anwendungsgebiete sind in der Entwicklung. Ein Beispiel sind Krankenhäuser, in denen Ausrüstung, Patienten und Krankenakten gefunden werden müssen [LPLaY00a]. In [WeLa98] wird eine Lösung für dieses Problem vorgeschlagen, die auf der Verwendung von *radio frequency identification Tags* — kleine einfache Sender — basiert.

Ein großes Anwendungsgebiet für positionsbestimmende Systeme ist das sogenannte aktive Büro [HaHo92]. Natürlich ist auch hier das Wiederfinden von Akten ohne zu suchen eine reizvolle Nutzung. Wegen der üblichen Verwendung von Rechnern ergeben sich weitreichende Applikationen. Umfangreiche Versuche wurden dabei mit der *Active Badge* Technologie bei Olivetti Research Limited (ORL) gemacht. Die Technik stelle ich kurz in Abschnitt 4.3 vor. An dieser Stelle möchte ich einige der wichtigsten Anwendungen, die sich das verfügbare Wissen über Identität beziehungsweise Aufenthaltsort des Nutzers zu Nutze machen, nennen [HaHo92]:

- Auffinden von Menschen und Gegenständen:
  Ein grafisches System zeigt in der obersten Ebene die Position von Menschen und Gegenständen. Die Auswahl einer Person liefert dann weitere Informationen, wie eine Liste der Personen und Gegenstände die sich im gleichen Raum befinden und ein Kommunikationsfenster.

- Ortsabhängige Kommunikation:
  Das Kommunitkationsfenster enthält eine Übersicht über das Kommunitkationsprofil des Angerufenen. Die Regeln zur Erstellung berücksichtigen Identität, Position und Kommunikationsmöglichkeiten des Anrufers und des Angerufenen.

- Teleporting:
  Die X Window Desktops folgen den Nutzern. Durch Drücken des rechten Knopfes auf dem Active Badge wird der Wunsch, seinen eigenen Desktop auf einer sich in der näheren Umgebung befindenden Workstation angezeigt. Das System liefert alle kollokierten Rechner und ermöglicht so eine regelbasierte Auswahl.

- Kontextbewußte Laptops:
  Die mit Infrarot Netzwerkkarten ausgestatteten Notebooks kommunizieren direkt mit den Badge–Sensoren. So können beispielsweise kollokierte Drucker identifiziert und genutzt werden.

Eine weitere Notwendigkeit für Positionsbestimmung in verschiedenen Umgebungen ergibt sich durch die Integration heterogener Fest– und Mobilnetze. Um den richtigen Zeitpunkt für einen vertikalen Handover zwischen verschiedenen Netzarchitekturen zu bestimmen, können Lokationsinformationen genutzt werden [YlMP00].

4.2 Positionsbestimmung in lokalen drahtlosen Netzen

4.2.1 Möglichkeiten der TDOA Bestimmung

Auch in drahtlosen lokalen Netzen können zur Positionsbestimmung TDOA Messungen verwendet werden. Dabei sollte wie bei den Telefonnetzen versucht werden, die nötigen Informationen aus existierenden Systemen zu gewinnen [LPLaY00b], da dies eine Integration in vorhandene Installationen ohne größere Veränderungen ermöglicht. Mit der Verabschiedung der Standards *IEEE 802.11* und *ETSI HIPERLAN/2* wurden zwei leistungsfähige Systeme geschaffen. „Der IEEE–Standard 802.11 beschreibt die derzeit am weitesten verbreitete Familie drahtloser LANs, zu der auch bereits vielfältige Produkte verfügbar sind“ [Schi00].

Zur Bestimmung der Position eines mobilen Endgerätes benötigt man wiederum drei Basisstationen. Dabei kann es sich um bereits vorhandene Zugangspunkte (Access Point, AP) oder speziell installierte Stationen handeln (Geolocation Base Station, GBS). IEEE 802.11 verwendet zur Übertragung das Direct Spread Spectrum (DSSS) Verfahren. Damit sind grundsätzlich die für DSSS Systeme bekannten Methoden zur Messung der TDOA beispielsweise mit Hilfe der Korrelationsfunktion anwendbar. Dazu müssen jedoch Sender und Empfänger synchronisiert sein [LPLaY00a]. Eine intuitiver Ansatz wäre auch, eine Nachricht mit Zeitstempel zu versenden und den Zeitpunkt des Empfangs an drei verschiedenen Stationen zu bestimmen. Aus diesen Signallaufzeiten könnten dann an einer zentralen Stelle TDOA Werte berechnet werden — eine Synchronisation des mobilen Gerätes wäre nicht nötig. Es ist jedoch unwahrscheinlich, dass in der Praxis eine Synchronisation mehrerer unabhängig voneinander operierender Basisstationen leicht zu bewerkstelligen wäre. Diese Schwierigkeit lässt sich durch die Messung der doppelten Signallaufzeit umgehen. Beide Ansätze funktionieren jedoch nicht ohne Veränderungen der vorhandenen Infrastruktur und Signalisierungsabläufe.

Die IEEE 802.11 Medium Access Control (MAC) Schicht bietet Möglichkeiten dies zu umgehen. Dabei verwendet man die im Medienzugriffsverfahren auftretenden Zeitabstände, wie SIFS (Short inter-frame spacing), PIFS (PCF inter-frame spacing) und DIFS (DCF inter-frame spacing) [LPLaY00a]. Vor dem Senden von Daten muss das

Medium für die Dauer DIFS nicht belegt sein. Nach dem Empfang von Daten durch ein Gerät wartet dieses für die Dauer SIFS — es ist damit priorisiert, da $SIFS < DIFS$ — und schickt eine Bestätigung. Die doppelte Signallaufzeit ließe sich dann leicht berechnen. Die Genauigkeit der Rahmenabstände ist jedoch nur auf $2\mu s$ festgelegt, was einem Fehler von bis zu 600m entspricht. Solch ein System macht in normalen Gebäuden keinen Sinn. Eine wesentlich genauere Methode nutzt die Fähigkeiten der MAC Schicht und benötigt in den GBS präzise Timer. Es gilt folgender Zusammenhang [LPLaY00a]:

$$\begin{aligned} TDOA_{21} &= \tau_2 - \tau_1 \\ &= [(\tau_{20} + \tau_{21}) - \tau_{00}] - [(\tau_{10} + \tau_{11}) - \tau_{00}] \\ &= (\tau_{20} + \tau_{21}) - (\tau_{10} + \tau_{11}) \end{aligned} \tag{3}$$

Die GBS misst τ_{11}, τ_{21} als Zeitspannen zwischen einem Daten– und einem Bestätigungspaket das AP und Mobilgerät austauschen. Die Zeiten τ_{10} und τ_{20} sind die Laufzeiten der Signale von AP zu den GBS und können aus deren bekannter Entfernung recht genau abgeleitet werden. τ_1 und τ_2 sind die Signallaufzeiten vom mobilen Endgerät zu den beiden GBS. Dies Zusammenhänge werden nochmals in Abbildung 3 deutlich.

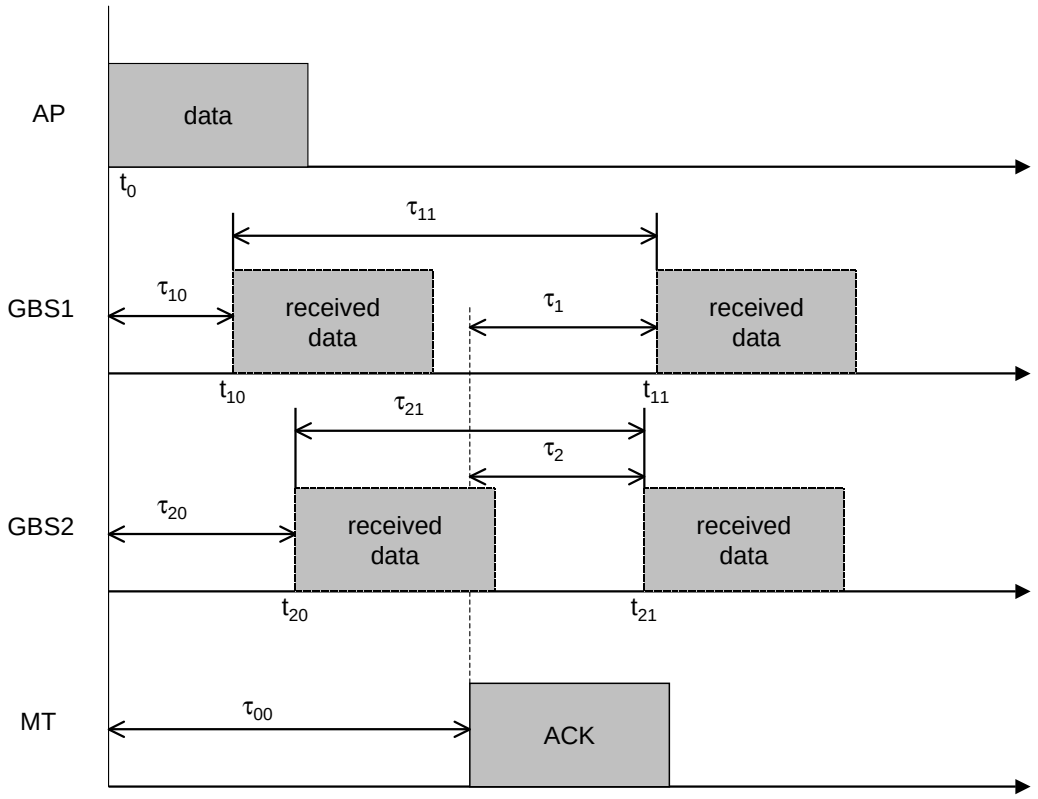

Abbildung 3: Berechnung der TDOA aus verschiedenen Ankunftszeiten

Sind die GBS mit stabilen und präzisen Timern ausgestattet, so genügt es nun die exakten Ankunftszeiten t_{10}, t_{11}, t_{20} und t_{21} zu bestimmen. Die Erkennung des Start Frame Delimiter (SFD) im IEEE 802.11 Rahmen dient als Auslöser für den Timer. Die Genauigkeit des Systems hängt dann von der Qualität der Synchronisation ab. IEEE 802.11 DSSS benutzt hierzu ein 128-Bit langes SYNC Feld und Korrelation. Durch die

Anwendung der Spreizband Technik können 128 Korrelationsspitzen berechnet und zur Synchronisation verwendet werden.

Leider gibt es meines Wissens noch keine praktischen Untersuchung bezüglich der Genauigkeit dieses Systems. Die in [LPLaY00a] durchgeführte Simulation zur Genauigkeit der Symbol–Synchronisation zeigte vielversprechende Ergebnisse. So sind Fehler im Bereich bis zu 20ns zu erwarten, was etwa 6m entspricht. IEEE 802.11 DSSS zeigte dabei eine größere Genauigkeit als HIPERLAN/2 mit OFDM (Orthogonal Frequency Division Mulitplexing), wofür die Autoren in [LPLaY00b] einen ähnlichen Ansatz vorstellen.

4.2.2 Positionsbestimmung durch Messung der Signalstärke

In [BaPa00] wird ein anderer Ansatz vorgestellt. Hier wird ohne eine Erweiterung der vorhandenen Infrastruktur versucht, die Position des mobilen Nutzers durch Messung der Signalstärke zu bestimmen. Dabei werden zwei Vorgehensweisen vorgestellt.

Im ersten Ansatz basiert die Auswertung der Empfangssignalstärke auf einem Vergleich mit vorher gewonnen Daten. Das Gebäude wird sozusagen empfangstechnisch vermessen. Auf einem Stockwerk $(22,5m \times 43,5m)$ sind drei Basisstationen untergebracht die sich gegenseitig überlappen und zusammen die ganze Fläche bedienen. Da die Signalstärke eine starke Abhängigkeit von der Orientierung des Benutzer aufweist, werden ca. 20 Werte für 70 Messpunkte in jeweils 4 Orientierunsrichtungen benötigt. Zur Positionsbestimmung können dann die Mittelwerte über die 20 Messungen benutzt werden. Anschließend kann durch Simulation die Leistungsfähigkeit bestimmt werden. Als Position wird der nächste Nachbar im *Signalraum* verwendet, das heißt die Koordinaten des Punktes, der zu dem zu bestimmenden Punkt die geringste Abweichung in den drei Signalstärken aufweist. Im Signalraum wird dabei die euklidsche Metrik verwendet. Für diese Vorgehensweise ergibt sich für den Median der Fehlerverteilung ein Wert von 2.94m. Werden als Referenzpunkte die stärkste Basisstation beziehungsweise ein zufällig gewählter verwendet, sinkt die Genauigkeit deutlich. Weitere Variationen ergeben, dass bereits ab etwa 40 Messpunkten ein gutes Ergebnis möglich ist—der mittlere Fehler ist im Vergleich zu 70 nur um 10% schlechter. Auch eine Verwendung mehrerer nächster Nachbarn bringt nur eine geringe Verbesserung, da es sich meist nur um die gleichen Koordinaten mit unterschiedlicher Orientierung handelt.

Der zweite gezeigte Ansatz kommt ohne vorherige Vermessung aus. Grundsätzliche ist eine Positionsbestimmung auch durch einfaches Messen der empfangenen Signalstärke möglich. Wird aus solchen Werten eine Entfernung bestimmt, so müssen Zusatzdämpfungen durch Wände und andere Hindernisse berücksichtigt werden. Kennt man den genauen Grundriss und damit für einen Punkt im Raum die Anzahl der Wände die auf dem direkten Weg zur Basisstation liegen, so kann eine Anpassung erfolgen. Die dazu verwendeten Wellenausbreitungsmodelle geben zum Beispiel Funktionen zur Berechnung der Zusatzdämfung durch Wände an [SeRa92]. Die Parameter dieser Funktionen können für ein konkretes Gebäude durch Regressionsanalyse geschätzt werden. Die Simulation dieser Technik in [BaPa00] ergibt für den Median der Fehlerverteilung einen Wert von 4,3m.

4.3 Infrarottechnik am Beispiel Active Badge

Die Technik des Active Badge Systems nutzt die Eigenschaften des Infrarotnetzes zur Positionsbestimmung. Die Räume des Gebäudes sind mit fest installierten Basissensoren ausgestattet, welche durch ein leitungsgebundenes Netzwerk verbunden sind. Die niedrige Datenrate von 9600 Baud ermöglicht kleine und energiesparende Geräte. Das zentrale mobile Gerät ist das Active Badge. Dieses sendet periodisch eine Infrarotnachricht, die von den Basissensoren empfangen wird. Da Infrarotlicht keine Wände durchdringen kann, ist damit eine Positionsbestimmung auf Raumniveau realisiert. Um Energie zu sparen, sind die Sendeintervalle 10s lang. Die wesentlichen Eigenschaften dieser Technologie bezüglich der Positionsbestimmung sind:

- Die Räumliche Auflösung korrespondiert mit natürlichen Grenzen
- Die Zeitliche Auflösung ist so fein, wie die Stromversorgung es zulässt

Eine feinere Auflösung wird durch zusätzliche Funktechnik realisiert. Funksender schwacher Leistung bilden Felder im Größenbereich um einen Meter. Nähert sich ein Benutzer mit Active Badge einem solchen Feld, so kann dieses mittels einer eingebauten Antenne die Funkwellen empfangen und die darin kodierten Daten auslesen. Diese werden dann an den Infrarotsensor im Raum übertragen. Damit merkt das System, wenn sich ein Benutzer einer mit Funksender ausgestatteten Workstation nähert.

Neben der Positionsbestimmung müssen die gewonnen Daten auch noch verwaltet werden. Sicherheitsaspekte und Fragen der Privatsphäre sind dabei zu berücksichtigen. Zusätzlich müssen die Basissensoren vernetzt und angesteuert werden. Eine genauere Übersicht über die verwendeten Ansätze und obige technische Eigenschaften findet sich in [HaHo92].

5 Zusammenfassung

Die Aktivitäten im Bereich Positionsbestimmung in drahtlosen Netzen haben vor allem in jüngster Vergangenheit stark zugenommen. Dies ist sicherlich größtenteils durch die E–911 Bestimmung begründet. Inwieweit für weitere Anwendungen der Technologie tatsächlich Märkte existieren, bleibt abzuwarten. Von technischer Seite aus scheinen im GSM Netz Lösungen, die die Anforderungen der FCC erfüllen, möglich. So hat sich beispielsweise ein großer Mobilnetzbetreiber für ein OTD Verfahren entschieden. Dennoch sind meiner Ansicht nach Lösungen die spezielle Technologien wie GPS benutzen für die Zukunft nicht völlig ausgeschlossen. Problematisch ist bei allen Ansätzen, dass die Positionsbestimmung ohne Veränderungen an den Mobilgeräten funktionieren muss.

Systeme zur Positionsbestimmung im WLAN haben, wie durch Active Badge gezeigt, vielfältige Anwendungsbereiche. Die Entwicklung steht hier scheinbar noch am Anfang. Wahrscheinlich erhält dieser Bereich durch konkrete militärische Aufträge einen Schub.

Es ist auch interessant zu beobachten, dass ein Großteil der Veröffentlichungen aus den letzten drei Jahren stammt und von einem relativ kleinen Kreis von Autoren verfasst wurde. Bestimmte Forschungsgruppen spezialisieren sich anscheinend auf ein gewisses Gebiet und beleuchten dies von verschiedenen Seiten, was auch in Zukunft interessante Fortschritte für die Positionsbestimmung in drahtlosen Netzen erwarten lässt.

Literatur

[Agen] Defense Advanced Research Projects Agency. Small Unit Operations Situation Awareness System. http://web-ext2.darpa.mil/baa/mda972-98-r0004.htm.

[BaPa00] P. Bahl und V. N. Padmanabhan. RADAR: An In–Building RF–based User Location and Tracking System. In *Proceedings of the Conference on Computer Communications (IEEE Infocom)*, März 2000.

[ChGa92] C.-K. Chen und A. Gardner. Signal–Selective Time–Difference–of–Arrival Estimation for Passive Location of Man–Made Signal Sources in Highly Corruptive Environments, Part II: Algorithms and Performance. *IEEE Transactions on Signal Processing* 40(5), Mai 1992, S. 1185–1197.

[DrNS98] C. Drane, M. Nacnaughtan und C. Scott. Positioning GSM Telephones. *IEEE Communications Magazine* Band April, 1998, S. 46–59.

[EHCD+93] S. Elrod, G. Hall, R. Costanza, M. Dixon und J. Des Rivieres. Responsive office environments. *Communications of the ACM* 36(7), Juli 1993.

[EnMi99] P. Enge und P. Misra. Special Issue on Global Positioning System. *Proceedings of the IEEE* 87(1), Januar 1999, S. 3–15.

[Fang90] B. T. Fang. Simple Solutions for Hyperbolic and Related Position Fixes. *IEEE Transactions on Aerospace and Electronic Systems* 26(5), September 1990, S. 748–753.

[Gett93] I. A. Getting. The Global Positioning System. *IEEE Spectrum*, Dezember 1993, S. 36–47.

[Grou99] The Strategis Group. Wireless Communication Services 1999, Oktober 1999.

[HaHo92] A. Harter und A. Hopper. A Distributed Location System for the Active Office. *IEEE Network*, Januar/Februar 1992.

[KrBR97] K. J. Krizmann, T. E. Biedka und T. S. Rappaport. Wireless Position Location: Fundamentals, Implementation Strategies, and Sources of Error. In *IEEE 47th Vehicular Technology Conference*, Band 2, 1997, S. 919–923.

[KrPB98] P. Krishnamurthy, K. Pahlavan und J. Beneat. Radio Propagation Modelling for Indoor Geolocation Applications. In *The Ninth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Band 1, 1998, S. 446–450.

[LPLaY00a] X. Li, K. Pahlavan, M. Latva-aho und M. Ylianttila. Comparison of Indoor Geolocation Methods in DSSS and OFDM Wireless LAN Systems. In *IEEE VTS Fall Vehicular Technology Conference 2000.*, Band 6, 2000, S. 3015–3020.

[LPLaY00b] X. Li, K. Pahlavan, M. Latva-aho und M. Ylianttila. Indoor Geolocation using OFDM Signals in HIPERLAN/2 Wirelss LANs. In *The 11th IEEE International Symposium on Personal, Indoor and Mobile Radio Communicaitons, 2000*, Band 2, 2000, S. 1449–1453.

[MaDo00] Z. Mao und C. Douligeris. High Throughput Database Structures for Location Management in PCS Networks. In *Proceedings of the Conference on Computer Communications (IEEE Infocom)*, März 2000.

[McCa99] E. McCabe. Location-based services offer a global opportunity for new revenue. *Telecommunications* 33(10), Oktober 1999, S. 99–102.

[RaRW96] T. S. Rappaport, J. H. Reed und B. D. Woerner. Position Location Using Wireless Communications on Highways of the Future. *IEEE Communications Magazine*, Oktober 1996, S. 33–41.

[RKWR98] J. H. Reed, K. J. Krizman, B. D. Woerner und T. S. Rappaport. An Overview of the Challenges and Progress in Meeting the E-911 Requirement for Location Service. *IEEE Communications Magazine*, April 1998, S. 30–37.

[Schi00] J. Schiller. *Mobilkommunikation: Techniken für das allgegenwärtige Internet*. Net.com. Addison–Wesley. 2000.

[SeRa92] S. Y. Seidel und T. S. Rappaport. 914 MHz path loss prediction models for indoor wireless communication in multifloored buildings. *IEEE Transactions on Antennas and Propagation* 40(2), Februar 1992, S. 207–217.

[Spir99] M. A. Spirito. Further results on GSM mobile station location. *Electroncs Letters* 35(11), Mai 1999, S. 867–869.

[SpMa98] M. A. Spirito und A. G. Mattioli. On the Hyperbolic Positioning of GSM Mobile Station. In *1998 URSI International Symposium on Signals, Systems, and Electronics (ISSSE 98)*, 1998, S. 173–177.

[WeLa98] J. Werb und C. Lanzl. Designing a positioning system for finding things and people indoors. *IEEE Spectrum*, September 1998, S. 71–78.

[YlMP00] M. Ylianttila, J. Mäkelä und K. Pahlavan. Geolocation Information and Inter-technology Handoff. In *2000 IEEE International Conference on Communications (ICC 2000)*, Band 3, 2000, S. 1573–1577.

[ZiWa88] I. Ziskind und M. Wax. Maximum Likelihood Localization of Multiple Sources by Alternating Projection. *IEEE Transactions on Accoustics, Speech, and Signal Processing* 36(10), Oktober 1988, S. 1553–1560.

[ZPBM98] J. M. Zagami, S. A. Parl, J. J. Bussgang und K. Devereux Melillo. Providing Universal Location Services Using a Wireless E911 Network. *IEEE Communicatios Magazine*, April 1998, S. 66–71.

Abbildungsverzeichnis

Kontextabhängige Dienste

Daniel Vogel

Kurzfassung

Der Begriff Kontext umfasst alle Eigenschaften der Umgebung, sowie die Situation in der ein Ereignis stattfindet. Kontextabhängige Dienste sind Dienste bzw. Anwendungen die Kenntnisse haben über den aktuellen Kontext, und die diese Kenntnisse nutzen um sich auf verschiedenste Art und Weise an die aktuelle Situation und Umgebung anzupassen. Die vorliegende Arbeit beschreibt zunächst den Begriff kontextabhängige Dienste genauer und zeigt wie Kontextinformation gewonnen werden kann. Es werden vielfältige Möglichkeiten aufgezeigt, Kontextinformation in zukünftigen Anwendungen und Systemen umzusetzen. Gleichzeitig wird gezeigt, dass es bereits heute zahlreiche Systeme gibt, die Kontextinformation erfolgreich nutzen. Dabei wird insbesondere der von der Universität Lancaster entwickelte interaktive Reiseführer GUIDE vorgestellt.

1 Was sind kontextabhängige Dienste?

1.1 Der Begriff Kontext

Für das Verständnis der nachfolgenden Ausführungen ist es zunächst entscheidend, den Begriff *KONTEXT* genau zu definieren. In einer allgemeinen Definition ([DUDE]) versteht man unter Kontext bzw. genauer unter situativem Kontext:

> Den inhaltlichen Gedanken - oder Sinnzusammenhang in dem ein Ereignis steht, und der Sach- und Situationszusammenhang, aus dem heraus es verstanden werden muss.

