University of Zurich
Department of Informatics

Burkhard Stiller
Thomas Bocek
Cristian Morariu
Peter Racz
Gregor Schaffrath
Martin Waldburger
(Eds.)

Mobile Systems II

March 2007

TECHNICAL REPORT – No. ifi-2007.04

University of Zurich
Department of Informatics (IFI)
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland

ifi

B. Stiller, T. Bocek, C. Morariu, P. Racz, G. Schaffrath, M. Waldburger (Eds.):
Technical Report No. ifi-2007.04, March 2007
Communication Systems Group (CSG)
Department of Informatics (IFI)
University of Zurich
Binzmühlestrasse 14, CH-8050 Zurich, Switzerland
URL: http://www.csg.uzh.ch

Introduction

The Department of Informatics (IFI) of the University of Zürich, Switzerland works on research and teaching in the area of communications. One of the driving topics in applying communications technology is addressing investigations on mobility aspects and support for mobile users. Therefore, during the winter term WS 2006/2007 a new instance of the Mobile Systems seminar has been prepared and students as well as supervisors worked on this topic.

Even today, the increasing number of mobile and wireless networks as well as their users or customers drive many developments of systems and protocols for mobile systems. The areas of underlying networking and development technology, of services assisting security or Quality-of-Service (QoS), and of mobility support determine an important part of future wireless networks. Therefore, this year's seminar addressed such areas in more depth. The understanding and clear identification of problems in technical and organizational terms have been prepared and challenges as well as weaknesses of existing approaches have been addressed in a mobile and wireless environment. All talks in this seminar provide a systematic approach to judge dedicated pieces of systems or proposals and their suitability.

Content

This second edition of the seminar entitled "Mobile Systems II" discusses a number of selected topics in the area of mobile communication. The first talk "MANET – Mobile Ad-hoc Networks" gives an introduction to mobile ad-hoc networks and discusses challenges. Talk two "Operating Systems for Mobile Devices" provides an overview of operating systems specially tailored to mobile devices. "WiMAX Wireless Network Technology" as talk three presents the new wireless network technology, WiMAX. The fourth talk addresses "Wireless Sensor Networks". Talk five "Delay Tolerant Networks – Challenges and Solutions" discusses networks, where application and network protocols have to deal with high network delays. The sixth talk "Push E-mail Systems" addresses e-mail systems providing push services.

Talk seven "Intrusion Detection in Wireless and Ad-hoc Networks" outlines intrusion detection systems specially focusing on wireless network environments. "Virus and Spam Threats on Mobile Devices" as talk eight discusses security issues as well, addressing viruses and spams on mobile devices. The ninth talk "Voice and Video Transmission over

Wireless Networks" presents protocols and mechanisms used for voice and video transmission in wireless networks. Talk ten "Routing in Multi-hop Mesh Networks" focuses on routing protocols in mesh networks. Talk eleven continues with "QoS-enabled MAC Schemes for Wireless Networks" and presents MAC protocols designed to improve QoS provisioning in wireless networks. Finally, talk twelve "Mobile Content Providers and Net Neutrality" discusses content providers vs. neutral network access.

Seminar Operation

Based on well-developed experiences of former seminars, held in different academic environments, all interested students worked on an initially offered set of papers and book chapters. Those relate to the topic titles as presented in the Table of Content below. They prepared a written essay as a clearly focused presentation, an evaluation, and a summary of those topics. Each of these essays is included in this technical report as a separate section and allows for an overview on important areas of concern, sometimes business models in operation, and problems encountered.

In addition, every group of students prepared a slide presentation of approximately 45 minutes to present his findings and summaries to the audience of students attending the seminar and other interested students, research assistants, and professors. Following a general question and answer phase, a student-lead discussion debated open issues and critical statements with the audience.

Local IFI support for preparing talks, reports, and their preparation by students had been granted by Thomas Bocek, Cristian Morariu, Peter Racz, Gregor Schaffrath, Martin Waldburger, and Burkhard Stiller. In particular, many thanks are addressed to Peter Racz for his strong commitment on getting this technical report ready and quickly published. A larger number of pre-presentation discussions have provided valuable insights in the emerging and moving field of Mobile Systems, both for all groups of students and supervisors. Many thanks to all people contributing to the success of this event, which has happened in a lively group of highly motivated and technically qualified students and people.

Zürich, March 2007

Contents

Kapitel 1

Mobile Ad-hoc Networks

Bielik Jan, Vonesch Christian, Wirz Franziska

Wichtige Aspekte der heutigen Gesellschaft bilden die Mobilität und die Kommunikation. Jedoch widersprechen sich diese zwei Eigenschaften oft. Auf dem Weg zur Arbeit, in den öffentlichen Transportmitteln und auch im öffentlichen Raum fehlt häufig ein Zugang zu einem schnellen Netzwerk, zum Beispiel zum Internet. Eine Lösung für dieses Problem liefern so genannte MANETs, welche im RFC 2501 [1] beschrieben werden. Diese Seminararbeit wird sich mit dieser neuen Netzwerktechnologie beschäftigen. Da in den letzten Jahren sehr intensiv an diesem Thema geforscht wurde, soll diese Arbeit in erster Linie einen state-of-the-art Überblick in diese Technik geben und daher die einzelnen Themenbereiche nicht bis ins letzte Detail besprechen. Fragen nach dem Stand der aktuellen Forschung werden hier beantwortet und auch die zukünftige Entwicklung wird beleuchtet. Vertieft behandelt werden die technischen Aspekte, wie auch die Herausforderungen, mit welchen diese Technologie noch zu kämpfen hat. Bis heute sind MANETs noch nicht sehr verbreitet. In dieser Arbeit sollen dennoch einige Beispiele aufgezeigt werden, in welchen man diese Technik einsetzen kann.

Inhaltsverzeichnis

1.1 Einführung

Ein Modewort, das von den IT-Fachleuten gerne in den Mund genommen wird, ist "Ubiquitous Computing". Dies bezeichnet die Allgegenwärtigkeit von Informationsverarbeitung und damit einhergehend der jederzeitige Zugriff auf Daten von beliebiger Stelle aus [2]. Mit dem Aufkommen der vielen kleinen mobilen Geräten (wie z.B. MP3-Player, Handy, Pocket-PC und Wearable-Computers) ist man diesem Ziel einen kleinen Schritt näher gekommen. Jedoch lassen sich diese Geräte nicht oder nur sehr schwer miteinander verbinden.

Wenn nun all diese Geräte sich automatisch miteinander verbinden und so selbstständig Informationen austauschen könnten, eröffnet dies für den Menschen einen erheblichen Mehrwert. Eine Person besitzt zum Beispiel einen MP3-Player und einen Notebook, welche miteinander verbunden sind. Dieser Benutzer wartet am Flughafen auf sein verspätetes Flugzeug und vertreibt sich die Zeit mit Musikhören. Gleichzeitig und vollautomatisch verbindet sich das Notebook mit einem Accesspoint im Terminal und empfängt eine neue Email. Dies kann das Notebook nun dem MP3-Player signalisieren und der Player kann diese Information akustisch dem Benutzer weiterleiten. Wenn nun diese Person noch einen PDA besitzt, welcher ebenfalls in dieses Netzwerk integriert ist, kann er die Email sogar auf seinem Pocket-PC anschauen und muss nicht einmal den Notebook aus der Tasche auspacken.

Ein solches Netzwerk, wie im oben beschriebenen Beispiel dargestellt, wird PAN, Personal Area Network, genannt. Die Technologie, um die Daten in einem PAN zu übertragen, ist von Vorteil drahtlos. Hierfür gibt es auf dem Markt schon einige etablierte und bewährte Funkstandards. Zu nennen sind hier Wireless LAN (IEEE 802.11), IrDA, UMTS, GSM, sowie auch Bluetooth [3]. Die meisten der heute gebräuchlichen mobilen Geräte haben einen oder mehrere dieser Standards bereits integriert. Ein grosser Teil der Infrastruktur für den Aufbau eines MANETs ist also bereits vorhanden. Es sind jedoch noch grössere Hürden zu nehmen, bis sich diese neue Technologie im alltäglichen Leben durchsetzten kann. So ist man in der Forschung noch bemüht einen effektiven und effizienten Routing-Algorithmus zu finden, der mit der stark dynamischen Topologie eines MANETs zurechtkommt. Des Weiteren muss der Betrieb sehr energieschonend sein, da die an einem MANET beteiligten Geräte mobil sind und daher mit einem Akku betrieben werden. Jedoch wird auf diese und einige andere Probleme später im Text noch detaillierter eingegangen. An dieser Stelle soll nur erwähnt sein, dass man im Moment noch einige Schritte von einem voll funktionsfähigen MANET entfernt ist.

MANET ist eine Abkürzung und steht für Mobile Ad hoc Netzwerk. Wie man aus diesen Wörtern herauslesen kann, ist ein MANET ein Netzwerk, welches spontan, also ad hoc gebildet wird und ohne feste Installation auskommt. Eine Definition könnte etwa wie folgt lauten:

Ein MANET ist ein sich selbst konfigurierendes, ohne zentrale Einheit auskommendes multihop-Netzwerk, welches aus mobilen Knoten besteht und drahtlos aufgebaut ist.

Der erste Teil dieser Arbeit beschreibt die technische Seite eines MANETs. Es wird aufgezeigt, wie die technische Implementierung eines solchen Netzwerks aussieht und welche

Tabelle 1.1: Erläuterung einiger Begriffe (In Anlehnung an [4])

Abkürzung	Begriff	Bedeutung
AODV	Ad hoc on-demand distance vector	Routing-Algorithmus, siehe Kapitel Routing
AP	Access Point	Zugangspunkt für Clients in einem WLAN
DSDV	Destination-sequenced distance vector	Routing-Algorithmus, siehe Kapitel Routing
DSR	Dynamic source routing	Routing-Algorithmus, siehe Kapitel Routing
GPRS	General packet radio service	Paketorientierter Übertragungsdienst im Bereich des Mobilfunks
IEEE	Institute of Electrical and Electronic Engineering	Berufsverband und Standardisierungsinstitut für den Netzwerkbereich
LAN	Local area network	Rechnernetz, mit maximal einigen 100 Metern Ausdehnung
MANET	Mobile ad hoc network	Mobiles, sich selbst konfigurierendes multihop Netzwerk
PAN	Personal area network	Rechnernetz, mit maximal einigen Metern Reichweite
PDA	Personal digital assistant	Kleiner tragbarer Computer
QoS	Quality of service	Gesamtheit aller Qualitätsmerkmale eines Dienstes
UMTS	Universal mobile telecommunications system	Mobilfunkstandard der dritten Generation
WLAN	Wireless LAN	Drahtloses Funknetz

Hardware dabei benötigt wird. Danach wird auf die Probleme und Herausforderungen dieser Technologie eingegangen. Da diese Technik noch sehr jung ist, gibt es in diesem Bereich noch einige Schwierigkeiten, die zu lösen sind. Natürlich werden auch die Lösungsansätze aufgezeigt, wie die Forscher diese Probleme behoben oder umgangen haben. Im hinteren Teil dieser Seminararbeit wird beschrieben, wie und wo man diese Technologie sinnvoll anwenden kann. Dies soll anhand einiger Anwendungsbeispiele illustriert werden. Am Schluss soll noch versucht werden, die Zukunftsaussichten dieser Technologie abzuschätzen.

1.2 Terminologie

Einführend sollen hier noch die wichtigsten Begriffe erläutert werden, die im Zusammenhang mit MANETs gebraucht werden. Nach [1] sind “Mobile Packet Radio Networking”, “Mobile Mesh Networking”, “Mobile Networking”, “Multihop Networking” und “Wireless Networking” Synonyme für Mobile Ad hoc Netzwerke. Die zwei erstgenannten Begriffe stammen aus den 70er und 80er Jahren und wurden hauptsächlich vom Militär benutzt

und sind daher heute nicht mehr sehr geläufig. Tabelle 1.1 zeigt die wichtigsten Abkürzungen auf, die im Gebiet der MANETs gebraucht werden. Sie soll als Nachschlagehilfe beim Lesen dieser Arbeit oder anderen Büchern auf diesem Gebiet dienen.

1.3 Technische Aspekte

MANETs unterscheiden sich in vielerlei Hinsicht von gewöhnlichen Netzwerken, welche heutzutage von der breiten Bevölkerungsschicht bereits genutzt werden. Im folgenden Abschnitt 1.3.1 werden die spezifischen Eigenschaften von MANETs hervorgehoben und anschliessend mit den gängigsten Netzwerktypen kurz verglichen. Unterschiede in der Infrastruktur und im Aufgabenbereich der einzelnen Endgeräte werden in diesem Absatz eine wichtige Rolle spielen. Danach werden in 1.3.2 die Komponenten präsentiert, welche für die Erstellung eines MANETs gebraucht werden. Abschliessend werden die prägnantesten Aussagen aller Abschnitte in einer Zusammenfassung wiedergegeben.

1.3.1 Charakteristiken

Ein mobiles ad hoc Netzwerk besitzt im Gegensatz zu herkömmlichen Netzwerken die vorteilhafte Eigenschaft, dass es keine stationäre Infrastruktur benötigt. Die teilnehmenden Geräte kommunizieren über drahtlose Verbindungen miteinander ohne auf eine zuvor fest installierte Basisstation zugreifen zu müssen [5]. Ein solches Netzwerk kann sich somit spontan bilden, sobald die Geräte in unmittelbarer Reichweite zueinander sind und Kontakt miteinander aufnehmen wollen. Bei einem MANET gibt es keine zentrale Instanz, weshalb sich die Geräte untereinander selbst organisieren müssen. Durch die drahtlose Kommunikation kann die Mobilität der Geräte unterstützt werden. Diese können sich demzufolge frei bewegen und die Topologie des Netzwerkes stark verändern [6].

Typischerweise sind die Geräte, die ein MANET bilden, heterogen. Diese Heterogenität beschränkt sich jedoch nicht nur auf verschiedene Geräte des gleichen Gerätetyps (z.B. Nokia und Sony Ericsson Handys), sondern beinhaltet auch den Aspekt, dass verschiedene Gerätetypen (z.B. Laptops, Handys und PDAs) miteinander verbunden werden können [4].

Oft werden die beteiligten Geräte eines MANETs auch als Knoten bezeichnet. Üblicherweise kann ein solcher Knoten nicht mit allen anderen Teilnehmern des MANETs eine direkte drahtlose Verbindung herstellen. Die Nachbarn eines Knotens A sind diejenigen Teilnehmer, welche sich in der Reichweite von A befinden und mit denen A dementsprechend eine direkte Verbindung herstellen kann. Um einen entfernten Knoten innerhalb des Netzwerkes trotzdem erreichen zu können, müssen die dazwischen liegenden Knoten als Router agieren und die Daten weiterleiten. Daraus wird ersichtlich, dass sich die Kommunikation der Geräte in einem MANET nicht nur auf das Senden und Empfangen von Daten reduzieren lässt, wie dies bei einem Endbenutzer eines klassischen Netzwerkes üblich ist. Eine Verbindung zwischen zwei Geräten in einem MANET führt gewöhnlich über mehrere intermediäre Geräte und setzt sich somit aus mehreren Teilstrecken zusammen. Häufig wird in der Literatur und im Sprachgebrauch für diese Art von Verbindung

der englische Begriff multihop und für eine direkte Verbindung zwischen dem Sender und Empfänger der englische Begriff singlehop benutzt. Ein hop bezeichnet dabei die direkte Verbindung zwischen zwei Geräten und entspricht dadurch einer Teilstrecke in einer Verbindung. Da einerseits alle Knoten das Netzwerk zu jeder Zeit betreten und auch wieder verlassen können und andererseits die Tatsache besteht, dass es in einem MANET keine zentrale Einheit gibt, liegt eine Gleichberechtigung aller Knoten sehr nahe [6].

Ein MANET bildet ein autonomes System, welches in isolierter Arbeitsweise operieren kann. Es besteht aber auch die Möglichkeit mit Hilfe eines teilnehmenden Gerätes eine Verbindung zum Internet oder zu einem Unternehmensnetzwerk herzustellen [1]. Diese Verbindung kann z.B. mit einem Laptop über einen LAN Accesspoint aber auch mittels eines GPRS/UMTS Handys, welches hierbei als Gateway eingesetzt wird, bereitgestellt werden [4].

Aus den vorher aufgeführten Eigenschaften wird erkennbar, dass ein Knoten den anderen Teilnehmern einen spezifischen Dienst zur Verfügung stellen kann. Die Weiterleitung des Datenverkehrs oder der Zugang zum Internet sind mögliche Ausprägungen eines Dienstes [7]. Ein Knoten in einem MANET kann sich nicht auf ein Netzwerk stützen, welches die Sicherheit und die Routing-Funktionen überwacht. Deshalb müssen die Routing-Algorithmen und die Sicherheitsüberprüfungen auf einer verteilten Basis operieren [4]. Die mobilen Geräte, die in einem MANET miteinander kommunizieren, werden immer vielfältiger, kleiner und leistungsfähiger. Jedoch besitzen sie begrenzte Eigenschaften in Bezug auf ihre Rechenleistung (CPU), Arbeitsspeicherkapazität und Energiereserven (Akku).

Die IETF MANET Working Group [8] befasst sich intensiv mit der Standardisierung der Funktionalität des IP Routing-Protokolls. Des Weiteren hat sie die Aufgabe ein effizientes Forwarding-Protokoll für MANETs zu entwickeln. Die Working Group hat bereits vier Request for Comments (RFCs) veröffentlicht, die einen Quasi-Standard oder die zurzeit besten Vorgehensweisen für MANETs beschreiben. Im RFC 2501 [1] werden die Eigenschaften von MANETs beschrieben, welche am meisten hervorstechen:

- Dynamische Topologien: Wegen der nahezu uneingeschränkten Mobilität der Knoten verändert sich die Topologie eines MANETs sehr schnell und zufällig. Da der Zeitpunkt dieser Veränderung nicht vorausgesagt werden kann, wird die Übertragung der Daten über eine multihop Verbindung beträchtlich erschwert.

- Bandbreitenbeschränkungen, variable Verbindungskapazitäten: Drahtlose Verbindungen werden wohl auch in Zukunft im Vergleich zu den fest verkabelten Verbindungen eine tiefere Kapazität aufweisen. Der realisierte Throughput ist zudem sehr viel tiefer als die maximale drahtlose Übertragungsrate. Gründe dafür sind z.B. die vielen Zugänge zu intermediären Knoten, die bei einer multihop Verbindung zwangsweise hergestellt werden müssen, die Abschwächung des Signals, welche mit zunehmender Distanz exponentiell ansteigt oder die Interferenzen nahe gelegener drahtloser Verbindungen, die das Signal zusätzlich abschwächen.

- Energiebeschränkte Operationen: Die Geräte eines MANETs werden gewöhnlich durch eine Batterie oder einen Akku betrieben. Die Energiereserven handlicher Geräte (wie z.B. Handys, Laptops, MP3-Players oder auch PDAs) reichen heutzutage

nur für einige Stunden aus, wenn diese intensiv benutzt werden, wie dies bei drahtlosen Übertragungen üblich ist. Folglich müssen die Operationen der Geräte auf ihren Energieverbrauch optimiert werden, um den Geräten eine möglichst lange Einsatzzeit im Netzwerk zu gestatten. Damit ein MANET jedoch erfolgreich funktionieren kann, müssen die Knoten auch dazu bereit sein, fremde Daten weiterzuleiten, welche ihnen keinen zusätzlichen Nutzen bringen. Ist diese Weiterleitungsoperation aber mit einem hohen Energieverbrauch verbunden, wird kaum jemand sein Gerät am Netzwerk anschliessen, ausser er möchte zur Zeit gewisse Informationen über das Netzwerk senden oder empfangen. Jeder Gerätebenutzer ist letztendlich darin bestrebt, dass sein Gerät noch genügend Energieressourcen besitzt um den eigentlichen Nutzen erfüllen zu können.

- Begrenzte physikalische Sicherheit: Fixe Netzwerke sind weniger anfällig gegenüber Sicherheitsbedrohungen als mobile drahtlose Netzwerke. Über die häufig vorkommenden multihop Verbindungen können weiterzuleitende Daten von den intermediären Knoten verändert oder abgehört werden. Ein intermediärer Knoten kann aber auch einen nachgefragten Dienst verweigern ohne dass der Dienstanbieter davon in Kenntnis gesetzt wird.

In Tabelle 1.2 werden die zuvor erläuterten Charakteristiken von MANETs mit denjenigen eines Mobilfunknetzes und des Internets verglichen.

1.3.2 Komponenten

Ohne die teilnehmenden Geräte würde ein MANET bereits bei der Entstehung scheitern. Denn diese Geräte stellen durch ihre Verbundenheit untereinander die Infrastruktur zur Verfügung, welche für die Kommunikation benötigt wird. Die Kommunikation wiederum wird durch eine drahtlose Mobilfunktechnik ermöglicht, die eine weitere zentrale Komponente eines MANETs darstellt. Der gewählte Mobilfunkstandard begrenzt die Bandbreite, die den mobilen Geräten zur Übermittlung der Daten bereitgestellt wird. Dabei kann die verfügbare Bandbreite nicht nur für die Übermittlung der eigenen Daten verwendet werden, sondern muss auch für die Weiterleitung von fremden Daten zur Verfügung gestellt werden. Für die Weiterleitung benötigt jedes MANET eine zusätzliche Komponente, nämlich ein Routing-Protokoll. Mit dessen Hilfe können zudem Knoten innerhalb des MANETs ausfindig gemacht werden und Änderungen in den Verbindungsrouten, aufgrund der dynamischen Topologie, rechtzeitig erkannt werden. Des Weiteren sind Algorithmen notwendig, die für das Auffinden von Diensten und für die automatische Zuweisung und Anpassung von Adressen eingesetzt werden können.

MANETs bestehen aus mobilen Endgeräten, welche über eine drahtlose Mobilfunktechnik miteinander kommunizieren. Hierfür benötigen die zumeist verschiedenen Gerätetypen keine stationäre Infrastruktur, da sie als Router agieren. Da in einem solchen Netzwerk keine zentralen Instanzen vorhanden sind, müssen sich die spontan zusammengeschlossenen Geräte untereinander selbst organisieren. Ein MANET kann über eines der teilnehmenden Geräte mit dem Internet verbunden werden, aber auch abgegrenzt von anderen Netzwerken operieren. Die Endgeräte verkörpern die wichtigste Komponente des MANETs. Diese

Tabelle 1.2: Gemeinsamkeiten und Unterschiede zwischen MANETs, Mobilfunknetzen und Internet

	MANET	Mobilfunknetz	Internet
Infrastruktur	keine stationäre Infrastruktur notwendig	stationäre Infrastruktur notwendig	stationäre Infrastruktur notwendig
Mobilität der Endgeräte	alle Endgeräte sind mobil	alle Endgeräte sind mobil	Endgeräte mehrheitlich stationär (PCs) und seltener mobil (Laptops, Handys, PDAs)
Kommunikation	drahtlos	drahtlos zur Infrastruktur im Netzwerk über Kabel und drahtlos	drahtlos zum Zugangspunkt möglich im Netzwerk häufiger über Kabel
Topologie des Netzwerkes	dynamisch	statisch	statisch
Netzwerkzugang	über jedes teilnehmende Gerät möglich	nur über zuvor installierte Infrastruktur möglich (Antennen)	nur über zuvor installierte Zugangspunkte möglich (WLAN Accesspoints)
Aufgabenbereich der Endgeräte	senden, empfangen und weiterleiten von Daten	senden und empfangen von Daten	senden und empfangen von Daten

benötigen einen geeigneten Mobilfunkstandard um Daten schnell untereinander austauschen zu können. Zudem sind für ein MANET ein Routing-Protokoll und verschiedene Algorithmen erforderlich.

Mit den oben erwähnten technischen Aspekten eines MANETs lassen sich nun im nächsten Kapitel die verschiedenen Herausforderungen aufzeigen.

1.4 Herausforderungen

Das folgende Kapitel beschäftigt sich mit den vielfältigen Herausforderungen die es im Bereich der MANETs zu lösen gilt. Zum einen muss bestimmt werden wie die Dienstgüte von mobilen ad hoc Netzen gemessen werden soll, was im Kapitel Quality of Service behandelt wird. Auch die dynamische Netzwerktopologie bereitet einige Schwierigkeiten, weil jeder Knoten beliebig hinzukommen oder aus dem Netz ausscheiden kann. Diese

Eigenschaft stellt das Routing vor ganz neue Probleme, was zum Teil das Entwerfen neuer Algorithmen erforderlich macht. Um die Performance in MANETs zu verbessern, gibt es vor allem einen Ansatz, der auf einer leicht veränderten Version des Transmission Control Protocols basiert. In einem weiteren Kapitel wird auf die Dienste in MANETs eingegangen und beschrieben wie diese Dienste erkannt werden können. Auch in MANETs spielt die Sicherheit eine wesentliche Rolle und darf bei den Herausforderungen nicht vergessen werden. Der Abschluss dieses Kapitels zeigt die Problematiken der Energieeffizienz, der Skalierbarkeit und der eindeutigen Adresszuweisung auf.

1.4.1 Quality of Service

Quality of Service (QoS) oder auf Deutsch Dienstgüte, bezeichnet im Allgemeinen das ordnungsgemässe Funktionieren eines Telekommunikationsnetzes. Dazu bestimmt man Fehlerparameter die ständig überwacht werden und die im Notfall als Ausgangslage für Reparaturmassnahmen verwendet werden. Alle Qualitätsmerkmale können zur Dienstgüte beitragen, die aus der Sicht des Nutzers wichtig für einen bestimmten Dienst sind. Das United Nations Consultative Commitee for International Telephony and Telegraphy (CCITT) definierte QoS wie folgt: "The collective effect of service performance which determines the degree of satisfaction of a user of the service" [9].

"Im Falle des Internets gibt es bereits einige Möglichkeiten die Dienstgüte zu bestimmen, doch ist es nicht möglich diese Lösungen auch auf MANETs zu übertragen" [10]. Dies liegt zu einem grossen Teil an der geringen Bandbreite der Geräte in MANETs und deren beschränkten Energiereserven, was völlig neue Methoden zur Dienstgütebestimmung nötig macht.

Modelle für QoS

Um die QoS für MANETs genauer zu beschreiben wurden Modelle entwickelt, welche über das Best-Effort-Prinzip hinausgehen, das bisher existierte. Die spezifischen Eigenschaften von MANETs wie beispielsweise die dynamische Topologie werden vermehrt in Betracht gezogen. Es gibt zwei unterschiedliche Modelle, die für das Internet von der Internet Engineering Task Force entwickelt wurden und die im Folgenden kurz vorgestellt werden.

IntServ

IntServ wird in den RFCs 1633, 2212 und 2215 [11] dokumentiert. Die Idee basiert darauf, auf einem per-flow Management, wo flussspezifische Zustandsinformationen direkt in den Routern gespeichert werden, das heisst die Router enthalten Informationen über beispielsweise Bandbreite oder Verzögerung. Es gibt zwei Service-Klassen in IntServ: Guaranteed-Service und Controlled-Load-Service. Im ersten Fall werden der Anwendung Garantien geboten was die Bandbreite und die Verzögerung betrifft. Beim Controlled-Load-Service wird der Anwendung ein erweiterter und zuverlässiger Best-Effort-Service geboten [10].

Die Implementation von IntServ erfolgt mittels vier Komponenten:

- einem Signalisierungsprotokoll
- der Zugangskontrolle
- dem Classifier und
- dem Paket-Scheduler.

Das Resource Reservation Protocol (RSVP) wird als Signalisierungsprotokoll verwendet. Damit können Anwendungen Ressourcen reservieren, um während der Übertragung darauf zuzugreifen. Kontrolliert wird diese Vergabe der Ressourcen von der Zugangskontrolle. Entlang des Weges überprüft jeder Router unabhängig, ob er genügend Ressourcen zur Verfügung hat. Der Classifier sorgt für die Identifizierung der Pakete und der Paket-Scheduler ist dazu da, um die Einhaltung der Parameter zu kontrollieren.

Zwei Nachteile ergeben sich somit bei IntServ: es ist kompliziert und die Skalierbarkeit ist schlecht. Für MANETs scheint IntServ nicht geeignet zu sein, da jeder Knoten in seinem ohnehin schon beschränkten Gerät noch etliche Zustandsinformationen speichern müsste. Ein zweites Problem würde sich durch die Mobilität ergeben, da ständig neue Reservierungen gemacht werden müssten, wäre der Protokoll-Overhead von RSVP enorm hoch [10].

DiffServ

In den RFCs 2474 und 2475 [11] wird nun ein weiteres Modell beschrieben. Dieses nennt sich Differentiated Services und setzt die Priorisierung der Datenpakete fest. Da ein mobiles Netzwerk nicht über die gleichen Kapazitäten verfügt wie ein herkömmliches IP-Netzwerk, müssen die Datenpakete nach ihrer Wichtigkeit sortiert werden, da ansonsten etliche Engpässe auftreten würden. Diese Priorität wird bereits vom Sender bestimmt, was einen Unterschied zu IntServ darstellt. Die Entscheidung über die Weiterleitung des Pakets wird alleine von den Routern auf dem Weg zum Empfänger vorgenommen. Diese stützen sich auf die Angaben des Senders. Um zu verhindern, dass einfach jeder seinen Paketen die höchste Priorität zukommen lässt werden an den Übergängen zu aktiven Netzwerkkomponenten sogenannte “Trust Boundaries” eingeführt, wo der Administrator selber überprüfen kann ob er den DiffServ Einstellungen Glauben schenkt oder nicht. Solche “Trust Boundaries” sind typischerweise die LAN Switchports, an welchen die Endgeräte angeschlossen sind.

Beide Modelle sind alleine jedoch nicht geeignet für MANETs. IntServ braucht viel zu viele Ressourcen, die in einem MANET nicht bereitgestellt werden können und DiffServ macht vor allem das Routing im Kern des Netzwerks einfacher, weil in einem MANET jedoch nicht genau bestimmt werden kann, was alles zum Kern gehört, kann auch dieses Modell nicht vollständig übernommen werden.

Die Lösung besteht in einem Modell, welches beide Ansätze verknüpft. Das Flexible QoS Model for MANETs “bietet die Möglichkeit von per-flow Management wie IntServ, als

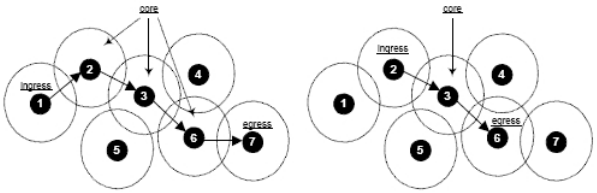


Abbildung 1.1: Dynamische Topologie eines MANETs [9]

auch Serviceklassen vergleichbar zu denen von DiffServ. Dabei werden Daten mit der höchsten Priorität per-flow gehandhabt, alles andere durch Serviceklassen abgedeckt. Es wird von der Annahme ausgegangen, dass nur ein kleiner Teil der Datenströme die höchste Priorität braucht. Ansonsten würden die Probleme von IntServ auftreten, also der Aufwand um jeden Strom einzeln zu verwalten zu groß werden." [10] FQMM unterteilt die Knoten in drei Arten: ingress nodes, core nodes und egress nodes. Also Ausgangsknoten, welche die Daten versenden, innere Knoten, welche die Daten weiterleiten und Endknoten, welche die Daten empfangen. Jedes Gerät kann dabei jede Position übernehmen, je nach dem wie sein Standpunkt im Netzwerk gerade ist. Im Bild 1.1 ist zu erkennen, dass zuerst der Knoten 1 den ingress-Knoten darstellt, 2, 3 und 6 bilden die core-Knoten und 7 ist der egress-Knoten. Im rechten Teil von Bild 1.1 wird der Knoten 2 zum ingress-Knoten, da er nun Daten sendet, Knoten 3 ist immer noch ein core-Knoten und der egress-Knoten ist nun der Knoten 6. Es soll gezeigt werden, wie sich die Geräte anpassen müssen und je nach Position im Netzwerk eine andere Funktion erhalten.

Jedoch ist auch FQMM noch nicht ganz ausgereift, beispielsweise können nicht beliebig viele Anwendungen das per-flow Management benutzen, da die gleichen Probleme der Skalierbarkeit auftreten würden wie bei IntServ. Dieses Modell stellt einen ersten Versuch dar, die Probleme von QoS in MANETs zu beheben, löst diese Probleme aber noch nicht ganz.

1.4.2 Netzwerktopologie

Die Netzwerktopologie in MANETs stellt eine der grössten Herausforderungen dar. Die Knoten sind mobil und ändern ihre Position ständig, was zur Folge hat, dass auch die Topologie immer in Bewegung ist. "Der Ansatz aus klassischen Netzen, dass eine einmal aufgebaute Verbindung auf absehbare Zeit verfügbar bleibt, trifft hier nicht zu. Es kann sich jederzeit ein Knoten aus der Reichweite eines Anderen bewegen." [10] Die Knoten müssen über Zwischenknoten miteinander kommunizieren, da sie durch die begrenzten Ressourcen auch eine limitierte Reichweite haben. Wie in der Abbildung 1.2 gezeigt wird kann Knoten A nicht direkt mit Knoten C kommunizieren.

Die Kommunikation zwischen A und C kann nur über den Knoten B stattfinden. Dazu benötigt A zuerst einmal die Zusicherung, dass die Daten von B an C weitergeleitet werden können. Wenn nun plötzlich auch B nicht mehr in der Lage ist, die Pakete weiterzuschicken, muss A über diesen Zustand informiert werden, da A sonst immer weiter

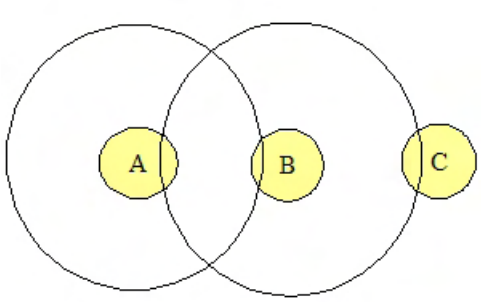


Abbildung 1.2: Begrenzte Reichweite der Knoten

versuchen wird, auf diesem Weg mit C zu kommunizieren. Um die Kommunikation in MANETs zu ermöglichen und die Knoten zu informieren über welche Zwischenknoten sie ihre Daten senden können, müssen Protokolle her, die das Routing regeln.

Einen sehr interessanten Ansatz bieten Anders Lindgren und Olov Schelén, zwei schwedische Wissenschaftler. In ihrem Paper über "infrastructured ad hoc networks" beschreiben sie die Möglichkeit gewisse Knoten in einem Netzwerk fix zu installieren. Ihrer Meinung nach könnten mobile Netze für militärische Zwecke, Katastrophenhilfe oder für Konferenzen gebraucht werden. Doch bei einer Konferenz zum Beispiel wäre der Aufwand eines festinstallierten Knotens gering und würde den Teilnehmern ermöglichen auch mit anderen hosts im Internet zu kommunizieren. Ausserdem sehen auch sie das Problem der fehlenden Anreize für Knoten die Pakete anderer Knoten weiterzuleiten. Mit ihrem Vorschlag eines hybriden Netzes würde der Vorteil von MANETs beibehalten werden ohne dass andere Knoten die Weiterleitung übernehmen müssen. So können grössere Gebiete abgedeckt werden ohne dass die Verantwortung des Routings und der Paket-Weiterleitung auch bei den Knoten liegt. Lindgren und Schelén führen den Begriff von "pseudo base station" (PBS) ein, welche eine ständige Energiequelle haben, nur beschränkt mobil sind und welche sich um den Hauptteil des Routings und der Weiterleitung kümmern. Der Hauptunterschied zu gewöhnlichen Knoten besteht in der eingeschränkten Mobilität. Diese PBSs müssen nicht mobil sein und können deshalb beispielsweise am Strom angeschlossen werden. Somit erfahren sie auch kein Energiedefizit. Die Abdeckung gewisser Gebiete würde realisiert werden indem solche PBSs an strategischen Punkten eingerichtet würden, wo sie beispielsweise eine Konferenz vollständig versorgen könnten. Um diesen Ansatz zu realisieren werden spezielle Routing-Protokolle benötigt. Lindgren und Schelén nennen ihr Protokoll ISAIAH [34]. Auf den genauen Mechanismus dieses Protokolls wird jedoch nicht weiter eingegangen. Die möglichen Routing-Mechanismen werden nun im folgenden Kapitel vertieft behandelt.

1.4.3 Routing

Ein grosser Teil der MANET-Forschung wird im Bereich des Routings getätigt. Die herkömmlichen Routing-Protokolle basieren alle auf einer fixen Topologie und können daher in MANETs nur beschränkt eingesetzt werden. Die Knoten in einem solchen Netz sind meist in Bewegung und verändern somit die Topologie permanent. Routingpfade die gerade noch funktioniert haben, können einige Zeit später nicht mehr existieren und es muss für

die Datenpakete ein ganz neuer Weg gesucht und gefunden werden. Diese Seminararbeit wird nicht die Details der Routing-Protokolle für Mobile Ad hoc Netzwerke beschreiben, sondern lediglich einen Überblick über die aktuellen Ansätze in der Forschung geben und die Probleme aufzeigen, mit welchen die Forscher noch zu kämpfen haben. Die detaillierte Funktionsweise solcher Protokolle kann in [12] nachgelesen werden.

Protokolle lassen sich in zwei Gruppen aufteilen. Zum einen sind dies die proaktiven und zum anderen die reaktiven Protokolle. Letztgenannten Protokolle kennen nicht zu jedem Zeitpunkt die Routen zu allen anderen Stationen. Will ein Knoten ein Paket zu einem anderen senden, muss zuallererst ein Weg zum Ziel gefunden werden. Dies geschieht meist durch Fluten des Netzwerks. Daher verzögert sich das Ankommen des ersten Pakets um einige 100 Millisekunden [13]. Der Vorteil hingegen ist, dass die Routing-Information stets aktuell ist, was bei dynamischen Netzen sehr wichtig ist. Proaktiv ermittelte Routen, also anhand einer Routingtabelle, sind hingegen nach kurzer Zeit nicht mehr aktuell und daher nicht mehr zu gebrauchen. Vor allem aus diesem Grund werden in MANETs reaktive Algorithmen bevorzugt.

Bis heute sind noch nicht sehr viele reaktiven Protokolle bekannt, da die meisten der heute gebräuchlichen Netzwerke, aufgrund ihrer fixen Topologie, auf dem proaktiven Ansatz beruhen. In Simulationen hingegen wurden schon einige reaktiven Protokolle getestet und überzeugen mit sehr guten Ergebnissen [13]. Im praktischen Bereich wurden hingegen noch nicht sehr viele getestet. Dies kommt daher, dass es für die Forscher viel einfacher ist ein Netzwerk mit einigen hundert Knoten zu simulieren, als ein solches exakt nachzubauen. Dieser Ansatz wird jedoch von einigen Leuten kritisiert. Die Kritiker bemängeln, dass in diesen Simulationen nicht alle wichtigen Faktoren, wie zum Beispiel dicke Mauern zwischen zwei Sendestationen oder auch metallische Gegenstände, welche die Funkwellen beeinflussen, mitberücksichtigt und Annahmen getätigt werden, die nicht der Wirklichkeit entsprechen. Kotz et al. [14] stellten eine Liste von Annahmen zusammen, welche in Simulationen oft zur Vereinfachung angenommen werden, in Wirklichkeit jedoch einen grossen Einfluss auf das Resultat haben. Diese Axiome sind:

- Axiom 0: The world is flat
- Axiom 1: A radio's transmission area is circular
- Axiom 2: All radios have equal range
- Axiom 3: If I can hear you, you can hear me (symmetry)
- Axiom 4: If I can hear you at all, I can hear you perfectly

Axiom 0 beschreibt den Sachverhalt, dass nicht alle Knoten auf der gleichen Höhe stehen müssen. In einem Gebäude, kann es vorkommen, dass zum Beispiel zwei Router in der genau gleichen Ecke installiert wurden, jedoch auf verschiedenen Stockwerken liegen. Liegt nun ein Knoten im Keller und einer auf dem Dach, kann sich ein mobiles Gerät, welches sich im Erdgeschoss befindet, zum Beispiel nur mit dem Knoten im Keller verbinden, da der Knoten auf dem Dach zu weit weg und durch die Stockwerke dazwischen, abgeschirmt ist. Die meisten heute gebräuchlichen Simulationen berücksichtigen jedoch diesen Sachverhalt nicht.

Bei Axiom 1 wird angenommen, dass die stärke des Funksignals, welches von einem Knoten ausgesendet wird, in gleicher Entfernung zu diesem Knoten immer gleich stark ist. Jedoch ist dies nicht der Fall, wie Kotz et al. [14] in einigen Experimenten herausgefunden haben. Gründe hierzu sind Hindernisse oder Störsignale, welche die Funkwellen abschwächen.

Axiom 2 widerlegt den Sachverhalt, dass alle Funkquellen eine gleich grosse Reichweite haben. Unterschiede in der Distanz können durch Hindernisse, wie Bäume und Häuser, zustande kommen.

Das Axiom 3 zeigt auf, dass es zwischen zwei Knoten auch unidirektionale Verbindungen geben kann und man diese daher auch in den Simulationen berücksichtigen muss. Die Gründe hierfür sind wie bei den vorhergehenden Axiomen bei Hindernissen und Störsignalen zu suchen.

Das letzte Axiom beschreibt den Sachverhalt, dass wenn zwei Knoten in Reichweite zueinander stehen, nicht unbedingt auch eine Kommunikation stattfinden kann. Die Signalqualität kann aufgrund von Störungen rapide absinken oder die Kommunikation kann wegen Framelosts oder Errors nicht möglich sein.

Einen weiteren Nachteil von reaktiven Protokollen gegenüber den proaktiven ist, dass diese nicht genügend Zeit und Ressourcen haben, den effizientesten oder schnellsten Weg zu suchen. Um die Linkeigenschaften zu bestimmen und zu optimieren, müssten mehrere Pakete verschickt und ausgewertet werden. Dies würde die ohnehin schon grosse Latenz zu Beginn des Datenaustauschs noch weiter erhöhen. Proaktive Protokolle hingegen haben hierfür genügend Zeit und können diese Eigenschaften in den Routingtabellen mitberücksichtigen.

Im Bereich der MANETs müssen also Routing-Protokolle eingesetzt werden, die mit der sich dauernd verändernden Topologie zurecht kommen, multihopfähig und energieschonend sind und dennoch einen effizienten Weg zum Ziel finden. Grosse Probleme bereitet den Forschern die zwei letztgenannten Eigenschaften, da sich diese zum Teil widersprechen. Aufgrund der Multihopfähigkeit des Netzwerkes müssen die Geräte auch Transitverkehr verarbeiten, wenn zwei Knoten nicht direkt benachbart sind und Daten austauschen möchten. Dies benötigt zusätzlich Energie, welche ohnehin schon recht rar ist. Um dieses Problem zu lösen, gibt es Routing-Algorithmen, welche den aktuellen Energiestand der einzelnen Geräte miteinbezieht und wenn möglich die Daten über andere Geräte leitet, die noch über mehr Energiereserven verfügen.

Nach [15] gibt es kein Routing-Protokoll, das für alle denkbaren Netzwerke gleich gut funktioniert. Der Algorithmus muss auf die vorherrschende Grösse des Netzes, wie auch auf den Abstand der Knoten zueinander, also der Dichte des Netzwerkes, abgestimmt werden. Es ist daher denkbar, dass verschiedene Protokolle in verschiedenen Netzen zum Einsatz kommen. Für den Fall, dass zwei Knoten aus verschiedenen Netzen miteinander kommunizieren möchten, müssen diese Protokolle jedoch zueinander kompatibel sein.

Um zu bestimmen, wie gut ein gewählter Weg im Routing, im Gegensatz zu den Alternativen ist, werden sogenannte Metriken benützt. Dies sind Werte, anhand deren man ablesen kann, welche Route zum Ziel besser ist. In [1] sind wünschenswerte Metriken aufgelistet, die die Protokolle von MANETs berücksichtigen sollen. Man kann sie in einen qualitativen und in einen quantitativen Bereich unterteilen. Die qualitativen Metriken sind:

- Distributed operation: In einem MANET gibt es keine zentrale Einheit, die für das Netzwerk verantwortlich ist, wie zum Beispiel ein DHCP-Server. Alle Knoten sind gleichberechtigt und für das funktionieren des Netzwerkes verantwortlich.
- Loop-freedom: Es muss verhindert werden, dass Pakete endlos in Netz umher kreisen. Eine Möglichkeit dies zu verhindern ist eine TTL (Time to live) einzuführen.
- Demand-based operation: Der Routing-Algorithmus soll sich bei Bedarf dem aktuellen Verkehrsaufkommen anpassen und Parameter wie Bandbreite und Energiereserven mitberücksichtigen.
- Proactive operation: Reaktive Protokolle haben den Nachteil, dass sie eine grosse Latenz beim Senden des ersten Pakets aufweisen. Wenn es energetisch und von der Bandbreite her zulässig ist, können auch proaktive Ansätze eingesetzt werden.
- Security: Das MANET sollte genügend gegen Attacken von aussen wie auch von innen abgesichert sein.
- "Sleep" period operation: Um Energie zu sparen, sollten die Geräte einen Sleep-Modus besitzen, welcher das Protokoll aktivieren kann, wenn das Gerät gerade nicht gebraucht wird. Von Vorteil ist, wenn sich die Geräte untereinander absprechen könnten und nur in den Sleep-Modus verfallen, wenn genügen andere Geräte in der Nähe aktiv sind um den Transitverkehr zu bewältigen.
- Unidirectional link support: In mobilen Netzwerken können durch die Dynamik des Netzwerkes unidirektionale Verbindungen entstehen. Aus diesem Grund sollten die MANET Routing-Algorithmen mit diesen Verbindungen umgehen können.

Die quantitativen Metriken, anhand deren die Protokolle eines MANET gemessen werden sollen, sind:

- End-to-end data throughput and delay: Der Throughput gibt an, wie viele Daten pro Zeiteinheit verarbeitet werden können. Delay steht für die Verzögerung.
- Route Acquisition Time: Gibt an, wie lange es dauert bis nach einem Request ein Weg zum Zielknoten gefunden wurde.
- Percentage Out-of-Order Delivery: Dieses Mass gibt die prozentuale Anzahl der Pakete an, die in einer anderen Reihenfolge ankommen, als sie abgeschickt worden sind. Dies geschieht dadurch, dass nicht alle Pakete den gleichen Weg zum Ziel hin nehmen.
- Efficiency: Dieser Messwert gibt an, wie effizient der zur Verfügung stehende Kanal zur Übertragung mit Nutzdaten genutzt wird, im Verhältnis zu den Kontrolldaten.

Es gibt also eine grosse Auswahl an verschiedenen Protokollen für das Routing. In [16] und [17] sind die wichtigsten heute gebräuchlichen Algorithmen beschrieben. Um einen groben Überblick zu erhalten, werden nachfolgend aus dieser Auswahl einige Protokolle kurz beschreiben. Abbildung 1.3 zeigt die wichtigsten Protokolle unterteilt nach den zwei

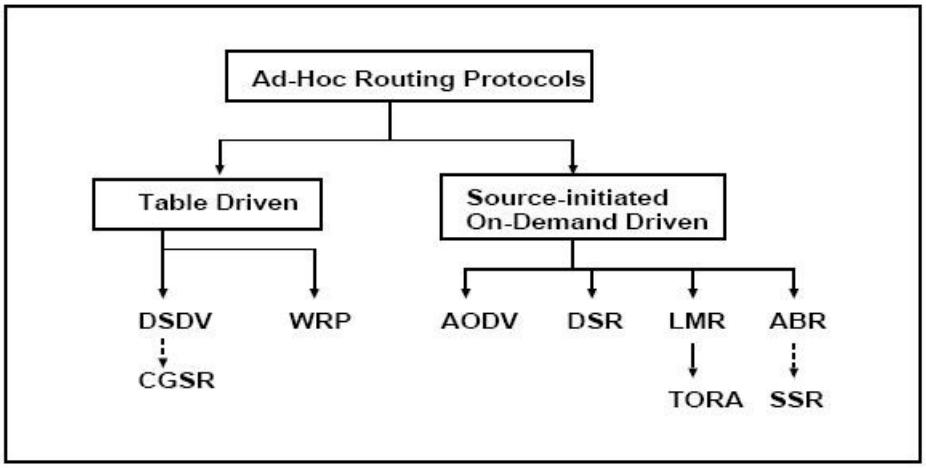


Abbildung 1.3: Übersicht der heute gebräuchlichen Routing-Protokollen [17]

Bereichen proaktiv und reaktiv. Aus dem Gebiet der proaktiven Protokolle hat der Destination Sequenced Distance Vector (DSDV) gute Ergebnisse geliefert. Dieser Algorithmus ist eine Weiterentwicklung des Distance Vector Routing für den Gebrauch in MANETs. Die Idee ist das Suchen des kürzesten Weges von einem Startknoten zum Ziel. Dabei wird als Metrik die Anzahl der Hops, also die Anzahl der Teilstrecken innerhalb einer Route, gemessen. Wie für proaktive Protokolle üblich, besitzen alle Knoten eine Tabelle, welche den kürzesten Weg zu allen anderen Knoten aufzeigt und über welchen Nachbarknoten sie dabei routen müssen. Für den Einsatz in einem ad hoc Netz wird nun noch zusätzlich eine Sequenznummer berücksichtigt. Diese gibt an, wie aktuell die Route ist. Bei den regelmässigen Updates zwischen den Knoten muss dieser Wert miteinbezogen werden. Empfängt eine Station eine Zeit lang kein Update mehr von seinem Nachbarn, löschte es diesen aus der Tabelle.

Bei den reaktiven Protokollen gibt es eine grössere Auswahl, die in MANETs zum Einsatz kommen. Zu nennen sind hier der Ad hoc On-Demand Distance Vector (AODV) und das Dynamic Source Routing (DSR). Diese zwei Ansätze sind sich sehr ähnlich und suchen, im Gegensatz zum DSDV, erst nach einer Route, wenn sie tatsächlich gebraucht wird. Sucht ein Knoten nach einer Route, schauen beide Protokolle zuerst in ihrer eigenen Tabelle nach, ob sie den Zielknoten und den Weg zu diesem von früher kennen. Zu erwähnen ist hier, dass die Tabellen in den Knoten nach einiger Zeit für ungültig erklärt werden. Damit wird der schnelllebigen Struktur eines MANETs Rechnung getragen. Ist also ein gesuchter Knoten nicht vorhanden, wird ein Paket an die Nachbarn gesendet. Diese wiederum schauen in ihren Tabellen nach und senden ein Paket weiter an ihre Nachbarn, wenn sie den Weg zum Ziel auch nicht kennen oder schicken ein Paket zurück, wenn sie den Weg zum gesuchten Knoten kennen. Den Rückweg findet das Paket durch einen temporären Eintrag in den Transitknoten, welcher beim Hinweg hinterlassen wurde. Im Unterschied zum AODV muss der Rückweg beim DSR nicht gezwungenermassen die gleiche Route, wie beim Hinweg nehmen. Bei diesem Verfahren wird der Rückweg wie der Hinweg wieder neu gesucht. Der Vorteil an diesem Protokoll ist, dass so auch unidirektionale Verbindungen zulässig sind. Ein weiterer wichtiger Unterschied liegt in der Art, wie die zwei

Protokolle ihren Weg speichern. Beim DSR wird der ganze Pfad im Header des Pakets gespeichert. Knoten, die ein Paket weiterleiten, können so aus dem Header neue Informationen über die Netzwerktopologie gewinnen. Dies reduziert in stark ausgelasteten Netzen das Datenaufkommen.

Als letzter Algorithmus soll noch der Temporally Ordered Routing Algorithms (TORA) erwähnt sein. Auch hier wird die Pfadsuche nur durchgeführt, wenn ein Knoten ein Paket senden möchte. Die Idee ist, dass jedem Knoten eine bestimmte Höhe im Netzwerk gegeben wird. Die Daten können dann nur von einem höher gelegenen Knoten zu einem tieferen wandern. Dies ist mit einem Kanalsystem vergleichbar. Bei der Wegwahl werden für einen bestimmten Zielknoten mehrere Wege bereitgestellt. Der Vorteil von TORA ist, dass dieser Algorithmus in stark dynamischen Netzwerken eingesetzt werden kann und im Vergleich zu anderen vergleichbaren Algorithmen, wenig Kontrollnachrichten benötigt. Diese Eigenschaft wird dadurch erreicht, dass nur dort Kontrollnachrichten versendet werden, wo sich auch die Topologie des Netzes verändert hat [16]. Aus diesem Grund müssen die einzelnen Knoten all ihre Nachbarn kennen. Der Nachteil dieses Algorithmus ist, dass die Uhren der einzelnen Geräte exakt aufeinander abgestimmt sein müssen. Ist dies nicht der Fall, leidet die Effizienz stark.

1.4.4 Performance

Die Performance von mobilen ad hoc Netzwerken hängt von etlichen Faktoren ab, wie zum Beispiel der Geschwindigkeit oder der Dichte der Knoten. Da jeder Knoten die Funktion eines Routers erfüllen muss, wird er somit auch gebraucht um die Pakete anderer Knoten weiterzuleiten. Man braucht keine zentrale Instanz und dadurch, dass sich die Knoten unabhängig voneinander bewegen, verändert sich die Netzwerktopologie ständig.

TCP, das Transmission Control Protocol, nimmt in der Transportschicht im Internet eine sehr wichtige Rolle ein. Doch auch in schnurlosen Netzwerken sollte TCP gut funktionieren, beispielsweise um Dateien zu verschicken. Forschung, die auf diesem Gebiet gemacht wurde, zeigt aber, dass die üblichen TCP Kontrollmechanismen für schnurlose Netzwerke nicht adäquat sind. Weil TCP ursprünglich für physisch vernetzte Netzwerke konzipiert wurde, nimmt dieses Protokoll an, dass wenn immer ein Paket verloren geht, ein Stau der Grund dafür ist. Aufgrund dieser Vermutung veranlassen die Kontrollmechanismen bestimmte Aktionen, wie beispielsweise die Grösse des Sendefensters zu verringern.

Doch bei mobilen Netzwerken sind meist andere Ursachen Schuld, wenn ein Paket verloren geht. Die Verbindung könnte zum Beispiel unterbrochen worden sein, weil ein Gerät ausser Reichweite gelangte. So gehen sowohl Bestätigungen wie auch Datenpakete verloren, jedoch nicht unbedingt aufgrund eines Staus. TCP interpretiert aber auch diese Fehler als staubedingt und reagiert falsch, was dazu führt, dass die Performance unnötig darunter leidet. Pakete, die nicht in der richtigen Reihenfolge ankommen, stellen einen weiteren typischen Fehler in mobilen Netzwerken dar. Einen solchen Fehler nennt man out-of-order-Fehler.

Um aufzuzeigen, wieso die Performance so sehr darunter leidet, muss genauer auf den Slow Start Mechanismus eingegangen werden. Solange keine Paketverluste auftreten, steigert

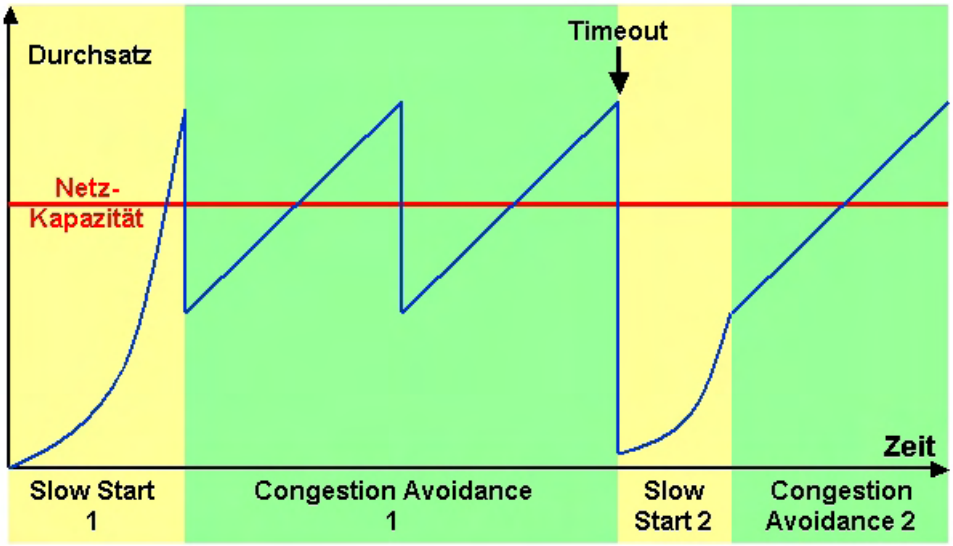


Abbildung 1.4: Slow Start-Mechanismus in TCP [35]

der Sender die Übertragungsrate exponentiell. Der Empfänger schickt eine Bestätigung für die erhaltenen Pakete und wenn diese positiv ist, wird der Sender die doppelte Anzahl Segmente schicken. Sobald die Kapazität aber überschritten wird, beispielsweise durch einen Paketverlust, halbiert der Sender die Fenstergrösse. "Wenn keine aussergewöhnlichen Stausituationen auftreten, verbleibt TCP in diesem Modus."[18] Die Übertragungsrate wird dann linear gesteigert, bis der Empfänger wieder überlastet ist. Sobald aber ein Timeout auftritt, bedeutet dies, dass keine Segmente mehr beim Empfänger ankommen oder alle Bestätigungen verloren gegangen sind, was auf eine ernste Stausituation hindeutet. Nun fängt TCP wieder von vorne an und zwar mit einem Slow Start. Die Übertragungsrate wird nur bis in die Hälfte des Fensters beim Timeout exponentiell gesteigert, ab da wird das Sendefenster wieder linear erhöht.

Feng Wang und Yongguang Zhang [37], zwei amerikanische Wissenschaftler, schlagen nun vor, dass TCP seinen Zustand einfrieren soll, sobald es merkt, dass eine Verbindung unterbrochen ist. Sobald dann eine neue Route gefunden wurde, kann TCP seinen Dienst wieder aufnehmen. Dies benötigt aber ein ständiges Feedback von allen Knoten zu TCP und von der Netzwerk-Schicht zur Transport-Schicht, was wiederum ziemlich schwierig zu implementieren ist. Die zwei Wissenschaftler zeigen mit ihrer Forschung, dass die TCP Performance erheblich gesteigert werden kann, ohne jedoch ständiges Feedback zu benötigen. Ihre Idee beruht darauf, dass out-of-order-Fehler als solche erkannt werden und die Grösse des Sendefensters nicht halbiert wird. Zu solchen Fehlern kann es kommen, wenn Paket 2 einen anderen Weg einschlagen musste, zum Beispiel aufgrund einer Routenänderung, und vor Paket 1 am Ziel ankommt. Wenn nun ein solcher Fehler aufgedeckt wird, kann TCP entsprechend reagieren. Sobald der Empfänger erkennt, dass die Reihenfolge nicht mehr korrekt ist, kann er dem Sender ein Acknowledgement schicken, beispielsweise mit einem Bit im Datenstrom, das einen out-of-order-Fehler anzeigt. Daraufhin kann der Sender sofort reagieren. Erstens kann er für einen Moment die Staukontrolle ausschalten und zweitens versuchen, wieder in den Zustand zu gelangen, den er vor der Routenänderung hatte ohne aber die Fenstergrösse zu verkleinern.

Mit diesem Mechanismus konnten Feng Wang und Yongguang Zhang zeigen, dass sich die TCP Performance in MANETs um bis zu 50% verbessern lässt. In ihrem Ansatz benötigt TCP kein Feedback, weder von den unteren Schichten noch vom Netzwerk. Nur die Endpunkte sind in den Prozess involviert. In ihrer Zusammenfassung schreiben sie, dass ein Ansatz welches Feedback erhält genauer wäre, jedoch nicht immer möglich ist in einem ad hoc Netzwerk, da die Geräte nur beschränkte Ressourcen haben und Overhead so weit als möglich vermieden werden sollte. In solchen Fällen könnte man mit ihrer Methode eine erhebliche Steigerung der Performance erreichen.

1.4.5 Dienste

Nicht nur die Daten, die der Sender schicken wollte werden in einem Netzwerk verschickt, sondern auch viel Informationen über Protokolle werden mitgesendet. Somit ist die effiziente Nutzung der Energie, die ein solches Gerät zur Verfügung hat von grösster Wichtigkeit. Das Auffinden von Diensten zur Versendung von Datenpaketen ist eine der Hauptherausforderungen, die es im Bezug auf die effiziente Nutzung von mobilen ad hoc Netzwerken gibt. Vor allem Funkverbindungen spielen eine wichtige Rolle in MANETs, wie beispielsweise Bluetooth oder WLAN. Da die Geräte eine mobiles Netzwerk bilden, dass sich jederzeit wieder verändern kann, kommt ihnen die Aufgabe zu, das Netzwerk selber zu konfigurieren. Dazu gehört zum Beispiel die Vergabe der IP-Adressen oder der Routing-Algorithmus. Dies wird jedoch von den meisten Betriebssystemen auf PDAs selbständig verrichtet [19].

Die einzelnen Geräte eines MANETs wissen weder genau wo sich die anderen Knoten befinden, noch was die anderen Knoten für Fähigkeiten haben. Ausserdem kommt hinzu, dass es ganz verschiedene Modelle gibt, die sich hinsichtlich ihrer Rechenleistung oder Ressourcenverwaltung enorm unterscheiden können. Es gibt viele Varianten, welche Dienste möglich wären. Das Bereitstellen der Musiksammlung für die andern Geräte im MANET wäre eine Möglichkeit. Oder während man irgendwo warten muss, könnte man mit anderen über das Netzwerk spielen [19]. Die Herausforderung liegt aber darin zu erfahren, von wem welche Dienste angeboten werden. Darauf wird im Folgenden näher eingegangen.

Zwei Ansätze existieren um das Auffinden von Diensten in MANETs zu ermöglichen. Der erste Ansatz geht davon aus, dass eine zentrale Station vorhanden ist, die ein Verzeichnis über alle verfügbaren Dienste und deren Anbieter führt. Die könnte eine statische oder dynamische Station sein. Statisch scheint aber für MANETs nicht geeignet zu sein, das sich solche Netzwerke ja sehr spontan bilden. Auch die dynamische Zuteilung ist schwierig zu implementieren, da jedes Gerät informiert werden müsste, welcher Knoten als Server bereitsteht. Dazu wäre ein erheblicher Kommunikationsaufwand von Nöten, welcher das Netzwerk behindern würde [19].

Der zweite Ansatz beruht darauf, dass die einzelnen Knoten ein lokales Verzeichnis haben, in welchem sie Dienstinformationen speichern. Funktionieren würde das so, dass jeder Dienstanbieter in regelmässigen Abständen eine Broadcast-Nachricht verschickt, in welcher er bekannt gibt welchen Dienst er anbietet. Doch auch da besteht das Problem in einem erhöhten Kommunikationsaufwand, obwohl dieser Ansatz für MANETs besser geeignet wäre.

Ein interessanter Ansatz dieses Problem zu bewältigen besteht darin, nur einige Knoten zu fragen was mit einem sogenannten Multicast erreicht werden kann. Dabei gibt es zwei Vorschläge, auf die aber nicht näher eingegangen wird, da sie den Rahmen unseres Themas sprengen würden. Der eine heisst Konark und stellt ein Protokoll bereit, welches das Auffinden und Anbieten von Diensten für die Geräte möglich macht. Der zweite Vorschlag wurde von einem Team der Universität in Maryland entwickelt und heisst "Distributed Service Discovery Protocol". Hier fungieren einige Knoten als Service Broker und ermöglichen eine effiziente Verteilung der Nachrichten über die Dienste, weil sie untereinander vernetzt sind. Der letztere Ansatz bedeutet zwar einen höheren Aufwand, was aber durch eine bessere Performance ausgeglichen wird [19].

Im Falle des ersten Vorschlags, muss man sagen, dass ein erhöhter Kommunikationsaufwand besteht, die Effizienz jedoch trotzdem nicht Schaden nimmt, weil jeder Knoten nur dann eine Nachricht verschickt, wenn er Dienste anbietet die noch nirgends erkannt wurden. So wird das Netz nicht mit unnötigen Nachrichten überflutet, sondern nur mit Aktualisierungen von Diensten die neu hinzukamen.

1.4.6 Sicherheit

Eine der wichtigsten Herausforderungen ist die Sicherheit von mobilen Netzwerken zu gewährleisten, da sie sehr anfällig für jegliche Attacken sind. Weil jedes Gerät auch ein Router ist, kann man nur schwer sogenannte bösartige Knoten ("malicious nodes") erkennen und verhindern, dass sie Schaden anrichten. Dazu kommt das Problem, dass grosse Anreize für Knoten bestehen, "sich nicht an der gemeinsamen Routing-Infrastruktur zu beteiligen, um ihre eigenen Ressourcen zu schonen" [20]. Denn jeder Knoten braucht eigene Energie (Bandbreite, Batterie) um die Datenpakete weiterzuleiten und vertraut darauf, dass die anderen Knoten seine Datenpakete auch weiterschicken. Doch je mehr egoistischer Knoten in einem Netz zu finden sind, desto schlechter ist die Gesamtleistung des MANETs. Auch wenn die Verlockung gross ist, nur eigene Pakete zu verschicken, wird sie die Leistung extrem mindern. Die vier Hauptmerkmale, um die Sicherheit in ad hoc Netzen zu beschreiben, sind laut Hohenberger die folgenden [21]:

- **Authentizität:** Der Kommunikationspartner ist der, den er vorgibt zu sein.
- **Integrität:** Eine allfällige Veränderung der Nachricht soll erkannt und möglichst verhindert werden.
- **Vertraulichkeit:** Niemand soll die Kommunikation abhören können.
- **Verfügbarkeit:** In herkömmlichen Netzen ist dieser Aspekt nicht so wichtig. Doch in MANETs sind alle Knoten aufeinander angewiesen und wenn Knoten ausfallen kann es sein, dass das Auswirkungen auf das gesamte Netz hat.

Um einen besseren Überblick zu erhalten, welche Möglichkeiten von Angriffen es gibt, sind sie hier nochmals erklärt. Man unterteilt die Attacken in zwei Gruppen, aktive oder passive Attacken.

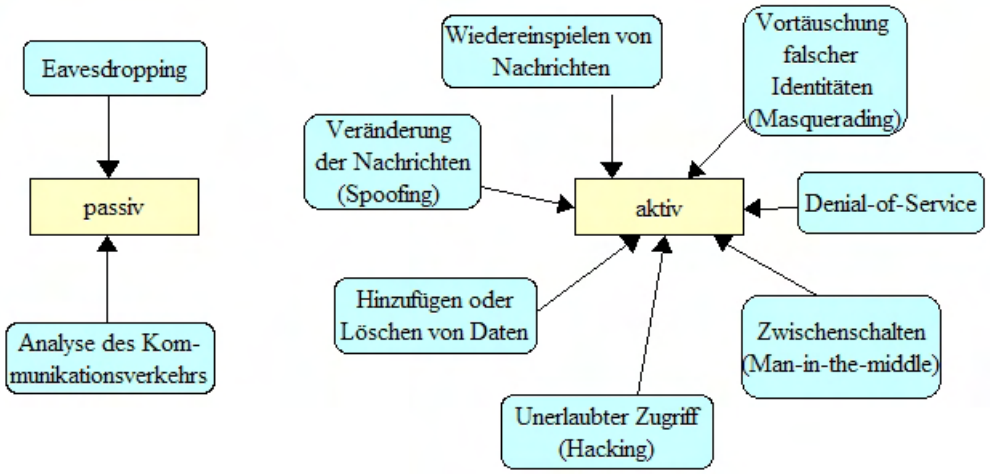


Abbildung 1.5: Aktive und Passive Attacken [22]

Der Gebrauch von drahtlosen Verbindungen ist anfällig auf passive, sowie auch aktive Attacken. Eavesdropping würde es einem Dritten ermöglichen an geheime Informationen heranzukommen und so das Prinzip der Vertraulichkeit zu verletzten. Aber auch aktive Attacken wie zum Beispiel Spoofing wären möglich, wenn man sein Netz nicht sicher schützt. "[..]Das Mithören einer Verbindung [ist] im Funknetz deutlich einfacher. Es ist nicht nötig, eine Leitung zu manipulieren oder einen der Kommunikationspartner oder Router zu kompromittieren, um an die übermittelten Daten zu kommen. Da die meist eingesetzte Funktechnik richtungslos ist, kann ein Angreifer einfach einen kompatiblen Empfänger innerhalb der Funkreichweite aufstellen und unbemerkt die Kommunikation mitschneiden. Auch DoS-Attacken sind hier leichter zu realisieren, da die einzelnen Knoten naturbedingt nur über eine vergleichsweise geringe Bandbreite und Rechenleistung verfügen und somit durch einen geschickten Angriff leicht an ihre Grenzen gebracht werden können. Im Allgemeinen existiert hier auch keine Firewall, da alle Knoten direkt im Netz hängen. Da jeder teilnehmende Knoten auch als Router fungiert, hat theoretisch auch jeder Knoten die Möglichkeit, die über ihn weitergeleiteten Daten zu kopieren oder zu verändern. Dadurch, dass die Routingtabellen der einzelnen Knoten davon abhängen, dass andere Knoten verlässliche Informationen über sich und ihre Umgebung beisteuern, kann ein einzelner manipulierter Knoten durch gezielte Fehlinformationen die Routingtabellen vieler anderer Knoten und damit den Verkehr in einem großen Teil des Netzes stark durcheinander bringen." [21]

Kann man diese Risiken nun minimieren? Ein Weg um Nachrichten effektiv zu schützen besteht darin sie zu verschlüsseln. Dies wiederum zwingt die Benutzer sich auf kryptographische Schlüssel zu verlassen ohne jedoch eine vertrauenswürdige Drittpartei anfragen zu können.

Weil ein solches Netzwerk sehr spontan entsteht, kann man sich nicht darauf verlassen, dass alle Teilnehmer den public key besitzen. Wenn jedoch zwei Teilnehmer schon einmal miteinander kommuniziert haben und eine sicher Verbindung besitzen, können sie diese auf das gesamte Netz ausweiten. Dies würde folgendermassen funktionieren. Knoten A nimmt die Position des Servers ein. A startet den Vorgang in dem er eine Start-Nachricht über das gesamte Netzwerk verschickt. Jeder Knoten, der nun diese Nachricht erhält,

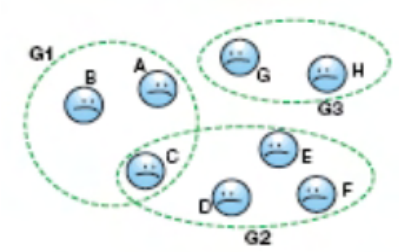


Abbildung 1.6: Drei Gruppen, die nicht miteinander Kommunizieren können [4]

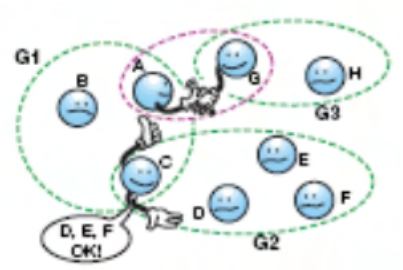


Abbildung 1.7: A erhält die public keys von C [4]

flutet das Netzwerk mit einer Nachricht die seine public keys enthält. A sammelt diese Antworten und identifiziert die sicheren Verbindungen im Netzwerk.

C sendet die public keys die er von D, E und F erhält weiter an Knoten A. Darauf erstellt A eine neue vertrauenswürdige Beziehung zu G.

G sendet nun den public key, den er von H erhält weiter an A.

A flutet nun das Netzwerk mit allen public keys und erstellt so eine neue Gruppe, welche alle Knoten beinhaltet und untereinander sicher kommunizieren kann.