In einer engeren Bedeutung können mit dem Begriff Kontext alle Eigenschaften der Umgebung sowie die Situation, in der ein Ereignis stattfindet, zusammenfassend beschrieben werden. Es wird gezeigt, dass für die Informationstechnologie Geräte interessant sind, die Kenntnisse haben über diese Art von Kontext, d.h. die Umgebung und die Situation in der sie, bzw. ihre Benutzer, sich aktuell befinden. Dabei liegt der Fokus dieser Arbeit auf *mobilen Geräten und Anwendungen.*

Wesentliche Kontextbestandteile sind:

- *Ort und Zeit* : Der Ort bzw. die Lokation ist ein wesentlicher Kontextbestandteil. Dabei kann man unter Lokation zum einen die *absolute geographische Position in Längen- und Breitengraden*, wie sie z.B. von GPS Systemen geliefert wird verstehen. Zum anderen meint Lokation aber auch die *Position relativ zu anderen Objekten* (deren absolute Position bekannt ist), z.b. die Position eines Geräts/Benutzers in einem Auto,in einem Konferenzraum oder in einer Stadt relativ zu Gebäuden. Es gibt zahlreiche Systeme, insbesondere Navigationssysteme für Autos auf Basis von GPS, die fast ausschließlich auf Ortsinformationen aufbauen. Es muss betont werden, dass die Lokation zwar ein ganz wesentlicher, aber eben *nur einer von vielen Bestandteilen* des viel allgemeiner zu verstehenden Begriffs Kontext ist (Man würde sonst von orstabhängigen Diensten/Systemen und nicht von kontextabhängigen sprechen !!). Ein weiterer Kontextbestandteil ist die *Zeit*.

- *Infrastruktur in der Umgebung* : Dazu gehören Geräte wie Drucker, Telefone, Faxgeräte, Scanner, PC´s etc. in der Umgebung. Für kontextbasierte Systeme sind Kenntnisse über Geräte in der Umgebung notwendig, um mit diesen Geräten zusammenarbeiten zu können.

- *Allgemeine, aktuelle und dynamische Informationen über Umgebung :* Dazu gehören z.b. Informationen über Gegenstände, Gebäude, Geschäfte und ähnliches in der Umgebung. Hierzu zählen auch Informationen wie Öffnungszeiten, Speisekarten, Veranstaltungshinweise u.a.. Insbesondere bei Geschäften sind aktuelle Sonderangebote, Produktneuheiten, Werbebotschaften etc. zu nennen. Dynamisch sind diese Informationen deshalb, weil sie sich im Zeitablauf mehr oder weniger schnell ändern.

- *Physikalische Bedingungen* : Ein weiterer Kontextbestandteil sind schließlich Parameter wie Temperatur, Vibrationen, Geräuschpegel, Lichtverhältnisse, etc..

Weiterhin lassen sich einige mehr auf den *Menschen* bezogene Kontextbestandteile nennen:

- *Kenntnis über Benutzer* : Meint Kenntnis über die Gewohnheiten, über die Verfassung, über Absichten und Ziele des Benutzers sowie die Aufgaben, die er verrichten möchte.

- *Kenntnis über sein Umfeld* : Bedeutet Kenntnis über andere Menschen/Benutzer in der Umgebung, Kenntnis über das Verhältnis zu ihnen und evtl. auch Kenntnis über gemeinsam zu erledigende Aufgaben.

Außerdem sollte hier noch der *Gerätezustand* (z.b. Akku) und vor allem der *Zustand der Netzwerkverbindung* genannt werden. Beide Punkte gehören zwar nicht zu den eigentlichen Kontextbestandteilen, sind aber trotzdem entscheidend für die Funktion der mobilen Geräte, die für die Bereitstellung der kontextbasierten Dienste benötigt werden.

Jetzt wird auch ersichtlich *was kontextabhängige Dienste sind.* Kontextabhängige Dienste sind Dienste bzw. Anwendungen die Kenntnisse haben über den aktuellen Kontext,

und die diese Kenntnisse nutzen um sich auf verschiedenste Art und Weise an die aktuelle Situation und Umgebung anzupassen.

Kenntnis des Kontext bedeutet dabei nicht nur Kenntnis der einzelnen Kontextbestandteile für sich alleine betrachtet. Man kann sich leicht vorstellen, dass Kontextbestandteile wie Temperatur, Helligkeit oder auch geographische Lokation für sich alleine heute relativ einfach zu ermitteln sind. Natürlich können diese dann auch direkt für kontextabhängige Anwendungen verwendet werden, z.b. um die Leuchtstärke eines Displays an die Helligkeit der Umgebung anzupassen (vgl. Abschnitt 2.2).Doch Kenntnis des Kontext bedeutet ja Kenntnis der Situation. D.h.vor allem auch komplexere Situationen, wie z.b. der Benutzer ist gerade in einem Meeting (d.h. er ist in Besprechungsraum, er ist dort mit mehreren Kollegen, er ist zu bestimmter Uhrzeit dort, ...), sollten erkannt werden. Dazu müssen verschiedene Kontextbestandteile gleichzeitig kombiniert betrachtet werden.

1.2 Motivation für kontextbasierte Dienste

Jetzt stellt sich die Frage warum kontextabhängige Dienste überhaupt sinnvoll sind.

Die heutige Situation ist dadurch gekennzeichnet, dass es *immer mehr extrem mobile Endgeräte*, wei z.b. ultraslim - und Sub - notebooks, Palmtops und PDA´s gibt. Sie sind in vielen Fällen eine große Hilfe im täglichen Leben. Insbesondere der Handy Markt boomt (knapp 50 Mio. Nutzer in BRD).
Die Nutzer dieser Geräte haben eines gemeinsam. Sie alle wollen, dass ihre Geräte möglichst *vielfältige Funktionen* haben und überall eingesetzt werden können. Im Kontrast dazu steht aber der Wunsch nach *einfacher und schneller Bedienbarkeit*. Die Geräte sollen ohne lange Einarbeitungszeiten, möglichst von jedermann bedient werden können. Erschwerend kommen die meist stark reduzierten Ein- und Ausgabeschnittstellen hinzu.

Vor diesem Hintergrund wird sofort der *zusätzliche Nutzen* von Geräten deutlich, die in der Lage sind die aktuelle Situation und Umgebung des Benutzers zu erkennen und sich daran anzupassen. Durch kontextabhängige Systeme kann die Unterstützung des Benutzers deutlich verbessert werden, es sind viel weniger explizite Benutzerinteraktionen notwendig (z.B. weil automatisch die aktuell benötigte Anwendung geöffnet wird ....) und die Bedienung wird stark vereinfacht. Aus dem nachfolgenden Abschnitt 1.3 und Kapitel 3 wird deutlich, dass *gleichzeitig* durch Verwendung von Kontext aber auch ganz neue Dienste und Anwendugen möglich werden. D.h. die Funktionalität der Geräte kann deutlich erweitert werden. Insgesamt liefern kontextabhängige Systeme dem Benutzer dadurch einen merklichen Zusatznutzen.

1.3 Abgrenzung möglicher Arten der Verwendung von Kontext

Die folgende Auflistung soll einen allgemeinen Überblick geben über die verschiedenen Möglichkeiten Kontextinformationen zu nutzen. Zu den einzelnen aufgeführten Punkten gibt es jeweils sehr viele Beispiele, von denen hier nur einige aufgeführt werden können. Der “Phantasie” sind hier keine Grenzen gesetzt. Dabei soll hier davon ausgegangen

werden, dass die Kontextinformation zur Verfügung steht. Wie Kontextinformation gewonnen wird, wird in Kapitel 2 genauer besprochen.

- *kontextsensitive Anpassung der Benutzerschnittstelle :* Es würde das Arbeiten mit mobilen Geräten viel angenehmer und einfacher machen, wenn sich die ohnehin *eingeschränkte Benutzerschnittstelle jeweils an die aktuelle Situation anpassen würde*. D.h. Parameter wie Schriftgröße, Helligkeit und Kontrast des Displays, Lautstärke des Klingeltons bzw. Vibrationsmodus etc. sollten sich automatisch an die momentane Situation anpassen. Viele Geräte, insbesondere Handys, unterstützen heute schon die Einstellung verschiedener Benutzerprofile. Diese müssen dann aber manuell ausgewählt werden. Unter Nutzung von Kontextinformationen könnte dies automatisch geschehen.

- *Gezielte/gefilterte Bereitstellung aktueller und dynamischer Informationen :* Im Informationszeitalter sind die Menschen einer immer größer werdenden Informationsflut ausgesetzt. Diese Entwicklung wird durch billigere und einfacher zu handhabende Geräte (ständige Erreichbarkeit), durch Vebreitung der Internettechnologie und zunehmende Automatisierung der Informationsverteilung (z.b. mailing Listen) unterstützt. *Kontextbasierte Systeme sind in der Lage diese Informationsflut sinnvoll zu reduzieren.* So könnten Benutzer ihre Präferenzen beispielsweise einmalig eingeben und kontextbasierte Systeme wären dann in der Lage aufgrund dieser Präferenzen und der aktuellen Situation Informationen und Kommunikationsinhalte gezielt zu *filtern*, d.h nur Informationen die für den Benutzer momentan wichtig sind würden angezeigt. Denkbar ist hier z.b. eine *kontextabhängige To - Do - List*. Ist der Benutzer gerade in seinem Büro zeigt sie ihm die hier zu erledigenden Aufgaben an, ist er dagegen daheim, so werden die hier anstehenden Aufgaben angezeigt.
  Andererseits können dem Benutzer mit Hilfe kontextbasierter Systeme ganz gezielt aktuelle und situationsbezogene Informationen gemäß seinen Wünschen bereitgestellt werden. Dazu gehört zum Beispiel die Übermittlung aktueller Nachrichten die für den Benutzer wichtig sind (z.b. starke Kursänderung seiner Aktien) oder die Benachrichtigung über Sonderangebote beim Betreten eines Supermarkts. *Durch die bessere Organisation der Informationsverteilung ist es also letzlich möglich dem Benutzter insgesamt viel mehr Informationen als bisher bereit zu stellen, ohne ihn dabei mehr zu belasten!*

- *kontextsensitive Auswahl von Anwendungen :* Eine Möglichkeit der Verwendung von Kontextinformation besteht in der *automatischen Auswahl der Anwendung die ein Benutzer mit größter Wahrscheinlichkeit in der aktuellen Situation benötigt*. Beispiele finden sich in Tabelle 1.

Kontext	*Wahrscheinlichste Anwendung*
Warten an Straßenbahnhaltestelle	Fahrplan
Meeting	Tabellenkalkulation oder Notizblock
Betreten einen Geschäfts	vorher vom Benutzer erst. Einkaufsliste öffnen

Tabelle 1: Automatische Anwendungsauswahl

- *Navigation* : Der Kontextbestandteil Ortsinformation kann direkt für Navigationssysteme genutzt werden. Aus der Sicht kontextabhängiger Dienste sind besonders Navigationssysteme interessant, die *außer der reinen* - z.b. über GPS gewonnenen - *Ortsinformation noch weitere Informationen nutzen.* So werden seit einiger Zeit im Traffic Message Channel (TMC) permanent und unhörbar über den UKW Rundfunk Verkehrsinformationen übertragen. Neuere Navigationssysteme können diese Informationen direkt in ihre Routenberechnungen mit einbeziehen, um so beispeilsweise Staus zu umgehen.

Für ergänzende Informationen zum Begriff Kontext und zu dessen Verwendung siehe auch [Gell].

2 Wie wird Kontextinformation gewonnen ?

2.1 Möglichkeiten der Kontextgewinnung

Entscheidend für die Realisierbarkeit kontextabhängiger Dienste ist die Frage, wie und in welchem Umfang Kontextinformation überhaupt gewonnen werden kann. Man muss sich immer vor Augen halten, dass ohne effiziente Verfahren der Kontextgewinnung auch keine kontextabhängigen Dienste möglich sind.

Zunächst gibt es *folgende grundsätzliche Möglichkeiten* Kontextinformationen zu gewinnen:

- *Den Benutzer befragen :* Eine Möglichkeit ist sicher, den Benutzer direkt zu fragen. Sinnvoll ist es beispielsweise den Benutzer einmalig aufzufordern seine *Präferenzen oder bestimmte Gewohnheiten* anzugeben. Die so gewonnene Information muss vom Gerät gespeichert werden und kann dann später genutzt werden. Es ist aber klar, dass es nicht sinnvoll ist, jede Art von Kontextinformation so zu gewinnen.

- *Clevere Umgebung :* Eine andere Möglichkeit ist es die Umgebung so auszurüsten, dass diese die mobilen Geräte mit Kontextinformation versorgt.
  Es ist üblich *Ortsinformationen* auf diese Art zu gewinnen. So erlaubt z.b. die Positionierung von *GPS* - Satelliten (Global Positioning System) im Weltall mobilen Geräten ihre Position nahezu weltweit auf 10 Meter genau zu bestimmen. Aufgrund der zu starken Dämpfung der Satelittensignale innerhalb von Gebäuden können GPS Systeme dort jedoch nicht genutzt werden. Daher kommen hier andere Systeme zum Einsatz, wie z.b. das vom Cambridge University Computer Laboratory (vgl. [Webs]) entwickelte *Active Badge System.* Das *Funktionsprinzip* ist recht einfach. Die mobilen Geräte senden alle 10 Sekunden via Infrarot Signal eine eindeutige Kennung aus. In jedem Raum gibt es einen Infrarot Sensor, der dieses Signal und damit die Kennung empfängt. Alle Sensoren sind miteinander und mit einem zenralen Rechner vernetzt. Auf diese Art und Weise kann also genau bestimmt werden in welchem Raum ein Benutzer gerade ist. Eine weitere clevere Umgebung bildet auch die *GSM Infrastruktur.* Hier kann die Zelle ermittelt werden in der ein Benutzer sich gerade befindet.

Es muss jedoch deutlich darauf hingewiesen werden, dass *nicht nur Ortsinformationen* durch eine clevere Umgebung bereitgestellt werden können. Es ist klar dass *zahlreiche andere Informationen* durch die Umgebung geliefert werden können, wie z.b. die oben erwähnte Fahrplaninformation an der Haltestelle, oder einfach der Speiseplan der Mensa wie an der Universität Saarbrücken praktiziert.

- *Kommunikation zwischen Geräten :* Eine weitere wichtige Möglichkeit Kontext - Informationen zu gewinnen ergibt sich aus der Kommunikation zwischen den Geräten selbst. Hier kann Information gewonnen werden über die in der Nähe *vorhandene Infrastruktur* (Computer, Drucker, Telefone, andere mobile Geräte ...), die ein mobiles Gerät nutzen könnte, über andere *Personen in der Nähe*, aber auch über die *exakte Position* innerhalb einer Umgebung, d.h. die Position relativ zu anderen Geräten oder Gegenständen (z.b. wenn der Benutzer gerade bei einem Drucker steht).

- *Clevere Endgeräte :* Natürlich können auch die Geräte selbst so ausgestattet werden, dass sie unabhängig von der Umgebung Kontextinformationen gewinnen können. Diese Art der Kontextgewinnung hat den Vorteil, dass sie *überall funktioniert*. Dieses Ziel kann unter anderem erreicht werden, indem man die Geräte mit Sensoren ausstattet. Insbesondere bestimmte *Umgebungsparameter* wie z.b. Temperatur, Helligkeit etc. können so ermittelt werden. Durch die Kombination vieler kleiner, einfacher und preiswerter Sensoren, ist es aber auch möglich *komplexere Situationen* zu erkennen. Dieser Ansatz ist ein wichtiges aktuelles Forschungsgebiet. In Abschnitt 2.2 wird hierauf ausführlicher eingegangen.

Aus dieser Auflistung wird ersichtlich, dass für die verschiedenen Arten von Kontext jeweils unterschiedliche Methoden besonders geeignet sind, um Informationen darüber zu gewinnen. Bei der Gewinnung von Kontextinformation muss jeweils beachtet werden wie komplex die zu erkennende Situation ist. Es ist beipielsweise relativ einfach, festzustellen ob ein Benutzer gerade in einem Raum ist oder nicht. Häufig sollen aber gerade auch komplexere Situationen erkannt werden, wie z.B. ob der Benutzer in diesem Raum gerade an einem Meeting teilnimmt. Um auch komplexe Situationen erkennen zu können, müssen Kombinationen aller oben aufgeführten Möglichkeiten der Kontextgewinnung verwendet werden. Ein Beispiel hierfür ist der in Kapitel 3 beschriebene Reiseführer GUIDE.

Im folgenden Abschnitt soll jetzt die Möglichkeit, Kontextinformation durch *Clevere Endgeräte* zu gewinnen, anhand eines Projekts näher beschrieben werden.

2.2 Beispielprojekt: Automatiche Benutzerprofilerkennung bei NOKIA 6110

Die meisten NOKIA Handys unterstützen verschiedene *Benutzerprofile*. Diese Profile legen z.b. die Lautstärke des Klingeltons fest, oder bestimmen welche Anrufe angenommen, welche dagegen direkt auf die Mobilbox weitergeleitet werden sollen. Sie müssen "per Hand" in einem Menü aktiviert werden. Ziel des Projekts (siehe dazu auch [Schm]) war es nun mit Hilfe von Kontextinformationen diese *Benutzerprofilauswahl zu automatisieren*. Die Kontextinformation wurde dabei durch Kombination verschiedener

Sensoren gewonnen. An dem Projekt waren insbesondere auch das TecO der Universität Karlsruhe und NOKIA beteiligt. Tabelle 2 zeigt welche verschiedenen Situationen automatisch erkannt werden sollten. Das Benutzerprofil ermöglichte bei dem Versuch ausschließlich die Anpassung des Klingeltons bzw. des Vibrationsmodus.

Situation	*zugehöriges Benutzerprofil*	*Bemerkungen*
Handy in Hand	Nur Vibrationsalarm	–
Handy auf Tisch	Das Handy klingelt leise	Benutzer in Meeting o.ä.
Handy im Rucksack	Telefon soll nicht klingeln	Ankommende Anrufe erscheinen später in Anrufliste
Handy im Freien	Klingeln auf voller Lautstärke und Vibrationsalarm	–
Allgemein	Normale Klingellautstärke, keine Vibration	Falls keine der anderen Situationen zutrifft.

Tabelle 2: Automatisch zu erkennende Situationen

2.2.1 Verwendete Hardware

Das Herz ist eine Platine, auf der 8 verschiedene Sensoren intergriert sind. Eine *Fotodiode* misst die Helligkeit, zwei *Beschleunigungssensoren* messen die Beschleunigung in 2 Raumrichtungen und erkennen Vibrationen, ein *passive Infrarotsensor* erkennt wärmeerzeugende Objekte - also z.b. Personen - in der Nähe, ein *Temperatursensor* misst die Temperatur und ein *Drucksensor* den Luftdruck in der Umgebung, ein *CO Gas Sensor* misst schließlich den Kohlendioxidgehalt der Luft. Die analogen Signale der Sensoren werden in einem *A/D Wandler* in digitale Signale umgewandelt. Ein *Mikroprozessor (PIC)* fragt die Sensoren nun reihum ab, und stellt die so gewonnenen Signale dem angeschlossenen Gerät (hier ein Notebook) zur Verfügung. Zusätzlich ist ein Mikrofon direkt an dieses Gerät angeschlossen.

Die *Versuchsumgebung* gestaltet sich wie folgt: Die Platine ist 17*10 cm groß. Für die Versuche ist sie an ein Notebook angeschlossen, auf dem die unten beschriebene *Kontext-Erkennungssoftware* installiert ist. Das Handy wiederum ist an dieses Notebook angeschlossen. Man kann sich heute leicht vorstellen, die Platine wesentlich kompakter zu gestalten und sie mitsamt der Kontext-Erkennungssoftware direkt im Handy zu integrieren.

2.2.2 Datenanalyse

Um später verschiedene Situationen erkennen zu können, ist es zunächst notwendig, die Daten der verschiedenen Sensoren in den verschiedenen Situationen zu analysieren. Dazu wird die Platine in verschiedene Umgebungen gebracht, und dort jeweils die Verläufe der Messergebnisse der verschiedenen Sensoren über die Zeit betrachtet und in Diagrammen dargestellt.
Das in Abbildung 1 beispielhaft aufgeführte Diagramm zeigt 3 Situationen: Zuerst lag die Platine für 100 Sekunden auf dem *Schreibtisch*, dann hielt sie der Benutzer 100

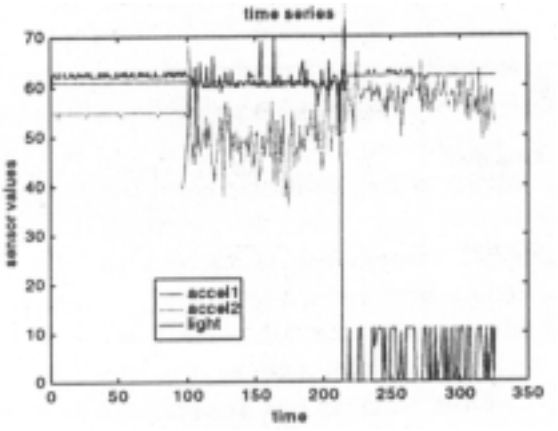

Abbildung 1: Messergebnisse der Sensoren

Sekunden in der *Hand* und schließlich wurde sie für weitere 100 Sekunden in einem *Rucksack* transportiert. In diesem Diagramm werden nur die beiden Beschleunigungssensoren und der Lichtsensor ausgewertet. Für die 3 Situationen erkennt man ganz typische Verläufe der 3 Sensormesswerte. Außerdem wird deutlich, dass sich die 3 Sensoren ergänzen. Es wäre z.b. schwer die Situation "Hand" und "Rucksack" zu unterscheiden, wenn es keinen Lichtsensor gäbe.
Um diese Verläufe quantitativ zu beschreiben werden nun aus den Messreihen *charakteristische Kennzahlen*, wie z.b. arithmetisches Mittel, Varianz und Standardabweichung, Quartilsabstand und die erste Ableitung etc., berechnet. Wie man sich leicht vorstellt ergeben sich für jede Situation ganz typische Kombinationen dieser charakteristischen Kennzahlen. Um die Situationen voneinander abgrenzen zu können, werden jetzt für die verschiedenen Situationen Grenzwerte der Kennzahlen ermittelt, für die sich der Benutzer/das Gerät *gerade noch* in einer bestimmten Situation befindet. Die so durch die Datenanalyse gewonnenen Grenzwerte werden später zur Erkennung des Kontext verwendet.

2.2.3 Vier - Schichtenarchitektur und Kontexterkennung

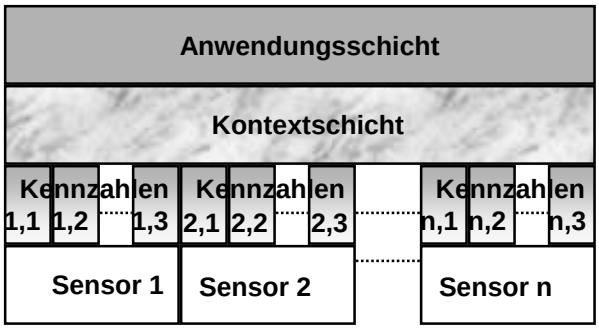

Abbildung 2: Die Schichtenarchitektur

Zunächst wird dazu die in Abbildung 2 gezeigte *Vier-Schichtenarchitektur* eingeführt. Die Funktion der einzelnen Schichten :

- *Sensorschicht :* Physikalische Schicht. Entspricht der beschriebenen Platine. Mit Hilfe der Sensoren werden zunächst die verschiedenen Messreihen ermittelt, und der Kontexterkennungssoftware im Notebook zur Verfügung gestellt.

- *Kennzahlen - Schicht :* Mit Hilfe der Erkennungssoftware werden nun jeweils aus der Messreihe eines Sensors die verschiedenen, oben beschriebenen, Kennzahlen berechnet. Diese Kennzahlen werden der Kontexterkennungsschicht zur Verfügung gestellt.