Mit dieser Methode haben alle Knoten einer Gruppe den selben key mit dem sie die Informationen verschlüsseln können. Eine Möglichkeit es allfälligen Lauschern schwerer zu machen, ist der Einsatz von Frequenzsprüngen (frequency hopping). Bei dieser Methode wird die Funkfrequenz in bestimmten Abständen gewechselt. Dies hilft auch Störungen zu beseitigen die entstehen, wenn mehrere Anwendungen gleichzeitig senden wollen. Die Sicherung der Routing-Informationen stellt einen sehr wichtigen Teil dar um ein Netz zu

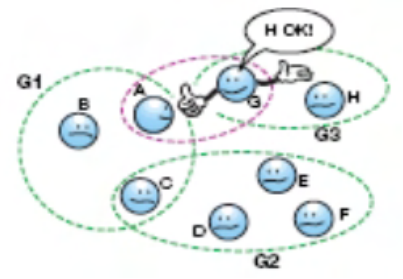


Abbildung 1.8: G sendet public key an A [4]

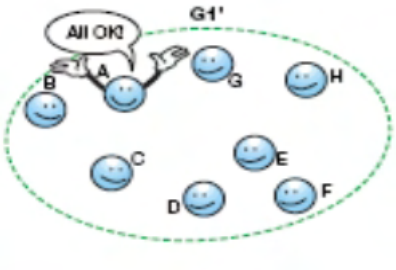


Abbildung 1.9: A flutet Netzwerk mit public keys [4]

schützen. Es sind verschiedene Verfahren möglich, die zum Teil sogar erkennen wenn sich ein Knoten egoistisch verhält.

Das Problem von nicht-kooperativen Knoten stellt eine grosse Herausforderung dar, denn solche Knoten sind nicht ohne weiteres zu erkennen. Eine Möglichkeit bietet der Einsatz eines Watchdogs. "Ein so genannter Watchdog wird eingesetzt, um die Weiterleitung von Paketen im Netz auch dann sicher zu stellen, wenn sich einer oder mehrere Knoten egoistisch oder falsch verhalten. Obwohl sie sich bei der Routenfindung zur Weiterleitung bereit erklärt haben schicken sie Pakete, die über sie geleitet werden sollen, nicht weiter. Dazu muss man davon ausgehen, dass auf jedem kooperativen Knoten ein Watchdog läuft und die Knoten in alle Richtungen gleichzeitig funken, wie es in den meisten Funknetzen üblich ist. Der Watchdog benötigt weiterhin eine Möglichkeit, auch nicht an seinen Knoten adressierten Datenverkehr zu belauschen, was in normalen Funknetzen ebenso gegeben ist. Wenn er nun an einen seiner Nachbarn eine Nachricht verschickt kann er kurz darauf feststellen, ob dieser sie auch weiterleitet, auch wenn er möglicherweise ausserhalb der Funkreichweite des nächsten Knotens auf der gefundenen Route ist. Wenn nicht jeder Knoten die weitergeleiteten Daten neu verschlüsselt kann der Watchdog sogar erkennen, wenn sein Nachbar die Daten verändert hat. Dazu speichert der Watchdog alle von ihm weitergeleiteten Pakete. Wenn er dann erkennt, dass sein Nachbarknoten das Paket korrekt weiterleitet, löscht er das Paket aus seinem Puffer. Geschieht dies innerhalb eines festgelegten Zeitraumes nicht, so erhöht der Watchdog seinen Fehlerzähler für den betreffenden Nachbarn. Übersteigt die Fehlerrate eines Knoten ein Toleranz-Niveau, so sendet der Watchdog eine Warnung an den Sender des Paketes, dass die gewählte Route kompromittiert ist." [21]

Doch auch ein Watchdog bietet keine hundertprozentige Sicherheit. Egoistische Knoten werden aber doch eher erkannt und bösartige Knoten können meist weniger Schaden anrichten. Ein sehr interessanter Ansatz egoistischen Knoten entgegenzuwirken bietet der Einsatz von virtueller Währung. Hier existieren zwei Ansätze wie es egoistischen Knoten schmackhaft gemacht werden kann, die Pakete anderer Knoten doch weiter zu schicken. Beim Packet Purse Model schickt der Sender einen Betrag mit, und jeder Knoten der das Paket weiterleitet darf eine Entschädigung behalten. Problematisch ist hier, dass der Sender schon im Voraus abschätzen muss, wie aufwändig die Übertragung des Pakets sein wird. Wenn er zu wenig Nuglets (so heisst diese Währung) mitgibt, dann erreicht das Paket sein Ziel nicht und das Geld ist verloren. Gibt er zu viel mit, dann bezahlt er die andern Knoten zu hoch, was für den Sender auch nicht befriedigend ist. Um die Verwaltung dieser Nuglets zu regeln wurde der Vorschlag gemacht, dass in jedem Endgerät

eine manipulationssichere Einheit installiert wird.

Die zweite Möglichkeit heisst Packet Trade Model. Der Sender muss nicht mehr im Voraus seinem Paket genügend Geld mitgeben, sondern das Paket wird von den weiterleitenden Knoten gekauft und dann gewinnbringend verkauft. Hier tritt aber schon der erste Nachteil auf, denn der Sender kann beliebig viele Pakete ins Netz schicken, da die Kosten für den Transport vom Empfänger getragen werden. Es gibt noch einen weiteren Ansatz, wo in den einzelnen Knoten ein Zähler installiert ist, der sich erhöht wenn der Knoten ein Paket weiterleitet und sich verringert wenn der Knoten ein Paket schicken will. Das Problem von egoistischen Knoten kann man wohl nur durch ein Anreizsystem in den Griff bekommen. Da jeder Knoten seine eigenen Ressourcen aufwenden muss um Pakete weiterzuschicken, ist die Verlockung sehr gross, diese Ressourcen zu schonen.

1.4.7 Energieeffizienz

Einem beteiligten Gerät eines MANETs stehen für dessen Aktivitäten nur sehr knapp bemessene Energieressourcen bereit. Diese nehmen bei der Benutzung einer drahtlosen Mobilfunktechnik drastisch ab. Somit resultiert ein sehr grosses Energiesparpotential, wenn einzelne Geräte des MANETs in einen Schlafmodus versetzt werden, welche aktuell für die Datenübermittlungen nicht notwendig sind. Dabei muss aber stets beachtet werden, dass keine Knoten ausgewählt werden, welche eine Teilung des Netzwerkes zur Folge hätten. Mittels eines Timers erwachen nach einer bestimmten Zeit die sich im Schlafmodus befindenden Knoten.

Ferner besteht die Möglichkeit, die Energiebestände der einzelnen Geräte zu vergleichen und auf diese Weise die Pakete über diejenigen Geräte zu senden, welche noch über reichlich Energie verfügen. Damit kann die Nutzungsdauer des gesamten Netzwerkes verlängert werden, weil sich die Energiebestände der verschiedenen Geräte auf einem ähnlichen Niveau einpendeln. Die Grösse eines Datenpaketes bestimmt die Zeitdauer, die für das Senden gebraucht wird. Beim Senden und Empfangen von Daten wird jedoch ein Vielfaches mehr an Energie als im Ruhemodus verbraucht. Durch die geschickte Wahl der Route, welche optimalerweise am wenigsten hops und dadurch auch am wenigsten Sende- und Empfangsvorgänge beinhaltet, kann der Energieverbrauch einer Übertragung minimiert werden. Ein Nachteil dieser Routenwahl ist hingegen, dass sich die Energiebestände der einzelnen Knoten dadurch stark voneinander unterscheiden können. Knoten, die häufig für die Weiterleitung von Daten eingesetzt werden, könnten keine Energie mehr haben und würden somit dem Netzwerk nicht mehr zur Verfügung stehen. Diese verschiedenen Möglichkeiten müssen einander gegenüber gestellt werden, da ihre gleichzeitige Realisierung nur bedingt erreicht werden kann. Welche Energiesparansätze bevorzugt und verfolgt werden, hängt auch vom Verwendungszweck des MANETs ab [23].

1.4.8 Skalierbarkeit

Jedes Gerät, welches sich innerhalb der Reichweite eines MANETs befindet, kann sich mit diesem verbinden und es auf Wunsch wieder verlassen. Diese Eigenschaft hat zur Folge,

dass ein MANET theoretisch beliebig gross werden kann, sofern die Dichte und die Verteilung der teilnehmenden Geräte dies zulässt. Ansonsten zerfällt das Netzwerk in einzelne Teilnetzwerke und die Kommunikation zu einigen Endgeräte wird dadurch nicht mehr möglich sein. In grossen Netzwerken können zwar viele Teilnehmer miteinander kommunizieren, jedoch steigt dadurch auch der Anteil fremder Daten, die weitergeleitet werden müssen. Wenn die übliche Verbindung zwischen zwei Endgeräten über mehrere Zwischengeräte verläuft, kann dieser Fremdanteil weit über 50% liegen. Die bereits stark begrenzte Bandbreite eines mobilen Gerätes würde somit vor allem für den Transitverkehr verwendet werden. Hieraus wird auch ersichtlich, dass sich MANETs nicht für Verbindungen eignen, die sich über eine grosse Distanz erstrecken und bei denen eine hohe Datenrate unterstützt werden muss [5].

1.4.9 Adresszuweisung

Damit in einem MANET das Routing von Datenpaketen reibungslos durchgeführt werden kann, müssen eindeutige Adressen den einzelnen Geräten automatisch zugewiesen werden. Da aber keine zentralen Instanzen vorhanden sind und alle Geräte innerhalb eines MANETs gleichberechtigt sind, müssen diese eindeutigen Adressen auf einer verteilten Basis zugeordnet werden können, sobald ein Gerät an einem solchen Netzwerk teilnehmen möchte. Eine besondere Herausforderung stellt in dieser Hinsicht auch das Vereinigen von mehreren Teilnetzwerken zu einem einzigen Netzwerk dar. Infolgedessen könnten einige Adressen im daraus entstandenen Netzwerk mehrfach vorhanden sein und zu falschen Weiterleitungsentscheidungen in den Knoten führen. Einerseits könnten in einem solchen Fall einige Geräte Daten empfangen, die fälschlicherweise zu ihnen weitergeleitet wurden und andererseits könnten auch Daten zwischen mehreren Knoten hin und her gereicht werden bis sie aufgrund ihrer Time to live aus dem Netzwerk entfernt werden.

Ein Verfahren, welches mit allen zuvor geschilderten Schwierigkeiten keine Probleme aufweist, wird PACMAN (Passive Autoconfiguration for Mobile Ad hoc Networks) genannt. Dieses nutzt den bereits vorhandenen Datenverkehr, welcher durch das eingesetzte Routing-Protokoll verursacht wird. Adresskonflikte werden dadurch erkannt, dass bestimmte Anomalien in diesem Datenverkehr auftreten. Die betroffenen Geräte können durch diese Anomalien erkannt und deren Adressen verändert werden. Durch die Nutzung des ohnehin vorhandenen Datenverkehrs wird die begrenzte Bandbreite eines MANETs nicht zusätzlich belastet. Den teilnehmenden Geräten werden mittels PACMAN zudem IP Adressen zugeordnet, welche komprimiert werden können und demzufolge den Verkehr des Routing-Protokolls, bis hin zu einem Faktor 10, verkleinern können. Ausserdem sind die Algorithmen, die zur Erkennung von doppelten Adressen verwendet werden, sehr vielfältig. Daher können diese Algorithmen auf den vorhandenen Routing-Protokollen angewendet werden so das letztere wenn überhaupt nur marginal geändert werden müssen [24].

Wie aus den vorhergehenden Abschnitten ersichtlich wird, gibt es noch einige Herausforderungen, welche zu lösen sind. Im Bereich QoS entstand das FQMM, welches sich den spezifischen Gegebenheiten der MANETs angenommen hat. Dieses ist jedoch noch nicht vollständig ausgereift. Der dynamischen Topologie muss mit neuartigen Routing-Protokollen Rechnung getragen werden. Hierzu eignen sich hauptsächlich die Algorithmen

aus dem reaktiven Bereich. Indem die Grösse des Sendefensters bei einer allfälligen Nichtübertragung nicht reduziert wird, kann mit einer leicht veränderten Version des TCPs die Performance im MANET um bis zu 50% gesteigert werden. Ein interessanter Ansatz besteht im Bereich der Dienstauffindung, wobei einige Knoten als Service Broker fungieren und somit die effiziente Verteilung der Nachrichten über die verfügbaren Dienste ermöglichen. Ein MANET stellt durch den offenen unkontrollierten Zugang grosse Anforderungen an die Sicherheit. Vor allem für die Problematik der egoistischen Knoten müssen noch besser Lösungsansätze entwickelt werden. Die Energiesparmöglichkeiten sind in einem MANET sehr vielfältig und müssen in Abhängigkeit des Verwendungszweckes umgesetzt werden. Nachfolgend wurde im Abschnitt Skalierbarkeit festgestellt, dass sich MANETs für die Kommunikation zwischen entfernten Knoten, welche eine hohe Datenrate erfordern, nicht eignen. Das PACMAN-Verfahren erfüllt die verschiedenen Anforderungen im Bereich der automatischen Adresszuweisung ohne ständig Kontrolldaten zu senden. Der Algorithmus benutzt den ohnehin vorhandenen Verkehr des Routing-Protokolls, um die geeigneten Entscheidungen zu fällen.

Im folgenden Kapitel werden nun einige Anwendungen gezeigt, welche sich in Zukunft wegen dieses neuen Netzwerktyps durchsetzen könnten.

1.5 Anwendungen

Die Technologie MANET wird bis heute noch fast nicht kommerziell eingesetzt. Hierfür müssten noch einige Probleme gelöst werden. Grosse Schwierigkeiten ergeben sich durch die kleine Grösse der mobilen Geräten und der daraus folgenden geringen Energiereserven. Ebenfalls sind noch nicht alle Fragen im Bereich des Routings gelöst. Die in dieser Arbeit vorgestellten Algorithmen bilden sicherlich eine solide Grundlage, können aber noch erheblich optimiert werden. Zu diskutieren ist auch, wie man egoistische Knoten erkennen kann und wie man mit diesen umgehen soll. Auch im Gebiet der Sicherheit gibt es noch einige Probleme, die zu lösen sind. Wie erkennt man böswillige Knoten und wie kann man sich gegen die Attacken dieser Knoten wehren?

Aus den oben genannten Gründen sind bis heute fast nur Anwendungen zu Forschungszwecken bekannt. Grossen Nutzen aus dieser Technologie erhofft sich das Militär, da es im Ernstfall nicht immer auf eine bestehende Infrastruktur zurückgreifen kann. Es ist darum sicherlich nicht verwunderlich, dass viele der bis heute bekannten Applikationen vom Militär in Auftrag gegeben wurden und von ihnen zu Testzwecken verwendet werden. Solche und auch einige Anwendungen aus dem zivilen Bereich sollen nun nachfolgend kurz vorgestellte werden (Anlehnung an [25]).

- Konferenzen: Um Dokumente bei einer Konferenz elektronisch austauschen zu können, muss heutzutage erst ein Netzwerk eingerichtet werden, bei welchem sich die Teilnehmer dann anmelden können. Sie sind also auf eine vorhanden Infrastruktur angewiesen. Mit einem ad hoc Netzwerk könnten die Teilnehmer ihr eigenes Netz aufbauen und so einfach ihre Präsentationen und Dokumenten austauschen.

- Heimnetzwerk: Der Absatz von mobilen Geräten, wie Notebooks, MP3-Player oder auch Handys hat in letzter Zeit rapide zugenommen. Dies ermöglicht uns auch anderswo als von zuhause aus mit dem Computer zu arbeiten. Jedoch muss der Notebook bei jedem Gebrauch in einem neue Netz neu konfiguriert und eingerichtet werden. Ein MANET könnte diesen Aufwand erheblich erleichtern, indem es einem diese Arbeit abnimmt.

- Personal Area Network (PAN): In diesem Gebiet ist der Einsatz von MANETs im zivilen Bereich schon am weitesten Fortgeschritten. Ein PAN ist ein Netzwerk, welches sich in einem Abstand von einigen Metern um eine Person herum befindet. So ist zum Beispiel ein Netzwerk mit einem Handy, Notebook und einem MP3-Player ein solches PAN. Wie im Einführungsbeispiel schon dargelegt, könnten diese Geräte miteinander kommunizieren und zum Wohle des Benutzers Daten austauschen und automatisch abgleichen. Dieser Datenaustausch muss nicht gezwungenermassen zwischen mobilen Geräten erfolgen. Denkbar ist dieses Szenario auch zwischen einem Computer am Arbeitsplatz und dem PDA. Wenn nun die Person mit dem PDA in das Büro eintritt, erkennt dies der Computer und kann zum Beispiel alle elektronische Geräte einschalten und die Termine und Emails abgleichen.

  Möglich ist auch der Einsatz dieser Technologie in Kleidungsstücken. Diese könnten dann automatisch Alarm schlagen, wenn die Farben der einzelnen Kleidungsstücken nicht aufeinander abgestimmt sind. Unterstützen kann uns das MANET auch beim Einkaufen. Dort sind Szenarien wie berührungs- und bargeldloses Bezahlen, automatische Routenführung des Einkaufwagens oder auch elektronische Werbedisplays denkbar. Dies kann nur erreicht werden, indem all diese Geräte miteinander per Funk verbunden werden und das Netzwerk flexibel genug ist, um sich auf Veränderungen einzustellen. In Deutschland wurde ein Einkaufszentrum eröffnet, welches für Versuchszwecke solch neue Technologie im praktischen Alltag ausprobiert. Details sind unter [26] zu finden.

- Emergency/Disaster: Dies ist wohl einer der bekanntesten Einsatzbereiche von einem MANET. Ist in einem Katastrophenfall, zum Beispiel aufgrund eines Erdbebens, die bestehende Infrastruktur zerstört, ermöglicht ein solches Netzwerk dennoch die Kommunikation für die Rettungskräften untereinander. Lebensrettend kann es für eingeschlossene Personen sein, die mithilfe des MANETs Kontakt zur Polizei aufnehmen kann.

- Verkehr: Vernetzt man alle bestehenden Fahrzeuge miteinander, eröffnet dies ein Feld vieler neuen Applikationen. Daten, wie die eigene Position, Geschwindigkeit und dem Zielort, lassen sich mit Staumeldungen von entgegenkommenden Autos verknüpfen und so einen optimale Route berechnen. Mit Geschwindigkeitsreduktionen oder Umleitungen lassen sich bei grossem Verkehrsaufkommen in Zukunft sogar Staus ganz vermeiden. Entgegenkommende Autos können auch vor Gefahren, wie Baustellen, scharfen Kurven oder Glatteis warnen.

- Electronic Dust: Electronic Dust, auch als Elektronische Wolke bekannt, ist eine Ansammlung von kleinster elektronischer Geräte, welche mit einer Menge von Sensoren ausgerüstet sind. Wird ein solcher Schwarm über einem Gebiet abgeworfen und besitzen diese Kleincomputer die Fähigkeit ein ad hoc Netzwerk aufzubauen,

könnten diese ihre Rechenleistung, wie in einem Cluster teilen. Ebenso müssen nicht mehr alle Geräte dieser Wolke die Leistung besitzen, nach aussen zu Kommunizieren und können so Energie sparen. Die Kommunikation könnte von einzelnen leistungsstarken Geräten übernommen werden, die dann als Gateway funktionieren.

- Militärische Nutzung: Die Forschung für die Nutzung eines MANETs ist in diesem Bereich schon am weitesten Fortgeschritten. Das Militär benötigt für seine Zwecke ein Netzwerk, das flexibel ist, auch nach dem Ausfall einiger Geräte noch funktioniert und wenn möglich ohne Infrastruktur zureicht kommt. MANETs erfüllen genau diese Anforderungen mit seiner dezentraler Struktur. Dies erlaubt dem Militär auch in feindlichem Gebiet, ohne Infrastruktur untereinander zu kommunizieren, sofern die benötigten Frequenzen nicht vom Gegner gestört werden.

1.6 Trends

Damit sich eine der zuvor geschilderten MANET Anwendungen erfolgreich etablieren kann, müssen noch einige technische Weiterentwicklungen realisiert werden.

Die heute üblichen Mobilfunktechniken (wie z.B. Wireless LAN (IEEE 802.11), IrDA, GSM, UMTS oder Bluetooth) werden in den aktuellen Endgeräten vermehrt zur Verfügung gestellt. Jedoch sind diese für die Übertragung einer grösseren Datenmenge noch nicht geeignet. Die Übertragung einer gewöhnlichen MP3-Datei zwischen zwei Handys mittels Bluetooth dauert beispielsweise immer noch einige Minuten. Aber auch bei einer Verbindung zum Internet mittels eines GPRS/UMTS Handys kann der Download von einigen Megabytes unangenehm lange dauern. Es gibt bereits eine Weiterentwicklung von UMTS, welche den Namen HSDPA trägt und um ein Vielfaches schneller ist als UMTS. Mit diesem Mobilfunkstandard sind Übertragungsraten von bis zu 1.8 Mbit/s möglich. Der Mobilfunkanbieter Swisscom Mobile bietet HSDPA bereits in einigen Grossstädten der Schweiz an, jedoch gibt es bis anhin noch fast keine Handys, die diesen Standard unterstützen [27].

Die in Abschnitt 1.4.3 beschriebenen Routing-Protokolle geben nur einen kleinen Überblick über die Vielfalt der heute vorhandenen Routing-Protokolle für MANETs. Eine Standardisierung dieser Vielfalt, wie sie mit dem TCP/IP Protokoll für das Internet erlangt wurde, könnte dabei helfen MANETs zu etablieren. Denkbar wäre in dieser Hinsicht auch, dass es in Zukunft mehrere Standardprotokolle gibt, da die Routing-Algorithmen auf die Dichte der Endgeräte innerhalb des Netzwerkes und auf die Grösse des Netzes selbst abgestimmt werden müssen [15].

Die Energiereserven eines mobilen Gerätes, welches aktiv genutzt wird, sind meistens nach weniger als einem Tag aufgebraucht. Bei einem Laptop sind es häufig sogar nur wenige Stunden. Mit der ständigen Nutzung einer Mobilfunktechnik werden diese ohnehin schon knappen Laufzeiten der Akkus noch einmal stark gesenkt. Eine zukünftige MANET Anwendung kann sich somit nur durchsetzten, wenn ihre Operationen energieeffizient durchgeführt werden können. Alternativ könnte auch eine neue revolutionäre Energiequelle die

heutigen Akkus ersetzten und so den Geräten eines MANETs eine längere Einsatzzeit ermöglichen.

Die rasante Entwicklung in der Technik und die Forschungsresultate, die im Bereiche von MANETs in den letzten Jahren erzielt wurden, deuten darauf hin, dass es möglicherweise in einigen Jahren bereits kommerzielle MANET Anwendungen geben wird. Diese Anwendungen werden wohl keinen kontinuierlichen Datenaustausch zwischen zwei Endgeräten benötigen. Es ist wahrscheinlicher, dass erste Anwendungen nur einen gelegentlichen Datentransfer zwischen den Geräten durchführen müssen (wie z.B. die in Abschnitt 1.5 erwähnte Verkehrsanwendung). Falls nun zu einem gewissen Zeitpunkt keine Daten mit entgegenkommenden und überholenden Fahrzeugen ausgetauscht werden können, ist die Hauptfunktionalität, nämlich das Autofahren, nicht davon beeinflusst. Dadurch wird auch ersichtlich, dass solche Anwendungen zu Beginn Zusatzfunktionen übernehmen werden, die den Alltag und dessen Probleme vereinfachen sollen.

Literaturverzeichnis

[1] S. Corson, J. Macker: Mobile Ad hoc Networking (MANET): Routing Protocol Performance Issues and Evaluation Considerations (RFC 2501), http://www.ietf.org/rfc/rfc2501.txt, 02.10.2006.

[2] Friedemann Mattern, Informatik-Lexikon, http://www.io-port.net/ioport2004/content/e45/e383/e598/e599/index_ger.html#dokanfp3, 23.11.2006.

[3] ETSI, http://www.etsi.org/, Online: Frühling 2007.

[4] Ericsson: Wireless ad hoc networking the art of networking without a network, http://www.ericsson.com/ericsson/corpinfo/publications/review/2000_04/files/2000046.pdf, 2000.

[5] Martin Mauve: Mobile ad hoc Netzwerke: Kommunikation ohne Infrastruktur, http://www.uni-duesseldorf.de/home/Jahrbuch/2003/PDF/Mauve.pdf, 2003.

[6] Renate Lechler: Entwicklung eines energieeffizienten Dienstauffindungsverfahrens für mobile ad hoc Netzwerke, ftp://ftp.informatik.uni-stuttgart.de/pub/library/medoc.ustuttgart_fi/DIP-2337/DIP-2337.pdf, 15.11.2005.

[7] Fabian Gremm, Sebastian Zöller: Sicherheit in Ad-Hoc Netzen: Protokolle und Anwendungen, http://www.sec.informatik.tu-darmstadt.de/pages/lehre/SS04/seminar_adhoc/ausarbeitungen/Service_Discovery.pdf, 08.06.2004.

[8] IETF Mobile Ad-hoc Networks (MANET) Working Group: Mobile Ad-hoc Networks (MANET) Charter, http://www.ietf.org/html.charters/manet-charter.html, 02.10.2006.

[9] Zeinalipour-Yazti Demetrios: A Glance at Quality of Services in Mobile Ad-Hoc Networks, http://www.cs.ucr.edu/~csyiazti/courses/cs260/html/manetqos.html, Herbst 2001.

[10] Dennis Gräff: Dienstgüte in mobilen ad hoc Netzen, http://www.ibr.cs.tu-bs.de/courses/ws0203/skm/articles/graeff-qos-adhoc.pdf, 2002/2003.

[11] IETF, http://www.ietf.org/rfc, Online: Frühling 2007.

[12] Daniel Dönni, Daniel Rickert, Andreas Bossard: Routing in Multi-hop Mesh Networks, UniZH, Frühling 2007.

[13] Guido Hiertz, Spiro Trikaliotis: Maschenfunk, ct magazin für computer technik, Heise Zeitschriftenverlag, 09.01.2006.

[14] David Kotz, Calvin Newport, Chip Elliott: The mistaken axioms of wireless-network research, http://pdos.csail.mit.edu/decouto/papers/kotz03.pdf, 18.07.2003.

[15] Jürgen Nagler, Frank Kargl, Stefan Schlott, MichaelWeber: Ein Framework für MA-NET Routing Protokolle, http://medien.informatik.uni-ulm.de/forschung/publikationen/wman02.pdf, Online: Frühling 2007.

[16] Andreas Mühling: Hauptseminar Routing-Protokolle für Adhoc-Netzwerke, http://wwwspies.in.tum.de/MVS/sem06/contents/AdHoc-Routing_Ausarbeitung.pdf, 31.05.2006.

[17] Elizabeth M. Royer, C-K Toh: A Review of Current Routing Protocols for Ad-Hoc Mobile Wireless Networks, http://www.ee.surrey.ac.uk/Personal/G.Aggelou/PAPERS/Adhoc_Review.pdf, Online: Frühling 2007.

[18] Reto Trinkler: TCP/IP in drahtlosen Netzen, http://www.basis06.com/media/pdf/speech_wexpo.pdf, 26.03.2004.

[19] Matthias Bier: Multicast Service Discovery in Mobile Ad Hoc Networks, http://www.inf.fu-berlin.de/inst/ag-tech/teaching/SS06/19554-S-TI/Matthias\%20Bier\%20-\%20Multicast\%20Service\%20Discovery.pdf, Online: Herbst 2006.

[20] Frank Kargl: Sicherheit in mobilen ad hoc Netzwerken, http://medien.informatik.uni-ulm.de/forschung/publikationen/gidisspreis.pdf, 2003.

[21] Daniel Hohenberger: Sicherheit in ad hoc Netzen, http://www.net.in.tum.de/teaching/WS02/security/securityUeb/15ausarbeit.pdf, 31.01.2003.

[22] Prof. Dr. Stiller: Vorlesungsunterlagen zu Kommunikation und verteilte Systeme, http://www.csg.unizh.ch, Frühling 2007.

[23] Suresh Singh, Mike Woo, C. S. Raghavendra: Power-Aware Routing in Mobile Ad Hoc Networks, http://www.cs.pdx.edu/~singh/ftp/mobicom98.pdf, Oktober 1998.

[24] Kilian Weniger: PACMAN: Passive Autoconfiguration for Mobile Ad hoc Networks, http://www.tm.uka.de/doc/2004/autoconf_jsac_epub.pdf, März 2005.

[25] Manuel Beetz, Marcus C. Gottwald: Ad-hoc-Netzwerke und Routing in Ad-hoc-Netzwerken, http://cheers.de/uni/mobkomm/adhoc.pdf, 01.02.2002.

[26] Metro AG: Future Store Initiative, http://www.future-store.org, Online: Frühling 2007.

[27] Felix Raymann: Kaufberatung: Handys für Geschäfts- und Multimedia-Anwendungen, Online PC Zeitung, Neue Mediengesellschaft Ulm mbH, Dezember 2006.

[28] National Institute of Standards and Technology (NIST), Wireless Communications Technology Group: Wireless Ad Hoc Networks Bibliography, http://w3.antd.nist.gov/wctg/manet/manet_bibliog.html, Online: Frühling 2007.

[29] Mobile Metropolitan Ad hoc Network (MobileMAN) Project: MobileMAN; Project Web Site, http://cnd.iit.cnr.it/mobileMAN/index.html, Online: Frühling 2007.

[30] Florian Bahr: Seminar Verteilte Systeme und Netzwerkmanagement, http://www.ibr.cs.tu-bs.de/courses/ws0102/svs/bahr.pdf, Winter 2001/2002.

[31] Guido Krause: Mobile Ad-hoc-Netze Multicast in Ad-hoc-Netzen, http://net.informatik.uni-tuebingen.de/fileadmin/RI/teaching/seminar_mobil/ws0405/slides/slides-krause.pdf, Winter 2004/2005.

[32] Matthias Grünewald, Christian Schindelhauer, Stefan Rührup, Klaus Volbert: Projektgruppe Mobile und Drahtlose Netzwerkkommunikation, http://wwwcs.uni-paderborn.de/cs/ag-madh/vorl/MobDlKomm/unterlagen/PG-MoDraNet-5.pdf, 23.12.2002.

[33] Urs Frey: Topologie und Positionsbestimmung in Mobilen Ad-hoc Netzwerken, http://www.tik.ee.ethz.ch/~beutel/projects/sada/2002ws_frey_btnode_pos.pdf, März 2003.

[34] Anders Lindgren und Olov Schelén: Infrastructured ad hoc networks, http://www.sm.luth.se/~dugdale/publications/lic.pdf, September 2003.

[35] TEIA: TCP slow start and congestion avoidance, http://www.teialehrbuch.de/KIT/16220-TCP-Slow-Start-und-Congestion-Avoidance.html, Online: Frühling 2007.

[36] Srinath Perur and Leena Chandran-Wadia and Sridhar Iyer: Improving the Performance of MANET Routing Protocols using Cross-Layer Feedback, http://www.it.iitb.ac.in/~sri/papers/rCLF-cit03.pdf, Online: Frühling 2007.

[37] Feng Wang and Yongguang Zhang: Improving TCP Performance over Mobile AdHoc Networks with OutofOrder Detection and Response, http://www.iks.inf.ethz.ch/education/ss04/seminar/52.pdf, 2002.

[38] University of Singapore: Performance of mobile ad hoc network in constrained mobility pattern, http://www.ntu.edu.sg/home/bcseet/ICWN2003.pdf, 2003.

[39] Fabius Klemm: TCP-Performance in schnurlosen ad hoc Netzwerken, http://www.informatik.unibw-muenchen.de/mmb/mmb42/4d_DA_Klemm.pdf, Online: Frühling 2007.

[40] Universität Ulm: Intergrierte Sicherheit für mobile ad hoc Netzwerke, http://medien.informatik.uni-ulm.de/forschung/publikationen/wman2004.pdf, Online: Frühling 2007.

[41] Creighton T. Hager: Mobile ad hoc network security, http://www.irean.vt.edu/research_workshop_may2002/06_Hager.pdf, Juni 2002.

[42] Maxim Raya and JeanPierre Hubaux: The security of vehicular ad hoc networks, http://lcawww.epfl.ch/Publications/raya/RayaH05C.pdf, Online: Frühling 2007.

[43] Lidong Zhou and Zygmunt J. Haas: Securing ad hoc networks, http://www.cs.cornell.edu/home/ldzhou/adhoc.pdf, Winter 1999.

[44] Fraunhofer Institut: Security in mobile ad hoc networks, http://www.igd.fraunhofer.de/igd-a8/publications/flyer/manet-security-flyer-english.pdf, Online: Frühling 2007.

[45] Sonja Buchegger: Coping with misbehaviour in mobile ad hoc networks, http://biblion.epfl.ch/EPFL/theses/2004/2935/2935_abs.pdf, 2004.

[46] Christopf Lindemann, Oliver P. Waldhorst: Effective Dissemination of Presence Information in Highly Paritioned Mobile Ad Hoc Networks, http://rvs.informatik.uni-leipzig.de/de/publikationen/papers/SPEED-SECON06.pdf, Online: Frühling 2007.

[47] Universität Karlsruhe: Mobiles Internet, http://doc.tm.uka.de/tr/TM-2005-2.pdf, April 2005.

[48] Gerhard Kadel: Wireless ad hoc & mesh networks, http://www.feldafinger-kreis.de/download/2005/WS1_Impulsref_Kadel.pdf, Januar 2005.

[49] Gunnar Karlsson, Vincent Lenders, Martin May: Delay-Tolerant Broadcasting, http://www.csg.ethz.ch/people/lenders/chants06.pdf, September 2006.

[50] Jörg Roth: Mobile Computing, http://www.dpunkt.de/leseproben/3-89864-165-1/Kapitel_1.1.pdf, Online: Frühling 2007.

Kapitel 2

Operating Systems for Mobile Devices

Ueli Hofstetter, Philippe Hunberbühler, Anil Kandrical

Diese Seminararbeit beschäftigt sich mit Betriebsystemen für Kleingeräte. In einem ersten Schritt wird die Funktion des Betriebsystems erklärt und ein Überblick über die verschiedenen mobilen Betriebsysteme gegeben. In einem zweiten Schritt werden die Anforderungen an ein mobiles Betriebsystem herausgearbeitet und zu guter letzt werden die Programmiersprachen für Applikationen auf Kleingeräten vorgestellt.

Inhaltsverzeichnis

2.1 Einführung

In diesem Abschnitt werden die Begriffe Betriebssystem, Embedded System, Personal Digital Assistent (PDA) und Smartphone eingeführt.

2.1.1 Betriebsystem

Die Aufgabe eines Betriebsystems besteht grundsätzlich darin, den Applikationen eine einfache Grundvoraussetzung zu bieten, um eine Ausführung überhaupt möglich zu machen ohne dabei viel Kenntnis von der Hardware besitzen zu müssen. Die Hauptaufgaben eines Betriebsystems bestehen grob gesagt darin den Prozessor, den Speicher, die Peripherie und die Geräte zu verwalten [30]. Somit werden diese komplizierten Bereiche abstrahiert, so dass die Anwendungsprogramme keinen hardwarenahen Programmcode benötigen, um lauffähig sein zu können.

2.1.2 Kleingeräte für mobile Betriebsysteme

Früher kamen die ersten Rechner ganz ohne Betriebsystem aus, was auf ihre mechanische Bauweise zurückzuführen war. Heutzutage besitzt jedes neuere Mobiltelefon ein Betriebsystem. Nun stellt sich die Frage, warum ein Betriebsystem eigentlich in so einem Kleingerät überhaupt benötigt wird. Dieser Umstand ist damit zu erklären, dass die ursprüngliche Funktionalität des Telefonierens dem Benutzer bei weitem nicht mehr genügen. Dank dem technologischen Fortschritt ist es möglich die neuen Bedürfnisse zu befriedigen, so dass heutzutage das Hören von Music, Fotografieren, Aufnehmen von Videos, Surfen im Internet, und das Speichern von Terminen auf einem mobilen Kleingerät zur Selbstverständlichkeit geworden ist. Die Funktionalität benötigt wie auf dem PC eine Vielfalt von Applikationen wie Audio-Player, Video-Player, Textverarbeitungsprogramm etc. die wiederum nur unter bestimmten Bedingungen arbeiten können. Das wiederum verlangt nach einem Computer mit Betriebsystem, der die Grundvoraussetzung dafür bietet und gleichzeitig mit Hardware auskommt, die klein aber leistungsfähig sein muss. Hier kommen die mobilen Betriebsysteme zum Zug.

Embedded System

Unter dem Begriff Embedded System versteht man einen Computer, dessen Funktionen beschränkt sind, somit müssen wenig Hardware und Software entwickelt und verwendet werden als bei einem herkömmlichen PC werden und es lassen sich Kosteneinsparungen vornehmen, als wenn man einen herkömmlichen Computer zum Einsatz bringen würde [2]. Der iPod nano von Apple kann als Beispiel erwähnt werden. Dessen Display ist auf ein monochromes Bildschrim beschränkt und auch die Leistungen des Prozessors sowie auch seiner Hardware sind stark reduziert im Gegensatz zu einem PC. Weitere solcher embedded Systeme finden in Flugzeugen, Autos, DVD-Playern etc. ihren Einsatz.

Personal Digital Assistent (PDA)

PDA ist die Kurzform für Portable Device Assistant. Psion gilt als Erfinder dieser Geräte, da sie 1984 den ersten elektronischen Organizer entwickelt haben. Dieses Gerät konnte nur einzeilige Information einlesen, speichern und ausgeben und die Uhrzeit auf einem kleinen schwarz-weiss LCD (Liquid Crystal Bildschirm) Bildschirm [12] anzeigen. Heutige PDAs sind multifunktional, besitzen hochauflösende Farbdisplays und haben nur noch wenige Gemeinsamkeiten mit dem ersten PDAs.

Abbildung 2.1: Erster PDA von Psion [12]

Smartphone

Unter dem Begriff des Smartphones versteht man grundsätzlich die Vereinigung eines Mobiltelefons mit dem PDA, die eine Vielzahl von Funktionalität aufweist. Als Beispiel für eines der bekanntesten und ersten seiner Art ist der Nokia Communicator zu erwähnen. Typische Funktionen solcher Geräte: Personal Information Manager, Mobile Telefonie und diverse zusätzliche Applikationen. Personal Information Manager ist eine Software, die persönliche Daten wie Kontakte, Aufgaben, Termine, Notizen und im erweiterten Verständnis auch Dokumente wie Briefe, Faxe und E-Mails verwaltet [31]. Mobiltelefonie bezeichnet das mobile Telefonieren mittels eines Smartphones und eines Handys.

2.2 Überblick über die Betriebsysteme für Kleingeräte

In diesem Kapitel werden die vier bekanntesten Betriebsysteme für mobile Geräte vorgestellt, miteinander verglichen und mittels einer SWOT Analyse betrachtet. SWOT steht für Strengths, Weaknesses, Opportunities und Threats. Diese Analyse wird üblicherweise in Unternehmen vorgenommen um eine neue Technologie oder ein neues Produkt genauer analysieren und einschätzen zu können hinsichtlich der Entwicklungschancen.

2.2.1 Embedded Linux

Linux wurde im Jahr 1991 vom 21-jährigen Studenten Linus Torvald entwickelt. Seine Intention war es ein eigenes Betriebsystem zu entwickeln, dass nach öffentlich zugänglich war, und das jeder nach seinem Belieben verändern konnte [6]. Im Jahr 2000 wurde das Embedded Linux Consortium gegründet, das zum Ziel hat, die technologische Entwicklung sowie auch die Verbreitung von Linux auf dem Markt für mobile Betriebsysteme zu fördern [7].

Smartphones, PDA und WebPads gehören zu den Produkten, die Embedded Linux als Betriebsystem verwenden. Innerhalb von 13 Monaten sind an die 52 mobile Kleingeräte auf den Markt gekommen, die Linux verwenden [8]. Die bekanntesten Marken sind: Panasonic, Samsung, Motorola und Siemens. Ein grosser Teil dieser 52 Produkte stammt von unbekannten Herstellern, die aus Asien stammen. Das lässt sich vielleicht darauf zurückführen, dass das Betriebsystem Open Source Software ist, und somit Kosten eingespart werden können und gleichzeitig ein schneller Einblick und Einstieg in diese Branche ermöglich wird, als wenn man sich die ganze Technologie von Grund auf selber entwickeln müsste.

Strengths

- Open Source Software ermöglicht unabhängige Entwicklungen und Anpassungen.
- Es findet zunehmende Verbreitung und Anwendung in mobilen Kleingeräten.

Weaknesses

- Fehlende Attraktivität den Code für mobile Betriebssysteme zu entwickeln und dann öffentlich zugänglich machen zu können durch General Public License (GPL), da die Konkurrenten ohne Aufwand die Inovationen übernehmen können.
- Nachahmung durch Konkurrenz gross, da Open Source allen die gleiche Ausgangsbedingung verschafft.

Opportunities

- Starkes Marktwachstum prognostiziert für die Zukunft.
- Firmen können leicht in den Markt für mobile OS einsteigen.
- Customized mobiles OS bedeutet, dass Zusammenstellung der Betriebsystemkomponenten ermöglicht wird.

Threats

- Konkurrenz durch Symbian ist gross.
- Keine oder nur beschränkter Patentschutz für die Produkte.

Diese Auflistung zeigt unter anderem, dass die Wahl der Lizenz einen starken Einfluss auf die SWOT Analyse hat.

2.2.2 Palm OS

1992 wurde die Firma Palm Computing durch Jeff Hawkins gegründet. Palm entwickelte Anwendungssoftware für Sharp, Casio und Apple. Zu dem bekanntesten Feature zählt das Texteingabe-Software Graffiti. 1994 beginnt die Entwicklung des Handheld-Computer Palm Pilot. 1995 kauft U.S. Robotics Palm Computing. 1996 entsteht der erste PDA der Firma: Palm Pilot. 2002 wird Palm OS als Tochterfirma gegründet und noch im gleichen Jahr zu PalmSource umbenannt. 2003 wird das erste Betriebsystem für Smartphones entwickelt [3].

Für PalmOS sind gegenwärtig über 25000 Anwendungen erhältlich. Stärkster Konkurrent ist einzig Windows Mobile Pocket PC, welches ca. über 15000 Anwendungen anbieten kann. Es ist ein Real-Time Operation System, welches lange zeit nicht Multithreading fähig war. Kadak, welche die Rechte auf dieses System besass, beharrte darauf, keine Änderung zuzulassen ohne seine Genehmigung. Dies hat die Entwicklung von Palm OS um Jahre zurück geworfen [4].

Palm verfügt als einer der wenigen Hersteller über eigene mobile Kleingeräte. Palm stellt vor allem eigene PDAs aber auch Smartphones her. Zu diesen Modellen zählen Treo 680, Treo 750v und Treo 650 (Abbildung 2.2).

Abbildung 2.2: Palm Treo 750v, Palm Treo 650 und Palm Treo 680 [15]

Palm OS hat einst durch seine innovativen Erfindungen, wie der Graffiti Eingabe überzeugen können. Gegenwärtig kämpfen sie aber mit der starken Konkurrenz, die durch Microsoft und Symbian repräsentiert wird. Zudem hat PalmOS keine effektive Verwaltung von Rechten und Separierung von Prozessen. Dies führt dazu, dass mittels einfachen Methoden Zugriff auf dem Palm beschafft werden kann [13].

Strengths

- Die Firma ist sehr innovativ wie Erfindung der Texteingabe-Software Graffiti zeigt.
- Vielfalt an Programmen ist verfügbar im Gegensatz zu Symbian und Windows. Mobile [4]
- Palm ist Marktführer im Segment der PDAs.
- Gehörten zu den Early Mover.

Weaknesses

- Bedarf nach PDA sinkt dank der Erhöhung der Funktionalität der Smartphones.
- Palm verfügt über ein sehr schwaches Rechtemangement.
- Single User, Multi Task Technologie ist erst spät verfügbar.

Opportunities

- Die Applikationsvielfalt kann noch mehr ausgeweitet werden.

Threats

- Starker Druck von Smartphones Betriebsystemhersteller Symbian und Microsoft ist spürbar, da beide stärkeres Wachstum im Markt verzeichnen.

Wie aus der SWOT Analyse zu sehen ist, wird Palm aus dem Markt herausgedrängt, da seine Produkte durch die der Konkurrenz ersetzt werden. Die Firma muss sich neuorientieren im Markt und neue Innovationen entwickeln, um sich gegen die Konkurrenz behaupten zu können.

2.2.3 Symbian

Symbian ist ein privates englisches Unternehmen, das im Jahr 1998 gegründet wurde und grosse Teile von Psion übernahm. Seine Joint Venture Partner sind Nokia, Motorola und Sony Erriccson. Seine wichtigsten Lizenznehmer sind Nokia, Sony Erricson, Motorola, Panasonic, Samsung, BenQ, Fujitsu und Siemens [5].

Symbian basiert auf dem Betriebsystem EPOC, welches von Psion für PDAs entwickelt wurde. Entscheidend wurde die Weiterwicklung 1998 von Nokia, Ericcsson (heute Sony Erricson) und Motorola vorangetrieben, die Bluethooth (mobile kabellose Datenübertragung über Funk), WAP (kabelloses Applikations-Protokoll), Kompatibilität von Java und die Architektur des Prozessors massgeblich verbesserten. Im Bereich der Anwendungen hinkt Symbian den beiden Konkurrenten Microsoft Mobile und PalmOS hinterher. Dies könnte auch damit zu tun haben, dass die Grundfunktionalität des Smartphones, die aus dem PIM und der Mobiltelefonie besteht, reichen und eher ein PDA verwendet wird, wenn leistungsfähigere Applikationen benötigt werden.

Die folgenden Kommunikations- und Nachrichtenprotokolle werden von Symbian unterstützt: TCP/IP, WAP 2.0, Bluetooth, USB, QoS für UMTS und GPRS, HTTP, WSP, SMS, EMS, SMTP, IMAP4 und POP3. Von den Entwicklungssprachen werden C++ und Java unterstützt, wobei es für Java J2ME gibt.

Symbian findet vor allem bei den Mobiltelefonen Einsatz. Gegenwärtig verwenden 64 Modelle von Herstellern wie Nokia, Sony-Ericsson, Motorola und Samsung dieses Betriebsystem [14].

Symbians kann der zukünftigen Entwicklung auf dem Markt gelassen entgegen sehen. Sie gehören zu den führenden Herstellern auf dem Markt für mobile Betriebsysteme. Als einzige echte Bedrohung kann die übermässige Funktionalität der Geräte genannt werden, die sich darin äussert, dass die Benutzer die meisten Applikationen, die durch das Betriebsystem bereitgestellt werden, nicht nutzt. Dies wird in der nachfolgenden Auflistung ersichtlich.

Strengths

- Symbian ist Marktführer auf dem Markt der Smartphones.
- Es findet grosse Hersteller Unterstützung im Smartphone Markt.
- Die Komponenten des Betriebsystems sind frei zusammenstellbar.
- Es zeichnet sich durch seine einfache Bedienung aus.

Weaknesses

- Wenig Applikationsprodukte sind verfügbar.
- Kleine Sicherheitsschwachstellen, die aber zu grossen Schäden führen können.

Opportunities

- Die Wahrung Marktführerschaft ist als Ziel definiert.

Threats

- Zu hohe Funktionalität, die nicht vollständig genutzt wird, ist vorhanden.

Symbian ist führend auf dem Markt für Smartphones und baut diese Stellung weiterhin aus. Es muss nur darauf achten, sich auf die echten Konsumerbedürfnisse zu fokusieren und somit der zu hohen Funktionalität entgegenzuwirken.

2.2.4 Windows

Microsoft brachte 1996 Windows CE für die mobilen Betriebsysteme auf den Markt, welches vornehmlich für Kleincomputer, Industriegeräte und Autogeräte gedacht war. Später erfolgte dann die Aufspaltung von Windows CE in Windows Pocket PC und Windows Smartphone. Diese wurde dann aber wiederum verworfen und es folgte die Vereinheitlichung zu Windows Mobile im Jahre 2003. Windows Smartphone besitzt neben den Gemeinsamkeiten, die sich im Design und der Bedienung äussern, auch Unterschiede, die darin bestehen, dass die Auflösung weniger hoch ist und keine Touchscreen besitzt. Im Jahr 2005 wurde Windows Mobile 5.0 Magneto herausgegeben. Die Vorzüge diesen Betriebsystems bestehen darin, dass es .NET Framework verwendet. Im Mitte 2007 soll das Windows Mobile Crossbow und im 2008 Windows Mobile Photon auf den Markt kommen.

Gegenwärtig sind ca. 15000 Applikationen für Windows Mobile Pocket erhältlich.

QTek und HP gehören zu den prominentesten Beispielen im Windows Mobile Pocket PC Markt [15]. Windows Betriebsysteme sind für Windows Mobile Pocket PC und Windows Mobile Smartphones erhältlich. Grosser Nachteil ist, dass Programme auf dem einem Betriebsystem auf dem anderen nicht laufen, obwohl sie vom gleichen Hersteller sind.

Folgende SWOT Analyse zeigt die Aussichten für Windows Mobile auf dem Markt.

Strengths

- Es verfügt einen hohen Bekanntheitsgrad durch das Desktop Betriebsystem.

Weaknesses

- Mobilität der Programme ist nicht vorhanden.

Opportunities

- Starkes Wachstum wird prognostiziert.
- Benutzer gewöhnen sich ans Betriebsystem, wie beim PC. Ein Möglicher Aufbau einer Monopolstellung wird angestrebt.

Threats

- Ein Mangelnder Sicherheitsstandard wird Windows Mobile vorgeworfen [13].

Windows Mobiles grosse Stärke besteht darin eine hohe Vielfalt an Applikationen aufzuweisen. Die mangelnden Sicherheitsstandards werden aber weiterhin für Schwierigkeiten sorgen.

2.3 Überblick über den mobilen Betriebsystem-Markt

Aus der Abbildung 2.3 wird ersichtlich, dass Symbian dominierender Marktführer ist, gefolgt von Linux, Palm und Microsoft. Dies ist unter anderem darauf zurückzuführen, dass Symbian in vielen neuen Smartphones von Nokia vorzufinden sind und dass das Unternehmen als Branchenleader seine Produkte besser absetzen kann als die Konkurrenz. Microsoft Mobile wächst auch, aber dessen Betriebsystem ist längst nicht so weit verbreitet wie die der Mobiltelefone von Nokia.

Seit der Einführung von Symbian auf Mobiltelefonen sind Produkte wie der Palm Handheld überflüssig geworden, da seine Kernfunktionalität, das Personal Information Management, auf den Mobiltelefonen auch vorhanden ist. Zudem werden PDAs, die beispielsweise für den Aussendienst benötigt werden, meistens mit einem Windows Betriebsystem vertrieben, da die Mitarbeiter dieses Betriebsystem kennen und somit weniger Lernkosten haben.

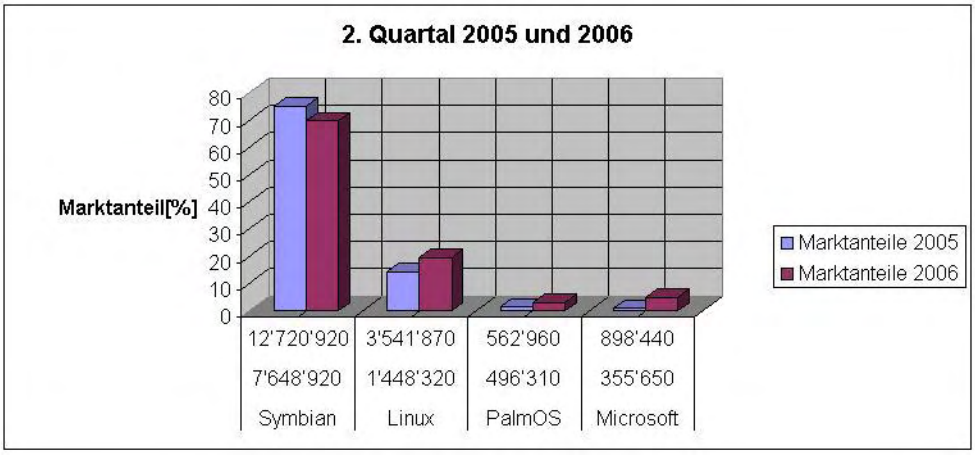


Abbildung 2.3: Marktanteile der Betriebssystemhersteller im 2.Quartal 2005 und 2006 [1]

2.4 Anforderungen an ein Betriebssystem für mobile Geräte

2.4.1 Einleitung

Ein Betriebssystem für mobile Geräte hat grundsätzlich dieselben Aufgaben wie ein Betriebssystem für nicht mobile Endgeräte. Zu den Aufgaben von üblichen Betriebssystem zählen [16]:

- Speicherverwaltung: Das Betriebssystem stellt die Überwachung, Zuweisung und Freigabe des Speichers sicher.
- Prozess Verwaltung: Die einzelnen Prozesse / Programme werden durch das Betriebssytem gesteuert.
- Geräte, Dateiverwaltung: Erkennung und Initialisierung von Ein- und Ausgabegeräten.
- Rechteverwaltung: Das Betriebssystem garantiert, dass Benutzer nur Daten und Programme lesen beziehungsweise ausführen, die für sie bestimmt sind .

Jedoch hat ein Betriebsystem, das für mobile Geräte konzipiert wurde, noch andere Herausforderungen zu meistern. Die Ausgangslage und die Voraussetzungen sind aber ganz unterschiedlich. Als erstes werden auf diese Unterschiede eingangen: Nicht nur die Ein- und Ausgabegeräte sind stark miniaturisiert, sondern auch Prozessor, Speicher und Hauptspeicher von mobilen Geräten sind leistungsschwächer als in stationären Computern. Die Mobilität hat auch Auswirkungen auf die Integrität der Systeme. Als nächstes werden die besonderen Bedrohungen von mobilen Geräten beachtet. Im Weitern werden die speziellen Funktionalitäten von Betriebssystem für Mobile Geräte, die für effizientes Arbeiten nötig sind einerseits und andernseits die Eigenarten eines mobilen Devices nutzen kann, eingegangen. Als letztes wird dieses Kapitel mit den aktuellen Forschungsthemen beschäftigen,

die in einer mobilen Umgebung vorkommen, insbesondere mit den Problem der beschränkten Energie.

2.4.2 Ausgangslage

Beschränkte Ressourcen

Aufgrund der viel kleineren Abmessungen von mobilen Geräten müssen an einigen Punkten in der Konstruktion gespart werden. In der heutigen Zeit ist der Trend zu immer kleineren Geräten deutlich sichtbar: Hatten die ersten Mobiltelefone noch die Ausmasse eines Aktenkoffers, sind die aktuellen Geräte kaum grösser als eine Kreditkarte. Wie es der Name bereits sagt, sollen mobile Devices tragbar sein.

Obwohl die Industrie ständig leistungsfähigere Kleingeräte produziert, sind die Leistungsunterschiede zu den stationären Geräten nicht von der Hand zu weisen. Der Prozessor ist bei highend PDAs immer noch knapp zehn Mal schwächer als in Desktop Computern. Sowohl der Arbeitsspeicher, als auch der Speicher sind auch verhältnismässig klein. So müssen PDS oft mit weniger als 32 Megabyte Hauptspeicher auskommen; Mobiltelefone haben noch weniger. Auch die Ein- und Ausgabegeräte sind sehr begrenzt. Zwar haben heutzutage fast alle mobilen Geräte farbige Displays jedoch misst die Diagonale kaum mehr als 3.5 Zoll. Eingaben müssen entweder über virtuelle Tastaturen auf dem Bildschirm, per Schrifterkennung oder mit einer Tastatur mit 10 Tasten getätigt werden. Zwar gibt es Geräte mit einer QWERTZ-Tastatur, die Tasten sind jedoch schwer zu bedienen. Mobile Geräte beziehen ihre Energie von Akkus. Dadurch sind mobile Geräte für eine gewisse Zeit vom Stromnetz unabhängig, müssen aber sobald sie leer sind wieder geladen werden.

Sicherheit

Immer mehr mobile Geräte sind ständig in einem Netzwerk: bei Mobiltelefonen ist die sogar eine Voraussetzung, aber auch moderne PDAs besitzen die Möglichkeit sich mit andern Geräten drahtlos zu verbinden. Es ist deshalb erforderlich, dass ein Betriebssystem für mobile Geräte trotz der knappen Ressourcen sicher ist.

Bereits 2005 existierten verschiedene Viren für Mobiltelefone, wovon Leavitt [18] mehrere ausführlich beschreibt. Der *Cabir* Virus beispielsweise war einer der ersten Viren für Mobiltelefone. Er wurde von einer Hacker Gruppe als *proof of concept* geschrieben und befiel ausschliesslich Telefone mit Symbian als Betetriebsystem. Er verbreitete sich via Bluetooth, richtete jedoch kaum Schaden an. Ganz anders *Mquito*: Dieser Virus versendete auf befallenen Mobiltelefonen - es waren wieder nur Symbiangeräte betroffen - SMS auf gebührenpflichtige Nummern. Nicht nur für Geräte, auf denen Symbian läuft, werden Viren geschrieben. Jedoch sind die am weitesten verbreitesten Betriebssysteme, welche von Hackern bevorzugt werden [1]. Das Ziel vieler Hacker ist, es mit ihren Viren einen möglichst grossen Schaden anzurichten und schreiben ihre Viren deshalb für die am häufigsten verwendeten Betriebssysteme.

In Zukunft, so Leavitt [18], werden Viren versuchen sich nicht nur über Bluetooth zu verbreiten, da Bluetooth eine Reichweite von maximal 10 Metern hat, ist die Verbreitung nicht so schnell möglich, wie beispielsweise bei einer Ausbreitung via MMS. Es wird so sein, dass sie sich über selbstgenerierte MMS verbreiten.

Yang et al. [17] beschreibt 3 verschiedene Umgebungen, in denen es unterschiedliche Anforderungen gibt um ein System sicher zu machen. Für diese Arbeit sind jedoch nur die 2 von Bedeutung.

- *Single Computer Security*:Ein nicht vernetztes Computersystem. Dabei muss sicher gestellt werden, dass nur befugte Personen sich einloggen können und das Betriebssystem muss dafür sogen, dass die Benutzer auf keine fremden Daten zugreifen können.
- *Network Security*:In einem vernetzten System wird grundsätzlich davon ausgegangen, dass der Angriff von aussen kommt. Firewalls arbeiten deshalb alle nach dem gleichen Prinzip: Sie blockieren Versuche, die von ausserhalb des System eine Verbindung aufbauen wollen.

Es werden auch verschiedene Sicherheitsmassnahmen für mobile Geräte beschrieben [17]: Authentifikation, Verifikation und Autorisierung. Die Authentifikation ist dafür zuständig, dass nur befugte Zugriff auf das Gerät haben und findet bei Mobiletelefonen mittels SIM-Karte und PIN-Eingabe statt. Die Verifikation überprüft, ob der Code keine verbotenen Befehle ausführt. Dass Programme nur auf genehmigte Ressourcen (read, write, SMS versenden etc.) zugreifen, stellt die Autorisierung sicher.

2.4.3 Funktionalitäten zukünftiger Betriebssystem für mobile Geräte

Die Anforderungen die an mobile Systeme gestellt werden haben sich verändert, sodass man die Grundprinzipien von Betriebssystemen überdacht werden müssen. Zukünftige mobile Geräte werden folgende Funktionalitäten implementiert haben [19]: Reconfigurability, Context-Awareness, Adaptability und Personalization.

Reconfigurability ist die Fähigkeit einen unterbruchsfreien Betrieb auf OS-Level während der Laufzeit - auch bei Veränderungen in Software oder Hardware - zu gewährleisten. Dabei sucht das Gerät ständig nach passenden peripheren Geräten, beispielsweise eine externe Tastatur, oder nach Geräten mit denen es ein Ad-hoc Netzwerk aufbauen kann. Sobald ein Gerät verschwindet oder neu hinzukommt, muss dies durch eine effiziente Geräteerkennung registiert werden. Das Betriebssystem muss dann immer entscheiden, ob es sich neukonfigurieren muss. Sobald sich das Gerät neu konfiguriert, muss es seine Integrität sicherstellen und Schutz vor unautorisiertem Code bieten.

Ein Gerät, das *context-sensitve* ist, kennt die Informationen aus seiner Umgebung und weiss auch sinnvoll damit umzugehen. Ein Mobiltelefon, das um sich noch einige andere Mobiltelefone, mehrere Notebooks und einen Beamer hat, merkt beispielsweise, dass es

sich in einem Meeting befindet und stellt seinen Rufton auf lautlos. Context Sensitivität kann auch genutzt werden um Energie zu sparen. Befindet sich eine Mobiltelefon beispielsweise in einem Funkloch (Tunnel beispielsweise) und ist sich dessen bewusst, ist es besser wenn es die Sendeleistung auf ein Minimum reduziert, anstatt ständig nach einem Netz zu suchen.

Adaptability hat eine ähnliche Funktionalität wie die Reconfigurability: eine Applikation passt sich an ändernde Umgebungen an. Dabei findet die Anpassung aber nicht auf OS-Ebene statt, sondern auf Applikationsebene. Eine Software für Videotelefonie auf mobilen Geräten beispielsweise wechselt von der normalen Videotelefonie auf textbasierte Kommunikation, sobald der Durchsatz zum andern Teilnehmer zu klein ist (beim Wechseln von UMTS auf GPRS etwa).

Personalization wird als Schlüssel für künftige mobile Geräte angesehen. Ein *personalisierbares* Gerät muss in der Flut von Informationen entscheiden, welche wichtig für den Benutzer sind. Eine grosse Herausforderung spielen dabei die kulturellen Unterschiede, die bei der Realisierung beachtet werden müssen. Wichtige Inforamtionen dürfen dabei auf keinen Fall verloren gehen.

2.4.4 Research

Raatikainen [19] verlangt einen Paradigmenwechsel für die Requirements eines Betriebssystems.

> Forget the end-user terminal and start thinking about end-user systems!

Es müssen demnach die Systeme als Ganzes betrachtet werden und nicht nur die Endgeräte als Einzelstücke. Es müssen wie oben erwähnt die Konzepte für Betriebssystem für die geänderten Anforderungen überdacht werden. Er nennt vier Punkte: Self-Awareness, Detection and Notifactions, System Integrity und Power Management.

Self-Awareness ist die Voraussetzung für Reconfigurability und Adaptability. Damit ein System rekonifgurierbar sein kann, muss das Betriebssystem den eigenen Konfiguration und Zustand sowie den Zustand von Geräten, das es verwaltet, kennen.

Auch *Detection and Notifications* sind erforderlich, damit ein System *reconfigurable* sein kann. Es muss erkennen, wenn neue Geräte/Subsysteme oder Service vorhanden sind. Selbstverständlich muss es auch erkennen, wenn solche verschwinden. Wenn Speichersubsysteme plötzlich verschwinden, darf es zu keinen Datenverlusten kommen, auch nicht bei zu dieser Zeit bearbeiteten Dokumenten.

Der Erfolg vom zukünftigen Internet hängt vom Vertrauen der Konsumenten ab [19]. Das gegenwärtige Internet ist befallen von Viren, Spam und Fälschungen. Dieses Problem zeigt, dass von Anfang an ein Konzept vorhanden sein muss, um vor solchen Schäden vorzubeugen. Wenn man sich erst nach dem Auftauchen dieser Probleme damit befasst, ist es zu spät. Die *System Integrity* muss auf jeden Fall sicher gestellt werden! Es dürfen

demnach nur berechtige Personen auf die private Inhalte Zugriff haben. Eine Grundvoraussetzung dafür ist die Authentifikation. Eine Lösung ist das Konzept der *chains of trust*: Bei jeder Neukonfiguration wird eine neue *chain of trust* erstellt, das heisst es werden nur Verbindungen zu vertrauenswürdigen anderen oder zu bereits bekannten Devices aufgebaut. Auch bei dieser Lösung stellt sich das Problem der Effizienz, da die Erstellung einer *chain of trust* Energie benötigt. Ein anderer Ansatz verfolgt die Technik der Sandboxen. Dabei werden den Programmen verschiedene Sandboxen zugewiesen und ihre Befehle/Anweisungen haben nur Einfluss auf diese Umgebung und nicht auf das ganze System.

Die beschränkte Strom-unabhängige Laufzeit von mobilen Geräten ist ein grosser Bereich ([20], [22]), an dem zurzeit geforscht wird. Wie bereits erwähnt, verbrauchen mobile Device Energie für Aufgaben, die feste Geräte nicht haben: wiederholtes Auf- und Abbauen von Netzwerkverbindungen, Suchen und Erkennen von anderen Geräten etc. Durch effizienteres Power Management, das heisst die genannten Schwierigkeiten mit möglichst wenig Energie zu erledigen und eine effiziente Ressourcen Verwaltung, hat ein mobiles Device eine längere Laufzeit. Aus diesem Grund wird nachfolgend genauer auf Power Management eingegangen.

2.4.5 Power Management

Batterien und Akkumulatoren sind oft die schwersten und grössten Teile von mobilen Geräten. Während die restlichen Komponenten immer kleiner, leichter und leistungsfähiger werden, trifft dies bei den Batterien nur sehr beschränkt zu. Zwar könnten durchaus kleinere Akkus verbaut werden, dabei wird jedoch eine bewusste Reduktion der Kapazität der Akkus vorgenommen. Dies resultiert wieder in einer kürzeren Laufzeit des Gerätes [20].

Die Tabelle 2.1 zeigt den Energieverbrauch einzelner Komponenten eines Notebooks. Auf der Grafik fehlen noch die Komponenten für die drahtlose Kommunikation. Auffallend ist jedoch, dass das Basissystem, auch bei gedrosselter CPU Frequenz, der grösste Energieverbraucher ist. Zum Basissystem gehören die CPU und der Arbeitsspeicher. Der nächst grössere Energiekonsument ist die Rückbeleuchtung des Displays. Der Motor der Festplatte verbraucht auch einen wesentlichen Teil der gesamten Energie.

Für die Komponenten CPU, Harddisk und für Drahtlose Kommunikation werden im Folgenden Lösungen und Ansätze ein effizientes Power Management beschrieben.

Der Prozessor ist der grösste Energieverbraucher eines mobilen Systems. Dementsprechend sind die Bemühungen in diesem Gebiet besonders gross. Nicht nur auf Betriebssystmebene wird versucht den Verbrauch zu minimieren, sondern auch auf Hardwarelevel. So sind die Mikroprozessor Hersteller versucht möglichst energiesparende Chips für Notebooks zu entwickeln. In dieser Ausarbeitung wird nur ersteres genauer betrachtet.

Der Verbrauch des Prozessor ist proportional zur Spannung im Quadrat mal der Frequenz: $V^2 * F$. Moderne so genannte System-On-a-Chip Prozessoren haben mehrere stromsparende Zustände [22]. Demnach können durch Senken der Frequenz lineare Einsparungen

Tabelle 2.1: Energieverbrauch einzelner Komponenten (in Anlehnung an [20])

Component	Power (Watts)
base System (2MB, 25 MHz CPU)	3.650
base System (2MB, 10 MHz CPU)	3.150
base System (2MB, 5 MHz CPU)	2.800
screen backlight	1.425
hard drive motor	1.100
math co-processor	0.650
floppy drive	0.500
external keyboard	0.490
LCD screen	0.315

gemacht werden; setzt man die Spannung herunter können sogar quadratische Einsparungen erfolgen.

Der allgemeine Ansatz für einen stromsparenden CPU Algorithmus ist die Last möglichst auf ein tiefes Niveau zu verteilen, anstatt den Prozessor erst voll zu belasten und später ist er in einem untätigen Zustand zu versetzten. Das heisst, solange eine Aufgabe bspw. in 100 ms mit einer konstanten, niedrigen CPU-Belastung durchgeführt werden kann, statt den Prozessor zuerst 50 ms voll zu belasten und danach ist der Prozessor 50 ms untätig, wird die erste Variante gewählt. Die Bedienbarkeit der einzelnen Applikationen (Ladezeiten, Startzeiten etc) sollten nicht oder nur geringfügig schlechter werden [20].

Mobile Kommunikation

Bei der mobilen Kommunikation haben sich bis heute zwei Standards durchgesetzt: der IEEE Standard 802.11 und Bluetooth. Da beide Standards vorwiegend in mobilen Geräten eingesetzt werden, haben beide Stromsparmassnahmen implementiert. Insbesondere, wenn die Geräte oft in einem untätigen Zustand sind, ist es wichtig, ein gutes Power Management zu haben.

Das Prinzip hinter dem Power Management von *IEEE 802.11* ist simpel. Der Transceiver wird ausgeschaltet, solange dieser nicht benötigt wird. Das Power Management kann jedoch nicht im Voraus wissen, wann etwas empfangen werden soll. Der Transceiver muss deshalb zu gewissen Zeitpunkten eingeschaltet sein. Je grösser das Intervall zwischen den Zeitpunkten ist, desto mehr Strom kann gespart werden, jedoch wird dadurch auch der Durchsatz kleiner. Stromsparende Algorithmen gibt es nicht nur für Netzwerke mit einer Infrastruktur, sondern auch für ad-hoc Netzwerke (Ad-hoc Netzwerke verfügen über keine Infrastruktur, sondern setzten sich lediglich aus den einzelnen Geräten zusammen). Jedoch wird hier nur auf die Netzwerke mit einer festen Basisstation eingegangen.

Es gibt zwei Zustände beim Power Management von *IEEE 802.11*: *sleep* und *awake*. Die *Timing Synchronization Function* sorgt dafür, dass alle Teilnehmer zum gleichen Zeitpunkt im Status awake sind. Der Access Point versendet zu diesem Zweck Beacon

Frames. Hat der Access Point Pakete für eine bestimmte Station, wartet er eine zufällige Zeit und versendet anschliessend eine Traffic Indication Map (TIM) oder eine Delivery Traffic Indication Map (DTIM). Eine TIM enthält alle Empfänger für unicast Nachrichten, die DTIM hingegen alle Empfänger für multicast oder broadcast Nachrichten. Erhält eine Station weder eine TIM noch eine DTIM oder ist in keiner als Empfänger vermerkt, geht sie nach einer gewissen Zeit wieder in den Status sleep. Möchte eine Station etwas versenden, wartet auch sie eine zufällige Zeit nach dem Beacon Frame und versendet anschliessend ihr Packet. Der Access Point puffert und versendet diese (siehe Abbildung 2.4) [21].

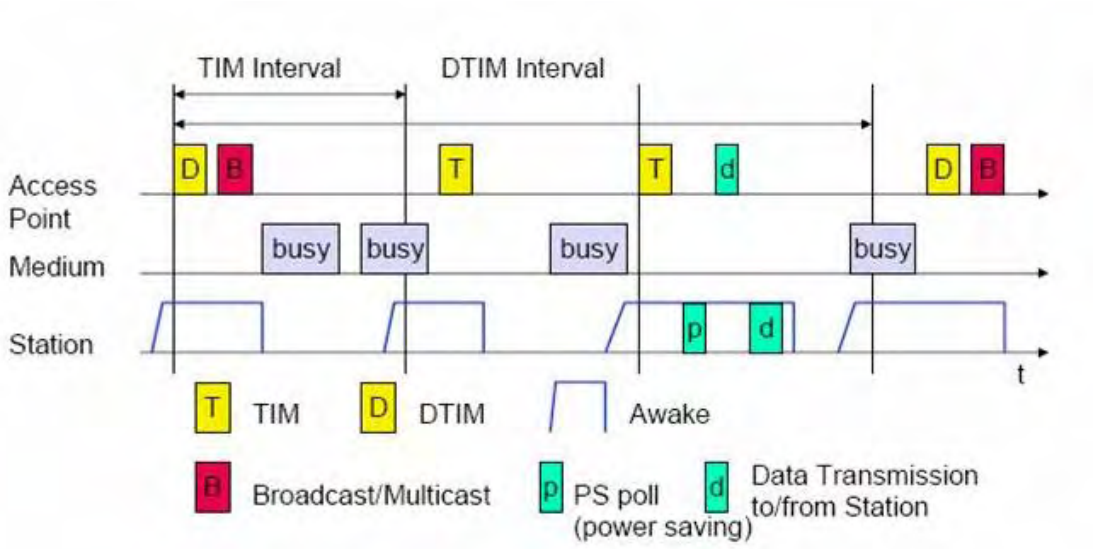


Abbildung 2.4: Energiesparmassnahmen in IEEE 802.11 [21]

Der *Bluetooth* Standard kennt 3 Stromsparzustände: Hold, Sniff und Park. Ausserdem misst der Empfänger die Signalstärke und teilt dem Sender mit, ob er das Signal verstärken oder schwächen soll.

Im *Hold* Modus behält das Gerät seine *active member adress*, kann aber keine Asynchronous Connectionless Link Pakete mehr empfangen. Dabei wird dem Master mitgeteilt, dass das Gerät für eine gewisse Zeit nicht mehr zuhört. Findet im Piconet (Bluetooth ad-hoc Netzwerk) keine Aktivitäten statt, kann das Gerät entweder den Stromverbrauch senken oder in einem anderem Piconet teilnehmen.