- *Kontexterkennungsschicht :* Hier wird jetzt mit der Hilfe der Kennzahlen der aktuelle Kontext bestimmt. Dazu werden zunächst *verschiedene sich ausschließende Zustände* definiert:

  - Zustand 1: Handy in der Hand vs. Handy im Rucksack vs. Handy auf dem Tisch
  - Zustand 2: Nutzung innerhalb eines Gebäudes vs. Nutzung im Freien

  Mit Hilfe der in der *Datenanalyse* (siehe Abschnitt 2.2.2) gewonnenen Grenzwerte kann nun festgestellt werden, welche Kombination von Zuständen vorliegt. Also z.b. Handy in der Hand *und* Nutzung im Freien oder Handy auf dem Tisch *und* Nutzung innerhalb eines Gebäudes... . Für die Erkennung der vorliegenden Zustandskombination werden, wie in Tabelle 3 für die Erkennung des ersten Zustands (Hand, Rucksack oder Tisch) gezeigt, einfache Regeln implementiert. Dabei werden die aktuellen Ergebnisse der Beschleunigungssensoren (accelX und accelY) und der Photodiode (light) ausgewertet und mit den in der Datenanalyse gewonnenen Grenzwerten Dx, Dy, D, L, Q der Kennzahlen verglichen. Aus der

Hand(t) :	Standardabweichung(accelX,t) $>$ Dx, Standardabweichung(accelY,t) $>$ Dy, Bem.: Gerät bewegt sich in X- und Y-Richtung Durchschn.(light,t) $>$ L , Bem.: Es ist also nicht dunkel
Tisch(t) :	
Rucksack(t) :	Durschn.(light,t) $<$ L , Bem.: Es ist völlig dunkel

Tabelle 3: Regeln für die Bestimmung der Zustände

  so ermittelten Zustandskombination folgt unmittelbar die für die Anpassung der Klingeltöne benötigte Kontextinformation. Diese Information wird der Anwendungsschicht übergeben.

- *Anwendungsschicht :* Entspricht in diesem Fall dem Handy Nokia 6110. Mit Hilfe der Kontextinformation wird hier das passende Benutzerprofil eingestellt.

2.2.4 Ergebnisse

Das wichtigste Resultat ist, dass im Experiment in *90% aller Fälle der Kontext richtig erkannt* wurde. Probleme kann es geben aufgrund von *eventuellen Mehrdeutigkeiten* (z.B. wird wenn es draußen sehr dunkel ist, wird immer angenommen, das Handy sei

im Rucksack). Außerdem können nur *einige wenige Situationen* (vgl. Tabelle 2) erkannt werden, und es gibt daher Situationen, die nicht definiert sind. Das Handy weiß dann nicht wie es reagieren soll, bzw. geht fälschlicherweise von einer der definierten Situationen aus.
Je mehr verschiedene Kennzahlen jedoch verwendet werden, und je mehr verschiedene Zustandskombinationen zur Erkennung vorgesehen werden (mit Hilfe komplexerer Erkennungsregeln), desto mehr verschiedene Situationen können dann auch *zuverlässig* erkannt werden. Das ist also eine Frage der Komplexität der verwendeten Software.

Bis zu dieser Stelle wurde allgemein darauf eingegangen was Kontext ist, wie Kontext genutzt werden kann und wie er gewonnen wird. Im Folgenden soll nun das GUIDE-Projekt besonders auch im Hinblick auf diese Aspekte betrachtet werden.

3 Das GUIDE - Projekt

GUIDE (siehe auch [Lanc]) ist ein tragbarer, interaktiver und vor allen Dingen kontextsensitiver Reiseführer. Er wurde im Rahmen eines Projekts der Universität Lancaster und der Stadt Lancaster von April 1997 bis Juli 1999 entwickelt und erprobt. Wer heute die Stadt Lancaster besucht, kann ein GUIDE-Endgerät im Tourismusbüro ausleihen und es verwenden. GUIDE unterstützt verschiedene, später genauer beschriebene, Dienste wie z.b. Berechnung einer Tour und Führung der Touristen, Information über Sehenswürdigkeiten, Bereitstellung aktueller dynamischer Informationen (Wetter, Verkehr, Veranstaltungshinweise ...), Interaktive Dienste, Kommunikation ... .

3.1 Hardware - Architektur

Als *Endgerät* kommt das *Fujitsu Teampad 7600* (vgl. auch [Fuji]) zum Einsatz. Es misst 213*153*15 mm und wiegt 850g. Das Teampad ist stiftbasiert, es besitzt einen Pentium 166 MMX Prozessor. Außerdem ist es mit einer WaveLAN PC-Karte ausgerüstet. Mit einer Batterieladung kann es bei gleichzeitiger Benutzung der Netzwerkkarte 2 Stunden betrieben werden (was ein Problem ist). Das Display ist so ausgelegt, dass es auch in direktem Sonnenlicht gelesen werden kann.

Abbildung 3: Screenshot der Benutzeroberfläche

Abbildung 3 zeigt die *Benutzeroberfläche* von GUIDE. Das GUIDE System ist *webbasiert*. Auf jedem Endgerät ist ein *HotJava Webbrowser* installiert. Der Großteil der Displayfläche dient zur Darstellung der Informationen. Diese sind *HTML basiert*. In der Abbildung ist insbesondere auch ersichtlich, dass die verschiedenen Seiten durch Links miteinander verbunden sind. Der webbasierte Ansatz wurde im Sinne einer einfachen, intuitiven Bedienbarkeit gewählt, und insbesondere auch weil man davon ausgehen kann, dass ein Großteil der Benutzer heutzutage weiß, wie man einen Webbrowser bedient.

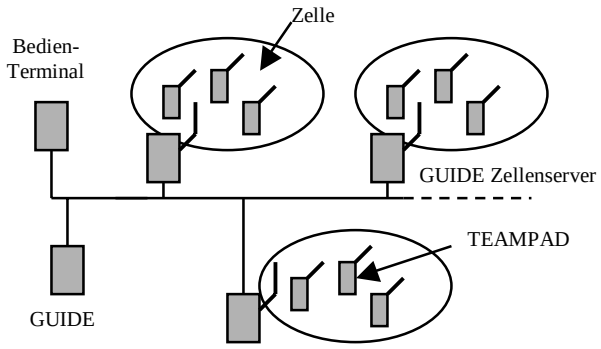

Abbildung 4: Systemarchitektur

Im Gegensatz zu anderen elektronischen Reiseführern basiert GUIDE auf einem *drahtlosen Netz* gemäß dem IEEE 802.11 Standard. Nur so ist es möglich aktuelle, an die Situation angepasste *dynamische Informationen* bereitzustellen und *interaktive Dienste* zu unterstützen. Außerdem gewährleistet die Netzwerkinfrastruktur die Gewinnung von *Ortsinformationen*.

Abbildung 4 zeigt die *Systemarchitektur*. Die Stadt wird dabei in *WaveLAN Zellen* aufgeteilt. Zu jeder Zelle gehört ein mit einer WaveLan Karte ausgestatteter *Zellen-Server* (Proxy Cache). Pro Zelle steht eine Datenrate von *2 MBit/s* (geteilt !) zur Verfügung. Die Zellen haben einen Durchmesser von *bis zu 300m*, sind also relativ groß. Die einzelnen Zellen-Server sind untereinander und mit dem *GUIDE WebServer* sowie einem *Bedienterminal* verbunden. Um Zugriffszeiten zu verringern, unterstützen die Zell-Server *Caching Mechanismen*, d.h. sie speichern die Informationen, die in der zugehörigen Zelle besonders häufig nachgefragt werden. Weil die Zugriffsmuster der Benutzer innerhalb einer Zelle bei einem Reiseführer naturgemäß sehr ähnlich sind, muss der zentrale Webserver also nur selten befragt werden.

Da die Zellen bei gleichzeitig beschränkter Bandbreite relativ groß sind, und daher pro Zelle bis zu 100 Benutzer verkraftet werden müssen, würde ein *CSMA/CA Ansatz überhaupt nicht skalieren*. Stattdessen wurde ein *Broadcast Ansatz* gewählt. Dabei sendet der Zellen-Server (Proxy Cache) periodisch die in seiner Zelle am häufigsten benötigten Informationen in die Zelle. Diese werden dann lokal auf dem Fujitsu Teampad, das einen relativ großzügigen *8MB Speicher* besitzt, zwischengespeichert. Außerdem wird ein *Index* mitgeschickt, aus dem hervorgeht, welche Informationen bereits geschickt wurden, und welche als nächstes geschickt werden. Möchte der Benutzer nun auf bestimmte Informationen zugreifen, wird zunächst überprüft, ob diese bereits im Speicher vorliegt. D.h. auch der Speicher in dem Teampad fungiert als Proxy Server. Falls die gewünschten Informationen nicht vorliegen, wird anhand des Index geprüft, ob sie in nächster Zeit übertragen werden. Nur falls dies nicht der Fall ist, wird ein *explizites*

Request beim Zellen-Server eingeleitet. Würden alle Benutzer innerhalb einer Zelle auf unterschiedliche Informationen zurückgreifen wollen, würde auch dieser Ansatz nicht skalieren. Wie bereits erwähnt hat sich aber gezeigt, dass die Zugriffsmuster innerhalb einer Zelle sehr ähnlich sind. Außerdem kann Dank des großen Speichers praktisch die gesamte für eine Zelle relevante Information abgespeichert werden. In den allermeisten Fällen liegt die Information also bereits lokal auf dem Teampad vor, bevor ein Benutzer sie überhaupt nachfragt. Dies wird auch dadurch unterstützt, dass ein Benutzer meist erst dann darüber informiert wird, dass er eine neue Zelle betreten hat, wenn der Speicher bereits gefüllt ist. *Explizite Requests sind daher äußerst selten.*

Durch dieses Vorgehen skaliert das System auch bei vielen Benutzern sehr gut, und die Zugriffszeiten bleiben gering. Da beim Empfangen über eine WaveLAN Karte wesentlich weniger Energie benötigt wird, als beim Senden, wird dadurch außerdem sehr viel der für mobile Geräte sehr knappen Ressource *Energie gespart.* Außerdem kann GUIDE so auch falls die Verbindung kurzzeitig unterbricht weiterarbeiten.

3.2 Software - Architektur

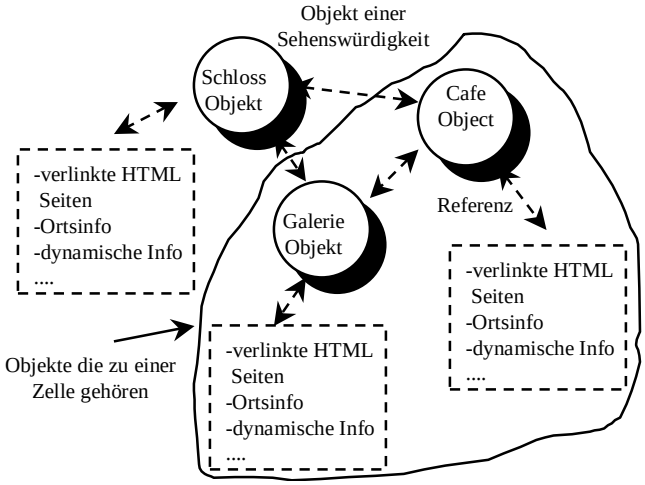

Abbildung 5: Objektmodell

Bei GUIDE kommt ein *objektorientierter Ansatz* zum Einsatz (Vergleiche hierzu Abbildung 5). Dieser ist die *wesentliche Grundlage* für die Bereitstellung kontextsensitiver Informationen. Dabei ist das gesamte Objektmodell zunächst auf dem GUIDE Webserver implementiert.
Für jede Attraktion in der Stadt gibt es ein Objekt (z.b. das Cafe Objekt oder das Schloss Objekt). Jedes dieser Objekte hat nun Referenzen auf verschiedene Informationen, wie folgende Auflistung zeigt:

- *Referenzen auf allgemeine und aktuelle Informationen :* Hierzu gehören die Öffnungszeiten, Eintrittspreise, Veranstaltungshinweise oder z.b. die aktuelle Tageskarte (Cafe Objekt) und vieles mehr. Außerdem Informationen zum aktuelle Status des Objekts, also z.b. ob es gerade geöffnet hat. Zusätzlich gibt es Zähler die z.b. angeben wie oft der Benutzer das entsprechende Objekt bereits besucht hat.

- *Referenz auf die Beschreibung des Objekts :* Zu dieser Beschreibung gehören alle verfügbaren geschichtlichen, architektonischen, kulturellen, etc. Informationen zu diesem Objekt.

- *Referenzen auf geographische Information :* Hierzu gehören Informationen über den Ort des Objekts innerhalb der Stadt, bzw. auf dem Stadtplan. Außerdem sind auch die Positionen der nächsten Nachbarn gespeichert. Es ist weiterhin bekannt welche Möglichkeiten es gibt diese Objekte zu erreichen (zu Fuß, Buslinie xy, ...), *und wie* diese Objekte am besten erreicht werden können. Jedes Objekt hat außerdem eine Referenz auf alle anderen Objekte in der Nachbarschaft.

Vor allem die Informationen zur Beschreibung von Objekten liegen in Form von miteinander verlinkten HTML Seiten vor.

Eine Besonderheit von GUIDE ist es nun, dass diese HTML-Seiten ihrerseits mit Hilfe von *GUIDE TAGS* auf die im Objektmodell gespeicherten Informationen zurückgreifen können, also *dynamisch* sind. Ein kurzes Beispiel für eine solche HTML Seite :

....
Welcome to <GUIDETAG INSERT POSITION>
From here you can visit the neighbouring locations of:
<GUIDETAG INSERT NEIGHBOURS>
.....

Beim ersten GUIDETAG würde also die aktuelle Position (z.b. Schloss) in die HTML-Seite eingefügt, beim zweiten TAG eine Liste der nächsten Nachbarn. Das Cafe könnte z.b. je nach Tageszeit oder Wochentag eine andere Speisekarte darstellen. Man erkennt, dass diese dynamischen Seiten auf einfache Art und Weise die Darstellung kontextsensitiver Informationen erlauben.

Das gesamte Objektmodell ist nun auf dem GUIDE Webserver gespeichert, und kann vom Bedienterminal aus verwaltet werden (vgl. Abbildung 4). Betritt ein Benutzer nun

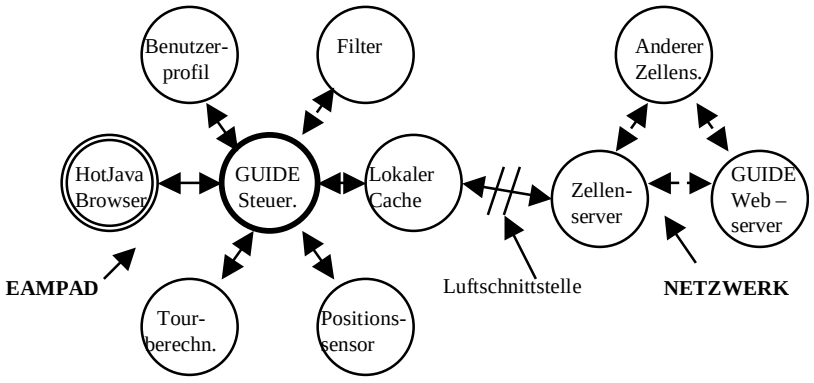

Abbildung 6: Software Architektur

eine Zelle, so wird, wie im Abschnitt 3.1 über die Hardware Architektur beschrieben, der für diese Zelle relevante Teil des Objektmodells (d.h. die in dieser Zelle vorhandenen Objekte mitsamt der referenztierten Information und die für die Tourberechnung

notwendigen Informationen) vollständig und ungefiltert auf das Teampad übertragen. Die gesamte Verarbeitung erfolgt auf dem Teampad. Die einzelnen dazu notwendigen und in Abbildung 6 gezeigten Komponenten sollen kurz erläutert werden:

Lokaler Cache Speicher : Ist der Speicher des Teampad. Hier sind alle zu der aktuellen Zelle gehörenden Informationen gespeichert. Diese werden vom *Zellen-Server* (Proxy Cache) und falls dort nicht vorhanden vom *GUIDE Webserver* beim Betreten der Zelle hierher übertragen.

HotJava Browser : Dieser stellt die Informationen dar. Manche Informationen werden automatisch (durch GUIDE Steuerung) angezeigt, andere auf Anfrage des Benutzers.

Benutzerprofil : Zu Beginn einer Tour wird der Benutzer aufgefordert bestimmte Fragen zu seiner Person (Name, Alter, Sprache ...) und zu seinen Interessen (Architektur, Geschichte, Essen ...) zu beantworten. Daraus wird ein *Benutzerprofil* erstellt, das ebenfalls lokal auf dem Teampad gespeichert ist.

GUIDE Steuerung : Ist die zentrale Komponente. Alle Komponenten werden durch sie gesteuert und abgefragt. Sollen mit dem Browser bestimmte Inhalte angezeigt werden (entweder auf Anfrage des Benutzers oder weil die GUIDE Steuerung es für sinnvoll hält) holt die Steurung sich zunächst die entsprechenden HTML Seiten aus dem lokalen Cache-Speicher, bzw. dem dort vorliegenden, zur Zelle gehörigen Objektmodell. Diese HTML Seiten werden nun der *Filter Komponente* zugeführt.

Filterkomponente : Hier werden die Seiten jetzt auf GUIDETAGS hin untersucht. Die an der Stelle der GUIDETAGS einzufügenden *aktuellen und dynamischen Informationen* werden jetzt ebenfalls aus dem Proxy Cache geholt. Außerdem werden die HTML Seiten jetzt mit Hilfe des Benutzerprofils an die Wünsche des Benutzers angepasst. Interessiert sich ein Benutzer z.b. nicht für Architektur werden diese Informationen herausgefiltert bzw. ganz an das Ende des Dokuments gestellt.

Tourenberechner : Dieser enthält Algorithmen zur Berechnung der Touren aus den Informationen die das Objektmodell im Proxy Cache bereitstellt.

Positionssensor : Dieser ermittelt (unten beschrieben) die aktuelle Position innerhalb der Zelle und stellt diese der GUIDE Steurung zur Verfügung.

Ein Beispiel : Möchte ein Benutzer z.b. über andere Sehenswürdigkeiten in der Nähe informiert werden, wird über die GUIDE Steuerung und den Proxy Cache zunächst ermittelt welche Sehenswürdigkeiten in der Nähe sind. Das Filter ordnet diese nun gemäß den Informationen über die Öffnungszeiten (im lokalen Cache Speicher), den Benutzer (Benutzerprofil) und die Erreichbarkeit (lokaler Cache Speicher + Tourenberechner) in einer Liste an. Die GUIDE Steuerung gibt das Dokument an den Browser weiter, wo es dargestellt wird.

Der *objektorientierte Ansatz* und die beschriebene *Software-Architektur* machen es so möglich dem Benutzer kontextsensitive, dynamische und aktuelle Informationen bereitzustellen.

3.3 Kontextbezug im GUIDE-Projekt

3.3.1 Genutzte Kontextinformation und Gewinnung dieser Kontextinformation

GUIDE nutzt verschiedene Kontextbestandteile (vgl. jeweils Abschnitt 1.1: Der Begriff Kontext und Abschnitt 2.1: Möglichkeiten der Kontextgewinnung) :

- *Ort und Zeit :* Zunächst liefert die Zellenstruktur selbst eine grobe Ortsinformation. Wird eine neue Zelle betreten, empfängt das Teampad zusätzlich zu den anderen Informationen, vom Zellenserver eine Benachrichtigung. Die aktuelle Zelle wird unter *Your location cell* im Display eingeblendet (vgl. Abbildung 3. Eine Zelle kann aber bis zu 300 m groß sein. Daher werden zusätzlich an strategischen Positionen innerhalb der Zelle kleine Sender angebracht, die periodisch Informationen über die Lokation innerhalb der Zelle aussenden (z.b. Benutzer steht vor dem Schloss). Diese Signale werden von dem *Positionssensor* (Empfänger) des Teampad empfangen. Das Teampad empfängt vom Zellenserver zusätzlich eine *Karte* der Zelle, die dem Benutzer über das *Button Map* auch angezeigt werden kann. Ein Punkt auf dieser Karte kennzeichnet die aktuelle Position. Diese Art der Positionsgewinnung wurde hier gegenüber GPS vorgezogen, weil sie mit weniger zusätzlichem Aufwand (sonst hätte ein GPS Empfänger in Teampad integriert werden müssen) realisiert werden konnte und weil es innerhalb einer Stadt nicht immer gewährleistet ist ausreichend viele Satelliten für die Positionsbestimmung zu empfangen. Die Zeit wird dem Teampad über den Zellenserver mitgeteilt. Kenntnisse über Ort und Zeit werden bei GUIDE also durch einen *Clevere Umgebung* gewonnen.

- *Allgemeine, aktuelle und dynamische Informationen über Umgebung :* Dynamische Informationen spielen bei GUIDE eine große Rolle. Beispiele sind: Die *Öffnungszeiten* der einzelnen Attraktionen. Da z.b. das Schloss von Lancaster gleichzeitig als Gerichtshof dient, sind Besuche hier nur möglich, falls keine Verhandlung stattfindet. Die Besuchszeiten können sich somit relativ kurzfristig ändern. Weiterhin zählen lokale *Wetter - und Verkehrsberichte*, aktuelle *Veranstaltungshinweise*, *Warteschlangenlängen* oder *Speisekarten* der verschiedenen Restaurants zu den von GUIDE genutzten dynamischen Informationen. Diese Informationen werden zum größten Teil über das Bedienterminal auf dem GUIDE Webserver gespeichert, und gelangen von dort wie beschrieben zum Teampad. Hier können diese Informationen nun über den bereits beschriebenen Mechanismus der dynamischen Webseiten und den GUIDETAGS dem Benutzer in ansprechender Weise verfügbar gemacht werden. Dynamische Informationen werden also ebenfalls über eine *Clevere Umgebung* in Form der GUIDE Netzwerkinfrastruktur dem Teampad zur Verfügung gestellt.

- *Kenntnis über Benutzer :* Wie bereits erwähnt, wird der Benutzer zu Beginn einer Tour aufgefordert bestimmte Fragen zu seiner Person (Name, Alter, Sprache

...) und zu seinen Interessen (Architektur, Geschichte, Essen ...) zu beantworten. Daraus wird ein *Benutzerprofil* erstellt, das ebenfalls lokal auf dem Teampad gespeichert ist. Kontextinformation wird in diesem Fall also durch *direktes Befragen des Benutzers* gewonnen.

- *Kenntnis über Benutzer in der Umgebung :* Hat ein Benutzer eine neue Zelle betreten sendet das Teampad seinerseits in einer kurzen Nachricht den Namen des Besuchers über den Zellenserver an den Webserver. Dort ist also jederzeit bekannt, welcher Benutzer sich gerade in welcher Zelle befindet. Diese Information wird für den unten beschriebenen *Messaging Service* benötigt, mit dem Besucher miteinander kommunizieren können (z.b. interessant für eine Touristengruppe, die sich trennt um verschiedene Attraktionen zu besuchen).

Bezüglich der *Netzwerkverbindung* lässt sich schließlich noch folgendes anmerken: In der aktuellen Zellenstruktur von GUIDE kann es vorkommen, dass die Netzwerkverbindung vorübergehend unterbrochen wird. Da ein Großteil der für die Zelle relevanten Information lokal gespeichert ist, scheint dies zunächst kein Problem zu sein. GUIDE baut aber in hohem Maße auf aktuellen dynamischen Informationen und auf Ortsinformationen auf, und unterstützt weiter unten beschriebene interaktive Dienste. Ist die Netzwerkverbindung unterbrochen, kann es passieren, dass die lokal gespeicherte Information nicht mehr mit der aktuell auf dem Zellenserver vorliegenden Information übereinstimmt.
Ein nicht besonders befriedigender Ansatz wäre es in Zeiten der getrennten Netzwerkverbindung den Betrieb ganz einzustellen. Ein anderer einfacher Ansatz ist es, die Verbindungsqualität vor dem Benutzer *zu verstecken* und das Gerät offline weiterzubetreiben. Dies kann, insbesondere bei häufigem Wechsel zwischen dem Zustand verbunden und nicht verbunden, dazu führen, dass die Benutzer verwirrt werden und das Vertrauen in GUIDE verlieren. Daher wurde für GUIDE ein anderer Ansatz gewählt. *Der Benutzer wird auf vielfältige Weise über die Verbindungsqualität informiert.* Ohne groß nachzudenken, soll er sich darüber im Klaren sein, dass bei eingeschränkter Verbindungsqualität auch nur eingeschränkte Informationen zur Verfügung stehen. Dies geschieht durch verschiedene Maßnahmen.
Zunächst wird eine aus dem Mobilfunk allseits bekannte und akzeptierte Technik eingesetzt. Oben im Display wird die Empfangsstärke mit Hilfe einiger *Balken* eingeblendet (Abbildung 3). Außerdem gibt es unten im Display ein *Status Zeile*, aus der hervorgeht ob und was GUIDE gerade empfängt.
Bezüglich der Aktualität der Ortsinformationen wird weiterer Aufwand betrieben. In einer Zeile namens *Location* wird normalerweise die aktuelle Zelle eingeblendet. Ist die Netzwerkverbindung unterbrochen, ändert sich die Überschrift der Zeile von Location nach *Last known location cell* und in der Status Zeile wird Auskunft darüber gegeben, wann das letzte mal Ortsinformation empfangen wurde (z.b. Last location update received over 10 minutes ago). Der Benutzer wird so unterrichtet, dass die aktuelle Ortsinformation nicht mehr zuverlässig ist. Um trotzdem, wenn auch eingeschränkt, weiterhin den Benutzer entlang der Tour führen zu können, stellt dass System dem Benutzer *Ja/Nein Fragen* wie: Können Sie das Schloss sehen ? Wir waren zuletzt im Gebiet des Touristenbüros. Wissen Sie ob wir dort immer noch sind ? ... .

Zusammenfassend läßt sich sagen, dass der Großteil der Information dem Teampad über eine *Clevere Infrastruktur* zur Verfügung gestellt wird. Allerdings sind diese Informationen bis dahin *völlig ungefiltert.* Erst im Teampad werden die Informationen über den

Filter an die individuellen Wünsche des Benutzers angepasst. So werden ungewünschte Informationen ausgefiltert, für den Benutzer wichtige Informationen erscheinen vor unwichtigen und die benötigten aktuellen und dynamischen Informationen werden eingefügt. Insofern kann man durchaus auch von Kontextgewinnung durch *Clevere Endgeräte* sprechen. Interessant ist, dass dabei nur ein Sensor, nämlich der Positionssensor, zum Einsatz kommt.

3.3.2 Umsetzung der Kontextinformation

Mit Hilfe der gewonnen Kontextinformationen bietet GUIDE die folgenden Dienste an (vgl. Abschnitt 1.3 :Abgrenzung möglicher Arten der Verwendung von Kontext) :

- *Ortsbestimmung :* Lokal gespeichert sind einfache Karten der jeweiligen Zelle und der ganzen Stadt. Mit Hilfe der *Zellenstruktur* (d.h. dem Wissen in welcher Zelle der Benutzer gerade ist) des *Ortssensors* (Position innerhalb der Zelle) und der *Karten*, kann der Benutzer jederzeit darüber unterrichtet werden, wo er sich gerade befindet. Dazu muss er nur das *Map Button* (vgl. Abbildung 3) drücken. Der Besucher kann sich außerdem jederzeit über das *itshape Info Button* darüber informieren welche Attraktionen gerade in der Nähe sind. Es wird ihm dann eine gemäß seinem *Profil* und den *Öffnungszeiten* angepasste Liste angezeigt. Bereits besuchte Attraktionen erscheinen am Ende der Liste.

- *Gezielte Präsentation statischer Informationen :* Das sind Informationen über Geschichte, Architektur und ähnliches, die sich in der Regel nicht ändern. Es ist bekannt in welcher Zelle sich der Benutzer gerade aufhält. Die zur Zelle gehörende Information befindet sich im lokalen Speicher des Teampad. Durch den *Ortssensor* ist weiterhin bekannt vor welcher Attraktion der Besucher sich aufhält. Die dazu relevante Information wird nun gemäß dem *Benutzerprofil* aufbearbeitet (gefiltert und sortiert) und dem Benutzer *automatisch* präsentiert. Der Benutzer kann nun wie vom WWW gewohnt mit Hilfe der Links und der *Back und Forward Buttons* durch diese gefilterte und wohl sortierte Informationen "surfen". Die wichtigsten Informationen werden dabei sofort auf den ersten Seiten dargestellt. Der Besucher kann nun je nach Interesse mehr oder weniger tief in die Materie einsteigen.

- *Gezielte Präsentation dynamischer Informationen :* Die statischen Informationen werden zusätzlich mit den zahlreich zur Verfügung stehenden dynamischen Informationen (Öffnungszeiten, Veranstaltungen, Speisekarten etc) angereichert. Wesentlich hierfür sind die beschriebenen *dynamisch programmierten Webseiten.*

- *Informationen auf Anfrage :* Der Benutzer kann über das *Info Button* jederzeit auf alle auf dem GUIDE Webserver gespeicherten Informationen zurückgreifen. Also auch auf solche, die nicht die Zelle betreffen in der er sich gerade aufhält.

- *Geführte Touren :* Mit Hilfe der *Öffnungszeiten*, des *Benutzerprofils* und dem Wissen über die Abstände zwischen den Sehenswürdigkeiten wird eine *individuelle Tour* berechnet. Der Benutzer wird von GUIDE dann mit Hilfe der vorhandenen Ortsinformation von Attraktion zu Attraktion geführt. Der Weg wird im schriftlich und auf der Karte angezeigt. GUIDE macht auch Bemerkungen zu interessanten Stellen zwischen zwei Sehenswürdigkeiten. Bei einer Attraktion

angekommen stehen dem Benutzer dann die oben beschriebenen Informationsmöglichkeiten zur Verfügung. Im Gegensatz zu einem "menschlichen Reiseführer" kann der Besucher solange wie er möchte bei den einzelnen Attraktionen verweilen, oder die Tour auch für längere Zeit unterbrechen. Erst durch drücken des *Show Next Instruction* Buttons wird er zur nächsten Attraktion geführt. Andererseits kann er sich über das *Repeat Last Instruction* Button jederzeit nochmal über bereits besuchte Orte informieren lassen.

- *Interaktive Dienste :* An vielen Stellen entlang seines Weges wird der Benutzer über *aktuelle Veranstaltungshinweise* informiert. Er hat dann die Möglichkeit über das *Ticket Button* direkt Tickets zu reservieren. Außerdem wird Besuchern die Möglichkeit geboten interaktiv eine *Unterkunft in einem Hotel* zu buchen. An manchen Attraktionen wird der Benutzer auch aufgefordert an einer *Umfrage* zu diesem Ort teilzunehmen. Es wird derzeit über weitere Interaktive Dienste nachgedacht, z.b. das *Rufen eines Taxis*.

- *Messaging Service :* Außerdem kann der Benutzer über das *Message Button* jederzeit eine Nachricht an andere Benutzer und an das Tourismusbüro verschicken und umgekehrt natürlich auch Nachrichten von diesen empfangen. Dies geht, weil wie oben beschrieben bekannt ist in welcher Zelle sich welcher Benutzer gerade aufhält. Ausnahme: Benutzer die im Benutzerprofil angeben, dass sie *anonym* bleiben wollen können an diesem Service nicht teilnehmen.

- *Kommentare :* An manchen Stellen hat der Besucher die Möglichkeit ein Kommentar zu hinterlassen. Andere Besucher die später vorbeikommen können diese Kommentare dann lesen, eigene Kommentare abgeben oder via Messaging Service direkt mit dem Schreiber eines Kommentars in Verbindung treten.

Bemerkung: Bei *unterbrochener Netzwerkverbindung* funktionieren die interaktiven Dienste, das Messaging System und der Kommentarmodus nicht. Die zugehörigen Buttons können dann nicht mehr angeklickt werden. Dies ist eine weitere Maßnahme den Benutzer über den *Status des Netzwerkverbindung* zu unterrichten.

3.4 Erkenntnisse aus dem GUIDE Projekt

Verschiedene Untersuchungen haben ergeben, dass die große Mehrheit (über 95 %) der Besucher von Lancaster mit GUIDE sehr zufrieden ist. Die Möglichkeit zur Navigation bzw. Führung durch die Stadt, die flexible Art und Weise der Informationsbereitstellung und die Verfügbarkeit dynamischer Informationen wird als sehr positiv empfunden. Die meisten Benutzer, insbesondere auch solche ohne Weberfahrung, haben (nach fünfminütiger Einarbeitung) dank der einfachen webbasierten Oberfläche keine großen Probleme GUIDE zu bedienen. Etwas skeptisch stehen die Benutzer den Interaktiven Diensten gegenüber. Viele möchten z.b. beim Buchen einer Unterkunft lieber persönlich vor Ort sein. Kritisiert wird auch, dass durch GUIDE derzeit keine explizite Bestätigung für interaktiv abgeschlossene Geschäfte erfolgt.

Zwei wesentliche Erkenntnisse die während der Entwicklung und Erprobung des kontextsensitiven GUIDE Systems gemacht worden sind und die auch für den Entwurf anderer kontextbasierter Systeme relevant sind sollen hier noch angeführt werden.

Für die Akzeptanz kontextsensitiver Systeme ist es zum einen *wichtig , dass diese flexibel sind*. Verschiedene Benutzer handeln sehr unterschiedlich und haben unterschiedliche Wünsche und Absichten. Kontextsensitive Systeme müssen all diesen Anforderungen zugleich gerecht werden. Sie dürfen deshalb nicht zu eng auf bestimmte Verhaltensweisen ausgelegt sein, und "abstürzen" falls sich jemand unvorhergesehener Weise anders verhält.

Je mehr verschiedene Situationen nun aber berücksichtigt werden sollen, desto komplexer und aufwendiger wird logischerweise die Implementierung eines kontextsensitiven Systems. Tritt eine unvorhergesehene Situation ein und soll das System trotzdem weiterfunktionieren, muss *letztendlich immer der Benutzer eingreifen*. Bei Guide: Fällt unvorhergesehenerweise die Netzwerkverbindung aus, soll der Besucher dennoch zur nächsten Attraktion geleitet werden. Dies ist wie oben beschrieben nur möglich, indem das System dem Benutzer Fragen über die aktuelle Position stellt.

Zusammenfassend heißt das: Der Benutzer wünscht Flexibilität, sonst ist er unzufrieden mit dem System. Ein Teil dieser Flexibilität kann je nach getriebenem Aufwand sicher durch eine *intelligente Implementierung* des Systems gewährleistet werden. Der andere Teil muss aber *vom Benutzer selbst* getragen werden. Bei der Entwicklung eines kontextbasierten System muss man sich also genau überlegen wie flexibel es sein soll, und wie diese Flexibilität gewährleistet werden kann ohne die Benutzer zu überfordern.

Die zweite Erkenntnis betrifft den Einsatz in mobilen Umgebungen. Für die Akzeptanz, bzw. das Vertrauen dass der Benutzer einem mobilen kontextbasierten System entgegenbringt ist es wie bereits im Abschnitt 3.3 beschrieben wichtig, den Benutzer durch geeignete Maßnahmen über den Zustand der Netzwerkverbindung zu unterrichten. Untersuchungen bei GUIDE wiederum haben ergeben, dass die große Mehrheit der Benutzer sich darüber bewusst ist, dass GUIDE, ähnlich wie beim Mobilfunk, auf einem drahtlosen Netz mit wechselnder Empfangsstärke beruht. Die meisten Benutzer akzeptieren dies und können mit der benutzten Symbolik umgehen. Sie ziehen ein solches System einem System vor, bei dem sie nicht über die Empfangsqualität unterrichtet werden und daher nie wissen ob die Informationen noch gültig sind.

Die Intention dieser Arbeit war es zunächst, den Begriff Kontext genau zu definieren, und die typischen Merkmale und Eigenschaften kontextabhängiger Dienste allgemein, d.h. unabhängig von konkreten Anwendungen, abzugrenzen. Die so gewonnenen Erkenntnisse können nun dazu dienen, bereits existierende und zukünftige kontextbasierte Systeme besser zu verstehen und klassifizieren zu können.
GUIDE vereint zahlreiche der typischen Eigenschaften kontextbasierter Dienste und war deshalb besonders gut geeignet, die gewonnenen Erkenntnisse anhand eines bereits funktionierenden Systems zu veranschaulichen. Näher beschrieben wurden auch die automatische Benutzerprofilerkennung beim NOKIA 6110 sowie das Active Badge System.
Kontextbasierte Systeme werden heute meist in Form von Forschungsprojekten realisiert. Die Einführung solcher Systeme ist mit einem hohen technischen und finaziellen Aufwand verbunden. Ob sich solche Systeme in Zukunft auch in der Praxis stärker behaupten können hängt damit im wesentlichen von der Akzeptanz der Benutzer, d.h. auch deren Bereitschaft für solche Dienste zu zahlen, ab. Da kontextbasierte Systeme aber viele Vorteile für die Benutzer mit sich bringen, werden nach Meinung des Autors in Zukunft nach und nach immer mehr solcher Systeme am Markt erscheinen.

Literatur

[DUDE] DUDEN. *Deutsches Universalwörterbuch A-Z.* Begriff Kontext nachschlagen!

[Fuji] Fujitsu. *Fujitsu Teampad 7600 Technical Page.* http://www.fjicl.com.

[Gell] H.W. Gellersen. *There is more to Context than Location.* Environment Sensing Technologies for Adaptive Mobile User Interfaces.

[Lanc] Uni Lancaster. *The GUIDE PROJECT.* http://www.guide.lancs.ac.uk/overview.html.

[Schm] A. Schmidt. *Advanced Interaction in Context.* mail to : albrecht@teco.edu.

[Webs] Webseite. *The Active Badge System.* http://www.cam-orl.co.uk/ab.html.

Abbildungsverzeichnis

Tabellenverzeichnis

Middlewaresysteme in partitionierbaren Netzen

Stefan Sellschopp

Kurzfassung

In diesem Aufsatz wird der Jgroup Ansatz der Universität Bologna vorgestellt, der auf eine differenzierte Transparenz für die Teilnehmer in Netzen mit verteilten Anwendungen abzielt. Hervorzuheben ist, dass Jgroup partitionierbare Netze betrachtet und spezielle Tools entwickelt hat, die durch Partitionen entstehenden Komplikationen zu lösen. Es werden drei Komponenten beschrieben: Der Gruppenkommunikationsdienst, die Methode der Anfragestellungen und der Dienst zum Verschmelzen von Partitionen. Durch die Verwendung replizierter und über dem Netz verteilter Dienste wird eine höhere Leistungsfähigkeit (bezüglich Zuverlässigkeit und Verfügbarkeit) erreicht, als in den bisher eingesetzten Middleware Systemen wie z.B. CORBA. Im Anschluss an die Vorstellung des Jgroup Konzepts werden Vergleiche zu anderen verteilten Systemen gezogen und die Anstrengungen der Object Management Group für ein drahtloses CORBA beschrieben, das sich mit einer in Teilaspekten vergleichbaren Aufgabenstellung beschäftigt.

1 Einleitung

Es wird immer selbstverständlicher jederzeit und überall Zugang zu Daten und Anwendungen zu haben. Oft werden diese Dienste nicht am Ort des Anwenders bereitgestellt, sondern weit entfernt. Traditionelle dokumentenbasierte Systeme sind nicht gut geeignet, solche modernen Netzwerkdienste und Anwendungen auszuführen und bereitzustellen ([MoDB99]). Protokolle wie TCP/IP und nachrichtenbasierte Middleware Systeme (z. B Isis, PVM) haben diese Anwendungen in der Vergangenheit für Netzwerke geöffnet, bieten aber für die Anwendungsentwicklung ein unzureichendes Abstraktionsniveau. Objektorientierte Middleware Systeme wie CORBA (1991 in der ersten Version definiert [Grou01b]), haben die Leistungsfähigkeit seitdem deutlich erhöht ([MoDB99]). Hier war aber die Zuverlässigkeit und die Verfügbarkeit der Systeme, aufgrund der mangelnden multicast Kommunikation, nicht im ausreichenden Maße gegeben. Multicast Kommunikation ist jedoch die übliche Kommunikationsform beim Beantragen replizierter Dienste und bei der Erhaltung der Konsistenz von replizierten Diensten([MoDB99]). Diesen Problemen begegnet Jgroup mit dem Konzept der Replikation der Dienste und deren Verteilung über das Netz, wie in der Publikation *Middleware for Dependable Network Services in Partitionable Distributed System* ([MoDB99]) beschrieben wird, das Ausgangspunkt dieser Ausführungen ist. Es wird von Jgroup auf Teilgruppen (Partitionen) eingegangen, deren Server mit den Servern der jeweils anderen Partition nicht kommunizieren können. Diese Partitionen entstehen durch instabile Verbindungen, welche die Gruppe der Server in Untergruppen aufteilt. Diese

Teilgruppen bestehen aus Servern, die untereinander kommunizieren können und existieren so lange bis wieder eine Kommunikationsverbindung zwischen den Partitionen besteht. Erstmals wird vorgeschlagen nicht nur eine Hauptpartition, sondern mehrere gleichberechtigte Partitionen zu erlauben. Ein Beispielsnetz (ohne Partitionen) kann man in Abbildung 1 betrachten. Die Server sind vernetzt und bieten den angeschlossenen Clients Dienste transparent (ohne dass diese wissen, wo der Dienst erbracht wird) an.

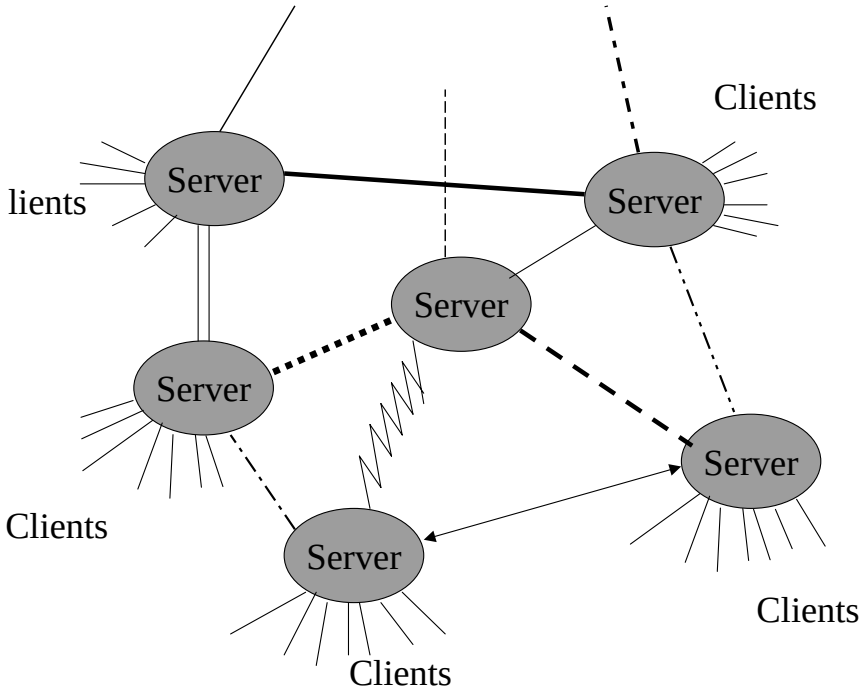

Abbildung 1: Dienste werden von verschiedenen Rechnern in einem verteilten System erbracht.

Es wird RMI (Java Remote Method Invocation) [Micr01] zur Kommunikation zwischen den Netzteilnehmern (Clients/Server) verwendet und so die Anfragen durchgeführt. Hier wird zwischen dem externen Aufruf (Dienstnehmer beantragt Service) und dem internen Aufruf (zur Abgleichung der internen Dienstanbieter/Server) unterschieden. Es wird ein Gruppenkommunikationsdienst eingeführt, der Wissen über die Teilnetze (PGMS: partition aware group membership service) verwendet und so einem weiteren Dienst, dem State-Merging-Service (SMS), erlaubt verschiedene Partitionen verschmelzen zu können. Dieser Fall tritt ein, sobald die Kommunikation zwischen den Partitionen wieder hergestellt ist.

2 Was ist Transparenz?

Eine wichtige Eigenschaft verteilter Systeme ist die der Transparenz. Üblicherweise ist die Transparenz für den Nutzer ein sehr wichtiges Merkmal bei der Nutzung von Applikationen in verteilten Rechnernetzen. Jgroup ermöglicht eine Differenzierung bei der Transparenz, die bei internen (eine Applikation ist Nutzer der Dienste des Netzes) und bei externen Aufrufen erzielt werden soll. Jede Applikation soll in der Lage sein den Zustand im Netz, und die eigene Fähigkeit einen Dienst anzubieten, einzuschätzen.

Als Konsequenz aus diesem Wissen entscheidet die Applikation, ob eine Anfrage erfüllt werden kann. Im Netz (intern) ist also keine vollständig Transparenz (Nebenläufigkeitstransparenz) gegeben. Um den Begriff Transparenz zu klären, werden anschließend kurz die wichtigsten Formen der Transparenz aufgezählt und beschrieben, wie sie in der Lehre aktuell behandelt werden ([Schl99], [Kräm99]) und für diese Arbeit von Relevanz sind.

- Ortstransparenz
  Der Benutzer ist sich des Ortes eines Objektes im Netz nicht bewußt; der Zugriff erfolgt über einen Namen, der keine Ortsinformation enthält;

- Zugriffstransparenz
  Auf alle Objekte wird in ein und derselben Weise zugegriffen;

- Replikationstransparenz
  Der Benutzer greift auf replizierte Objekte zu, als seien sie nur einmal vorhanden;

- Fehlertransparenz
  Fehler im Netz werden bis zu einem bestimmten Grad vom System maskiert;

- Nebenläufigkeitstransparenz
  Mehrere Benutzer oder Anwendungsprogramme können gleichzeitig auf gemeinsame Objekte (z.B. Daten) zugreifen ohne sich gegenseitig zu beeinflussen oder sich ihrer gegenseitigen Existenz im Rechner bewußt zu sein (fehlende *group awareness*);

- Prozeßtransparenz
  Der Ausführungsort eines Prozesses hat keinen Einfluß auf die Durchführbarkeit;

- Sprachtransparenz
  Die Interaktionen zwischen Teilkomponenten ist unabhängig von der Programmiersprache, die für die Implementierung der jeweiligen Teilkomponente benutzt wurde;

Die großen Vielfalt der Begriffe erfordert eine genaue Spezifizierung, wenn man sich auf Transparenz bezieht. In der Literatur wird dies jedoch nicht konsequent umgesetzt und es werden die einzelnen Transparenzen auch noch unterschiedlich, zum Teil mit leicht abweichenden Bedeutungen, bezeichnet (Sprachtransparenz - Netztransparenz, Zugriffstransparenz - Benutzertransparenz, Fehlertoleranz - Ausfalltoleranz, Ortstransparenz - Daten/Verarbeitungstransparenz). Es besteht daher die Gefahr von Missinterpretationen in der Argumentation, die in dieser Arbeit vermieden werden soll. Auf die Vorteile und Nachteile verschiedener Transparenzen wird in Kapitel 3.2 noch eingegangen.

3 Das Modell der verteilten Objekte

Jgroup verwendet Replikation in Form von Objektgruppen, um die Bereitstellung von zuverlässigen Anwendungen zu ermöglichen. Anwendungen und Daten werden auf verschiedenen Servern gespeichert und ausgeführt. Wenn Clients den Dienst beantragen,

sind sie sich der Replikation dieser Dienste nicht bewusst (Replikationstransparenz Kapitel 2) sondern sie erscheinen transparent. Wie in Abbildung 2 zu erkennen ist, werden von dem angefragten Server die weiteren Rechner der Gruppe in die Diensterbringung mit einbezogen und es werden ihnen Anfragen geschickt.