Der *Sniff* Modus empfängt das Gerät nur noch zu programmierbaren Zeitpunkten vom Piconet. Die Intervalle für das Zuhören kann programmiert werden und ist Applikationsabhängig. Von den Stromsparmodi verbraucht der Sniff Modus am meisten Strom.

Der *Park* Modus ist der sparsamste Modus. Das Gerät verliert dabei die *active member adress* und erhält dafür aber eine *parked member address*. Das Gerät gehört zwar immer noch zum Piconet, kann aber keine Daten versenden, solange es keine *active member adress* besitzt. Dafür muss es in einen der zuvor genannten Sparmodus oder in den aktiven Modus wechseln.

Bei der Kommunikation ist es wichtig, dass auch Geräte von unterschiedlichen Herstellern miteinander kommunizieren können. Um diese Interkompatibilität zu gewährleisten, wurden Standards geschaffen. Bis Implementationen einen Hersteller-übergreifenden Standard durchgesetzt haben, müssen sie durch ein kompliziertes und langwieriges Prozedere. Es ist deshalb schwer Verbesserungen oder Neuerungen für die Power Management Funktion zu machen. Nokia wagt jedoch einen Versuch mit WiBree. WiBree benutzt das Frequenzband von Bluetooth, hat etwa dieselbe Reichweite aber hat mit 1 Mbps eine kleinere Übertragungsrate. Der Stromverbrauch soll aber um einiges kleiner sein, als bei Bluethooth. Dementsprechend soll WiBree vor allem in mobilen Kleingeräten eingesetzt werden. Ob sich WiBree durchsetzen kann, wird sich noch zeigen. Wenn Nokia aber keine Partner findet, die diese Technologie verwenden, wird WiBree sich zu einer Nokia exklusiven Technologie entwickeln und nicht zu einem herstellerübergreifenden Standard werden [23].

Festplatte

Eine Harddisk verbraucht durch ihren mechanischen Aufbau einiges mehr an Strom als beispielsweise ein Flashspeicher, der ganz ohne bewegliche Teile auskommt. Es versteht sich von selbst, dass eine Disk, die sich dreht, mehr Energie braucht als eine die nicht dreht. Der Spinup, also die Disk wieder zum Drehen zu bringen, benötigen jedoch eine gewisse Zeit bis die Disk wieder einsatzbereit ist und benötigt zusätzliche Energie. Es ist deshalb je nach Situation effizienter die Disk keinen Spindown vollziehen zu lassen, da der Wiederanlauf mehr Strom braucht, als durch den Spindown gewonnen wurde. Douglis et al. [24] haben sich genauer mit dem Zeitpunkt beschäftigt, wann diese Spindowns vollzogen werden.

Die Autoren führen sechs Klassifizierungen ein: Sowohl für den Zeitpunkt des Spindowns als auch für jenen des Spinups legen sie einen Algorithmus fest. Vier Algorithmen werden eingeführt: OPTIMAL, DEMAND, THRESHOLD und PREDICTIVE. Der Spindown beziehungsweise Spinup wird bei den Algorithmen OPTIMAL und DEMAND, wie es der Name schon vermuten lässt, im optimalen Fall ausgeführt respektive sobald die Festplatte nicht mehr gebraucht wird. Der erstere ist ein theoretischer Algorithmus und dient als Benchmark. Der THRESHOLD Algorithmus wartet eine gewisse Zeit ab, in der die Disk nicht bebraucht wird, bis eine Schwelle überschritten wurde und führt dann den Spindown aus. Der PREDICTIVE Algorithmus versucht einen optimalen Zeitpunkt aus vergangen Daten zu finden. Diese Algorithmen lassen sich kombinieren: einer für den Spindown und einer für den Spinup. Beispielsweise bedeutet die Klasse PREDICTIVE_DEMAND, dass der Zeitpunkt des Spindowns auf Grund vergangener Daten berechnet wird und der Spinup, sobald die Disk wieder gebraucht wird, stattfindet.

OPTIMAL_OPTIMAL führte zu Reduktionen des Energie Verbrauches von 35% bis 50%. Die THRESHOLD_DEMAND Klasse, bei einem Schwellwert zwischen 1 und 10 Sekunden, führte zwar zu kurzen Verzögerungen, sparte jedoch 14% bis 44% der Energie. Sie erwies sich als die effizienteste Klasse.

2.5 Programmiersprachen für mobile Kleingeräte

2.5.1 Einleitung

Heute bieten die meisten Programmiersprachen für den Desktop- und Serverbereich dermassen viel Funktionalität, gute Compiler und Laufzeitumgebung, dass sich fast jede Aufgabenstellung damit lösen lässt. Die Auswahl der optimalen Programmiersprache ist dadurch sicherlich nicht einfacher geworden. Der Trend bei Anwendungen für mobile Kleingeräte geht in eine ähnliche Richtung. Deshalb sollen in diesem Kapitel nicht die verfügbaren Sprachen für mobile Kleingeräte untersucht werden, sondern vielmehr Kriterien erarbeitet werden, die eine Programmiersprache für diesen Bereich prädestinieren. Anhand dieser Kriterien sollte es dann möglich sein, bestehende Programmiersprachen für mobile Kleingeräte zu beurteilen und den Nutzen einer möglichen Portierung einer bereits existierenden Programmiersprache zu bewerten, was allerdings nicht Gegenstand dieser Arbeit sein soll. Dabei gilt es sowohl die hardwaretechnischen Einschränkungen der Kleingeräte als auch das Einsatzgebiet der erstellten Anwendungen zu beachten. Ausserdem muss auch die Wirtschaftlichkeit der Verwendung einer Programmiersprache miteinbezogen werden. Denn nur wenn Applikationen schnell und kostengünstig hergestellt und angeboten werden können, haben sie auf dem Markt eine Chance.

2.5.2 Anforderungen

Die folgenden Abschnitte basieren auf einem ACM Bericht von Tommy Thorn [32] aus dem Jahre 1997. Er beschreibt darin kurz die Anforderungen an mobilen Code und wie diese von einigen Programmiersprachen umgesetzt werden. Obwohl mobiler Code, also Programmcode der über über das Netz auf das eigene System übertragen und dann unmittelbar ausgeführt wird (z.B. ActiveX Dll's, Applets etc.), ja nicht zwingend auf Kleingeräten ausgeführt werden muss, treffen viele der beschriebenen Anforderungen auch auf Programmiersprachen für mobile Kleingeräte zu. Diese sollen im Folgenden erläutert und um aktuell gewordene Anforderungen ergänzt werden.

Portabilität

Ein sehr wichtiges Kriterium für die optimal auf Kleingeräten einzusetzende Programmiersprache ist sicherlich die Portablität der damit geschriebenen Applikation. Es scheint als sei diese Fähigkeit sogar noch entscheidender als auf gewöhnlichen Systemen. Denn erstens ist die Anzahl der unterschiedlichen Modelle von mobilen Kleingeräten tendenziell höher und die Modelle habe einen kürzeren Produktlebenszyklus (was eine häufigere Portierung erfordern würde). Ein einfaches Mittel Portabilität zu erreichen ist wie auf gewöhnlichen Systemen der Einsatz virtueller Maschinen. Man könnte jetzt anführen, dass durch die eingeschränkten Betriebsmittel deren zusätzlicher Ressourcenverbrauch zu hoch sei. Wie sich allerdings am Erfolg von Java sowohl auf traditionellen wie auch auf mobilen Systemen zeigt, überwiegt der Vorteil der Portabilität in diesem Falle ganz klar. Dies nicht

zuletzt durch die optionale Möglichkeit Java Code auch native zu kompilieren oder einen JIT Compiler zu verwenden.

Safety

Unter Safety versteht man die Sicherheit, die ein Programm oder dessen Benutzer bei der Benutzung bietet. Also z.B. die Ausfallsicherheit eines computergesteuerten Bremssystems oder eines Mobiltelefones (in Notfallsituationen). Die Gefahr, dass ein Programmfehler zu einem Systemabsturz führt und damit eben diese Sicherheit gefährdet, besteht überall. Sei es durch einen unerlaubten oder ungültigen Speicherzugriff, einen Bufferoverflow, eine unerlaubt Typkonvertierungen oder etwas ähnliches. Dies durch Sprachkonstrukte, Laufzeitumgebungen und andere Massnahmen verhindern zu können, ist natürlich nicht nur für Sprachen, die auf mobilen Geräten zum Einsatz kommen, sehr wichtig. Trotzdem gibt es einige Gründe dafür, weshalb diesem Kriterium bei Anwendungen hier besonderen Stellenwert zugeordnet werden kann. Vor der Auslieferung werden die Anwendungen höchstwahrscheinlich weniger intensiv getestet, d.h. der Fehler tritt eventuell erst beim Anwender auf. Durch den tiefen Preis und die einfach Installation ist die Applikation dann aber schon weit verbreitet, was zu hohen Folge- und Rückrufkosten führen kann. Auch gilt es zu beachten, dass solche Applikationen evt. auch in Geräten installiert werden, die in kritischen Bereichen eigesetzt werden. Deshalb sollte eine Sprache/Laufzeitumgebung für mobile Endgeräte Konzepte wie die Typprüfung während der Kompilierung, Fehlerbehandlung, Garbagecollection und Möglichkeiten verwenden, um mögliche Programmfehler zu verhindern.

Security

Das mobile Kleingeräte aufgrund ihres Einsatzgebietes das öffentliche Funk- und Telephonnetz zur Kommunikation verwenden müssen, sollte eine verschlüsselte Datenübertragung möglich sein. Um deren Einsatz möglichst flexibel zu gestalten, wäre es von Vorteil, dass die Programmiersprache bereits über die nötige, schlank implementierte Funktionalität in Form von austauschbaren Bibliotheken verfügen würde. Dabei sollten vor allem die gängigsten Protokolle, wie Secure Socket Layer (SSL) und dessen Nachfolger Transport Layer Security (TLS) unterstützt werden, die die verschlüsselte Datenübertragung im Internet ermöglichen. Dies nicht zuletzt deshalb, weil das sogenannte *mobile e-commerce* immer mehr an Bedeutung gewinnt. Da man die Applikationen einfach herunterladen und installieren sollte können, wäre ausserdem die Möglichkeit einer Codesignierung sehr wünschenswert.

Sparsamer Umgang mit beschränkten Hardwareressourcen

Ein der wohl offensichtlichsten Anforderungen an eine Programmiersprache ist der schonende Umgang mit den knappen Betriebsmitteln. Dazu bieten sich natürlich mehrere Ansatzpunkte. Aus Sicht der Prozessorbelastung wäre es wahrscheinlich sinnvoll, auf eine

virtuelle Maschine zu verzichten und nur plattformspezifischen Maschinencode zu verwenden. Wie aber schon erläutert, ist deren Einsatz als Voraussetzung für die Portabilität des Programmcodes allerdings unerlässlich. Es bestehen aber auch bei Einsatz von virtuellen Maschinen durchaus Optimierungsmöglichkeiten, wie z.B. die schon erwähnte JIT Kompilierung. Aus Sicht der Speicherausnutzung gibt es mehrere Kriterien: statisch betrachtet wären dies z.B. die Verwendung von platzsparenden Datentypen und schlanker darauf aufbauender Klassenbibliotheken. Während der Laufzeit kommt es dann unter anderem darauf an, den auf dem Heap allokierten Speicher so schnell wie möglich wieder frei zu geben. Beim Einsatz eines Garbage Collectors ist dabei auf dessen möglichst schlanke Implementierung und effektive Arbeitsweise zu achten.

Kurze Entwicklungszeit

Natürlich ist eine schnelle und damit kostengünstige Anwendungsentwicklung überall wünschbar. Da sich die Technologie im Bereich der Kleingeräte (wie z.B. die Entwicklung im Bereich der Mobiltelephone) mit rasanter Geschwindigkeit entwickelt und auch daran angepasste Anwendungen erfordert, kommt auch diesem Punkt eine spezielle Bedeutung zu. Um Rapid Application Development, hier auf die möglichst schnelle Umsetzung der geforderten Funktionalität und Stabilität bezogen, zu ermöglichen, sind neben leistungsstarken Bibliotheken integriete Entwicklungsumgebung nötig, die ein automatisiertes Erstellen der Anwendung aus dem Sourcecode und das anschliessende Inbetriebnehmen und Testen auf den Endgeräten ermöglichen. Entscheidend ist ausserdem, dass das Programmiermodell nicht zu stark von demjenigen der Muttersprache abweicht. Dies erlaubt es Entwicklern, ihre Kenntnisse auch in anderen Umgebungen sofort produktiv einzusetzen, was ebenfalls zu kürzeren Entwicklungszeiten beiträgt.

Mächtige Multimedia- und Streamingbibliotheken

Durch die stetig steigenden Bandbreiten ergeben sich auch für Applikationen auf mobilen Geräten ganz neue Anwendungsmöglichkeiten. Um diese allerdings auch nutzen zu können, muss die Programmiersprache über Bibliotheken zur Unterstützung der Protokolle verfügen. Ähnlich sieht es bei der Benutzerschnittstelle aus. Die angeboten Funktionalität der immer leistungsfähigerern Displays können ohne entsprechend Funktionen nicht voll ausgenutzt werden.

Auch wenn die Aufzählung der Anforderungen keineswegs vollständig war, zeigt sie doch klar, dass nicht jede auf konventionellen Systemen etablierte Programmiersprache auch automatisch für den Einsatz auf mobilien Kleingeräten geeignet ist.

2.6 Die Java Micro Edition

Einleitung

In diesem Kapitel soll die Java Micro Edition (Java ME) näher vorgestellt werden. Sie hat sich zur führenden Plattform für die Anwendungsentwicklung für mobile Kleingeräte entwickelt. Dies hat mehrere Gründe. Zum einen erfüllt die sie fast alle erwähnten Anforderungen. Dazu kommt Sun's offene Politik in den Bereichen Lizenzierung, Spezifikation und Dokumentation.

Übersicht

Die Java ME Plattform wurde ins Leben gerufen, um Java Applikationen und eventuell sogar Applets auf Kleingeräten und in integrierten Systemen mit eingeschränkten Hardwareressourcen (CPU, Memory, I/O) auszuführen. Durch die Verwendung der Java Standard Edition (Java SE) als Grundlage, verfügt die Java ME auch über deren sprachlichen Features. Da die verfügbaren Ressourcen und Anforderungen der Kleingeräte aber sehr unterschiedlich sind, wurde ein System eingeführt, dass es erlaubt, die Java ME speziell an das jeweilige Kleingerät anzupassen. Das Java ME Runtime Environment wird deshalb für die unterschiedlichen Geräte aus 3 Komponenten, der Configuration, dem Profile und Packages, individuell zusammengesetzt (Abbildung 2.5). Die Hauptkomponente(Configuration) besteht aus einer virtuellen Maschine (VM) und die grundlegenden Klassenbibliotheken. Zur Zeit sind zwei solche Basiskomponenten, die Connected Device Configuration (CDC) und die Connected Limited Device Configuration (CLDC) verfügbar. Während die CDC mit ihrem größeren Funktionsumfang und höheren Ressourcenverbrauch vor allem in integrierten Systemen und highend PDAs eingesetzt wird, wird die hardwaretechnisch gesehen weniger anspruchsvolle CLDC hauptsächlich in mobilen Kleingeräten verwendet. Auf diesen Konfigurationen setzten dann die Profile auf. Dabei handelt es sich um zusätzliche Klassenbibliotheken, die weitere Funktionalität zur Verfügung stellen. Um noch mehr Funktionalität der einzelnen Geräte zu nutzen können weitere Klassenbibliotheken, Packages hinzugefügt werden(z.B. für Bluetooth oder Multimediafunktionalität [28]).

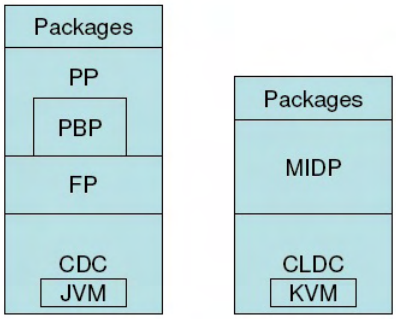


Abbildung 2.5: Möglicher Aufbau von Java ME Laufzeitumgebungen. Links auf der CDC aufbauend, rechts auf der CLDC

Die Konfigurationen im Detail

Die CDC

Wie bereits erwähnt benötigt die CDC mehr Speicher (2MB RAM und 2MB ROM) und einen schnelleren, leistungsfähigeren Prozessor (im Gegensatz zur CLDC läuft die CDC nur auf 32-Bit Prozessoren; höhere Prozessorleistung ist ausserdem meist mit einem höheren Stromverbrauch verbunden). Dafür ist neben dem höheren Funktionsumfang auch die Verarbeitung grössere I/O - Streams möglich und es sind Möglichkeiten für das Powermanagement vorhanden. Durch die technische Nähe zur Java SE (gleiche VM-Spezifikationen) können außerdem bestehende Desktop Applikationen leichter darauf portiert werden (im folgenden wird der Begriff VM gleichbedeutend mit deren Spezifikationen/ Referenzimplementierung verwendet). Eingesetzt wird die CDC deshalb hauptsächliche in Home Entertainment- und ähnlichen Geräten. Dadurch, dass die CDC die selbe VM-Spezifikation wie die Java SE verwendet, bietet sie gegenüber der CLDC 1.0 ausserdem folgende Vorteile: Gleitkommaberechung, Definition eigener Klassenlader sowie Threadunterstützung. Die CLDC 1.1 bietet bis auf die Möglichkeit der Implementierung eigener Klassenlader all diese Funktionen allerdings ebenfalls schon an.

Die CLDC

Die CLDC kommt mit langsameren 16- und 32-Bit Prozessoren und weniger Speicher (min. jedoch 160KB; 128KB für die KVM und die Klassenbibliotheken und 32KB für die Runtime Objekte der KVM) aus. Neben der maximal möglichen Verabeitungsgeschwindigkeit im I/O-Bereich (9,6 Mbit/s) ist auch die von den Klassenbibliotheken gebotenen Funktionalität geringer [29]. Dafür verbraucht die CLDC Configuration natürlich weniger Strom, was einer der Gründe dafür ist, dass sie immer mehr auch in grösseren mobilen Geräten eingesetzt wird. Der Funktionalitätsumfang lässt sich dann durch zusätzliche Packages immer noch an die spezifischen Anforderungen anpassen. Als VM kommt die von der Java SE-Spezifikation abweichende Kilobyte Virual Machine (KVM zum Einsatz). Bis auf ein paar kleine Ausnahmen ist diese aber mit den Jave SE Spezifikationen kompatibel [26].

Zum Schluss muss noch angefügt werden, dass die Spezifikation von den Herstellern nicht immer gleich gut umgesetzt wird.

Die Profile im Detail

Zur Zeit sind offiziell vier verschiedene Profile spezifiziert. Das Foundation Profile (JSR 46), das Personal Basis Profile(JSR 129) sowie das Personal Profile(JSR 62) werden zusammen mit der CDC verwendet. Nur das Mobile Information Device Profile 2.0 (MIDP JSR 118) wird zusammen mit der CLDC eingesetzt [28].

Das Foundation Profile (FP) stellt die Basisfunktionalität sowie Netzwerk- und I/O Support zur Verfügung. Jedoch bietet es keine Unterstützung für das Erstellen graphischer Benutzeroberflächen.

Das Personal Profile (PP) baut auf dem FP auf und stellt zusätzlich die vollständige AWT- und Appletbibliothek sowie eine unvollständige Implementierung der JavaBeans Spezifikation zur Verfügung. Durch die mächtige GUI- und Netzerkunterstützung ist es daher geeignet für Communicators (leistungsfähige Mobiltelefon mit 10-Fingertastatur) und Spielkonsolen.

Das Personal Basis Profile (PBP) baut ebenfalls auf dem FP auf, bietet aber zusätzlich noch Xlet Unterstützung, also mit der Applettechnolgie der Java SE vergleichbare Funktionalität.

Das Mobile Information Device Profile wurde speziell für die Zusammenarbeit mit der CLDC auf mobilen Kleingeräten konzipiert und liegt heute in der Version 2.0 vor. Da bei mobilen Geräten zur Zeit fast ausschliesslich diese Kombination eingesetzt wird, werden deren Konzepte und Architektur weiter unten noch ausführlich erläutert. Die folgende Aufzählung soll nur einen Überblick über die Hauptkomponenten und deren Funktionen geben.

- Bibliothek zur einfachen und ressourcensparenden GUI Erstellung und Ereignisbehandlung
- Gameframework zur Unterstützung der Spielprogrammierung (Animationsklassen, Screenmanager etc)
- Eine Teilmenge der als Package verfügbaren umfangreichen Mobile Media API zur Bild- und Tonverabeitung
- Unterstützung von HTTP, HTTPS (falls der Request über ein WAP Gateway erfolgt), Datagrams und Sockets.
- Over-the-Air Provising: Unterstütztzung zum einfachen Herunterladen, Installieren und Updaten von MIDlets
- Schnittstelle zu einer von der Plattform implementierten persistenten Datenbank
- Möglichkeit die digitale Signaturen von heruntergeladenen MIDlets zu verifizieren (X.509 PKI)

2.7 Das MIDP im Detail - Konzept und Entwicklungsmodel

Nach der Betrachtung der von der Java ME zur Verfügung gestellten Infrastrukutur wird in diesem Abschnitt auf die Programmierung eingegangen. Als Ausgangspunkt dient das MIDP.

Mit dem MIDP ist es möglich, gleichzeitg mehrere Applikationen, sogenannte MIDlets, auszuführen. Die Spezifikation (JSR 118) definiert dabei unter anderem wie ein solches MIDlet verpackt werden muss, damit es von der Laufzeitumgebung erkannt, ausgeführt,

verwaltet und auch wieder beendet werden kann. Ausserdem ist das Format und der Inhalt des sogenannten JAD-File festgelegt, das optional weitere Informationen über das zu deployende Midlet enthält [25].

2.7.1 Der Lebenslauf eines MIDlets

Die Basisklase jedes MIDlets muss von der Klasse MIDlet abgeleitet werden, damit die Laufzeitumgebung das MIDlet starten und verwalten kann. Anders als Desktopapplikationen werden Midlets nicht durch den Aufruf einer Main-Methode gestartet, sondern durch das Instanzieren der Basisklasse des MIDlets. Diese Instanz wird danach vom Applikations Management System der Laufzeitumgebung der CLDC übergeben, die die startApp()-Methode aufruft. Damit beginnt der eigentliche Lebenslauf eines MIDlets. Nach der Instanzierung kann sich ein MIDlet dabei nur in einem der drei Zustände Paused, Active und Destroyed befinden. Diese Zustände wurden zur besseren Verwaltbarkeit des MIDlets durch die Laufzeitumgebung eingeführt. Im Paused Zustand, den das MIDlet unmittelbar nach dem Start einnimmt, ist es noch inaktiv (und sollte deshalb auch noch nicht auf gemeinsam verwendete Ressourcen zugreifen). Soll es dann verwendet werden, wird von der Laufzeitumgebung die startApp()-Methode des MIDlets aufgerufen und das MIDlet ändert seinen Zustand in Active. In diesem Zustand führt das MIDlet dann auch seine eigentliche Funktion aus. Von diesem Zustand kann es von der Laufzeitumgebung oder vom Programmierer durch den Aufruf der pauseApp()-Methode wieder in den Paused-Zustand versezt werden, wenn es gerade nicht mehr verwendet wird aber noch nicht beendet werden soll. Definitiv beendet werden kann das MIDlet dann durch den Aufruf der destroyApp()-Methode (vgl. Abbildung 2.6) . Dabei können noch alle notwendigen Daten persistent gespeichert und belegte Ressourcen freigegeben werden [26]. (vgl. Abbildung 2.7)

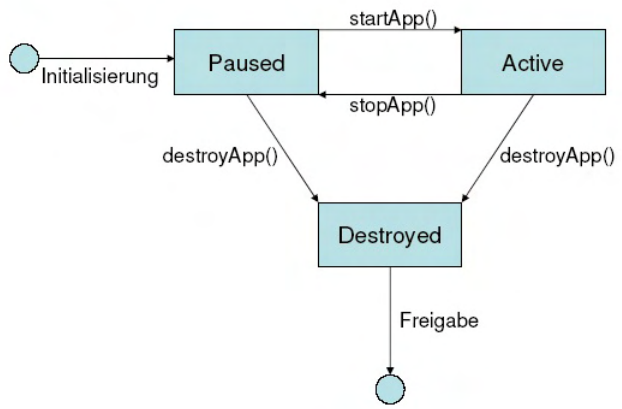


Abbildung 2.6: Der Lebenslauf eines MIDlets [26]

2.7.2 Die Erzeugung eines MIDlet

Um ein lauffähiges MIDlet zu erhalten, muss der Quellcode zuerst kompiliert werden. Danach wird der erzeugte Bytecode von einem Verifier überprüft. Im Gegensatz zur Java SE

```
import javax.microedition.midlet.*;
public class MyMidlet extends MIDlet
{
   public MyMidlet(){...}
   public void startApp()throwsMIDletStateChangeException{...}
   public void pauseApp(){...}
   public void destroyApp(boolean unconditional){...}
}
```

Abbildung 2.7: Codeskelett eines MIDlets

wird diese Überprüfung allerdings nich direkt vor der Ausführung vorgenommen, sondern normalerweise direkt nach der Kompilierung (in der Entwicklungsumgebung). Dadurch werden die knappen Betriebsmittel auf dem mobilen Gerät weniger belastet. Danach wird der Code in ein Java Archiv gepackt. Um den Inhalt des JAR der Application Management Software, die für die Installation, die Deinstallation sowie den Lebenszyklus des MIDlets verantwortlich ist, bekannt zu machen, wird noch eine sogenanntes Java Application Descriptor (JAD) Datei erstellt, die weitere Informationen enthält [25]und um ein digitales Zertifikat ergänzt werden kann.

Das JAR kann dann zusammen mit dem JAD File auf das mobile Gerät übertragen werden (Bluetooth, GPRS, etc). Meist wird der Benutzer mithilfe eines Wizards, der allerdings auf jedem Gerät anders implementiert sein kann, durch die Installation geführt.

2.8 Conclusion

Die technische Entwicklung mobiler Betriebssysteme war und ist geprägt durch die zur Verfügung stehenden, beschränkten Hardwareressourcen. Dabei spielt vor allem die Batterie-Abhängigkeit eine entscheidende Rolle. Neben stromsparender Hardware muss deshalb auch die besonders effiziente Ressourcennutzung durch das Betriebssystem sichergestellt werden. Gut und auch kommerziell erfolgreich scheinen dies vor allem Symbian mit SymbianOS und Microsoft mit WindowsMobile gelungen zu sein.

Viele der dabei angebotenen Funktionalitäten bleiben dabei aber von den Konsumenten unbeachtet. Deshalb stellt sich die Frage inwiefern es wirtschaftlich ist Funktionalität für komplexe Anwendungen, die heute durch zahlreiche Programmiersprachen entwickelt werden können, zur Verfügung zu stellen.

Literaturverzeichnis

[1] Fast facts: Canalys worldwide total smartphone device market - market shares 2006 Q3 2005 / Q3 2006. http://www.symbian.com/about/fastfacts/fastfacts.html, zuletzt abgerufen am 29. Oktober 2006.

[2] Embedded System. http://en.wikipedia.org/wiki/Embedded_system, zuletzt abgerufen am 29. Oktober 2006.

[3] History of Palm OS. http://www.palmsource.com/de/press/fact_sheet.html, zuletzt abgerufen am 29. Oktober 2006.

[4] Why PalmOS. http://www.palmsource.com/palmos/whyPalmOS.html zuletzt abgerufen am 29. Oktober 2006.

[5] History of Symbian. http://www.symbian.com/about/overview/history/history.html, zuletzt abgerufen am 29. Oktober 2006.

[6] History of Embedded Linux. http://www.linuxjournal.com/article/9065, zuletzt abgerufen am 29. Oktober 2006.

[7] Embedded Linux Consortium. http://www.lynuxworks.com/de/corporate/press/2003/elc.php, zuletzt abgerufen am 29. Oktober 2006.

[8] Embedded Linux: Kleingeräte. http://en.wikipedia.org/wiki/Embedded_Linux, zuletzt abgerufen am 29. Oktober 2006.

[9] Embedded Linux Architecture. http://www.palmsource.com/img/cms_mfone_arch.gif, zuletzt abgerufen am 29. Oktober 2006.

[10] Embedded Linux: RTOS. http://www.lynuxworks.com/rtos/rtos.php, zuletzt abgerufen am 29.10.2006.

[11] Java Mobile Operating System. http://java.about.com/b/a/256849.htm, zuletzt abgerufen am 29. Oktober 2006.

[12] Psion Organiser. http://en.wikipedia.org/wiki/Image:Psion_Organiser_II_-_270404.jpg, zuletzt abgerufen am 29. Oktober 2006.

[13] Murmann, T., Rossnagel, H.: How Secure are current mobile OS?. http://www.wiiw.de/publikationen/HowSecureareCurrentMobileOpera968.pdf, zuletzt abgerufen am 29. Oktober 2006.

[14] Symbian Phones. `http://www.symbian.com/phones/index.html`, zuletzt abgerufen am 29. Oktober 2006.

[15] Windows devices. `http://www.microsoft.com/windowsmobile/5/devices/default.mspx`, zuletzt abgerufen am 29. Oktober 2006.

[16] Tanenbaum, Andrew Stuart. - Moderne Betriebssysteme / Andrew S. Tanenbaum ; übers. von Uwe Baumgarten. - München : Pearson Studium, 2002.

[17] Yang Kun, Guo Xin, Liu Dayou: Security in mobile agent system: problems and approaches, ACM SIGOPS Operating Systems Review, Jan. 2000.

[18] Leavitt, Neal: Mobile phones: the next frontier for hackers?, Computer Volume 38, Issue 4, April 2005.

[19] Raatikainen, Kimmo: Operating system issues in future end-user systems, Personal, Indoor and Mobile Radio Communications, PIMRC 2005. IEEE 16th International Symposium on, Volume: 4, Sept. 2005.

[20] Welch, Gregory: A survey of power management techniques in mobile computing operating systems, ACM SIGOPS Operating Systems Review, Volume 29 , Issue 4, Oct. 1995.

[21] Schiller, Jochen. - Mobile communications / Jochen H. Schiller. - London : Addison-Wesley, 2003.

[22] Olsen, Michael, Narayanaswarni, Chandra: PowerNap: an efficient power management scheme for mobile devices, Mobile Computing, IEEE Transactions on, Volume: 5 Issue: July 2006.

[23] Nokia: The Technologie of WiBree, `http://www.wibree.com/technology/`, zuletzt abgerufen am 20. Januar 2007.

[24] Douglis, Fred, P. Krishnan, Marsh, Brian: Thwarting the Power-Hungry Disk, USENIX Winter 1994 Technical Conference Proceedings, Jan 1994.

[25] Jode de, Martin: Programming Java 2 Micro Edition for Symbian OS: A Developer's Guide to MIDP 2.0, John Wiley & Sons, 2004.

[26] Schmatz, Klaus-Dieter: Java Micro Edition, Entwicklung mobiler JavaME-Anwendungen mit CLDC und MIDP, dpunkt, 2006.

[27] Esmertec: `http://www.esmertec.com/`, zuletzt abgerufen am 26 .November 2006.

[28] Java Specification Request: `http://jcp.org/en/jsr/all`, zuletzt abgerufen am 26. November 2006.

[29] Java Specification Request 139: `http://jcp.org/en/jsr/detail?id=139`, zuletzt abgerufen am 26. November 2006.

[30] Operating System. `http://en.wikipedia.org/wiki/Operating_system`, zuletzt abgerufen am 26. November 2006.

[31] Personal Information Manager. `http://en.wikipedia.org/wiki/Personal_Information_Manager`, zuletzt abgerufen am 26. November 2006.

[32] Tommy Thorn. Programming languages for mobile code. ACM Computing Surveys, 29(3):213-239, 1997.

Kapitel 3

WiMAX Wireless Network Technology

Marc Hämmig, Sonja Näf, Martina Vazquez

WiMAX (Worldwide Interoperability for Microwave Access) [10] ist eine Funktechnik für breitbandige Hochgeschwindigkeitsübertragungen im Anschlussnetz. Im Jahre 2001 wurde das WiMAX-Forum gegründet, das die Aufgabe hat die Kompatibilität und Interoperabilität der Produkte, die nach dem IEEE-802.16-Standard produziert werden, zu zertifizieren [1]. In demselben Jahr wurde der erste Standard auf den Markt gebracht. Der grosse Erfolg von WiMAX ist bis jetzt aber ausgeblieben. Möglicherweise wird der neuste Standard IEEE 802.16e, der Mobilität ermöglicht, den Durchbruch erreichen. WiMAX operiert auf dem Frequenzbereich von 2 bis 66 GHz und hat eine Reichweite bis zu 100km und eine Datenübertragungsrate bis zu 100MB/s [2]. Speziell am ISO/OSI-Referenzmodell von WiMAX ist es, dass WiMAX im Vergleich zu anderen Funktechnologien durch eine flexible Kanalbreite und drei Funkschnittstellen ausgestattet ist. Dies ermöglicht den Netzbetreibern die verfügbare Bandbreite möglichst zielgerecht einzusetzen. Als Modulationsverfahren werden OFDMA (orthogonal frequency division multiple access) oder sOFDMA (scalable orthogonal frequency division multiple access) verwendet, die es ermöglichen, dass mehrere Benutzer gleichzeitig auf das Medium zugreifen können. Zur Untersützung von Multimediaanwendungen wurde in WiMAX das Konzept des Quality of Service integriert. Die in der MAC-Schicht beinhaltete Privacy Sublayer soll bei der Datenübertragung Sicherheit gewähren. WiMAX eröffnet neue Möglichkeiten. So ist es nun beispielsweise möglich weitabgelegene Regionen mit einem Internetzugang zu versorgen oder gemütlich in einem Park sitzend an einer Videokonferenz teilzunehmen. Viele grosse Unternehmen sehen, dass sich mit WiMAX ein neuer attraktiver Markt öffnen könnte und investieren darum viel, um die neue Technologie voranzutreiben.

Inhaltsverzeichnis

3.1 Einleitung

WiMAX (Worldwide Interoperability for Microwave Access) [10] ist definiert als eine Funktechnik für breitbandige Hochgeschwindigkeitsübertragungen im Anschlussnetz. Durch den Standard IEEE 802.16 wird WiMAX spezifiziert.

Der Standard IEEE 802.16 wurde im Jahre 2001 auf den Markt gebracht und wurde in den nachfolgenden Jahren weiterentwickelt [10]. Heutzutage, fast fünf Jahre später, ist der ursprüngliche Standard IEEE 802.16 veraltet und die Standards IEEE 802.16-2004 [4] und IEEE 802.16e [4] sind diejenigen, die die meiste Bedeutung haben [4]. IEEE 802.16-2004 wird auch WiMAX fixed genannt und IEEE 802.16e WiMAX mobile [9]. Neben diesen zwei WiMAX Arten gibt es weitere IEEE Standards, die in Anwendung oder in Entwicklung sind. IEEE 802.16a [4], IEEE 802.16b[4], IEEE 802.16c [4] und IEEE 802.16d [4] werden heutzutage benutzt und zwei weitere Standards sind in Entwicklung [4]. Der Standard IEEE 802.16a spezifiziert den lizenzierten und unlizenzierten Frequenzbereich zwischen 2 und 11 GHz, der Standard IEEE 802.16b beschäftigt sich mit dem Frequenzbereich zwischen 5 und 6 GHz und den Quality of Service. Der Standard IEEE 802.16c repräsentiert den Frequenzbereich zwischen 10 und 66 GHz und der Standard IEEE 802.16d ist eine Verbesserung des Standards IEEE 802.16a.

WiMAX ist eine Weiterentwicklung der WLAN Technologie und erreicht höhere Reichweiten und Übertragungsraten als WLAN. Indem WLAN üblicherweise Reichweiten von 30 bis 100 Metern hat, sind die Reichweiten von WiMAX bis zu 50 Kilometern. WLAN operiert auf unlizenzierten Frequenzen und kann von jedem gebraucht werden. Jedermann kann in ein Geschäft gehen und einen Access Point erstellen und dann zu Hause die Möglichkeiten von WLAN nutzen. WiMAX hingegen operiert sowohl auf unlizenzierten wie auch auf lizenzierten Frequenzen.

Abbildung 3.1: Fieberkurve [5]

Die heutige Positionierung von WiMAX wird von Marktanalytikern in einer Fieberkurve betrachtet [5] . Es ist eine Graphik, bei der die x-Achse den Reifegrad der entsprechenden Technologie bezeichnet und die y-Achse den Aufmerksamkeitsgrad. Gemäss den Marktanalytikern hat WiMAX momentan einen grossen Aufmerksamkeitsgrad und befindet sich im Reifegrad in der Phase der Euphorie. Das Tal der Tränen sagt man, würde noch bevorstehen. Mit dem Tal der Tränen meint man, dass die Begeisterung für WiMAX nicht

mehr in dem Masse vorhanden ist, wie zu Beginn. In der Abbildung 3.1 ist die Kurve graphisch dargestellt.

Anhand zweier Beispiele soll die Funktionsweise von WiMAX fixed und WiMAX mobile erläutert werden. Beim ersten Beispiel geht es um die Funktionsweise von WiMAX fixed. In der Abbildung 3.2 wird die Funktionsweise des WiMAX fixed dargestellt.

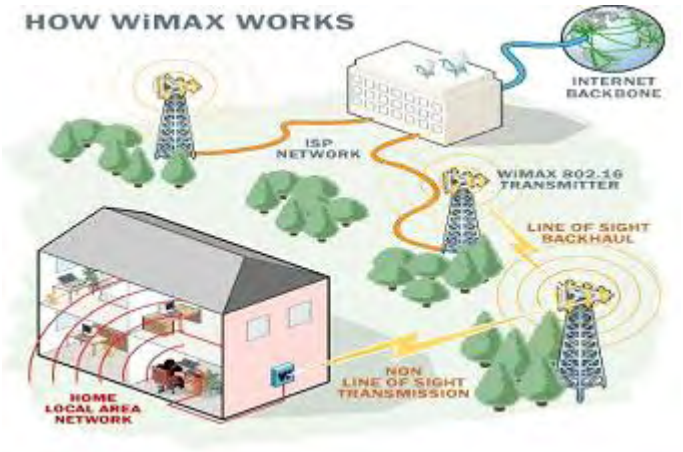


Abbildung 3.2: WiMAX fixed [3]

Ein Internetbenutzer schickt einem Geschäftpartner eine Nachricht. Die Nachricht wird an einen WiMAX 802.16 Transmitter geschickt mittels einer Non-Line-Of-Sight Übertragung. In der drahtlosen Kommunikation ist die Non-Line-Of-Sight eine nicht direkte Sichtverbindung zwischen dem Sender und Empfänger, d.h. es kann Hindernisse wie Bäume, Gebäude zwischen dem Sender und dem Empfänger haben und die Nachricht kommt dennoch an. Nachdem die Nachricht beim WiMAX 802.16 Transmitter angekommen ist, wird sie an einen weiteren WiMAX 802.16 Transmitter geschickt mittels Line-Of-Sight Backhaul. Die Line-of-Sight ist die direkte Sichtverbindung zwischen dem Sender und dem Empfänger, dies heisst, dass es keine Hindernisse zwischen dem Sender und Empfänger geben darf. Der Begriff Backhaul bezeichnet die Anbindung von einem Netzknoten an einen zentralen Netzknoten. Die Nachricht wird in das ISP Netzwerk, das Internet Service Provider Netzwerk, geschickt. Dann wird sie weitergeleitet an die Internet Backbone. Da wird der Empfänger der Nachricht gesucht und lokalisiert. Der soeben beschriebene Transfer zwischen den Transmitter spielt sich nun in der umgekehrten Reihenfolge ab. Die Nachricht wird nun auf einem ähnlichen Weg wie sie vom Sender geschickt wurde zum Empfänger gelangen.

Das zweite Beispiel basiert auf der Unterscheidung zwischen WiMAX fixed und WiMAX mobile. In der Abbildung 3.3 wird dies durch die drei blauen Felder dargestellt. Das linke und das mittlere Feld repräsentieren WiMAX fixed und das mittlere und das rechte Feld WiMAX mobile.

Der Unterschied zwischen diesen beiden Arten von WiMAX ist derjenige, dass bei WiMAX fixed der Anwender während der Anwendung nur innerhalb einer Zelle eine Verbindung zum AP (Access Point) haben kann. Die Zelle wird eigentlich dadurch definiert. Begibt sich der Benutzer ausserhalb der Funkzelle, dann wird der Dienst unterbrochen. Beim

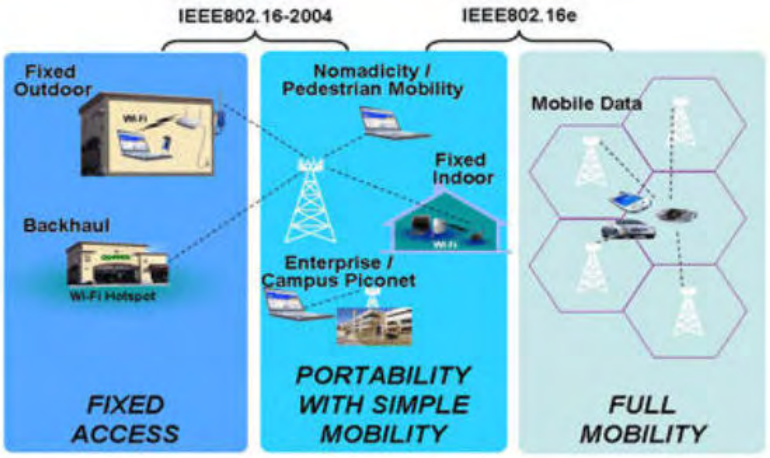


Abbildung 3.3: WiMAX fixed und mobile [7]

WiMAX mobile kann sich der Benutzer während der Anwendung von einer Funkzelle in die andere bewegen und es geschieht kein Unterbruch.

3.2 Architektur

Die Architektur vom Standard IEEE 802.16 basiert auf der Kommunikation zwischen einer BS (Basisstation) und mindestens einer SS (Subscriber Station). Wie im Folgenden gesehen werden kann, werden verschiedene Aspekte in Betracht gezogen.

3.2.1 Überblick über die verschiedenen Standards

Die Tabelle 3.1 teilt den Standard IEEE 802.16 in drei Spalten, die einzelne Standards repräsentieren [4].

In der ersten Spalte ist der Standard IEEE 802.16, in der zweiten Spalte die Standards IEEE 802.16a, IEEE 802.16REVd, IEEE 802.16-2004 und in der dritten Spalte der Standard IEEE 802.16e zu sehen. Der Standard IEEE 802.16 ist der ursprüngliche Standard, der heutzutage der älteste ist. Die Standards IEEE 802.16a, IEEE 802.16REVd und IEEE 802.16-2004 sind Standards des WiMAX fixed. Der Standard IEEE 802.16e repräsentiert den Standard für WiMAX mobile, der heutzutage noch in Bearbeitung steht. Diese drei Spalten werden anhand der Charakteristiken Fertigstellung des Standards, Verfügbarkeit der Produkte, Frequenzbereich, Übertragungsart, maximale Datenrate, Bandbreiten, Modulationsarten, Empfänglichkeit und maximale Reichweite verglichen.

Es kann festgestellt werden, dass der IEEE 802.16 im Dezember 2001, die Standards IEEE 802.16a, IEEE 802.16REVd und IEEE 802.16-2004 zwischen Januar 2003 und Juli 2004 und der Standard IEEE 802.16e im Dezember 2005 fertig gestellt wurden.

Tabelle 3.1: Überblick über die verschiedenen Standards [4]

	IEEE 802.16	IEEE 802.16a, IEEE 802.16 REVd, IEEE 802.16-2004	IEEE 802.16e
Fertigstellung	im Dezember 2001	zwischen Januar 2003 und Juli 2004	im Dezember 2005
Produkte	verfügbar	ab Mitte 2005	ab 2006
Frequenzband	10 bis 66 GHz	2 bis 11 GHz	0,7 bis 6 GHz
Übertragung	LOS	Near LOS, Non LOS	Non LOS
max. Datenrate	32 bis 134 Mbit/s	bis zu 75 Mbit/s	bis zu 15 Mbit/s
Bandbreiten	20, 25 und 28 MHz	skalierbar von 1,5 bis 20 MHz	28 MHz
Modulationsarten	QPSK, 16QAM, 64QAM	OFDM256, OFDMA, 64QAM, 16QAM, QPSK, BPSK	OFDM256, OFDMA, 16QAM, 64QAM
Empfänglichkeit	fest	feste Außenantenne und Innenanwendung	mobil
max. Reichweite	variabel, ? 100km	bis zu 50 km	bis zu 5km, typisch 1,5km

Zu der Verfügbarkeit der Produkte, der älteste Standard ist verfügbar, die aktuellen Standards sind ab Mitte 2005 verfügbar und der mobile Standard wird ab 2006 verfügbar sein. Das Frequenzband liegt beim ältesten Standard zwischen 10 und 66 GHz, die aktuellen Standards liegen im Bereiche von 2 bis 11 GHz und der mobile Standard wird zwischen 0,7 und 6 GHz liegen.

Bei der Charakteristik Übertragungsart gebraucht der älteste Standard LOS (Line-of-Sight), die aktuellen Standards gebrauchen Near LOS, NLOS (Non-Line-of-Sight) und der mobile Standard wendet NLOS an. Zu den maximalen Datenraten, beim ältesten Standard liegt der Bereich zwischen 32 bis 134 Mbit/s, bei den aktuellen Standards ist es bis zu 75 Mbit/s und beim mobilen Standard bis zu 15 Mbit/s. Die Unterscheidung der Bandbreiten zeigt folgendes auf: der älteste Standard liegt bei 20, 25 und 28 MHz, die aktuellen Standards sind skalierbar von 1, 5 bis 20 MHz und der mobile Standard liegt bei 28 MHz.

Die nächste Charakteristik sind die Modulationsarten, auf die in einem weiteren Kapitel eingegangen wird. Dann zur Empfänglichkeit, der älteste Standard ist fest die aktuellen Standards haben eine feste Aussenantenne und Innenanwendung und der mobile Standart hat mobile Empfänglichkeit. Zur maximalen Reichweite, beim ältesten Standard ist sie variabel, bis zu 50 km, typisch sind es mit einer Aussenantenne 15km und bei einer Innenantenne 5km, bei den aktuellen Standards sind es bis zu 5km, typisch 1,5 km beim mobilen Standard.

3.2.2 Netzwerkarchitektur

Die ursprüngliche Idee von WiMAX war es, dass auf den Dächern von Häusern und Gebäuden Sende- und Empfangsgeräte in Form einer Basisstation mit dem Internet verbunden sein würden. Jede Basisstation sollte die WiMAX Technologie benutzen, um Daten

zu senden und zu empfangen. Geforscht wird heutzutage, ob die Antennen auch innerhalb des Hauses sein könnten. Um den Sender mit dem Empfänger zu verbinden gibt es mehrere Netzwerkarchitekturen, die angewendet werden können. Point-to-Point, Point-to-Multipoint und Mesh sind drei solche Netzwerkarchitekturen [11].

- Point-to-Point

  Die Point-to-Point Architektur ist definiert als eine Architektur, die sich mit einer direkten, unmittelbaren Verbindung zwischen zwei Punkten oder Orten beschäftigt. Die Point-to-Point Architektur ist an sich keine Netzwerkarchitektur, da sich dieses Netz nur auf zwei Parteien beschränkt. Früher wurde dieses Netzwerk für weit entfernte Telefongespräche und Datenübertragung benutzt.

- Point-to-Multipoint

  Die Netzwerkarchitektur Point-to-Multipoint ist die Architektur, die sich mit einer Verbindung beschäftigt, die mehrere Benutzer hat. Für Netzwerke mit tieferen Microwave Frequenzen wird die Point-to-Multipoint Architektur angewendet. Mittels der Point-to-Multipoint Architektur erhält der Provider die grösste Anzahl von Kunden, zahlt dafür wenig und begrenzt die Anzahl von Router und Switches, die er benötigt, um ein Netzwerk aufzustellen. Problematisch ist aber die Topologie der grossen Städte. Eine Basisstation kostet mindestens 100 000 Dollar und könnte nicht von allen Bewohnern einer Grosstadt benutzt werden, weil sie sonst überfordert sein würde. Der Provider müsste somit mehrere Basisstationen in einer Grosstadt aufstellen, was zu sehr hohen Kosten führen würde.

- Mesh

  Mesh bezeichnet die Struktur der Verbindungen mehrerer Geräte untereinander, um einen gemeinsamen Datenaustausch zu gewährleisten. Die Geräte, die hier verbunden werden, haben nicht mehr eine Basisstation von welcher die ganze Information kommt, sondern die Verbindungen, die hier existieren, sind zwischen Geräten, die die gleiche Wichtigkeit im Netz haben. Fällt ein Gerät aus, kann die Information über andere Wege weitergeleitet werden. Die Netzwerkarchitektur ist somit robust. Es stellte sich jedoch heraus, dass es das teuerste Verfahren ist, ein Mesh Netzwerk aufzubauen. Da es bis anhin nur wenige Anwendungen der Mesh Architektur gab, ist es schwierig zu sagen, in welchen Situationen es besser sein würde ein Mesh Netzwerk zu implementieren und nicht eine Point-to-Multipoint Architektur.

3.3 Bitübertragungsschicht (PHY)

Die 1. Schicht im ISO/OSI-Referenzmodell von WiMAX zeichnet sich im Vergleich zu anderen Funktechnologien durch eine flexible Kanalbreite (1.75MHz - 20MHz) und drei Funkschnittstellen aus. Aufgrund dieser flexiblen Kanalbreite kann ein Netzanbieter die verfügbare Bandbreite möglichst zielgerecht einsetzen. Die Abbildung 3.4 gibt einen Überblick über die zwei untersten Schichten von WiMAX, und man erkennt, dass die MAC Schicht aus mehreren Teilschichten besteht. Zuerst wird jedoch die physikalische Schicht und später im darauf folgenden Kapitel die MAC Schicht erläutert.

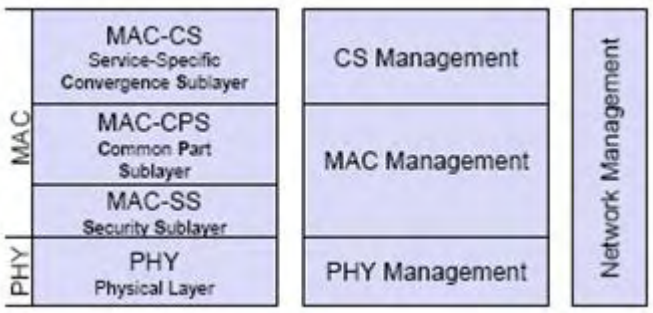


Abbildung 3.4: Überblick PHY und MAC Layer [12]

In der physikalischen Schicht gibt es verschiedene Arten von drahtlosen Breitband-Luftschnittstellen wie z.B. Single Carrier (SC), Orthogonal Frequency Division Multiplexing (OFDM), Orthogonal Frequency Division Multiple Access (OFDMA) und scalable Orthogonal Frequency Division Multiple Access (sOFDMA), welche im folgenden Kapitel ausführlicher erklärt werden. Im Markt hat man sich für OFDMA (oder sOFDMA) entschieden, weil dieses Modulationsverfahren die größten Vorteile mit sich bringt.

Tabelle 3.2: Überblick Funkschnittstellen [37]

802.16-2004 (fixed)	**802.16e (mobile)**
Single Carrier	Single Carrier
OFDM 256 FFT	OFDM 256 FFT
OFDM 2048 FFT	OFDM 2048 FFT
	sOFDMA 1024 FFT
	sOFDMA 512 FFT
	sOFDMA 128 FFT

3.3.1 Modulationsverfahren

In einem single carrier System werden digitale Signale in Form von Bits oder Ansammlungen von Bits (Symbole) nur auf einen Träger moduliert. Diese Methode hat den großen Nachteil, dass der Datenstrom aufgrund der hohen Symbolrate sehr störanfällig auf Mehrwegausbreitung ist. Um dieses Problem zu lösen, wendet man OFDM an.

OFDM ist eine Übertragungstechnik, die ideal für eine schnelle drahtlose Kommunikation von Daten ist. Sie wurde erstmals in den sechziger Jahren angewendet, ist aber erst kürzlich sehr populär geworden. Die Daten werden gleichzeitig übertragen, wobei mehrere modulierte Träger sich gegenseitig überlappen. Das ist nur möglich, wenn die Subfrequenzen orthogonal gehalten werden. D.h. wenn eine Subfrequenz ihr Amplitudenmaximum erreicht hat, alle anderen einen Nulldurchgang aufweisen. Demzufolge sind spektrale Überlappungen der einzelnen Subfrequenzen möglich. Diese Technologie wird sowohl in ADSL (Asymmetric Digital Suscriber Line) als auch in drahtlosen Systemen wie IEEE 802.11a/g

(Wi-Fi) und IEEE 802.16 (WiMAX) angewendet [15] [16]. WiMAX bietet entweder 256 oder 2048 Teilfrequenzen (Unterträger) an, was deutlich mehr sind als beim Standard 802.11, der nur über 64 Teilfrequenzen verfügt. Eine wichtige Eigenschaft von OFDM ist, dass die Datenübertragung durch die Aufteilung des Datenstroms auf mehreren Trägern sehr robust ist und speziell bei NLOS Verbindungen eingesetzt wird. Die Empfängerseite kann somit ein Signal besser auswerten, weil es auf die Mehrwegausbreitung (Reflexionen, Streuung oder Beugung) aufgrund der tiefen Symbolrate weniger anfällig ist.

OFDMA ist ein Medienzugriffsverfahren, das mehreren Benutzern gleichzeitig einen Zugriff auf das Medium ermöglicht. Eine weitere Kanalaufteilung wird mit Hilfe von Time Divison Multiplexing (TDM) als Downlink und Uplink eingesetzt. Die Basisstation kann so einem Teilnehmer einen Zeitschlitz zur Verfügung stellen, indem er Daten senden oder empfangen kann. Da es mehrere Teilnehmer in einer Zelle gibt und eine Basistation teilen müssen, werden mit TDMA die Zugriffe verwaltet.

Scalable OFDMA ist eine weitere Variante von OFDM, welche im mobilen WiMAX angewendet wird. Mit ihr können verschiedene Kodierungsarten (BPSK, QPSK, QAM) für die unterschiedlichen Teilkanäle angeboten werden. Die Funktionsweise dieser verschiedenen Kodierungsarten werden nun ein wenig genauer betrachtet.

- PSK (Phase Shift Keying)

  Zwei wichtige Modulationsverfahren werden für eine robuste Datenübertragung eingesetzt. Das BSPK erreicht eine Datenübertragungsrate von 1Mbit/s. In der Abbildung 3.5 sieht man, wie genau zwei Symbole (0 und 1) anhand der Phasenverschiebung, die entweder 0 Grad oder 180 Grad beträgt, kodiert werden können. Beim QPSK können vier Zustände (Symbole) dargestellt werden, die jeweils zwei Bits enthalten. Mit dieser Technik kann die Kapazität der Datenübertragung auf 2Mibt/s gesteigert werden, wie man in Abbildung 3.6 sehen kann. Beide Verfahren sind robuster als das Frequency Shift Keying (FSK), weil Geräusche oder Lärm die Energie des Trägers, also die Amplitude, verändern könnten.

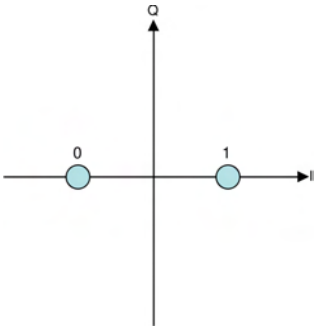


Abbildung 3.5: BPSK [17]

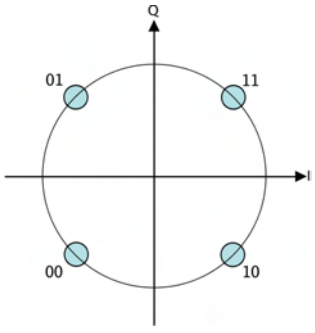


Abbildung 3.6: QPSK [18]

- QAM (Quadrature Amplitude Modulation)

  Dieses Modulationsverfahren ist eine Kombination von Amplituden- und Phasenmodulation und die häufigsten Formen sind 16QAM und 64QAM. In Abbildung 3.7 wird gezeigt, dass die Symbole bei QAM in einem Quadrat angeordnet sind.

Die zwei signifikantesten Bits werden in QPSK kodiert, welches wiederum in QAM eingebettet ist.

Ein gutes Beispiel, um die Funktionsweise von QAM zu erklären, ist die Fernsehübertragung. Die Standardauflösung eines Fernsehers wird in QPSK kodiert und wird mit einer hohen Priorität eingestuft, die hohe Auflösung jedoch mit einer niedrigen. Die Daten mit der niedrigen Priorität befinden sich in QAM-Teil und sind in den letzten 4 Bits kodiert. Ist bei der Datenübertragung ein schlechter Empfang vorhanden, wird nur der QPSK-Teil dekodiert und die Daten für die hohe Auflösung werden verworfen. Mit diesem Vorgehen kann trotz der schlechten Verbindung ein Bild mit einer Standardauflösung übertragen werden [13].

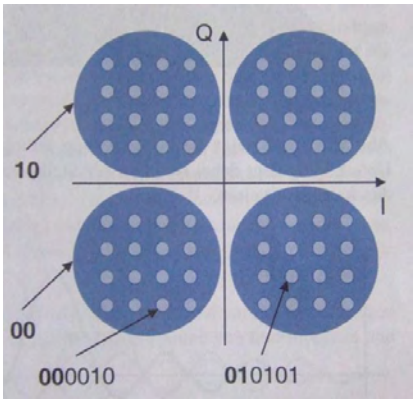


Abbildung 3.7: Quadrature Amplitude Modulation QAM64 [13]

3.3.2 Frequenzen

Der Standard 802.16 bietet einen Frequenzbereich zwischen 2 und 66 GHz [10]. Denjenigen Frequenzbereich zu wählen, der am besten dem entsprechenden Gebiet zusagt, ist schwierig. Der Grund dafür ist, dass die Frequenzbereiche des Standards 802.16 nicht auf den Eigenschaften der Umwelt basieren, sondern auf den Frequenzbereichen des Deployments der reicheren Staaten.

Der Standard IEEE 802.16 teilt die Frequenzbereiche in drei verschiedene Bereiche [4]:

- Frequenzen zwischen 10 und 66 GHz

  Der erste Bereich ist derjenige von 10 bis 66 GHz. Er bietet eine Umgebung mit kurzen Wellenlängen an. Benötigt wird dabei eine direkte Sichtlinie zwischen dem Sender und dem Empfänger. Die Mehrwegausbreitung wird somit reduziert. Die Reduktion von der Mehrwegausbreitung erhöht die Bandbreite. In diesem Frequenzbereich sind Bandbreiten von 25 MHz oder 28 MHz typisch.

- Frequenzen zwischen 2 und 11 GHz

  Der zweite Frequenzbereich liegt zwischen 2 und 11 GHz. Dieser tiefere Frequenzbereich wird im Standard IEEE 802.16a, IEEE 802.16REVd und IEEE 802.16-2004

geregelt. Frequenzen unter 11 GHz haben längere Wellenlängen und können durch Hindernisse gehen. Sie stellen somit eine physikalische Umgebung zur Verfügung, in welcher LOS zwischen Sender und Empfänger nicht notwendig ist und die Mehrwegausbreitung eine wichtige Rolle spielen kann. Die Fähigkeit Near LOS und NLOS zu unterstützen, erlaubt es flexiblere WiMAX Implementierungen zu machen.

- Frequenzen zwischen 0.7 und 6 GHz

  Der dritte Frequenzbereich ist zwischen 0.7 und 6 GHz und wird im Standard IEEE 802.16e spezifiziert. Auch bei diesem Frequenzbereich wird die Fähigkeit NLOS zu unterstützen gegeben.

Es gibt sowohl lizenzierte wie auch unlizenzierte Frequenzen [11]. Bei den lizenzierten Frequenzen haben die Provider einen zugeteilten Frequenzbereich, auf dem sie operieren. Die unlizenzierten Frequenzen basieren auf dem Konzept des Open Spectrum. Das Konzept des Open Spectrum besagt, dass das Spectrum, welches existiert nicht privat sein soll, sondern, dass es alle nutzen können sollen. Jeder, der bestimmte Kriterien erfüllt, kann sie gebrauchen. Diese bestimmten Kriterien berufen sich auf die Fähigkeit Beeinträchtigungen zu tolerieren und flexibel zu sein, so dass nicht ein bestimmter Teil des Spectrums nur von einem einzigen Anwender besetzt wird.

Bei den Anwendungen der Frequenzbereiche werden drei Arten unterschieden: die Möglichkeit des Static Use mit LOS und NLOS, die Möglichkeit des Nomadic Use und die Möglichkeit des Mobile Use [8].

- Möglichkeit A: Static Use (LOS, NLOS)

  In der Anwendung des Static Use wird unterschieden zwischen den beiden Frequenzbereichen 2-11 GHz und 11-66 GHz. Der Bereich von 2-11 GHz verlangt NLOS Operationen und Richtantennen. Dazu wird eine Set-Top Box auf dem entsprechenden Gebäude installiert mit einer Antenne. Wegen des Errichtens der Anmeldestation ist die Mobilität nicht richtig vorhanden, aber man kann die Einrichtung innerhalb der Abdeckung der Zelle neu positionieren. Der Frequenzbereich von 11-66 GHz hingegen gebraucht LOS Operationen und Richtantennen. Auch hier wird eine Set-Top Box installiert aber ausserhalb des Gebäudes. Obwohl die Basisstation einen Sendebereich von 50 km hat, ist keine Änderung der Anmeldestation möglich, da die Richtantenne meistens ausserhalb des Gebäudes angebracht wird und die Position nicht ändern kann.

- Möglichkeit B: Nomadic Use

  Die Nomadic Anwendung ist nur möglich bei NLOS Operationen und ist in den Standards IEEE 802.16a, IEEE 802.16REVd und IEEE 802.16-2004 spezifiziert. Nomadic Anwendungen erlauben dem Benutzer die Zelle zu wechseln. Die Zellgrösse ist gleich oder wenigstens vergleichbar zum vorherigen Fall und beträgt 500m-15km. Roaming ist möglich, aber es gilt zu bedenken, dass die IP- Konnexion verloren gehen kann, wenn die dienende Zelle verlassen wird. Nach der Registrierung in der neuen dienenden Zelle, können wieder Dienstleistungen bezogen werden.

- Möglichkeit C: Mobile Use

  Die Mobile Anwendung wird spezifiziert durch den Standard IEEE 802.16e. Sie ist heutzutage noch nicht in Anwendung. Am Anfang wird sie auf den Frequenzbereichen 2.3 GHz, 2.5 GHz, 3.3 GHz und 3.4 GHz bis 3.8 GHz operieren. Weitere Frequenzbereiche sollen je nach Marktnachfrage dazugetan werden. Bezogen werden kann der Frequenzbereich nicht auf lizenzierte Bandbreiten. Die Mobile Anwendung erlaubt es dem Anwender sich in einer unbegrenzten Region zu bewegen. Die gezielte Maximalgeschwindigkeit der Übertragung beträgt bis zu 125 km/h und der Zellradius liegt zwischen einigen hundert Metern und 15 km. Die Grösse für einen typischen Zellradius ist 7 km.

3.4 Medienzugriffsschicht (MAC)

Da in einem Funknetzwerk, ein Medium geteilt wird, braucht es dafür einen Mechanismus für die Zugriffssteuerung. Im Unterschied zu anderen Funktechnologien wie z.B W-LAN bietet die 2. Schicht des OSI-Referenzmodells bei WiMAX ein wichtigen zusätzlichen Dienst an, nämlich den Quality of Service. Zuerst wird der Aufbau der MAC Schicht aufgezeigt. Die Abbildung 3.8 zeigt welche Teilschichten in der MAC-Schicht existieren.

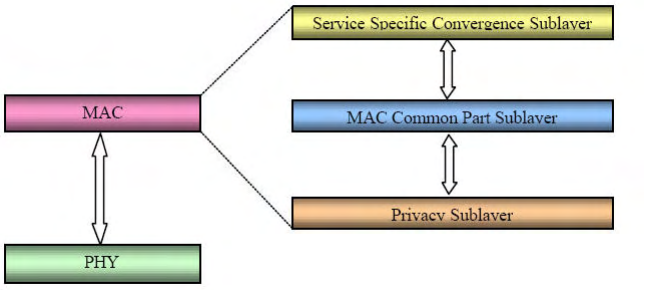


Abbildung 3.8: 3 Teilschichten der MAC-Schicht [46]

Der MAC Service-Specific Convergence Sublayer (MAC-CS), ist die Konvergenzschicht für den Transport von Schicht-2 bzw. Schicht-3 Dateneinheiten und dient als Adapter zwischen WiMAX und darüber liegenden Schichten.

Der MAC Common Part Sublayer (MAC-CPS) ist zuständig für folgende Aufgaben wie z.B. die Kanalzugriffssteuerung, die Fehlererkennung und -behebung mittels ARQ-Verfahren, die Bereitstellung mehrerer Dienstgüteklassen (Endgerät kann Übertragungskapazität bei Basisstation anfordern) oder die Verbindungsverwaltung.

Zu guter letzt gibt es noch den MAC Security Sublayer (MAC-SS), der die Aufgabe der Authentifizierung und der Verschlüsselung übernimmt. Bei der Authentifizierung meldet sich jeder Benutzer mit dem vom Hersteller signierten Zertifikat bei der Basisstation an und kann im erfolgreichen Fall eine Verbindung herstellen [14].

Die angewendeten Verschlüsselungsverfahren sind der Tripel Data Encription Standard (TDES) und der Advanced Encryption Standard (AES), der als Nachfolger von DES

bzw. TDES bekannt gegeben wurde. DES galt mit einer Schlüssellänge von 56 Bit nicht mehr ausreichend sicher gegen Brute-Force-Attacken. Mit der dreifachen Anwendung von DES, eben TDES, kann man die Schlüssellänge auf 112 Bit erhöhen, jedoch hat dies einen negativen Einfluss auf die Geschwindigkeit [45]. Ein zusätzlicher wichtiger Dienstgüteparameter, der von der MAC Schicht angeboten wird, ist QoS (Quality of Service). In der heutigen Zeit, in der viele Multimedia-Anwendungen übers Internet angeboten werden, ist QoS notwendig und wird daher im Verlaufe der Arbeit genauer beschrieben, in dem die Funktionsweise und die Anwendung von QoS aufgezeigt wird.

3.4.1 ARQ

Automatic Repeat Request (ARQ) ist ein Dienst, der eine zuverlässige Datenübertragung innerhalb eines Netzwerkes ermöglicht und wird in drei verschiedenen Verfahren angewendet. Der ARQ Mechanismus ist ein optionaler Teil der MAC-Schicht in WiMAX. Falls er zum Einsatz kommt, wird die Selective Repeat Methode defaultmässig angewendet.

Beim Stop-und-Wait Verfahren wird ein Paket verschickt und der Sender wartet solange bis dieses vom Empfänger quittiert wird, bevor weitere Pakete gesendet werden können. Nach einer erfolgreichen Übertragung wird das nächste Paket verschickt. Trifft der Fall ein, dass ein Paket unterwegs verloren gegangen ist, findet ein Timeout statt und der Sender schickt dasselbe Paket nochmals.

Eine weitere Methode wird Go-Back-N genannt. Der Sender schickt kontinuierlich anstatt nur ein Paket mehrere Pakete nacheinander innerhalb einer festgelegten Fenstergröße. Wenn ein Paket verloren gegangen ist, muss der Sender alle Pakete ab dem verloren gegangenen nochmals senden. Ein grosser Nachteil ist hier die verschwendete Bandbreite. Gewisse Pakete werden u.U. wiedeholt gesendet, obwohl sie gar nicht verloren gegangen sind.

Beim letzen Verfahren, dem Selective Repeat ARQ, sind alle Pakete in einem Fenster nummeriert und werden auch kontinuierlich übertragen. Wenn ein Paket verloren geht oder fehlerhaft beim Empfänger ankommt, dann werden nur die fehlerhaften Pakete innerhalb eines Fensters wiederholt.

Wie man sehen kann ist das Selective ARQ und Go-Back-N Verfahren ähnlich, aber mit dem Unterschied, dass die dritte Methode effizienter in ihrer Arbeitsweise ist [19].

3.4.2 APC

Automatic Power Control ist ein weiterer Dienst, der einen Kanal optimiert. Ein Algorithmus ist in der Basisstation implementiert und wird verwendet, um die Leistung des Systems gesamthaft zu steigern und den Energieverbrauch zu senken. Bei den Endgeräten wird die Sendeleistung an ein vorgegebenes Niveau der Basisstation angepasst. Auf diesem Weg kann ein minimaler Energieverbrauch erzielt werden, denn dieser ist bei OFDMA ziemlich hoch. Die Energie, die von einem Benutzer gebraucht wird, ist abhängig von der Mehrwegausbreitung und sollte deshalb angepasst werden [20].

3.4.3 Adaptive Modulation

Bei der adaptiven Modulation werden die Übertragungsmethoden anhand des Kanalzustands ermittelt. Befindet sich eine Person weit entfernt von einer Basisstation, ändert sich das Modulationsschema. Die Verbindung wird robuster, aber die Durchsatzrate sinkt. In diesem Falle wird BSPK benützt. In der Nähe einer Funkantenne, wo der Empfang grundsätzlich gut ist, können Modulationsverfahren, die eine schnellere Datenübertragung erlauben, eingesetzt werden.

Die verfügbare Durchsatzrate hängt aber auch noch von anderen Schlüsselelementen ab, wie z.B. von der verfügbare Bandbreite oder dem Backhaul. WiMAX ist abhängig von der adaptiven Modulation, um eine robuste Abdeckung zu gewährleisten [6]. Die Abbildung 3.9 zeigt in welchen Zonen die verschiedenen Modulationsverfahren angewendet werden.

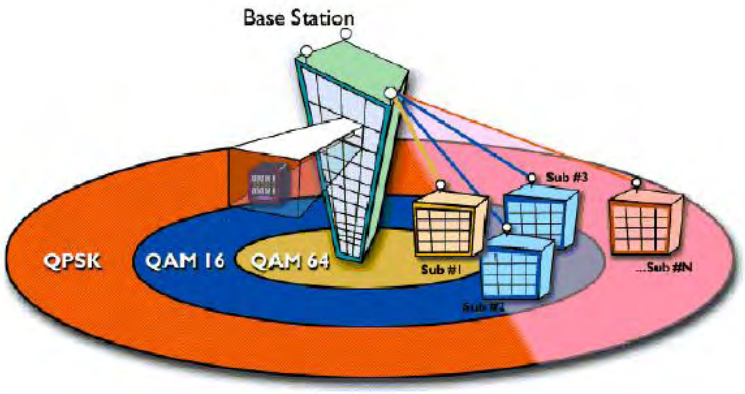


Abbildung 3.9: Adaptive Modulation [6]

3.4.4 QoS

Warum QoS? Herausforderungen Dank neuer Technologien, die hohe Übertragungsraten versrpechen, werden drahtlose Bandbreitanschlüsse als Alternative zu Kabelanschlüssen wie xDSL angesehen. Es wird erwartet, dass Applikationen wie VoIP, Video Streaming oder Videokonferenzen einwandfrei funktionieren, ohne dass man sich mit flimmernden Bildern oder stockenden Gesprächen begnügen muss. In kabelgebunden LANs ist dies weniger ein Problem, da man sehr grosse Bandbreiten hat und das Kabel ein relativ zuverlässiges Medium mit einer kleinen Paketsverlustrate ist. Doch in drahtlosen Netzwerken ist die Bandbreite limitiert und die Übertragung über Luft ist viel unzuverlässiger. Die Qualtität wird stark von der Umgebung beeinflusst. So können das Wetter, Gebäude, andere Geräte usw. Störungen hervorrufen. Ausserdem sind die Gegebenheiten in drahtlosen Netzwerken auch nicht konstant und ändern sich über die Zeit und je nach Ort.

Wie eben geschildert, tauchen bei der Unterstützung von Multimediaanwendungen in drahtlosen Netzen einige Probleme auf. Darum braucht es speziell dafür vorgesehene Mechanismen. Das Konzept des Quality of Service stellt Mechanismen zur Verfügung, die den

Zugang und Nutzung des Mediums kontrollieren, abhänging von der Art der Applikation [21].