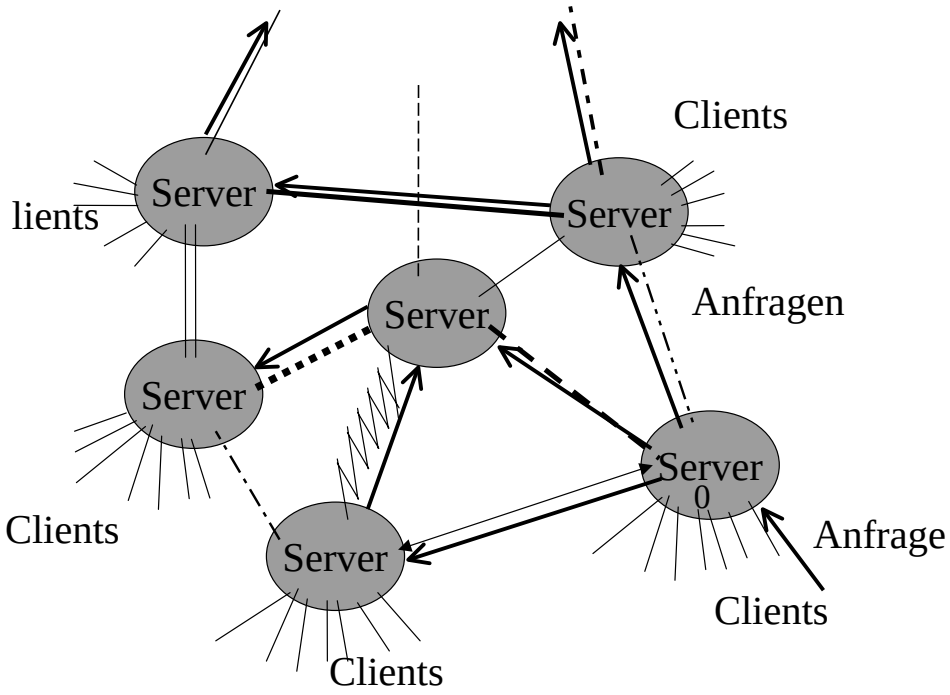

Abbildung 2: Von einem Client wird eine Anfrage an einen Server gestellt, der weitere Anfragen an das Netz stellt.

Für Clients ist bei dem Jgroup Modell zunächst kein Unterschied zu Anfragen mittels normalem RMI (Java Remote Method Invocation) [Micr01] erkennbar. Das Gruppenmodell ist jedoch insofern erweitert, als das auch partitionierbare Gruppen berücksichtigt werden. Dazu werden folgende drei weitere Systemkomponenten eingeführt, die in den nächsten Kapiteln ausführlicher beschrieben werden:

- Ein Gruppenkommunikationsdienst mit Einbeziehung des Wissens über Teilnetze (PGMS: partition aware group membership service),
- einem Gruppenaufrufdienst (GMIS: group method invocation service) und
- dem Dienst um Partitionen zu verschmelzen (SMS: state merging service).

3.1 Gruppenkommunikationsdienst mit Einbeziehung des Wissens über Teilnetze

Gruppen sind Ansammlungen von Servern, die gemeinsam verteilte Dienste anbieten. Da in verteilten Systemen oft Server ausscheiden (unerreichbar werden) oder neu hinzukommen (wieder erreichbar werden), besteht der Zwang für die Server immer zu wissen, mit welchen Servern sie aktuell zusammen arbeiten können. Eine solche Partitionierung

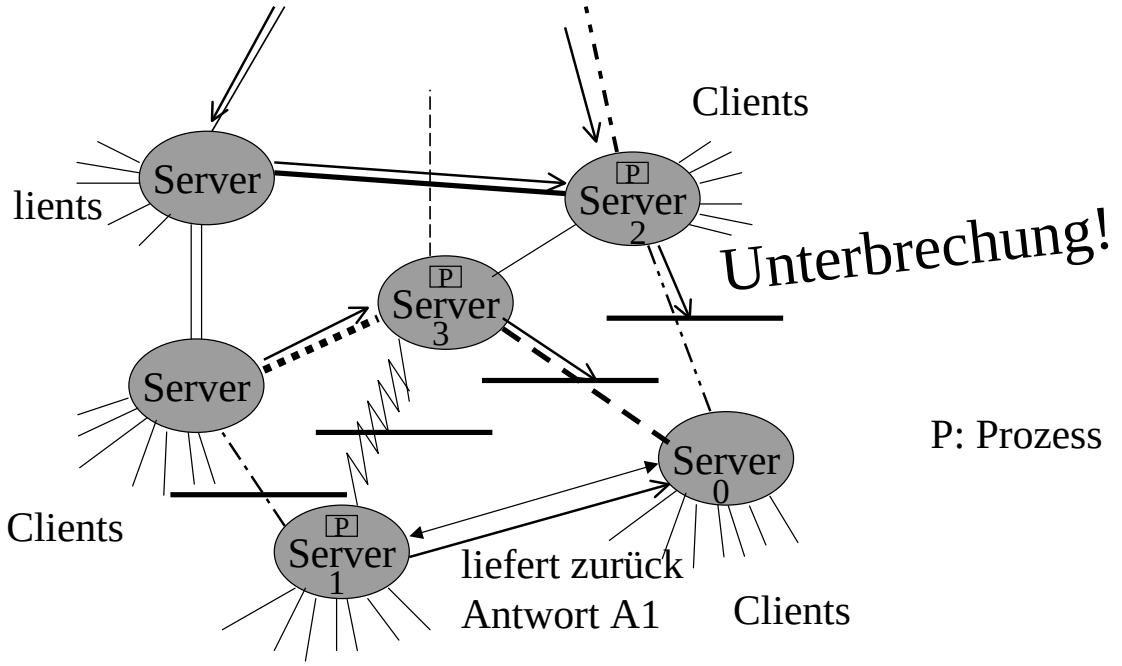

Abbildung 3: Nicht alle Antworten der Server können gesendet werden, wegen Unterbrechung (Unterteilung in Partitionen).

ist in Abbildung 3 abgebildet, in der unter anderem die angefragten Server 2 und 3 ihre Antworten nicht zurückgeben können. Wenn die Server nicht wüssten, mit wem sie zusammenarbeiten könnten, entstünden unterschiedliche Auffassungen, welche Systeme noch nicht freiwillig oder unfreiwillig aus der Gruppe bzw. Partition ausgeschieden sind. Der PGMS (Partition aware group membership service) verfolgt diese Ein- und Austritte und gewährleistet einen einheitlichen *View* der Server in jeder Partition. Ein View entspricht der Sicht des Servers auf seine Umgebung und beinhaltet alle Server, mit denen er kommunizieren kann. Jeder View wird durch eine Identifikation eindeutig gekennzeichnet. Der PGMS kommuniziert diesen View an die Mitgliedsserver in dem View, so dass alle den gleichen Wissensstand haben. Ein neuer View, der die Veränderung in der Gruppe widerspiegelt, kann nur installiert werden, wenn alle in dem View enthaltenen Server einig sind, welche Server erreichbar sind und welche nicht.

In diesem Entscheidungsprozess scheiden unter Umständen Server aus dem View aus, falls keine Einigung erzielt werden kann. Jgroup führt folgende formale Kriterien für die Views auf:

- View Accuracy
  Wenn ein Prozess für eine bestimmte Zeit von einem anderen erreichbar ist, gehört dieser dauerhaft in dessen aktuellen View.

- View Completeness
  Wenn alle Prozesse in einer bestimmten Partition p von dem Rest der Gruppe unerreichbar sind, haben alle diese anderen Prozesse (nicht in p), die Prozesse der Partition p nicht im aktuellen View.

- View Coherency
  Wenn ein Prozess einen neuen View installiert, müssen alle in diesem View enthaltenen Prozesse ebenfalls diesen View übernehmen oder der Prozess installiert

einen Nachfolge-View. Alle der von einem View in den nächsten View überlebenden Prozesse müssen ebenfalls vorher den alten View installiert haben.

- View Order
  Alle Views werden, im Falle der Installation, von allen Prozessen in der gleichen Reihenfolge installiert.

- View Integrity
  Jeder View der von einem Prozess installiert wird, beinhaltet den Prozess selber.

In älteren Publikationen ([BDMS97]) spricht Jgroup statt von View Coherency von dem intuitiveren View Agreement, das besagt, dass wenn ein Prozess p einen View v und dessen Nachfolger w installiert die beide Prozess q enthalten, Prozess p den Nachfolgeview w erst installiert, nachdem q den View v übernommen hat. Außerdem muss ein Prozess einen neuen View installieren, wenn ein Prozess aus dem alten View dauerhaft unerreichbar wird.

3.2 Gruppenanfragedienst

Es gibt verschiedene Arten der Kommunikation in einem verteilten System: Clients stellen Anfragen an die Gruppe von Servern und Server kommunizieren mit anderen Servern. Clients nehmen Dienste in Anspruch ohne sich darüber im klaren zu sein, dass sie tatsächlich mehrere Rechner ansprechen (Replikationstransparenz) und ob Fehler in Teilbereichen des Netzes auftreten (Fehlertransparenz). Dazu muss eine Funktionalität bereitgestellt werden. Die Server hingegen müssen miteinander kommunizieren, um die Zusammensetzung der Gruppe zu klären. Wenn Server zum Zwecke der Organisation der Dienste kommunizieren, wird von interner Kommunikation gesprochen. Anfragen von Dienstnehmern (Clients) an die Servergruppe heißen externe Kommunikation. Das besondere an dem Konzept von Jgroup ist, dass sowohl für die interne als auch für die externe Kommunikation einheitliche Gruppenanfragen (group method invocations) verwendet werden. Auf diesem Wege werden die Vorteile der Objektorientierung auch für die interne Kommunikation genutzt. Trotzdem wird zwischen den internen und den externen Anfragen unterschieden, was folgende Vorteile mit sich bringt ([MoDB99]):

- Sichtbarkeit
  Methoden der Diensterbringung und replizierte Dienste, sollten für den Anfrager nicht sichtbar sein (Visibility). Nur eine öffentliche Schnittstelle steht zur Verfügung, während die Ausführung dem Server vorbehalten bleibt.

- Transparenz
  Das Ziel von Jgroup ist, dass Anfrager nicht merken sollen, dass sie einen Aufruf bei einer Gruppe von Servern tätigen und sie sollen keinen Unterschied zu normalem RMI bemerken.

- Effizienz
  Wenn die selben Spezifikationen für interne und externe Anfragen umgesetzt worden wäre, hätten Clients der Gruppe beitreten müssen. So wäre aufgrund der höheren Anzahl der Clients die Skalierbarkeit und die Kosteneffizienz geopfert worden. Die leicht schwächere Semantik der externen Anfragen, bewirkt, dass der Großteil (viel weniger Server als Clients!) der Anfragen billiger ist.

Hier wird deutlich, dass bei den externen Anfragen auf Transparenz Wert gelegt wird, um eine hohe Benutzerfreundlichkeit zu gewährleisten. Der Benutzer muss nicht überlegen, wie der Dienst erbracht wird. Im Gegensatz dazu ist bei der internen Kommunikation für den anfragenden Server wichtig, welcher andere Server erreichbar ist und welche Dienste dieser erbringen kann. Hier ist das Wissen über die Erbringung ein entscheidender Baustein in der Fähigkeit des beschriebenen Gruppenkommunikationsdiensts PGMS (3.1), die Dienste bereitzustellen. Unter Benutzung der in Kapitel 2 vorgestellten Definitionen, spricht man von selektiver Transparenz. Orts- und Zugriffstransparenz werden unterstützt, wohingegen keine strikte Nebenläufigkeitstransparenz umgesetzt wird. Eine entsprechende Aussage findet sich in der Argumentation der TU München [Schl99] über die Anwendung von selektiver Transparenz für CSCW (Computer supported Cooperated Work).

3.2.1 Interner Gruppenkommunikationsdienst

Im Gegensatz zur traditionellen Java Remote Method Invocation, liefert der interne Gruppenkommunikationsdienst IGMI (Internal Group Method Invocations) ein Array mit den Antworten zurück. Bei den internen Anfragen wird noch einmal unterschieden zwischen synchronen und asynchronen Anfragen. Nach einer *synchronen Anfrage* bleibt der anfragende Server so lange blockiert, bis alle angefragten Rechner geantwortet haben. Alle Antworten werden dann in dem oben erwähnten Array zurückgeliefert. Diese Art der Blockade ist für viele Nutzungsszenarios zu teuer, da ein Server erst wieder freigegeben wird, wenn der letzte Rechner geantwortet hat. Zudem müssen hier die Programmierer auf die Gefahr eines Deadlock achten, der bei gegenseitiger Anfrage entstehen kann. Im Falle einer *asynchronen Anfrage* wird vom Anfrager ein Callback Objekt spezifiziert. Dieses wartet auf die Benachrichtigungen, dass die angefragten Server fertig sind. Im Falle einer leeren Rückgabe, kann der Anfrager auswählen, ob er erfahren will, wann die Anfrage beendet war oder dass er nicht daran interessiert ist. Die Beantwortung einer Anfrage durch eine Gruppe erfüllt die *View Synchrony* Eigenschaft, die verlangt, dass zwei Server die das gleiche Paar aufeinanderfolgender Views installieren, die selbe IGMI im ersten der zwei Views beantwortet. Diese Eigenschaft hat sich als ein wichtiges Kriterium erwiesen, wenn über Zuverlässigkeit in nachrichtenbasierten Systemen argumentiert wird ([MoDB99], [ScRi93]). Eine Eigenschaft von IGMI in Verbindung mit View Synchrony ist, dass jede Anfrage genau in einem View bearbeitet wird, was entscheidend bezüglich der Konsistenz ist. Details können in ([MoDB99]) nachgelesen werden.

3.2.2 Externer Gruppenkommunikationsdienst

Externe Anfragen von Clients an die Servergruppe sind vollkommen transparent für die Anfrager, die Dienste nutzen als seien sie normale RMIs. Wenn ein Entwickler einen Dienst entwickelt, muss er zwischen zwei Anfragetypen wählen:

- Anycast Anfrage
  Eine anycast Anfrage muss mindestens von einem Server beantwortet werden, solange zumindest ein Server in der Partition des Anfragers operativ ist. Dieser Anfragetyp ist für Anfragen geeignet, die den replizierten Serverzustand nicht verändern, wie zum Beispiel bei Informationsanfragen auf einer Datenbank.

- Multicast Anfrage
  Alle Server in der Partition des Anfragers bearbeiten die Anfrage. Dieser Anfragetyp ist geeignet für Operationen, die möglicherweise den Zustand des Servers verändern.

Zusammenfassend wird hier das Prinzip *read once - write all* umgesetzt, um eine gute Konsistenz zu erreichen und die Verfügbarkeit so hoch wie möglich zu halten.

3.3 Dienst um Partitionen zu verschmelzen

In bisherigen Lösungsansätzen für zuverlässige verteilte Systeme gibt es eine *Hauptpartition*, die bei einer Verschmelzung von Partitionen bestimmt, welcher Zustand als korrekt angesehen wird. Jgroup versucht die parallele Existenz mehrerer gleichwertiger Partitionen zu erlauben, um Ergebnisse der Rechner in kleineren Partitionen nicht zu verwerfen und so einen höheren Servicegrad (eine höhere Anzahl Anfragen kann beantwortet werden) zu bieten. Durch verschiedene Partitionen und deren Verschmelzung kann es jedoch zu Schwierigkeiten kommen. Diese Probleme rühren aus dem Verfügbarkeit - Konsistenz Konflikt her und sind unabhängig von den Annahmen die bei Jgroup getroffen wurden. Der Dienst zur Verschmelzung von Partitionen (SMS: State merging service), muss einen neuen Zustand im Server (server state) erzeugen, nachdem Kommunikationsstörungen behoben sind, und dafür soweit wie möglich die differierenden Partitionen in Einklang bringen, wie in Abbildung 4 zu erkennen ist. In der Abbildung werden von verschiedenen Servern Antworten aus unterschiedlichen Partitionen an den Server 0 zurückgegeben, der von dem Client angefragt worden war.

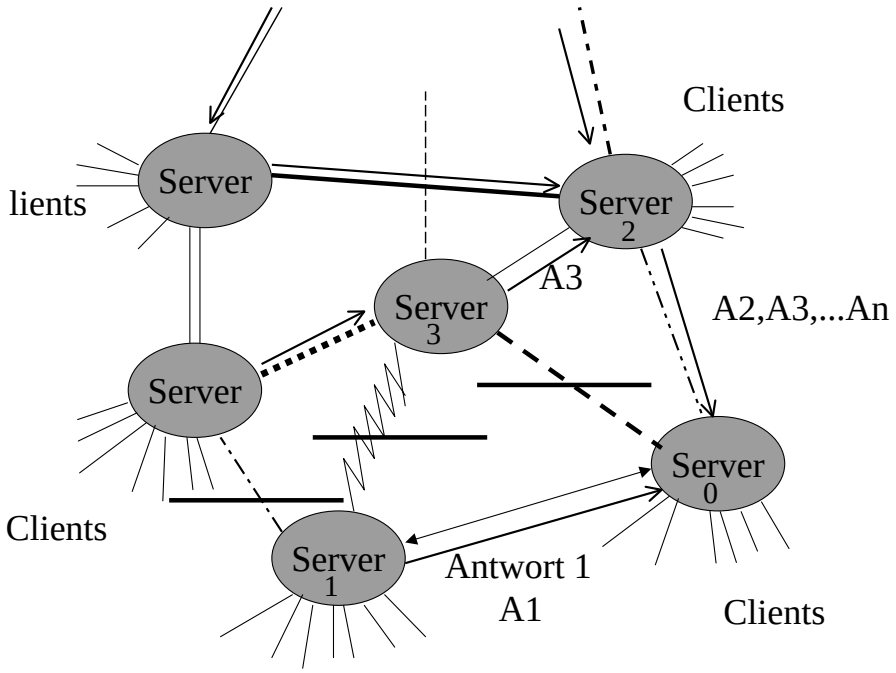

Abbildung 4: Es besteht wieder Verbindung zwischen den Partitionen und die Kommunikation wird wieder aufgenommen.

Alle nicht in konfliktstehenden Updates werden in den neuen Zustand übernommen und Konflikte erkannt. Eine völlige Automatisierung der Konfliktbehandlung ist nicht mög-

lich, jedoch bietet der SMS Unterstützung bei der Generierung von anwendungsbezogenen Protokollen zur Ausnahmefallbehandlung. Ein kompletter Informationsaustausch ist der Grundstein von SMS: Von den Servern werden in jeder Partition *Koordinatoren* ausgesucht, die denjenigen Servern, die bisher in einer anderen Partition waren, die Änderungen zusenden. Diese Informationen werden dann von diesen in den jeweiligen lokalen Serverzustand übernommen. SMS regelt den Nachrichtenaustausch, die Wahl der Koordinatoren und die Verteilung der Information, wie schematisch in Abbildung 5 zu erkennen ist. Alle Informationen werden in der jeweils anderen Partition ausgetauscht und die anschließende Aktualisierung der Zustände von SMS unterstützt.

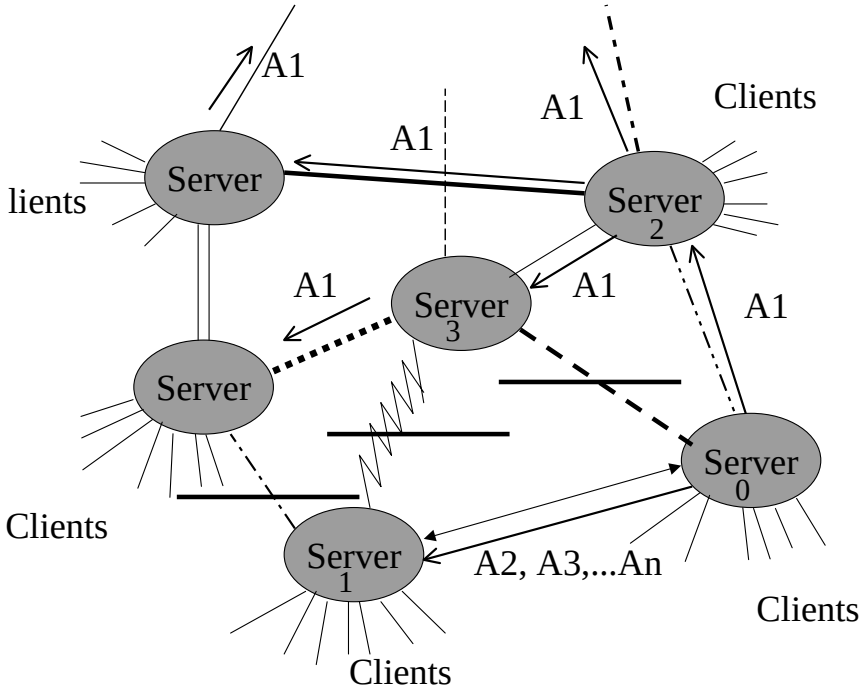

Abbildung 5: Der State Merging Service (SMS) verteilt die Informationen in den Partitionen und bietet Hilfe zur Konfliktlösung.

Voraussetzung eines Servers für die Teilnahme ist, dass der Server selbst Koordinator sein kann und dass er alle einkommenden Updates auf seinen lokalen Serverzustand (Konfiguration) anwenden kann. Dies kann die Einsetzbarkeit des SMS einschränken, da bei konsistenzkritischen Anwendungen eine Installation von Updates zu Konflikten führen kann und SMS daher nicht benutzt werden kann. Am Ende eines Anfrageprozesses und den oben beschriebenen netzinternen Vorgängen, steht die Rückgabe der Antwort an den Client. Er hat den Dienst transparent genutzt und von den Netzinterna nichts mitbekommen, wie in Abbildung 6 verdeutlicht wird.

4 Jgroup Implementation

Services für Clients (externe Anfragen) werden über sogenannte Stubs angefordert, die als eine Art Proxy fungieren und die Anfrage weiterleiten. Für die internen Anfragen dienen die Gruppenmanager (genau einer für jedes Serverobjekt). Der Stub ist alleine

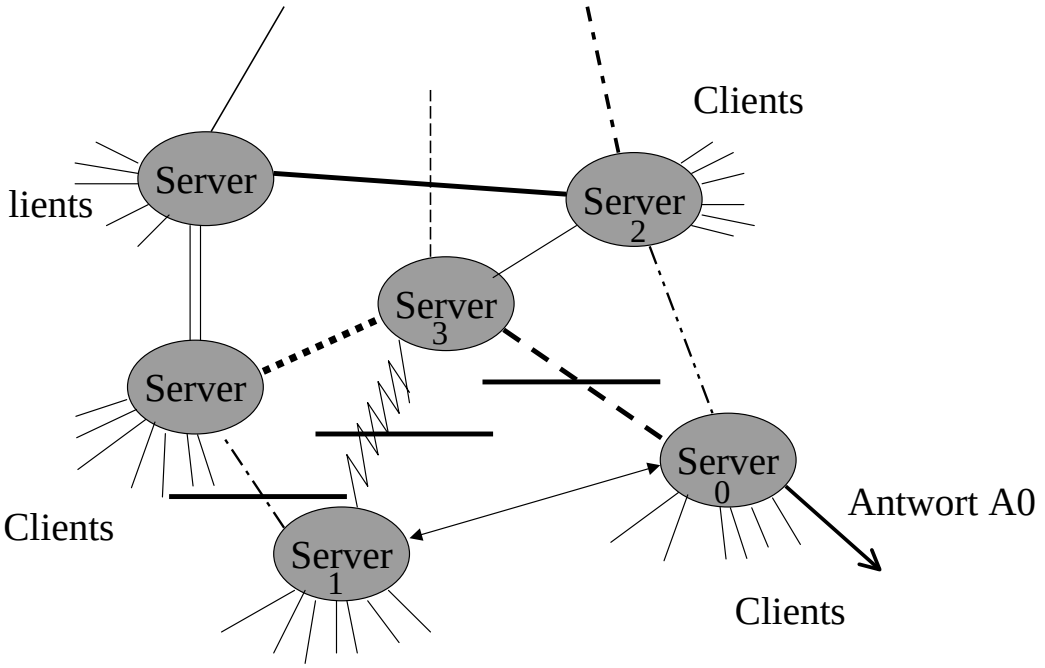

Abbildung 6: Der Client bekommt die Antwort von dem angefragten Server (transparent).

verantwortlich für die Behandlung von externen Anfragen. Die Gruppenmanager implementieren nur einen Teil der Funktionen der Organisation der Gruppen. Einen wichtigen Teil dieser Dienste, wie die Gruppenzugehörigkeit und Multicast-Kommunikation erfüllt ein weiteres Modul: Der Jgroup Daemon. In jeder Java Virtual Machine (JVM) wird ein Daemon ausgeführt. Durch den Daemon wird erreicht, die Anzahl der nötigen Nachrichten in der Gruppenkommunikation zu reduzieren. Basisdienste wie die Fehlererkennung werden nur einmal pro JVM ausgeführt und es erlaubt eine Unterscheidung zwischen den für die JVM lokalen und entfernten Objekten. Dies ist wichtig bezüglich der Verwaltung von Mitgliedslisten der Gruppen im Falle einer Kommunikationsunterbrechung. Der Jgroup Daemon ist aus Schichten aufgebaut. Auf der obersten Schicht des Daemons wird ein zuverlässiger nachrichtenbasierter Multicast-Dienst angeboten. Die Gruppenmanager sind, genau wie der Daemon, schichtenförmig und jeweils den Anforderungen der Server entsprechend aufgebaut. Die bisher verfügbaren Komponenten sind die schon angesprochenen Gruppenanfragedienste, der Dienst zum Verschmelzen von Partitionen und verschiedene Ordnungseigenschaften (ermöglichen Abarbeitung der Anfragen in der richtigen Reihenfolge bzw. im richtigen Zeitfenster im Bezug auf andere Partitionen). Weitere Komponenten können leicht ergänzt werden, um Entwicklern neue Möglichkeiten zu geben ([MoDB99]). Der Stub an dem die externe Anfrage gestellt wird, gibt der Anfrage eine Identifikationsnummer, mit der Replikate und doppelte Ausführungen erkannt werden können. Diese Unterscheidung ist besonders wichtig, wenn Anfragen in verschiedenen Partitionen bearbeitet werden und diese dann verschmelzen oder ihre Ergebnisse zurückgeben. Diese ID besteht aus der IP Adresse des Hosts (auf dem die JVM ist), einer *incarnation* Nummer, um verschiedene JVM auf einem Server zu differenzieren und aus einer fortlaufenden Nummer. Der SMS kontrolliert, dass alle Server bezüglich der anderen auf dem aktuellen Stand sind und führt, im Falle dass dies nicht zutrifft, ein Protokoll zum Abgleich der Serverzustände (state reconciliation protocol) aus. Normalerweise funktioniert der SMS, indem

Koordinatoren ausgesucht werden und diese den Servern in der jeweils anderen Gruppe die Ergebnisse der abgearbeiteten Anfragen mitteilen. So wird der gleiche Serverzustand in allen Servern erzeugt und erhalten. Details können in ([MoDB99]) nachgelesen werden. Um die Zweckmäßigkeit des Konzepts zu demonstrieren wurde ein *zuverlässiger Registrierungsservice* als Ersatz des standardmäßigen Registrierdiensts der Java RMI erstellt und in ([MoDB99]) vorgestellt. Ein Punkt der in der Ausarbeitung des Konzepts von der Jgroup vernachlässigt wurde, ist der Aspekt der Sicherheit. Bei dem Wiedereintritt von Servern in die Gruppe, muss die Authentizität geprüft werden, um Missbrauch zu verhindern.

5 Verwandte Arbeiten

Jgroup betont in [MoDB99], warum zum Beispiel der Object Transaction Service (OTS) von CORBA nicht zum Aufbau zuverlässiger verteilter Systeme geeignet ist: Die fehlende einer-an-viele (multicast) Kommunikation schließt aus, dass verlässliche Systeme auch gleichzeitig eine ausreichende Verfügbarkeit erreichen. Es werden drei Wege CORBA zu modifizieren erwähnt: Der *Integration Approach*, der den Object Request Broker modifiziert und erweitert. Anfragen werden über das Gruppenkommunikationssystem per multicast an die replizierten Server geschickt. Der *Interception Approach* beruht darauf, dass Nachrichten auf niedriger Ebene mitgehört oder abgefangen werden und dann an den Gruppen Kommunikationsdienst weitergeleitet werden. Oder man kann Gruppenkommunikation auch als weiteren Dienst anbieten, was im *Service Approach* getan wird. Dieser Weg wurde oft gewählt und beeinflusste die Antworten auf ein *Request for Proposals* von OMG (Object Management Group)[t1398] mit dem Ziel eines fehlertoleranten CORBA (kurz erwähnt in Kapitel 6.1). In Java existiert der Ansatz Filterfresh [BCHR+98], der die selben Ziele hat wie Jgroup. Es ist aber ein weniger weitreichender Vorschlag der weder einen externen Funktionsaufruf mit Multicast oder einen internen Methoden Aufruf kennt. Javagroups [Ban98] ist ein Toolkit, dass zuverlässige multicast-Kommunikation erlaubt. Der Methodenaufruf ist nicht transparent. Die iBus [Maff97] Architektur (ebenfalls in Java) ist ein kommerzielles Produkt, das die Anwendungen fehlertolerant machen soll. Es verbindet nicht die Gruppenkommunikation mit Java RMI, sondern verwendet das Konzept von mehreren Kanälen über die Informationen verteilt werden (IP multicast group). Dienstnehmer abonnieren einen Kanal und es besteht keine Sicherheit, dass eine Pull-Operation (Anforderung neuer Information) synchron mit einer Veränderung der Situation in der Gruppe geschehen.

6 CORBA Modifikationen

6.1 Fault tolerant CORBA

Die Object Management Group hat Versuche unternommen, CORBA fehlertolerant zu machen. Im Dezember 2000 wurde dazu der Bericht der Finalization Task Force (ptc/2000-12-06) eingereicht und hier soll kurz die Kernidee präsentiert werden um den Vergleich zu dem Jgroup Ansatz zu ermöglichen. In geographischen Regionen werden jeweils Applikationen repliziert und so die Verfügbarkeit erhöht. Das Ziel ist, keinen

single point of failure im System zu haben. Die angesprochenen Regionen (vergleiche Abbildung 7) werden lokal konsistent gehalten, da zu große Netze, nach Ansicht der diesen Vorschlag ursprünglich unterstützenden Firmen, zu große Schwierigkeiten machen würden. Eine Instanz eines *Replication Managers* je Domain, (auch sie liegen repli-

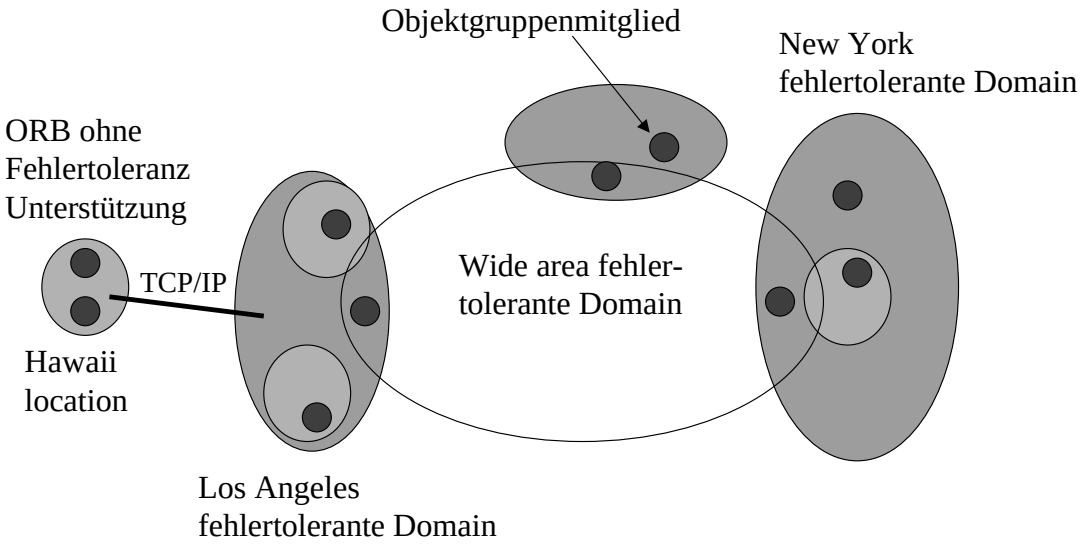

Abbildung 7: Der Client bekommt die Antwort von dem angefragten Server (transparent).

ziert vor) sorgt für den Abgleich der Serverzustände. Er besteht aus dem Object-group Manager und dem Property Manager dessen Kriterien vom Nutzer spezifiziert werden. Der Objektgruppen manager ist verantwortlich für neue Gruppenobjekte oder deren Entfernung. Details können in [Grou00] nachgelesen werden.

6.2 OMG RFP: Wireless Access and Terminal Mobility

In den Arbeiten von Jgroup [MoDB99][BDMS97] wird nicht explizit von drahtlosen Verbindungen gesprochen. Jedoch läßt die als Annahme vorausgesetzte hohe Ausfallwahrscheinlichkeit eine Verwendung auch für solche Zwecke als wahrscheinlich erscheinen. Die schon erwähnte Object Management Group hat im Mai 1999 ein Request for Proposal (RFP telecom/99-05-05) veröffentlicht, mit der Aufforderung Ideen und Umsetzungsvorschläge für ein drahtloses CORBA einzureichen. Die Zielsetzungen waren sehr ähnlich, wie bei dem beschriebenen Jgroup Ansatz: Es soll ein System von Servern ermöglicht werden, die zusammen Dienste anbieten, die jedoch auf Grund ihrer Anbindung (drahtlos, mobil) nicht mit stabilen Verbindungen arbeiten können. Es wird nicht explizit auf die von Jgroup avisierte hohe Verfügbarkeit eingegangen, die aber durch ein vergleichbares Konzept der replizierten Applikationen umgesetzt werden müßte. Die Abgleichung der Serverzustände wäre auch hier entscheidend für die Konsistenz der Server. Zu diesem RFP gingen bis Mai 2000 (dem Einsendeschluss) zwei Antworten ein: Nokia und Vertel führen ein Konsortium [Raat00], das eine sehr detaillierte Ausarbeitung eingereicht hat. Der weitere Vorschlag wurde von Inprise und Highlander eingereicht [JiCu00]. Die beiden Vorschläge unterscheiden sich, wie nachfolgend erklärt, fundamental.

6.3 Inprises Vorschlag für drahtloses Corba

Inprise hält das existierende CORBA mit TCP Verbindungen für am besten geeignet und zitiert in seiner Argumentation immer wieder die Ausschreibung von OMG. Es wird zum Beispiel betont, dass TCP höhere Paketverluste als bisher bei Festnetzübertragungen auftreten. Inprise schließt daraus, dass TCP eine ausreichende Fehlerkorrektur bietet, um auch die im Vergleich zum Festnetz bei drahtlosen Anwendungen höheren Fehlerraten zu kompensieren und auch kurzzeitige Abbrüche das System nicht instabil werden lassen [JiCu00]. Hier wird deutlich, dass die Fehlertoleranz relativ ist und dem Anfrager unter Umständen Wartezeiten (z.B. wenn der mobile Server gerade in einem Tunnel ist) zugemutet werden. Es muss TCP nicht explizit bekannt sein, dass einzelne Server mobil sind, da die Fragen der Verbindungen (terminal attachment) heute bei Verwendung von Mobil IP auf der Netzwerkschicht oder bei Mobilfunk auf der Verbindungsschicht (link layer) geregelt werden. Somit ist die Anbindung für CORBA transparent und muss nicht separat gelöst werden, zumal andere Anwendungen (Web, RMI, EJB(Enterprise Java Beans)) ähnliche Probleme haben und Sonderlösungen nicht wünschenswert sind.

6.4 Nokias Vorschlag für drahtloses CORBA

Im Gegensatz zu dieser Meinung vertritt Nokia ein neues Konzept, das sich an dem Aufbau eines Mobilfunknetzes orientiert. Als erster grundsätzlicher Unterschied ist die Wahl von UDP als Kommunikationsprotokoll zu erkennen. Im Gegensatz zu TCP ist UDP datagramm- und nicht verbindungsorientiert, so dass die Zuverlässigkeit und die übrigen Funktionalitäten von einer höheren Schicht übernommen werden müssen. TCP wird als nicht zuverlässig und ungeeignet für Terminal Mobilität bezeichnet und dazu auf verschiedene Arbeiten [telecom/98-11-09] verwiesen, in denen auch bei Verwendung von Mobile IP unzulängliche Ergebnisse erzielt werden. Da so auch TCP eine weitere Schicht benötigt hätte, um die Verbindungen zuverlässig zu machen, wurde UDP gewählt, so dass nur auf einer Schicht die Sicherung stattfindet. Jeder einzelne Server wird als ORB Domäne behandelt, die durch WABs (Wireless Access Bridges) an die anderen Domänen angebunden werden. Der ganze Aufbau erinnert an ein Mobilfunknetz, bei dem jeder mobile Rechner über ein Home Location Register (HLR) verfügt. Temporär können fremde WABs heimatlosen mobilen Rechnern eine HLR zur Verfügung stellen und ermöglichen so erst die Mobilität. Die WABs entsprechen den Basisstationen (Infrastruktur!) und das System benutzt als Kern ein fest verdrahtetes Netz an das sich die mobilen Rechner über die WABs anschließen. Dies unterscheidet diesen Ansatz sehr von dem Jgroup Konzept, welches auch verschiedene Partitionen und nicht nur eine kennt. Andererseits kann die Annahme, dass ein Teil des Netzes relativ stabil verfügbar ist (wie das feste Netz im Vorschlag Nokias) als realistisch bezeichnet werden. So kann mit großer Wahrscheinlichkeit ein hohe Servicequalität erreicht werden (Kommentar des Autors). Falls der Ansatz von Nokia als zu infrastrukturintensiv angesehen würde und es sich nicht als praktikabel erweisen sollte alles beim Status quo zu belassen, könnte sich der Ansatz von Jgroup ein wichtiger Diskussionsbeitrag erweisen.

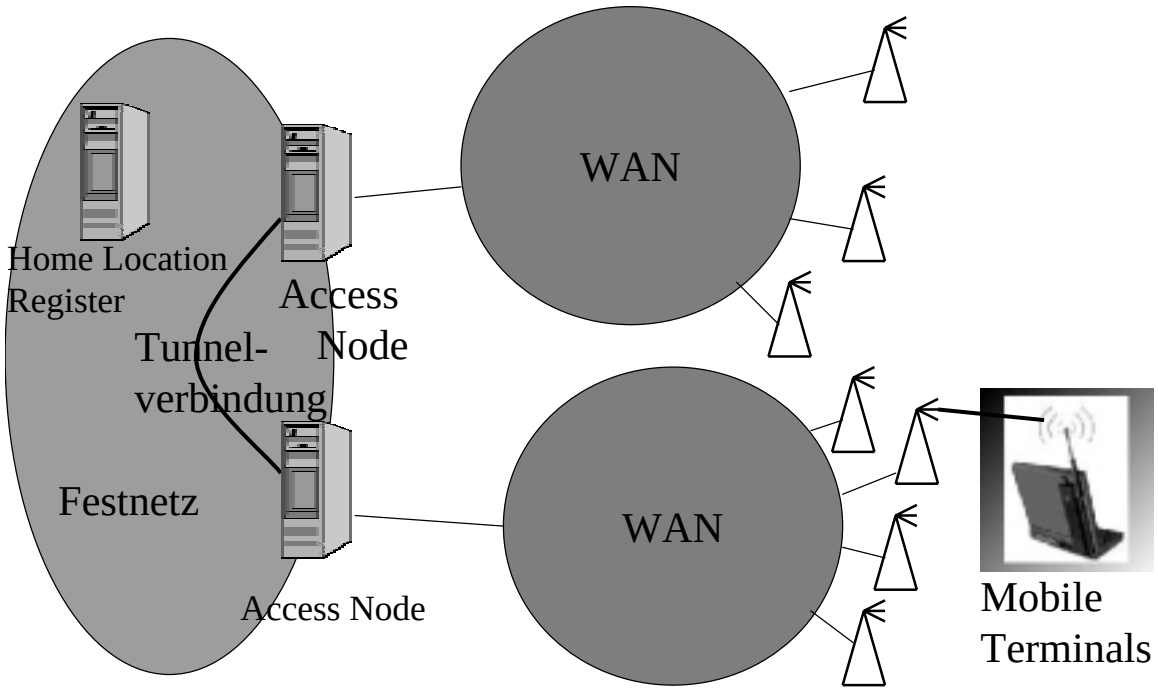

Abbildung 8: Der Client bekommt die Antwort von dem angefragten Server (transparent).

7 Zusammenfassung

Jgroup hat einen interessanten Ansatz zur Organisation von verteilten Systemen aufgezeigt. Applikationen wissen aufgrund eines teilgruppensensitiven Kommunikationsdiensts, welche Rechner im Netz erreichbar sind und entscheiden so, welche Anwendungen zur Verfügung stehen. Die Bearbeitung von Anfragen läuft auf allen erreichbaren Rechnern, die sich unter Umständen zeitweise in verschiedenen Partitionen befinden. Die Partitionen werden, nach der Wiederherstellung der Kommunikationsverbindungen zwischen den Partitionen, mit Hilfe der *View* Eigenschaft verschmolzen. Hierzu löst der Dienst zum Verschmelzen von Partitionen aus dem lokalem Wissen heraus Konflikte auf, die in den parallelen Partitionen aufgrund von unterschiedlichen abgearbeiteten Prozessen entstanden sind. Der skizzierte Vorschlag von Nokia für ein drahtloses CORBA beruht auf dem Konzept eines festverdrahteten und zuverlässigen Kernnetzes, an das sich die Server und Clients anschließen können und das die fehleranfälligen drahtlosen Übertragungen minimiert. Eine hohe Verfügbarkeit von Anwendungen müßte auch hier durch Replikation erreicht werden. Dieser Ansatz könnte aufgrund der zu installierenden Infrastruktur sowohl zuverlässiger aber auch teurer in der Umsetzung sein
.

Literatur

[BaDM98] Ö. Babaoglu, R. Davoli und A. Montresor. Group Communication in Partitionable Systems: Specification and Algorithms. *Technical Report UBLCS-98-01*, April 1998.

[Ban98] Bela Ban. JavaGroups - Group Communication Patterns in Java. Technischer Bericht, Cornell University, July 1998.

[BCHR+98] A. Baratloo, P. Chung, Y. Huang, S. Rangarajan und S. Yainik (Hrsg.). Filterfresh: Hot Replication of Java RMI Server Objects. Proc. of the 4th conf. on object-oriented technologies and systems, April 1998.

[BDMS97] Ö. Babaoglu, R. Davoli, A. Montresor und R. Segala. System Support for Partition-Aware Network Applications. *Technical Report UBLCS-97-08*, Oktober 1997.

[dtDi00] d-tec Distributed Technologies GmbH. http://www.d-tec.ch, 2000.

[Grou99] Object Management Group. Request for Proposal;Wireless Access and Terminal Mobility. Omg rfp telecom/99-05-05, Mai 1999.

[Grou00] Object Management Group. Fault Tolerance updated revised submission. orbos/00-01-19, Marz 2000.

[Grou01a] Object Management Group. http://www.omg.org, 2001.

[Grou01b] Object Management Group. http://www.corba.com, 2001.

[JiCu00] Ke Jin und Jon Currey (Hrsg.). Response to Request for Proposal;Wireless Access and Terminal Mobility. Omg rfp telecom/99-05-05 submission, Inprise Corporation and Highlander Engineering, Mai 2000.

[Kräm99] Krämer. Prof. http://www.et-online.fernuni-hagen.de/lehre/k02521/ke/ke01/node24.html, 1999.

[Maff97] S. Maffeis. iBus - Java Intranet Software Bus. Technischer Bericht, Olsen and Associates, April 1997.

[Micr01] Sun Microsystems. http://www.sun.com, 2001.

[MoDB99] A. Montresor, R. Davoli und Ö. Babaoglu. System Support for Partition-Aware Network Applications. *Technical Report UBLCS-99-19*, Oktober 1999.

[Mont00] Alberto Montresor. Jgroup Tutorial and Programmer's Manual. *Technical Report UBLCS-00-13*, Oktober 2000.

[Orac00] Oracle. http://oradoc.photo.net/ora81/DOC/server.815/a67784/dsch1.htm, 2000.

[Raat00] Kimmo Raatikainen (Hrsg.). Response to Request for Proposal;Wireless Access and Terminal Mobility. Omg rfp telecom/99-05-05 submission, Nokia and Vertel, Mai 2000.

[Schl99] J. Schlichter. Prof. http://wwwschlichter.informatik.tu-muenchen.de/lehre/lectures/ss2000/va/extension/html-kurz/va-2000-full3.2.html, 1999.

[ScRi93] A. Schiper und A. Ricciardi. Virtually-synchronous Communication Based on Weak Failure Suspector. Proc. of the 23rd Int. Symp. on Fault-Tolerant Computing, June 1993.

[t1398] Request for Proposal;Fault Tolerant CORBA Using Entity Redundancy. Omg rfp 98-04-01, Object Management Group, April 1998.

Abbildungsverzeichnis

Mensch Maschine Interaktion im Bereich Mobiler Endgeräte

Jürgen Schäck

Kurzfassung

Der Einsatz mobiler Geräte lässt die Verwendung verschiedener Anwendungen an jedem Ort und zu jeder Zeit zu. Voraussetzungen sind jedoch eine hohe Leistungsfähigkeit und kleine Ausmaße der Geräte. Da einige Ein-/Ausgabemechanismen nicht beliebig verkleinerbar sind, ist der Einsatz neuer und angepasster Mechanismen notwendig. Dieser Artikel beschreibt zunächst einige dieser Mechanismen. Er stellt Verfahren zur Beurteilung vor und vergleicht ausgewählte Eingabemechanismen.

1 Einleitung

Der große Erfolg des Internets hat auf die Gesellschaft weitreichende Auswirkungen. Die Abwicklung von alltäglichen Aufgaben mit Hilfe des Computers ist mittlerweile für viele Menschen eine selbstverständliche Angelegenheit. E-Mail, Online-Banking und Online-Shopping erfreuen sich einer großen Beliebtheit. Diese Aktivitäten waren aber lange Zeit an den heimischen PC mit Internetverbindung gebunden. Diese Einschränkung wird nicht mehr lange existieren. Der Benutzer kann mit Hilfe mobiler Geräte an jedem Ort und zu jeder Zeit seine Tätigkeiten durchführen. Diese Tätigkeiten umfassen nicht nur die oben genannten, wie Online-Banking usw, sondern auch eine Reihe neuer Anwendungen, die erst durch die neu gewonnene Mobilität ermöglicht werden. Dadurch entstehen neue Geschäftsmodelle. Drei Faktoren treiben diese Entwicklung voran.

1) Die Verbreitung von neuer Kommunikationtechnologie und neuen Protokollen, die dem mobilen Szenario optimal angepasst sind. UMTS als Technologie und WAP als Protokollsuite sind hierfür Beispiele.

2) Die Standardisierung dieser Technologien.

3) Die fortlaufende Verbesserung der mobilen Geräte.

Die Weiterentwicklung der mobilen Geräte ist von neuen Innovationen in verschiedenen Bereichen abhängig. Eine heutige Lithium-Ionen-Batterie hat zum Beispiel kleinere Ausmaße als eine Nickel-Kadmium-Batterie, aber eine höhere Energiedichte. Neue Prozessoren haben bei gleicher Leistung einen geringeren Energieverbrauch. Die günstigen Preise für Speicher machen es möglich den Speicherbedarf heutiger Anwendungen zu decken. Diese Entwicklungen machen eine weitere Steigerung der Leistung und der

Betriebsdauer möglich. Die weitere Verringerung der Ausmaße der Geräte ist nicht ohne die Verwendung neuer Ein-/Ausgabemechanismen möglich. Ein Beispiel ist die QWERTY Tastatur. Heutige Laptops besitzen eine verkleinerte Variante einer PC-Tastatur. Die Eingabegeschwindigkeit nimmt bereits deutlich ab. Wird diese Tastatur noch kleiner, ist eine normale Benutzung nicht mehr möglich. Aus diesem Grund ist eine Vielzahl neuer Ein-/Ausgabemechanismen entstanden, die an die Größenverhältnisse angepasst sind. Ein kleines Display lässt zum Beispiel nur bestimmte Dialogelemente zu. Verschiedene Eingabegeräte sind für diese Dialogelemente besonders geeignet.

Der erste Teil des Berichtes widmet sich der Beschreibung einiger Ein-/Ausgabemechanismen. Der zweite Teil vermittelt Vorgehensweisen für ihre Bewertung und stellt daraus resultierende Ergebnisse vor.