Beispiele für unterschiedliche QoS-Anforderungen:

- bei Sprachübertragungen ist man auf kleine Latenzzeiten angewiesen –> Applikation hat höhere Priorität das Medium zu nutzen
- bei Videokonferenzen braucht man hohe Bandbreiten –> es wird eine längere Übertragungszeit gewährt
- bei Datenübertragungen ist die Zuverlässigkeit wichtig –> es wird eine kleine Paketsverlustrate garantiert

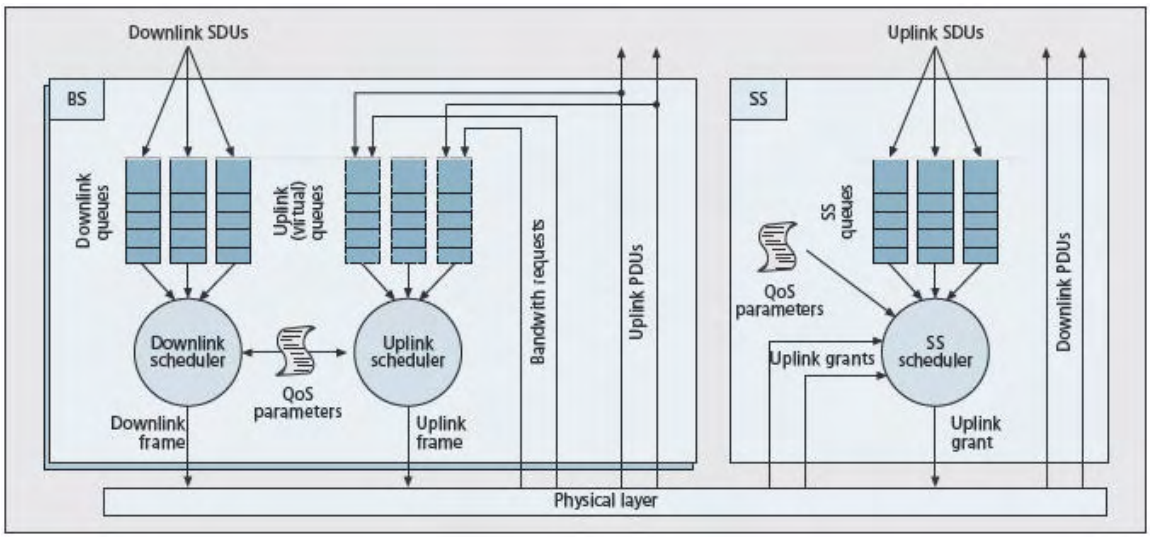


Abbildung 3.10: Architektur QoS [21]

Request and Grant Mechanismen Um QoS in IEEE 802.16 zu garantieren, wurden sogenannte Service Flows definiert. Diese ermöglichen es Pakete unidirektional in upload- oder downlink Richtung zu transportieren und zwar so, dass gewisse QoS-Paramter eingehalten werden [22]. Abbildung 3.10 zeigt die Architektur, die es ermöglicht QoS zu garantieren. Jede Downlink-Verbindung hat in der Basistation eine Queue. Mittels des Downlink Schedulers ermittelt die Basisstation welches Service Data Unit als nächstes gesendet werden soll. Die Wahl des nächsten Packets wird anhand der verhandelten QoS Parametern ermittelt. In der Download Richtung erhalten alle Teilnehmer die gleichen Pakete und wählen die gewünschten aus. Damit sie aber Daten zur Basisstation senden können (Upload Richtung), müssen sie erst Bandbreite reservieren. Dies ist nötig, damit die Basisstation abschätzen kann, wie viel Bandbreite diese Verbindung braucht und somit so anhand der QoS - Parametern sowie der Bandbreitenanfragen Bandbreiten garantieren. Im Standard IEEE 802.16 wurden vier verschiedene Bandwidth-Request-Mechanismen spezifiziert:

- Unsolicited Granting

  Während der Verbindungsphase wird für die Upload-Verbindung eine fixe Bandbreite, die periodisch zur Verfügung steht, garantiert. Nach dieser Phase muss der Teilnehmer nicht mehr explizit Bandbreite nachfragen.

- Unicast Poll

  Die Basisstation fragt einen Teilnehmer, ob er Bandbreite braucht. Damit er diese Anfrage beantworten kann, stellt ihm die Basisstation die dafür notwendige Bandbreite zu Verfügung. Wenn der Teilnehmer schon Bandbreite reserviert hat oder wenn er nichts zu übertragen hat, gibt er keine Anwort.

- Broadcast Poll

  Die Basistation fragt nicht nur einen Teilnehmer, ob Bandbreite benötigt wird, sondern alle. Der Nachteil dieser Variante ist, dass wenn mehrere Teilnehmer auf die Anfrage antworten, Kollisionen entstehen.

- Piggybacked

  Der Teilnehmer kann auch eine Bandbreitenanfrage einem Packet anhängen. Dies ist aber nur möglich, falls diesem Teilnehmer schon Bandbreite zur Verfügung steht.

Wie oben erwähnt, muss für eine Upload-Verbindung Bandbreite reserviert werden. Die Basisstation garantiert die Bandbreite aber nicht für die einzelne Verbindung, sondern stellt die Bandbreite dem Teilnehmer zur Verfügung zu welchem die Verbindung gehört. Darum braucht jeder Teilnehmer selber noch einen Uplink Scheduler, damit die zur Verfügung stehenenden Bandbreiten gerecht auf die einzelnen Verbindungen aufgeteilt werden können. Wie der genaue Scheduling Algorithmus aber funktioniert ist im Standard IEEE 802.16 nicht spezifiziert und ist den Herstellern überlassen. Dies hat Vor- und Nachteile. Ein Vorteil ist zum Beispiel, dass sich so die Hersteller diversifizieren können. Ein Nachteil könnte jedoch sein, dass die Kunden bei der Auswahl für ein Produkt überfordert sind, da sie auch abwägen müssen, welcher Scheduling Algorithmus am meisten Vorteile für sie hat [21].

Scheduling Services [10] Wie oben schon einmal erwähnt weisen Applikationen unterschiedliche Charakteristiken auf und haben somit verschiedene QoS-Anforderungen. Dafür werden im MAC - Layer vom Standard IEEE 802.16 vier verschiedene Scheduling Services angeboten. Diese sind unten kurz beschrieben.

- Unsolicited Grant Service (UGS)

  UGS unterstützt Real-Time Applikationen wie zum Beispiel Voice over IP, die Pakete mit fixer Datengrösse periodisch senden. Für Uplink-Garantien wird als Bandwidth-Request Mechanismus der Unsolicited Granting benutzt. Die garantierte Bandbreite wird dabei anhand der minimal reservierten Übertragungsrate berechnet. Die in diesem Service obligatorischen Parameter sind Maximum Sustained Traffic Rate, Maximum Latency, Tolerated Jitter und Request/Transmission Policy.

- Real-time Polling Service (rtPS)

  rtPS ist ebenfalls für Real-Time Applikationen konstruiert, unterstützt aber solche, die Datenpakete mit variabler Grösse periodisch versenden. Ein Beispiel dafür ist MPEG. Da der Bandbreitenanspruch nicht fix ist, muss die Basisstation merken, wieviel Bandbreite gebraucht wird. Darum versendet sie periodisch unicast Polls, die es dem Teilnehmer ermöglichen Bandbreitennachfragen zu senden. Die wichtigen QoS-Parameter bei diesem Service sind Maximum Sustained Traffic Rate, Minimum Reserved Traffic Rate, Maximum Latency und Request/Transmission Policy.

- Non-real-time Polling Service (nrtPS)

  nrtPS ist für Anwendungen gedacht, die keine speziellen Verzögerungsanforderungen haben. Ein Beispiel dafür ist FTP. Bei diesem Service gewährt die Basisstation Bandbreitenachfragen, indem sie den Teilnehmern per unicast polls oder broadcast polls die dafür benötigte Bandbreite zur Verfügung stellt. Die notwendigen QoS-Parameter sind Minimum Reserved Traffic Rate, Maximum Sustained Traffic Rate, Traffic Priority und Request/Transmission Policy.

- Best Effort (BE)

  BE bietet nur ein Minimum an QoS-Parametern an. Daten können nur versendet werden, falls es genügend Bandbreite hat. Die obligatorischen QoS-Parameter sind Maximum Sustained Traffic Rate, Traffic Priority und Request/Transmission Policy.

3.5 Sicherheit

Die drahtlose Kommunikation gewinnt auch noch heute immer mehr an Bedeutung. Es werden neue, verschiedene Anwendungen entwickelt, die nicht nur für belanglose Dinge verwendet werden, sondern immer mehr auch für Geschäftszwecke. Darum nimmt der Aspekt der Sicherheit eine wichtige Stellung ein. Kann WiMAX nicht genügend Sicherheit bieten, wird diese Technologie kaum Erfolg bei ernsthaften Anwendungen haben [24].

3.5.1 Sicherheitsarchitektur

Sicherheit ist im Standard IEEE 802.16 im MAC Protokoll im Privacy Sublayer vorgesehen. Diese Sicherheitsarchitektur besteht aus fünf Komponenten, die unten vorgestellt werden [24].

- Sicherheitsassoziationen

  Die Sicherheitsassoziationen enthalten Informationen zum Sicherheitsstatus, also zum Beispiel Schlüssel und ausgewählte Sicherheitsalgorithmen. Es gibt zwei verschiedene Arten von Sicherheitsassoziationen. Dies sind die Datensicherheitsassoziationen, welche im Standard expilizit definiert sind und die Authorisierungsassoziationen, die nicht explzit definiert sind. Die Datensicherheitsassoziation sichert die

Transportverbindung zwischen den Teilnehmern und den Basisstationen. Die Authorisierungsassoziation ist ein von der Basisstation und einem Teilnehmer geteilter Status. Die beiden Stationen halten den Authorisierungsschlüssel als Geheimnis. Die Basisstationen benötigen die Authorisierungsassoziation, um die Datensicherheitsassoziation beim Teilnehmer zu konfigurieren.

- X.509 Zertifikatsprofil

  X.509 ist ein Standard, der von ITU-T für Public-Key-Infrastrukturen entwickelt wurde [27]. Das X.509 Zertifikat identifiziert die teilnehmenden Kommunikationspartner. In WiMAX gibt es zwei verschiedene Zertifikate, nämlich Zertifikate, die den Hersteller eines Gerätes identifizieren sowie die Teilnehmerzertifikate. Der Hersteller kann sein eigenes Zertifikat entweder selber erstellen oder von einer Drittpartei ausstellen lassen. Das WiMAX-Forum hat VeriSign als Herausgeberin von Zertifikaten ausgewählt [25]. Indem die Basisstation das Herstellerzertifikat mit dem Teilnehmerzertifikat vergleicht, kann das Gerät eines Teilnehmer identifiziert werden.

- Privacy Key Management (PKM) Authorization

  Für die Authorisierung müssen drei Nachrichten zwischen dem Teilnehmer und der Basisstation ausgetauscht werden. Als erster Schritt sendet der Teilnehmer der Basisstation eine Authentifikationsnachricht, die das Zertifikat X.509 seines Herstellers enthält. So kann die Basisstation entscheiden, ob das Gerät vertrauenswürdig ist. Danach schickt der Teilnehmer eine Authorisierungsanfragenachricht, die das Zertifikat des Teilnehmers, die unterstützten Verschlüsselungsalgorithmen sowie der Verbindungs-ID des Teilnehmers beinhaltet. Wenn die Basistation den Teilnehmer identifizieren kann, generiert sie einen Authorisierungsschlüssel. Dieser wird in einer Nachricht, die auch die Lebensdauer des Authorisierungschlüssel sowie eine Liste von Beschreibungen der zugelassenen Sicherheitsassoziationen beinhaltet, an den Teilnehmer sendet [23].

- Datenschutz und Schlüsselverwaltung

  Das Privacy and Key Managment-Prokotoll (PKM) erstellt die Datensicherheitsassoziationen zwischen der Basisstation und dem Teilnehmer. Um so eine Assoziation aufbauen zu können, sendet der Teilnehmer der Basisstation eine ID der Assoziation, die aufgebaut werden soll und einen Keyed-Hash Message Authentication Code (HMAC), den es der Basisstation erlaubt Fälschungen zu entdecken. Falls der Code gültig ist und die gesendete ID zu einer Assoziation des Teilnehmers gehört, erstellt die Basisstation die Asssoziation.

- Verschlüsselung

  Der MAC-Layer ist verbindungsorientiert. Es gibt Managementverbindungen, die Verwaltungsnachrichten, wie zum Beispiel Bandbreitenachfragen transportieren. Weiter gibt es Secondary Managementverbindungen für den Austausch von Internetprotokoll Verwaltungsnachrichten und als dritte Verbindungsart gibt es nocht die Transportverbindungen. Verschlüsselt werden nur die Secondary Managementverbindungen und die Transportverbindungen. Zur Verschlüsselung wird AES oder 3DES verwendet.

3.5.2 Notwendige Schritte für Netzwerkeintritt

1. Als erstes scannt der Teilnehmer, der ein Netz betreten will, nach einem passenden Signal einer Basisstation.

2. Es wird einen Primary Managementkanal zur Basisstation erstellt, damit Nachrichten zur Authorisierung, Schlüsselverwaltung sowie Verhandlungen bezüglich sonstigen Parametern ausgetauscht werden können.

3. Die Basisstation authorisiert den Teilnehmer mittels des PKM-Authorisierungs-Protokolls.

4. Aufbau der Secondary Managementverbindung zum Versenden von Verwaltungsnachrichten.

5. Zum Schluss wird zwischen der Basisstation und dem Teilnehmer die Transportverbindung erstellt. [24]

3.5.3 Analyse der Sicherheit in 802.16

WiMAX bietet Sicherheit an, doch kann das Sicherheitssystem durchbrochen werden.

Die eigentlichen Daten abzuhören ist eher schwierig, da diese mit AES bzw. 3DES verschlüsselt sind. Doch die Verwaltungsnachrichten sind es nicht und diese können leicht abgehört oder gar modifiziert werden. Ein solcher Angriff könnte zu schwerwiegenden Problemen im Netz führen. Forscher konnten auch schon beweisen, dass es möglich ist zwei X.509-Zertifikate mit gleichem Hash-Wert zu erzeugen. So besteht für Betrüger die Möglichkeit ein echtes Zertifikat vorzugaukeln [28]. Würde dies jemandem gelingen, könnte natürlich einen verheerenden Schaden angerichtet werden. Ein Problem ist auch, dass es zwar Zertifikate für den Hersteller sowie für den Teilnehmer gibt, die Basisstation sich aber nicht identifizieren muss. Eine weitere Gefahr sind sind Denial of Service Attacken.

Die obigen Sicherheitsgefahren sind die grössten. 100-prozentiger Schutz kann nie gewährleistet werden, aber es muss sicher noch geforscht und Erfahrungen gesammelt werden, damit gegen bekannte Angriffe effiziente Vorkehrungen getroffen werden können.

3.6 Evaluation

In der Evaluation wird zuerst das Marktpotenzial anhand von drei Gebieten (USA, Asien und Europa) untersucht. Darauf folgt eine Auflistung der Stärken und Schwächen von WiMAX und zum Schluss wird noch auf die Konkurrenz und auf laufende Projekte, welche einen großen Einfluss auf die Zukunft von WiMAX haben werden, eingegangen.

3.6.1 Marktpotenzial

Wie WiMAX funktioniert wurde in den vorherigen Kapiteln erklärt. Nun stellt sich die Frage, wo WiMAX eingesetzt werden kann. Da Funkfrequenzen ein rares Gut sind, wird die Einführung von WiMAX in gewissen Gebieten langsamer vonstatten gehen, als man sich erhofft hat. Die Regulierung der Funkfrequenzen weicht zwischen den einzelnen Ländern ab und bereits vorhandene Mobilfunktechnologien hemmen den weltweiten Durchbruch von WiMAX. Es wird untersucht, wie sich die unterschiedliche Entwicklungen in den folgenden Regionen erklären. Diese Regionen wurden ausgewählt, weil sie die wichtigsten Märkte vorweisen, in welchen die Entwicklung von WiMAX vorangetrieben werden könnten.

- USA

  Ein großer Mobilfunk-Provider in den USA hat erst vor kurzem angekündigt, dass er gut drei Milliarden Dollar bis Ende 2008 für ein landesweites WiMAX Netz investieren möchte. Möglich wird diese Entwicklung, weil der Provider laut seinen eigenen Angaben genügend Frequenzen im 2.5GHz Band besitzt. Es sind aber noch andere Telekommunikationsunternehmen vorhanden, die wiederum auf die Entwicklung von UMTS/HSDPA setzen. Branchenkenner sind jedoch der Meinung, dass sich UMTS in Zukunft nicht durchsetzen wird, weil diese Technologie aus Europa komme und weltweit weniger verbreitet sei [41]. Es ist sicher sinnvoller in eine Technologie zu investieren und sie zu fördern, wenn sie eine schnellere Durchsatzrate und eine bessere Reichweite verspricht, aber insgesamt wird die weltweite Vernetzung der Wirtschaft durch die Inkompatibilität der Standards ausgebremst.

  Eigentlich könnte man sich erhoffen, dass die beteiligten Länder aus den Problemen mit den GSM-Netzen etwas gelernt haben. Durch die mangelhafte Kommunikation und Kooperation jedoch bleibt diese Hürde immer noch nicht als übersprungen und die Leute müssen sich entweder ein teures Universalgerät, das auf beiden Frequenzbänder arbeitet, kaufen oder sich sogar mit zwei verschiedenen Endgeräten herumschlagen. Die USA ist ein großer und wichtiger Markt und anscheinend wird sich der mobile WiMAX Standard 802.16e in diesem Lande durchsetzen.

- Asien

  In Asien wird sich die Verbreitung von WiMAX am schnellsten entwickeln, da sind sich einige Leute ziemlich sicher. Laut einer Studie eines Marktforschungsunternehmens wird der asiatische Raum ungefähr in drei Jahren fast die Hälfte aller WiMAX Benutzern beanspruchen [42]. Weshalb zeichnet sich diese rasante Entwicklung gerade in Asien ab und gibt es vielleicht auch noch andere Alternativen zu WiMAX in so einer Hitech-Kultur?

  Ein einfacher Grund, weshalb die asiatisch-pazifische Region uns immer eine Nasenlänge voraus ist, besteht darin, dass die Nachfrage nach neuen Technologien u.a. nach mobilen Breitbandzugängen viel höher ist, als bei uns in Europa. Seit Mitte 2006 wurde in Südkorea erstmals den Einsatz einer neuen Breitband-Funktechnologie WiBro (Wireless Broadband) kommerziell genutzt. Am Anfang war man überzeugt, dass es sich dabei um eine südkoreanischen Sonderlösung handeln würde, aber im

Verlaufe der Zeit hat sich herausgestellt, dass WiBro durchaus einen potentielle Alternative zum WiMAX Standard 802.16e werden könnte.

Am 4G-Forum 2005 demonstriert Samsung einen erfolgreichen Handover bei einer Geschwindigkeit von 60 km/h, mit der sich ein Gefährt von einer Zelle in eine andere bewegt. WiBro könnte sogar Zellübergaben mit Geschwindigkeiten von bis zu 120 km/h bewerkstelligen, bestätigt Samsung. Zusätzlich soll WiBro kompatibel zum mobilen WiMAX-Standard IEEE 802.16.e sein, was den koreanischen Standard für weltweite Netzbetreiber in Bezug auf Wirtschaftlichkeit und Schnelligkeit sehr interessant macht [44]. Es findet dort eine rasante Entwicklung statt und bei uns werden Technologien erst zum Einsatz kommen, die im asiatischen Raum bereits wieder veraltet sind. Diese Umstände machen es auch nicht einfacher einen weltweiten Standard festzulegen.

Als zusätzliche Information ist noch zu erwähnen, dass im Moment in Pakistan am weltweit größten WiMAX Netz gearbeitet wird. Sowohl der fixe WiMAX-Standard 802.16-2004 für die Überbrückung der letzten Meile, wie auch der mobile WiMAX-Standard 802.16e kommen dort zum Einsatz. Mehrere Millionen Teilnehmer sollen dann zuhause mit einem Breitband-Internet versorgt werden und zugleich mit dem mobilen WiMAX-Standard auch Zugang ins Internet mittels Hotspots erhalten [43]. WiMAX ist also nicht nur graue Theorie, sondern findet schon in der Praxis ihre Anwendung. In den industrialisierten Ländern schreitet die WiMAX-Entwicklung leider, vor Allem wegen den raren Frequenzbändern und den bereits eingekauften UMTS-Lizenzen, nur langsam voran. In der folgenden Region sind genau diese zwei Punkte schuld daran für die langsame Entwicklung von WiMAX.

- Europa

In Europa hat WiMAX einige Startschwierigkeiten, die auf mehrere Ursachen zurückzuführen sind. Im Unterschied z.B. zu den Asiaten sind die Europäer zurückhaltender, was die Begeisterung für neue Technologien anbelangt. Die Nachfrage für eine neue Breitbandbandtechnologie scheint bei uns nicht so stark zu sein, denn der etablierte DSL-Anschluss stillt unsere Bedürfnisse.

Die bereits erwähnten UMTS Lizenzen, die für viel Geld in vielen Teilen von Europa erworben wurden, hemmen den Einsatz von WiMAX und es ist klar, dass die Leute Angst haben, dass sich wieder so ein Szenario mit den Lizenzen abspielt. Da UMTS eigentlich überall eingesetzt werden kann, machen die Kunden jedoch von diesem Dienst kaum Gebrauch. Vielen Leuten ist das Angebot z.B. über das mobile Telefon fern zu sehen noch zu teuer oder es erscheint ihnen nicht als notwendig. Nichtsdestotrotz sind in Deutschland schon einige erfolgreiche Feldversuche mit WiMAX abgeschlossen worden und es gibt dort Gebiete, die mit einem kleineren WiMAX-Netzen abgedeckt sind. Auch in Europa scheint WiMAX für unerschlossene Gebiete oder Sonderlösungen geeignet zu sein, aber als Ersatz für bestehende Technologien, sehen die Erfolgschancen für WiMAX nicht gut aus.

3.6.2 Stärken und Schwächen

Weshalb soll nun WiMAX angewendet werden? Was sind die Stärken und Schwächen von WiMAX?

Durch die Einführung von WiMAX, werden einige Vorteile aufkommen wie zum Beispiel: Mittels WiMAX wird es keine Festlegung über das Gebiet geben, wo WiMAX eingesetzt werden soll [11]. Die Provider werden sowohl die WLAN-Hotspots, dies sind Orte, an denen der Provider gegen Entgelt das Netzwerk den Kunden zur Verfügung stellt, wie auch die Basisstationen des Mobilfunks mit der WiMAX Technologie verbinden können. Ein Kunde kann in diesen Cafes oder Restaurants einen Laptop mitnehmen und surfen. Der Kunde hat somit den Vorteil, dass er nicht an einen festen Internetanschluss gebunden ist. Der Zugang des Internetanschlusses wird mobil.

WiMAX könnte aber auch eine Lösung sein für die Verbindung von ländlichen oder schwer erreichbaren Zonen, die bis heutzutage keinen Netzanschluss hatten. Einen anderen Vorteil, den WiMAX mit sich bringt, ist, dass die Kunden und die Provider, die WiMAX benutzen nicht von einer einzigen Hardware abhängig sind. Sie können die Hardware kaufen, die ihnen zusagt. Der Wettbewerb zwischen den Produkte Hersteller wird somit angekurbelt und es ist zu erwarten, dass die Preise gesenkt werden.

Schliesslich wurde für die Zertifizierung von WiMAX Produkten speziell das WiMAX Forum [29] gegründet, welches heutzutage mehr als 200 Firmen beinhaltet, die WiMAX zur Nummer 1 der Mobilen Systeme machen wollen.

Neben den zahlreichen Stärken von WiMAX gibt es auch Aspekte, die gegen eine Benutzung von WiMAX sprechen. Ein Problem, mit welchen sich WiMAX auseinandersetzen muss, ist die Kompatibilität zwischen den Varianten von WiMAX. Obwohl die Spezifikation der WiMAX Arten schon vorhanden ist, muss noch erforscht werden, ob WiMAX fixed und WiMAX mobile kompatibel sind. Wird beispielsweise ein Notebook gebraucht, welches die Technologie des Standards IEEE 802.16e aufweist, ist noch nicht klar, ob ein Mail des Notebooks mittels einer Basisstation des Types WiMAX fixed übertragen werden kann. Obwohl sowohl das Notebook wie auch die Basisstation WiMAX anwenden, gebrauchen sie durch verschiedene Arten von WiMAX eine andere Technologie. Viele Provider wollen heutzutage noch nicht die existierenden Versionen von WiMAX fixed installieren, da sie auf die Version IEEE 802.16e, WiMAX mobile warten [9]. Würden sie WiMAX heutzutage installieren und würden dann in einigen Jahren auf WiMAX mobile umsteigen wollen, müssten sie vieles neu installieren. Dies wäre schlicht ein zu grosser Aufwand.

Ein weitere Schwäche von WiMAX sind die lizenzierten Frequenzen. Die Frequenzen des WiMAX Standards werden versteigert. Schliesslich ist es von Nachteil, dass der Erfolg von WiMAX von den Strategien von den Providern und der Frequenzpolitik von den Regulierungsbehörden abhängen und nicht von der Technik an sich.

3.6.3 Laufende Projekte

Projekte, die heute lanciert werden, haben womöglich grossen Einfluss auf die Zukunft. In diesem Abschnitt möchten wir ein paar aktuelle Projekte um WiMAX vorstellen, die unserer Meinung nach eine wichtige Bedeutung für den Erfolg von WiMAX haben könnten.

- Intel

  Intel hat sich zum Ziel gesetzt Notebooks mit einem Chip zu versehen, die sie WiMAX tauglich macht. Im Oktober 2006 hat Intel einen Chip, INTEL WiMAX Connection 2250, mit dem Codenamen Rosedale 2 angekünkdigt. Dieser soll im Dualmodus laufen. Das heisst, er wird kompatibel zu IEEE 802.16d-2004 und 802.16e-2005 sein und somit die fixe sowie die mobile Variante von WiMAX unterstüzten. Damit WiMAX Erfolg haben kann, müssen die Bestandteile kostengünstig beziehbar sein. Darum wird dieser Chip speziell für kostengünstige Modems optimiert sein. Dieser Chip kann nur erfolgreich werden, wenn es auch andere Unternehmen gibt, die an ihm interessiert sind. Grosse Unternehmen wie Motorola, Siemens oder Alcatel haben bereits angekündigt diesen neuen Chip in ihre Produkte zu integrieren [30].

- Motorola

  Motorola plant den Chip von Intel in ihre Produkte zu integrieren. Diese sollen bereits ab 2007 auf den Markt kommen. Motorola will ebenfalls eigene Chips herstellen. Diese sind für Mobilfunkgeräte vorgesehen und sollen Sprach-, Video- und Daten-Kommunikation unterstützen. Im Gegensatz zum Chip von Intel soll dieser Chip nur die mobile Variante von WiMAX unterstützen. Auf einer Webseite von Motorola findet man die untenstehende Abbildung (3.11). Wahrscheinlich haben sie den Entscheid, nur das mobile WiMAX zu unterstützen, auf diese Grafik gestützt. Dort sieht man, dass in Zufkunft vor allem die Variante IEEE 802.16e Erfolg haben wird.

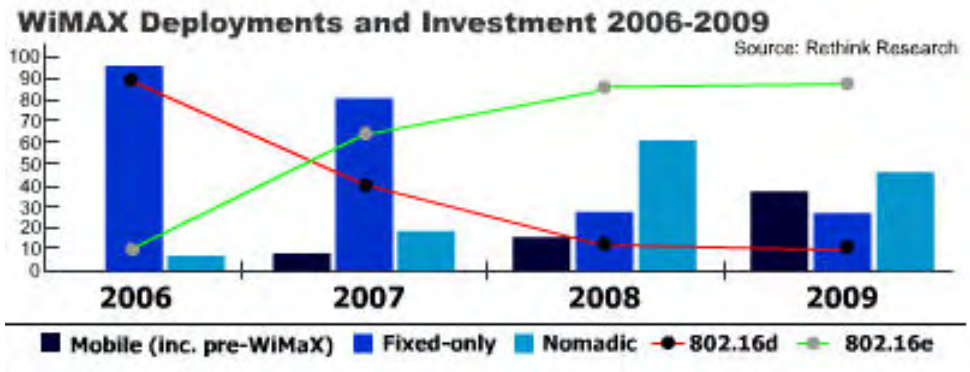


Abbildung 3.11: Development WiMAX [26]

Es ist beabsichtigt, diese eigene Chips ab dem Jahr 2008 auf den Markt zu bringen. Motorola möchte die führende Rolle für das Engagement von WiMAX sein. Darum investieren sie sehr viel in WiMAX. Weitere Projekte betreffend WiMAX laufen

in Pakistan, wo ein landesweites WiMAX-Netz aufgebaut werden soll. Weiter ist Motoroloa auch im Projekt von Sprint Infrastruktur involviert, die ein landesweites Netz in der USA aufbauen wollen. In diesem Projekt liefern sie die Endgeräte [31].

- Nokia

  Auch Nokia sieht grosses Potenzial in WiMAX und hat darum auf Ende 2007 eine Basisstation für WiMAX angekündigt. Wie Intel, schaut auch Nokia, dass das Produkt möglichst günstig zu erwerben sein wird und ohne grossen Aufwand in Betrieb genommen werden kann. Ein Jahr später soll dann ein von Nokia entwickeltes WiMAX-fähiges mobiles Endgerät auf den Markt kommen [32].

- Ghana

  In Ghana ist die DSL Abdeckung noch gering und die Distanzen zwischen Dienstanbietern und Kunden sind gross. Doch ist die Nachfrage nach schnellen Internetzugängen gross. In so einem Land kommt eine Technologie wie WiMAX, die grosse Flächen abdecken kann und dazu noch schnelle Verbindungen verspricht, wie gerufen. Ghana soll nun das erste Land werden, in welchem ein landesweites WiMAX-Netz aufgebaut wird. Man hat sich entschieden erst einmal die mobile Variante von WiMAX (IEEE 802.16e) in der Hauptstadt zu installieren [33].

3.6.4 Konkurrenz

Werden diese Projekte ihre Ziele erreichen, sieht die Zufkunft von WiMAX schon mal vielversprechend aus. Doch die Konkurrenz schläft nicht und wird es WiMAX sicher nicht einfach machen. Darum wollen wir hier jetzt auch mögliche Konkurrenten erwähnen.

- IEEE 802.20 (Mobile Broadband Wireless Access)

  Mobile Broadband Wirless Access (IEEE 802.20 [34]) wird wie der Standard 802.16 von der Organisation amerikanischen Instituts der Elektro- und Elektronikingenieure (IEEE) entwickelt und soll mobile Breitbandzugänge per Funk ermöglichen. WiMAX war zuerst für fixe Anwendungen gedacht, doch mit der Variante 802.16e gehen die beiden Standards in die gleiche Richtung. WiMAX hat jedoch einen Vorsprung. Von 802.20 wurde erst dieses Jahr einen Entwurf anerkannt. Im Gegensatz zu 802.20 wird WiMAX stark von der Industrie unterstützt, was für einen Erfolg einer neuen Technologie äusserst wichtig ist. Ausserdem gab es in der Abreitsgruppe 802.20 diesen Sommer interne Streitigkeiten. Diese führten soweit, dass der Vorstand abgesetzt werden musste, was für die Technologie sicher nicht förderlich war [36].

- DSL

  Mit DSL hat WiMAX einen starken Konkurrenten. DSL ist in den weitentwickelten Industriestaaten weit verbreitet. Dessen Bandbreiten wurden in letzter Zeit immer wieder erhöht. Deshalb werden die jetztigen DSL-Kunden, die das Internet vor allem zu Hause oder im Büro verwenden, eine Technologie wie WiMAX nicht vermissen. Darum wird WiMAX auch eher Erfolg in Drittweltländern haben, wo DSL sich noch nicht etabliert hat oder an abgelegenen Orten, wo sich eine Verkabelung nicht lohnt.

Tabelle 3.3: Vergleich 802.20 mit 802.16e [35]

	802.16e	802.20
max Geschwindigkeit	bis zu 120 km/h	bis zu 250km/h
Spektrum	lizenzierte Bänder: 2.0 - 6.0GHz	lizenzierte Bänder <3.5GHz
Mobility Features	Local/Regional Mobility, Handoff und Roaming Support	Global Mobility, Handoff und Roaming Support
Spitzendatenrate DL	70Mbps mit 14MHz	16Mbps mit 5MHZ

- UMTS

  UMTS ist ein Mobilfunkstandard der dritten Generation und könnte den Erfolg von WiMAX einschränken. Telekommunikationsanbieter haben für die Konzessionen von UMTS viel Geld ausgegeben, doch der Erfolg von UMTS ist bis jetzt ausgeblieben. Gut möglich, dass dies Investoren hemmt, schon wieder in eine neue Technologie zu investieren, deren Erfolg ungewiss ist.

- WiBRO (Wireless Broadband)

  WiBRO ist ebenfalls eine Funktechnologie, die mobilen Zugang zum Internet ermöglicht. WiBRO wird von der koreanischen Telekommunikationsindustrie entwickelt und wurde 2004 standardisiert. Mit dieser Technologie sind Datenübertragungsraten von 30 bis maximal 50 MBit/s im Umkreis von 1-5km um die Basisstation möglich [38].

3.7 Zusammenfassung

WiMAX besitzt ein grosses Potenzial, weltweit ein akzeptierter Mobilfunkstandard zu werden und öffnet neue Toren für Privatanwender und Unternehmen. U.a. kann die die letzte Meile drahtlos überbrückt werden und in nicht industrialisierten Ländern werden Millionen von Menschen mit relativ kostengünstigen Hispeed-Internetzugängen versorgt. Grosse Firmen wie z.B. Motorola und Intel haben das Potenzial von WiMAX erkannt und treiben die Entwicklung voran. Ein WiMAX-Chipsatz (WiMAX Connection 2250), der für stationäre und mobile WiMAX Anwendungen geeignet ist, wurde bereits produziert. Nach Motorola hat auch Nokia WiMAX-kompatible Endgeräte angekündigt, die aber erst 2008 erscheinen werden. Auch wenn schon bereits Alternativen zu WiMAX auf dem Markt sind, denken wir, dass der Standard 802.16d und 802.16e erfolgreich sein wird. Dass es für WiMAX nicht für einen weltweiten Standard reicht, ist klar. Die fehlende Kommunikation und Kooperation der beteiligten Ländern hat diese Entwicklung verhindert, trotzdem werden wir noch viel von WiMAX zu hören bekommen.

Literaturverzeichnis

[1] WiMAX-Forum, http://de.wikipedia.org/wiki/Special:Search?search=wimax+forum&go=Go, 12.01.2007

[2] Datenübertragungsrate, http://de.wikipedia.org/wiki/DatenC3BCbertragungsrate, 11.11.2006

[3] WiMAX fixed, http://library.thinkquest.org/04oct/01721/wireless/wimax.htm, 12.11.2006

[4] Grundlagen WiMAX - Anwendung, Architektur und Aufbau IEEE802.16 und seine Bestandteile, http://www.computerwoche.de/knowledge_center/wireless/570194/index2.html, 22.10.2006

[5] Fieberkurve, http://www.heise.de/mobil/artikel/62460/0, 23.10.2006

[6] Adaptive Modulation, http://www.lnt.e-technik.tu-muenchen.de/mitarbeiter/oikonomidis/Seminar/SeminarReports/WiMAX.pdf, 20.10.2006

[7] Mobilität, http://developer.intel.com/technology/itj/2004/volume08issue03/art01_g/, 23.11.2006

[8] Heine, Gunnar: WiMAX from A-Z, INACON, 2005

[9] S. J. Vaughan-Nichols, Achieving wireless broadband with WiMax, Computer, Volume 37, Issue 6, June 2004

[10] IEEE 802.16-2004. IEEE standard for Local and Metropolitan Area Networks – Part 16: Air Interface for Fixed Broadband Wireless Access Systems, October 2004

[11] Sweeney Daniel: LinkWiMax operator's manual : building 802.16 wireless networks, LinkBerkeley, Calif. : Apress, 2004

[12] PHY Layer, www.tm.uka.de/itm/uploads/folien/115/MK-04-DrahtloseMANs-4up.pdf, 05.10.2006

[13] Jochen Schiller: Mobilkommunikation 2., überarbeitete Auflage, Addison-Wesley, 2003

[14] MAC Layer, http://www.computerpartner.de/sonstiges/640856/index.html, 06.10.2006

[15] OFDM, http://www.wi-fiplanet.com/tutorials/article.php/10724_3557416_1,01.10.2006

[16] WLAN IEEE 802.11g/n, http://wwwspies.in.tum.de/MVS/sem06/contents/WLAN_Ausarbeitung.pdf, 01.10.2006

[17] BPSK, http://en.wikipedia.org/wiki/Image:BPSK_Gray_Coded.svg, 01.10.2006

[18] QPSK, http://en.wikipedia.org/wiki/Image:QPSK_Gray_Coded.svg, 01.10.2006

[19] ARQ, http://de.wikipedia.org/wiki/ARQ-Protokoll, 02.11.2006

[20] APC, http://www.elecdesign.com/Articles/Index.cfm?AD=1&ArticleID=9096&pg=2, 06.11.2006

[21] Claudio Cicconetti, Luciano Lenzini, and Enzo Mingozzi, University of Pisa Carl Eklund, Nokia Research Center, Quality of Service Support in IEEE 802.16 Networks, IEEE Network, March/April 2006

[22] Carl Eklund, Nokia Research Center Roger B. Marks, National Institute of Standards and Technology Kenneth L. Stanwood and Stanley Wang, Ensemble Communications Inc., IEEE Standard 802.16: A Technical Overview of the WirelessMAN Air Interface for Broadband Wireless Access, IEEE Communications Magazine, June 2002

[23] Sen Xu, Manton Matthews, Chin-Tser Huang, Security Issues in Privacy and Key Management Protocols of IEEE 802.16

[24] David Johnston and Jesse Walker, Overview of IEEE 802.16 Security, http://ieeexplore.ieee.org/Xplore/login.jsp?url=/iel5/8013/29015/01306971.pdf, 26.10.2006

[25] VeriSign, http://www.itseccity.de/?url=/content/markt/invests/060720_mar_inv_verisign.html, 06.01.2007

[26] Entwicklung WiMAX, http://www.connectwithcanopy.com/, 26.10.2006

[27] Wikipedia X.509, http://en.wikipedia.org/wiki/X.509, 01.11.2006

[28] Forscher erzeugen unterschiedliche X.509-Zertifikate mit gleichem MD5-Hash, http://www.heise.de/newsticker/meldung/mail/57038, 01.12.2006

[29] WiMAX Forum Overview, http://www.wimaxforum.org, 20.10.2006

[30] Intel, http://www.intel.com/netcomms/technologies/wimax, 20.10.2006

[31] Motorola entwickelt Chips für mobiles WiMAX, http://www.wireless-zuerich.ch/modules.php?op=modload&name=News&file=article&sid=858&mode=thread&order=0&thold=0, 01.11.2006

[32] Erstes mobiles WiMAX-Gerät von Nokia kommt 2008, http://www.wireless-zuerich.ch/modules.php?op=modload&name=News&file=article&sid=875&mode=thread&order=0&thold=0, 01.11.2006

[33] Ghana bekommt erstes landesweites WiMAX-Netz, http://www.wireless-zuerich.ch/modules.php?op=modload&name=News&file=article&sid=862, 01.11.2006

[34] IEEE 802.20, http://grouper.ieee.org/groups/802/20/, 06.01.2007

[35] BakkArbeit, http://stud3.tuwien.ac.at/~e0225546/wimax/BakkArbeit_WiMAX.pdf, 01.11.2006

[36] Vorstand der Arbeitsgruppe 802.20 abgesetzt, http://www.heise.de/newsticker/meldung/78502, 01.11.2006

[37] Tabelle Funkschnittstellen, http://www.webopedia.com/DidYouKnow/Computer_Science/2005/OFDMA.asp, 20.10.2006.

[38] Wikipedia WiBro, http://de.wikipedia.org/wiki/Wibro, 01.11.2006

[39] WiMAX - die neue letzte Meile, http://www.wimax.ch, 20.10.2006

[40] ATM, http://de.wikipedia.org/wiki/Asynchronous_Transfer_Mode, 03.01.2007

[41] WiMAX in den USA, http://www.dsltarife.net/news/1691.html, 25.10.2006

[42] WiMAX in Asien, http://www.teltarif.ch/arch/2005/kw40/s4534.html, 27.10.2006

[43] Anwendung von WiMAX in Pakistan, http://www.dsltarife.net/news/1344.htm, 27.10.2006

[44] Handover mit WiBro, http://neasia.nikkeibp.com/dailynewsdetail/002104, 11.11.2006

[45] Triple DES, http://de.wikipedia.org/wiki/Advanced_Encription_Standard, 03.01.2007

[46] Multimedia preformance of IEEE 802.16 MAC, http://de.wikipedia.org/wiki/Advanced_Encription_Standard, 03.01.2007

Kapitel 4

Wireless Sensor Networks

Daniel Eisenring, Nora Kleisli, Tobias Wolf

Diese Seminararbeit beschäftigt sich mit dem noch jungen Gebiet der Wireless Sensor Networks. Zuerst wird das Gebiet eingeführt, angefangen mit der traditionellen Definition, einer Charakterisierung verschiedener Attribute welche die grosse Spannweite der heutigen Möglichkeiten aufzeigen soll, und anschliessend verschiedene Anwendungsmöglichkeiten kurz skizziert. Der erste Teil wird abgeschlossen mit einem Vergleich verwandter Technologien. In einem zweiten Teil werden technische Herausforderungen sowohl den Aufbau von Sensorknoten als auch mögliche Kommunikationsmethoden betreffend diskutiert. In einem dritten Teil wird zunächst der aktuelle Stand der Technik dargestellt, bezogen auf Hardware und auch Software, und anschliessend Beispiele konkreter Anwendungen im Detail vorgestellt. Die Arbeit wird abgeschlossen durch einen Vergleich der ursprünglichen Erwartungen mit der tatsächlichen, heutigen Situation.

Inhaltsverzeichnis

4.1 Einleitung

In diesem Kapitel werden wir zur Einführung in das Gebiet zunächst die traditionelle Definition eines Wireless Sensor Networks geben. Anschliessend soll die breite Spanne der heutigen Möglichkeiten anhand einer Reihe von Attributen charakterisiert werden. Abgeschlossen wird das Kapitel durch eine kurze Übersicht möglicher Einsatzszenarien für Wireless Sensor Networks.

4.1.1 Definition

Die ersten Forschungsprojekte Ende der 90er Jahre haben zu folgender de facto Definition geführt: Sie verstanden unter einem Wireless Sensor Network ein grossflächiges, nicht partitioniertes, drahtloses ad-hoc Netzwerk bestehend aus vielen kleinen, vorwiegend homogenen, mehrheitlich unbeweglichen Sensorknoten mit beschränkten Resourcen. In Bezug auf die Kommunikation wurde von multi-hop Routing ausgegangen, bei welchem Nachrichten von Knoten zu Knoten propagiert werden bis sie an ihr gewünschtes Ziel gelangen. Ferner wurde davon ausgegangen, dass eine Vielzahl dieser Knoten zufällig über ein Einsatzgebiet verstreut werden würden.

Diese Definition ergab sich aus den damals primär ins Auge gefassten Anwendungen, welche bedingt durch die Quelle der Forschungsgelder in erster Linie militärischer Natur waren. In den Folgejahren hat sich das mögliche Anwendungsgebiet dieser Technologie jedoch rasant verbreitert, und mittlerweile deckt die oben genannte Definition nur noch einen - wenn auch wichtigen - Teil der Gesamtmöglichkeiten ab und ist entsprechend nicht mehr zeitgemäss.

Das Gebiet der Wireless Sensor Networks ist heute so breit gestreut, dass eine umfassende Definition über die Kernelemente (ein drathloses Netzwerk bestehend aus mit Sensoren bestückten Knoten) hinaus praktisch unmöglich wird. Aus diesem Grund möchten wir auf solch einen Versuch verzichten, und die breite Spanne des Gebietes anhand einer Reihe von Attributen charakterisieren.

4.1.2 Charakteristika

Wie im letzten Abschnitt motiviert, sollen im Folgenden einige wichtige Attribute von Wireless Sensor Networks (kurz auch *WSNs* genannt) näher charakterisiert werden. Dies geschieht in Anlehnung an [25] und soll das weite Feld von Möglichkeiten aufzeigen welche das Gebiet heute bietet.

Aufstellung

Die Aufstellung von Sensorknoten eines WSN kann aus zwei Sichten betrachtet werden. Als erstes kann die *Art* der Aufstellung unterschieden werden. Auf der einen Seite gibt es

Anwendungen, in denen alle Knoten manuell *gezielt* an genau vorgegebenen Orten platziert werden. Auf der anderen Seite können Sensorknoten auch *zufällig* platziert werden, in dem sie beispielsweise von einem Flug- oder Fahrzeug abgeworfen werden. Die zweite Variante ist besonders dann interessant, wenn das WSN aus vielen kleinen Knoten besteht und ein grösseres Gebiet möglichst flächendeckend überwacht werden soll, die exakte Positionierung der Knoten jedoch keine Rolle spielt.

Als zweites lässt sich die *zeitliche Dimension* der Aufstellung unterscheiden. Der Prozess kann je nach Anwendung *einmalig* sein, oder *schrittweise* erfolgen. Während in vielen Anwendungen der erste Fall die Regel ist, gibt es durchaus Gründe die für den zweiten Fall sprechen können. So ist es denkbar, dass ausgefallene Knoten kontinuierlich ersetzt werden sollen, oder ein bestehendes WSN nach und nach um weitere Knoten erweitert wird um die Dichte oder Grösse des Netwerks zu erhöhen.

Mobilität

Ein nächstes Unterscheidungsmerkmal ist die Mobilät von Sensorknoten nach Aufstellung des WSN. Dabei können drei Dimensionen betrachtet werden. Die erste ist die *Art* der Mobilität. Knoten können sich enweder *passiv* bewegen, in dem sie beispielsweise direkt an mobilen Trägern angebracht werden, oder durch externe Kräfte wie Wind oder Wasser bewegt werden. Im Kontrast dazu sind auch Anwendungen denkbar in denen sich die Knoten *aktiv* selbst fortbewegen können.

Nach der Art der Mobilität lässt sich als zweites der *Grad* der Mobilität unterscheiden. Das gesamte Netzwerk kann vollständig *unbeweglich* sein (kein Knoten ist mobil), oder es kann *teilweise* (einige Knoten) oder sogar *vollständig* (alle Knoten) mobil sein.

Als drittes lässt sich auch bei der Mobilität eine *zeitliche Dimension* betrachten. Das Netzwerk kann entweder *gelegentlich* mobil sein (z.B. periodische Überwachung verschiedener Punkte einer Route) oder *regelmässig* in Bewegung sein (z.B. wenn Sensorknoten an lebenden Tieren angebracht werden).

Knotengrösse

Je nach Anforderungen der Anwendung kann die Grösse von Sensorknoten beträchtlich variieren. Angefangen von der Grössenordnung eines Schuhkartons reicht die Spannweite bis hinunter zu der eines Reiskorns, und in der Forschung werden selbst mikroskopisch kleine Partikel in Betracht gezogen. Die meisten Projekte die über blosse Forschungsprototypen hinaus gereift sind verwenden derzeit Sensorknoten in der Grössenordnung einiger Kubikzentimeter und aufwärts. Das selbe gilt auch für kommerziell erhältliche Produkte wie zum Beispiel [7].

Kosten

Ähnlich vielfältig wie die Grösse der Knoten eines WSN fallen auch die Kosten pro Knoten aus. Auf der einen Seite des Spektrums liegen Netzwerke mit wenigen mächtigen (viele

verschiedene Sensoren, besonders leistungsstarke Kommunikationsvorrichtungen, grosse Energiereserven) Knoten, die drei- oder sogar vierstellige Beträge kosten können. Auf der anderen Seite liegen Netzwerke bestehend aus einer Vielzahl kleiner, einfacher Knoten mit Stückkosten jenseits des Dezimalpunktes.

Resourcen

Schranken in Bezug auf die Grösse und Kosten der Sensorknoten sowie deren geplante Lebensdauer beinflussen direkt die Resourcen welche einem Sensorknoten zur Verfügung stehen. Im allgemeinen sind Sensorknoten in Bezug auf Speicher, Prozessorleistung sowie sonstige Resourcen stark eingeschränkt, und liegen meistens um Grössenordnungen unterhalb von dem was einem heutigen Mobiltelefon oder PDA zur Verfügung steht.

Energie

Ein zentrales Attribut von Sensorknoten ist die ihnen zur Verfügung stehende Energie. Da Knoten eines WSN üblicherweise keine externe Stromversorgung besitzen, bleiben zwei Möglichkeiten für die Stromversorgung. Die erste und in den meisten Fällen angewandte Methode ist *gespeicherte* Energie in Form von Batterien zu verwenden. Grösse, Kosten, Resourcenausstattung und geplante Lebensdauer der Knoten beeinflussen dabei die zur Verfügung gestellte, gespeicherte Energie.

Die zweite Möglichkeit ist, Energie in den Sensorknoten selbst zu *erzeugen*. Hierbei wird die Umgebung zur Energiegewinnung genutzt, und beispielsweise das Sonnenlicht über Solarzellen, oder mechanische Energie in Form von Vibrationen in elektrische Energie umgewandelt. Es sind natürlich auch Mischformen von gespeicherter und erzeugter Energie möglich.

Heterogenität

In der ursrünglichen Definition von WSNs wurde davon ausgegangen, dass das Netzwerk *homogen* ist, also nur aus gleichartigen Knoten besteht. Dies ist in den heutigen Anwendungen mittlerweile nur noch selten der Fall, meistens ist das Netzwerk nun *heterogen* und besteht aus Knoten mit verschiedener Ausstattung und Funktionalität (und sei es nur der Fall dass es einen „Master“ Knoten gibt).

Kommunikationsform

Eine fundamentale Eigenschaft von WSNs ist die gewählte Kommunikationsform. Hierbei gibt es eine Reihe von Möglichkeiten. Am meisten Verwendung findet hierbei die Kommunikation über *Funk*. Diese Art der Kommunikation hat den Vorteil, dass keine Sichtverbindung benötigt wird, und dass mittlere Kommunikationsdistanzen mittlerweile

mit relativ geringem Energieaufwand und kleinen Antennen realisiert werden können. Die Mindestgrösse der Antennen begrenzt allerdings auch den Grad der Miniaturisierung.

Vorrichtungen für die optische Kommunikation über *Licht* können platzsparender realisiert werden. Hierbei kann die Kommunikation entweder *aktiv* über Laserdioden erfolgen, oder *passiv* über Reflektion von Licht einer externen Quelle (ein Ansatz den z.B. das „Smart Dust“ Projekt [20], [34] verfolgt). Nachteile optischer Kommunikation sind die benötigte Sichtverbindung, und die Anfälligkeit auf Störungen durch andere Lichtquellen.

Weder Funk noch Licht eignen sich besonders für die Kommunikation unter Wasser, für WSNs die in solch einer Umgebung eingesetzt werden sollen bietet sich als weitere Variante die Kommunikation über *Schall* an. Aber auch in trockener Umgebung lässt sich diese Form der Kommunikation nutzbringend einsetzen, wenn es z.B. um die Bestimmung von Distanzen geht.

Netzwerkinfrastruktur

Die Netwerkinfrastruktur in einem WSN lässt sich analog zur Situation in einem Wireless LAN charakterisieren. Die erste Möglichkeit ist der *Infrastrukur* Modus, hierbei läuft die gesamte Kommunikation über eine oder mehrere „Basisstationen“, sowohl zwischen den Knoten selbst als auch zu potentiellen externen Zielen. Da der Aufbau einer Infrastruktur häufig mit höheren Kosten und grösserem Aufwand verbunden ist, wird diese Variante häufig nicht favorisiert, es sei denn eine bestehende Infrastruktur kann mitbenutzt werden.

Im Kontrast zum Infrastruktur Modus steht der *ad-hoc* Modus, in welchem alle Knoten gleichwertig sind und direkt miteinander kommunizieren. Dies kann entweder durch Punkt zu Punkt Kommunikation zwischen zwei beliebigen Knoten geschehen, oder durch multihop Routing bei welchem Nachrichten von einem Knoten zum nächsten weiter gereicht werden bis sie das gewünschte Ziel erreichen. Diese Form ist die grundsätzlich präferierte Variante in WSNs.

Schlussendlich finden sich aus Notwendigkeit in vielen Anwendungen *hybride* Mischformen, in welchen die Knoten beispielsweise ad-hoc untereinander kommunizieren, Messergebnisse aber von einer Basisstation weiter verarbeitet und über eine bestehende Infrastruktur zu einem externen Ziel geleitet werden.

Netzwerktopologie

Ein weiteres wichtiges Attribut von WSNs ist die logische Topologie welche die Sensorknoten bilden. In der einfachstens Form findet die Kommunikation single-hop statt, jeder Knoten kommuniziert direkt mit jedem anderen. In diesem Fall liegt eine vollständig *vermaschte* Netzwerktopologie vor.

Kommunizieren die Knoten nicht direkt untereinander sondern nur über eine oder mehrere Basisstationen, findet die Kommunikation über jeweils maximal zwei Hops statt und die Topologie entspricht einem *Stern*.

Wie im letzten Abschnitt schon beschrieben ist die bevorzugte Kommunikationsform in vielen WSNs ad-hoc mit multi-hop Routing. In solch einem Fall können die Knoten eine beliebige *Graphstruktur* bilden, häufig wird jedoch eine vereinfachte Form wie ein *Baum* oder eine *Menge verbundener Sterne* gewählt.

Konnektivität

Abhängig von der konkreten Anwendung und anderen Faktoren wie Kommunikationsreichweite und physische Position der Sensorknoten, lassen sich verschiedene Formen der Netzwerkkonnektvität in WSNs unterscheiden. Wenn zu jedem Zeitpunkt eine Verbindung zwischen zwei beliebigen Knoten möglich ist, wird das Netzwerk als verbunden und die Konnektivität als *permament* bezeichnet.

Wenn das Netzwerk zeitweise partitioniert ist, ist die Konnektivität *periodisch*.

Schlussendlich kann es auch sein, dass die Knoten die meiste Zeit isoliert voneinander sind, die Konnektivität *unregelmässig* ist. Dies kann beispielsweise der Fall sein wenn Sensorknoten auf mobilen Entitäten angebracht sind, die nur ab und zu in der Reichweite anderer Knoten sind.

Abdeckung

Auch die Abdeckung des Einsatzgebietes durch die Sensorknoten ist ein Unterscheidungsmerkmal verschiedener WSN Anwendungen. Das Spektrum fängt hier an mit *spärlicher* Abdeckung, bei der die Knoten ziemlich verstreut sind nur eine Untermenge des Einsatzgebietes überwachen können. Da für viele Anwendungen keine vollständige Überwachnung notwendig ist, reicht dies in den meisten Fällen bereits aus.

Wenn die Sensorknoten das Einsatzgebiet vollständig oder nahezu vollständig überwachen können, lässt sich von einer *dichten* Abdeckung sprechen.

Das Ende des Spektrums bildet schliesslich eine *redundante* Abdeckung. Hierbei werden verschiedene Teilgebiete von mehreren Sensorknoten gleichzeitig abgedeckt. Dies ist insbesondere dann interessant, wenn eine lückenlose Überwachung des Einsatzgebietes auch beim Ausfall einiger Sensorknoten noch gewährleistet sein soll.

Netzwerkgrösse

Die Netzwerkgrösse (Anzahl Knoten im Netzwerk) wird prinzipiell durch den gewünschten Grad der Abdeckung, die Grösse des Einsatzgebietes sowie der maximalen Kommunikationsreichweite der Knoten bestimmt. So kann die Netwerkgrösse je nach Anwendung von einer Hand voll Sensorknoten bis hin zu mehreren Tausend variieren.

Lebensdauer

Die Lebensdauer der Sensorknoten hängt in erster Linie vom Energieverbrauch sowie der vorhandenen Energiemechanismen eines Knotens ab, andererseits aber auch von der Robustheit der Knoten und der Einsatzumgebung. Je nach Anwendung ist eine Lebensdauer von wenigen Stunden bis hin zu mehreren Jahren denkbar.

4.1.3 Anwendungsgebiete

Im folgenden sollen einige mögliche Anwendungsgebiete für Wireless Sensor Networks kurz illustriert werden. Diese Liste ist natürlich bei weitem nicht abschliessend. Grundsätzlich lassen sich die Anwendungen in drei Kategorien unterteilen, *räumliche* Beobachtungen, Beobachtungen von *Dingen* und schliesslich die Beobachtung der *Interaktion* zwischen Dingen und/oder ihrer Umgebung [4].

Umweltbeobachtung WSNs können für eine Vielzahl an Umweltbeobachtungen eingesetzt werden. Angefangen von allgemeinen Klimabeobachtungen, detaillierten Beobachtungen von Gletschern, Ozean- oder sonstigen Phänomenen bis hin zu Frühwarnsystemen für Erdbeben oder Tsunamis gibt es in diesem Anwendungsgebiet eine breite Spanne von Möglichkeiten.

Lebensraumbeobachtung Auch für die Beobachtung der Lebensräume von wildlebenden Tieren bietet sich die Technologie der WSNs an. So wurden beispielsweise schon Projekte durchgeführt die das Nistverhalten von Vögeln aus der Nähe erforschen sollten, oder die Bewegung von Wildtieren in grossflächigen Lebensräumen verfolgen.

Prozessüberwachung Ein weiteres mögliches Einsatzgebiet ist die Überwachung von Fertigungsprozessen. Sensorknoten können dabei z.B. die Vibrationsmuster oder Laufgeräusche von Maschinen registrieren, und bei Abweichungen von Standardmustern (welche auf Verschleiss oder sonstige Fehlfunktionen hindeuten können) Techniker alarmieren welche die Maschinen dann näher untersuchen.

Medizinische Überwachung WSNs lassen sich auch für die medizinische Überwachung einsetzen. Denkbar sind hier einerseits Anwendungen welche Patienten in ihrem täglichen Leben überwachen, als auch andererseits Systeme die Patienten innerhalb eines Krankenhauses kontinuierlich beobachten.

Intelligente Räume Der Einsatz in Gebäuden eröffnet auch mehrere Möglichkeiten. Angefangen bei der Einbruchserkennung oder Messung von Stromverbrauch verschiedenster Komponenten bis hin zur automatischen Steuerung von Temparatur, Licht oder anderen Umgebungsvariablen lassen sich hier etliche Anwendungen realisieren.

Militärische Anwendungen Grosses Interesse an WSNs zeigt auch das Militär, welches insbesondere in den Anfangsjahren der Forschung viele Projekte finanziert hat. Mögliche Anwendungen sind hier generelle grossflächige Überwachungen, automatische Lokalisierung von Heckenschützen, Verfolgung von radioaktivem Material über grosse Distanzen oder auch „intelligente“ Minenfelder die versuchen sich selbständig zu reparieren.

4.2 Abgrenzung zu verwandten Gebieten

Nachdem das vorherige Kapitel eine Einführung in das Gebiet der Wireless Sensor Networks gegeben hat, sollen in diesem Kapitel nun verwandte Technologien und deren Beziehung zu WSNs betrachtet werden.

MANETs

Mobile Ad-hoc NETworks sind drahtlose ad-hoc Netzwerke, welche sich selbst konfigurieren und in denen die Kommunikation über multi-hop Routing von einem Knoten zum anderen abgewickelt wird. Die Knoten sind dabei meist mobil und entsprechend kann sich die Netzwerktopologie sehr rasch verändern. MANETs können dabei entweder alleine stehen, oder über Gateways mit einem grösseren Netz wie z.B. dem Internet verbunden sein.

Anhand der im vorherigen Kapitel aufgezeigten Charakteristiken von WSNs lässt sich sehen, dass MANETs eine mögliche Ausprägung eines WSN sein können. Insbesondere die traditionellen Anwendungen von WSNs gehen von einem drahtlosen, ad-hoc Netzwerk aus in dem die Knoten über muli-hop Routing kommunizieren. Wie aber auch schon beschrieben, ist das in den heutigen Anwendungen nicht mehr immer der Fall. Ein Hauptunterschied selbst zur traditionellen Sicht von WSNs ist die Mobilität. Während bei MANETs grundsätzlich von ständig mobilen Knoten ausgegangen wird, ist dies bei WSNs nur in wenigen Anwendungen der Fall.

Mobile Grids

Zentraler Gedanke des Grid Computing ist die gemeinsame Nutzung verteilter Resourcen. In erster Linie Rechenzeit, aber auch anderer Resourcen wie Speicherkapazität, Bandbreite usw. Mobile Grids erweitern diesen Kerngedanken (der fest vernetzte Knoten im Zentrum hatte) auf mobile Knoten die drahtlos miteinander oder einem grösseren, externen Netz kommunizieren.

Ein WSN könnte grundsätzlich die Aufgabe eines Mobile Grids übernehmen oder ein bestehendes Mobile Grid ergänzen, es gibt allerdings eine Reihe von Problemen. Die Resourcen von WSNs und insbesondere die Energievorräte sind stark beschränkt. Wenn Knoten nicht für ihren Hauptzweck aktiv sind, ist es meist wünschenswert dass, sie „schlafen“ und ihre Resourcen während dieser Zeit nicht für andere Zwecke verschwenden. Ein weiteres

grosses Hindernis ist, dass im Grid Computing Resourcen organisationsübergreifend zur Verfügung gestellt werden sollten. WSNs hingegen sind meist für einen spezifischen Einsatzzweck entworfen, und ihr Eigentümer dürfte nur in den seltensten Fällen bereit sein die Resourcen gegenüber Fremden zu öffnen.

RFID

Radio Frequency Identification bietet eine automatische Identifzierung von Objekten die mit RFID Tags versehen wurden. In ihrer *passiven* Form enthalten RFID Tags keine integrierte Energiequelle, Strom der durch ein externes Signal induziert wurde wird genutzt um in einer kurzen Antwort die auf dem RFID Tag gespeicherten Daten zurück zu übermitteln. In den meisten Fällen wird nur ein Identifikator zurück geliefert der das Objekt eindeutig identifiziert, es können aber grundsätzlich beliebige Daten auf dem RFID Tag gespeichert werden. Diese Form bietet nicht wirklich eine Verwandtschaft zu WSNs.

RFID Tags gibt es allerdings auch noch in einer *aktiven* Variante, welche eigene Energiequellen besitzen und entsprechend selbständig kommunizieren können. In einigen Fällen werden solche aktiven RFID Tags auch mit Sensoren bestückt, was uns schon zu einer ziemlich nahen Verwandschaft mit dem Gebiet der WSNs bringt. Der Fokus liegt jedoch bei beiden Technologien unterschiedlich. Während WSNs eingesetzt werden um Phänomene übergreifend beobachten zu können, beschränkt sich die Anwendung von aktiven RFID Tags meist nur auf die Überwachung des einzelnen Objektes an welchem sie befestigt sind.

4.3 Aufbau und Kommunikationsmethoden

Dieses Unterkapitel beschäftigt sich mit dem Aufbau und den Kommunikationsmethoden von Sensornetzwerken. Im ersten Unterkapitel wurden Definitionen und Charakteristiken von Sensornetzwerken vorgestellt. Hier wird vor allem auf die Umsetzung und das Funktionieren dieser Netzwerke eingegangen.

4.3.1 Generelle Problematik

Die Natur der Wireless Sensornetzwerke führt zu spezifischen technischen Herausforderungen. In diesem Abschnitt werden die für die Umsetzung speziellen Eigenschaften zusammenfassend genannt und es wird dabei kurz erläutert, wo die Problematik bei der Umsetzung von Sensornetzwerken liegen. Der Abschnitt geht auf [2, 1] zurück.

Energieeffizienz

Da drahtlose Sensornetzwerke ihre Energie aus kleinen endlichen Energiequellen beziehen, stellt die Energieeffizienz die wichtigste Herausforderung dar. Energieeffizienz ist in

Sensornetzwerken allgegenwärtig. Sowohl beim Einsatz der Sensoren, als auch bei der Verarbeitung der gemessenen Daten und bei der Datenvermittlung. Jeder Aspekt ist von der Energieeffizienz betroffen. Da der Prozessor und allen voran die Kommunikation die grössten Energieverbraucher sind, ist jedoch das Interesse an diesen beiden Orten zu sparen, am grössten.

Ad hoc Platzierung und Verwendung von Knoten

Die meisten Knoten werden in Gebieten platziert, in denen wenig oder gar keine Infrastruktur vorzufinden ist. Unter solchen Voraussetzungen wird es in der Verantwortung der Knoten sein, den Initierungsprozess durchzuführen. Das heisst, sie müssen nach der Platzierung eine Verbindung zu einander aufbauen und ihre Lokalität feststellen können.

Dynamische Änderungen

Es ist ausserdem wichtig für ein Sensornetzwerk anpassbar zu sein. Das heisst, es muss reagieren können auf Einflüsse unterschiedlicher Art. Beispielsweise muss es stabil bleiben, auch wenn Knoten ausfallen oder hinzugefügt werden. In diesem Fall sollte das Netzwerk fähig sein, sich neu zu sammeln und neu zu konfigurieren. Spezifische Netzwerke, die nur unter bestimmten Voraussetzungen arbeiten sollten, wie zum Beispiel nur am Tag, sollten ausserdem fähig sein, in der Nacht auf Standby zu schalten. Damit kann Energie gespart und ihre Lebenszyklus verlängert werden.

Kommunikation

Oft haben Sensorknoten keine globale ID, weil die Anzahl Knoten sehr hoch ist und ein grosser Overhead entstehen kann. Diese Eigenschaft führt dazu, dass die Knoten keine Point-to-Point-Kommunikation verwenden können. Statt dessen wird für die Kommunikation die Attribut- oder Datenbasierte Adressierung verwendet. Dabei wird eine Abfrage an ein Sensornetzwerk gerichtet, welches Sql-Abfragen bei relationalen Datenbanken ähnlich ist. Beispielsweise kann die Abfrage an ein Sensornetzwerk, dessen Knoten die Temperatur messen, vereinfacht lauten: „An welchen Orten weist die Temperatur einen Wert über 20 Grad Celsius auf?“

4.3.2 Sensorknoten

Die neusten technischen Errungenschaften haben dazu geführt, dass die Knoten in WSN immer kleiner und günstiger wurden und dabei auch noch weniger Strom benötigen. Des weiteren sind sie fähig lokale Prozesse durch den Einsatz eigener Miniprozessoren durchzuführen und kabellos miteinander zu kommunizieren. In den folgenden Abschnitten wird detailierter auf den Aufbau und die Eigenschaften der Knoten in einem typischen Sensornetzwerk eingegangen.

Bausteine von Knoten

Knoten bestehen im Wesentlichen aus vier Bausteinen [26]. Dem Mikroprozessor, dem Kommunikationsmittel, den Sensoren und der Energieressource. In Abbildung 4.1 werden sie dargestellt und sie werden im folgenden Abschnitt näher beschrieben.

Rechnereinheit Die Rechnereinheit besteht aus einem Mikroprozessor. Dieser ist verantwortlich für die Steuerung der Sensoren und für die Ausführungen der Kommunikationsprotokolle. Normalerweise arbeiten die Rechnereinheiten unter verschiedenen Modi. Damit lässt sich einrichten, dass die Einheit auf Energieverbrauch optimiert, das heisst der Konsum minimiert werden kann.

Kommunikationseinheit Die Kommunikationseinheit besteht aus einem Sender und Empfänger für Funkwellen, der für kurze Distanzen ausreicht. Die Distanz muss von einem Knoten zu seinen Nachbarknoten reichen. Eine solche Einheit kann sich in verschiedenen Zuständen befinden. Entweder sie übermittelt (*transmit*), empfängt (*receive*), wartet oder schläft (*idle*). Es ist entscheidend für die Energieeffizienz, dass die Kommunikationseinheit nicht im Schlafmodus verweilt, wenn sie nicht gebraucht wird, sondern, dass sie in dieser Zeit ausgeschaltet werden kann.

Sensoreinheit Ein Knoten kann ein oder mehrere Sensoren haben. Die Sensoren bilden die Verbindung der Knoten zur Aussenwelt. Sie können Licht, Wärme, Feuchtigkeit, Temperatur oder andere Phänomene in irgend einer Weise wahrnehmen. Um den Energiekonsum klein zu halten, sollten jedoch die Sensorkomponenten möglichst leistungsschwach ausgewählt werden, so dass sie ihre Aufgaben noch erfüllen können.

Energiespeichereinheit Die Energiespeichereinheit besteht aus einer Batterie, welche den Knoten mit Energie versorgt. Beim Konsum von Energie ist darauf zu achten, dass er nicht in kurzer Zeit mit grosser Intensität verläuft, sondern mit schwacher Intensität über längere Zeit. Damit kann die Lebensdauer der Batterie und damit des Sensornetzwerkes verlängert werden.

4.3.3 Kommunikationsmethoden

Während in anderen Netzwerken die Protokolle in erster Linie auf einen hohen Durchsatz, eine kleine Latenz und im speziellen Ad hoc-Netzwerke auf Knotenmobilität optimiert werden, steht in drahtlosen Sensornetzwerken vor allem die Energieeffizienz im Vordergrund.

Ausserdem sind adressbasierte Routingverfahren, wie sie beispielsweise im Internet verwendet werden, für drahtlose Sensornetzwerke nicht anwendbar. Netzwerke, in denen die Knoten keine Adressen haben, verwenden für die Kommunikation *Flooding* (Fluten) [1]. Beim Flooding wird mit Broadcast eine Nachricht in das Netzwerk gesendet und alle Knoten, welche die Nachricht empfangen, senden diese weiter. Das Fluten wird dann beendet,

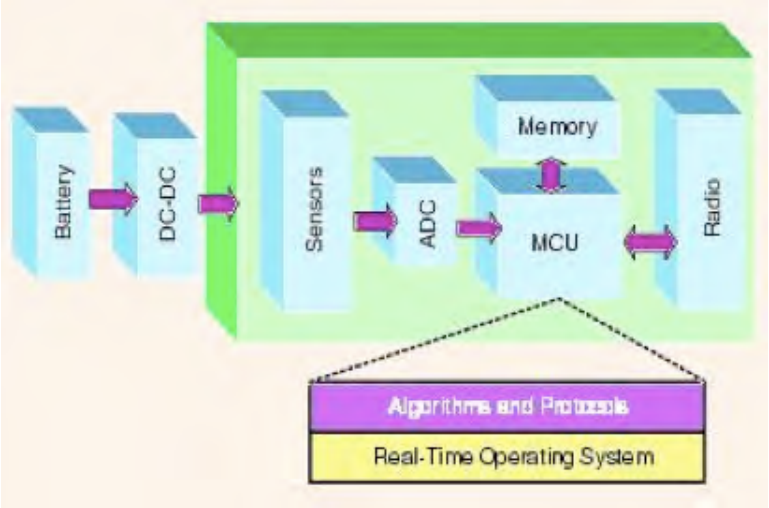


Abbildung 4.1: Systemarchitektur eines typischen Knoten [2]

wenn eine bestimmte Anzahl Hops hinter sich gelassen worden sind oder wenn das Ziel erreicht ist. Das grundsätzliche Problem des Fluten ist die zu grosse Anzahl Nachrichten, die versendet werden. Damit wird der Energieverbrauch unnötig vervielfacht. In [19] wird dabei zwischen *Implosion* und *Overlap* unterschieden. Implosion entsteht in Situationen, bei denen mehrere Nachrichten an einen Knoten gesendet werden, welche alle dieselben (redundanten) Informationen haben. Weil eine Nachricht in den meisten Anwendungen ausreichen würde, wird unnötig viel Energie durch die redundante Nachrichteneingänge verschwendet. Overlap entsteht dadurch, dass mehrere Knoten dieselben Werte mit ihren Sensoren messen, wenn sie beispielsweise räumlich nahe sind. Die Temperatur von zwei Knoten ist sehr ähnlich, wenn sie nicht weit genug voneinander entfernt sind. Eine Verbesserung ist das *Gossiping* (Tratschen) [1]. Dabei wird eine Nachricht nicht mehr allen Nachbarknoten gesendet, sondern nur zufällig ausgewählten Knoten. Dadurch soll die Anzahl versendeter Nachrichten reduziert werden. Auch dieses Verfahren ist jedoch für Sensornetzwerke nur suboptimal, weil es nicht eben systematischer vorgeht. Ein weiterer Ansatz besteht in der Aggregation [18], bei dem die Daten bereits im Netz beim Transportieren von einem Knoten zum anderen zu verarbeitet werden. Um Energie zu sparen, wird dabei versucht die Anzahl Nachrichten zu minimieren. Nehmen beispielsweise mehrere räumlich nahe liegende Knoten mit den Sensoren dieselben Messdaten auf (bsp. die Temperatur), so können diese Daten konsolidiert werden und es braucht nur noch eine (konsolierte) Nachricht weitergeleitet zu werden.