2 Beschreibung der Ein/Ausgabemechanismen

2.1 Displays

Die Vorzüge von LCDs (Liquid Crystal Display) bezüglich Gewicht, Größe und Energieverbrauch machen sie im Bereich der Mobile Devices unverzichtbar. Es gibt sie in verschiedenen Varianten. PDAs verwenden meistens DSTN(Double layer SuperTwist Nematic) Displays, die auf der Passive Matrix Technologie [Kiy00] basieren. Dabei können die einzelnen Pixel nur über die entsprechende Zeile und Spalte gesteuert werden. Im Multiplex Betrieb wird sequentiell auf die einzelnen Zeilen Strom gegeben. Der Treiberbaustein gibt die Pixeldaten für die entsprechende Zeile auf die Spaltenleitungen. Es leuchtet also immer nur eine Zeile mit voller Leistung. Das komplette Bild muß also immer wieder frisch aufgebaut werden, ähnlich der Fernsehtechnik. Aufgrund ihrer Bauart eignen sich diese Displays nicht für die Darstellung sich schnell verändernder Bilder. Ihr Vorteil ist ihr Preis, der deutlich geringer ist als der von TFT(Thin-Film-Transistor) Displays. Ein Grund für deren hohen Preis ist die verwendete Aktive Matrix Technologie. Hier wird jedes Pixel mit seinen eigenen Transistoren gesteuert. Kondensatoren sorgen für einen Speichereffekt, d.h der momentane Pixelzustand wird solange beibehalten, bis neue Pixelinformationen geliefert werden. Das Bild steht still. Insgesamt führt diese Technik zu einer geringeren Leistungsaufnahme. Die hohe Bildqualität und die flimmerfreie Technik machen die TFTs im Officebereich zur Konkurrenz für Röhrenmonitore.

In der Zukunft wird es mit Hilfe von OLEDs (Organic Light Emitting Devices) noch dünnere Displays geben, die eine größere Helligkeit bei geringerem Energieverbrauch haben. Diese Displays werden dann sogar biegsam sein. [Kiy00]

Auch Mikrodisplays dürften in Zukunft eine große Rolle spielen. [Kuhl00] Die einzelnen Pixel haben eine Größe von 10-12 Micrometern. Die Pixeldichte dieser Displays ist um vielfaches höher als die Pixeldichte heutiger LCDs. Mikrodisplays werden zusammen mit einer passenden Vergrößerungsoptik verwendet. Sie übertreffen die heutigen Displays bei der Fläche und Auflösung der gelieferten Bilder bei weitem. Ein weiterer Pluspunkt ist der geringe Stromverbrauch. Bild 1 zeigt eine Brille mit integriertem Mikrodisplay der Firma MicroOptical. Das Display hat eine Auflösung von 320*240 Punkten mit einer Farbtiefe von 16bit. [Micr]

Abbildung 1: Microdisplay von MicroOptical

2.2 Navigierende Geräte

Die momentanen Anwendungen der Mobilen Telefone verwenden Texteingabe und Auswahl aus Listen für den Dialog. Andere Dialogelemente, wie zum Beispiel Schieberegler oder ähnliches sind aufgrund der kleinen Displays nicht verwendbar. Navigierende Eingabegeräte geben dem Benutzer die Möglichkeit die verwendeten Dialogelemente schnell zu bedienen.

2.2.1 NaviRoller&JogDial

Der NaviRoller von Nokia [Navi] und JogDial von Sony [JogD] basieren beide auf dem selben Prinzip. Durch Drehen des Eingabeelementes kann der Benutzer die Listen durchlaufen und einzelne Listeneinträge durch Drücken des Elementes auswählen. Beim NaviRoller ist dieses mit Hilfe einer drehbaren Walze realisiert, Sony verwendet ein Drehrad. Die Elemente werden mit dem Daumen bedient.

2.2.2 Haptische Schnittstellen

Der programmierbare Drehregler mit haptischer Rückkopplung von VDO [VDO] ist ein weiterer Vertreter dieser Eingabegeräte. Im Gegensatz zu den anderen Geräten kann er eine frei programmierbare Rückkopplung erzeugen. Das Gerät erzeugt dazu einen für den Benutzer spürbaren Widerstand. So emuliert es wahlweise einen stufenlosen Drehregler bzw. einen der in einer frei bestimmbaren Zahl von Positionen einrastet. Der Drehregler kann also den Dialogen des Systems angepaßt werden. Die Rückkopplung gibt dem Benutzer die Möglichkeit bekannte Dialogabläufe auszuführen, ohne auf das Display zu achten.

2.3 Zeigegeräte

Viele Anwendungen verwenden eine Vielzahl verschiedener Dialogelemente. (Menüs, Listen , Knöpfe, Drehregler etc.) Der Benutzer benötigt hierfür flexible Eingabegeräte,

um die einzelnen Dialogelemente schnell und korrekt bedienen zu können. Hier kommen die Zeigegeräte zum Einsatz. Diese sind entweder direkt oder indirekt. [Shne98] Direkte Eingabegeräte (Lightpen, Stylus, Touchscreen) profitieren davon, daß das direkte Zeigen auf ein Dialogelement für den Menschen eine äußerst einfache Aufgabe ist. Lernaufwand existiert also nicht. Indirekte Zeigegeräte (Maus, Trackpoint, Trackball etc.) stellen auf dem Bildschirm Zeiger dar, um eine Rückkopplung für den Benutzer zu erzeugen. Die Koordination von Zeiger, zu wählendem Dialogelement und Eingabe ergibt einen zusätzlichen Aufwand, der dem Benutzer nach einer Weile jedoch nicht mehr auffällt. Im Gegensatz zu den indirekten Zeigegeräten ist also ein gewisser Lernaufwand aufzubringen. Die heutigen PDAs verwenden fast ausschließlich direkte Zeigegeräte. Laptops werden hauptsächlich mit indirekten Zeigegeräten ausgestattet, viele haben allerdings auch einen Touchscreen. Im Gegensatz zu den PDAs, ist eine Verwendung der indirekten Zeigegeräte aber unangenehm, da die Hand immer zum Bildschirm bewegt werden muß und dadurch schnell ermüdet. Im folgenden wird noch der Trackpoint, als Vertreter der in Laptops verwendeten Zeigegeräte, vorgestellt.

2.3.1 Trackpoint

Das Trackpointsystem [IBM] besteht aus einem isometrischen Joystick, den Auswahlknöpfen und einem Rechenelement mit PS/2 Schnittstelle. Der Joystick liegt zwischen den G, B und H Tasten der Tastatur und wird mit der Fingerspitze gesteuert. Seine Kappe ist ungefähr 1mm höher als die umliegenden Tasten. Die Hände bleiben beim Steuern in der selben Position wie beim Tippen. Der Daumen betätigt die Auswahlknöpfe. Da die Schnittstelle mit der einer PS/2 Maus kompatibel ist, können bereits vorhandene Treiber und Software unverändert übernommen werden. Der Joystick verarbeitet auch Kräfte, die in Richtung der Achse wirken, die senkrecht auf der Tastaturebene steht. Dadurch kann der Benutzer Dialogelemente ohne Verwendung der normalen Auswahlknöpfe anwählen.

2.4 Tastaturen

Die Eingabe von Text wird auch bei den mobilen Geräten oft benötigt. Kommen bei Laptops noch normale Tastaturen zum Einsatz, werden bei PDAs oft On-Screen- Tastaturen eingesetzt. PDAs haben meistens Displays, die gleichzeitig auch als Eingabegerät verwendet werden. Die On-Screen-Tastaturen werden in die Anwendung eingeblendet und mit dem Stylus bedient. Oft werden die Tastaturlayouts auf Overlayfolien gedruckt und auf eine vom Display getrennte Eingabefläche aufgebracht. Dies hat den Vorteil, daß der Anwendung das ganze Display zur Verfügung steht. Bei der Bedienung hat der Benutzer den Blick auf die Tastatur gerichtet, da diese im Gegensatz zu normalen Tastaturen keine Rückmeldung liefert. Ein anderer Weg Platz zu sparen ist der Einsatz von Tastaturen, die einer Taste mehrere Zeichen zuordnen.

2.4.1 QUERTY Tastatur

Christopher Latham Shole entwarf 1870 das QWERTY Layout für die Schreibmaschine. [Shne98] Er plazierte die oft verwendeten Zeichen weit auseinander, um die Eingabegeschwindigkeit zu reduzieren. Dadurch wurden Verklemmungen in der Mechanik

stark reduziert und seine Schreibmaschine zu einem Erfolg. Durch diesen Erfolg ist das QWERTY Layout heute immernoch der Standard. Andere Layouts, die den Fingerweg reduzieren und eine weitaus höhere Eingabegeschwindigkeit erlauben (Dvorak Layout 1920), führen ein Schattendasein oder sind ganz verschwunden. QWERTY Tastaturen gibt es in verschiedenen Größen und auch als ON-Screen-Tastaturen.

2.4.2 Fitaly

Unter der Annahme, daß alle Tasten die gleiche Größe haben und die Folgewahrscheinlichkeiten der einzelnen Buchstaben gegeben sind, führt MacKenzies Modell(siehe 3.1) unter dem Gesichtspunkt der Optimierung zu 2 Forderungen:

1) Der Abstand von Tasten mit hoher Folgewahrscheinlichkeit muß sehr gering sein.

2) Der Abstand aller Tasten zueinander muß gering sein.

Die Fitaly-Tastatur von Textware Solutions [Fita] erfüllt beide Forderungen in einem hohen Maß. Tabelle 1 zeigt die Anordnung der einzelnen Zeichen und die zugehörigen Häufigkeiten in einem englischen Text mit einem Umfang von 10000 Zeichen.

Z	**V**	**C**	**H**	**W**	**K**
20	**77**	**230**	**415**	**138**	**49**
F	**I**	**T**	**A**	**L**	**Y**
176	**551**	**701**	**615**	**319**	**133**
SPACE		**N**	**E**	**SPACE**	
1741		**550**	**976**	**1741**	
G	**D**	**O**	**R**	**S**	**B**
147	**305**	**590**	**497**	**497**	**110**
Q	**J**	**U**	**M**	**P**	**X**
20	**20**	**210**	**187**	**150**	**20**

Tabelle 1: Fitaly-Tastaturbelegung

Die Buchstaben T, A, N, E, O, R und die Leertaste umfassen bereits über 50% aller in einem englischen Text vorkommenden Zeichen und haben einen maximalen Abstand von einer Taste zur Mitte (N und E Tasten). Nimmt man die Buchstaben I, L, D, S hinzu hat man schon 73% aller Zeichen und mit C, H, U, M bereits 84%. Alle mit einem maximalen Abstand von 2 Tasten zur Mitte. Die Tasten sind außerdem so angeordnet, daß die Buchstaben mit hoher Folgewahrscheinlichkeit auch sehr nahe beieinander liegen. (Der Leser sollte mit einem Stift ein englisches Wort schreiben und sich selbst überzeugen).

Die Fitaly-Tastatur kann als On-Screen-Tastatur in die Anwendung eingeblendet werden oder als Overlayfolie auf ein separates Eingabefeld gelegt werden (Fitaly Stamp). Die Anwendung hat dann das ganze Display zur Verfügung. Bild 2 zeigt ein Beispiel auf dem Palm Pilot.

Die Fitaly-Tastatur stellt alle 220 Zeichen des ISO/ANSI Latin 1 Zeichensatzes bereit. Wenig vorkommende Zeichen werden durch Umschalten in einen anderer Tastaturmodus erreicht. Es ist möglich Zeichen zu erzeugen, indem man den Stift von der Mitte eines Zeichens in einer bestimmten Richtung aus dem Tastenfeld zieht.(Bild 3). Dies führt zu einer weiteren Zeitersparnis.

Abbildung 2: Fitaly für den Palm

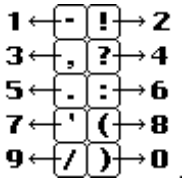

Abbildung 3: Die Bildung von Zeichen mit Sliding

2.4.3 Intelligente Tastaturen

Ist der Platz für die Tastatur sehr klein, sind die normalen Tastaturen, echt oder On-Screen, die eine 1:1 Zuordnung von Tasten und Zeichen haben, nicht mehr verwendbar. Ab einer bestimmten Größe nimmt die Eingabegeschwindigkeit stark ab (Fitt's Law) bis die Bedienung schließlich unmöglich wird. Eine Lösung für dieses Problem ist die 1:n Zuordnung von Tasten und Zeichen. Dadurch ist die Eingabe aber nicht mehr eindeutig. Bei alten Mobiltelefonen muß der Benutzer diese Mehrdeutigkeit selbst auflösen, indem er zum Beispiel die 'ABC' Taste 2 mal betätigt, um ein 'B' zu erhalten. Modernere Mobiltelefone nehmen dem Benutzer diese Arbeit ab. Das System löst die Mehrdeutigkeiten auf. Notfalls muß es den Benutzer um Hilfe bitten. Im Bereich der Texteingabe gibt es folgende Ansätze zur Auflösung der Mehrdeutigkeiten. [Kush, GoTe]

1) Buchstabenebene: Das System versucht nach jedem Tastendruck aufzulösen. Es verwendet statistische Untersuchungen über bestimmte n-Gramme als Vorhersagebasis. n-Gramme sind Gruppen von n Buchstaben, die in dieser Reihenfolge in Wörtern vorkommen. Da die Anzahl der n-Gramme gering ist, ist der benötigte Speicherplatz auch gering. Dieses Vorgehen hat folgenden Nachteil. Der Benutzer muß nach jedem Tastendruck überprüfen ob die richtige Vorhersage getroffen wurde.

2) Wortebene: Das System versucht nach der Eingabe eines Wortes aufzulösen. Es verwendet ein Wörterbuch. Um gute Ergebnisse zu erzielen, muß es fast alle möglichen Wörter enthalten. Dies führt zu einem größeren Speicherbedarf.

Aufgrund geringer Speicherkosten und der Verwendung von Datenkompressionsverfahren sind Datenbanken mit 130000-160000 Wörtern günstig realisierbar. Der größte

Nachteil des zweiten Ansatzes verringert sich dadurch und Systeme wie Tegics T9 verwenden daher das Wörterbuch.

T9

T9 von Tegic [T9] ist eines der bekanntesten Systeme dieser Art. Mehrere Funktelefonhersteller haben sich dafür entschieden und es ist für den Palm und andere PDAs erhältlich.(Bild 4) Der Name kommt von dem Ausdruck 'Typing with 9 Keys' und dementsprechend ist die Tastatur auch aufgebaut. Eine Taste steht für eine Gruppe von Buchstaben. Die Unterteilung der einzelnen Buchstaben in Gruppen richtet sich nach dem Alphabet. Die Leertaste ist eindeutig und bestimmt das Wortende.

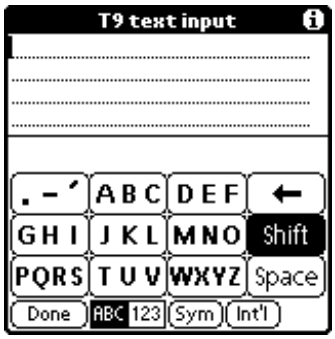

Abbildung 4: Die T9 Tastatur für den Palm.

Der Benutzer drückt die Taste der Gruppe, die den gewünschten Buchstaben enthält, nur noch einmal. Nach jedem Tastendruck werden alle Kombinationen der bisherigen Eingabe mit dem Wörterbuch verglichen und ein Vorschlag ausgegeben. Gibt es mehrere gültige Möglichkeiten, wird das Wort gewählt, das in der gewählten Sprache häufiger verwendet wird als alle anderen möglichen Lösungen. Hat der Benutzer ein Wort komplett eingegeben und ist der Vorschlag nicht das gewünschte Wort, kann er die anderen gefundenen Wörter anschauen und gegebenenfalls eines wählen. Ist das gesuchte Wort überhaupt nicht im Wörterbuch vorhanden, muß es erst eingetragen werden. Der Benutzer gibt die einzelnen Buchstaben des Wortes dann durch mehrfaches Drücken der jeweiligen Tasten auf eindeutige Weise ein.

Octave

Octave von E-Acute [Octa] ist ein weiteres System, das Mehrdeutigkeiten auf der Wortebene auflöst. Es ist in verschiedenen Varianten erhältlich. Für den Palm ist eine kleine Folie von ungefähr einem Quadratzentimeter auf die Eingabefläche aufzubringen. In der Mitte der Folie ist ein Stern mit acht Ecken ausgeschnitten, der zur Führung des Stiftes dient. Das ganze kann aber auch mit einem Joystick gesteuert werden. Es gibt wie bei T9 acht Buchstabengruppen. Die Gruppenzugehörigkeit eines Buchstaben richtet sich aber nicht nach dem Alphabet, sondern nach Symbolen, die alle Buchstaben einer Gruppe gemeinsam haben (vgl. Bild 5). Diese Aufteilung soll das Lernen erleichtern.

Um ein Wort einzugeben, setzt der Benutzer den Stift in der Ecke des Anfangsbuchstabens an und fährt dann die jeweils folgenden Ecken ab, ohne den Stift abzuheben. Das System zeigt ihm dabei immer die wahrscheinlichste Lösung an. Ist es die richtige, hebt der Benutzer den Stift ab und der Vorschlag wird an die Anwendung geschickt.(Bild 6)

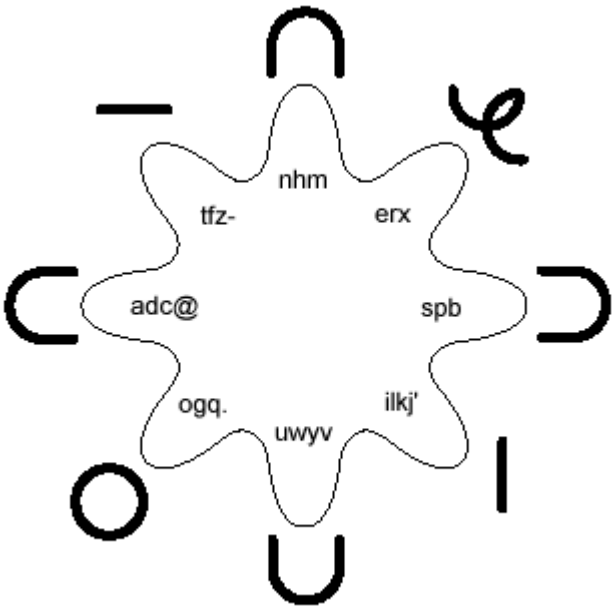

Abbildung 5: Der Octave Stern mit Buchstabeneinteilung

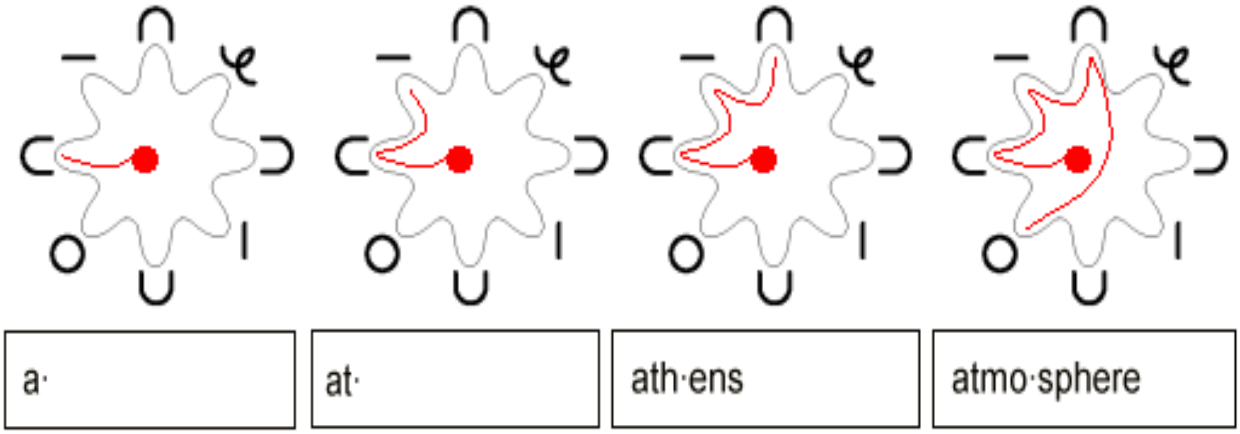

Abbildung 6: Eingabe des Wortes atmosphere

Nicht vorhandene Wörter werden auch hier in einer eindeutigen Art und Weise in das Wörterbuch eingefügt. Mit Octave lassen sich auch viele andere Zeichen des normalen ASCII-Zeichensatzes darstellen. Dazu wird mit Hilfe von Octave-Befehlen in einen bestimmten Eingabemodus gewechselt.

2.5 Handschriftenerkennung

Die heutigen PDAs haben bereits genügend Rechenleistung, um auch rechenintensivere Aufgaben zu lösen. Die Erkennung von handgeschriebenen Wörtern und Buchstaben gehört dazu. Die einzelnen Methoden zur Erkennung unterscheiden sich deutlich in

Rechenaufwand und Erkennungsgenauigkeit. An die letztere werden besonders hohe Anforderungen gestellt, da die Benutzerakzeptanz bei zu hoher Ungenauigkeit verloren geht. [MaZh97]

2.5.1 Worterkennung

Die Geschwindigkeit von Schreibschrift beträgt ungefähr 33 Wörter pro Minute und die Erkennung handgeschriebener Wörter ist die komplexeste aller Methoden. Letztere wird trotzdem durchgeführt, da die Lernphase entfällt. Eine für die Hersteller wichtige Eigenschaft. Die hohe Komplexität führt jedoch zu einer gewissen Fehlerrate, die von den Benutzern oft nicht akzeptiert wird.

2.5.2 Zeichenerkennung

Ein anderer Ansatz ist es sich auf die Erkennung von einzelnen Zeichen zu beschränken. Der Rechenaufwand und die Fehlerrate werden dadurch geringer. Bei den meisten Buchstaben wird der Stift zwischen den einzelnen Strichen abgehoben. Dadurch ist es schwierig festzustellen, wann die Eingabe eines Buchstabens beendet ist. Ein 'l' wird zum Beispiel erst durch einen vertikalen Strich zum 't'. Es sind verschiedene Lösungen denkbar, wie zum Beispiel zwei getrennte Eingabeflächen, die alternierend verwendet werden oder das Drücken einer Taste nachdem die Eingabe des Buchstabens beendet ist. Da viele Buchstaben gleiche Symbole enthalten, (vgl. Octave) ist die Unterscheidung schwierig. Eine Methode, die beide Probleme umgeht, ist die Einführung neuer Zeichen, die hinreichend verschieden sind und ohne Anheben des Stiftes gezeichnet werden können. Die Designentscheidung richtet sich nach Erlernbarkeit und Eingabegeschwindigkeit der einzelnen Zeichen. Unistroke und Graffiti sind zwei Beispiele hierfür, die jeweils eines der beiden Kriterien übergewichten. Im folgenden werden nur die Zeichen für Buchstaben beachtet.

Unistroke

Unistroke richtet sich vor allem an Expertenbenutzer, die für eine schnelle Eingabe einen höheren Lernaufwand in Kauf nehmen. Bild 7 zeigt die Abbildung des Alphabets auf die einzelnen Unistroke-Zeichen. Beim Design wurde hauptsächlich auf die Einfachheit der einzelnen Zeichen geachtet und nicht auf die Ähnlichkeit zu den Buchstaben die sie repräsentieren. Buchstaben mit hohen Häufigkeiten, wie zum Beispiel ' E, A, T, I, R ' sind sehr einfache Striche zugeordnet. Die einzelnen Zeichen sind vom System leicht unterscheidbar und machen den Erkennungsaufwand gering.

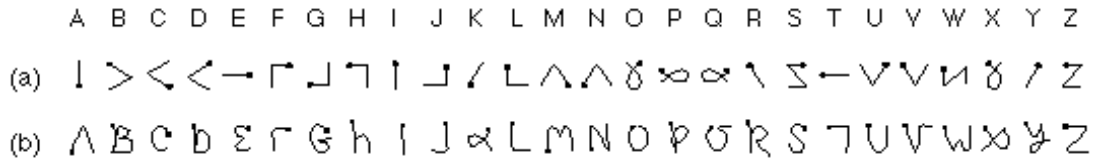

Abbildung 7: Zuordnung von Buchstaben und Zeichen bei (a) Unistroke (b) Graffiti

Graffiti

Graffiti wird auf den Palm Pilots eingesetzt. Die Zuordnung von Graffiti-Zeichen und Buchstaben ist ebenfalls in Bild 7 zu sehen. Das Ziel ist eine leichte Erlernbarkeit. Die Graffiti-Zeichen haben meistens eine hohe Ähnlichkeit mit den ihnen zugeordneten Buchstaben. Sie sind dadurch aber auch schwieriger zu zeichnen und sie gleichen sich untereinander mehr. Der Buchstabe ' L ' wird zum Beispiel durch eine zu starke Krümmung des vertikalen Striches zu einem ' H '. Der Aufwand für die Erkennung steigt also.

2.6 Eingabegeräte für tragbare Computer

2.6.1 Thumbcode

Thumbcode [Prat] ist eine Eingabemethode, die speziell für den Bereich Wearable Devices entwickelt wurde. Ein besonderes Ziel ist die Unabhängigkeit von der Erkennungshardware. Deshalb ist die Hand selbst das Eingabewerkzeug. Egal welche Erkennung später gewählt wird, Thumbcode bleibt gleich. Ein weiteres Ziel ist zeichenorientiert zu arbeiten und dabei alle Zeichen des ASCII Zeichensatztes und viele andere Zeichen einer normalen PC Tastatur darzustellen.

Bei der Eingabe wird folgendermaßen vorgegangen. Der Daumen dient als Eingabewerkzeug, die anderen 4 Finger als Tastatur. Die einzelnen Finger werden entsprechend ihrer 3-Segmente aufgeteilt in Spitze, Mitte und Basis. Jedes dieser Segmente ist bei einem Erwachsenen ungefähr 2.5 cm lang. So lassen sich bereits 12 Daumenzustände unterscheiden. Die Stellung der einzelnen Finger zueinander in dem Moment, in dem der Daumen die Eingabe betätigt, spielt auch eine Rolle. Es geht dabei um die Paare von Kleinem Finger und Ringfinger, Ringfinger und Mittelfinger und Mittelfinger und Zeigefinger. Die jeweiligen Paare können geöffneten oder geschlossenen Zustand einnehmen. Dadurch ergeben sich 8 weitere Zustände. Diese werden in Shifted und Unshifted unterteilt. Sie heissen Open, Pair, Trio, Closed. Bei Unshifted Closed und Unshifted Open sind alle Finger zusammen bzw offen. Bei Unshifted Pair ist nur Zeige und Mittelfinger zusammen und bei Unshifted Trio sind Mittel, Ring und kleiner Finger zusammen. Die entsprechenden Shifted Zustände werden durch Invertierung der Stellung des kleinen Fingers erreicht. Insgesamt gibt es also 96 verschiedene Zustände. Bild 8 zeigt die Stellungen zusammen mit der Zeichenbelegung der einzelnen Fingersegmente. Bis auf das Control-Zeichen sind dies alles ASCII-Zeichen. Um die restlichen Zeichen darzustellen, arbeitet man mit Steuercodes, die zwischen verschiedenen Tastaturmodi umschalten.

2.6.2 Twiddler

Der Twiddler von Handkey Corporation [Twid] ist eine Kombination aus Tastatur und Zeigegerät. Er ist für den einhändigen Gebrauch konzipiert und deswegen sind seine Ausmaße begrenzt. Das Gerät wird mit einem Gurt an der Hand befestigt. Der Twiddler besitzt 12 Tasten an der Vorderseite und 6 weitere an der Oberseite. Die Vorderen 12 sind in 4 Zeilen mit 3 Spalten unterteilt. Die Tasten haben kleine Einwölbungen, die

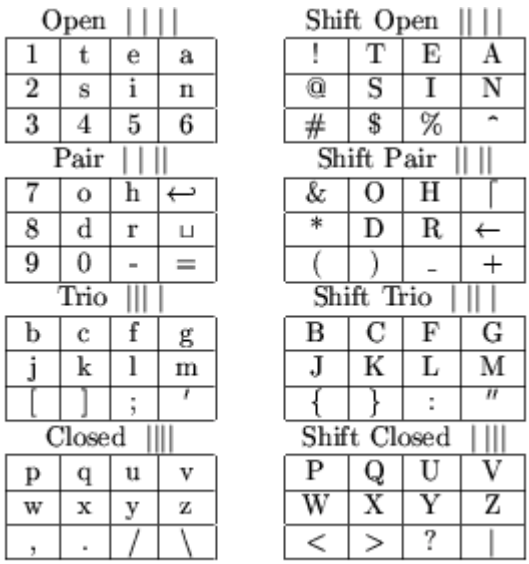

Abbildung 8: Die Bedeutung der einzelnen Fingersegmente in Unshifted und Shifted Zustand

dem Benutzer die Orientierung erleichtern. Jede dieser Zeilen ist einem Finger zugeordnet. Der Twiddler besitzt eine sogenannte Akkordtastatur. Bei der Eingabe von Zeichen wird zwischen Einzelzeichen und Akkorden unterschieden. Für ein Einzelzeichen wird nur eine Taste verwendet und ein Akkord wird durch das gleichzeitige Loslassen mehrerer Tasten gebildet. Bild 9 zeigt die Vorderseite des Twiddlers. Die Beschriftungen auf den Tasten sind die Einzelzeichen. Die Akkordzeichen stehen daneben. Jeder Taste sind mehrere davon zugeordnet und mit verschiedenen Farben markiert. Über den oberen Tasten befinden sich Punkte mit den 3 verschiedenen Farben. Ein Zeichen wird dann durch die Kombination einer dieser oberen Tasten und einer der anderen Tasten mit entsprechender Farbe gebildet. Die Tasten and der Oberseite haben die Wirkung der ALT, CTRL ... Tasten einer normalen PC-Tastatur. Der Benutzer kann die Belegung der Tastatur selbst konfigurieren und es ist möglich Tasten mit ganzen Wörtern zu belegen. Dadurch sind hohe Eingabegeschwindigkeiten möglich. Der Twiddler hat einen internen Bewegungssensor der zur Emulation der Maus eingesetzt wird. Um in den Maus-Modus zu kommen und darin zu bleiben, muss der Benutzer mit dem Daumen permanent eine Umschalttaste drücken. Die Bewegung der ganzen Hand wird von dem Sensor aufgenommen und in Mausbewegungen umgerechnet. Die Tasten an der Vorderseite dienen als Mausknöpfe. In der zweiten Version des Twiddlers wird die Maussteuerung von einem IBM Trackpoint übernommen.

3 Bewertung

Eine allgemeine Bewertung der einzelnen Eingabemechanismen quantitativ durchzuführen fällt schwer. Gibt es überhaupt Zahlen, ist meistens nicht zu erfahren, wie diese gewonnen wurden. Unabhängige Untersuchungen sind oft theoretischer Natur oder verwenden eine geringe Anzahl von Versuchspersonen. Ergebnisse, die aus verschiedenen

Abbildung 9: Der Twiddler

Untersuchungen stammen, sind nicht unbedingt vergleichbar. Oft sind die Versuchsbedingungen zu unterschiedlich. Zu vielen Fragen sind gar keine Ergebnisse vorhanden. Wie bereits erwähnt, werden bei der Untersuchung oft Versuche zur Ermittlung empirischer Daten und theoretische Modelle verwendet. Im folgenden wird ein theoretisches Modell vorgestellt und eine Versuchsdurchführung skizziert. Im Anschluß werden einige der oben vorgestellten Eingabemechanismen verglichen.

3.1 Modell

Ein Modell zur Bestimmung der Eingabegeschwindigkeit kann bei der Entwicklung einer Tastatur sehr hilfreich sein. Der Entwickler kann seine Entwürfe ohne großen Aufwand testen und damit den Bau eines Prototypen zeitlich nach hinten verschieben. MacKenzie & Soukoreff [MaSo99] entwickelten ein solches Modell für On-Screen -Tastaturen mit 1:1 Zuordnung von Buchstaben und Tasten. Annahme ist, daß ein Stylus zur Eingabe verwendet wird. Eine Komponente des Modells ist eine Variante von *Fitts Gesetz*. Der Psychologe Fitt stellte es bereits 1954 vor. Zunächst wird der Abstand der Tasten i und j zueinander bestimmt.

$$A_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$$

Dann wird die Zeit MT (in Sekunden) bestimmt, die der Benutzer benötigt um den Stylus von Taste i zur Taste j zu bewegen. MacKenzies Variante von Fitts Gesetz lautet allgemein:

$$MT = a + b \log_2(\frac{A}{W} + 1)$$

W ist dabei das Minimum von Höhe und Breite der Tasten. a und b sind Parameter, die mit Hilfe von empirischen Daten an die Versuchsanordnung angepaßt werden. Für das Tippen mit einem Stylus auf einer On-Screen-Tastatur folgt

$$MT_{ij} = 0.204 \log_2(\frac{A_{ij}}{W_j} + 1)$$

Das Hick-Hymen Gesetz (Hick 1952, Hymen 1953) beschreibt die Reaktionszeit eines ungeübten Benutzers, um eine Taste zu finden. Der Wert 27 resultiert aus der Hinzunahme der Leertaste zu den anderen Buchstaben des Alphabetes.

$$RT = 0.200 log_2(27) = 0.951\mathbf{s}$$

Für fortgeschrittene Benutzer wird RT zu 0 gesetzt.

Die Durchschnittszeit (in Sekunden) über alle möglichen Kombinationen zweier aufeinanderfolgender Zeichen, ist folgendermaßen definiert:

$$\overline{MT} = \sum_i \sum_j (P_{ij} \times MT_{ij}) + RT$$

Pij ist die Wahrscheinlichkeit, daß die Taste j der Taste i folgt. Die Eingabegeschwindigkeit v (in wpm (Wörtern pro Minute)) ist dann

$$v = (\frac{1}{\overline{MT}}) \times \frac{60}{5}$$

Ein Wort ist als 5 Zeichen inklusive Leertaste definiert.

Es gibt noch andere Modelle, zum Beispiel um die Eingabegeschwindigkeiten verschiedener Eingabemethoden von Funktelefonen vorherzusagen. [SiKo00]

3.2 Versuchsablauf

Thomas, Tyerman und Grimmer stellen den Test in ihrem Papier 'Evaluation of three Input Mechanisms for Wearable Devices ' vor. [ThGr97] Sie vergleichen 3 verschiedene Eingabemechanismen für einen tragbaren Computer bezüglich des Lernaufwandes und der Geschwindigkeit. Bei der folgenden Beschreibung liegt der Schwerpunkt bei dem Versuchsablauf und nicht bei der anschließenden Bewertung. Der tragbare Computer ist an der Hüfte befestigt. Ein am Gürtel befestigter Joystick dient als Mausersatz. Ein Minidisplay (Private Eye) ist vor dem Auge angebracht. Die drei Eingabemechanismen sind eine am Vorderarm befestigte Tastatur (Forearm Keyboard), eine On-Screen-Tastatur und eine Akkord-Tastatur. Das Forearm Keyboard hat auseinander liegende Tasten, die einen festen Druckpunkt haben. Der Benutzer weiß dadurch, daß eine Taste erfolgreich betätigt wurde. Die Tastatur hat ein QWERTY Layout. Die On-Screen-Tastatur wird mit dem Joystick bedient und hat ebenfalls ein QWERTY Layout. Die Akkord Tastatur ist ein handgroßes Gerät mit 5 Knöpfen.

Die drei Autoren verwenden folgenden Versuchsablauf. Das gesamte Experiment dauert 3 Wochen. Eine Woche vor dem Beginn findet eine Trainingsphase statt. In dieser Phase machen

sich die Versuchspersonen mit der Hard- und Software vertraut. Sie geben mehrere Texte ein. Danach beginnt das Experiment. In den 3 Wochen absolviert die Versuchsperson sechs Versuche. Sie dauern eine Stunde und sind folgendermaßen aufgebaut:

1) Die Versuchperson legt mit Hilfe anderer den Computer und die Eingabegeräte an. 2) Die Versuchsperson löst 9 Aufgaben. 3 für jedes Eingabegerät. 3) Die Versuchsperson legt den Computer ab. 4) Die Versuchsperson löst eine Aufgabe am PC mit einer QWERTY Tastatur. 5) Die Versuchsperson beantwortet einen Fragenkatalog.

Die Aufgabenstellungen sind an verschiedenen Wänden angebracht. Die Versuchspersonen lösen die Aufgaben im Stehen und laufen dann zu der Wand mit der nächsten Aufgabe. Die Autoren wollen dadurch den mobilen Einsatz simulieren. Die Versuchspersonen müssen 3 verschiedene Aufgabentypen bearbeiten. 1) Öffnen und Schließen: Die Versuchsperson öffnet eine Anwendung durch Drücken eines Icons. Sie gibt ein beliebiges Zeichen ein und schließt die Anwendung wieder. 2) Nachricht 1: Die Versuchsperson öffnet die Anwendung, gibt eine Textnachricht ein und schließt die Anwendung wieder. 3) Nachricht 2: Die Versuchsperson wiederholt die 'Nachricht 1' Aufgabe. Die 'Öffnen und Schließen' Aufgabe dient zur Feststellung der für die Bedienung der Stopuhr benötigte Zeit. Die Textnachrichten haben eine Länge von 50 bis 65 Zeichen und verwenden die Zahlen 1 bis 9. 'the 3 children were late at 845pm to 60 people street in 1972' ist eine davon. Fehler bei der Eingabe werden nicht verbessert. Ein Fehler ist ein fehlendes Zeichen, ein zusätzliches Zeichen oder ein falsches Zeichen.

Die 12 Versuchspersonen sind in einem Alter zwischen 19 und 42 Jahren. Der Durchschnitt ist 23.8 Jahre. Alle Personen haben bereits Computerkenntnisse. Sie können Hände und Füße ohne Probleme bewegen und besitzen volle Sehstärke.

Für die Beurteilung werden die Durchführungszeiten und die Anzahl der Fehler festgehalten. Die Autoren stellen folgende Ergebnisse fest. Bei der 'Öffnen und Schließen'-Aufgabe ist das Forearmkeyboard am schnellsten. Bei der On-Screen-Tastatur ist eine deutliche Abhängigkeit von der Versuchsperson aber kein Trainingseffekt festzustellen. Die Ergebnisse der Akkord-Tastatur lassen auf einen Trainingseffekt schliessen. Das Forearm Keyboard ist auch bei der 'Nachricht I'-Aufgabe am schnellsten. Hier ist bei allen Geräten ein Trainingseffekt vorhanden. Eine Abhängigkeit von der Versuchsperson ist bei der Forearm und Akkord-Tastatur erkennbar. Die 'Nachricht 2'-Aufgabe liefert ähnliche Ergebnisse. Bei den gemessen Schreibfehlern hat dir Forearm Tastatur ebenfalls die besten Werte.

3.3 Vergleich

Folgende Kriterien können für einen Vergleich von Eingabemechanismen verwendet werden.

Eingabegeschwindigkeit : Die Eingabegeschwindigkeit ist eines der wichtigsten Kriterien. Sie wird in wpm (Wörtern pro Minute) angegeben.

Größe : Die physikalischen Abmessungen bzw. der benötigte Anteil eines anderen Eingabegerätes (zum Beispiel Display und Eingabefläche eines Palm Pilots).

Lernaufwand : Der Aufwand den der Benutzer aufbringen muß, um ein durchschnittliches Niveau der Eingabegeschwindigkeit zu erreichen und auf diesem Niveau zu bleiben.

Fehlerrate : Die Fehlerrate die das System verursacht. Besonders wichtig bei Handschriftenerkennung und Spracheingabe.

Rechenleistung : Wieviel Prozessorleistung benötigt der Eingabemechanismus um seine Arbeit zu verrichten.

Art der Rückkopplung : Muß der Benutzer seinen Blick auf die Tastatur richten oder nicht. Dieser Punkt ist im Bereich der Wearable Devices wichtig.

Ein Experte erreicht mit einer normalen Qwerty Tastatur eine Eingabegeschwindigkeit von 150 wpm, ein Durchschnittsbenutzer um die 50 wpm. Vorraussetzung ist die beidhändige Bedienung. Bei einem Laptop fällt die Eingabegeschwindigkeit bereits. Eine On-Screen Version hat noch schlechtere Werte aufgrund der großen Stiftwege. Optimierte Tastaturen wie Fitaly schlagen die On-Screen Qwerty deshalb bei weitem. MacKenzie bestimmt mit seinem Modell(vgl 3.1) eine Eingabegeschwindigkeit von 43.2 wpm für eine On-Sreen Qwerty Tastatur. Fitaly erreicht einen Wert von 55.9. Tabelle 2 zeigt die Ergebnisse des Dom Perignon Testes der Firma Textware Solutions. Aufgabe war es 40 Zeichen einzugeben. Diese Werte unterscheiden sich von denen des theoretischen Modelles, zeigen aber auch, daß sich dieses für einen Vergleich eignet. Die Intelligenten Tastaturen liegen im Bereich der Fitaly Tastatur. Die 1:n Zuteilung von Tasten und Buchstaben wirkt sich allerdings vorteilhaft auf den benötigten Platz aus. E-Acute gibt für Octave einen Wert von mehr als 60 wpm an. Der Lernaufwand für die intelligenten Tastaturen hält sich aufgrund ihrer geringen Anzahl von Tasten in Grenzen. E-Acute gibt eine Zeit von 2 Stunden an, um die Lokation der Buchstaben zu kennen. Fitaly und Querty haben einen deutlich größeren Lernaufwand. Fitaly benötigt auf einem Palm 50 Kybte Speicherplatz, T9 170 Kbyte. Dieser Unterschied ist allerdings auf den meisten Systemen von keiner Bedeutung.

Tastatur	**Vmax**	**V Durchschnitt**	**Anzahl Teilnehmer**
Fitaly Stamp	81.74	57.73	18
Fitaly	74.34	51.03	13
Graffiti	49.44	28.19	19
Qwerty	47.05	36.14	5

Tabelle 2: Ergebnisse des Dom Perignon Tests

Unistroke und Graffiti sind trotz ihrer Gemeinsamkeiten bezüglich der Eingabe sehr unterschiedlich zu bewerten. Die einfachen Unistrokezeichen bringen einen Vorteil bei der Eingabegeschwindigkeit und erlauben einen einfacheren Erkennungalgorithmus. Der Rechenaufwand ist deshalb geringer als bei einem Graffiti System. Dies wirkt sich auch positiv auf die Fehlerrate aus. Das große Problem von Unistroke ist die Erlernbarkeit. Da der Benutzer 26 Zeichen lernen muß, die sich für einen Menschen teilweise nur wenig unterscheiden, hat Graffiti in diesem Punkt die besseren Ergebnisse.

Vergleicht man Unistroke und Graffiti mit den oben behandelten Tastaturen haben die letzteren bezüglich der Eingabegeschwindigkeit einen klaren Vorteil. Es stellt sich natürlich die Frage ob die Lage der Zeichen schneller zu lernen ist, als die verschiedenen Zeichen. Genaue Daten hierzu sind nicht verfügbar. Beachtet man zusätzlich die benötigte Rechenleistung scheint mir eine On-Screen-Tastatur wie Fitaly die bessere Alternative zu sein.

Zu Thumbcode sind keine Daten auffindbar. Einzelne Personen haben mit dem Twiddler Werte über 50 wpm erreicht. Tests mit mehreren Personen waren hier auch nicht zu finden. Der Lernaufwand für beide Systeme dürfte ähnlich hoch sein.

Vergleiche mit den anderern Eingabemechanismen in Bezug auf die Eingabegechwindigkeit sind nicht direkt durchführbar, da die einzelnen Methoden für ihren Aufgabenbereich spezialisiert sind. Der Twiddler ist zum Beispiel ohne Probleme mit einem Mikrodisplay kombinierbar. Eine On-Screen- Tastatur wie Fitaly kann zwar angezeigt werden, aber durch den Umstieg von einem direktem Zeigegerät (Stylus) zu einem Indirekten (z.B Joystick) geht einiges an

Geschwindigkeit verloren. Dies wird auch durch die Ergebnisse des oben beschriebenen Versuchs deutlich. Eine mögliche Lösung wäre hier vielleicht die Kombination mit einem System, das die Augenbewegungen misst. Die Ausführungen von T9 als echte Tastatur und Octave als Joystick sind eher vergleichbar. Allerdings sind diese beiden Mechanismen auf die Eingabe von normalen Text optimiert. Wörter die nicht im Wörterbuch stehen verursachen bereits eine Verminderung der Geschwindigkeit. Der Twiddler macht hier keine Unterschiede.

4 Abschließende Bemerkungen

Es wurde die Funktionsweise verschiedener Ein-/Ausgabemechanismen beschrieben. Ein Teil dieser Mechanismen wird kommerziell eingesetzt, andere wiederum sind eher Gegenstand der Forschung. Die einzelnen Ein-/Ausgabemechanismen verwenden dabei verschiedene Ansätze, um die an sie gestellten Anforderungen bezüglich Größe, Bedienbarkeit, Lernaufwand usw. zu erfüllen. Allen gemein ist eine starke Spezialisierung auf die zu erfüllenden Aufgaben. Dadurch ist es nicht immer sinnvoll alle Mechanismen direkt miteinander zu vergleichen, und dies spiegelt sich auch in dem abschließenden Vergleich wider. Es wurden auch zwei Methoden für die Bestimmung der Eingabegeschwindigkeit von Geräten vorgestellt. Es wird deutlich, wie stark die veröffentlichten Werte und besonders deren Aussagekraft von den verwendeten Methoden abhängen.

Die weitere Verringerung der Geräteabmessungen wird auch in Zukunft das Ziel der Entwickler sein. Der heutige Ansatz, alle Komponenten (Recheneinheit, Ein-/Ausgabegeräte) in einem Gerät zu integrieren, wird zunehmend Konkurrenz durch den Ansatz der Wearable Devices erhalten. Hier werden die einzelnen Systemkomponenten auf einzelne Geräte verteilt, die am Körper befestigt oder in Kleidung integriert werden. Unabhängig von dem gewählten Ansatz ist auch weiterhin mit interessanten Lösungen zu rechnen.

Literatur

[Fita] Fitaly. www.fitaly.com.

[GoTe] Alsio Goldstein, Book und Tessa. Ubiquitous Input for Wearable Computing : Qwerty Keyboard without a board.

[IBM] IBM. Trackpoint Engineering Specification Version 4 www.ibm.com.

[JogD] JogDial. www.sony.com.

[Kiy00] Michael Kiy. Molekulares Leuchten. *c't* Band 20, 2000, S. 110–113.

[Kuhl00] Ulrike Kuhlmann. Kleinanzeigen. *c't* Band 4, 2000, S. 300–303.

[Kush] Kushler. AAC Using a reduced Keyboard. CSUN 98 Papers.

[MaSo99] Zhang MacKenzie und Soukoreff. Text entry using soft keyboards. *Behaviour and Information Technology* Band 18, 1999, S. 235–244.

[MaZh97] MacKenzie und Zhang. The Immediate Usability of Graffiti. In *Proceedings of Graphics Interface 97*. Canadian Information Processing Society, 1997, S. 129–137.

[Micr] MicroOptical. www.microopticalcorp.com/HomePage.html.

[Navi] NaviRoller. www.nokia.com/phones/7110/phone/new/roller.html.

[Octa] E-Acute Octave. www.e-acute.fr.

[Prat] Vaughan R. Pratt. Thumbcode: A Device-Independent Digital Sign Language.

[Shne98] Ben Shneiderman. *Designing the User Interface*. Addison-Wesley. 1998.

[SiKo00] MacKenzie Silfverberg und Korhonen. Predicting Text Entry Speed on Mobile Phones. In *Proceedings of the ACM Conference on Human Factors in Computing Systems - CHI 2000*. ACM, 2000, S. 9–16.

[T9] Tegic T9. www.tegic.com.

[ThGr97] Tyerman Thomas und Grimmer. Evaluation of Three Input Mechanisms for Wearable Computers. In *First International Symposium on Wearable Devices*, 1997.

[Twid] Twiddler. www.handkey.com.

[VDO] VDO. www.vdo.de.

Abbildungsverzeichnis

Tabellenverzeichnis