Aus diesen Beschreibungen geht hervor, welche Anforderungen ein Kommunikationsprotokoll für drahtlose Sensornetzwerke genügen muss. Die Nachrichten sollten möglichst effizient, dh. energiesparend versendet werden können. Dabei muss sowohl die Quantität als auch die Qualität (wenig Redundanzen und kurze, direkte Wege) der Nachrichtenflüsse berücksichtigt werden. Im Folgenden werden drei Routingverfahren vorgestellt, welche in Sensornetzwerken zum Einsatz gelangen. Diese Protokolle bedienen sich grundsätzlich unterschiedlicher Ansätze. Dabei können sie nur bedingt miteinander verglichen und gewertet werden. Der Grund liegt darin, dass es sehr von der Anwendung abhängt, welcher Ansatz der geeignetste ist.

Negotiation-basiertes Routing

Das meist verwendete Routingverfahren dieser Kategorie heisst SPIN (Sensor Protocols for Information via Negotiation) [19]. Dieses Verfahren ist eine Variante kontrollierten Floodings. Es werden Daten verbreitet, indem sie jeweils an alle Nachbarknoten versendet werden. SPIN löst die Probleme des Flooding (Implosion und Overlap) mit einem Handshake-Protokoll. Bevor Daten von einem Knoten zu anderen übermittelt werden, wird geprüft (verhandelt), ob der Knoten die Daten bereits hat oder nicht. Dazu werden *Metadaten* verwendet, welche normalerweise nur einen Bruchteil der Grösse der eigentlichen Daten haben. Der Inhalt zweier Metadaten ist unterschiedlich, wenn auch deren eigentliche Daten unterschiedlich sind. Und umgekehrt. Das Protokoll ist dreistufig, wobei drei verschieden Nachrichtentypen gebraucht werden.

- *ADV:* wird benötigt, um Daten anzubieten. Wenn ein Knoten neue Daten erhält oder selber erzeugt, dann kann er dies all seinen Nachbarn mit einer ADV-Nachricht ankündigen.
- *REQ:* wird benötigt, um Daten anzufordern. Sobald ein Knoten die Daten möchte, die er mit einer ADV-Nachricht angekündigt erhalten hat, sendet er diese REQ-Nachricht.
- *DATA:* sind die eigentlichen Daten. Eine DATA-Nachricht erhält die tatsächlichen Daten und einen Header in Form von Metadaten.

Das Ziel beim SPIN-Protokoll ist, die Grösse der versendenden Daten klein zu halten und somit Energie sparen zu können. Das gelingt solange die Metadaten um Grössenordnungen kleiner sind als die eigentlichen DATA-Nachrichten.

In der Abbildung 4.2 wird die Verwendung der Nachrichtentypen illustriert. Der Knoten A sendet eine ADV-Nachricht an Knoten B. Der Knoten B möchte die offerierte Nachricht erhalten und teilt dies mit einer REQ-Nachricht mit. Nachdem die Antwort von B in Knoten A angekommen ist, wird die Nachricht (DATA) von A nach B versendet. In selben Stil fährt B mit seinen Nachbarn fort.

Flaches Routing

Das *Flat networks routing*-Verfahren ist dadurch gekennzeichnet, dass alle Knoten gleichberechtigt sind. Es besteht also keine Hierarchie. Der bekannteste Vertreter dieses Verfahrens ist das *Directed Diffusion* [17]. Voraussetzung dieser Methode ist das Wissen eines jeden Knoten über seinen Ort. Denn die Nachrichten werden mit dem Ort und optional mit weiteren Attributen versehen. Der Ort muss enthalten sein, damit die Richtung des Flooten kontrolliert werden kann. Wenn der Ort in einer Nachricht enthalten ist, kann ein Knoten, der die Nachricht erhält entscheiden, ob er die Nachricht weitersendet oder nicht. Befindet sich der empfangende Knoten auf dem Weg zum Zielknoten, so wird er die Nachricht weitersenden, andernfalls nicht. Dabei kommt ein Gradientenverfahren zum Einsatz, bei dem ein Gradient eines Knoten umso höher festgesetzt wird, je direkter der

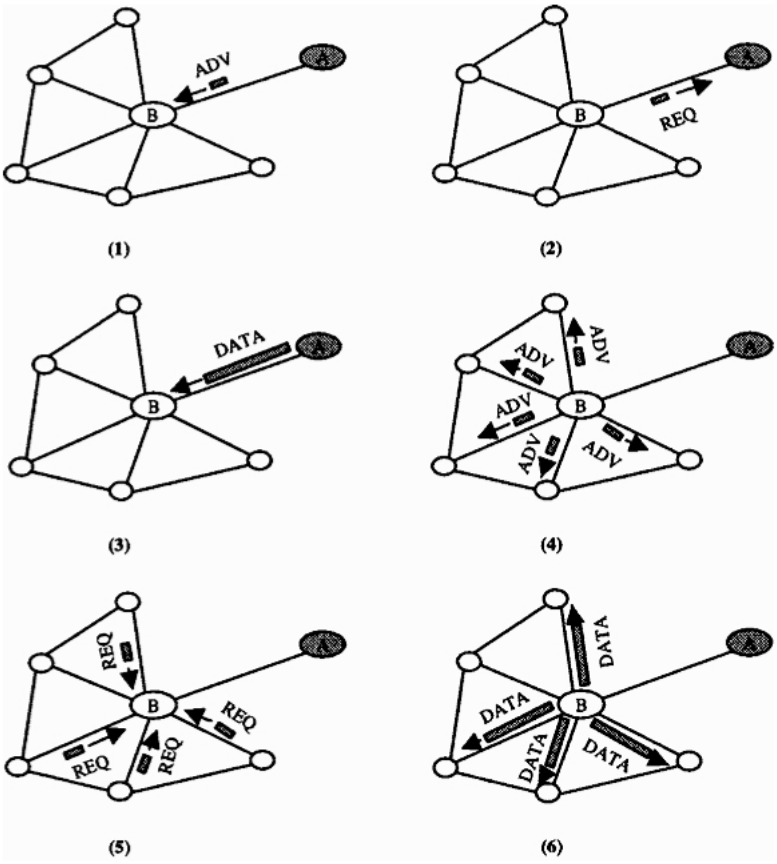


Abbildung 4.2: SPIN-Protokoll [19]

Weg zwischen dem Zielknoten und des Quellknotens des Interesses liegt. Dadurch wird das Fluten auf eine Richtung beschränkt, weil so ein Pfad von der Quelle zur Senke aufgebaut wird, entlang dessen die Daten geroutet werden können.

Hierarchisches Routing

Das bekannteste Protokoll des hierarchischen Routingverfahrens ist LEACH (Low-Energy Adaptive Clustering Hierarchy) [13]. Das hierarische Routing hat einerseits *Runden* und andererseits *Phasen*, die innerhalb einer Runde durchlaufen werden. Eine Runde besteht aus einer Konfigurations- und einer Betriebsphase. In der Konfigurationsphase werden zuerst Clusterführer auserkoren. Die Wahl wird mit einem Zufallsgenerator vorgenommen. Jeder Knoten generiert dabei eine Zufallszahl zwischen 0 und 1. Liegt nun die Zahl unterhalb eines Schwellwerts, dann wird der Knoten zum Clusterführer erwählt, andernfalls nicht. Sind die Clusterführer gewählt, dann sendet dieser eine Meldung an seine Umliegenden Knoten. Die Knoten, welche keine Clusterführer sind, empfangen diese Meldungen. Sie wählen den Clusterführer, dessen Meldung die höchste Signalstärke aufweist und senden diesem eine Bestätigung, dass sie sich ihm angeschlossen haben. Dabei werden die Clusterführerschaften gleichmässig unter allen Knoten verteilt, damit der Energiestand im Netzwerk möglichst gleichverteilt bleibt. In der Betriebsphase nimmt Clusterführer alle Nachrichten seiner Kindsknoten auf und kommuniziert mit einer Basisstation. Er

verbraucht entsprechend mehr Energie. Damit sich unter die Kindsknoten bei der Kommunikation mit dem Clusterführer nicht gegenseitig stören, wird vom Clusterführer ein TDMA-Schedule erzeugt, wodurch allen Kindsknoten ein fixes Zeitfenster für die Kommunikation mit ihm zugeteilt wird.

4.3.4 Lokalisierung und Ortung

In Sensornetzwerken werden die Knoten ungeordnet in einem Gebiet verstreut. Die Knoten haben damit kein *a priori* Wissen über ihren Standort. Es muss nach der Platzierung festgestellt werden, wo sich die Knoten befinden. Dabei gibt es verschiedene Kriterien, welche zuerst in Betracht gezogen werden müssen:

Physikalische Position vs. symbolische Ortsangabe Die physikalische Positionierung ist die Beschreibung der Position eines Knoten durch Koordinaten. Die symbolischen Ortsangaben werden Namen von Gebietsausschnitten angegeben. Beispielsweise ist der Raum 2.B.11 in der Binzmühlstrasse 14 in Zürich eine symbolische Ortsangabe.

Absolute vs. relative Koordinaten In einem absoluten Koordinatensystem gibt es eindeutige und statische Referenzpunkte und das ganze Koordinatensystem richtet sich nach diesen Referenzpunkten. In relativen Koodinatensystemen gibt es keine Referenzpunkte. Deshalb können für ein Knoten mehrere relative Koordinaten zutreffen.

Genauigkeit Die Genauigkeit spielt in einem Ortungssystem eine grosse Rolle. Es gibt zwei Arten, wie die Genauigkeit verstanden und gemessen werden kann. Einerseits kann die Abweichung der gefundenen Position im Vergleich zu der realen Position damit verstanden werden. Andererseits kann die prozentuelle Angabe für die Häufigkeit für das Erreichen einer bestimmten Genauigkeit damit gemeint sein. Beispielsweise soll bei der Lokalisierung eine Genauigkeit von 0.5m in 95% aller Fälle erreicht werden.

Topologie Es gibt Sensornetzwerke, die vollständig ad hoc aufgebaut sind. Diese haben überhaupt keine Infrastruktur und auch keine Referenzknoten. Eine andere Variante wäre der Aufbau einer gewissen Infrastruktur. Diese bietet einige Referenzknoten, welche die anderen Knoten helfen, sich zu orientieren.

Einsatzbeschränkungen Ein Sensornetzwerk sind selten uneingeschränkt verwendbar. Es gibt Einschränkungen bei der Funkreichweite, bei der Skalierbarkeit oder bei der Umgebung. Während ein System in Gebäuden geeignet ist, kann es im Freien nicht umgesetzt werden. Als bekanntes Beispiel kann hier GPS angeführt werde, welches sich im Freien sehr eignet, jedoch in Gebäuden mit Schwächen versehen ist.

GPS

Die naheliegenste Herangehensweise, um das Problem der Positionsbestimmung zu lösen, wäre der Einsatz von GPS (Global Positioning System) einzuführen. Es gibt jedoch auch

gute Gründe, die dagegen sprechen. Zum einen ist GPS auf den Einsatz unter freiem Himmel eingeschränkt. Und zum anderen sind die Empfänger teuer, was nicht den Eigenschaften der Knoten für Sensornetzwerke entspricht, die möglichst kleine und günstige Knoten brauchen. Des weiteren begibt man sich in die Abhängigkeit fremder Infrastrukturen, was je nach Anwendung und Situation wenig wünschenswert ist.

Um die Position der Knoten ausfindig zu machen werden geometrische Konzepte verwendet. Dabei gibt es Lösungswege über die Entfernung zwischen Knoten (Lateration) und über Winkel (Angulation).

Trilateration

Die Methode, die am häufigsten verwendet wird, um die Position der Knoten ausfindig zu machen, ist Trilateration. Es handelt sich um eine Lateration, bei der man drei Referenzknoten heranzieht. Bei Multilateration werden mehr als drei Referenzknoten gebraucht. Bei der Trilateration sind also drei Referenzknoten nötig, die ein Signal aussenden. Diese Referenzknoten sind oft komplexer aufgebaut als die anderen. Ihre Standorte müssen bekannt sein, möglicherweise durch GPS. Bei der Trilateration werden die Abstände zwischen einem Knoten zu den Referenzknoten ermittelt. Dabei misst man beispielsweise die Laufzeit der Signale. Ist die Ausbreitungsgeschwindigkeit der Signale bekannt, kann daraus die Distanz abgeleitet werden. Mit der Ermittlung der Distanzen zu allen Referenzstationen, kann die Position eines Nicht-Referenzknoten ermittelt werden, indem Kreise mit der ermittelten Distanz als Radius um die Referenzpunkte gezogen werden. Wie in Abbildung 4.3 dargestellt wird, ist der Schnittpunkt der drei Kreise die Position eines Knoten.

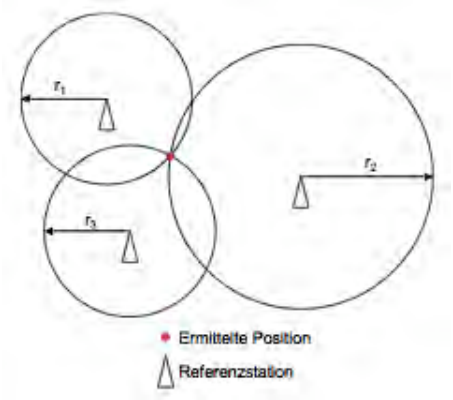


Abbildung 4.3: Verfahren der Trilateration [2]

Es gibt erweiterte Varianten, bei denen die Knoten, deren Position ermittelt wurde, wiederum für weitere Knoten als Referenzknoten herangezogen werden können.

Mit dem Einbeziehen der Winkel (Angulation) kann aus nur zwei Referenzknoten die Position eines Knoten ermittelt werden. Hat man zwei Winkel eines Dreiecks und die Länge zweier Seiten, dann kann die Position des dritten Punktes zu bestimmen.

Das Prinzip der Trilateration ist nicht nur für den zweidimensionalen Raum anwendbar, sondern auch für den dreidimensionalen Raum. Da eine Dimension hinzukommt, braucht es nicht mehr nur drei Referenzknoten, sondern deren vier. Das Prinzip wird unter anderem auch für GPS verwendet, wie in Abbildung 4.4 zu sehen ist.

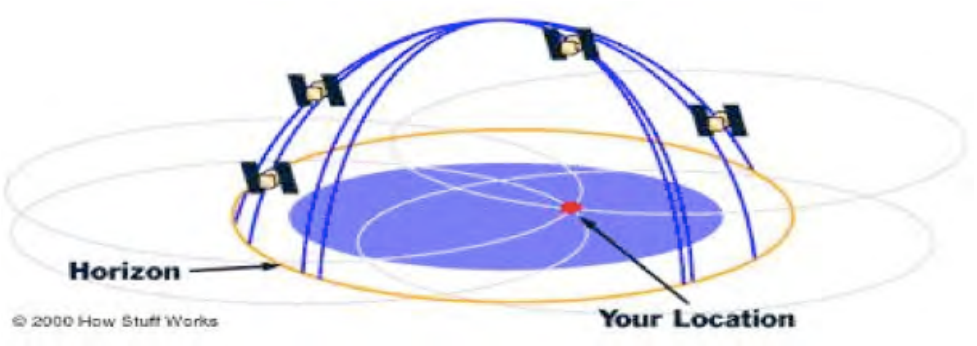


Abbildung 4.4: Verfahren der Trilateration im dreidimensionalen Raum [2]

Sind keine Referenzknoten vorhanden, dann kann zwar die Position der Knoten nicht mehr an das Weltkoordinatensystem ausgerichtet werden. Aber es können Verfahren angewendet werden, welche die Position der einzelnen Knoten innerhalb des Netzwerkes bestimmen. Es wird in diesem Falle ein lokales Koordinatensystem entwickelt. Ob das ausreicht oder nicht, hängt von der Anwendung ab.

4.4 Aktueller Stand der Technik

Durch die extrem unterschiedlichen Einsatzgebiete unterscheiden sich auch die Techniken sehr stark voneinander. Einen Überblick über den Stand der Technik zu geben ist dementsprechend schwer und kann nur mit entsprechenden Abstraktionen in einem sinnvollen Ausmass realisiert werden. Das folgende Kapitel ist in die Bereiche Hardware, Software und Anwendungen unterteilt. Im ersten Unterkapitel „Hardware“ werden typische Eigenschaften betrachtet welche Sensorknoten, dem heutigen Stand der technischen Entwicklung entsprechend, aufweisen. Im zweiten Abschnitt wird zuerst ein Überblick gegeben über verschiedene Ansätze für die Entwicklung von Software für WSN, danach werden verschiedene Softwarestandards vorgestellt und erklärt. Zum Schluss wird im dritten Teil dieses Kapitel ein kurzer Überblick gegeben über Forschungsprojekte und kommerziell vertriebene Systeme die diese Technologie einsetzen.

4.4.1 Hardware

In WSN-Systemen wird je nach Anwendungsgebiet und speziellen Anforderungen oft eigene Hardware entwickelt oder bestehende angepasst, daher existiert eine Vielzahl an unterschiedlichen Sensorknoten. Es gibt allerdings eine Reihe von standardisierten Sensorknoten welche flexibel angepasst und eingesetzt werden. Sie werden oft von Forschungsinstitutionen entwickelt und vertrieben.

Nachfolgend werden nun kurz die gängigsten Eigenschaften beschrieben welche diese heutzutage aufweisen. Danach werden zwei bekannte Sensorknoten genauer vorgestellt um einen Einblick in die konkrete Ausgestaltung von Sensorknoten zu bekommen.

Gängige Eigenschaften

Aufbau: Die Sensorknoten basieren meist auf einem RISC (Reduced Instruction Set Computer) Mikrokontroller mit einem relativ kleinen Programm und Datenspeicher (um die 100KB). Sie sind entweder modular oder integral aufgebaut. Ein modularer Aufbau ist dynamischer, da verschiedene Sensoren auf eine Basiseinheit aufgesetzt werden können und er so vielseitig eingesetzt werden kann. Da er aus mehreren einzelnen Modulen besteht, ist die Konstruktion aber entsprechend instabiler im Aufbau. Ein Sensorknoten der integral aufgebaut ist ist stabiler und besser geeignet um in einem rauem Umfeld, wie zum Beispiel der Natur, eingesetzt zu werden, besitzt aber den Nachteil, dass er eine fixe Struktur besitzt und nicht flexibel ist.

Kommunikation: Die Kommunikation erfolgt in den meisten Fällen über Funk. Visuelle Kommunikation ist zwar billiger, konsumiert weniger Energie und ist auch einfacher in der Konstruktion als eine Funkeinheit, bietet aber einen wesentlichen Nachteil: Sie benötigt direkten visuellen Kontakt zwischen zwei Knoten um eine Kommunikation zu ermöglichen und ist so weniger flexibel einsetzbar.

Energieversorgung: Die Batterie stellt die gängigste Art der Energieversorgung dar. Alternative Ansätze, welche Energie aus der Umwelt beziehen, wie zum Beispiel die Nutzung von Solarenergie oder natürlichen Temperaturschwankungen, konnten sich bis heute nicht an breiter Front durchsetzen. In der Interaktion mit der Umwelt lauern viele Schwierigkeiten für die bis heute noch keine umfassenden, zufriedenstellenden Lösungen entwickelt wurden. So stellt die Batterie oft den limitierenden Faktor für die Energieressourcen eines Sensorknotens dar und determiniert gleichzeitig mit ihrem Umfang auch wesentlich die Grösse desselben.

Grösse: Die Grösse der Sensorknoten bewegt sich im Normalfall im Rahmen von einigen Kubikzentimetern. Da der Umfang eines Sensorknotens grösstenteils von der Batterie bestimmt wird, müssten für eine relevante Verkleinerung der Knoten andere Technologien für die Energieversorgung genutzt werden. Die Grösse von einigen Kubikzentimetern ist aber für die meisten Anwendungen genügend klein.

Modelle von Sensorknoten

Wenn in einem Projekt keine spezielle Hardware entwickelt wird, so werden oft standardisierte Produkte verwendet. Einige der bekanntesten davon sind der „BTnode“ der Eidgenössischen Technischen Hochschule (ETH) Zürich, „Mica2“ der Crossbow Inc., „EYES“

der University of Twente, „Imote“ von Intel und „Telos A“ von Moteiv. Diese Knoten entsprechen alle in etwa den gängigen Eigenschaften wie sie im vorhergehenden Abschnitt ausgeführt wurden, wenn sie sich auch natürlich hinsichtlich der konkreten Ausgestaltung voneinander unterscheiden.

Um eine Vorstellung davon zu bekommen wie so ein Sensorknoten konkret aussieht werden nun zwei dieser Modelle als Beispiele genauer unter die Lupe genommen.

BTnode [8]: Der BTnode wurde an der ETH Zürich entwickelt und wird auch von ihr vertrieben. Er dient als Demonstrations- und Prototypplattform für die Forschung mit mobilen und Ad-Hoc Netzwerken. Das Gerät kommuniziert über Bluetooth oder über eine zweite, alternative Funkeinheit. Die zwei Kommunikationseinheiten können simultan benutzt oder individuell ausgeschaltet werden um den Energieverbrauch minimieren zu können. Die Energie bezieht das Gerät von zwei Batterien welche an das Gerät angeschlossen werden. Der Knoten läuft mit einem ATmega128l Mikrokontroller und besitzt 128kB Programmspeicher, 64kB Datenspeicher und 128kB Storage Speicher. Der Sensorknoten besitzt eine Grösse von 1890 mm2.

Imote [24]: Dieser Sensorknoten wurde von Intel entwickelt um den speziellen Anforderungen von Anwendungen im Bereich der industriellen Überwachung gerecht zu werden. Diese stellt durch bestimmte Messungen wie Beschleunigung oder Vibration, höhere Anforderungen an die Bandbreite und Samplingintervalle. Grössere Datenmengen sollen effizient übertragen werden können. Die aus diesen Anforderungen resultierende Architektur besteht aus einem ARM7 Mikrokontroller und einer Bluetooth Funkeinheit die Nachrichten innerhalb eines Radius von ca 30m verschicken kann. Sie beinhaltet weiter 512kB Programmspeicher und 11kB Datenspeicher und ist insgesamt 9000mm2 gross.

4.4.2 Software

Wie schon bei der Hardware, zeigt sich auch im Bereich der Software ein durchaus heterogenes Bild. Daher werden nun im ersten Teil dieses Abschnittes verschieden Programmierparadigmen vorgestellt, die einen Eindruck vermitteln sollen welche verschiedenen Ansätze bei der Entwicklung von Software für WSN angewendet werden können. Danach wird das TinyOS vorgestellt, das einen Quasi Standard für Betriebssysteme auf Sensorknoten darstellt. Zum Schluss werden die zwei Datenübertragungsstandards IEEE 802.15.4 und ZigBee erläutert.

Programmierparadigmen für WSN

Die Vielzahl der unterschiedlichen Einsatzgebiete von WSN zeigt sich auch in den verschiedenen Umsetzungen. Um eine Übersicht über die verschiedenen Ansätze von Software für WSN zu schaffen haben Salem Hadim and Nader Mohamed [15] ein Schema entwickelt.

Dieses teilt die verschiedenen Programmiermodelle zuerst in die zwei Klassen der Abstraktion (programming abstractions) und der Unterstützung der Programmierung. Die Ansätze der ersten Klasse beschäftigen sich mit der Frage wie wir WSN wahrnehmen und bieten dazu Konzepte und Abstraktionen an. Die Zweite beschäftigt sich dagegen mit dem System und den Mechanismen, welche eine Software dem Netzwerk zur Verfügung stellt. Sie definieren Algorithmen für die Übertragung, Verbreitung oder die Ausführung von Code.

In Abb. 4.5 sehen wir die Anordnung der verschiedenen Modelle die nachfolgend vorgestellt werden.

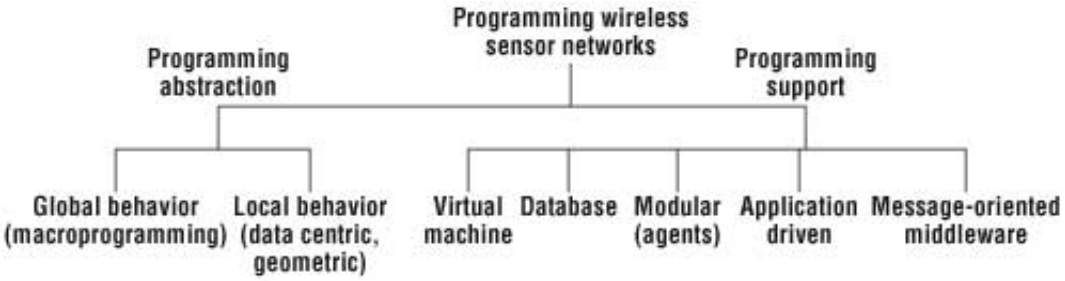


Abbildung 4.5: ProgrammierModelle nach Hadim und Mohamed [15]

Globales Verhalten (global behavior):

Der Ansatz des globalen Verhaltens wird auch als Makroprogrammierung bezeichnet. Es erlaubt dem Programmierer das Verhalten des WSN als Ganzes zu programmieren. Das Verhalten der einzelnen Knoten wird danach automatisch generiert, so braucht sich der Programmierer nicht mit der low-level Spezifikation auseinanderzusetzen.

Ein Beispiel für die Entwicklung einer solchen Software ist Kairos [11]. Dieses System bietet dem Programmierer im Grunde drei wichtige Abstraktionen an um das Verhalten des Sensornetzes zu beschreiben. So werden, erstens, die Knoten mit IDs identifiziert welche allerdings keinen Zusammenhang mit der Topologie des Netzwerkes aufweisen müssen sondern lediglich einer logischen Zuteilung entsprechen. Hinzu kommt die getNeighbours() Funktion, welche es dem Entwickler erlaubt alle Knoten, die über genau einen Hop mit einem bestimmten Knoten verbunden sind anzusteuern. Die dritte Abstraktion ermöglicht einen direkten Zugriff auf aggregierte Daten der einzelnen Knoten.

Anhand dieser Abstraktionen können Algorithmen für das WSN programmiert werden. Der Code wird dann zuerst von einem Präprozessor, einer Erweiterung des Compilers, bearbeitet um einen kommentierten Quellcode zu erzeugen. Danach wird dieser Quellcode kompiliert und das Programm, welches nun die Algorithmen enthält, an die einzelnen Knoten übertragen.

Lokales Verhalten (local behavior):

Dieser Ansatz basiert auf der Idee, dass in vielen Anwendungen von WSN der Fokus nicht darauf gelegt wird alle gemessenen Informationen zu sammeln und auszuwerten sondern lediglich Ausnahmesituationen zu registrieren. So sind zum Beispiel bei einem System

das Waldbrände aufspüren soll, lediglich die Messwerte relevant, welche Temperaturen anzeigen die über einem gewissen Grenzwert liegen. Diejenigen, welche Messwerte im normalen Bereich aufweisen interessieren nicht und können somit ignoriert werden.

VigilNet ist ein Projekt, das sich das Aufspüren und Verfolgen bestimmter Objekte, z.B Fahrzeuge oder Feuer, zum Ziel gemacht hat. Ein Programm das speziell für das Projekt VigilNet entwickelt wurde und eine Umsetzung dieses Ansatzes darstellt ist Enviro Track [14].

Das Routing und die Adressierung der verschiedenen Sensorknoten erfolgt in diesem Fall basierend auf den gesuchten Ausprägungen der Messwerte. Der Programmierer definiert diese Eigenschaften oder Kombinationen von Eigenschaften, welche Indikatoren für das gesuchte Objekt darstellen.

Werden nun diese Werte von Sensorknoten gemessen, werden diese Knoten gruppiert eingesetzt um das gesuchte Objekt zu orten. Die Knoten werden jeweils nur vorübergehend zu Gruppen zusammengeschlossen und nur für die Zeitspanne in welcher der auslösende Event, das Erkennen des gesuchten Objektes, vorhanden ist. Durch diese dynamische Gruppierungsmechanismen können auch mobile Objekte verfolgt werden indem entlang der Bewegung des Objektes jeweils die Sensoren eine Gruppe bilden, die das Objekt zu einem bestimmten Zeitpunkt wahrnehmen.

Die Gruppen haben jeweils einen Leiterknoten, der die gemessenen Daten der ganzen Gruppe aggregiert und diese zu einer Basisstation schickt.

Virtuelle Maschine (virtual machine):

Der Entwickler programmiert Applikationen in kleinen separaten Modulen, diese werden im Netzwerk verteilt und dann auf den einzelnen Knoten von einer virtuellen Maschine interpretiert. Programme für Virtuelle Maschinen (VM) sind weniger umfangreich. Dadurch müssen weniger Daten im Netzwerk übermittelt werden. Zusätzlich bietet eine VM auch eine grössere Sicherheit beim Ausführen der Programme, da sie allfällige Fehler abfangen kann.

Das Projekt Maté WSN [21] hat diesen Ansatz umsetzt. Es stützt sich auf das Bedürfnis nach neuen Programmierparadigmen um die Beschränkungen in Bezug auf Bandbreite und den enormen Energieverbrauch bei Netzwerkaktivitäten zu umgehen.

Maté ist eine Architektur, welche die Erstellung verschiedener virtueller Maschinen ermöglicht. Sie wurde entwickelt um die Reprogrammierbarkeit innerhalb eines WSN, vom Verändern einzelner Parameter bis hin zum Updaten ganzer Programmcodes, einfacher zu gestalten. Um die Übertragung von Daten zu vereinfachen werden alle Programme in 24 Byte lange Instruktionen unterteilt und erst dann ins Netzwerk eingeschleust. Dadurch werden grössere Programme bei der Übertragung im Netzwerk nicht benachteiligt.

Kombiniert mit Versionsnummern der einzelnen Instruktionen wird zum Updaten eines Netzwerkes einfach ein Knoten mit den neuen Versionen der Programme in das Netzwerk eingeführt. Sobald ein Nachbarknoten bemerkt das eine neue Version eines Programmes verfügbar ist, kopiert er dieses von seinem Nachbar. So wird nach und nach das ganze Netzwerk mit dem neuen Programmcode versorgt.

Datenbank (database):

Das WSN wird als virtuelle Datenbank betrachtet in der Sensordaten über eine Benutzerschnittstelle anhand Queries aus dem System abgefragt werden können. Ein solches System bietet aber keine Unterstützung für Real-Time Applikationen, da die Daten erst nach einer Abfrage zur Verfügung stehen.

Cougar [12] ist ein solches System. Die aggregierten Daten werden als relationale Datenbank betrachtet und die Sensordaten können anhand einer SQL-ähnlichen Sprache abgefragt werden. Die gespeicherten Daten werden als Relationen repräsentiert und die Sensordaten als Zeitreihen um die Abfrage zu vereinfachen. Im Netzwerk gibt es drei Arten von Sensorknoten. Die reinen Sensorknoten, die nur Daten von ihren Sensoren sammeln und diese weitergeben. Die Leiterknoten, die die Messwerte mehrerer Sensorknoten abholen und diese für eine Abfrage aufbereiten und die Gatewayknoten, welche die Schnittstelle zwischen Netzwerk und Anwendersystem definieren.

Für jede Anfrage wird vom Server ein Anfrageplan erstellt. Dieser Plan bestimmt dann den Ablauf der gesamten Anfrage, welche Rollen und Verantwortlichkeiten die einzelnen Knoten haben, wie die relevanten Knoten koordiniert werden und wieviele Berechnungen im Netzwerk durchgeführt werden.

Modulare Programmierung:

Diesem Ansatz liegt die Idee zugrunde: Je kleiner die Module eines Programmes, die in einem Netzwerk übertragen werden, desto weniger Energie wird gebraucht.

Das Projekt ZebraNet [27] hat ein WSN zur Beobachtung von Zebras in freier Wildbahn eingesetzt. Zu diesem Zweck musste das Netzwerk über längere Zeit autonom funktionieren um die Verhaltensweisen der Zebras möglichst unverfälscht messen zu können.

Impala ist eine Middleware die speziell für dieses Projekt entwickelt wurde. Impala besteht aus zwei Schichten. In der unteren Schicht befinden sich Eventfilter, Adapter und Updater und auf der darüberliegenden Schicht die Applikationen und Protokolle. Der Eventfilter veranlasst Reaktionen auf auftretende Events. Der Adapter modifiziert Applikationen nach verschiedenen Szenarien wie Energieknappheit oder Wichtigkeit der Applikation für das System oder den Anwender. Der Updater kümmert sich um die Umstände und die Durchführung der Updates zwischen den einzelnen Knoten.

Um die Updates effizient zu halten besitzen alle Module Versionsnummern. Bevor neue Versionen eines Moduls ausgetauscht werden, werden die Versionsnummern verglichen, so können überflüssige Datenübertragungen vermieden werden.

Anwendungsorientiert (application driven):

Ein weiterer Ansatz besteht darin die Netzwerkverwaltung nicht wie üblich getrennt von den einzelnen Applikationen zu implementieren, sondern den Anwendungen Zugriff auf den Netzwerkprotokollstack zu gewähren. So kann der Programmierer das Netzwerk auf spezifische Anforderungen einer Applikation anpassen.

MiLAN (Middleware Linking Applications and Networks) [16] ist eine Middleware, die diese Idee umgesetzt hat. Man geht davon aus, dass in einem WSN die angestrebte Qualitiy

of Service (QoS) davon abhängt, wie gut sich die Netzwerkaktivitäten an die Umstände eines dynamischen Netzwerkes anpassen können. Allerdings genügt diese Anpassung allein nicht um die QoS über längere Zeit in dem Masse zu erfüllen wie sie die Applikationen benötigen. Die Idee: Die Anwendungen sollen proaktiv sein und das Netzwerk aktiv beeinflussen.

MiLAN sammelt die Anforderungen der Anwendungen an die QoS des Netzwerkes, die relative Wichtgkeiten der einzelnen Anwendungen innerhalb des Gesamtsystems und Informationen des Netzwerkes über die verfügbaren Ressourcen. Anhand dieser Informationen passt MiLAN die Netzwerkkonfigurationen ständig an, um die Anforderungen der Anwendungen möglichst gut zu erfüllen und so deren Lebendauer zu verlängern.

Nachrichtenorientiert (message oriented):

Bei dem nachrichtenorientierten Ansatz handelt es sich hauptsächlich um ein Kommunikationsmodel für ein verteiltes Sensor Netzwerk. Es wird ein Publish-Subscribe Mechanismus eingesetzt, um den Austausch von Nachrichten zwischen Knoten zu vereinfachen. Eine Stärke dieses Algorithmus ist die Unterstützung von asynchroner Kommunikation, die sich in einem WSN, in dem die Applikationen oft ereignisbasiert funktionieren, geradezu anbietet.

Umgesetzt wurde dieser Ansatz in [28]. Der Publish-Subscribe Mechanismus wurde in diesem System sogar noch erweitert. So werden Themen definiert denen die einzelnen Nachrichten zugeordnet und so klassifiziert werden. Sensorknoten können nun ein Thema abonnieren und erhalten dann alle Nachrichten die zu diesem Thema im Netzwerk publiziert werden. Nachrichten müssen folglich wenn sie publiziert werden einem Thema zugeteilt werden, um von anderen Knoten beachtet zu werden. Dadurch, dass nur diejenigen Nachrichten übermittelt werden die ein Knoten respektive eine Anwendung tatsächlich abonniert hat, kann der Energieverbrauch für die Datenübertragung wesentlich reduziert werden. Neben dem Publish-Subscribe Service besteht die Architektur dieser Software aus weiteren Komponenten, die für das Routing oder weitere Services wie zum Beispiel die Datenaggregation verantwortlich sind. Der Publish-Subsribe Service dient sowohl der Kommunikation zwischen den verschiedenen Komponenten als auch der Bereitstellung einer Liste mit den abonnierten Themen und den zu veröffentlichenden Nachrichten.

TinyOS [10] als Quasi Standard

Das Betriebssystem TinyOS wurde an der Universität von Berkeley entwickelt und hat sich zu einem de facto Standard in Anwendungen von WSN entwickelt. Es wird von einer breiten Allianz aus Entwicklern, Firmen, Universitäten und Regierungsorganisationen eingesetzt, weiterentwickelt und verbreitet. Ursprünglich generell für eigebettete, vernetzte Systeme entwickelt wird es heutzutage vor allem für Anwendungen im Bereich der Sensornetze eingesetzt. Um den Anforderungen nach Energieeffizienz und geringem Speicherplatzverbrauch gerecht zu werden, haben die Entwickler vom klassischen Aufbau eines Betriebssystem abgesehen und einen neuen Ansatz gewählt. Anstatt wie beim klassischen Aufbau Kernel und Anwendungen strikte zu trennen wurden sie hier sehr eng miteinander verbunden. Sowohl die Anwendungen als auch das Betriebssytem sind in Komponenten

aufgegliedert. Beim Kompilieren werden zusätzlich zu der Anwendung auch die benötigten Komponenten des Betriebssystems miteinbezogen. So enthält das kompilierte Programm sowohl die Anwendung, als auch Teile des Betriebssystems. Die Programmierung der Systemkomponenten und der Anwendungen erfolgt mit NesC [23], einer Erweiterung der Sprache C, die speziell für die Umsetzung des TinyOS entwickelt wurde.

Die einzelnen Komponenten können Daten austauschen indem sie die *Command*-Handler anderer Komponenten aufrufen und Anfragen(*Commands*) stellen. Dies sollten aber nur kurze Anfragen sein, welche keinen grossen Rechenaufwand erfordern. Für aufwändigere Berechnungen werden *Tasks* eingesetzt. Diese werden zentral von einem Scheduler durch einen FIFO Stack koordiniert um die Rechenleistung des Systems gerecht zu verteilen. Zudem besitzen die Komponenten *Event*handler. Diese reagieren auf *Events*, welche von der Hardware oder von anderen Komponenten ausgelöst werden. Auf die wiederum mit *Commands* oder *Tasks* reagiert werden kann.

IEEE 802.15.4 und Zigbee

Der ZigBee Standard [37] ist ein, von einen Zusammenschluss von kommerziellen Unternehmen entwickelter, Standard um eine gemeinsame Basis für kommerzielle Produkte im Bereich der kabellosen Sensornetzwerken zu erstellen. Dieser baut auf dem öffentlichen Standard des Institute of Electrical and Electronics Engineers (IEEE) 802.15.4 auf.

IEEE 802.15.4: Entwickelt von der Taskgroup IEEE 802.15.4 [30], wurde dieser Standard für low rate wireless private area networks (LRWPAN) im Mai 2003 veröffentlicht. Er definiert, basierend auf dem OSI/ISO Modell, die Bitübertragungs- und die MAC Schicht.

Bitübertragungsschicht:

Diese Schicht stellt auf drei verschiedenen Frequenzbändern insgesamt 27 Kanäle für die Kommunikation zur Verfügung. Das 2.4GHz-Band steht weltweit mit 16 Kanälen zur Verfügung. Die hohe Frequenz ist aber insofern ein Nachteil, als dass für die Übertragung mehr Energie verbraucht wird. Weiter gibt es noch 10 Kanäle im Frequenzband um 915MHz, die aber nur in Amerika nutzbar sind. In Europa und Asien steht daher noch ein Frequenzband mit einem Kanal bei 868MHz zur Verfügung.

Die Leistungsregelung ist eine weitere wichtige Funktion der Bitübertragungschicht. Sie kontrolliert laufend die Qualität ihrer Verbindungen um diese mit dem minimalen Energiebedarf gerade noch aufrecht zu erhalten.

MAC Schicht:

Die MAC Schicht unterscheidet zwischen zwei unterschiedlichen Sensorknoten, die Reduced Function Devives (RFD) und die Full Function Devices (FFD). Die RFD sind Endgeräte die mit Sensoren ausgestattet sind, aber ausschliesslich mit einem einzigen FFD kommunizieren können. Die FFDs dagegen sind die Knoten, welche das eigentliche Netzwerk bilden und mit mehreren RFDs sowie FFDs kommunizieren. Im 802.15.4 Standard werden zwei Netzwerktopologien betrachtet. Die Sterntopologie und die Peer-to-Peer Topologie. Im ersten Fall wird ein Master-Slave Netzwerkmodell verwendet. Der Master

ist in diesem Fall der PAN-Koordinator und trägt die Verantwortung für Adresszuweisung und -verwaltung. Die Peer-to-Peer Topologie ermöglicht es allen FFDs mit anderen FFDs innerhalb ihrer Funkreichweite zu kommunizieren und über diese, Nachrichten für weiter entfernte FFDs zu verschicken. Auch in dieser Topologie wird ein Knoten die Rolle des PAN-Koordinator übernehmen. Der PAN-Koordinator kann das Netzwerk mit oder ohne Superframe verwalten. Wie auf Abbildung 4.6 zu sehen beginnt ein Superframe mit einem aktiven Zeitfenster in dem die Kommunikation stattfindet. Darauf folgt eine inaktive Phase in dem die Funkeinheit abgeschaltet wird um Energie zu sparen. Das aktive Zeitfenster wir aufgeteilt in eine Contention Access Period (CAP) und eine Contention Free Period (CFP). Während der CAP findet die Kommunikation wettbewerbsbasiert statt, in der CFP können dagegen einzelne Knoten beim PAN-Koordinator garantierte Zeitslots (GTS) anfordern die, wenn sie vom PAN-Koordinator gewährt werden, alleine von einem Gerät genutzt werden können.

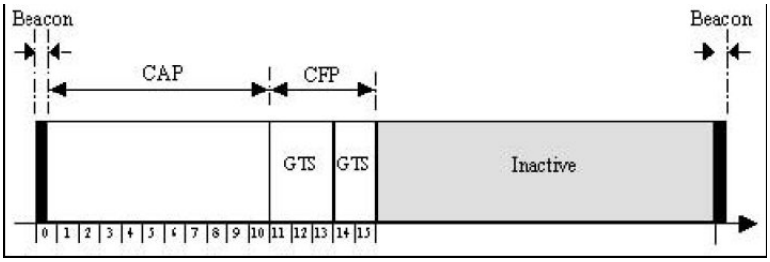


Abbildung 4.6: MAC Superframe [3]

Die MAC Schicht bietet verschiedene Möglichkeiten für die Sicherheit an. So kann das Netzwerk ohne Sicherheitsvorkehrungen betrieben werden oder es werden Zugriffsbeschränkungen anhand von Listen der akzeptierten Geräteadressen, sogenannten access control lists (ACL) gemacht. Der Standard unterstützt aber auch eine symmetrische Verschlüsselung der Daten durch den AES-Algorithmus mit einer Schlüssellänge von 128 Bits.

ZigBee:

Die ZigBee Allianz [37] stellt ein Zusammenschluss von verschiedenen Unternehmen dar die sich zum Ziel gesetzt haben Produkte zu entwickeln die zuverlässige, kosteneffektive, kabellos vernetzte Systeme ermöglichen die sich auf globale offene Standards stützen. Unter anderen beteiligen sich Philips, Motorola und Samsung in dieser Allianz. Der gemeinsam erarbeitete Standard baut auf dem IEEE 802.15.4 Standard auf und definiert die Vermittlungschicht, die sich um das Routing in einer Multihopnetzwerktopologie kümmert und die Anwendungsschicht, die ein Rahmen für die Entwicklung von Anwendungen bildet und die Kommunikation zwischen den einzelnen Anwendungen koordiniert.

Netzwerkschicht:

ZigBee unterscheidet zwischen 3 verschiedenen Gerätetypen. Die Endgeräte, die den RFD aus dem IEEE Standard entsprechen. Es können aber auch FFDs als Endgeräte eingesetzt werden. Ein ZigBee-Router ist ein FFD mit Routingfähigkeiten und ein ZigBee-Koordinator entspricht weitgehend dem PAN-Koordinator, welcher das Netzwerk administriert. Allerdings werden im ZigBee Standard nicht nur die Stern- und die Peer-to-Peer

Topologie unterstützt sondern auch komplexere Topologien wie Baum- oder Meshnetzwerke. Die Netzwerkschicht bietet unter anderem Funktionalitäten wie das Multihop Routing, die Identifizierung von Routingpfaden innerhalb des Netzwerkes und die Koordination von ins Netzwerk ein- oder austretenden Knoten.

Anwendungsschicht:

ZigBee Anwendungen bestehen klassischerweise aus mehreren Anwendungsobjekten die über verschiedene Knoten verteilt sind. Die einzelnen Objekte werden über Nummern identifiziert die zusammen mit den Adressen der Geräten eine eindeutige Identifikation ermöglichen. Ein spezielles Anwendungsobjekt ist das ZigBee Device Object (ZDO), dieses muss in jedem Sensorknoten implementiert sein und ist verantwortlich für das Auffinden von anderen Knoten und den Messwerten oder Services die diese bereitstellen. Der Aplication Sub Layer (ASL) stellt Datenübertragungsservices für die Anwendungsobjekte zur Verfügung und bildet so die Schnittstelle zu den unteren Schichten des OSI/ISO Modells.

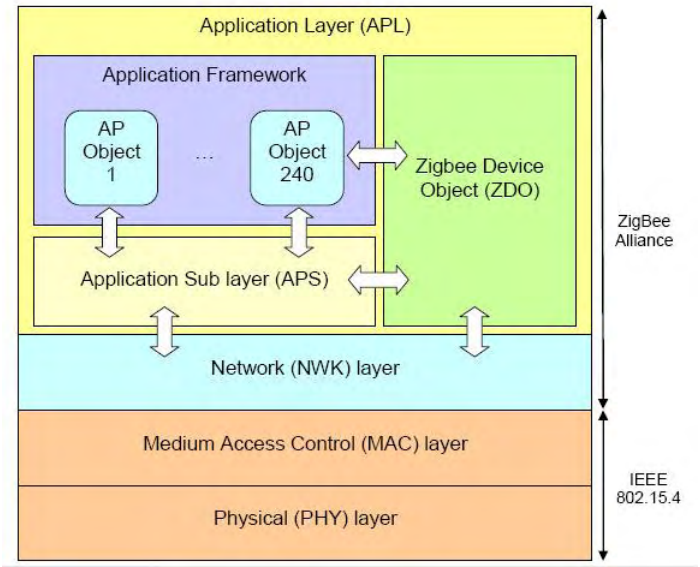


Abbildung 4.7: Modell der IEEE 802.14.5 und Zigbee Standards [3]

4.5 Anwendungen

Wie schon erwähnt wird die Technologie von WSN heute in einem breiten Feld von Anwendungsgebieten eingesetzt die von der Beobachtung von Tieren in freier Wildbahn über die Messung von seismologischen Schwingungen an einem aktiven Vulkan bis hin zu kommerziell einsetzbaren System für die Überwachung des Eigenheims oder Geschäftsräumen reicht. Wir werden hier nun unterscheiden zwischen Forschungsprojekten die diese Technologie einsetzen und kommerziellen Systemen die auf dem Markt bereits vertrieben werden um der Bandbreite der verschieden WSN annähernd gerecht zu werden.

4.5.1 Projekte

Die Forschungsprojekte, welche sich im Bereich der WSN ansiedeln sind so zahlreich, dass es unmöglich ist, diese umfassend zu beschreiben. Schon die Breite der Themen dieser Projekte ist enorm. So gibt es Forscher, welche sich mit konkreten Anwendungen von WSN beschäftigen, sei das der Einsatz der Technologie um aktive Vulkane zu überwachen [33] oder um die Ausbeute eines Weingutes zu optimieren [32]. Ein weiteres Forschungsgebiet ist auch die medizinische Überwachung [6]. Aber es gibt auch im Hardware- im Architektur- und im theoretischen Bereich Forschunsprojekte.

4.5.2 Kommerzielle Systeme

Systeme auf der Basis von WSN werden bisher vor allem in den Bereichen Überwachung und Sicherheit in und um Gebäude kommerziell vertrieben. Dieses Einsatzgebiet bietet den Systemen ein Umfeld, das verhältnismässig geschützt ist und absehbare Eigenschaften aufweist. So sind Sensorknoten innerhalb eines Gebäude nicht der Witterung ausgesetzt, können gezielt platziert werden, damit die Kommunikation gut funktioniert, die Temperaturen bewegen sich im Normalfall innerhalb eines definierbaren Rahmens, die Geräte sind für Reparaturen oder Batteriewechsel einfach zu erreichen und auch die Grösse der Geräte von einigen Quadratzentimetern ist völlig ausreichend um den Anforderungen der Kunden gerecht zu werden.

Sicherheit zu Hause

Ein Anwendungsgebiet für WSN ist die Sicherheit zu Hause. Verschiedene Anbieter haben bereits Systeme mit einer WSN Technologie auf dem Markt welche die Sicherheit im eigenen Heim verbessern sollen. In diese Kategorie fallen Bewegungsmelder, die zum Teil sogar so entwickelt wurden, dass sie nicht auf Haustiere reagieren. Daneben gibt es Rauch- oder Hitzesensoren um mögliche Brände zu erkennen, Flutwarnsysteme die bei Überschwemmungen Alarm auslösen oder Kontaktsensoren die registrieren ob eine Türe offen oder geschlossen ist. Spezielle Geräte die man auf dem Körper trägt können mit Panicbuttons ausgestattet älteren Menschen in Notsituationen helfen Hilfe zu holen oder tun dies, mit Sturzsensoren bestückt, in gewissen Fällen sogar automatisch.

Die Lusora Inc. [22] hat sich speziell auf das Gebiet der alleinlebenden Senioren spezialisiert und bietet ein individuell konfigurierbares System mit verschiedenen Sensorbgeräten und einem zentralen Hub an. Ihr System, LISA, beinhaltet einen Personal Pendant, der mit einem Panic Button und einem Sturzdetektor ausgestattet ist. Weiter gibt es sogenannte Heimsensoren, die an Wänden und Türen platziert, „unnormales“ Verhalten entdecken sollen. Leider werden diese Sensoren auf der Homepage der Lusora Inc. nicht genauer beschrieben. Allfällige Alarme oder Warnungen werden über den zentralen Hub, der in die Telefonbuchse eingesteckt wird, an Familienangehörige oder das Care Team der Lusora Inc. übermittelt. Diese können dann mittels Telefon Rücksprache nehmen oder unter Umständen den Notruf alarmieren.

4.6 Vergleich und Schlussfolgerungen

Am Anfang der Wireless Sensor Networks Forschung stand die Vision von „Smart Dust“, mikroskopisch kleiner Sensorknoten die in hoher Anzahl breit gestreut verschiedenste Arten von Anwendungen finden würden. Die Realität sieht doch etwas mehr ernüchternd aus. Das gleichnamige Smart Dust Projekt der University of California Berkeley [20] hat bereits vor mehreren Jahren versucht die Grenzen der Miniaturisierung auszuloten. Ziel war ein Sensorknoten in der Grösse eines Kubikmilimeters, welchem das Ergebnis schon recht nahe kam.

Dennoch sind die heute mehrheitlich verwendeten Sensorknoten, sowohl in Forschungsprojekten als auch im kommerziellen Umfeld, noch markant grösser. Ein kommerzieller Ableger des Smart Dust Projektes, Dust Networks [7], vertreibt beispielsweise Knoten in der Grössenordnung mehrerer Kubikzentimeter, was auch die Untergrenze für eine Vielzahl der Forschungsprototypen darstellt. Für die meisten heutigen Anwendungen scheint eine weitere Miniaturisierung auch gar nicht notwendig. Der Haupttreiber für weitere Forschung in der Richtung dürfte in erster Linie das Militär sein, welches unauffällig kleine Sensorknoten möchte.

Während kein Mangel an Ideenreichtum für die möglichen Anwendungen von WSNs herrscht, hat sich die Technologie bis jetzt doch nicht grossflächig etabliert. Sie wird in einigen Nischenanwendungen bereits kommerziell genutzt, aber den grössten Anteil machen nach wie vor Forschungsprojekte aus. Von der Vision einer Alltagstechnologie sind wir noch einen weiten Weg entfernt.

Ein Hauptproblem dabei mag sicherlich auch fehlende Standardisierung von Hardware sein, aber durch die breit gefächerten Anforderungen und Anwendungen erscheint es fragwürdig ob ein oder mehrere gemeinsamee Standards überhaupt sinnvoll wären. Im Bereich der Software sieht die Situation schon ein wenig besser aus, mit TinyOS hat sich hier immerhin schon ein de facto Standard etabliert der vielfältig eingesetzt wird.

Abschliessend lässt sich sagen, dass die Forschung zwar schon an vielen Aspekten erfolgreich gearbeitet hat, aber nach wie vor ungelöste Probleme verbleiben. Insbesondere der kommerzielle Einsatz muss noch weiter voran getrieben werden, bevor man von Wireless Sensor Networks wirklich als Alltagstechnologie betrachten kann.

Literaturverzeichnis

[1] Akyildiz, I. F., W. Su, Y. Sankarasubramaniam und E. Cayirci: *Wireless sensor networks: a survey*. Comput. Networks, 38(4):393–422, 2002.

[2] Bharathidasan, Archana und Vijay Anand Sai Ponduru: *Sensor Networks: An Overview*. Technischer Bericht, Department of Computer Science, University of California, Davis, CA 95616, 2002.

[3] Baronti, Paolo, Prashant Pillai, Vince Chook, Stefano Chessa, Alberto Gotta und Y. Fun Hu: *Wireless Sensor Networks: a Survey on the State of the Art and the 802.15.4 and ZigBee Standards*. Technischer Bericht, University of Bradford, United Kingdom, Wireless Networks Laboratory, Istituto di Scienza e Tecnologie dell'Informazione, Pisa, Italy and Department of Computer Science, University of Pisa, Pisa, Italy, 05 2002.

[4] Culler, David, Deborah Estrin und Mani Srivastava: *Guest Editors' Introduction: Overview of Sensor Networks*. IEEE Computer, 37(8):41–49, 2004.

[5] Culler, David E., Jason Hill, Philip Buonadonna, Robert Szewczyk und Alec Woo: *A Network-Centric Approach to Embedded Software for Tiny Devices*. In: *EMSOFT '01: Proceedings of the First International Workshop on Embedded Software*, Seiten 114–130, London, UK, 2001. Springer-Verlag.

[6] *Wireless Sensor Networks for Medical Care*. `http://www.eecs.harvard.edu/~mdw/proj/codeblue/` (letzter Abruf im Januar 2007).

[7] Dust Networks, Inc.: *Dust Networks: Embedded Wireless Sensor Networking for Monitoring and Control*. `http://www.dust-inc.com/` (letzter Abruf im November 2006).

[8] ETH Zürich: *BTnode rev3 - Product Brief*, 2005.

[9] *FM Electronics Ltd*. `http://www.fmelectronics.co.uk/` (letzter Abruf im Januar 2007).

[10] Gauger, Matthias: *Komponenten in TinyOS*, 2005.

[11] Gummadi, Ramakrishna, Omprakash Gnawali und Ramesh Govindan: *Macro-programming Wireless Sensor Networks using Kairos*. In: *Proceedings of the International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2005.

[12] Group, The Cornell Databse: *The Cougar Database Project*. `http://www.cs.cornell.edu/database/cougar/` (letzter Abruf im Januar 2007).

[13] Heinzelman, Wendi Rabiner, Anantha Chandrakasan und Hari Balakrishnan: *Energy-Efficient Communication Protocol for Wireless Microsensor Networks*. In: *HICSS '00: Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 8*, Seite 8020, Washington, DC, USA, 2000. IEEE Computer Society.

[14] He, Tian, Sudha Krishnamurthy, John A. Stankovic, Tarek Abdelzaher, Liqian Luo, Radu Stoleru, Ting Yan, Lin Gu, Gang Zhou, Jonathan Hui und Bruce Krogh: *VigilNet: An Integrated Sensor Network System for Energy-Efficient Surveillance*. ACM Transactions on Sensor Networks, 2004.

[15] Hadim, Salem und Nader Mohamed: *Middleware Challenges and Approaches for Wireless Sensor Networks*. IEEE Distributed Systems Online, 7(3), 2006.

[16] Heinzelman, W., A. Murphy, H. Carvalho und M. Perillo: *Middleware to Support Sensor Network Applications*. IEEE Network Magazine Special Issue, 2004.

[17] Intanagonwiwat, Chalermek, Ramesh Govindan, Deborah Estrin, John Heidemann und Fabio Silva: *Directed diffusion for wireless sensor networking*. IEEE/ACM Trans. Netw., 11(1):2–16, 2003.

[18] Kalpakis, Konstantinos, Koustuv Dasgupta und Parag Namjoshi: *Efficient algorithms for maximum lifetime data gathering and aggregation in wireless sensor networks*. Comput. Networks, 42(6):697–716, 2003.

[19] Kulik, Joanna, Wendi Heinzelman und Hari Balakrishnan: *Negotiation-based protocols for disseminating information in wireless sensor networks*. Wirel. Netw., 8(2/3):169–185, 2002.

[20] Kahn, J. M., R. H. Katz und K. S. J. Pister: *Next century challenges: mobile networking for Smart Dust*. In: *MobiCom '99: Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*, Seiten 271–278, New York, NY, USA, 1999. ACM Press.

[21] Levis, Philip und David Culler: *Maté: A Tiny Virtual Machine for Sensor Networks*. In: *Proceedings of the 10th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS X)*, 2002.

[22] *Lusora Healthcare Systems Inc.* `http://www.lusora.com/` (letzter Abruf im Januar 2007).

[23] *nesC: A Programming Language for Deeply Networked Systems*. `http://nescc.sourceforge.net` (letzter Abruf im Januar 2007).

[24] Nachman, L., R. Kling, R. Adler, J. Huang und V. Hummel: *The Intel mote platform: A Bluetooth-based sensor network for industrial monitoring,*. In: *Proc. 4th Int'l Conf. Information Processing in Sensor Networks (IPSN '05)*, 2005.

[25] Römer, Kay und Friedemann Mattern: *The Design Space of Wireless Sensor Networks*. IEEE Wireless Communications, 11(6):54–61, 2004.

[26] Raghunathan, V., C. Schurgers, Park. S und M. B. Srivastava: *Energy-aware wireless microsensor networks*. IEEE Signal Processing Magazine, 19(2):40–50, 2002.

[27] Schulze, Astrid: *Seminararbeit: The ZebraNet Wild Life Tracker*, 2006. `http://www.ibr.cs.tu-bs.de/courses/ws0506/ssvm/papers/Schulze-ZebraNet.pdf` (letzter Abruf im November 2006).

[28] Souto, Eduardo, Germano Guimaraes, Glauco Vasconcelos, Mardoqueu Vieira, Nelson Rosa, Carlos Ferraz und Judith Kelner: *Mires: a publish/subscribe middleware for sensor networks*. Personal and Ubiquitous Computing, 10, 2005.

[29] Srisathapornphat, Chavalit, Chaiporn Jaikaeo und Chien-Chung Shen: *Sensor Information Networking Architecture*. In: *ICPP '00: Proceedings of the 2000 International Workshop on Parallel Processing*, Seite 23, Washington, DC, USA, 2000. IEEE Computer Society.

[30] *Taskgroup IEEE 802.15.4*. `http://www.ieee802.org/15/pub/TG4.html` (letzter Abruf im Februar 2007).

[31] `http://webs.cs.berkeley.edu/tos/` (letzter Abruf im November 2006).

[32] *Camalie Net Wireless Sensor Network at Camalie Vineyards*. `http://camalie.com/WirelessSensing/WirelessSensors.htm` (letzter Abruf im Januar 2007).

[33] *Monitoring Volcanic Eruptions with a Wireless Sensor Network*. `http://www.eecs.harvard.edu/~mdw/proj/volcano/` (letzter Abruf im Januar 2007).

[34] Warneke, B., M. Last, B. Liebowitz und K. S. J. Pister: *Smart Dust: communicating with a cubic-millimeter computer*. IEEE Computer, 34(1):44–51, 2001.

[35] Wang, Lan und Yang Xiao: *A survey of energy-efficient scheduling mechanisms in sensor networks*. Mob. Netw. Appl., 11(5):723–740, 2006.

[36] *YaleLocks*. `http://www.yalelock.com` (letzter Abruf im Januar 2007).

[37] *Zigbee Alliance*. `http://www.zigbee.org` (letzter Abruf im Februar 2007).

Chapter 5

Delay Tolerant Networks - Challenges and Solutions

Daniel Heuberger, Ronny Kallupurackal, Marcel Lanz

Abstract Messaging in a challenged environment e.g. inter-planetary communication differs from the traditional Internet communication in that way, that it must cope with long delays that are caused by the long distances of the communication partners. This seminar paper highlights the DTNRG's network solution, Delay Tolerant Network (DTN). DTNRG's aim is to design and implement architectures and protocols for networks, where no end-to-end connectivity can be assumed and which differ in the characteristics of the known Internet. After giving an overview of DTNRG's DTN the focus will be on Delay Tolerant Mobile Networks (DTMN). The difference to DTN is that all nodes in the network are now mobile. At last three other different concepts called Plutarch Triad and Metanet, that allows communication through and between heterogeneous networks, which do not use the TCP/IP protocol stack, are discussed.

Contents

5.1 Introduction

The goal of Delay Tolerant Networking (DTN) is to solve the issues, such as high delays, high error rates and extremely asymmetric up- and downloads, arising in mobile or extreme environments using the traditional TCP/IP approach. With its help it is possible to ensure the delivery of messages although there is not always continuous network connectivity. Using store-and-forward, messages, also called bundles, are sent form node to node, even if the nodes are not in the same, homogeneous network. The DTN architecture is an overlay architecture over the existing Networking Architecture.

In the following paper the terms ”homogeneous” and ”heterogeneous” networks are often used. For there is no misunderstanding in the meaning of these terms, a definition for their use in this context is given now [2]:

Figure 5.1 shows three different kinds of networks:

- Internet with IPv4
- GPRS
- sensor network

These three networks are heterogeneous to each other and homogeneous in themselves. You might object that the Internet is not homogeneous at all. ATM uses, for example, not the same addressing schema on the data link layer as ethernet does. But the IP layer makes them homogeneous. Recapitulating, a network is homogeneous in itself when there is a addressing of all members in the network possible without any manipulation or translation of the packets sent and when the package formats, transport protocols and naming services are compatible.

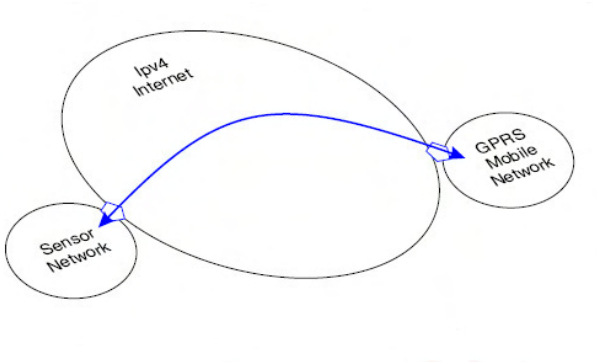


Figure 5.1: Homogeneous vs Heterogeneous Networks

Section 5.2 highlights the details of DTN’s: what it is exactly for, how it solves the problems that arise in mobile or extreme environments. Technical details can also be found in this chapter. Further it points scenarios out in which DTN is used. Section 5.3 is all about Delay Tolerant Mobile Network (DTMN). DTMN is a DTN with special attention to the mobile aspect of communication. In section 5.4, some specific projects that are related to DTN are discussed. Finally, the conclusion in chapter 5.4.

5.2 Delay Tolerant Network (DTN)

In challenging communication environments like inter-planetary communication it is possible that long delays occur because of the long disconnections and long distances between the communication partners. A symmetric communication or a continuous communication path between communication partners can not be expected in these networks. Consequently, delay tolerant applications require delay tolerant networks to be run successfully.

5.2.1 DTNRG

There is a research group called Delay Tolerant Network Research Group (DTNRG) [1] and its aim is to design and implement architectures and protocols for networks, where no end-to-end connectivity can be assumed and which differ in the characteristics of the known Internet. This section focuses on DTNRG's concept of a possible Delay Tolerant Network (DTN). First of all the need to design and implement Delay Tolerant Networks shall be explained.

5.2.2 Why do we need a DTN?

In a world of vast amounts of existing networks and networks either in planning or in the realisation phase, there is an increasingly need to enable communication between those heterogeneous networks. These heterogeneous networks are not using the IP protocol and are therefore incompatible with other heterogeneous networks. Each heterogeneous network is good in transmitting messages within the network, but it is quite hard to transmit a message from one heterogeneous network to another. People wrongfully assume that it is easy to technically implement it. Reasons indicating the difficulties of communication with and through heterogeneous networks are:

- the networks are incompatible
- there is no end-to-end connectivity
- long delays or variable delays are usual
- there exist asymmetric data rates that will not let conversational protocols work properly

Challenged by these problems the DTNRG has proposed a new high-level network (DTN) that supports the communication between different networks with different characteristics.

5.2.3 Example

Nowadays there are a lot of networks that differ from the traditional Internet. In the literature they are called challenged networks (e.g. the ad-hoc network in armed forces). A deeper characteristic of challenged network is given in Section 5.2.4.

Troops are connected with other troops and the strategic head quarter gives them advice how to act in the hostile environment. Of course it is not possible that they are always online or reachable, because the enemies are listening and observing all communication. That is why such networks must be delay tolerant. Delay tolerant can be defined as following: a message that is sent on date A has to be stored in the network till the receiver gets online (on date B) and is able to receive the message. This example is extended with another sensor network on the moon. Somehow the armed forces on the battlefield are interested in the sensor data on the moon. It should be noted that both networks are challenged, because they are not using the IP protocol. Therefore it is not possible to interchange messages from one network to the other, without a change in the network topology. Since the sensor network on the moon is restricted in resources there is not always a connection to the battlefield and because of the long distance between the communication partners delays in getting send messages are inevitable. What a DTN now does is to provide translational services between these incompatible networks and to support delay tolerant applications.

5.2.4 Characteristics of Challenged Networks

What a DTN does is to build a homogeneous network out of heterogeneous networks. As explained before heterogeneous networks are called challenged networks, because of not using the IP protocol. This section gives a short overview about the characteristics of such networks [5].

High Latency and Low Data Rates

Usually a DTN has got a high latency and a low data rate. Inter-planetary communication can be used as a topic to visualise this fact. For instance there shall be communication between a person on the Earth with someone on the planet Mars. It is clear that there is a long distance between the Earth and Mars and that it takes a long time till the message will arrive at the person on the Mars. Because of the long distance between the two planets there must be a DTN gateway or router that can store the message and resend it if it could not be delivered. Of course one can not expect that the bandwidth of the connection is at 10Mbps.

Disconnections

There are phases of long periods of disconnections. In the inter-planetary communication the router on the Earth has to wait till there is a line of sight with the router on the Mars.

In the military network - as mentioned above - there will be several disconnections due to security reasons (the soldiers do not want to be unmasked by their enemies). There could be also other reasons why a node is not connected to the network (e.g. saving power). In that connection the addresser gets only online if he has got something to send.

Persistent storage

Due to the fact that messages can not be delivered instantly they have to be stored in the network. Usually DTN routers have got persistent storage devices where they have to store the messages till they can be forwarded.

Interoperability

If a DTN should support communication across multiple networks the network must offer translation services, because not every network uses the IP protocol. That's why in heterogeneous networks messages have got different structures and why it is not possible to exchange messages from a heterogeneous network to another (problems based on differing formats).

Security

If messages sent over a DTN have to be encrypted challenges or keys have to be exchanged between the sender and the receiver. This approach is not so attractive, because of the intermittence of the network. Another drawback is that the bandwidth is already low and by adding authentication and access control information to messages the throughput will decrease further. These drawbacks do not mean that is not possible to send encrypted messages in a DTN. The solution of the problem is that messages can be signed by a forwarding node and the receiver can check the nodes identity by consulting the Certificate Authority. After confirming the identity of the forwarding node the node exchanges the signature of the message by his own and forwards it to the next node in direction to the receiver.

Low Duty Cycle Operation

Devices in a DTN can have limited resources like small battery packs, low cpu power etc. To save energy they go offline and only go online if they have to collect data or transmit data. This approach implies that the receiver of the data has to know when the device goes online, so that it can collect the data.

5.2.5 Difference to the Internet

A quick overview of the characteristics of the Internet shall be provided before analysing the differences between the traditional Internet and a DTN. The Internet consists of networks and subnetworks that run the TCP/IP protocol stack. These protocols are used for routing and reliable message exchange. Apart of the satellite and wireless communication few of the connections are made by wired links using the telephone network. This results in a continuously connected end-to-end path between the source and the destination. Since the end-to-end connection is wired delays and error rates are decreased and a symmetric bidirectional communication can be assumed. Apart of the internet there are heterogeneous networks. These heterogeneous networks have special characteristics and differ from the traditional internet (for characteristics of heterogeneous network refer to Section 5.2.4). Communication in these networks is specialised for these networks (a particular region) so that communication between two heterogeneous networks is hard to manage. In contrast to the Internet the wireless heterogeneous networks have to cope with high error rates, long delays and asymmetric bidirectional communication [11]. The military wireless network described earlier is a heterogeneous wireless network that has the mentioned properties. In [11]the DTN is described as follows:"A delay-tolerant network (DTN) is a network of regional networks. It is an overlay on top of regional networks, including the Internet." What a DTN does is to support communication between heterogeneous networks, by translating messages into the appropriate format of the network and support long delays.

5.2.6 Structure of a DTN

This section describes the structure and main processes in a DTN and is based on [11].

Bundle Layer

In a DTN a new protocol layer, a so-called bundle layer is put on the protocol stack of the different heterogeneous networks. This allows applications to communicate over different heterogeneous networks. The main task of the bundle layer is to store and forward bundles, also called messages, that are send in a DTN. The bundle layer is used in every heterogeneous network and is built on the region specific layers, which are responsible for communication and transport of messages inside the heterogeneous network.

On the top of Figure 5.2 the protocol layers in a DTN are illustrated. On the top of the region specific layers is the bundle layer that holds together the underlieing layers. On the bottom of the figure the traditional Internet layers are compared to the DTN layers. The bundle layer is situated in a DTN protocol stack between the application layer and the transport layer. In contrast to the Internet layers the transport, network, link and physical layers of a DTN can vary according different DTN regions.

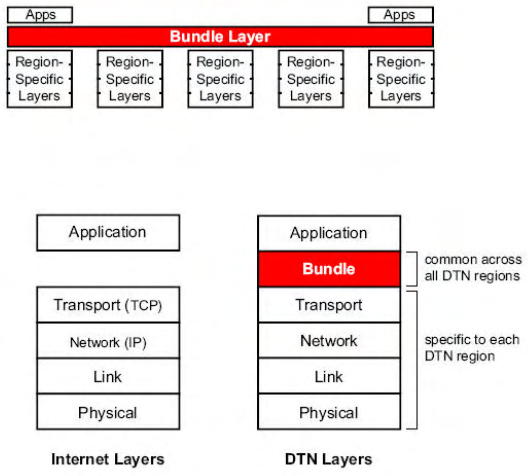


Figure 5.2: The bundle layer on the protocol stack [11]

Bundle and Bundle Encapsulation

According to [11] the bundle, the message send through a DTN, is divided into three parts:

- User data
- Control information
- Header

In the user data part the data that the source application produces is packed in. It is followed by control information that the source application is provided to the receiving application and in which it is described how to handle and to process the user data. Like every protocol layer the bundle adds his header to the bundle when it is forwarded to the lower layers of the protocol stack.

Figure 5.3 shows how the application data is encapsulated and fragmented while forwarded in the protocol stack.

Non-conversational protocol

TCP is a conversational protocol, several messages are exchanged between the sender and the receiver before a connection is established. Using TCP to establish a connection in a DTN is impractical, due to the characteristics of the network. If there is limited connectivity the exchange of messages will take long time and the possibility of failure is high. That is why conversational communication between bundle layers is kept at a minimum and acknowledgement of messages are optional.

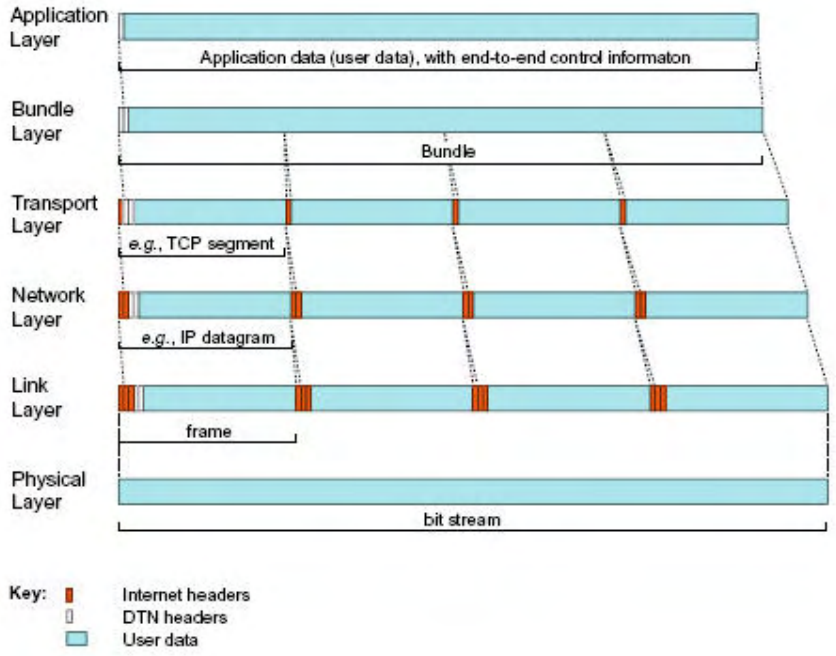


Figure 5.3: How bundle encapsulation works [11]

Figure 5.4 shows two node's protocol stack. The bundle layers communicate with each other with non-conversational protocols that mean that for connection establishment there is no need to exchange several messages that are acknowledged. Lower layer protocols can of course be conversational like TCP but due to intermitted connectivity and long delays it is beneficial to implement non-conversational or minimal-conversational protocols.

DTN nodes

According to [11] in a DTN nodes have got the bundle layer on the top of their protocol stack. Nodes can act as hosts, routers and gateways. A host sends and receives bundles from others, but does not forward bundles. Like other nodes a persistent storage device is needed to store bundles. In contrast to hosts a router takes bundles from hosts and forwards them inside the heterogeneous network. The router too needs a persistent storage device to store bundles till links are available to send them forward to the destination. Gateways are placed between heterogeneous networks and their main duty is to forward bundles between different networks. One can say that they act as translators because they have to transform the bundles from one network into the appropriate format of the other.

Custody Transfer

In [11] it is described how retransmission of corrupt bundles or lost bundles between two neighbouring bundle layers can be achieved. The procedure is called custody transfer. If a node wants to make sure that the bundle is correctly received at the neighbouring

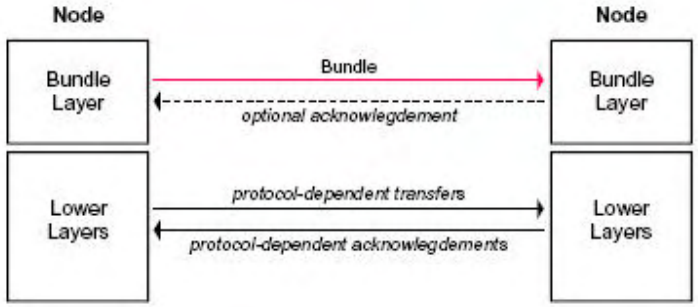


Figure 5.4: How non-conversational protocol works [11]

node it can request a custody transfer and send the bundle. If the receiving node does not accept custody and the time-to-acknowledge period expires, the sender will send the bundle again. The bundle is stored either at the sender till a neighbouring node accepts custody transfer or till the bundle's time to live expires.

Classes of Bundle Service

In [11] the authors distinguishe six classes of services that a bundle layer provides for the bundle. One of them is already described under the section custody transfer. Another service is called return receipt, where the destination node sends a message to the source that the bundle has been received. If the source node receives notifications that the nodes on the way to the destination have got accepted custody transfer, then this services is called custody transfer notification. The only difference between custody transfer notification and bundle-forwarding notification is that the source is notified whenever a bundle is forwarded to a node in the direction of the destination node without applying custody transfer. Then of course one can decide the priority of bundle delivery. There are three options: bulk, normal and expedited delivery. The last service that the bundle layer offers is the authentication of the sender's identity by digital signature.

Figure 5.5 shows four classes of services supported by the bundle layer. In the custody transfer every node that accepts custody transfer sends an acknowledgment message to the sending node (this is pictured through the dashed arrow towards the sending node). In the second service 'return receipt' the source node managed to send a bundle to his destination node, which is represented in the figure with an arrow from the source node to the destination. The destination node acknowledges the reception of the bundle (dashed arrow to the source). In custody transfer notification the source is notified about the acceptance of custody transfer of nodes that forward the bundle to the destination node. The custody transfer notification is represented by the dashed arrows showing to the source node. The last service pictured is the bundle-forwarding notification. Bundles are forwarded towards the destination node without custody transfer, but the source gets to know when the bundle has been forwarded from a node to another.

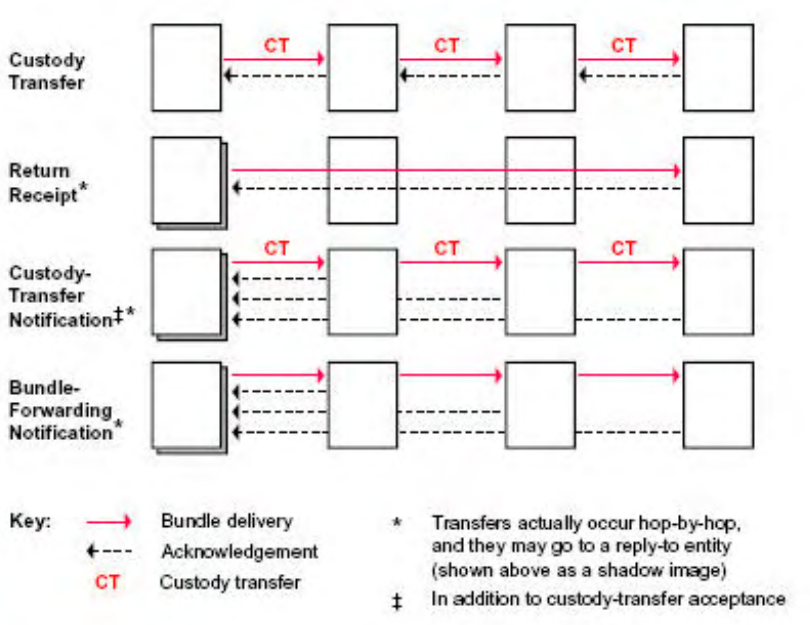


Figure 5.5: Bundle service classes [11]

DTN regions, names and addresses

In [11] the authors state that a DTN consists of several heterogeneous networks, where each network has its own region. To distinguish the networks it was introduced a region ID. These region IDs are known by the entire DTN and are part of every node's address. A DTN node's address consists of two identifiers. The region ID and the entity ID. The entity ID is used for routing inside a heterogeneous network, whereas the regional ID is used for routing between networks."An entity may be a host (a DTN node), an application instance, a protocol, an URL, a port (used to find the bundle service on a host) and potentially a token (used to find a particular application instance that is using the bundle service), or something else [11]."

Figure 5.6 shows how the region and entity identifiers are used for routing. The source is situated in the region 1 and the destination in the region 2. The gateway uses the region identifier to decide where to route the bundle. The gateway is the only node that has got two or more region IDs, because they belong to several heterogeneous networks. In the region 2 the router uses the entity ID to route the bundle to the destination.

Security

How to make sure the authenticity and integrity of a bundle is described in [11]. Most of the networks check, if a security method is used, the authenticity of the user and the integrity of the message, but they do not check the authenticity of forwarding routers and gateways. In a DTN forwarding routers and gateways are authenticated too.

Figure 5.7 shows how a bundle is secured delivered to the destination. If a host wants to send a bundle on a secure way it sends the bundle with its signature to the next node.

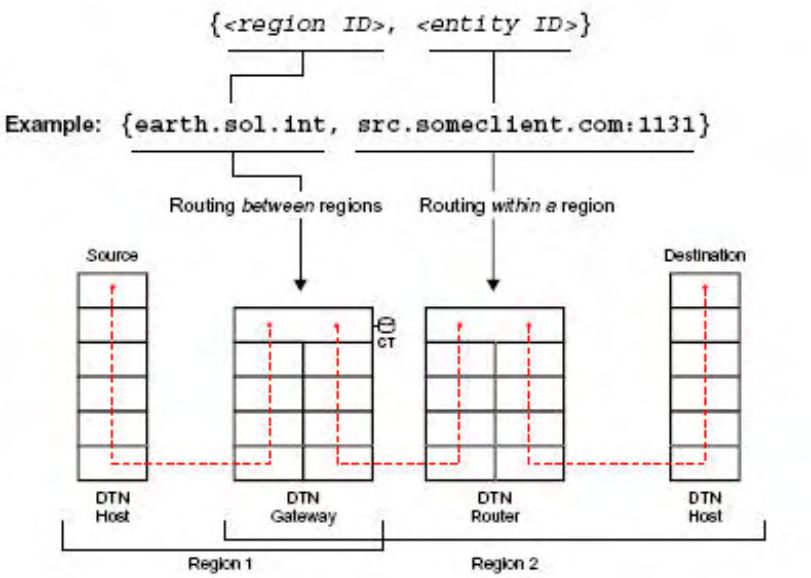


Figure 5.6: Region and entity ID and routing [11]

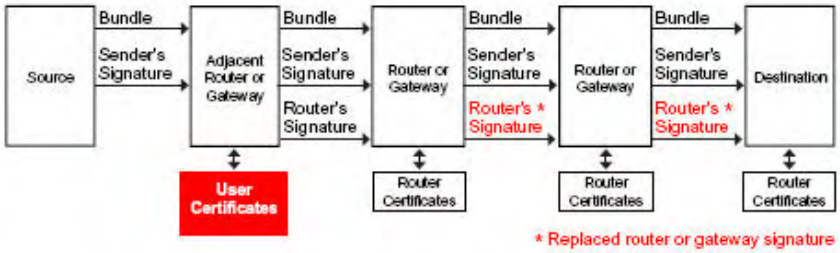


Figure 5.7: Secured transmission of a bundle [11]

The receiving node can check the integrity and the authenticity of the sender by applying the public key, which it gets from a Certificate Authority, on the digital signature. If with the bundle is something fishy, the bundle can be dropped. If everything is alright the forwarding node replaces the source's digital signature with its digital signature and forwards it to an adjacent node. This procedure is done till the bundle finally arrives the destination node.

5.3 Delay Tolerant Mobile Network (DTMN)

This section will present a short overview about Delay Tolerant Mobile Networks.

5.3.1 What is a DTMN

A Delay Tolerant Mobile Network (DTMN) is a DTN where all nodes in the network are mobile. Further one assumes that there is no end-to-end connection between any two

nodes.

Two key items characterize a DTMN:

- Node Blindness, meaning that "the nodes in the network do not know any information regarding the state, location or mobility patterns of other nodes" [9]
- Node Autonomy, "each node has independent control over itself and its movement." [9]

There are two ways of looking at a DTMN. On the one hand, it can be viewed as a special kind of a classical DTN, as a single DTN region with multiple nodes, or on the other hand, it can be considered as a DTN in itself, where each node is simultaneously a DTN region and a DTN gateway.

5.3.2 Operation of a DTMN

As described in [9], there are three types of nodes in a DTMN. A sender type, the node that starts the transmission, a forwarder type, the node that relays a bundle and the destination type, the last node that receives the bundle.

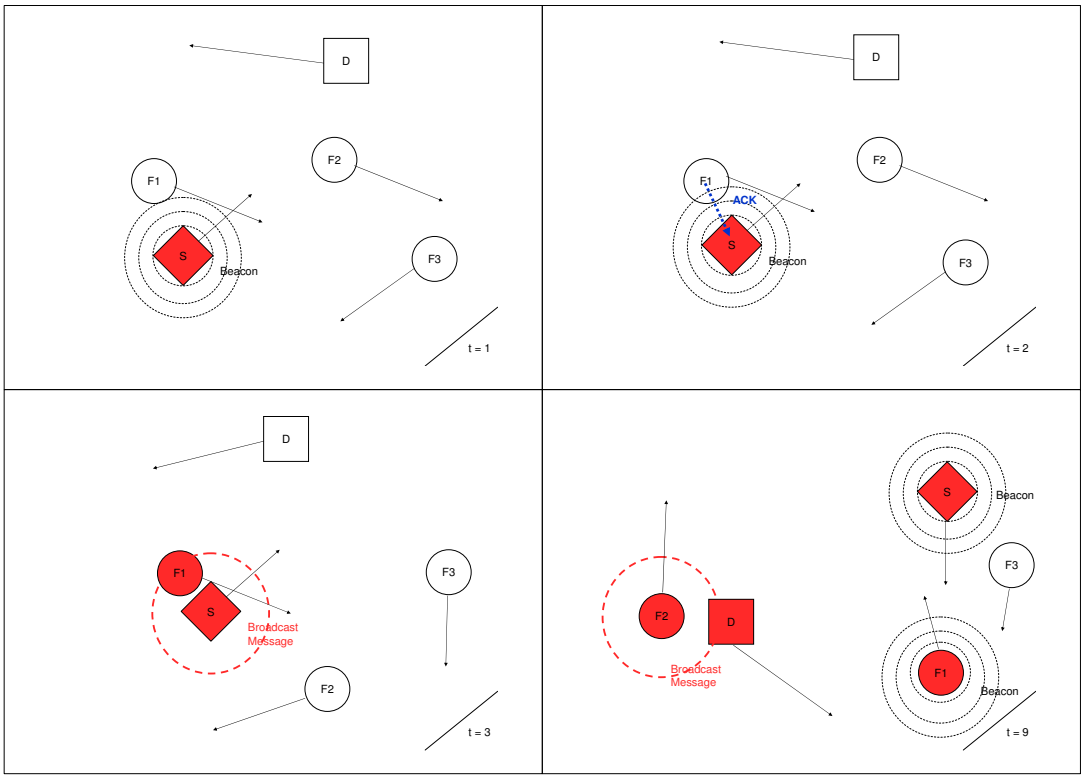


Figure 5.8: Some time shifted snapshots showing the operation of a DTMN

Figure 5.8 shows some time shifted snapshots of a sparse network with one sender S, multiple forwarders F1-F3 and one ultimate destination D. The arrows in the graphic show the direction of movement.

At point in time t=1, the sender S starts sending a beacon for neighbour discovery and announces that it has something to send. The beacon will be periodically retransmitted. As soon as one or more forwarders F or even the ultimate destination D receive the beacon, they send if they haven't already received the message an *ack* back to the sender S. This is illustrated in t=2. After the sender S has received the *ack* it starts at t=3 broadcasting the message. Having received the message, the forwarders F respectively the ultimate destination D start to send their own beacon to announce that they have something to send as well. With this procedure the message successively propagates through the network till it eventually reaches the ultimate destination D. In our example this can be observed at t=9. Note that after the destination is reached, the message will continue propagating till eventually each node in the network is infected. This approach results in an overuse of the network through the iterant flooding of messages, but it also has the advantage that a high delivery rate and a small delay can be achieved.

5.3.3 Modelling Node Willingness in a DTMN

To limit the amount of messages sent and to restric the overuse of the network a metric is used to define how hard a given node tries to infect other nodes. Such a metric is stated in [9] and is called node willingness which is there defined "as the degree at which a node actively engages in trying to re-transmit a message". Node willingness can be expressed with regard to three parameters:

- Beacon Interval means "the amount of time a sender or forwarder node waits before sending a new beacon"[9].
- Times-to-Send is "the number of times a node successfully forwards a message to other nodes in the network before it stops forwarding the message." [9]
- Retransmission Wait Time denotes "the amount of time a node waits without beaconing before it tries to resend the message to other nodes in the network" [9]

The sender includes the values of these three parameters in the message header, so the forwarder nodes can then adapt their node willingness suitably.

5.3.4 Controlled Flooding Models

In highly mobile networks there is often only a short time when nodes come within range of each other. Because of the very limited time nodes have to communicate simple and smart algorithms are needed. The following three schemes are based on such kind of algorithms.

Basic Probabilistic (BP)

In the previous section we assumed that all nodes have the same node willingness. Choosing a uniform distribution probabilistic function for modelling the node willingness will provide a much more realistic view. With this addition and based on the result of the probabilistic function a node can choose for example to forward a message at half of the sender's willingness or forward it at the same level of willingness as the sender or even choose to not forward a message at all.

Time-to-Live (TTL)

The Time-to-Live is usualy added on top of the basic probabilistic scheme. TTL here determines "how many times the message is forwarded before it is discarded"[9].

Kill Time

This scheme is also usualy added on top of the basic probabilistic scheme. A time stamp is added to each message describing "the time after which the message should no longer be forwarded"[9].

5.3.5 Four Reliability Approaches

K. Harras and K. Almeroth present in their paper about Transport Layer Issues in Delay Tolerant Mobile Networks [8] four different reliability approaches. This section will give a short overview about these four issues by means of an illustrative example.

Hop-by-hop Reliability

Figure 5.9 explains the most basic reliability approach named hop by hop reliability. To apply Occam's razor, the beacon signals and the acknowledgement that a beacon was received are not displayed in the figure.

At point in time t=2 the forwarder node F3 comes into the range of S and receives the message M as a result of the sender's broadcast. If the message is delivered successfully an acknowledgement is sent back to the deliverer. According through their node willingness level the sender and forwarder nodes infect as many nodes as they can and if there is enough time and mobility the ultimate destination D will eventually receive the message. This state is depicted at t=3. Hop-by-hop acknowledgement can not provide end-to-end reliability but the advantage of this approach is that it minimizes the time that the message M remains in the buffer. Hop-by-hop reliability can be viewed as a kind of fundamental approach over which the other approaches will be built.

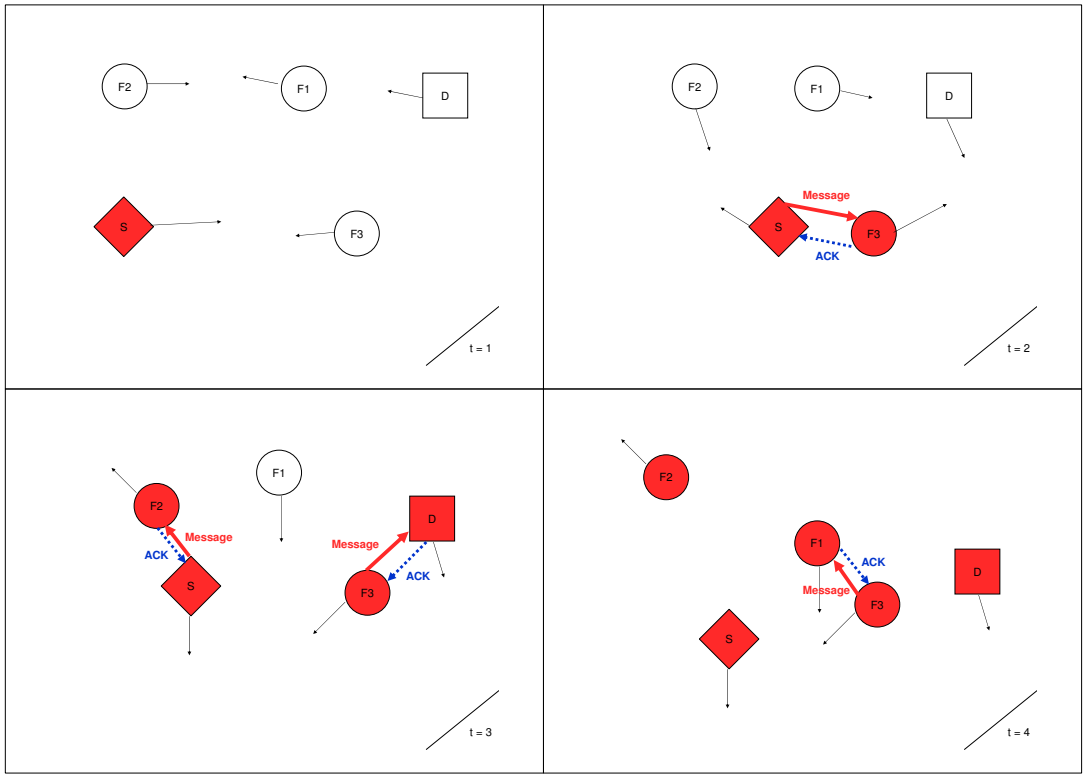


Figure 5.9: A scenario visualizing the hop-by-bop reliability approach

Active Receipt Reliability

In some situations hop-by-hop reliability is not sufficient and a more reliable approach is needed. The active receipt reliability provides an end-to-end acknowledgment as after a reception at the ultimate node D a receipt is dispatched actively. Actively means that the receipt is treated like a message that has to be sent back to the sender S.

In Figure 5.10 we see the procedure of an active receipt. T=1 shows the situation where D has just received the message and is sending an active receipt. The receipt is forwarded till it reaches the sender S as demonstrated at t=3. The task of the active receipt is to "heal" the network by means of stopping the infected nodes sending their messages. The active receipt R is kept according to the node's level to prevent re-infection. The drawback of this approach is that the active receipt now is spread periodically until a timeout or the TTL is reached. However it reduces the cost of storing and transmitting due to the smaller size of the receipt compared to the message.

Passive Receipt Reliability

As seen in the previous section the cost of active receipt reliability can still be very high in some cases because active receipt reliability reaches a state where two messages are simultaneously infecting the nodes in the network. The passive receipt approach eliminates this insufficiency. Passive here means reactive, so a node only sends his passive receipt when an infected node trying to pass his message to him is into reach. Figure 5.11 visualizes the

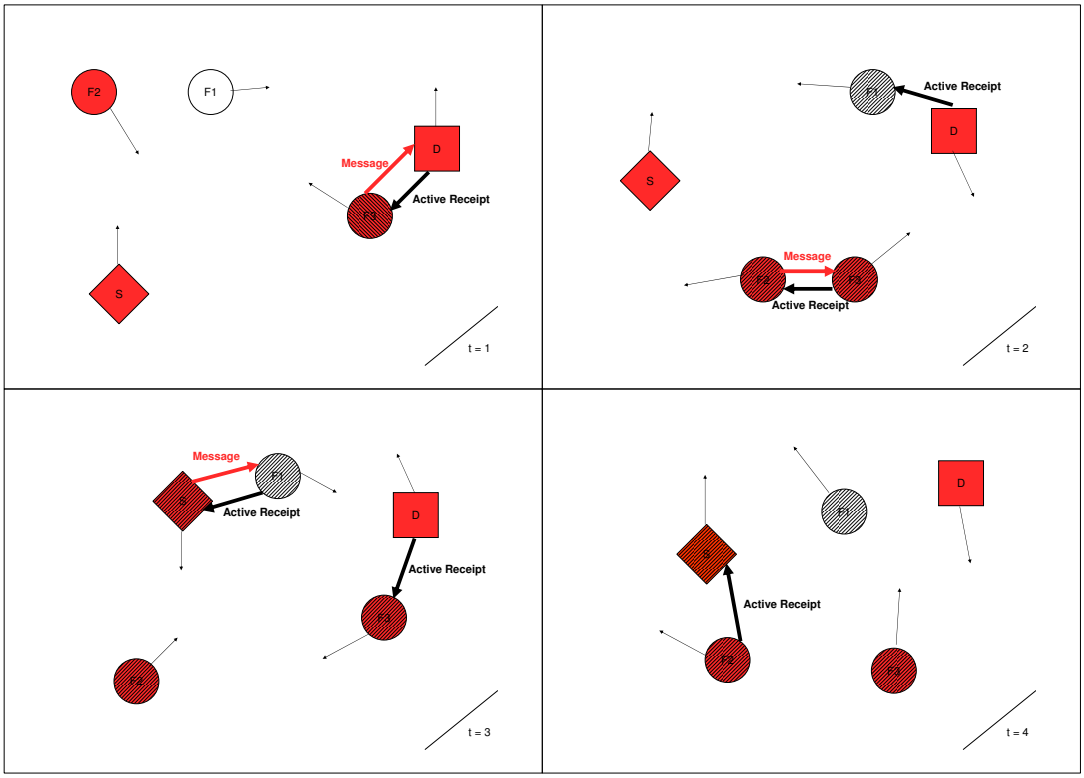


Figure 5.10: Active Receipt Reliability

passive receipt approach. Starting with the same situation where the ultimate destination D has just received the message, we saw that at t=2 with the active receipt approach a receipt message was sent to both an infected as well as to an uninfected node. In the case of the passive approach the cured node F3 sends a receipt to F2 because this node tried to infect F3. Note that it doesn't send a receipt to node F1. We saw that in the active approach at t=4 F4 still sends a receipt to S although the sender S is already cured. This would not be the case in the passive receipt approach. If we compare the figure 5.10 with the Figure 5.11 we see that this reduction in cost doesn't come without a disadvantage. The passive approach results in a slower spreading of the receipt. We see that in the active approach the receipt is received by the sender S in the third snapshot compared to the forth snapshot in the case with the passive approach. This slower dispersion also means that the chance of having an infected node still sending some messages after the sender S has already received the receipt is much higher in the passive approach than in the active.

Network-bridged Receipt

The forth reliability assumption combines all the previous approaches and also adds a new element, namely a cell network, into the architecture.

Figure 5.12 shows the idea of the network bridged receipt approach. For the sake of simplicity, the nodes displayed in the graphic are not shown as mobile nodes visualized with time shifted snapshots but of course we still address a DTMN scenario. In the

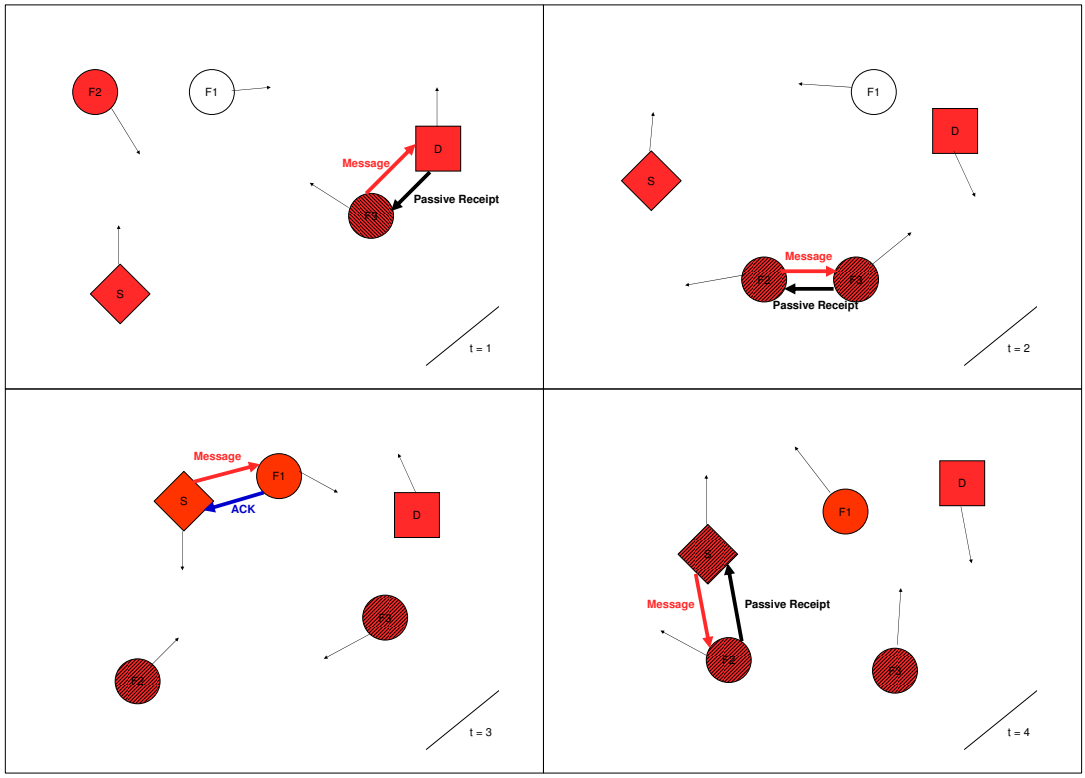


Figure 5.11: Passive receipt reliability

upper section of Figure 5.12 we see a DTM network. This section is earmarked by its discontinuous non-end-to-end path and its high bandwidth. This extremely contrasts the continuous end-to-end path and low bandwidth characteristics of a cell network.

To bring the best of the two different network properties together we assign the message delivery task to the DTMN and assign the end-to-end reliability task to the cell network. Because reliability is delegated to the cell network, the least reliable assumption, scilicet the hop-by-hop reliability approach can be used for the DTMN. (steps 1-6) The cell network is used as an alternative end-to-end path for signalling purpose only. We see that in step 7 an active receipt is sent to a nearby cell phone which than sends a network bridged receipt (step 8) to another cell phone nearby the originally sender. Because this receipt is only sent reactively it looks similar to the passive receipt approach. The cell phone that receives the receipt then sends a passive receipt to the sender (step 9). With the network bridged approached the sender can be contacted directly, so the receipt has not to travel all the way back to the sending source. If also all the other nodes have access to a cell phone, this approach could be used to inform immediately all the other nodes about the successful reception of the message and to heal the whole network in a few seconds.

The advantage of this approach is that it reduces the round trip time between the sender S and the ultimate destination D approximately by half. The drawback lies in the added complexity due to the use of the cell network as a bridge. But, however, it seems to be an auspicious approach for the future.

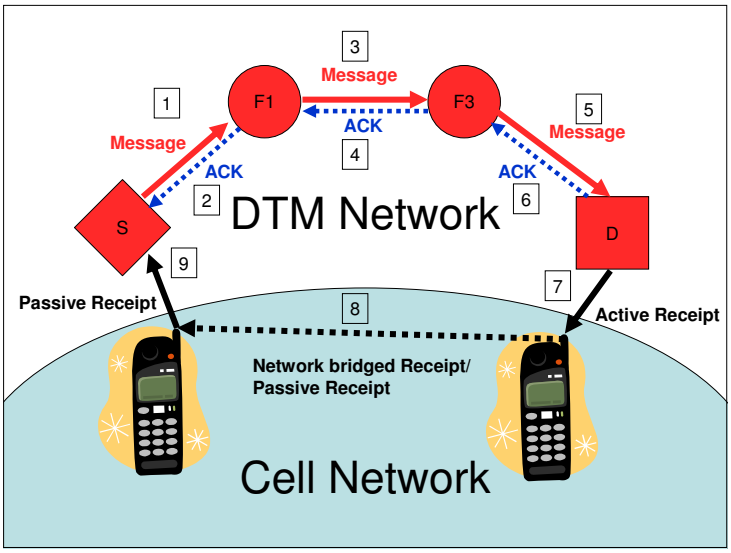


Figure 5.12: The network bridged receipt reliability approach

5.4 Related projects

The focus of this chapter lies on related projects to DTN.

5.4.1 Plutarch

[2] DTN and DTNM both support and connect networks, which differ in the used protocols. Communication between participants, who are in different networks, is made possible by an overlay network architecture. A homogeneity is thus produced by an additional layer. The heterogeneity is hidden .

Plutarch is in the beginnings and shows a direction in which the future development could/should go. Programmatic details are not yet relevant.

Idea

The idea of Plutarch is that the heterogeneity is accepted. A network, which is homogeneous in itself forms a context. A context consists of hosts, routers, switches, network links and much more. Homogeneity refers to addresses, package formats, transport protocols and naming services. The application types are similar to DTN and DTNM, where the employment of IPv4 or IPv6 is not possible (sensor networks) or not desired (intradomain functionality which ignores IP).

A scenario

A student with current location in Switzerland wants to check the status of a specific sensor in a sensor network located in Vancouver. The sensor network is linked to the Internet through a host. The host communicates with the sensors by using their MAC address. The student sets up a connection to the Internet with his GPRS capable cellphone attached to his notebook. Because the IP service of the GPRS provider places its service at the disposal by an obfuscated gateway and the sensors don't implement an IP stack in order to save energy neither of the two endpoints are directly addressable from the Internet.

With the help of Plutarch a connection end-to-end can be set up anyhow in three phases:

Phase 1: Naming

Because one cannot use global names the host or service must be searched for by a query. The query for the name takes place "epidemic-style-gossip" [2]. That means the naming service has to query across contexts. The answer on the query includes a chained-context description which is nothing else than a description of the contexts and the interstitial functions between the end-to-end points.

Phase 2: Chained context instantiation:

A context chain is chosen and the query is sent to a host at the edge of the context of the GPRS network where the interstitial function is configured. The query is forwarded to a host at the edge to the sensor network where the interstitial function is configured.

Phase 3: Communication.

Concept

Communication between contexts, that means different networks, is thus made possible by using interstitial functions. The functions naming, addressing, routing and transport should be made possible end-to-end over heterogeneous networks by explicit cooperation at the edges of the networks (see 5.13).

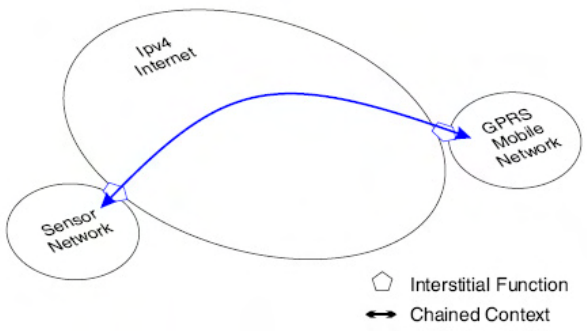


Figure 5.13: Interstitial functions

The Plutarch group believes that thereby the network model illustrates the reality better and that thereby the network model is easier expandable (new services integration).

Important components of Plutarch: Interstitial functions / context / end-to-end naming and addressing

- Interstitial functions: In order to establish an end-to-end communication between two users in different contexts, the data must be manipulated at the edge of the context. A bridge between these two different networks is thus provided. Interstitial functions bridge two networks and consist of two interfaces, whereby one interface each represents a network. Internal mechanisms translate then the arriving packet of a context in the other.

- Context: Contexts serve two purposes:

  - They describe communication mechanisms, which are contained in the different networks. Thus a endpoint can set up a session.
  - They serve as description, thus end-to-end services can be developed.

  A context, which describes a local LAN environment, would specify that the used protocol is Ethernet and that the link supports a speed of 100 Mb/s. An end point can be simultaneous in 2 contexts. 2 interfaces: 1x Ethernet 1x ATM (for example: a computer with a local LAN network card and a dial-up modem). The affiliation to a context is therefore dynamic and it should mechanisms be made available with which the entry and leaving of a context are regulated.

- End-to-end naming and addressing: From the view of the Plutarch group the functions naming and addressing should not be separated. Exactly this is the case with the today Internet:

  - addressing: IPv4
  - Naming: DNS

  That leads to the fact that globally bound addresses and names are necessary. With this system radically different networks can not be supplied with addresses and names (IPv6 is only one global addressing schema. i.e. Sensors are also not supported because they don't have an IP layer implemented). Therefore one would like a heterogeneous naming schema: No global names and addresses. Each network exists in an explicit context with own names and addresses. That leads to a flexibility in the end system. In order to make the naming and addressing between contexts possible, again interstitial functions are used.

  Possible methods for resolutions of the problems addressing, naming, routing and transport are:

  addressing: Mapping between two different contexts is a common problem (e.g. network address translation NAT). The Plutarch group proposes that programmatic interfaces with accordant functionality should be offered. With its help mappings should be built up automatically and administered by the network users.

naming: Momentarily DNS offers only one, global namespace. The Plutarch group forecasts that blossoming services such as VoIP and personal area networking will favor an alternative beginning of mapping between several naming systems. Reasons are scalability and minimize administrative overhead.

routing: for efficiency different networks need different routing protocols. By Interstitial functions routing information is to be translated of a context in the other.

transport: Only one transportation protocol will not be optimal for all network technologies. An example: wireless implementation of TCP, which supports splitting and proxying and enhances the transmission ratio. The optimization of transportation protocols for specific network types is needed. Again interstitial functions guarantee interoperability [2] .

Advantages

The advantages of Plutarch are evident:

- No changes at existing Internet are needed
- Contexts can be gradually added
- Harmonious living together of different networks

5.4.2 Metanet

The basic idea of Metanet is similar to the one of Plutarch. Analog to Plutarch Metanet presumes that the Internet today consists of uniform end-to-end connections and that there are many separate regions. A region in Metanet represents a context in Plutarch. The difference to Plutarch is the concept of membership of a region. A region doesn't only have to represent the technique how its members exchange messages but it can also represent a region of trust, physical proximity or regions of payment. These regions can exist all at the same time and provide great flexibility [12].

5.4.3 Triad

Triad is:

> ".. a new next generation Internet Architecture, which defines an explicit content layer that provides scalable content routing, caching, content transformation and load balancing, integrating naming, routing and transport connection setup. TRIAD is an acronym, standing for Translating Relaying Internet Architecture integrating Active Directories." [13]

With increasing use of content distribution (i.e. Websites) over the Internet some providers can be flooded with requests while resources are scarce. To overcome this problem some techniques are used such as:

- geographically replicating of content:

  > ".. specialized name servers that redirect DNS lookups to nearby (to the client) sites based on specialized routing, load monitoring and Internet "mapping" mechanisms, so-called content routing."[13]

- transparent caching of content by proxies
- load balancing switches which make it possible to route content requests accordant to the load of the server (content provider).

These methods are not in the terms of the original Internet design. I.e Using content routing the violation is that it needs a DNS server which is able to access routing information so it can locate the nearest content location but the users still have to contact a centralized DNS server in order to access the webcontent. It seems that the limitation in performance is becoming the roundtrip contacting the centralized server to obtain the IP address. By using transparent caching extra delays occur when the content is not cachable at all. Using load balance switches the original meaning of end-to-end connection is violated because network address translation on which load balancing switches rely on need to rewrite every incoming packet.

Triad can solve these issues. IP addresses are reduced to routing tags while the end-to-end identification is based on names and URLs. Content routing is becoming more efficient and unneeded roundtrip times to a centralized server become obsolete because of the content layer [13].

5.4.4 Further projects

Interesting but not discussed projects are:

- IP Next Layer (IPNL)
- 4+4
- AVES
- SelNet project
- NewArch
- TIER
- Saami

5.5 Conclusion

Connectivity on the Internet is based on constantly connected end-to-end paths with low delay, low error rates and relatively equal up and down speed. But there are also networks with completely other characteristics. There are for example independent, specialized networks, in which sensors with very limited power supply implement no IP stack, satellites, reachable only every 3 hours for 15 minutes or inter-planetary communication with very high delays. When there are high delays/long round trip times, high error rates, missing end-to-end paths or extremely different up and down speeds TCP/IP just fails. DTN aims off to enable communication in such networks which do not meet the assumptions of TCP/IP. Support for these networks is provided by putting an overlay architecture over the existing architecture, the so called bundle layer, by using global compatible names for the addressing, by using store and forward message switching and by deploying custody transfers.

A specific DTN architecture is the DTMN architecture. One assumes that all nodes in the network are mobile and that still no end-to-end connection exists. Each node is regarded as a region. Two fundamental assumptions characterize a DTMN: nodes are blind, so they do not know anything about the other nodes, and nodes are autonomous. Further there are four reliability approaches: Hop by hop, active receipt, passive receipt and network bridged receipt.

Plutarch has similar appendages as DTN. However Plutarch does not try to produce homogeneity on the basis of a new architecture layer but sets on the heterogeneity of the different network types. Communication is established by interstitial functions at the borders of different contexts.

Bibliography

[1] DTNRG Website: `http://www.dtnrg.org/`, Last visited: June 2006.

[2] J. Crowcroft et al, "Plutarch: An Argument for Network Pluralism", SIGCOMM FDNA Workshop, August 2003.

[3] M. Demmer, E. Brewer, K. Fall, S. Jain, M. Ho, R. Patra, "Implementing Delay Tolerant Networking", IRBTR-04-020, Dec. 28, 2004.

[4] R. Durst, "Delay-Tolerant Networking: An Example Interplanetary Internet Bundle Transfer", October 2003.

[5] K. Fall, "A Delay-Tolerant Network Architecture for Challenged Internets", IRB-TR-03-003, February 2003.

[6] K. Fall, "Messaging in Difficult Environments", Intel Research Berkeley, IRB-TR-04-019, Dec. 27, 2004.

[7] S. Farrell, S. Symington, and H. Weiss, "Delay Tolerant Networking Security Overview", draft-irtf-dtnrgsec-overview-01.txt, September 2006.

[8] K. Harras and K. Almeroth. "Transport Layer Issues in Delay Tolerant Mobile Networks". IFIP Networking. Coimbra, Portugal, May 2006.

[9] Khaled Harras, Kevin Almeroth and Elizabeth Belding-Royer. "Delay Tolerant Mobile Networks (DTMNs): Controlled Flooding Schemes in Sparse Mobile Networks", IFIP Netwoking. Waterloo, Canada, May 2005.

[10] S. Jain, K. Fall, R. Patra, "Routing in a Delay Tolerant Networking", SIGCOMM, Aug/Sep 2004.

[11] Forest Warthman."Delay-Tolerant Networks (DTNs): A Tutorial", May 2003.

[12] John T. Wroclawski. "The Metanet", MIT Laboratory for Computer Science, `http://www.cra.org/Policy/NGI/papers/wroklawWP` January 2007.

[13] Standford University "Triad" `http://www-dsg.stanford.edu/triad/index.html`, January 2007.

Kapitel 6

Push Email Systems

Adrian C. Leemann, Amir Sadat

Push E-Mail Systeme finden nach Nordamerika auch in Europa eine grosse Anhängerschaft. Reine PDA Geräte werden immer seltener und bei Neukäufen in der Regel durch leistungsfähige Smartphones und PocketPC's ersetzt. Diese Arbeit stellt aktuelle auf dem Markt agierende Konkurrenten detailliert vor und arbeitet die Unterschiede im Aufbau und Betrieb der Systeme heraus. In der Folge werden die Systeme hinsichtlich technischem Aufbau, Benutzergruppe, Sicherheit und Kosten miteinander verglichen und je nach Einsatzgebiet und Kundensegment die optimale Lösung vorgestellt. Nicht zu vergessen sind die sozialen Konsequenzen, die sich durch den Einsatz der Push E-Mail Technologie ergeben. Auf diese Probleme und Lösungsansätze wird am Ende der Arbeit eingegangen.

Inhaltsverzeichnis

6.1 Einleitung

Hier wird ein kurzer Überblick über die verwendete Technik im Push E-Mail Ansatz gegeben um ein Grundverständnis für die folgenden Analysen zu garantieren. Weiter wird der historische Einsatz der Technik kurz beleuchtet und die Motivation für den heutigen Einsatz von Push E-Mail Systemen, im täglichen Gebrauch, verdeutlicht.

6.1.1 Abgrenzung Push Technik

Im Gegensatz zur Pull Technologie, bei der der Benutzer gezwungen ist, eine aktive Rolle einzunehmen und gewünschten Dienste periodisch zu aktivieren, um Daten abzurufen - so genanntes Polling - funktioniert der Push Ansatz durchgängig angestossen durch den Sender. Dies impliziert, dass Nachrichten und Information augenblicklich und ohne Steuerung des Benutzers auf dem Gerät eintreffen und nicht die Gefahr besteht, wichtige Daten durch manuelle, periodische Abrufe zu verfehlen. Als Voraussetzung dafür gilt, dass mit Push-Technologie ausgerüstete Geräte permanent online und verfügbar sind. Letzteres birgt dieses Push-Verfahren ein Potential für Missbrauch, denn die Kontrolle über den Erhalt von Nachrichten wird vom Benutzer faktisch aus der Hand gegeben. Den Benutzern können irrelevante Informationen (Spamming) zugestellt oder wichtige Nachrichten vorenthalten werden.

Push-Technologien manifestieren sich im Alltag in verschiedensten Formen. So kann man beispielsweise das Telefon sowohl als Pull als auch Push implementierendes Gerät verstehen. Einen Anruf zu erhalten stellt bereits einen Push Ansatz dar, denn das Gerät ist permanent 'online' und wird direkt angesprochen. Andererseits lässt sich der Pull Ansatz dann erkennen, wenn ein Anruf getätigt wird.

6.1.2 Fallbeispiel Point Cast

1996 verfolgte die Firma PointCast einen Push Ansatz mit ihren Produkten. Die Pointcast Software versah den Bildschirmschoner des Benutzers mit Benutzer spezifizierten Informationen wie Börsenkursen oder aktuellen Wetterberichten. Diese wurden bei Änderungen direkt auf das Gerät des Benutzers weitergeleitet und aktualisiert. Der Benutzer konnte gemäss seinen Interessen und Vorlieben seinen Individuellen PointCast Screen zusammenstellen. So war es möglich, mit einem Blick Informationen zu erschliessen, für die üblicherweise der Aufruf verschiedener Datenquellen angefallen wäre. Dieses Push-Konzept führte zu jener Zeit zu einem Hype und man sah diesen Ansatz als Lösung gegen den Informationsüberfluss im Internet. Slogans wie:

> Push! Kiss your browser goodbye, Remember the browser war between Netscape and Microsoft? Well forget it. The Webbrowser itself is about to croak. And good riddance.

machten die Runde. PointCast stand sogar kurz vor einer Einbettung in das Betriebssystem Windows.

Von der Idee her liess sich das Angebot jedoch nicht mit der Philiosophie des Internet vereinbaren, denn es führt zu einer 'lay-back' Benutzung des Internets, im Sinne dass keine aktiven Recherchen mehr nötig sind. Alle relevanten Informationen würden direkt gepusht, doch aufgrund der Fülle des Internets und den stetig ändernden Wünschen und Anlieger der Benutzer ist das nur schwer denkbar. Benutzer beklagten teilweise die Überschüttung mit Informationen, während andere Quellen nicht genügend Nachrichten lieferten. Ein weiteres Problem zu dieser Zeit stellten auch die beschränkten Bandbreiten dar. Pointcast führte zu enormer Ressourcenverschwendung, da eine Mehrheit der Informationen nie aktiv genutzt wurde.

Aufgrund des daraus resultierten Rückgangs der Nachfrage wurde der Dienst 1999 ganz eingestellt [32].

Die Technologie selber konnte sich in dieser Form nicht behaupten. Nichtsdestotrotz hat die Idee überlebt und findet beispielsweise in, heutzutage sehr beliebten RSS Feeds ihre Verwendung, selbst wenn sie technisch streng gesehen keinen Push implementieren.

6.1.3 Einsatzgebiete Push E-Mail

Push E-Mail Applikationen lassen sich besonders im professionellen Umfeld gewinnbringend einsetzen. Am geläufigsten ist der Mobile Einsatz, bei dem typischerweise E-Mails, wie auch andere Daten auf ein Smartphone oder einen PocketPC weitergeleitet werden. Der Einsatz ist besonders interessant für Mitarbeitern, die häufig ausserhalb der Firma tätig sind. So stellen die meisten namhaften Unternehmen im nordamerikanischen und zunehmend auch im mitteleuropäischen Raum den höheren Kadern Blackberry Systeme zur Verfügung. Die Erreichbarkeit dieser Personen kann somit noch weiter gesteigert werden und Koordinationsaufwand wird minimiert. Agiert wird hier nach dem Prinzip, dass möglichst alle Kommunikationswege zu jeder Zeit offen sind und somit Rückrufe oder Voicemail-Anfragen eingespart werden können. In New York gehören U-Bahn Passagiere, die bereits auf dem Arbeitsweg erste E-Mail beantworten, zum Stadtbild. Hier besteht eine grosse Chance, Zeit effizienter zu nutzen.

6.2 Marktübersicht: Push E-Mail Anbieter

Der Markt für Push E-Mail Geräte und Dienstanbieter befindet sich Ende 2006 in einer starken Wachstumsphase. Viele Anbieter versuchen mit ihren Produkten, Fuss zu fassen und Marktanteile für sich zu beanspruchen. Dementsprechend umkämpft sind die Positionen und es vergeht kaum eine längere Zeit, ohne dass ein neuer Anbieter mit einer eigenen Push E-Mail Lösung ins Geschäft einsteigt. In diesem Kapitel werden die Lösungen der Anfangs 2007 grössten und etabliertesten Anbieter vorgestellt. Dabei wird insbesondere auf die zugrunde liegende Architektur, den Aufbau der Sicherheitsmechanismen und die Zielgruppe (Grosskunden gegenüber Privatbenutzern) eingegangen.

6.2.1 Research In Motion

Das Blackberry Endgerät [33] dürfte für jeden ein Begriff sein und wird jeweils als erstes mit der Push E-Mail Technik in Verbindung gebracht. Am treffendsten könnte man einen Blackberry definieren als: 'Von Grund auf E-Mail getrimmter PDA mit Mobiltelefonfunktionen'. Besonders die Anwendung der Push E-Mail Funktion verhalf dem Blackberry zum Durchbruch, obwohl andere Hersteller von PDA Geräten die Organizer Funktionen teilweise erheblich Benutzerfreundlicher anboten.

Research in Motion [34] gilt quasi als Erfinder des Push-Ansatzes im Bereich Mobiler E-Mail Empfang. Keines der grossen Mobilfunkunternehmen war an der Technologie interessiert, was den Entwickler veranlasste das Gerät selber unter dem Namen Blackberry auf den Markt zu bringen. Research in Motion als Marktführer und Pionier stellt daher seit 1999 die bekannten Geräte mit dem Scrollwheel auf der Seite her und vermochte bis Ende 2006 trotz Konkurrenzdruck weiter Marktanteilen zu gewinnen. Es existiert bereits eine sehr grosse Verbreitung im nordamerikanischen Raum, da der Blackberry über Jahre hinweg die einzige adäquate Lösung für mobilen E-Mail Empfang anbieten konnte. Der Blackberry zeichnet sich vor allem durch seine hohe Benutzerakzeptanz und eine über Jahre aufgebaute Reputation aus. Die Kunden betonen besonders das Scrollwheel auf der Gehäuseseite, welches den Einhandbetrieb gewährleistet. Es hat etliche Jahre gedauert bis andere Hersteller diese Ausstattung übernommen haben. Die Hersteller von Blackberry schwören auf eine vollständige Tastatur unterhalb des Displays im Vergleich zu der Konkurrenz, die häufig auf Touchscreen und Palm Schrift setzt. Neuere Geräte aus dem Hause RIM sind auch mit Touchscreens und etwas eingeschränkterer Tastatur auf den Markt gekommen.

Funktionsweise: Neben dem regulären Mobilfunkvertrag benötigt man bei den Blackberry Geräten noch eine sog. Blackberry Option, die den Datentransfer in der Regel pauschal abrechnet. E-Mails werden nicht per POP oder IMAP auf das Gerät geladen sonder direkt mit der Push Technologie auf das Endgerät weitergeleitet. Eine Funktion wie 'Neue Nachrichten abfragen' oder 'Nachrichten senden/empfangen' sucht man auf einem Blackberry vergebens.

Klar zu unterscheiden ist die Blackberry Lösung für Firmen und Grosskunden, die eigene Server betreiben und für Private und Kleinunternehmen. Für letztere ist ein Server Betrieb in der Regel unmöglich oder unrentabel und daher werden diese Dienste auslagert. Ursprünglich versteht man jedoch unter dem Blackberry Dienst die Anwendung für Grosskunden und professionelle Benutzer.

Varianten des Blackberry Dienstes

Architektur Blackberry Enterprise Solution Hierbei wird eine Anschaffung der sogennanten Blackberry Enterprise Server (BES) Software nötig, Installation letzterer auf einem firmeneigenen Server, und eine vorhanden E-Mail Lösung mit Microsoft Exchange Server, Lotus Domino oder Novell Groupwise.

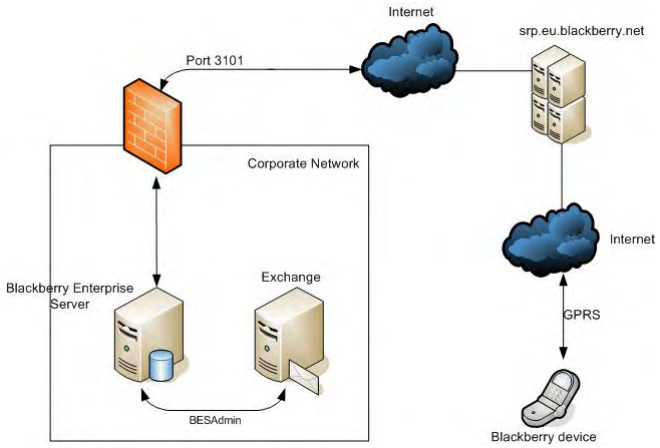


Abbildung 6.1: Blackberry Enterprise Edition Architektur [33]

Auf der Client Seite muss ein Blackberry Handgerät oder ein kompatibles Gerät (sogenannt 'Blackberry enabled') mit entsprechender Software vorhanden sein. Der Blackberry Enterprise Server registriert einen E-Mail Eingang auf dem Mailserver und leitet die neue Nachricht über das Internet an das Endgerät. Diese Weiterleitung läuft über einen der drei Weltweiten Research in Motion Server. Im Falle Europa steht dieses Mobile Routing Center in England, was zuletzt zu hitzigen Diskussionen Anlass gab. Abbildung 6.1 stellt diesen Vorgang am Beispiel eines Exchange Servers und des Routing Centers in England übersichtlich dar.

Das Blackberry Handgerät ist via GPRS, oder bei neueren Infrastrukturen auch via WLAN oder UMTS ununterbrochen mit diesem Server zur Synchronisation verbunden. Das Endgerät meldet sich beim Mobile Routing Center an und stellt so eine Konnektivität zum Enterprise Server her. Das Routing Center leitet Pakete des Enterprise Servers lediglich an das Endgerät weiter. Der Synchronisationsprozess wird vom Blackberry Enterprise Server angestossen. Letzterer überwacht den Mailserver und registriert Veränderungen. Trifft ein neues Mail ein oder wird ein Kalendereintrag aktualisiert wird der Synchronisatinsprozess mit dem Endgerät ausgeführt Eine neue Nachricht wird somit unverzüglich auf das Gerät weitergeleitet und eine quasi Realtime-Zustellung kann garantiert werden.

Geschätzt wird bei dieser Lösung von den Kunden besonders die Tatsache, dass vom Enterprise Server bis zum Endgerät alles durch einen Hersteller angeboten wird. Somit können Schnittstellenprobleme und Konversionen minimiert werden und ein schneller produktiver Einsatz ist wahrscheinlicher. Auch ermöglicht die gute Komprimierung der E-Mails einen minimalen Datenaustausch, womit Übertragungskosten gesenkt werden können.

Vorteile der Blackberry Enterprise Architektur

- Eine einzelne Verbindung durch die Firewall des Unternehmens
- Enterprise Server kann schlank gehalten werden
- Ganze Architektur abgedeckt durch RIM

Architektur Blackberry für Kleinunternehmen, Anwälte und Privatpersonen
Hier ist der Push-Ansatz im eigentlichen Sinne gegeben, aber die Idee eines Realtime zugestellten E-Mails ist hier nicht mehr durchgehend vorhanden. In dieser Lösung steht der Blackberry Enterprise Server direkt beim Mobilfunkanbieter welcher die Überwachnung der E-Mail Konten für den Kunden übernimmt. Das bedeutet jedoch, dass Zugangsinformationen zu E-Mail Konten an Drittunternehmen, in diesem Falle der Mobilefunkanbieter, weitergegeben werden. Das ganze wird in der Regel über periodisches POP Anfragen einer oder mehrerer E-Mail Konten realisiert. Ist seit der letzten Anfrage eine Nachricht hinzugekommen wird diese vom entsprechenden Server geladen und via GPRS auf das Endgerät weiter gepushed. Das bedeutet also de facto, dass eine E-Mail auch mit der Verzögerung der definierten Pollingzeit beim Empfänger ankommen kann.

Vorteile der Blackberry Prosumer Edition

- Anschaffung eines eigenen Servers wird hinfällig
- Auslagerung von Wartung und Betrieb des Servers
- Relativ reibungslose Konfiguration und sofortiger Betrieb

Sicherheit

Hersteller Research in Motion setzt bei den Blackberry Geräten auf eine Ende zu Ende Triple-DES (3DES) [41] oder AES [42] Verschlüsselung. Der vereinbarte Key wird bei der Initialisierung über die Dockingstation einmalig übertragen und ermöglicht so eine sichere Verbindung. Gemäss Research in Motion existiert kein universaler Schlüssel zum Dechiffrieren der Nachrichten. Ein solcher würde beispielsweise das Zwischenlesen am Routing Center ermöglichen. Die privaten Schlüssel werden ausschliesslich von den Unternehmen generiert und verwaltet.

Abbildung 6.2 illustriert die Ende zu Ende Verschlüsselung über die einzelnen Netzwerkabschnitte. Die 3DES und AES Verschlüsselungsmechanismen gelten heute als relativ sicher und sollten auch bei fortschreitender CPU Leistung, nach Moore's Law [43], noch einige Jahre resistent gegenüber Angriffen sein.

Durch einen Bericht der „Computerwoche“ [35] wurden im Juni 2005 etliche Blackberry Benutzer und Sicherheitsverantwortliche aufgeschreckt. Es war die Rede davon, dass AUDI [44] Aufgrund Sicherheitsbedenken - sprich: Angst vor Industriespionage - die Verbannung der Blackberry Geräte aus dem Unternehmen plante. Als problematisch wurde dabei die Tatsache, dass der gesamte Verkehr über den Blackberry Europa Server in Grossbritannien läuft betrachtet. In der Folge würde bei einem Streitfall englisches Recht zur Anwendung kommen. Hier war der sogenannte RIP Act [1] Stein des Anstosses. Gemäss diesem Gesetz könnte der englische Geheimdienst bei Verdacht auf schädliche Handlungen gegen Grossbritannien, oder zum Schutz des englischen Volkes, Research in Motion auf Herausgabe eines Dechiffrierungs-Schlüssels belangen. Es wäre somit denkbar, das Audi in England, zum Schutze der eigenen Wirtschaft, beziehungsweise der Automobilindustrie, ausspioniert würde. Research in Motion dementierte in einem Communiqué [45] aber, wie

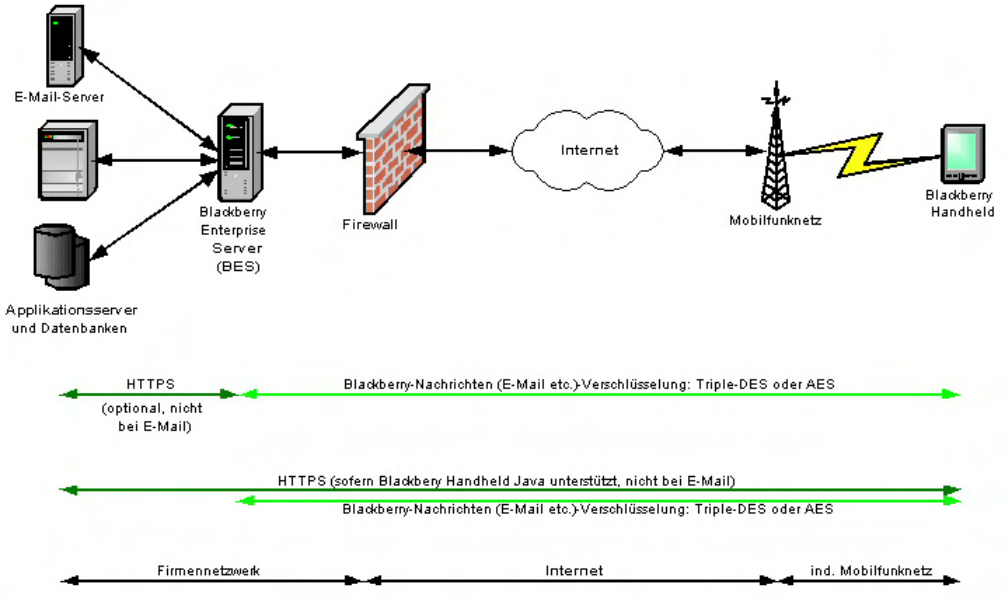


Abbildung 6.2: Blackberry Enterprise Edition Verschlüsselung [33]

erwähnt, die Existenz eines solch Dechiffrierungs-Schlüssels. Ebenfalls dementiert wurde das Gerücht, dass Research in Motion auf jedem BES über Administrator Rechte verfüge und theoretisch alle Nachrichten einsehen könnte. (Laut Gerüchten nur durch eine Vereinbarung zwischen Research in Motion und dem Abnehmer verboten, aber technisch möglich)

Nichtsdestotrotz untersuchen viele Unternehmungen ihre mobilen E-Mail Geräte und verfassen Reglemente über den Einsatz der Technologie. So empfiehlt zum Beispiel das Bundesamt für Sicherheit in der Informationstechnik (BSI) [3] aus Deutschland nur Geräte zu verwenden die einen eigenen Kryptoalgorithmus einbinde lassen. Für als 'streng geheim' deklarierte Bereiche, wie sichere Kommunikation zwischen Botschaften, empfiehlt das BSI auf den Einsatz von 3DES oder AES zu verzichten und auf das Produkt SINA-VPN [2] mit dem Libelle [4] Algorithmus zurückzugreifen.

Wie bei vielen Smartphone und PocketPC's lässt sich beim Blackberry das Gerät mit einem Timeout sperren. Das bedeutet: Wenn keine Bedienung stattfindet, wird nach einem wählbarem Timeout automatisch eine Pingeschützte Tastensperre aktiv wird. Der Code lässt sich bis 14 stellig setzen und schwache Kennwörter wie 1234567 etc. lehnt das System ab. Das Passwort wird als SHA-1 Hash im Gerät gespeichert. Es kann definiert werden, dass nach zehnmaliger Falscheingabe des Codes der gesamte Speicherinhalt automatisch gelöscht wird. Somit ist das Risiko erheblich minimierbar, dass Drittpersonen an sensible Geschäftsdaten gelangen können. Bis anhin waren allerdings die Nachrichten und Attachments auf den Blackberry Geräten im Plaintext gespeichert. Neuere Blackberry unterstützen nun eine AES Verschlüsselung für diese sensiblen Daten. Dem Administrator des BES ist auch ein sogenannter 'Remote Wipe' möglich. Er kann das Gerät veranlassen, seinen Inhalt zu vernichten und somit Datendiebstahl zu verhindern. Dies ist keinesfalls nur eine Spielerei, sonder tagtäglich in Einsatz. In Chicago wurden 2005 alleine in Taxis 85000 Mobiltelefone liegen gelassen oder verloren [5]. Allerdings ist zu bedenken, dass die Remote Wipe Funktion nur funktioniert, wenn eine Verbindung zum Gerät hergestellt wer-

den kann. Es ist jedoch relativ leicht den Empfang eines Blackberry abzuschirmen oder die Sendeeinheit zu beschädigen. Selbstverständlich sind die Geräte gegen mechanische Eingriffe ziemlich machtlos. In Anbetracht dieser, nicht unerheblichen Schwachstellen, ist es durchaus sinnvoll, den Einsatz zu überdenken und gegebenenfalls zu reglementieren.

Ein Benutzer der Blackberry Lösung für den privaten Bereich, bei der der Blackberry Enterprise Server zum Beispiel bei Orange steht, muss natürlich beachten, dass hier zwar eine Ende zu Ende-Verschlüsselung vom Blackberry Enterprise Server zum Endgerät besteht, jedoch die Administratoren bei Orange nach Belieben die Mail zwischenspeichern und lesen könnten. Grundsätzlich besteht diese 'Gefahr' bei allen E-Mail Anbietern wie GMX, Hotmail, GMail etc. ohne dass sich jemand darüber brüskiert. Es ist aber zu beachten, dass diese Kanäle immer ein Risiko darstellen können, besonders für sensible Daten wie sie beispielsweise bei Anwälten oder Ärzten entstehen.

Auch Research in Motion kämpft mit den Gefahren der IT Branche. Es kursieren immer wieder Meldungen über Firmen welche aufgrund von Sicherheitsbedenken oder Fehlern des Blackberry Enterprise Servers angeblich ihren Mitarbeitern den Einsatz der BlackBerry's verbieten. Am 27.10.2006 soll es mit dem Service Pack 4.02 ein Problem gegeben haben, das bei diversen Blackberry Kunden auftraten. Angeblich sollen Fragmente von Nachrichten an andere E-Mail Empfänger gelangt sein als der Sender ursprünglich eingegeben hatte [7].

6.2.2 Good Technologies

Good Technologies und Motorola

Good Technologies ist ein weiterer bedeutender Anbieter von Push E-Mail Lösungen. Nach einer Darstellung der organisatorischen Aspekte wird hier analog zu Blackberry die Architektur und die Sicherheitsfragen beleuchtet.

Good Mobile Messaging [39] Die im Silicon Valley ansässige Good Technologies [36] stellt Push E-Mail Dienste für verschiedene Geräte und Hersteller zur Verfügung. Grössen wie Nokia, Palm und einige Hersteller Windows Mobile basierter Geräte setzen die Lösung von Good Technologies bereits ein. Ende 2006 begann die Übernahme von Good Technologies, welche über sehr gute Kontakte zu verschiedensten Mobilfunkanbiertern rund um die Welt verfügen, durch den Mobilfunkriesen Motorola. Anfangs 2007 Jahres sollte die Akquisition vollzogen und Good Technologies in den Motorola Konzern integriert sein.

Einerseits könnte das zu einem Problem werden, denn nun ist es sicherlich schwieriger, Partnerschaften mit anderen Geräteherstellern einzugehen, andererseits beteuert die neue Leitung des Unternehmens, dass weiterhin eine Mehrgeräte-Strategie gefahren wird. Motorola will weiterhin Lösungen für Fremdhardware entwickeln und betreiben. Die Partnerschaften mit anderen Marken sollen gepflegt und ausgebaut werden. Nokia hat bekanntlich Intellisync übernommen und setzen diese Software bereits auf ihren Geräten ein. Daher

wird eine weitere Kooperation von Nokia und Good Technologies, bzw. Motorola als sehr unwahrscheinlich angenommen. Um eine breitere Grundlage für den Dienst zu schaffen, wurde nun auch eine Unterstützung von Lotus Notes und Microsoft Exchange Servers implementiert. Im Hause Motorola erhofft man sich durch das gute Renomée für stabile Geräte, dem Rivalen RIM Marktanteile abtrotzen zu können.

Es lässt sich sehr schwer abschätzen wie sich der Markt für diese Geräte entwickelt wird und wer sich hier durchsetzen kann. Daher ist die Akquisition durch Motorola für Good Technologies sicher ein Vorteil. Motorola, einer der Keyplayer im Mobilfunkmarkt, vermochte im 2005 einen Umsatz von 35.3 Milliarden US Dollar verbuchen und dürfte daher über die nötigen Reserven verfügen um gegen RIM oder die Lösung aus Redmond bestehen zu können.

Die Zielgruppe von Goodtechnologies sind Grosskunden und Firmen, für Privatpersonen und Kleinuntenehmen wird keine Lösung angeboten.

Good Mobile Internet Good Technologies setzt ähnlich wie Blackberry auf einen Transformationsansatz. Das bedeutet, dass nur 'relevante' Informationen einer Webpage angezeigt werden. Werbung und Bilder, die die Übertragung nur verlangsamen würden, werden effizient gefiltert und können bei Bedarf nachgeladen werden. Dies ist klar ein Vorteil gegenüber den Microsoft Mobile Geräten, die für diese Arbeit getestet wurden. Good Mobile Internet versucht weiter auch mit geschickter Zwischenspeicherung auf den Geräten nur aktualisierte Daten neu zu laden (Caching). Besteht eine WLAN Verbindung fällt dieser Vorteil natürlich geringer aus, doch ist eine flächendeckende, frei zugängliche WLAN Versorgung noch nirgends gegeben, und daher haben diese Massnahmen durchaus ihre Berechtigung.

Neben dem Einsehen von Webseiten aus dem öffentlichen Bereich ermöglicht Good Mobile Internet auch den Zugriff auf das gesicherte Intranet. Internet Zugriffe werden allerdings bei Good Technologies nicht direkt über die GPRS oder WLAN Verbindung realisiert, sondern umgeleitet über den Firmenserver. Dies hat zum Vorteil, dass eine gezielte Filterung des Inhaltes greifen kann und das Gerät von schädlichem Einfluss aus dem Internet geschützt wird. Der Seitenaufbau einer Webpage erfolgt immer abschnittweise, so dass nur so viele Daten übertragen werden müssen wie der User gerade sieht.

Aus der Firma können jegliche Arten von Daten direkt einen Mitarbeiter auf sein Gerät gepusht werden. Somit lässt sich praktisch jede Applikationen der Unternehmung auf das Mobilgerät auslagern, was ganz neue Ideen und Möglichkeiten schafft. Dem Administrator, ist es sogar möglich seine Aufgaben über Good Mobile Internet wahrnehmen und Benutzer online zu verwalten oder Updates auslösen ohne direkt am Good Link Mobile Messaging Server präsent zu sein. Remote Monitoring des Servers oder der Benutzer wird so in jedem versorgten Gebiet möglich.

Architektur Good Technologiers

Im Aufbau setzt Good Technologies quasi eine Kopie der Blackberry Lösung ein. Auch hier erfolgt die Kommunikation mit den Endgeräten über ein Operations-Center, das durch

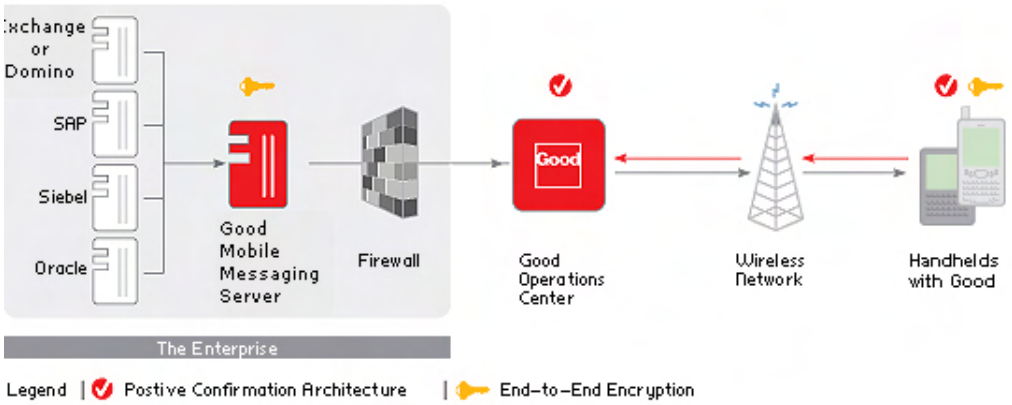


Abbildung 6.3: Good Technologies Architektur [36]

die Firewall mit der Firma verbunden ist. Die Endbenutzer kommunizieren mit dem Good Security Operations Center [38] und sind via Mobilfunkbetreiber oder WLAN mit letztem verbunden. Abbildung 6.3 zeigt den Aufbau von Good Technologies und Parallelen zum Ansatz von RIM sind deutlich zu erkennen

Sicherheit Good Technologies

Mit der Sicherheit eines Push E-Mail-Systems steigt und fällt der Erfolg der Lösung. Keine Firma wird ein Produkt einsetzen das nur den dritt- oder viertbesten Sicherheitskriterien genügt. Der Betrieb mit Anwendungen aus dem Bereich Customer Relationship Management und Enterprise Resource Planning verlangt nach einer hoch vertraulichen Behandlung der übertragenen Daten. In der Folge wird nur der Zugriff auf Daten des Unternehmens gewährt, deren Sicherheit man garantieren kann. Die Sicherheit sollte zur Dauer der Übertragung, wie auch bei der permanenten Speicherung auf dem Gerät jederzeit vollumfänglich gegeben sein.

Oberstes Gebot ist dabei, dass der Sicherheitsstandard der Firma nicht herabgesetzt wird durch den Einsatz der mobilen Geräte. Insbesondere muss sichergestellt werden, dass keine Möglichkeit besteht via einem mobilen Gerät auf die Daten hinter der Firewall zuzugreifen.

Auf dem Weg vom Firmenserver zum Endgerät ist besonders auf effiziente Verhinderung von man-in-the-middle [47] Angriffen zu achten. Es muss davon ausgegangen werden, dass Pakete auf der Luftschnittstelle abgefangen werden und Entschlüsselungsversuche stattfinden. Starke Verschlüsselungsverfahren greifen hier in der Regel sehr effizient und bieten eine sichere Ende zu Ende-Verschlüsselung.

Good Technologies setzt bei der Ende zu Ende-Verschlüssleung auf einen AES [42] Key, via SSL [46], der bei der ersten Inbetriebnahme des Gerätes ausgetauscht wird. Gemäss Good Technologies verfügt das Good Security Operations Center nicht über diesen Key oder ist in der Lage die Nachrichten mit einem universellen Schlüssel zu dechiffrieren. Um die Sicherheit noch zu erhöhen können die Schlüssel mit Verfalldatum versehen werden, die bei Ablauf über die verschlüsselte Verbindung neu generiert werden. Der AES Key wird in einer Länge von 192Bits unterstützt.

Good Technologies beherrschen auch die Verschlüsselung nach dem FIPS 140-2 Standard [8], wie von vielen Organisationen und öffentlichen Stellen gefordert.

Der Aufbau der Architektur lässt wie bei RIM die 'Auslagerung' ins Operations-Center als problematisch erachten. Das Operations-Center ist ein Single Point of Attack und muss daher zum Beispiel entsprechend gegen Denial of Service Attacken [48] geschützt sein. Im Gegenzug ermöglicht das Operations-Center eine sehr einfache Konfiguration der Endgeräte, da sie fest vorprogrammierte Kommunikationspartner haben. Ein weiterer Vorteil ist, dass der Good Mobile Messaging Server sehr schlank und effizient gehalten werden kann. Komplexe Aufgaben wie das Zwischenspeichern von Nachrichten oder erneute Übertragung bei Fehlern, wenn sich das Gerät in einen nicht versorgten Gebiet befindet, übernimmt das Good Security Operations-Center. Weiter ist nur eine Verbindung (physisch) durch die Firewall zum Operations-Center nötig um allen Geräten den Zugang zu ermöglichen.

Sind die Daten nun einmal auf dem Gerät angelangt, muss sichergestellt werden, dass sie dort auch adäquat verwaltet werden. Vorkehrungen gegen lokale Programme und Viren, die Informationen weiterschicken können, müssen getroffen werden. Wireless- und Bluetoothfähige Geräte müssen besonders gegen Manipulationen geschützt werden. Oftmals werden die genannten Schnittstellen wie auch die Kamerafunktion oder der Lautsprecher blockiert.

Über das Administrator-Tool lässt sich die Verschlüsslung des Gerätes definieren oder die Speicherkarte chiffrieren. Einen Remote Wipe, wie von Windows Mobile bekannt, kann der Administrator ausführen; ebenso lässt sich auch das automatische Löschen des Speichers bei mehrmaliger Falscheingabe des PIN definieren.

Elementar für die Sicherheit der Daten ist, dass ein Gerät zu jedem Zeitpunkt seine Authentizität beweisen kann um so zu verhindern, dass andere Geräte eine Identität vortäuschen können um Nachrichten abzufangen. Es muss sichergestellt sein, dass keine Drittperson eine Partnerschaft vortäuschen und mit den Firmenservern zu kommunizieren kann.

Dazu verwendet Good Technologies eine im ROM einprogrammierte Seriennummer oder die Seriennummer der SIM Karte. Auf dem Server werden diese Informationen ausgewertet und der Zugriff verwaltet. Auf vielen Handhelds ist das ROM aber mehr oder weniger beliebig austauschbar und Manipulationen sind mit entsprechenden Kenntnissen ein Kinderspiel. Bekannt sind solche Manipulationen eventuell aus dem Bereich der 'branded Devices' beispielsweise von Orange, die nur mit Orange Simkarten funktionieren sollen. Innerhalb weniger Minuten ist es möglich, den Simlock zu entfernen und gemäss Gerüchten ist es auch kein Problem, eine IMEI zu spoofen.

Ein Unlock Tool für solche Orange HTC Geräte, das Modifikationen am ROM vornimmt, das scheinbar sehr gut funktioniert, ist frei zugänglich und wird relativ leicht gefunden [27].

Manipulationen von Simkarten sind zwar komplex, aber durchaus realistisch. Ein andere Möglichkeit wäre die Simulation einer Simkarte [9]. Daher sind Bedenken an der Identifikation mit diesen Parametern gerechtfertigt.

6.2.3 Microsoft ActiveSync

Microsofts neustes Mobiles Betriebssystem Windows Mobile 5.0 feierte seinen Launch im Mai 2005. Diese neue Version ermöglichte mobilen Windows Geräten erstmals den Empfang von E-Mails per Push Technologie, Microsoft Direct Push genannt. Damit zog Microsoft schliesslich - gut 6 Jahre nach dem Pionier Research in Motion - mit ihrer eigenen Push E-Mail Lösung nach. Microsoft möchte in diesem Markt, das sich im Wachstum befindet, Marktanteile für sich gewinnen. Ein Unterfangen, das Erfolg haben kann, sieht man sich die Entwicklung von Windows Mobile an [10].

Die breite Funktionalität von Windows Mobile macht Microsofts Alternative attraktiv:

- Office Mobile (bestehend aus Word, Excel, Powerpoint)
- Outlook Mobile
- Internet Explorer Mobile
- Pocket MSN (MSN Messenger, MSN Hotmail)

Da viele Unternehmungen bereits Microsoft Office für ihre Desktop-Systeme einsetzen, kann Microsoft hier Netzwerkexternalitäten realisieren. Office für Desktopsysteme und Office Mobile sind weitestgehend kompatibel, so kann sich ein Mobile Worker beispielsweise im Flugzeug letzte Änderungen an einer Powerpoint Präsentation vornehmen. Dies spiegelt auch Microsofts Strategie wieder: „One of the top priorities of our salesforce around the globe will really be to drive Windows mobile penetration into the business market," [12] so Microsoft CEO Steve Ballmer. Anders als andere Anbieter vermarkten sie nicht nur ihre Push E-Mail Lösung sondern ihre ganze Plattform, bestehend aus diversen Software-Lösungen.

ActiveSync

Die Push Funktionalität wird von ActiveSync geliefert, einer Software, die es Endgeräten ermöglicht, die Mailbox mit einem anderen System zu synchronisieren. Klassisch wurden Windows Mobile Geräte durch ActiveSync mit Desktop Systemen synchronisiert. Konnektivität konnte beispielsweise über USB, Bluetooth oder Infrarot hergestellt werden. Neue ActiveSync Versionen erlauben auch drahtlosen Remote-Zugriff (Fernzugriff) und somit eine Synchronisation mit Exchange Servern über Drahtlosnetzwerke. Es gibt verschiedene Möglichkeiten, das mobile Gerät mit dem Server zu synchronisieren:

- Polling des Servers durch Outlook Mobile (Minimalintervall 5 Minuten). Dieses Verfahren beruht auf Pull Technologie
- Windows Mobile AUTD (Always Up To Date): Microsoft Exchange Server 2003 implementierte eine sogenannte „SMS-up-to-date" Funktionalität[1]. Trafen neue Informationen auf dem Server ein (beispielsweise eine neue E-Mail), wurde eine SMS an das mobile Gerät gesendet, welches darauf Synchronisationsprozess initiierte.

[1] AUTD wurde vor Service Pack 2 eingesetzt und ist inzwischen veraltet

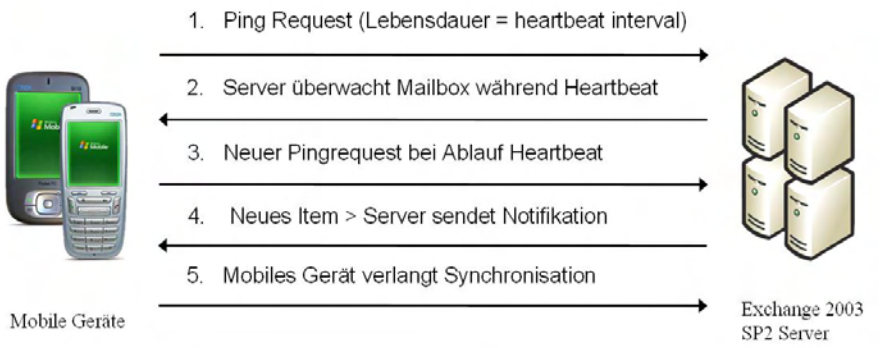


Abbildung 6.4: Synchronisationsablauf zwischen Mobilem Endgerät und Exchange Server [Quelle: Eigene Darstellung]

- Die neueste Möglichkeit ist der auf Microsoft Direct Push basierende Empfang von Daten, bei dem neue Outlook Informationen (E-Mail, Kalender, Kontakte und Tasks) vom Server automatisch auf den Client gestossen werden.

AUTD war Microsofts erster Schritt in Richtung Push E-Mail (genau genommen handelt es sich bei AUTD nicht um Push E-Mail [13]). Das Konzept litt jedoch unter Schwächen: Einerseits wurden die versandten SMS dem Benutzer vom Mobilfunkanbieter auf Basis empfangener Notifikationen in Rechnung gestellt. Dies stellte für einen Benutzer mit hohem E-Mail Verkehr eine äusserst unökonomische Lösung dar [13]. Ausserdem gibt es bei SMS keinerlei Garantie, dass die Nachrichten auch tatsächlich beim Empfänger bzw. beim Gerät ankommen (mögliche Gründe: Überlastetes Mobilfunknetz, etc). In diesem Fall bemerkte das Endgerät nichts von der Veränderung der Mailbox, die Synchronisation bliebe somit aus.

Microsoft Direct Push kann die Schwächen von AUTD beseitigen. Zwei Voraussetzungen müssen jedoch erfüllt sein, um von dieser neuen Funktionalität Gebrauch zu machen.

Den mobilen Geräten muss entweder Windows Mobile 5.0 als Betriebssystem mit dem Messaging and Security Feature Pack (MSFP) Update dienen oder es muss sich um „ActiveSync enabled Devices“ handeln. Kurz, das Gerät muss das ActiveSync Protokoll unterstützen. Bestimmte Partner von Microsoft haben ActiveSync lizenziert, die namhaftesten sind etwa Nokia, Sony Ericsson, Symbian, DataViz, Motorola und Palm [14]. Es gibt also eine breite Auswahl an Handsets, die mit Microsofts Lösung kompatibel sind.

Die restriktivere Voraussetzung betrifft den eingesetzten Server seitens der Unternehmung. Nur ein Microsoft Exchange Server 2003 SP2[2] als Frontend kann die gewünschte Funktionalität liefern[3] (nachfolgend wird „Exchange Server“ stellvertretend für „Exchange Server 2003 SP2“ verwendet).

Abbildung 6.4 demonstriert den Synchronisationsablauf: Das mobile Gerät sendet dem Exchange Server einen Ping Request, damit signalisiert es die Bereitschaft, Nachrichten

[2]DirectPush wurde erst mit dem Service Pack 2 integriert
[3]Es ist natürlich möglich, andere Mail-Server ausser Exchange im Backend zu betreiben

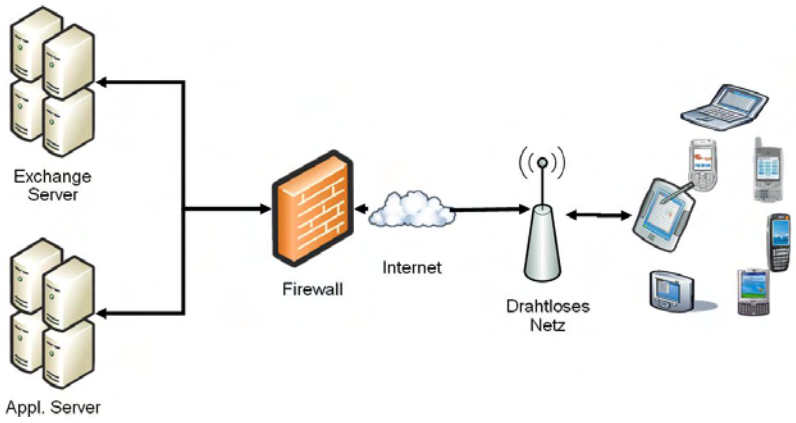


Abbildung 6.5: Microsofts Push E-Mail Architektur [16]

oder Daten zu empfangen. Danach fällt es in einen passiven Modus, um Energie zu sparen. Der Ping Request hat eine bestimmte Lebensdauer, auch „Heartbeat“ genannt. Während dieser Zeitspanne überwacht der Exchange Server das Postfach auf Veränderungen. Wird keine Veränderung festgestellt, läuft der Heartbeat nach einer gewissen Zeit ab[4], das mobile Gerät sendet erneut einen Ping Request. Wird irgendwann eine Veränderung durch Exchange Server festgestellt, beispielsweise durch den Eingang einer neuen E-Mail, notifiziert es das Endgerät darüber. Dieses verlangt daraufhin die Synchronisation.

Zusammenfassend die Push E-Mail relevanten Features von Windows Mobile 5.0 oder kompatiblen Geräten [14] [15]:

- Synchronisation aller Outlook Items zwischen mobilen Geräten und Exchange Server 2003: E-Mails, Kalender, Kontakte, Tasks
- Synchronisationmethoden: Manuell, (AUTD), Microsoft Direct Push
- Mehrere E-Mail Ordner simultan unterstützt
- GAL (Global Address Lookup): durchsuchen firmeninterner Adressbücher für eine vereinfachte E-Mail Zustellung
- Partielles Downloaden von E-Mail Nachrichten, der Client kann festlegen, wie viel er erhalten will oder das ganze Item laden
- Attachments: Automatischer Download (nach Dateigrösse oder Dateityp), on-demand Download oder manueller Download
- Weiterleitung und Beantwortung von E-Mails direkt auf dem Server ohne Zwischenlagerung auf mobilem Gerät
- Features zur Performancesteigerung auf Mobilen Geräten: Timelimits für die Speicherung von Items, gzip Kompression, usw.

[4]Typischerweise werden neue Ping Requests alle 5-20 Minuten gesendet

Architektur

Die Architektur von Microsoft Direct Push unterscheidet sich grundlegend von der Architektur anderer Anbieter. Typischerweise fand man einen mobilen Mailserver[5] vor, der für das Monitoring der Mailserver im Backend verantwortlich ist und die Push Funktionalität implementiert. Ausserdem wird häufig ein Relay Server bzw. Operations Center[6] eingesetzt, welches für das Routing, die endgültige Zustellung der Inhalte auf das Endgerät zuständig ist. Beide Instanzen fallen bei Direct Push weg (Abbildung 6.5). Der pushfähige Exchange Server stösst die Inhalte direkt über das Internet hin zum Endgerät. Das bedeutet, Middleware eines Drittanbieters in Form eines Servers hinter der Firmenfirewall ist unnötig [16]. Eingriffe in die bestehende IT-Infrastruktur sind somit überflüssig. Ausserdem können so Lizenzkosten gespart werden. Natürlich setzt dies voraus, dass die Unternehmung bereits über einen Exchange Server verfügt[7], ansonsten ist auch hier eine Anschaffung eines (Exchange) Servers und somit Lizenzgebühren fällig. Das Wegfallen des Relay Servers hat besonders sicherheitsrelevante Vorteile. Daten müssen nämlich nicht mehr über die Server eines Dritten geleitet werden, was bei Unternehmungen häufig aus Datenschutzgründen auf Bedenken stösst.

Sicherheit

Microsofts Massnahmen, um Sicherheit im mobilen E-Mail Verkehr zu garantieren, lassen sich in folgende Kernbereiche gliedern [14] [15]:

- HTTPS/SSL Verbindung von zwischen Outlook Mobile und Exchange Server
- Lokales Passwort und Timeoutmechanismus auf Mobilem Gerät
- Remotewipe-Funktionalität

Um Outlook Items zu pushen wird seit Outlook 2003 die End-zu-End Verbindung zwischen Exchange Server und Client durch SSL (Secure Socket Layer) verschlüsselt[8]. SSL verschlüsselte Verbindungen gelten heutzutage als relativ sicher. Während die Konkurrenz aber ausnahmslos für die Verschlüsselung u.a. den sehr sicheren AES (Advanced Encryption Standard) Algorithmus verwendet, setzt Microsoft hier noch auf RC4 (ARCFOUR) oder 3DES (Triple DES). Beide Algorithmen gelten für heutige Verhältnisse als nur noch bedingt sicher, weshalb von ihrem Einsatz abgeraten wird.

Lokal auf dem mobilen Gerät ist ein Passwortschutz implementiert, was vor unbefugtem Zugriff schützen soll. Der Passwortschutz lässt sich so konfigurieren, dass nach mehrmaliger Falscheingabe alle Daten und Einstellungen auf dem Gerät automatisch gelöscht

[5] Bsp.: Blackberry Enterprise Server

[6] Bsp.: Blackberry Routing Center

[7] Der Marktanteil soll sich Mitte 2005 auf ca. 50%, laut Microsoft sogar 75%, belaufen haben

[8] Bevor End-zu-End Verschlüsselung unterstützt wurde, musste die Sicherheit durch ein VPN (Virtual Private Network) oder RAS (Remote Access Service) gewährleistet werden.

werden, was eine Sicherheitsmassnahme gegen Diebstahl darstellt [15]. Der Timeoutmechanismus funktioniert wie man es auch von Desktop Systemen gewohnt ist: Nach einer bestimmten Inaktivitätsperiode des Gerätes sperrt es sich und ist erst nach Eingabe des Passworts wieder betriebsbereit. Sollten alle Stricke reissen, gibt es auch bei Microsofts Lösung die Möglichkeit, die Daten eines Endgerätes durch einen Fernzugriff zu löschen. Durch ein Webtool können Administratoren oder Help-Desk angestellte den Prozess initieren. Ausserdem ist es ihnen möglich, Security Policies der mobilen Geräte festzulegen, beispielsweise Minimallänge des verwendeten Passwortes, die Anzahl erlaubter Passwort-Fehlversuche, usw.

Alternativ zur Anmeldung mit Usernamen und Passwort am Exchange Server, wird auch eine zertifikatbasierte Authentifikation unterstützt.

Kritisiert wird, dass als lokale Sicherheit bloss ein Passwortschutz dient. Windows Mobile verfügt über keine Datenverschlüsselungsfunktion, was dazu führt, dass Daten im Klartext auf das Gerät geschrieben werden [17]. Dies gilt als grosser Sicherheitsmangel. So wird zwar der over-the-air Datenstrom zwischen Exchange Server und Endgerät via SSL verschlüsselt, aber gelingt es jemandem, das Passwort zu hacken, zu umgehen oder die Daten durch einen mechanischen Eingriff auszulesen, liegen sensible Daten offen. Bei Verlust oder Diebstahl eines Windows Mobile Gerätes trennt also lediglich das Passwort vor dem Offenliegen des Inhalts. Microsoft jedoch weist die Kritik zurück und ist der Meinung, dass eine solche Funktionalität kein Kernanliegen der Kunden darstellt, denn die Funktionalität kann durch Software eines Dritten gewährleistet werden, falls erwünscht [18].

6.2.4 SEVEN

Die Firma SEVEN wurde im Juni 2000 gegründet und ist somit früh auf den Push E-Mail Markt gestossen. Da Research in Motion den US Markt weitestgehend beherrschte, enschied sich SEVEN Ende 2005, die in Europa ansässige Firma Smartner zu übernehmen, um global schneller wachsen zu können[9] [19]. Mit diesem Schritt avancierte SEVEN zu einem der grössten Anbieter von Push E-Mail.

Weltweit operiert SEVEN inzwischen in über 60 Ländern und hat Verträge mit über 100 Telekommunikationsanbietern[10] in Amerika, Europa, Japan, Asien und Afrika. Unter anderem bestehen Verträge mit Orange, O2, ePlus, NTT DoCoMo, Yahoo!, mtc Vodafone, usw.. Der Push E-Mail Dienst wird aktiviert durch eine SEVEN Software, die auf dem mobilen Gerät installiert wird. Über 200 mobile Geräte, basierend auf den wichtigsten Mobilen Betriebssystemen, werden derzeit unterstützt[11], darunter: BREW, J2ME, Microsoft Pocket PC, Microsoft Smartphone 2003, Microsoft Windows Mobile 5.0, Palm OS und Symbian basierte Geräte, PDAs und gängige Internet Browser. SEVENs Kernkompetenz liegt in ihrer globalen Präsenz und der Kompatibilität ihrer Dienstleistungen mit Geräten verschiedenster Hersteller und Betriebssystemen.

[9] Smartner war besonders tätig in Europa, aber auch weiten Teilen Asiens
[10] Stand 3. Quartal 2006
[11] Stand 3. Quartal 2006

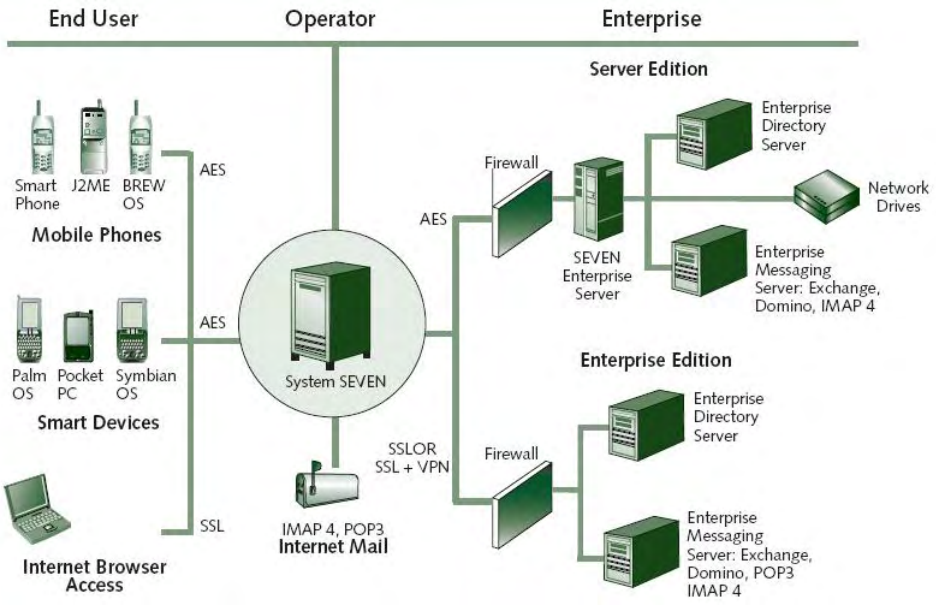


Abbildung 6.6: SEVENs Architektur [20]

„Our goal at SEVEN is to make mobile E-Mail available on every mobile phone at an affordable price."

Im Gegensatz zur Konkurrenz beschränken sie sich nicht auf Grossunternehmen als Zielgruppe sondern bieten Push E-Mail auch für den Privatanwender oder SOHO Bereich an, was sich in der angebotenen Produktlinie zeigt.

Consumer Edition

Die Consumer Edition erlaubt Push E-Mail Zustellung von POP3- oder IMAP basiertem Internet-E-Mail, wie es beispielsweise von Google Mail, Yahoo!, Hotmail, AOL Mail, usw. angeboten wird [21]. Wie der Name Consumer Edition bereits verrät, richtet sich das Angebot speziell an Privatkunden oder SOHO Angestellte, die ihre E-Mails von einem Internetanbieter beziehen. Dementsprechend einfach gestaltet sich der Dienst. Man installiert eine SEVEN Applikation auf dem Mobilen Gerät und meldet sich mit den E-Mail Kontodaten an. Anstatt die E-Mails nun über Pull, also durch Polling des Mailservers, manuell zu empfangen, werden sie automatisch auf das Gerät gestossen. Dies übernimmt der beim Mobilfunkanbieter stehende System SEVEN Server (Abbildung 6.6).

Da die Zustellung der E-Mails eine recht hohe Verzögerung von bis zu 5 Minuten mit sich zieht[12], kann davon ausgegangen werden, dass es sich dabei nicht um reines Push E-Mail handelt. Es ist durchaus möglich, dass der System SEVEN Server, welcher beim Mobilfunkanbieter steht, das Konto beim Internetanbieter periodisch nach neuen Mails durchsucht und sie dann auf das mobile Gerät stösst. Eine Kombination aus Pull und Push also.

[12]Der Dienst wurde getestet

Personal Edition, Workgroup Edition

Bei der Personal Edition handelt es sich um eine Desktop Client Solution für Kleinunternehmen ohne IT Fachkompetenz. SEVENs Software wird auf einem beliebigen Desktop System installiert, das eine Verbindung mit dem Mailserver des Unternehmens hat [22]. Grundsätzlich funktioniert das Desktop System analog zu den bekannten Enterprise Servern (BES, Exchange Server, etc.), indem es als eine Art Proxy fungiert und die Synchronisation mit dem Mailserver erlaubt. Diese Lösung soll lediglich eine weniger komplexe, low-cost Alternative zu Diensten mit einem Server sein. Die Workgroup Edition funktioniert analog, unterstützt jedoch eine grössere Anzahl Benutzer und erlaubt den Zugriff auf mehrere Mailboxen.

Server Edition

Bei der Server Edition handelt es sich um eine „behind-the-firewall“ Lösung. Ein SEVEN Enterprise Server wird hinter der Firewall in die vorhandene Unternehmensinfrastruktur integriert und übernimmt die Synchronisation zwischen Endgerät und Mailbox (Abbildung 6.6) [23]. Der Einsatz des Enterprise Servers erlaubt im Vergleich zur Desktop-Lösung mehr administrative Kontrolle über das System. Das IT Involvement seitens der Unternehmung ist dementsprechend höher, diese Edition dürfte daher eher Grossunternehmen ansprechen.

Ein wichtiger Unterschied zu Research in Motion ist, dass der System SEVEN Server nicht von SEVEN selbst verwaltet wird sondern jeweils beim Mobilfunkanbieter steht. Die Funktionalität, nämlich eine Art Operations Center, ist jedoch dieselbe.

Enterprise Edition

SEVEN bietet zusätzlich die Möglichkeit, den Enterprise Server zum Mobilfunkanbieter auszulagern (Abbildung 6.6). Diese „carrier-hosted“ Lösung hat zur Folge, dass keine Veränderungen in der Infrastruktur der Firma getätigt werden müssen, jedoch greift der Server des Mobilfunkanbieters so jeweils auf die Firmenserver zu [23]. Ausserdem basiert diese Enterprise Edition nicht auf über HTTP gesendete Notifikationen sondern auf SMS Notifikationen, wie es bei Microsofts AUTD der Fall war. Anders als bei Microsoft wird jedoch nicht jede Notifikation einzeln verrechnet, sondern über monatliche Pauschalen an den Mobilfunkanbieter geregelt. Im Vergleich zur Server Edition ist diese Lösung weniger komplex für die Firma und entlastet die IT, von der Sicherheit her betrachtet aber stellt es eine äusserst bedenkliche Alternative dar, denn man erlaubt dem Mobilfunkanbieter Zugriff auf die eigenen Server.

Sicherheit

End zu End-Verbindungen zwischen mobilen Geräten und System SEVEN Servern (Relay Servern) werden mit dem 128bit AES (Advanced Encryption Standard) verschlüsselt [20].

Die Schlüsselgrösse von AES ist skalierbar auf 192 oder sogar 256bit, was SEVEN zu gegebener Zeit durch transparente Updates zu unterstützen gedenkt. HTTP Verbindungen sind 128bit SSL Verschlüsselt. Da bei der Enterprise Edition die Telekommunikationsgesellschaft für Zugriffe auf Firmenserver zuständig ist, ist unter Umständen zusätzliche Sicherheit angebracht. Optional kann deshalb die Verbindung zwischen den Servern der Unternehmung und dem SEVEN Server der Telekommunikationsgesellschaft durch ein IPSec VPN geschützt werden. Ausserdem lässt sich die Anmeldung am System SEVEN Server analog zu Microsofts Ansatz der Authentifikation mit Zertifikaten bei SEVEN für über sogenannte Token bewerkstelligen. Die Möglichkeit für IT Administratoren, Daten auf mobilen Geräten durch Fernzugriff zu löschen, ist auch bei SEVEN gegeben.

6.2.5 Nokia IntelliSync

Ursprünglich handelte es sich bei IntelliSync um eine 1993 gegründete, amerikanische Firma (Gründungsname: Pumatech), die Synchronisationssoftware für Mobiltelefone und PDAs herstellte. Anfang 2006 wurde IntelliSync von Nokia akquiriert. Nokia entschied sich zu diesem Schritt, um im umkämpften mobilen Markt Fuss fassen zu können.

Nokias Intellisync Mobile Suite stellt eine Lösung für den mobilen Arbeitnehmer von heute dar. Sie vereint diverse Anforderungen wie z.B. PIM Synchronisation und Mobiles E-Mail. Die Mobile Suite besteht aus folgenden Kernkomponenten [25]:

- Intellisync Wireless E-Mail: Empfang und Versand von E-Mails auf mobilen Geräten
- Intellisync PIM Sync: Synchronisiert Kontakte, Kalender, Telefon Directories und andere PIM Informationen von Groupware-Plattformen auf das mobile Gerät
- Intellisync Device Management: IT-Management der Mobilen Plattform (u.a. ist Management von Security Policies, Konfiguration von Geräten, Verteilung von Software, Inventar & Reporting, etc möglich)
- Intellisync File Sync: Automatisierung der Verteilung von Dateien innerhalb eines Netzwerks auf Mobile Geräte
- Intellisync Data Sync: Erweiterung von Daten oder Applikationen auf andere Geräte

Nokia addressiert ihre Lösung ausschliesslich an Unternehmen und nicht an Privatkunden.

Architektur & Sicherheit

Wie Abbildung 6.7 demonstriert, stellt die Intellisync Mobile Suite eine „behind-the-firewall“ Lösung dar. Die Intellisync Mobile Suite Platform übernimmt als Server die Synchronisation von Mailbox und mobilem Gerät. Auffallend ist, dass das System ohne

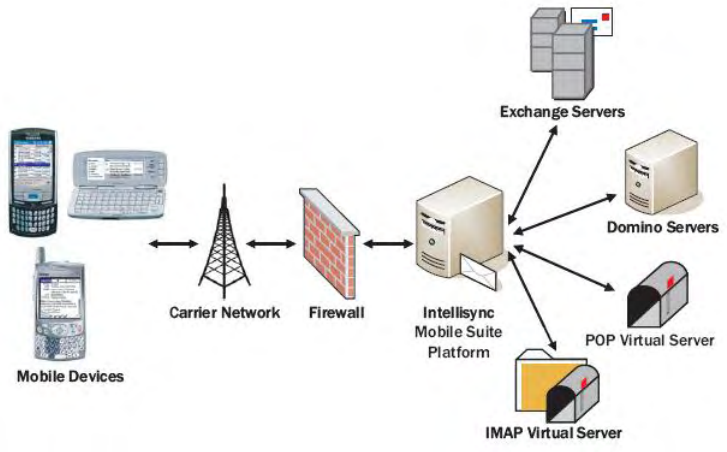


Abbildung 6.7: Nokias Intellisync Architektur [26]

Relay Server bzw. ein Operations Center auskommt. Man könnte es deshalb als eine Art Hybridlösung zwischen Microsoft ActiveSync und beispielsweise Blackberry sehen[13].

Damit die Synchronisation auf Push Basis stattfinden kann, stellt der Intellisync Mobile Suite Server eine ausgehende TCP/IP Verbindung zum Netzwerk der Mobilfunkanbieter her. Erkennt der Server neue, zu synchronisierende Informationen, sendet er dem mobilen Gerät über diesem Wege eine Push Notifikation, welches sich daraufhin beim Intellisync Server meldet, um den Synchronisationsprozess zu starten. Dies geschieht serverseitig auf einem dedizierten TCP Port, der - um die Sicherheit zu erhöhen - nur sehr selektiv spezifische Pakete akzeptiert [26]. Für den drahtlosen Datenverkehr gibt es mehrere over-the-air Verschlüsselungsoptionen: Daten vom mobilen Gerät zum Server werden per AES oder 3DES verschlüsselt via HTTP oder HTTPS. Werden Daten zu einem mobilen Gerät geschickt, welches über das Web Portal verbunden ist, kommt eine SSL Verschlüsselung zum Zuge [26].

Für Unternehmen, die um zusätzliche Sicherheit bemüht sind, kann optional ein Intellisync Secure Gateway eingesetzt werden (Abbildung 6.8), der in der DMZ (demilitarized zone / demilitarisierte Zone: Zone, welche sich zwischen dem internen Netzwerk einer Organisation und einem externen Netzwerk - hier dem Netzwerk des Mobilfunkanbieters - befindet) angebracht wird. Diese auf Java basierende Software stellt sicher, dass sämtlicher Datenverkehr des Intellisync Servers durch ausgehende TCP Verbindungen stattfindet. Will sich ein mobiles Gerät also mit dem Intellisync Server synchronisieren, muss es mit dem Secure Gateway Kontakt aufnehmen, dieser routet dann die Pakete nach einer Inspektion (packet filtering) zum Intellisync Server weiter. Da der Intellisync Server stets eine ausgehende Verbindung zum Gateway offenhält, wird diese auch für den anschliessenden Datenverkehr verwendet [26]. Fernzugriff von Angestellten auf das interne Firmennetzwerk ist mit dieser Lösung jedoch auch ausgeschlossen. Besitzt ein Unternehmen bereits selbst einen Reverse-Proxy dieser Art, kann er in den Synchronisationsprozess eingebunden werden und die Rolle des Intellisync Secure Gateways übernehmen.

Da Synchronisation zwischen mobilem Gerät und Server ein häufiges Vorkommnis ist, wäre

[13]Bei Microsoft ist ebenfalls kein Relay erforderlich und bei Research in Motion kommt ebenfalls ein mobiler Mailserver zum Einsatz

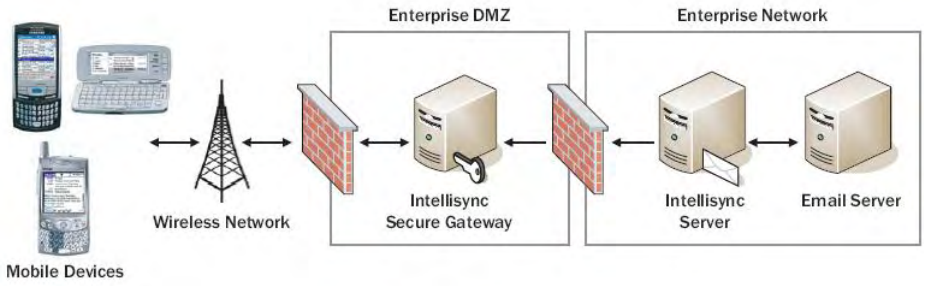


Abbildung 6.8: Erweiterte Sicherheit durch den Intellisync Secure Gateway [26]

es umständlich, sich jedes Mal manuell beim Server zu authentifizieren (Passworteingabe). Andererseits aber ist es gefährlich, zur Authentifikation benötigte Informationen direkt auf dem Gerät zu speichern, um eine automatische Anmeldung zu gewährleisten. Um dies zu umgehen, verwendet Nokia sogenannte „authentification token“. Meldet sich ein Benutzer beim Server an, kann dieser so konfiguriert werden, dass er ein Token auf dem mobilen Gerät ablegt. Dieses mit dem 160bit Blowfish Algorithmus verschlüsselte Token wird künftig dazu verwendet, um sich automatisch zu authentifizieren, die Gültigkeitsdauer kann dabei vom IT Administrator festgelegt werden. Jedem Token wird ausserdem eine eindeutige Geräte ID zugeordnet, damit eine Übertragung auf andere Geräte verunmöglicht wird [26].

Durch das Device Management hat der IT Admin die Möglichkeit bei Verlust oder Diebstahl eines Gerätes Benutzer, Benutzergruppen oder Geräte gezielt zu sperren. Reicht dies nicht aus, kann er per Fernzugriff stufenweise weitere Sicherheitsmechanismen in Gang setzen. Ihm stehen verschiedene Optionen beim Fernzugriff offen [26]:

- Ein Gerät vom Synchronisationsprozess ausschliessen
- Nur PIM und E-Mail Daten eines löschen
- Bestimmte Applikationen, Dateien und sonstige Daten löschen
- Daten von einem Speichermedium löschen
- Kill (hard reset) des Gerätes, alle Daten und Applikationen werden eliminiert

Diese Vorgänge können manuell an Geräte adressiert werden oder automatisch ausgelöst werden, wenn bestimmte Zustände eintreten. Ausserdem ist es durch das Device Management möglich, Antivirus Software eines Dritten auf die Geräte zu verteilen oder zu aktualisieren, um zusätzlichen und einheitlichen Schutz für alle Geräte zu bieten.

Daten werden ausserdem auf dem Gerät verschlüsselt. Sensible, unternehmensinterne Informationen sind so geschützt, selbst wenn das lokale Passwort umgangen werden sollte oder ein Auslesen der Daten durch mechanischen Eingriff erfolgen sollte. Auch Nokia hat hier - im Gegensatz zu Microsoft - umgesetzt, was von Sicherheitsexperten gefordert wird [17].

6.3 Aussichten Push E-Mail Markt

Nachdem die grössten Push E-Mail Anbieter nun vorgestellt wurden, stellt sich die Frage, wie deren zukünftige Chancen am Markt aussehen, wo ihre Stärken und Schwächen liegen und wie sich der Push E-Mail Markt generell entwickeln wird. Nach einem kurzen Review des vorhergegangenen Abschnittes wird versucht, ein Ausblick auf das Marktgeschehen zu geben.

6.3.1 PDA Markt

Er erwies sich als äusserst schwierig verlässliche Zahlen über die Konkurrenten im Markt zu erhalten. Je nach dem ob der PDA Markt inklsive Smartphone und PocketPC analysiert wird oder nur der Markt der Push E-Mail Geräte, weisen die Zahlen grosse Abweichungen auf.

Als gesichert gilt, dass der PDA Markt sehr stark rückläufig ist, man spricht von Einbrüchen im Bereich von 30% [28] für Hersteller wie HP oder Palm. Alte Geräte werden nicht mehr mit neune PDA erstzt; die Benutzer kombinieren ihr Mobiltelefon mit dem PDA und kaufen einen Blackberry oder einen PocketPC ein. Der reine PDA dürfte in geraumer Zeit vom Markt verschwinden und ganz durch die oben erwähnten Geräte abgelöst werden. Daher bestehen für alle Hersteller sehr gute Marktchancen, sofern sie den Sicherheitskriterien der Unternehmen und den Ansprüchen der Benutzer gerecht werden können.

6.3.2 Review der Anbieter

RIMs Lösung ist für viele Kunden attraktiv, da sie aus einer Hand kommt. Sowohl auf Software- wie auch auf Hardwareseite entwickeln und administriert die Unternehmung ihre Server und kann so eine hohe Verfügbarkeit und Konsistenz gewährleisten. Ausserdem hat Research in Motion als Pionier in der Branche Reputationsvorteile. Nach wie vor stösst der Blackberry auf grosse Akzeptanz. Durch die frühe Akquisition von Grosskunden bestehen für diese Wechselkosten, was einen Wechsel auf alternative Anbieter unattraktiv erscheinen lässt. In der Vergangenheit Negativschlagzeilen gemacht haben RIMs Routing Center, die sowohl technische Bedenken[14] als auch Datenschutzbedenken[15] ausgelöst haben. Auch fallen für ihre Lösung im Vergleich zur Konkurrenz hohe Kosten an. Die hohen Total Cost of Ownership scheinen aber bei den Kunden nicht das primäre Entscheidungskriterium zu sein. Frost and Sullivan [30] haben zu diesem Kriterium folgende Grafik 6.9 in ihrer Studie veröffentlicht:

Auch bietet RIM wohl die kleinste Auswahl an Handsets, die den Ruf haben, nicht unbedingt „trendy“ zu sein - im Vergleich zur Konkurrenz, die sich deutlich vielfältiger und

[14]Vgl. Fehlleitung von E-Mail Fragmenten
[15]Vgl. RIP Act

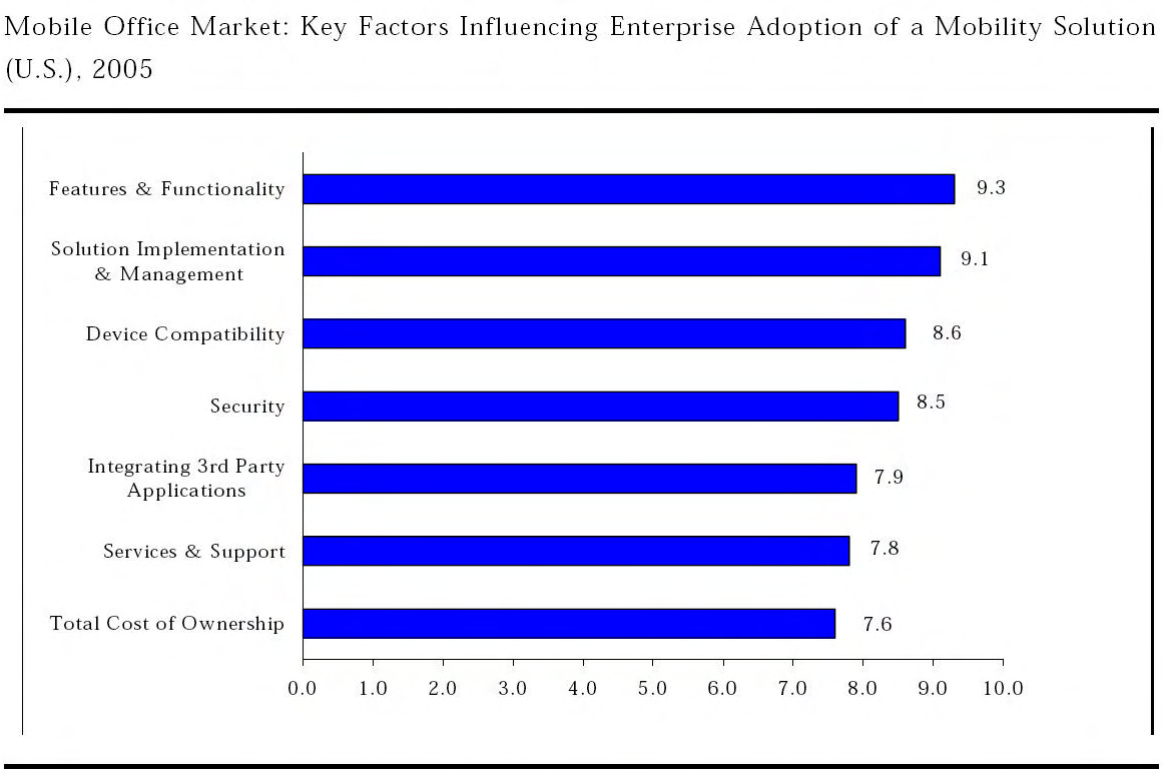


Abbildung 6.9: Studie Frost + Sullivan für Good Technologies [30]

offener zeigt. Diese Schwäche scheint Research in Motion jedoch erkannt zu haben, der Trend zeigt seit kurzem auch bei ihnen in Richtung Vielfalt und Design.

Good Technologies als weiterer grosser Anbieter kann ihre Lösung auf praktisch jedem verfügbaren Endgerät realisieren. Ihre Lösung bietet hohe Kompatibilität und ist bestens in jedes Mobile Betriebssystem integriert. Die Akquisition durch Motorola stellt ihnen ausserdem ein hohes Wachstumpotential in Aussicht. Die Studie [30] hält fest, dass Good Technologies und RIM die besten komplett Lösungen aus einer Hand anbieten. Als Vorteil gegenüber RIM werden besonders die Over the Air Management Funktionen der Good Technologie Lösung die, IT Kosten und Prozesszyklen reduzieren sollen, herausgestrichen. Durch den fast identischen Aufbau der Lösung, sind die Schwächen ähnlich wie bei RIM anzusiedeln. Es stellt sich die Frage, inwiefern sich Good Technologies von Research in Motion zu differenzieren vermag. Durch die junge Übernahme durch Motorola ist es schwierig, Aussagen über die zukünftige Strategie dieser Unternehmung zu machen.

Microsofts Stärke liegt besonders in der verwendeten Architektur. Exchange Server verfügt über einen enorm hohen Marktanteil, zudem ist die ActiveSync Kompatibilität der Handsets sehr hoch. Viele Unternehmungen verfügen also bereits über die benötigte Infrastruktur, um Push E-Mail zu realisieren. Ist dies der Fall, fallen weder Anschaffungskosten für Middleware (in Form eines Servers) an noch Kosten für den Push E-Mail Dienst selbst. Netzwerkexternalitäten durch die weite Verbreitung ihrer Produkte ist Microsoft auch als grosser Vorteil anzurechnen. Immer wieder gerät Microsoft jedoch wegen Sicherheitsaspekten in die Kritik, so auch bei Push E-Mail. Die zur Verschlüsselung verwendeten Algorithmen sind nicht state-of-the-art, über eine lokale Verschlüsselung verfügt Windows

Tabelle 6.1: Stärken und Schwächen der Push E-Mail Dienstleister [Quelle: Eigene Darstellung]

	Stärken	**Schwächen**
Research in Motion	Aus einer Hand, Reputation	Kosten, Datenschutz, Endgeräte
Good/Motorola	Multiplattform	Kosten, Datenschutz, Differenzierung?
Microsoft ActiveSync	Architektur, Netzwerkexternalitäten, (Kosten)	Sicherheit
SEVEN	Privatbenutzer	Datenschutz
Nokia IntelliSync	keine Besonderheiten	keine Besonderheiten

Mobile 5.0 nicht. Obwohl ActiveSync FIPS 140-2 zertifiziert ist, haben sie gegenüber der Konkurrenz im Vergleich das Nachsehen bei den Sicherheitsmechanismen. Auf Grossunternehmen mag das negative Auswirkungen haben.

SEVEN bietet als einziger Anbieter Lösungen an, die auf den Privatanbieter oder kleine Unternehmungen zugeschnitten sind. Sie sind ausserdem global tätig und haben Verträge mit vielen, grossen Mobilfunkanbietern und verfügen über einen enorm grossen addressierbaren Benutzerstamm [24]. Die Auslagerung vieler Dienste an den Mobilfunkanbieter vermindert zwar administrativen Aufwand seitens der Unternehmer, kann jedoch Bedenken in Sachen Datenschutz auslösen.

Nokia als recht junge Partizipantin in der Branche bietet bereits eine solide Lösung, die durch die starke Sicherheit besonders Grossunternehmen ansprechen dürfte. Diese Push E-Mail Lösung fiel bisher jedoch weder besonders positiv noch negativ auf.

Die Tabelle 6.1 gibt einen Kurzüberblick über das erläuterte Stärken-/Schwächenprofil der verschiedenen Anbieter.

6.4 Soziale Faktoren

Dieses Kapitel wird kurz die aus der neuen Technologie erwachsenden Probleme, besonders in sozialer Hinsicht beleuchten und mögliche Lösungen für einen sozial verträglichen Einsatz der Push E-Mail Geräte aufzeigen.

6.4.1 Einfluss der Push E-Mail Geräte im Alltag

Auch im europäischen Raum beginnen die Geschäfte ihre Öffnungszeiten auszudehnen, was in Übersee schon lange Standard ist. Ökonomisch gesehen ist das natürlich ein fragliches Unterfangen, da sich somit nicht die Gesamtverkaufsmenge steigern lässt, jedoch die

Verkäufe besser über die ganze Öffnungszeit verteilt werde. Diese gesteigerte Verfügbarkeit scheint mit den neuen Technologien auch auf andere Berufsgruppen ausgeweitet zu werden. Besonders in den Vereinigten Staaten legen die Mitarbeiter eine hohe Bereitschaft an den Tag jederzeit für die Unternehmung zur Verfügung zu stehen.

Den Anfang machte der Pager, der es einem Arbeitgeber ermöglichte einen Mitarbeiter zu einem Rückruf aufzufordern. Letztere kam diesem Anliegen mehr oder weniger motiviert nach. Darauf folgten die Mobiltelefone, die die Erreichbarkeit der Mitarbeiter noch einmal drastisch steigerten. Notebooks nehmen heute im öffentlichen Leben eine sehr dominante Stellung ein. So wird während Geschäftsreisen nicht mehr im Flugzeug die Zeitung gelesen oder mit dem Nachbar die wirtschaftliche Situation des Landes besprochen, sondern, sobald das Anschnallzeichen erloschen ist, das Notebook aufgeklappt und an der bevorstehenden Präsentation gearbeitet. Die in dieser Arbeit vorgestellten Push E-Mail Geräte sind seit der Einführung 1999 die neuste Errungenschaft in diesem Bereich.

Ein modernes PDA Gerät verfügt über die Möglichkeit E-Mails quasi Real Time dem Empfänger zuzustellen und man erwartet in der Regel darauf auch eine unverzügliche Bearbeitung. Liegezeiten von mehreren Tagen werden nicht mehr toleriert. Weiter implementieren einige dieser Geräte sogar die Möglichkeit direkt auf Firmendatenbanken zuzugreifen und mit Workflowprogrammen zu interagieren.

Auf der einen Seite ist das hervorragend für die Unternehmungen, denn frühere sog. tote Zeit auf Reisen oder Ähnlichem kann nun aktiv genutzt werden. Zu klären wäre, ob wirklich die Produktivität steigt oder ob der Mitarbeiter so lange benötigt, um alle Geräte zu beherrschen, so dass die Einsparung neutralisiert wird.

Mit einem Blackberry verwandten Gerät ist nun der Mitarbeiter theoretisch 24h für das Unternehmen verfügbar. In den USA gehören Personen zum Tagesbild, die im Fitnesscenter zwischen Übungen immer wieder zum Blackbberry greifen um noch Mails zu beantworten oder Mitarbeiter, die bereits auf dem Heimweg von der Arbeit noch schnell im Stau oder der Strassenbahn die letzten Aufgaben per Push E-Mail zu verteilen oder noch ihre Korrespondenzen erledigen. Diese gesteigerte Verfügbarkeit stellt teilweise einen Einbruch in das Privatleben des Arbeitnehmers dar. Klare Abgrenzungen und Selbstkontrolle müssen hier helfen, um eine Grenze zu ziehen und zu definieren wann der Arbeitstag aufhört.

Betreffend des internen Konkurenzkampfes und Aufstiegschancen sind hier auch wirtschaftliche Überlegungen in Betracht zu ziehen.

Auf der andern Seite ist ein viel genannter Grund für den Einsatz dieser Systeme, welche die die Verfügbarkeit so zu steigern vermögen, die Tatsache, dass anderenfalls, bei tiefer Verfügbarkeit, viel Koordinationsaufwand nötig wird, da Rückrufe getätigt werden müssen oder auf E-Mails geantwortet wird, die sich auf bereits veraltete Tatsachen beziehen. Möglichst alle Kanäle zu öffnen und immer verfügbar sein, vermag diese Probleme zu minimieren, bringt aber oben diskutierte Nachteile ein.

Drastisch kann dies am Statement eines IBM Angestellten zum Thema Blackberry verwendent werden.

Statement 1

> „Increased Customer Face Time: My on-demand job requires I be reachable by my customers 24x7x365 and my BB enables me to fulfil this requirement = improved customer satisfaction. My BB has better coverage than my cell phone too. My BB allows me to communicate and delegate faster than my cell phone (my opinion)[16].“ [37]

Hier wird die Gefahr solcher Systeme deutlich, denn Personen mit Hang zu unverhältnismässigen Arbeitszeiten werden sich leicht für ein solches Gerät begeistern.

Besonders selbsttändige Unternehmer und Manager in höchsten Positionen sind gewillt die 42 Stunden Woche deutlich überschreiten. Wozu solche Arbeitsexzesse führen können, ist bekannt, doch diese Personengruppe nimmt die Risiken auf sich, da sie eine mehr oder weniger angepasste Entlöhnung erfährt. Mit Worten: Sie bezahlen den Mehrverdienst mit ihrer Gesundheit oder mit dem Verlust von sozialen Kontakten und Kompetenzen.

Mit breitem Einsatz der Push E-Mail-Systeme besteht aber nun die Gefahr, dass auch Arbeiter der unteren und mittleren Lohnklassen, die bis anhin um 1630 das Büro verliessen und bis 0815 am nächsten Morgen nichts mehr mit dem Unternehmen zu tun hatten, so stark eingebunden warden, dass sie ihren Arbeitstag nicht an der Firmenporte abschliessen können, sonder immer wieder im Privatleben angestossen werden. Hier stellt sich dann vor allem ganz klar das Problem der Anrechenbarkeit: Verrechnung von Zusatzarbeit in der Freizeit muss klar geregelt sein. In Europa wird ein E-Mail in der Freizeit für den Arbeitgeber noch klar als Arbeit definiert, in den USA sind hier die Grenzen bereits stark verwischt, werden doch sogar interne Weiterbildungen häufig privat beglichen in Europa momentan ein undenkbares Szenario.

Gefordert ist für die Zukunft ein starkes individuelles Zeitmanagement, ein hohes Mass an Eigendisziplin und die Fähigkeit auch mal zu einer Aufgaben „nein“ sagen zu können.

6.4.2 Workaholic

Als Workaholic werden Arbeiter relativ schnell bezeichnet. Wo effektiv eine Grenze gezogen werden muss zwischen Fleiss und Karrierestreben und einer krankhaften Arbeitssucht ist sehr schwer zu definieren. Workaholic ist keinefalls ein zu vernachlässigendes Problem und ist als Krankheit akzeptiert. In Japan existieren bereits über 300 Behandlungszentren zur Therapie Süchtiger. Die Krankheit wird vor allem bei höheren Kadern auf und Personen in vorgesetzten Positionen, die sich ihre Arbeit mehr oder weniger frei einteilen können und nicht an die Geschäftszeiten gebunden sind, beobachtet. Selten fällt ein solches Verhalten bei Personen mit geregeltem Dienstplan auf. Einer Studie von Leitz zu Folge arbeitet eine Führungskraft in Deutschland im Schnitt 70 Stunden pro Woche, in Grossbritannien 60h, USA 58h, Frankreich 56h und Schweden 54h [31].

[16] Statement aus dem Blackberry Forum von IBM

Einen Viertel der Manager soll sich sogar mehr als 100 Stunden pro Woche für das Unternehmen einsetzen; nicht schwer zu verstehen, dass da eigentlich keine Zeit mehr für ein geregeltes Privatleben bleibt. Es muss bedacht werden, dass diese arbeitssüchtigen Angestellten in der Regel keinen verhältnismässigen ökonomischen Mehrwert schaffen. Besonders gegen Ende des Krankheitsverlaufes nimmt die Arbeitsleistung drastisch ab und Burnout Syndrome, Konzentrationsstörungen und schwerwiegende körperliche Probleme, die bis zum Tode reichen können, zeichnen sich ab. Push E-Mail als Verursacher zu sehen ist sicher unverhältnismässig, doch angesichts der folgenden Ausführungen eines Anwenders muss man solche Systeme kritisch betrachten und versuchen mit geeigneten Regeln einem Missbrauch vorzubeugen.

Statement 2

> „With the Blackberry, I am able to start my work day at home (without having to bootup my notebook) or on the road by checking overnight mail and responding to mail that needs immediate response. This keeps a task moving through our matrixed organization and complicated processes. Moving a purchase order through the crediting process at quarter end can be tough when you only check the mail at the end of the day (so you maintain facetime). Otherwise you loose selling days as you shepard the order through. Face time can be made more effective with the Blackberry. I have on numerous occations been at a customer meeting, sent a question to a specialist with the Blackberry, and received the answer while at the meeting. This keeps the opportunity moving...getting more done faster. This increases the „pace“ of workflow. There are less delays waiting for someone to respond to a note, then moving onto the next person [17].“ [37]

6.4.3 Fragmentierung der Arbeitszeit

Ein weiteres Problematik, die aus der immer ausgereifteren Technik und den damit verbundenen neuen Möglichkeiten erwächst, ist besonders für Personen aus den IT Bereich relevant. Neu muss nun beispielsweise ein Administrator, der für die E-Mail Server und Clients verantwortlich ist, sich auch noch um die ganze Anbindung der mobilen Geräte kümmern. Teilweise sind hier so komplexe Dienste möglich, dass ohne Probleme 2 weitere Stellen ausgeschrieben werden könnten. Grosse Firmen werden sicher auch diesen sinnvollen Weg beschreiten, doch für IT Personen in kleinen Unternehmen kann das bedeuten, dass sie neben ihrer anderen Tätigkeit sich auch noch die Fähigkeiten erarbeiten müssen, um diese Bereiche abzudecken. Allgemein könnte diese gesteigerten Anforderungen dazu führen, dass neben den Servern auch vom IT Personal eine 99.9999

Das bedeutet, dass die IT Fachfrau beispielsweise am Morgen viele Wartungsarbeiten an mobilen Geräten in Übersee ausführen kann und am Nachmittag, wenn alles rund läuft, ihren Hobbies nachgehen kann. Am frühen Abend sind dann wieder Einsätze in der

[17] Statement aus dem Blackberry Forum von IBM

Firma oder bei Kunden denkbar, die sich auch über die Nacht ausdehnen können. Mit anderen Worten: Es kann vorkommen, dass die Arbeitszeit von 10 Stunden einfach auf die 24 Stunden des Tages fragmentiert werden. Hier bestehen sicherlich Chancen, aber auch erhebliche Risiken für die Zukunft.

6.5 Zukunft Push E-Mail Dienste – Schlussfolgerungen

Push E-Mail birgt ein enormes Potential und hat sicherlich sehr gute Marktchancen. Der Einsatz im professionellen Umfeld ist in absehbarer Zeit der Bereich, in dem sich die Technologie schnell stark verbreiten wird und technische Verfeinerungen erleben wird. Die verschiedenen Hersteller sind dabei, den Markt unter sich aufzuteilen und die Tatsache, dass jeder ein positives Wachstum ausweisen kann, deutet darauf hin, dass noch genug Lücken vorhanden sind um Wachstumspotential an den Tag zu legen ohne andere Marktteilnehmer zu tangieren. Ähnlich wie die Mobiltelefone, die Mitte der 90er Jahre anfänglich sehr kostspielig, doch mit breiterer privater Nutzung drastisch im Preis gesunken sind, wird sich auch der Markt für die Push E-Mail Geräte entwickeln. SMS Nachrichten, ein auf Datenvolumen bezogen sehr kostspieliger Dienst, wird wohl im Zuge des Einsatzes von Push E-Mail verdrängt werden. Wie geschildert, birgt die Technik bei all ihren Vorteilen in ökonomischer Sicht auch auch entscheidende Gefahren und Risiken, besonders auf sozialer Ebene. Solche Probleme sind jedoch schon bei Mobiltelefonen zu beobachten und liegen weniger an der Technologie, als am Benutzer. Die weitere Entwicklung darf gespannt beobachtet werden und lässt sich angesichts der technischen Möglichkeiten, die sich aus der Idee ergeben, nur schwer prognostizieren [6].

Literaturverzeichnis

[1] Wikipedia: „Regulation of Investigatory Powers Act 2000“; http://en.wikipedia.org/wiki/RIP_Act, zuletzt besucht 20.12.2006 .

[2] Secunet: „SINA“; http://www.secunet.de/content.php?ln=1&text=k_sina_familie, zuletzt besucht 2.1.2007 .

[3] BSI: „Bundesamt für Sicherheit in der Informationstechnik“; http://www.bsi.bund.de/, zuletzt besucht 2.1.2007 .

[4] BSI: „Der Kryptoprozessor PLUTO“; http://www.bsi.de/literat/faltbl/F18PlutoKrypto.pdf, zuletzt besucht 1.1.2007 .

[5] Microsoft: „Demonstration Remote Wipe“; MMS://wm.microsoft.com/ms/Isn/windowsmobile/Final-MobilityV7-5Mbps.wmv, zuletzt besucht 2.11.2006 .

[6] Paula Rice: „Maintaining Work-Life Balance in 2010“; Paper, http://ieeexplore.ieee.org/iel5/9439/30282/01390866.pdf, Dezember 2005

[7] Jo Best und Jason Curtis: „BBC sperrt Blackberry-Dienst wegen Fehler“; http:www.zdnet.de/security/news/0,39029460,39137757,00.htm, October 2005

[8] TÜViT: „Prüfung kryptographischer Module nach FIPS 140-1/140-2“; http://www.tuvit.de/XS/c.010302/sprache.DE/seite.01/SX/ November 2005

[9] HackWatch: „GSM Simcard Emulator Released“; http://ireland.iol.ie/~kooltek/simcard.html, April 1998

[10] Volker Briegleb: „Microsoft sieht Wachstumschancen bei Windows Mobile“; http://www.heise.de/newsticker/meldung/79873, Oktober 2006

[11] Greg H. Gardella: „How Not to Get Squeezed“; Paper, http://ieeexplore.ieee.org/iel5/6/27087/01203088.pdf, Juni 2003

[12] Steve Ballmer: „3GSM World Congress“; Congress Keynote https://www.microsoft.com/presspass/exec/steve/2006/02-143GSM.mspx, zFebruar 2006

[13] msmobile: „Push technology in Windows Mobile 2003 : The Lie Revealed“; http://msmobiles.com/news.php/1870.html, Dezember 2003

[14] Microsoft: „Mobile Access Using Microsoft Exchange Server 2003“; White Paper, http://www.microsoft.com/exchange/techinfo/outlook/MobileAcc.doc, Juli 2003

[15] Microsoft: „Mobile Messaging with Exchange ActiveSync“; White Paper, http://www.microsoft.com/exchange/evaluation/features/mobileaccesswp.mspx, November 2006

[16] Johan Huss: „Windows Mobile Business Opportunity“; Präsentation, http://download.microsoft.com/download/9/f/4/9f45e30e-cff4-4991-8939-5b9d2013c543/MOB102.ppt , Datum Unbekannt

[17] Matthew Broersma: „Analyst blasts Windows Mobile security“; http://www.techworld.com/mobility/news/index.cfm?newsID=7223, Oktober 2006

[18] David Meyer: „Microsoft push-email row escalates“; http://news.zdnet.co.uk/communications/0,1000000085,39284628,00.htm, November 2006

[19] Seven: „Seven Acquiers Smartner, Increasing Global Presence and Positioning to Accelerate Growth“; Press release , http://www.seven.com/newseven/press_release.html, April 2005

[20] Eugene Signorini: „SEVEN's Serviced-Based Wireless Solutions Enable Enterprises to Untether E-Mail“; Survey, http://www.seven.com/downloads/YankeeGroup_SEVEN.pdf, Oktober 2004

[21] SEVEN: „Products: Consumer Edition“; http://www.seven.com/products/consumers/consumer_edition.html, zuletzt besucht 28.11.2006 .

[22] SEVEN: „Products: Personal Edition“; http://www.seven.com/products/consumers/personal_edition.html , zuletzt besucht 28.11.2006 .

[23] SEVEN: „Products: Server und Enterprise Edition“; http://www.seven.com/products/enterprises/enterprise_edition.html , zuletzt besucht 28.11.2006 .

[24] Timo Poropudas: „SEVEN takes two thirds of push mail operators“; http://www.mobilemonday.net/mm/story.php?story_id=4921, August 2006

[25] Nokia: „Mobility: An IT Perspective“; White Paper, http://europe.nokia.com/NOKIA_BUSINESS_26/Europe/Products/Mobile_Software/sidebars/pdfs/Mobility_An\%20IT\%20Perspective.pdf , 2006

[26] Nokia: „Security of the Intellisync Mobile Suite Platform“; White Paper, http://nds2.ir.nokia.com/NOKIA_COM_1/Press/Materials/White_Papers/pdf_files/Security_of_the_Intellisync_Mobile_Suite_Platform.pdf, Oktober 2005

[27] xdadev: „xdadev all unlock“; http://wiki.xda-developers.com/index.php?pagename=xdadev_all_unlock, Dezember 2005

[28] Silicon.de: „Palm OS verlier gegen Microsoft Mobile Edition“; http://www.silicon.de/enid/mobile_wireless/23650, November 2006

[29] Aaron Johnson: „Blackberry Market Share Quadrubles“; http://blackberryblog.com/2004/11/12/blackberry_market_share_quadrubles.html, November 2004

[30] Frost und Sullivan: „US Mobile Office Markets F650-65“; Whitepaper, http://www.goodmobilesecutity.com/Includes/F&S_Competitive_Whitepaper.pdf, 2005

[31] Roland Karle: „Auf dem Weg zum Workaholic“; http://inhalt.monster.de/2646_de-DE_p1.asp, Juli 2002

[32] Linda Himelstein: „PointCast: The Rise and Fall of an Internet Star“; http://www.businessweek.com/1999/99_17/b3626167.htm, April 1999

[33] Blackberry Homepage, http://www.blackberry.com/de/, zuletzt besucht 1.1.2007 .

[34] Research in Motion Homepage, http://www.rim.net/, zuletzt besucht 1.1.2007 .

[35] Martin Bayer: „Wirft Audi seine Blackberrys raus?“; http://www.computerwoche.de/index.cfm?pid=380&pk=557254, Juni 2005

[36] Good Technologies (Motorola) Homepage, http://www.good.com/corp/index.php, zuletzt besucht 12.12.2006 .

[37] Blackberry Forum IBM (Intranet), Kommentare von Benutzern zum Blackberry Dienst.

[38] Good Technology, Inc.: „Good Security“; Whitepaper, http://www.good.com/corp/uploadedFiles/Documentation/WhitePapers/Good_Security_WP.pdf, 2006

[39] Good Technology, Inc.: „Good Mobile Messaging“; Whitepaper http://www.good.com/corp/uploadedFiles/Documentation/WhitePapers/Good_Mobile_Messaging_WP.pdf, 2006

[40] Joachim Merz, Paul Böhm and Derik Burgert: „Timing, Fragmentation of Work and Income Inequality - An Earnings Treatment Effects Approach“; Paper, http://www.iariw.org/papers/2004/merz.pdf, 2004 .

[41] P. Karn et al.: „The ESP Triple DES Transformation“; http://rfc.net/rfc1851.html, September 1995

[42] J. Schaad: „Use of the Advanced Encryption Standard (AES) Encryption Algorithm in Cryptographic Message Syntax (CMS); http://www.rfc-archive.org/getrfc.php?rfc=3565, Juli 2003

[43] Gordon E. Moore; „Moore's Law“; http://www.intel.com/technology/mooreslaw/index.htm, Datum Unbekannt

[44] Audi: „Audi Deutschland“; http://www.audi.de/audi/de/de2.html, zuletzt besucht 5.1.2007 .

[45] Charmaine Eggberry: „Corporate Statement 11.Juni 2005“; Corporate Statement, http://www.wireless-technologies.de/srpa/bb-portal.nsf/Corporate_statement_13_June_2005%20_final_Deutsch.pdf, Juni 2005

[46] T. Dierks et al.: „The TLS Protocol Version 1.0“; http://www.ietf.org/rfc/rfc2246.txt, Januar 1999

[47] Wikipedia: „Man-in-the-Middle attack“; `http://en.wikipedia.org/wiki/Man_in_the_middle_attack`, zuletzt besucht 2.11.2007 .

[48] Wikipedia: „Denial-of-Service attack“; `http://en.wikipedia.org/wiki/Denial_of_service`, zuletzt besucht 18.1.2007 .

Kapitel 7

Intrusion Detection in Wireless and Ad-hoc Networks

Raphaela Estermann, Richard Meuris, Philippe Hochstrasser

Der stetig zunehmende Einsatz von kabellosen und mobilen ad-hoc Netzwerken stellt wesentlich höhere Herausforderungen an die heute eingesetzten Intrusion Detection Systeme (IDS), welche eine höhere Netzwerksicherheit gewährleisten sollen. Zuerst wird in dieser Arbeit eine kurze Übersicht über verschiedene Typen von IDS gegeben werden, dann werden konkrete Intrusion Detection Lösungsansätze angesprochen, welche versuchen die anfangs beschriebenen wireless- und MANET-spezifischen Probleme zu lösen.

Inhaltsverzeichnis

7.1 Einleitung

Seit es Computersysteme gibt, war auch stets deren Schutz ein essentielles Thema [17]. Insbesondere vernetzte (teilweise sogar vom Internet erreichbare) Systeme sind anfällig für Angriffe. Dies gilt vor allem für die im Kapitel "Wireless und Ad-hoc Netzwerke" besprochenen kabellosen und mobilen ad-hoc Netzwerke [3][30]. Zu schützende Eigenschaften eines Netzwerkes sind Daten-Sicherheit (information security), Kommunikations-Sicherheit (communications security) und Schutz des Equipments (physical security). Hierbei kann jeder dieser drei Bereiche wiederum hauptsächlich unter vier Aspekten betrachtet werden: Authentizität, Vertraulichkeit (confidentiality), Integrität (integrity) und Verfügbarkeit (availability).

Ein Sicherheitssystem für mobile ad-hoc Netzwerke z.B. enthält diverse Sicherheitsmechanismen: Die fehlertolerante Funkschnittstelle als erstes ist gegen physikalische Störungen der Kommunikation geschützt. Als zweites hat das abgesicherte Routing die Authentizität des Absenders und Korrektheit der Routinginformationen sicherzustellen. Das Intrusion[1] Detection System (IDS[2]) kann schliesslich komplementär z.B. parallel zur Firewall und der Systemsicherheit und nebst der vor allem gegen externe Attacken wirksamen Intrusion Prevention (Verschlüsselung, Authentifizierung...) als zweiter Verteidigunswall (second line of defense) operieren. Das IDS soll Angreifer aufspüren, indem verdächtige Aktivitäten an der Firewall[3] und im System genauer analysiert werden [30].

In verkabelten Netzwerken war es noch möglich, durch präventive Massnahmen eines Intrusion Prevention Systems (IPS) wie Verschlüsselung und Firewall ein gutes Mass an Sicherheit zu gewährleisten [26]. Diese präventiven Massnahmen wurden zusammenfassend auch als erster Defensivwall (first line of defense) bezeichnet. Der in den letzten Jahren stark gestiegene Einsatz kabelloser Netzwerke (wie z.B. wireless und mobile ad-hoc Netzwerke (MANET) oder noch ressourcenschwächerer kabelloser Sensor-Netzwerke (WSN)) stellt nun aber wesentlich höhere Anforderungen an die eingesetzten Sicherheitssysteme [30]: So reicht ein IPS alleine v.a. in MANETs nicht mehr aus, da z.B. MANETs im Gegensatz zu verkabelten Netzwerken aufgrund ihrer dynamischen Topologie grundsätzlich nicht über konstante zentrale Datenverkehrsknotenpunkte (wie Switches, Router, Gateways etc.) verfügen, welche mit Schutzsystemen wie Firewalls, Monitoring etc. ausgestattet werden könnten [29]. Ein daher eingesetztes IDS besitzt folgende Kernaufgaben: Als erstes werden die System- und Nutzeraktivität überwacht, indem sogenannte "Audit"[4]

[1] Andersons Publikation aus dem Jahre 1980 definierte "**Intrusion**" folgendermassen: "The potential possibility of a deliberate unauthorized attempt to access information, manipulate information or render a system unreliable or unusable." [2] Heberlein definiert dies sehr ähnlich: Eine Menge von Handlungen, deren Ziel es ist die Integrität, Verfügbarkeit oder Vertraulichkeit eines Betriebsmittels zu kompromittieren [12].

[2] Im grossen Stil wurden Intrusion Detection Systeme erstmals vor allem in den 80er Jahren z.B. von AT&T eingesetzt (AT&T setzte das IDS ein, um mit möglichst geringem Aufwand damals illegal eingesetzte "Blue Boxes" (welche AT&T-intern genutzte Signale erzeugen konnten) aufzuspüren, mit welchen rechtswidrig kostenlose Anrufe getätigt werden konnten [17].

[3] Die Aufgabe von **Firewalls** ist es, den Zugriff von aussen auf bestimmte, als sicher angesehene Applikationen zu erlauben und alle anderen Zugriffe zu verweigern. Firewalls nutzen hierbei das Konzept von TCP/IP, dass Applikationen über definierte Ports mit der Aussenwelt kommunizieren. Die Firewall erlaubt Zugriffe von aussen nur auf gewisse Ports.

[4] **Audit Daten** werden auf verschiedenen Ebenen im Netzwerk und im Computer gesammelt.

Daten verschiedener Quellen (Betriebssystem, Applikation, Netzwerk) gesammelt werden. Diese Audit Daten werden dann ausgewertet, um externe und interne (von einem Benutzer mit direktem Zugang zum System), passive (z.B. Datenspionage) und aktive Attacken (z.B. Veränderung von Daten) Angriffe auf das kabellose Netzwerksystem erkennen zu können [30].

Die Hauptaufgabe eines IDS besteht nun darin, kompromittierte Knoten (auch bösartige Knoten genannt) eines Netzwerkes zu identifizieren und gegebenenfalls mittels eines Intrusion Response Systems (IRS) Konsequenzen aus den entsprechenden Befunden zu ziehen [3] [30]. Das IDS wird (evtl. zusammen mit dem IRS) auch als zweiter Defensivwall (seconde line of defense) bezeichnet, welcher in einem kabellosen Netzwerk (vor allem gegen interne Attacken) viel wirksamer ist als das IPS allein und dieses somit komplementieren kann und in vielen Fällen die letzte Abwehrfront gegen Eindringlinge darstellt [10]. IDS lassen sich host- oder/und netzwerk-basiert realisieren. Hostbasierte IDS zeichnen sich dadurch aus, dass jedes zu überwachende System mit eigenen Sensoren ausgestattet wird. Bei netzwerkbasierten IDS werden hingegen meist eine geringere Anzahl von Sensoren im Netzwerk selber eingesetzt. Je nach Sicherheitsbedarf und Netzwerkfunktionsweise werden signatur-, anomalie- oder spezifikationsbasierte IDS oder Kombinationen davon eingesetzt, welche einerseits effektiv, aber andererseits aufgrund der genannten Einschränkungen der verfügbaren Ressourcen mobiler Knoten auch effizient sein müssen [20].

Auf diese angesprochenen verschiedenen Typen von IDS wird im allgemein gehaltenen Kapitel "Methodik und Architektur von IDS" inkl. Behandlung derer Vor- und Nachteile noch genauer eingegangen. Unabhängig von Funktionsweise und Aufbau sollte ein IDS fehlertolerant, skalierbar, korrekt reagierend, energiesparsam, interoperabel (mit anderen IDS) und ausserdem im Idealfall in der Lage sein, auch bis anhin unbekannte Angriffe identifizieren zu können. Ein IDS sollte die auf den mobilen Knoten eher geringen vorhandenen Ressourcen sparsam nutzen: Energiespeicher, Rechenkapazität, Speicherplatz und Kommunikation sollten vorwiegend für den Nutzer reserviert bleiben (bei geringer Dynamik eines MANETs sollte ein gut programmiertes IDS zusätzlich seinen Energiebedarf senken). Das IDS sollte ferner erweiterbar sein, da ständig neue Attacken auftreten können. Ferner sollte das IDS davon ausgehen, dass in der Startphase eines mobilen ad-hoc Netzwerkes (im Gegensatz zu einem verkabelten Netzwerk) keine initiale Vertrauensbeziehung zwischen den Teilnehmern vorliegt. Bis heute (2007) existieren noch keine ausgereiften IDS-Standardprodukte für MANETs auf dem Markt. Momentan wird noch viel mittels Simulationen geforscht, um die IDS zu verfeinern. Doch es existieren bereits gut funktionierende konkrete Lösungsansätze zur Behebung der im Kapitel "Wireless und Ad-hoc Netzwerke" beschriebenen durch wireless- und MANET-Attacken verursachten Probleme. Diese Lösungsansätze werden im Kapitel "Lösungsansätze" vorgestellt: Ansätze wie Watchdog, Pathrater und Routeguard stellen beispielsweise eine Erweiterung des ebenfalls in jenem Kapitel besprochenenen, in kabellosen Netzwerken sehr häufig eingesetzten Dynamic Source Routing (DSR) Protokolles dar. Watchdog soll unter anderem für die korrekte Weiterleitung von Paketen sorgen und egoistische Knoten zur Kooperation im Netzwerk motivieren [17]. Pathrater bewertet die einzelnen Knoten gemäss ihres Verhaltens mit einer Zahl, die sich über die Zeit ändern kann. In Routeguard wird eine etwas differenziertere Klassifikation der Knoten vorgenommen, welche insbesondere neu hinzugekommene mobile Knoten berücksichtigen soll. Zone-based Intrusion Detecion Systeme partitionieren insbesondere topologisch dynamische Netzwerke wie MANETs in Zonen und ernennen

darin Gatewayknoten, welche lokal registrierte verdächtige Aktivitäten aggregieren und auf diese Weise Alarme generieren. Das Kapitel "Ausblick" soll diese Arbeit mit einem Ausblick auf eine mögliche weitere Entwicklung von Intrusion Detection System in kabellosen Netzwerken abschliessen. Ferner wird darin auf die Problematik simulationsbasierter Tests eingegangen, welche zwecks IDS-Beurteilung durchgeführt werden.

7.2 Wireless und Ad-hoc Netzwerke

Drahtlose Netzwerke lassen sich grob in infrastrukturbasierte Netze und Netze ohne Infrastruktur (Ad-hoc Netze) einteilen. Es gibt jedoch auch Mischformen, welche beispielsweise einige Dienste über die Infrastruktur abwickeln, jedoch direkt miteinander kommunizieren. Wireless und Ad-hoc Netzwerke werden durch verschiedene Charakteristiken geprägt, die sich von Wired Netzwerken deutlich unterscheiden. Auf die resultierende Problematik sei in einem weiteren Abschnitt näher eingegangen[22].

7.2.1 Übersicht zu Wireless Netzwerken

In Infrastrukturbasierten Netzen findet die Kommunikation zwischen den drahtlos angebundenen Endgeräten und einem Zugangspunkt (access point) statt. Der Zugangspunkt steuert den Medienzugriff und ermöglicht eine Kommunikation auch zwischen verschiedenen Funknetzen. Ein Nachteil dieser Art von Netzen ist sicherlich die eingeschränkte Flexibilität. Ein Vorteil davon ist jedoch, dass die meiste Komplexität im Zugangspunkt liegt und nicht in den Endgeräten. Anwendung findet Wireless in Firmennetzwerken, an öffentlichen Plätzen oder in Heimnetzen.

7.2.2 Übersicht zu Ad-hoc Netzwerken

In Ad-hoc Netzwerken können die Endgeräte direkt miteinander kommunizieren, sofern sie innerhalb des Übertragungsbereichs liegen oder die Daten durch weitere Endgeräte weitergeleitet werden können. Es wird kein Zugangspunkt zur Zugriffssteuerung auf das Medium benötigt, das Routing erfolgt auf peer-to-peer Basis. Ein Nachteil von Ad-hoc Netzwerken liegt in der Komplexität der Endgeräte und deren eingeschränkten Ressourcen, wie beispielsweise Reichweite, Bandbreite, Batterielebensdauer, CPU oder Memory. Ein Vorteil ist jedoch die Flexibilität und die Möglichkeit eines raschen Aufbaus, beispielsweise in Katastrophenfällen. Anwendung findet diese Art von Netzwerken in Kampfeinsätzen, Rettungsmissionen, aber auch bei Konferenzen und im Klassenunterricht.

Ad-hoc Netzwerke können nach verschiedenen Aspekten unterteilt werden

- Kommunikation
- Netztopologie

- Knotenaustattung

Die **Kommunikation** kann aus single-hop oder multi-hop Verbindungen bestehen. Bei single-hop Verbindungen kommunizieren die Endgeräte direkt miteinander, bei multi-hop Verbindungen können Daten über zusätzliche Knoten weitergeleitet werden.

Die **Netztopologie** kann flach oder vielschichtig sein. In einer flachen Infrastruktur sind alle Knoten gleichwertig und beteiligen sich am Routing. In vielschichtigen Netzwerken werden die Knoten nicht als gleichwertig angesehen. Die Knoten innerhalb eines Übertragungsbereichs werden in Gruppen (Cluster) zusammengefasst. Die Knoten wählen einen Cluster-Head, der das Routing für das Cluster zentral regelt[3].

Die **Knotenausstattung** kann homogen sein, d.h. alle Knoten haben die gleichen Eigenschaften bezüglich Hardware, oder heterogen, wobei nicht alle Knoten die gleichen Dienste erbringen können[21].

Es gibt unterschiedliche Vertreter von Ad-hoc Netzwerken.

- Mobile Ad-hoc Netzwerke
- Wireless Sensor Netzwerke

MANETs bestehen aus mobilen Geräten wie Mobiltelefone, Laptops oder Kleincomputer. Die Knoten können Daten senden und empfangen und dienen gleichzeitig als Router. Die Topologie solcher Netze ist dynamisch und verlangt eine hohe Komplexität der Endgeräte. Anwendung finden MANETs in Katastrophenszenarien, Home-Office oder E-Learning[21].

Wireless Sensor Netzwerke bestehen aus kleinen oft batteriebetriebenen Knoten. Neben Funkschnittstelle und Prozessor bestehen die Knoten aus einer Anzahl von Sensoren, welche physische Faktoren oder Umgebungsbedingungen überwachen können. Die Knoten können sich selbstständig organisieren und drahtlos miteinander kommunizieren. Entwickelt wurden WSNs ursprünglich für militärische Einsätze, jedoch werden sie heute auch für viele andere Szenarien eingesetzt. Beispielsweise um seismische Aktivitäten zu erkennen, für die Logistik oder den Privatgebrauch.

7.2.3 Probleme von Ad-hoc und Wireless Netzwerken

Die Problematik von Wireless und Ad-hoc Netzwerken unterscheidet sich von der in Wired Netzwerken. Auf einige der Probleme wird hier näher eingegangen.

Probleme

- offenes Medium
- dynamisch ändernde Topologie

- kooperative Algorithmen
- kein zentrales Monitoring
- keine klare line-of-defense

Bei drahtlosen Systemen können keine Drähte zwischen den verschiedenen Sendern und Empfängern gezogen werden. Das Medium muss daher auf eine bestimmte Weise zwischen den Endknoten verteilt werden. Zusätzlich zu den Zugriffsverfahren ist auch der Sicherheitsaspekt nicht zu vernachlässigen, da das Medium einfacher und unbemerkter abgehört werden kann.

Nahezu alle Knoten in Wireless und Ad-hoc Netzen sind mobil und können sich beliebig bewegen, d.h. die Topologie verändert sich dynamisch. Es ist daher schwierig festzustellen, ob ein Knoten, der falsche Routinginformationen liefert, böswillig handelt oder lediglich nicht richtig synchronisiert.

Kooperative Algorithmen werden in Ad-hoc Netzwerken gebraucht um die Datenübertragung zwischen den verschiedenen Knoten zu koordinieren. Da es meist keine zentrale Instanz gibt, nutzen die Gegner diese Verwundbarkeit für neue Angriffe mit dem Ziel den Algorithmus zu brechen. Der Vorteil davon ist, dass bestimmte Knoten ihre Ressourcen schonen können und nicht oder nur bedingt am Routing-Prozess teilnehmen müssen.

Jeder Knoten kann nur die Übertragungen in Funkreichweite überwachen, d.h. es gibt kein zentrales Monitoring. Dies erschwert die Erkennung von Intrusions enorm, da an keiner Stelle alle Daten zusammengefasst werden können.

MANETs und Ad-hoc Netzwerke generell haben keine klar definierte physikalische Grenze und daher auch keine spezifischen Eingangspunkt, der entsprechend geschützt werden kann[30].

7.2.4 Attacken in Mobile Ad-hoc Netzwerken

Attacken sind Vorfälle oder Angriffe, die den sicheren Betrieb von Netzwerken gefährden. Die Anforderungen an ein sicheres Netzwerk werden nachfolgend aufgezählt und beschrieben.

Sicheres System

- Vertraulichkeit
- Verfügbarkeit
- Integrität
- Authentizität
- Nicht-Abstreitbarkeit

Vertraulichkeit garantiert die Nicht-Enthüllung von persönlichen Informationen, die in einem Knoten gespeichert sind oder übertragen werden.

Verfügbarkeit garantiert dass Netzwerk Services für authorisierte Entitäten regelmässig zugänglich sind.

Integrität garantiert die Korrektheit der Daten und dass sie während der Übertragung bzw. in gespeicherter Form nicht verändert werden.

Authentizität verifiziert die Identität eines Knotens. Dies kann über verschieden Ansätze, wie beispielsweise physikalische Token, Aufenthaltsorte oder Passwörter erreicht werden.

Nicht-Abstreitbarkeit ist die Fähigkeit zu beweisen, dass ein bestimmter Sender eine Nachricht verschickt hat[26].

Nachfolgend werden einige Attacken in MANETs kurz beschrieben.

Attacken in Manets

- Spoofing
- Packet Drop
- Tunneling
- Resource Depletion
- Selective Existence
- False Message Propagation
- Misrouting
- Man-in-the-Middle
- Sniffing

Spoofing dient der Vortäuschung einer falschen Identität. Es gibt sie in verschiedenen Formen, wie IP-Spoofing, MAC-Spoofing oder TCP-Spoofing. Diese Angriffe auf die Authentizität und Nicht-Abstreitbarkeit kann beispielsweise durch das Intrusion Detection Tool AODVSTAT erkannt werden. Sequenznummern wachsen normalerweise linear mit dem Ansteigen des Datenverkehrs, durch das Identifizieren von Anomalien bezüglich Sequenznummern für spezifische MAC Adressen, kann ein IDS MAC Spoofing erkennen.

Packet Drop steht für das Verwerfen von Datenpaketen oder Kontrollnachrichten, also ein Angriff gegen die Verfügbarkeit. Ein IDS führt eine Liste mit all ihren Nachbarn, welche ständig aktualisiert wird. Route Request Pakete sind Broadcast Pakete, auf welche benachbarte Knoten mit einer Route Reply antworten, oder den Route Request weiterschicken. Falls nun ein Knoten bei einem Broadcast nicht mit einer RREP oder einer RREQ antwortet, entdeckt das IDS die Attacke.

Bei einer **Tunneling**-Attacke werden Pakete abgefangen und versteckt an andere Knoten weitergeschickt. Diese Attacke greift die Verfügbarkeit des Systems an.

Bei einer **Resource Depletion** Attacke verschickt ein böswilliger Knoten unzählige Daten- und Kontrollpakete. Auch dieser Angriff geht gegen die Verfügbarkeit der Netzwerkdienste. Ein IDS zählt die Anzahl Pakete, die es von jedem Knoten erhält, für eine bestimmte Zeitdauer. Überschreitet diese Zahl ein gewisses Limit, wird ein Alarm signalisiert.

Selective Existence bedeutet, dass sich ein Knoten nur eingeschränkt am Routing beteiligt und daher die Verfügbarkeit des Netzwerks angreift. Der böswillige Knoten verhält sich egoistisch, indem sie das Netzwerk nur für ihre eigenen Bedürfnisse missbraucht. Da der Knoten keine HELLO-Messages verschickt und nur phasenweise in Erscheinung tritt, kann auch diese Attacke durch ein IDS erkannt werden.

False Message Propagation gibt es in verschiedenen Formen. Einerseits als Veränderung von Sequenznummern oder Hop-counts oder aber als Umleitungsattacke. Ein böswilliger Knoten leitet den Verkehr durch die Erhöhung von Sequenznummern in ihre Richtung, da jeweils die höchste Sequenznummer für eine Route gewählt wird. Jeder Sensorknoten eines IDS kann diese Attacke gegen die Integrität entdecken, wenn die Sequenznummer einen gewissen Wert überschreitet.

Misrouting bedeutet, dass ein böswilliger Knoten Datenpakete ans falsche Ziel schickt. Entweder durch das Weiterleiten an einen falsch benachbarter Knoten oder durch Verändern der Zieladresse im Paket. Dieser Angriff zielt gegen die Integrität und kann durch Testpakete vom IDS erkannt werden.

Man-in-the-Middle Angriffe können durch Kombination von Spoofing und Verwerfen von Paketen erreicht werden. Das Ziel ist die Kontrolle über die Kommunikation zweier Knoten im Netzwerk.

Bei **Sniffing**-Attacken werden Datenübertragungen abgehört und somit auf die Vertraulichkeit gezielt[26].

7.2.5 Lösungsansätze in Wireless und Ad-hoc Netzwerken

Um drahtlose Netzwerke gegen Angreifer zu schützen müssen die existierenden Sicherheitmechanismen erweitert werden. Neben Intrusion Prevention muss auch das Intrusion Detection System auf die drahtlose Umgebung angepasst werden.

Sicherheitsmechanismen

- Intrusion Prevention
- Intrusion Detection

Zu Intrusion Prevention gehören Firewalls und Encryption Software. Intrusion Detection Systeme werden im nächsten Kapitel noch detaillierter behandelt.

7.2.6 Zusammenfassung

In diesem Kapitel wurden drahtlose Netzwerke vorgestellt und charakterisiert. Neben infrastrukturbasierten Netzwerken, die über einen Access Point Daten austauschen, gibt es auch Ad-hoc Netzwerke, in welchen die Endgeräte direkt miteinander kommunizieren. Ad-hoc Netzwerke können nach unterschiedlichen Kriterien klassifiziert werden. Vertreter davon sind Wireless Sensor und Mobile Ad-hoc Netzwerke. Drahtlose Netzwerke unterscheiden sich bezüglich Kommunikationsmedium, Topologie und Routing deutlich von verdrahteten Netzwerken. Diese Unterschiede müssen auch bei der Entwicklung von Sicherheitsmechanismen berücksichtigt werden, da die drahtlosen Netze anfälliger gegenüber Störungen sind und entsprechend geschützt werden müssen. Zur Veranschaulichung der Problematik wurden Angriffe am Beispiel von Manets aufgezeigt, wie auch die Wirksamkeit der Intrusion Detection Systeme. Im nächsten Kapitel wird noch näher auf die Methodik, Architektur und Klassifikation von IDS eingegangen.

7.3 Methodik und Architektur von Intrusion Detection Systemen

In kabellosen Netzwerken kommt erschwerend hinzu, dass sich eine Differenzierung zwischen normalem und abnormalem Verhalten eines Knotens viel schwieriger gestaltet, denn es existieren auch andere Gründe für fehlbares Verhalten (Egoismus, Überlastung, Fehler etc.) [3] [30]. Ein in kabellosen Netzwerken eingesetztes IDS muss daher die durch seine Sensoren gesammelten Daten viel differenzierter analysieren können [10]: Ein IDS in kabellosen Netzwerken liefert so nebst IP-Adresse des Angreifers und Art des Angriffs noch weitere Informationen wie Sensor- und Wireless Access Point-Standort [27]. Die heute eingesetzten IDS lassen sich abhängig vom Ort der eingesetzten Sensoren in host- und netzwerk-basierte IDS klassifizieren: Bei host-basierten IDS wird auf jedem zu überwachenden System ein eigener Sensor eingerichtet. Bei netzwerk-basierten IDS hingegen werden meist eine kleinere Anzahl von Sensoren im Netzwerk selbst eingesetzt. Ferner können die heute eingesetzten IDS je nach Erkennungsmethode in signatur-, anomalie- und spezifikationsbasierte IDS klassifiziert werden, wobei auch hybride Varianten (i.d.R. Kombinationen aus signatur- und anomaliebasierten Methoden) zum Einsatz kommen können, welche in diesem Kapitel im Anschluss an die Beschreibung von host- und netzwerkbasierten IDS erläutert werden [8].

In flachen Netzwerkinfrastrukturen (flat network infrastructure) werden alle Knoten des Netzwerks als äquivalent betrachtet: Server und eine grössere Menge von Clients sind ohne jegliche Hierarchie beispielsweise über Hubs an einen Router angeschlossen. Bei vielgeschichteten hierarchischen Netzwerkinfrastrukturen (multi-layered and hierarchical network infrastructure) jedoch werden nicht alle Knoten als gleichwertig betrachtet, was vor allem bei grossen kabellosen Netzwerken vorteilhaft sein kann. Es werden in diesem Kapitel drei gängige IDS Architekturen beschrieben. Die Stand-alone IDS Architektur und die verteilt kooperative IDS Architektur basieren auf flachen Netzwerkinfrastrukturen, währenddem zentralisierte IDS Architekturen auf hierarchischen Netzwerkinfrastrukturen aufbauen: Individuelle Sensoren sammeln und leiten dabei ämtliche Audit-Daten an ein

zentrales Managementsystem weiter, in welchem die Daten gespeichert und verarbeitet werden [10]. Der Begriff IDS umfasst häufig beide Basisfunktionalitäten: Das IDS im engeren Sinne soll Attacken erkennen, das Intrusion Response System (IRS) stellt die definierbare Reaktion auf erkannte Angriffe dar. Bei Registrierung von Angriffen und nach der Alarmauslösung können durch das IRS die definierten Gegenmassnahmen eingeleitet werden. Möglichkeiten von Reaktionen des IRS werden am Ende dieses Kapitels kurz besprochen [18].

7.3.1 Klassifikation der IDS gemäss Einsatzort der Sensorik

Intrusion Detection Systeme (IDS) lassen sich gemäss des Standortes der zur Datensammlung (audit data collection) eingesetzten Sensoren klassifizieren: Je nach Einsatzort der IDS-Sensorik wird so zwischen host-basierten Intrusion Detection Systemen (HIDS) und netzwerk-basierten Intrusion Detection Systemen (NIDS) unterschieden [3] [1] [27] [8]. Bei HIDS werden dabei auf jedem Host eigene Sensoren eingerichtet. Bei NIDS hingegen ist die Sensorik nicht auf jedem Host vorhanden, sondern wird nur an einigen möglichst geeigneten Stellen im Netzwerk eingesetzt, an denen grundsätzlich besonders gut der Datenverkehr des Netzwerkes überwacht werden könnte. Jedoch gestaltet sich das Ermitteln solcher variabler Stellen besonders in MANETs aufgrund ihrer dynamischen Topologie als besonders schwer. Ein IDS muss wohlgemerkt nicht vollständig einer dieser beiden Varianten angehören, sondern es kann aus Sicherheitsgründen auch ein effektiveres hybrides IDS verwendet werden, welches aus sowohl host- als auch netzwerkbasierten Sensortypen besteht, welche an ein zentrales Management System angeschlossen sind.

Host-basiertes Intrusion Detection System (HIDS)

Die Sensoren des HIDS werden hier auf jedem zu überwachenden System installiert und müssen kompatibel zum jeweiligen Betriebssystem sein [3] [1] [27] [8]. Als Host wird in diesem Kontext jedes System, auf welchem ein Sensor installiert ist, bezeichnet. Solche auf jedem Host eingesetzten Sensoren untersuchen Daten des Betriebssystems, System- und Anwendungslogbücher usf., also den Zustand der entsprechenden Hosts. Zu den Vorteilen eines HIDS gehört die Möglichkeit einer umfassenden Überwachung von Hosts, da jeder über eigene Sensoren verfügt. Dementsprechend können genauere Aussagen über den Zustand von attackierten Systemen gemacht werden (z.B. können sehr gut verdächtige Systemdatenmodifikationen erkannt werden). Zu den Nachteilen zählt der hohe Aufwand der Installation und des Unterhalts eines eigenen Sensors auf jedem Host. Falls ferner ein Host kompromittiert wird, dann wird zwangsläufig die darauf installierte Sensorik ebenfalls kompromittiert, wodurch man sich nicht mehr auf die von den Sensoren retournierten Daten verlassen kann.

Netzwerk-basiertes Intrusion Detection System (NIDS)

IDS werden idealerweise an Stellen mit vergleichsweise hohem Datenverkehrsaufkommen eingesetzt (in MANETs wie erwähnt schwer bestimmbar, da keine konstant zentralen Da-

tentransferpunkte existieren) und haben die Aufgabe, im Idealfall alle Pakete im Netzwerk aufzuzeichnen (was aufgrund hoher Datenmengen und Verschlüsselungsanwendung in der Praxis aber zumeist nicht realisierbar ist), zu analysieren und verdächtige Aktivitäten zu melden [3] [1] [27] [8]. Mit einem einzigen Sensor können dabei Pakete eines ganzen Netzwerksegmentes überwacht werden, was kostengünstiger ist, als wie bei HIDS auf jedem Host einen eigenen Sensor zu installieren. Das NIDS versucht z.B. aus dem Datenverkehr Angriffsmuster insbesondere externer Attacken zu erkennen. Ein Vorteil ist, dass sich mit einem NIDS Attacken erkennen lassen, welche keine Modifikationen an den Hostdaten vornehmen (z.B. Veränderungen an Paketstatusinformationen) oder Netzwerkbandbreitenschmälerungen herbeiführen (z.B. Denial-of-service Attacken) [17].

7.3.2 Klassifikation der IDS gemäss verwendeter Methodik

Es existieren verschiedene Arten, wie Ereignisse erkannt werden können, welche im Zusammenhang mit unerlaubtem Eindringen stehen. IDS werden je nach verwendeter Methoden als signatur-, anomalie-, oder spezifikationsbasierte IDS klassifiziert. Bei signaturbasierten IDS werden die Angriffe, bei anomaliebasierten IDS der Normalzustand des Systems beschrieben. Die spezifikationsbasierte Erkennung ist der jüngste Ansatz der Intrusion Detection: Hier wird das zulässige Verhalten definiert und jede Abweichung davon als kritisches Ereignis bewertet. Es ist natürlich auch der Einsatz hybrider Systeme (engl. compound detecion) möglich, welche mehrere Methoden gleichzeitig einsetzen (in der Regel eine Kombination aus signatur- und anomaliebasierten IDS). Dies gestaltet sich zwar als aufwendiger, aber sicherer (bessere Erkennungsrate und höhere Qualität der Alarmentscheide). Es ist sogar so, dass ein rein anomaliebasiertes IDS relativ selten eingesetzt wird [3] [30] [26] [13] [20] [17]. Nachdem folgend auf die zwei verschiedenen Typen von vorkommenden Fehler bei der Angriffserkennung eingegangen wird, werden die drei genannten Methoden mit ihren Vor- und Nachteilen kurz vorgestellt.

Typ I/II Fehler

Es werden zwei Arten von Fehlern bezüglich Erkennung von Angriffen unterschieden. Typ I-Fehler stellen Falschalarme dar [23]. Die Gefahr bei diesen Falschalarmen liegt v.a. darin, dass bei einer langanhaltend hohen Falschalarmrate (false positive rate) aufgrund der vor allem bei MANETs problematischen, verschwendeten Rechenleistung z.B. vom Administrator die Sensitivität des IDS reduziert wird, was dann aber dazu führen kann, das die wirklich vorkommenden Angriffe nicht mehr erkannt werden. Typ II-Fehler sind wesentlich gefährlicher als Fehler des ersten Typs, da sie die nicht erfolgte Erkennung von Angriffen darstellen: Eine hohe Rate von Fehler des zweiten Typs (false negative rate) ist auf eine zu geringe Sensitivität zurückzuführen. Um die Performance bezüglich Erkennungsrate von Attacken zu vergleichen, ist es nötig die Systeme bezüglich ihrer Sensibilität so zu konfigurieren, dass die Häufigkeit von Typ I-Fehlern der Häufigkeit von Typ II-Fehlern entspricht (vergleiche Abbildung 7.1 unten). Dieses Gleichgewicht wird auch als Crossover Error Rate (CER) bezeichnet. Je nach sicherheitstechnischen Prioritäten kann nun z.B. das System mit dem niedrigsten CER-Wert ausgewählt werden. Wird verlangt, dass möglichst alle Attacken entdeckt werden, wird auf ein IDS mit möglichst niedriger

Typ II-Fehlerrate zurückgegriffen, also ein System mit sehr hoher Sensitivität, aber auch entsprechend hohem Rechenaufwand.

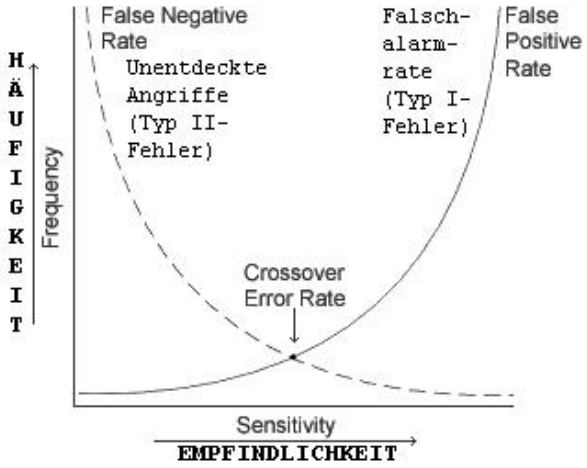


Abbildung 7.1: Typ I- und Typ II- Fehler sind invers proportional zueinander. Um die Effektivität bezüglich Angriffserkennung verschiedener IDS miteinander zu vergleichen, werden sie so konfiguriert, dass sich die Häufigkeiten beider Fehlertypen im Gleichgewicht befindet (sogenannter CER-Punkt) [23].

Signaturbasierte IDS

Signaturbasierte IDS (engl. signature- oder misuse-based IDS) sind in der Praxis (wohl auch wegen ihrer verglichen mit anderen Varianten einfacherer Methodik) am meisten verbreitet[5]: Ein Ereignis kann sich wegen auftretenden vordefinierten Signaturen bzw. Charakteristika (z.B. 4 fehlerhaften Passworteingaben) als Angriff entlarven. Bei signaturbasierten IDS werden grundsätzlich drei Schritte vollzogen: Im ersten Schritt werden mittels Sensoren Logdaten (bei Host-basierten IDS) oder Daten des Netzwerkverkehrs (bei Netzwerk-basierten IDS) gesammelt. Im einem zweiten Schritt werden die so gesammelten Daten überprüft, verarbeitet und mit den Signaturen aus einer Musterdatenbank (welches eine Teilmenge der möglichen Angriffe anhand deren Eigenschaften beschreibt) mittels eines modellierenden Algorithmus (z.B. ein regelbasiertes "Pattern Matching"[6] oder ein Zustandsautomat[7].) verglichen. Stimmt die Signatur eines Ereignisses mit derjenigen in der Musterdatenbank überein, kann im dritten und letzten Schritt ein Alarm ausgelöst werden, welche eine Reaktion (Intrusion Response, siehe weiter unten) nach sich

[5]**Signaturbasierte IDS** waren auch die ersten IDS, die eingesetzt wurden: Bereits 1980 wurde von Anderson ein erstes Pioniersystem vorgestellt, welches bereits imstande war, Audit-Daten automatisch auf vorkommende Angriffsmuster zu analysieren. Dieses System war der Anstoss für die Entwicklung weiterer IDS-Systeme im zivilen Bereich. Im militärischen Bereich hatte Anderson sogar bereits 1972 ein Intrusion Detection System geschaffen.

[6]**Pattern Matching:** Einfach ausgedrückt ist dies eine "brute force" (also triviale) 1-zu-1 Überprüfung auf Übereinstimmung (hier einer Attacke zu einer in der Musterdatenbank gespeicherten Signatur).

[7]Mit einem **Zustandsautomaten** lassen sich komplexere Angriffe erkennen. Für jeden bekannten Angriff wird hierbei ein Zustandsübergangsdiagramm verwaltet, wobei ein Erreichen eines Endknotens in diesem Angriffsbaum einen Angriff signalisiert [17].

ziehen kann (z.B. Meldung an Administrator oder Isolierung des offensichtlich kompromittierten Knotens) [26] [1] [29] [27]. Es soll noch kurz auf die Vor- und Nachteile von signaturbasierten IDS eingegangen werden:

Vorteile

- Niedrige Rate von Falschalarmen (false positive rate) bzw. Typ I-Fehlern, da mit in der Musterdatenbank abgespeicherten Signaturen übereinstimmende Angriffe mit ziemlich grosser Sicherheit auch tatsächliche Angriffe darstellen [26] [1] [17].
- Geringer Installations- und Wartungsaufwand, da vom Hersteller meist Updates mit aktuellen Signaturen bezogen werden können [17].
- Hohe Effizienz dank einfacher Algorithmen (z.B. das erwähnte Pattern Matching) [17].
- Signaturen für die Musterdatenbank können relativ leicht erzeugt werden, da der Grossteil an Attacken im Internet meist einem genau definierbaren Muster folgen (Grund: Ausführung von Skripten).

Nachteile

- Ungeeignet, um bis anhin unbekannte Attacken aufzuspüren. Es liegt also eine hohe Rate von nichterkannten Angriffen vor, also eine hohe falsche Negativrate (das ist der besonders gefährliche Typ II-Fehler). Bei Pattern Matching beispielsweise genügt eine leichte Variation des Angriffs und er wird aufgrund der Unflexibilität schon nicht mehr erkannt. Ein signaturbasiertes IDS ist also stark von der Datenbank-Signaturqualität und der Flexibilität des eingesetzten Erkennungsmodells abhängig [26] [1] [17].
- Die Ausgabe eines signaturbasierten IDS ist bloss ein boolscher Wert (Angriff erkannt oder nicht erkannt), was natürlich keine differenzierte Analyse des Angriffs darstellt. So lässt sich lediglich die Häufigkeit eines bestimmten Angriffs bestimmen, aber nicht dessen Tragweite [17].
- Sollte ein System stark vom Durchschnitt abweichen, gestaltet sich die Anpassung der vom Hersteller gelieferten Signaturdaten als sehr aufwendig.

Anomaliebasierte IDS

Die ersten anomaliebasierten IDS wurden bereits 1985 eingesetzt[8]. Anomaliebasierte IDS verwenden heuristische Methoden, welche mit Angriffswahrscheinlichkeiten arbeiten. Es

[8]Denning und Neumann schlugen 1985 ein System namens IDES vor. Dieses System wertete Audit-Daten des Systems aus und erstellte für jeden Benutzer ein Profil über die letzten 50 Tage [17].

wird davon ausgegangen, dass der Systemnutzer ein statistisch erfassbares und regelmässiges Verhalten an den Tag legt [17]. Jegliche signifikante Abweichung eines im Idealfall ständig aktualisierten Normalverhalten-Profils kann als unerlaubtes Eindringen registriert werden. Dieses Normalprofil wird mittels historischer Daten wie Systemaktivitäten (z.B. Nutzungszeit und -häufigkeit, Ressourcenverbrauch und regelmässig benützter Applikationen) und Spezifikationen des zulässigen Benutzerverhaltens definiert [26]. Dies hat zur Folge, dass vom anomaliebasierten IDS potentiell wesentlich mehr Alarme ausgelöst werden können, je nachdem wie hoch der Schwellenwert der kritischen Abweichung festgelegt wird. Im Gegensatz zu signaturbasierten System werden hier also nicht Angriffe beschrieben, sondern der Normalzustand (frei von Angriffen). Jede Abweichung von diesem Normalzustand wird vom anomaliebasierten IDS genauer untersucht, wobei zu beachten ist, dass solche Abweichungen in der Praxis sehr schnell auftreten können, obwohl es sich dann aber meist nur um Falschalarme (also nicht bösartigen Abweichungen) handelt. Es wird aber davon ausgegangen, dass natürliche, ungewöhnliche Aktivitäten zu kleineren Abweichungen führen als Angriffe [30] [10] [26] [1] [29] [27] [20]. Die anomaliebasierten IDS können folgendermassen unterteilt werden:

1. **Programmierte Systeme:** Die als normal angesehenen Parameter werden fest definiert. Nachteil: Bei sich änderndem Nutzerverhalten kann keine sinnvolle Erkennung mehr gewährleistet werden.

2. **Selbstlernende Systeme:** Diese Systeme versuchen das Benutzerverhalten statistisch zu erfassen, d.h. es findet eine Eichung (Training) mit "Intrusion"-freien Referenzdaten statt, deren Qualität die spätere Erkennungseffektivität bestimmt. Danach werden gezielt normale und abnormale Abweichungen induziert, um die Sensibilität des IDS (Abweichungsschwellenwert) festzulegen. Nachteil: Die selbstlernenden Systeme können durch einen Angreifer langfristig so trainiert werden, dass der eigentliche Angriff später dann nicht mehr erkannt wird.

Vorteile

- Anomaliebasierte IDS besitzen im Gegensatz zu signaturbasierten IDS die Fähigkeit bisher unbekannte Attacken aufzuspüren. Ein perfekt an das System angepasstes IDS könnte so theoretisch alle Angriffe erkennen, eben auch selbst solche welche ein signaturbasiertes IDS nicht erkennen könnte. Praktisch gesehen werden aber auch mit einem anomaliebasierten IDS nicht alle Angriffe registriert (siehe Nachteile) [26] [17].

- Die Ausgabe des anomaliebasierten IDS ist differenzierter als jene des vorher besprochenen signaturbasierten IDS: Es wird also nicht bloss das Stattfinden eines Angriffs gemeldet, sondern es kann ferner die Abweichungsintensität vom Normalprofil angegeben werden, aus welcher mehr oder weniger auf die Wahrscheinlichkeit eines tatsächlichen Angriffs geschlossen werden kann [17].

- Durch stete Aktualisierung des Normalprofils ist eine Anpassung des IDS an dynamische Systemkonfigurationen und dynamische Topologien (insbesondere von MANETs) möglich [17].

Nachteile

- Es besteht die Schwierigkeit, das System in einem stark dynamischen Netzwerk (welches häufig natürliche Abweichungen verursachen kann) wie z.B. einem MANET zu trainieren [26].
- Es stellt einen erheblichen Aufwand dar, das Normalprofil periodisch zu aktualisieren, was v.a. bei MANETs aufgrund ihrer dynamischen Typologie oder allgemein in Netzwerken, in welchen der Nutzungsverlauf nur wenige bis keine konstante Faktoren aufweist, nötig ist [20].
- Die Wahl der Abweichungsschwellenwerte für die Alarmauslösung gestaltet sich als schwierig, denn diese sind nicht deterministisch bestimmbar [17].
- "Langsame" Attacken haben die Eigenschaft, dass deren Abweichung vom Normalprofil kleiner ausfällt. So kann es sein, dass die Abweichung unter dem Schwellenwert bleibt, obwohl der verursachte Schaden beträchtlich sein kann [17].
- Die im Gegensatz zu signaturbasierten IDS viel aufwendigere Berechnung der Abweichung vom Normalprofil kann eine schwere Last für die eingeschränkten Ressourcenkapazitäten mobiler Geräte darstellen [20].
- Weiterer Aufwand wird durch die hohe Falschalarmrate (high false positive rate, Typ I-Fehler) verursacht, da es insbesondere in mobilen Netzwerken oft sehr schwer ist, zwischen gut- und bösartigen Abweichungen zu differenzieren. Auch die dynamische Natur mobiler Netzwerke kann häufig zu Falschalarmen führen, welche von tatsächlich gerechtfertigten Alarmen ablenken können: Etliche Angriffe nützen dies aus und provozieren viele Falschalarme um das IDS abzulenken. Die Charakterisierung normalen Verhaltens in mobilen ad-hoc Netzwerken stellt also eine erhebliche Schwierigkeit dar [1]. Um diesen Nachteil zu entschärfen können empirisch gesammelte statistische Daten bezüglich konkreter natürlicher Abweichungsfälle eingesetzt werden, um die Zahl dieser Falschalarme zu vermindern.

Spezifikationsbasierte IDS

Die spezifikationsbasierte Erkennung ist der jüngste Ansatz[9]: Dabei wird das zulässige Verhalten definiert und jede Abweichung davon wird als kritisches Ereignis gewertet. Es gilt also genau zu spezifizieren, was einer Applikation erlaubt ist und was nicht. Jedes feststellbare nicht spezifizierte Verhalten wird dabei als Angriff gewertet. Der genannte Ansatz ist zwar mit dem Verfahren der Abweichungsberechnung bei der Anomalieerkennung ähnlich, verlässt sich aber nicht auf schwer nachvollziehbare mathematischen Regeln, sondern auf einfach begreifbare Spezifikationen des Programms [17]. Spezifikationsbasierte IDS kombinieren unter anderem Merkmale von signatur- und vor allem von anomaliebasierten IDS, wobei die Normalzustände hier klarer definiert sind als bei anomaliebasierten

[9]Die signatur- und anomaliebasierten IDS wurden schon in den 80er Jahren in damals noch verkabelten Netzwerken eingesetzt.

IDS. Eine Menge von Einschränkungen definiert dabei die korrekte Operation eines Programmes/Protokolles [3]. Dessen Ausführung wird entsprechend überwacht, wobei bei Überschreitung vordefinierbarer Grenzwerte Alarme ausgelöst werden können. Für diesen Ansatz werden meist einfache, zum Teil kontextfreie Grammatiken verwendet, welche die genauen Rechte festlegen [3] [20] [17].

Vorteile

- Bei einem Angriff ist genau feststellbar, was unerlaubt versucht wurde, denn die Regeln sind gut nachvollziehbar [17].
- Vermag wie das anomaliebasierte IDS auch bis anhin unbekannte Angriffe aufzuspüren. Es ist also kein Wissen über Angreifer oder Angriff nötig für die Erkennung. Periodische Aktualisierungen sind im Gegensatz zu anomaliebasierten IDS jedoch nicht nötig, um die Effektivität des IDS zu gewährleisten [3].
- Tiefe Falschalarmrate, denn die Regeln sind an die im Netzwerk eingesetzten Applikationen anpassbar [3] [17].
- Es können vom spezifikationsbasierten IDS sogar nicht nur Angriffe, sondern auch Fehler der Software selbst erkannt werden.
- Verschiedenen Applikationen können unterschiedliche Profile zugewiesen werden.

Nachteile

- Für alle Applikationen im System sollten entsprechende Spezifikationen existieren, denn ohne diese kann ein rein spezifikationsbasiertes IDS die entsprechenden Applikationen nicht schützen [17].
- Eine Applikation so zu spezifizieren, bis sie ohne Falschalarme im System funktioniert, ist ein langwieriger und aufwendiger Weg [17].
- Die Granularität der möglichen Spezifikationen kann teilweise zu grob ausfallen (z.B. will der Administrator vielleicht eine abgespeicherte Datei nicht komplett für den Zugriff freigeben, sondern nur gezielte Teile davon) [17].

7.3.3 Klassifikation der IDS gemäss ihrer Architektur

Es werden drei gängige Grundkonfigurationen einer IDS-Architektur unterschieden: Die Stand-alone IDS Architektur und die verteilte kooperative IDS Architektur stellen dezentralisierte IDS Architekturen dar, wobei alle Knoten mit der dafür nötigen Sensorik oder bei Stand-alone IDS Architekturen gar mit einem ganzen IDS ausgestattet werden, was natürlich entsprechend aufwendiger ist als bei einer zentralisierten IDS Architektur, wo nur ein einziges IDS mit vielen Sensoren eingesetzt wird, was aber mit einem höheren Risiko (Ausfall der Zentrale!) verbunden ist [3].

Stand-alone IDS Architektur

Bei dezentralisierten Stand-alone IDS Architekturen kann auf jedem Host unabhängig ein IDS betrieben werden. Dieser Architekturtyp wird beispielsweise in flachen Netzwerkinfrastrukturen eingesetzt. Sämtliche Entscheidungen der IDS basieren auf den auf dem Host verfügbaren Informationen. Ein Hauptnachteil besteht darin, dass keine Kooperation zwischen den IDS möglich ist: Ein Host kann also nicht andere Hosts über entdeckte komprimittierte Knoten informieren. Diese fehlende Kooperation stellt insbesondere in MANETs einen erheblichen Nachteil dar, weil jedes IDS nur einen sehr begrenzten Überblick über die Situation hat. Somit stellen die folgend vorgestellten verteilten und kooperativen IDS Architekturen eine wesentlich bessere Lösung dar [3] [27].

Verteilte und kooperative IDS Architektur

Diese IDS Architektur ist insbesondere geeignet für ursprünglich flache Netzwerkinfrastrukturen (z.B. induziert das im Kapitel "Lösungsansätze" vorgestellte verteilt und kooperativ agierende Zone-Based Intrusion Detection System erst nachträglich ein gewisses Mass an Hierarchie). Jeder Knoten tätigt lokale Entscheidungen, kooperiert aber im Gegensatz zur Stand-alone IDS Architektur jedoch mit anderen Knoten zwecks globaler Angriffserkennung (global intrusion detection), was den mehr als kompensierten Nachteil eines höheren Aufwandes mit sich bringt. Es handelt sich also um ein weitgehend dezentralisiertes IDS. Verteilte und kooperative IDS Architekturen zeichnen sich durch Skalierbarkeit und Robustheit (sich über mehrere Host erstreckende Angriffe wie hohes Verkehrsdatenaufkommen können dank Kooperation leichter aufgespürt werden) aus. Ferner bestünde die Möglichkeit, die gesammelten Daten je nach Datentyp (z.B. gemäss Portnummer: http,ftp...) auf mehrere IDS zu verteilen, womit grössere Datenmengen bewältigbar wären. Ein Kernproblem stellt die Wahl des Abstraktionsgrades dar: Je niedriger die Abstraktion, desto mehr Alarme werden an andere IDS weitergeleitet, was eine höhere Falschalarmrate (höhere Häufigkeit von Typ I-Fehlern) mit sich bringen kann. Bei einer hohen Abstraktion hingegen werden nur ziemlich sichere Angriffe an andere IDS rapportiert. Aufgrund von Kosten- und Managementfragen ist eine solche Architektur vor allem für kleinere WLANs bzw. MANETs mit ein bis zwei WAPs (Wireless Access Points) geeignet. Ein Nachteil besteht u.a. darin, dass bösartige Knoten gutartige Knoten fälschlicherweise anzeigen können.

Zentralisierte IDS Architektur

Zentralisierte IDS werden in vielgeschichteten hierarchischen Netzwerkinfrastrukturen eingesetzt. Individuelle Sensoren sammeln und leiten sämtliche Daten an ein zentrales IDS Managementsystem weiter, in welchem die Daten gespeichert und verarbeitet werden. Der Vorteil ist dabei klar der geringere Koordinationsaufwand (deshalb auch hervorragende Eignung für den Einsatz in Krisengebieten) einer solchen zentralisierten Architektur. Diesem Vorteil stehen aber eine Vielzahl von erheblichen Nachteilen gegenüber, weswegen zentrale IDS trotzdem nur selten in kabellosen Netzwerken eingesetzt werden [3] [27] [17]:

- Erstes Problem: Übertragen der zentralen IDS Aufgabe an einen absolut vertrauenswürdigen Teilnehmer, welcher mit allen anderen Knoten vernetzt ist und hardwaremässig bereits mit einem IDS ausgerüstet ist. Dies gestaltet sich als schwer, denn wenn ein ad-hoc Netzwerk sich erstmals bildet, sind per se beim Eintritt eines Teilnehmers in die Verpflichtungen eines zentralen IDS noch keine vertrauensspezifischen Informationen über diesen verfügbar [17].

- Zweites Problem: Die Kommunikation des zentralen IDS mit den Teilnehmern führt aufgrund der Netzwerktopologie mit hoher Wahrscheinlichkeit zu Bandbreitenengpässen bei den unmittelbaren Nachbarn des zentralen IDS [17].

- Da ein Teil des Netzwerkdatenverkehrs immer ausserhalb der Empfangsreichweite des zentralen IDS geschieht, ist es de facto unmöglich, den gesamten Netzwerkdatenverkehr lückenlos zu analysieren, was ein erhebliches Sicherheitsmanko darstellt [17].

- Um zumindest den grössten Teil des Netzwerkdatenverkehrs analysieren zu können, müssen viele Datenströme aufwendig über das zentrale IDS umgeleitet werden, was aber kaum realisierbar ist [17].

- Das Nachfolgeproblem ist schwer lösbar: Wie ist vorzugehen, wenn der zentrale IDS Knoten das Netz verlässt oder zeitweise nicht erreichbar ist? Das IDS müsste dann auf einen anderen Teilnehmer platziert werden. Dieser Ersatzteilnehmer müsste dann aber möglichst mit dem bisher angesammelten Wissen versorgt werden, was sehr aufwendig wäre [17].

7.3.4 Intrusion response

Es wäre für das Sicherheitspersonal in den meisten Fällen ein Ding der Unmöglichkeit, in vernünftiger Zeit auf sämtliche durch das IDS ausgelösten Alarme manuell zu reagieren. Das IT-Personal ist ergo käumlich fähig ohne maschinelle Automatismen in effizienter Manier auf sämtliche Attacken zu reagieren. Es kann nun eine Effizienzsteigerung erreicht werden, wenn sich das Sicherheitspersonal vorwiegend auf die Beobachtung und Behandlung bis anhin unbekannter Angriffe konzentriert und das Reagieren auf bekannte Angriffe einem einsetzbaren Intrusion Response System (IRS) überlässt, welches sich entsprechend parametrisieren lässt [7]. Streng genommen bilden IDS und IRS zwei verschiedene Systeme (wobei die Verwendung lediglich eines Systems keinen Sinn macht), aber häufig umfasst der Begriff IDS auch gleichzeitig das darin eingesetzte IRS (z.B. ist ein IRS-Modul Bestandteil des in einem späteren Kapitel besprochenen lokalen IDS Agenten, welcher in gewissen IDS eingesetzt wird) [7]. Es existieren verschiedenste Möglichkeiten auf Attacken zu reagieren [24]. Die unterschiedlichen Reaktionen lassen sich grob in die zwei Hauptkategorien aktiv und passiv einteilen: Ein IRS sollte so wie das IDS im Idealfall verteilt und kooperativ funktionieren [29]. Die Art der Reaktion in kabellosen Netzwerken wie z.B. einem MANET hängt von Faktoren wie Angriffstyp, Netzwerkprotokoll, genutzte Applikationen und Angriffssicherheit ab.

Passive Reaktionen

Bei den häufiger eingesetzten passiven Reaktionen eines IRS auf die von IDS aufgespürten Attacken werden lediglich Informationen über den Angriff bzw. den/die Angreifer gesammelt und gemäss vorgenommenen Einstellungen darüber rapportiert [18]. Der Zweck passiver Reaktionen besteht aber nicht darin, weitere Attacken präventiv zu verhindern. Die zwei meistens verwendeten passiven Reaktionen sind Logging (Logbuchführung) und Reporting (Benachrichtigungen):

- **Logging:** Die Attacke wird protokolliert, beispielsweise mittels "IP Logging", wo netzwerkbasierte IDS Sensoren so konfiguriert werden, dass IP-Paketdaten protokolliert werden, nachdem ein Angriffsmuster (also eine Signatur) erkannt wurde. Die Sensoren werden so konfiguriert, dass die Grösse der Log-Dateien aus Performancegründen ein gewisses Mass nicht überschreitet. Die Log-Dateien werden dann automatisch an einen dedizierten Server transferiert, welcher diese Protokolldaten automatisch auf ein (aus Sicherheitsgründen) nichtlöschbares Medium wie CD schreibt. Es muss beachtet werden, dass das Protokollieren und vor allem das Analysieren der Log-Dateien stets einen beträchtlichen Aufwand darstellt [6].
- **Reporting:** Entweder wird in Echtzeit eine Benachrichtigung an einen Administrator verschickt (via Mail, Pager...) oder es können auch direkt bewanderte Endbenützer informiert werden, die dann selber weitere Nachforschungen anstellen und entsprechende Aktionen tätigen können [20]. Die in kabellosen Netzwerken generierten Rapporte fallen natürlich anders aus als in verkabelten Netzwerken und müssen wesentlich mehr Informationen enthalten [24]. Ein Rapport kann dabei Informationen wie Ziel, Zeit und Quelle der Attacke und Beschreibungen über verdächtigen Aktivitäten enthalten. Ein Reportingsystem sollte ferner in Echtzeit in gewissen Intervallen eine Trendanalyse liefern, welche u.a. über die vorgekommenen Hauptangriffe, die Anzahl aller Attacken und die Tageszeiten gehäufter Angriffsaktivität Bericht erstattet [6]. Das Reportingsystem sollte die Administration (v.a. in Fällen ausserhalb der Firewall, d.h. in Richtung Internet) nur benachrichtigen, falls ein gewisses Mass an Abweichung vom normalen Zustand auftritt oder neue, bis anhin unbekannte Angriffsmuster auftauchen, damit der Aufwand im Umgang mit Alarmen in Grenzen gehalten wird.

Aktive Reaktionen

Die wirksameren, aber in IDS noch selten eingesetzten aktiven Reaktionen versuchen gezielt, Angreifer aufzuspüren und mögliche Folgeschäden des Angriffs zu minimieren [24]. Evtl. wird sogar ein Gegenangriff lanciert, jedoch wird dabei auf Denial-of-Service Gegenattacken verzichtet, denn diese könnten dabei die Netzwerkleistungsfähigkeit erheblich in Mitleidenschaft ziehen. Es existieren eine Vielzahl aktiver Reaktionen, von denen einige wenige hier vorgestellt werden sollen:

- **Blocking** von IP-Adressen, TCP-Verbindungen, Ports etc. [6].

- **Neustarten** kompromittierter Systeme oder in besonders gravierenden Fällen des gesamten Netzwerkes [24].
- Anwendung des **Intrusion Response Modells (IRM)**: In diesem werden kompromittierte Knoten von einem Knoten identifiziert, falls ein Zähler eine gewisse Schwelle überschreitet.
- **Reauthentifizierung** aller Teilnehmer mittels diverser Mechanismen (z.B. könnte ein visueller Kontakt durch andere Knoten zur Reauthentifizierung verlangt werden). [20].
- Hierarchischer Ansatz mittels **"Forwarding Policy"**, welcher ein Weiterleiten von Paketen nur an authentifizierte Knoten erlaubt. Die **"Certificate Authority"** kann dabei verdächtige Knoten isolieren [3].
- **Eingeschränkte Nutzungsrechte** für Benützer, welche sich verdächtig verhalten [24].
- Identifikation komprimittierter Knoten und **Reorganisation des Netzwerkes** (z.B. Etablierung eines neuen Kommunikationskanals) um bösartige Knoten auszuschliessen [3] [20].
- **Expertensysteme** bei signaturbasierten IDS (siehe weiter oben in diesem Kapitel), welche mit streng kodierten Regeln in Form von if-then-else Konstrukten arbeiten. Die Qualität solcher Systeme steigt und fällt natürlich mit der Qualität dieser an das Netzwerk anzupassenden Regeln und der Aktualität der Signaturdatenbank.
- Einsatz von sogenannten **"Lockvögeln"**, wobei Eindringlinge durch Frames mit randomisierten, unsinnigen Daten verwirrt werden. Eine andere Möglichkeit besteht darin, verfälschte Management-Informationen an Eindringline zu versenden, welche z.B. nicht existierende IP-Adressen enthalten. Werden diese IP-Adressen dann vom Angreifer verwendet, ist er dadurch viel leichter durch das IDS aufspürbar [18].

7.3.5 Zusammenfassung

In diesem Kapitel wurden Intrusion Detection Systeme gemäss drei verschiedenen Kriterien klassifiziert. Einerseits lassen sich IDS bezüglich des Ortes der eingesetzten Sensorik unterscheiden: Bei hostbasierten IDS werden auf jedem zu überwachenden System eigene Sensoren eingerichtet, was etwas aufwendiger ist, jedoch eine umfassende Systemüberwachung zulässt. Bei netzwerkbasierten IDS werden meist eine kleinere Anzahl von Sensoren direkt im Netzwerk selbst eingesetzt, wobei es in einem MANET aufgrund der dynamischen Topologie relativ schwer ist, geeignete Einsatzstellen dafür zu ermitteln. Zweitens lassen sich IDS gemäss der verwendeten Erkennungsmethodik in signatur-, anomalie- und spezifikationsbasierte IDS einteilen: Bei signaturbasierten IDS werden Angriffsmuster in einer Datenbank mittels abgespeicherter Signaturen beschrieben. Anomaliebasierte IDS berechnen aufwendig Abweichungen vom insbesondere in MANETs schwierig definierbaren Normalzustand und sind im Gegensatz zu signaturbasierten IDS auch fähig, bis anhin unbekannte Attacken aufzuspüren. In spezifikationsbasierten IDS lässt sich das zulässige

Verhalten eingesetzter Applikationen parametrisieren. Als drittes Klassifikationskriterium wurden drei Grundtypen von IDS Architekturen genannt: In Stand-alone IDS Architekturen wird auf jedem System ein lediglich lokal und unabhängig funktionierendes IDS eingesetzt, was zu einer begrenzten Sichtweise der einzelnen IDS führt. Eine bessere Erkennungsrate (bei etwas höherem Koordinationsaufwand) stellen verteilte und kooperative IDS Architekturen dar, in welchen die eingesetzten IDS miteinander zusammenarbeiten, um nicht nur lokale, sondern auch globale Angriffe effektiv erkennen zu können. Zentralisierte IDS als dritte Variante bauen im Gegensatz zu den beiden vorherig genannten Varianten nicht auf flachen, sondern auf hierarchischen Netzwerkinfrastrukturen auf, was zur Folge hat, dass sie in kabellosen Netzwerken nur selten verwendet werden. Mit dem Intrusion Response System (IRS) kann auf vom IDS erkannte Angriffe passiv und aktiv reagiert werden. Bei passiven Reaktionen beschränkt sich das IRS darauf, Informationen über den Angriff und den Angreifer zu sammeln. Bei aktiven Reaktionen versucht das IRS den Angreifer aufzuspüren und mögliche Angriffsschäden zu minimieren.

7.4 Lösungsansätze

In diesem Abschnitt werden Ansätze und Lösungen besprochen, wie man MANET-spezifische Probleme lösen kann. Zuerst wird das Dynamic Source Routing (DSR) Protokoll vorgestellt, da es als Grundlage für die folgenden Ansätze dienen kann. Darauf wird eine Schwachstelle des DSR Protokolls anhand der Routing Disruption Attacke aufgezeigt. Der Watchdog / Pathrater Ansantz versucht diese Schwachstelle zu eliminieren. Dabei wird die Problematik von egoistischen Knoten beschrieben. Um diese zu entdecken werden Bewertungsverfahren wie CORE, CONFIDANT oder Routeguard eingesetzt. Mit dem Zone-Based IDS wird ein Ansatz von kooperierenden IDS vorgestelln. Das Kapitel wird mit der Beschreibung von verteilten Attacken und Vorschlägen, wie diese zu entdecken sind, abgeschlossen.

7.4.1 Dynamic Source Routing

Das Dynamic Source Routing (DSR) Protokoll ist kein IDS im engeren Sinne, jedoch bauen die nachfolgenden Mechanismen darauf auf. Deswegen wird seine Funktionsweise an dieser Stelle kurz erläutert. Das DSR Protokoll ist ein Routing Protokoll und löst das Problem in dem dynamischen Umfeld eines Manet eine Route von Sender zu Empfänger zu finden. Ein Sender der ein Paket an einen bestimmten Empfänger senden will, jedoch keine Route zu ihm kennt, sendet ein Route Request via Broadcast an alle seine Nachbarknoten. Das Route Request Paket besteht aus der Angabe von Quell- und Zieladresse sowie einer leeren Liste. Jeder Knoten der diese Nachricht erhält und nicht der Zielknoten ist, fügt sich in die Liste ein und sendet das modifizierte Paket weiter. Falls der betreffende Zielknoten das Paket erhält sendet er es (mit der Liste der zu ihm geführten Route) zurück an die Quelle. Die Quelle erhält somit die Route zum gewünschten Zielknoten.[19]

7.4.2 Routing Disruption Attacke auf das Dynamic Source Routing-Protokoll

Da Routingprotokolle wie das speziell für MANETs entwickelte, reaktive[10] Dynamic Source Routing-Protokoll[11] Eckpfeiler von MANETs darstellen, versuchen Angreifer häufig die gesendeten Routinginformationen zu verfälschen, so dass die Routingprotokolle nicht mehr funktionieren [25]. Eine derartig funktionierende aktive[12], speziell MANET-spezifische Attacke soll hier nun kurz beschrieben werden.

Angriffsszenario

Das hier beschriebene Beispielangriffsszenario bezieht sich auf die dargestellte Abbildung 7.2 unten [25]. Der Zweck dieser hier beschriebenen Attacke ist dabei, die Routinginformationen von möglichst vielen Opferknoten zu verfälschen. Der bösartige Knoten 1 antwortet dabei auf die von Knoten 3 gestellte Anfrage nach einem Pfad zu Knoten 1 mit einem zufällig konstruierten verfälschten Routing Reply (RREP) Paket[13], wobei das verfälschte Routing-Paket den gültigen Teilpfad (1,5,3) enthalten muss, um den Anfragerknoten 3 überhaupt erreichen zu können. Nun wird dieser Teilpfad noch mit randomisiert generierten, meist unsinnigen Pfaden ergänzt, wie z.B. (2,4,8,7,**1,5,3**). Wenn dieses RREP-Paket an Knoten 3 gesendet wird, sind einerseits alle Knoten entlang des Pfades (1,5,3), andererseits auch sämtliche Nachbarsknoten (also z.B. 2,4,7,8) dieses Pfades von der Attacke betroffen (Begründung siehe unten).

Unterschiede der Angriffswirkung in verkabelten Netzwerken und MANETs

Die Konsequenzen einer hier beschriebenen Routing Disruption Attacke sind nicht gleich für verkabelte Netzwerke und kabellose mobile ad-hoc Netzwerke, wie folgende kurz erläuterte Diskrepanzen aufzeigen sollen [25].

[10] Bei **reaktiven Routingprotokollen** wird ein Pfad nur dann gesucht, falls ein Quellknoten explizit Daten an einen Zielknoten zu senden wünscht (Route Discovery). Dieses Verfahren ist somit relativ effizient und unkompliziert. Ist so ein Pfad einmal etabliert, wird er aufrechterhalten (Route Maintenance) bis entweder nicht mehr auf den entsprechenden Zielknoten zugegriffen werden kann oder der Pfad nicht mehr länger erwünscht ist. Bei **proaktiven Routingprotkollen** sendet jeder Knoten im Gegensatz zu reaktiven Routingprotokollen seine Routing-Tabelle periodisch (time-driven updates) und wenn eine signifikante Änderung eingetreten ist (event-driven updates), egal ob momentan ein Bedarf am entsprechenden Pfad besteht.

[11] Das DSR-Protokoll ist deshalb so geeignet für MANETs, weil Netzwerke dadurch selbstorganisierend und -konfigurierend werden ohne dass eine zentrale Administration dafür nötig wäre.

[12] Grob gesagt, unterscheiden sich **aktive Attacken** von passiven Attacken dadurch, dass nicht nur Daten abgehorcht, sondern auch modifiziert werden.

[13] Würde der Angreiferknoten mit einem korrekten Routing Reply Paket (1,5,3) antworten, könnte er künftig an ihn geschickte Pakete auch schlichtweg wegwerfen. Der Angreiferknoten würde dann ein sogenanntes **Routing black-hole** darstellen, also eine Art schwarzes Loch, in welchem alle empfangene Pakete verschwinden. Dies ist bei der Routing Disruption Attacke jedoch nicht der Fall.

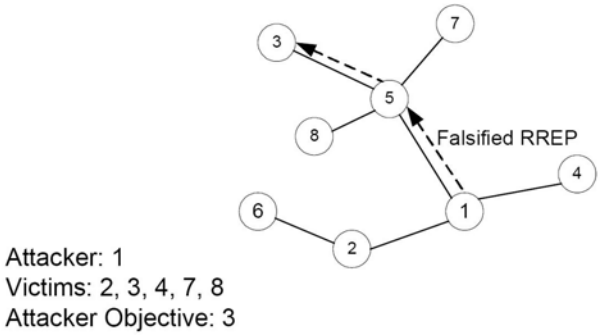


Abbildung 7.2: Vorgestelltes Angriffsszenario einer Routing Disruption Attacke [25].

Anzahl betroffener Knoten Die Ursache, dass auch die Nachbarsknoten zu Opfern werden, liegt darin, dass aufgrund der kabellosen und dynamischen Natur eines MANETs alle Knoten stets versuchen, durch ein hier legales Abhorchen an die aktuellsten Routinginformationen zwecks Aktualisierung ihrer Routing-Tabellen im eigenen Routing Cache zu gelangen [25]. Weil es sich beim DSR-Protokoll um ein reaktives Protokoll handelt, senden die Knoten nicht von sich aus ihre neuesten Routinginformationen. Durch das so verursachte Mithören wird also eine grössere Anzahl von Knoten des Netzwerkes zu Opfern verfälschter Routinginformationen, als dies in einem verkabelten Netzwerk der Fall wäre.

Angriffsdauer Die mobile Natur eines MANETs bietet für den Angreifer aber nicht nur Vorteile [25]: Die mobilen Knoten eines MANETs können aufgrund ihrer Bewegung auch lediglich partielle Opfer sein, da es unwahrscheinlicher wird, dass mobile Knoten ununterbrochen zum Opfer verfälschter Routinginformationen werden.

Auswirkung eines verfälschten Pakets Weil die Knoten ihren Routing Cache (dessen Inhalt die Mobilität des Netzwerkes reflektiert) viel häufiger aktualisieren müssen als in verkabelten Netzwerken, muss der Angreifer auch entsprechend eine grössere Anzahl verfälschter Routinginformationspakete verschicken um eine langanhaltende Wirkung erzielen zu können [25].

7.4.3 Watchdog / Pathrater

Neben der Routing Disruption Attacke gibt es weiter das Problem von egoistischen Knoten. Um Energie oder Rechenzeit zu sparen, verwirft ein egoistischer Knoten Pakete, die er weiterleiten sollte.[19] Ein solches Verhalten eines Knoten kann das Netzwerk erheblich stören. Sämtliche Routen, die über diesen Knoten verlaufen, können nicht mehr benutzt werden und es müssen andere, eventuell längere Routen benutzt werden. Dies hat einen höheren Energieverbrauch zur Folge (im gesamten Netz betrachtet). Je nach Topologie des Netzes können gewisse Knoten nicht mehr angesprochen werden, dies ist dann der Fall wenn die Route zwischen zwei Knoten über den egoistischen Knoten verläuft und keine alternative Route möglich ist.

Ansatz

Watchdog liefert einen Ansatz, um solche egoistischen Knoten zu entdecken. Watchdog basiert auf dem Promiscuous Mode, wobei alle Pakete die empfangen werden auch verarbeitet werden (im Gegensatz zu Unicast, wo die Knoten die Pakete nur dann verarbeiten, wenn die Pakete an sie adressiert sind.) Watchdog überprüft ob sein Nachbarknoten, dem er ein Paket zur Weiterleitung gesendet hat, dieses auch wirklich unverändert weitersendet. Um dies zu erreichen führt Watchdog eine Liste, die sämtliche Pakete enthält die an Nachbarknoten gesendet wurden und von diesen wieder versendet werden sollen. Der Knoten kann nun überprüfen ob das Paket unverändert weitergeleitet wurde und ob es überhaupt weitergeleitet wurde, denn wenn der Nachbarknoten das Paket weiterleitet, so wird es auch vom im Pfad vorangegangenen Knoten empfangen. Ist dies nicht der Fall oder es wurde eine Veränderung an dem betreffenden Paket festgestellt, kann der betreffende Knoten als böswillig identifiziert werden.[16] Der ursprüngliche Quellknoten des böswillig verworfenen Paketes wird über den betreffenden Knoten informiert und Pathrather wird für die zukünftige Routenfindung diese Informationen verwenden.[16] Was dazu führt, dass der egoistische Knoten nicht mehr für Routen berücksichtigt wird.[4]

Bewertung

Der Watchdog / Pathrater Ansatz ermöglicht es also die Zuverlässigkeit von Nachbarknoten zu überprüfen. In [4] wird gesagt, dass Knoten auf ihren Watchdog vertrauen müssen, hingegen wird in [16] beschrieben, dass die Informationen über böswillige Knoten weitergegeben werden. Dieser Widerspruch entsteht vermutlich dadurch, dass es unterschiedliche Implementationen von Watchdog gibt. Fakt hingegen ist, dass durch diesen Ansatz die egoistischen Knoten belohnt werden, da sie nicht mehr in Routen vorkommen und ihnen somit die Arbeit abgenommen wird Pakete zu verwerfen. Böswillige Knoten können jedoch ihre eigenen Pakete ungehindert versenden und allfällige Antworten darauf erhalten.

7.4.4 Bewertungsverfahren

Die nachfolgend beschriebenen Bewertungsverfahren sind Ansätze, welche den Watchdog / Pathrater Ansatz insofern verbessern, als dass ein detektierter egoistischer Knoten nicht mehr einfach toleriert wird. Den folgenden Ansätzen ist gemeinsam, dass sie andere Knoten bewerten. Die Knoten können sich somit ein Bild von seinen Nachbarknoten machen betreffend Zuverlässigkeit und Vertrauenswürdigkeit.

CORE Ansatz

Core (Collaborative Reputation Mechanism) [16] besitzt eine Watchdog Komponente, die mit einem Reputations-Mechanismus versehen ist.[4] Der Mechanismus verfügt über

verschieden gewichtete Bewertungen. Subjektive (durch Beobachtungen), indirekte (positive Berichte von anderen Knoten) und funktionelle Bewertungen (aufgabenspezifisch) ermöglichen es einen Knoten zu bewerten. Die Knoten beobachten einander und vergleichen die erwarteten Resultate mit den effektiv erhaltenen. Weiter tauschen sie positive Bewertungen untereinander aus.[4] Somit ist es nicht möglich, dass Knoten falsche negative Bewertungen versenden können, was zu Angriffszwecken missbraucht werden könnte. In [16] werden jedoch explicit DoS Nachrichten beschrieben, mit denen ein Knoten seinen Nachbarn mitteilen kann, dass er Anfragen von einem als bösartig gesehenen Knoten nicht mehr verarbeiten wird. Die Nachbarn des bösartigen Knotens überprüfen ihre Bewertungen des als bösartig bezeichneten Knotens. Sollten diese Bewertungen jedoch positiv ausfallen, werden sie den Knoten der die Anschuldigung gesendet hat schlechter Bewerten um ihn zu bestrafen. Es kann also auch vorkommen, dass ein Knoten der seine Nachbar-Knoten zu Recht warnt, dafür bestraft wird. Dies ist möglich, wenn der bösartige Knoten erkennt, dass er überprüft wird und sich deswegen normal verhält (was somit auch wieder als Attacke missbraucht werden kann). Um den schlechten Bewertungen mehr Gewicht zu verleihen, müssen diese deutlich stärker gewichtet werden als die Positiven.[16] Die betreffenden Bewertungs-Mechanismen sind in [16] genauer beschrieben.

CONFIDANT Ansatz

Das CONFIDANT (Cooperation Of Nodes, Fairness In Dynamic Ad-hoc NeTworks) Protokoll ist eine Erweiterung zu einem reaktiven Source-Routing Protokoll, wie DSR eines ist. CONFIDANT hat das Ziel bösartige Knoten zu erkennen und zu isolieren, um die Attraktivität zu senken die Mitarbeit und Kooperation in einem Ad-hoc Netz zu verweigern.[5] In [5] wird das Protokoll detailliert Beschrieben, Leistungsanalysen sind ebenfalls zu finden. Die Knoten sammeln Informationen über andere Knoten selber und verarbeiten Informationen (positive wie auch negative) die sie von anderen Knoten über dritte erhalten. Mit diesen Informationen bzw. Bewertungen über andere Knoten, können bösartige Knoten gemieden oder ganz isoliert werden.[4]

Bewertung CORE, CONFIDANT

CORE und CONFIDANT sind beides nützliche Erweiterungen des Watchdog / Pathrater Mechanismus. Nach [16] ist Core eines der wenigen Systeme die genauer spezifiziert sind. Bei CONFIDANT werden in [16] mehrere Mängel genannt, z.B. dass das Rating der Knoten nicht spezifiziert wird und dass Mechanismen fehlen, die das IDS selber vor Angriffen schützt. Weiter kommt hinzu, dass der verwendete Promiscous Mode viele Probleme mit sich bringt. [16] Es gilt auch zu beachten, dass ein Knoten mit den Ratings ”spielen” kann. Sinkt sein Rating durch böswillige Aktionen, so muss er sich nur eine gewisse Zeit korrekt verhalten, um wieder ein gutes Rating zu erhalten und kann sich somit tarnen um erneute Angriffe zu lancieren.

Routeguard Ansatz

Die bisher beschriebenen Bewertungsverfahren bewerten andere Knoten nach ihrer Funktionalität. Wird die geforderte Funktion erfüllt so steigt die Bewertung oder sie sinkt. Die Bewertung befindet sich in einem bestimmten stetigen Intervall. Verschiedene Knoten haben eine unterschiedliche Bewertung eines bestimmten Knotens. Routegard klassifiziert die Knoten zusätzlich. Die Knoten eines Netzes werden in die fünf Klassen: Fresh, Member, Unstable, Suspect oder Malicious eingeteilt. Dabei werden sie je nach ihrem Status unterschiedlich behandelt.[11]

Fresh Dieser Status wird vergeben, wenn ein Knoten neu im Netzwerk (z.B. durch Routensuche) entdeckt wird. Das Netzwerk behandelt den neuen Knoten mit Vorsicht und dieser darf nur eingeschränkt am Netz teilnehmen (z.B. Pakete weiterleiten). Ist sein Rating nach einer kurzen Periode grösser oder gleich 0, so wird er als *Member* klassifiziert. Ansonsten als *Suspect*.[11]

Member Dies ist der reguläre Status in dem senden, empfangen und weiterleiten von Paketen erlaubt ist. Das Rating wird auf 0 gesetzt und kann höchstens einen maximalen Wert erreichen. Sinkt das Rating jedoch unter einen Schwellenwert wird der Knoten auf *Unstable* zurück gestuft.[11]

Unstable Ein Knoten der sich in diesem Status befindet, darf für eine bestimmte Zeitperiode nur Pakete empfangen und weiterleiten. Schlechtes Verhalten wird in diesem Status stärker bestraft, hat er aber nach abgelaufener Überwachungsperiode ein positives Rating, so wird er zurück in den *Member* Status gehoben. Ist das Rating zum Überprüfungszeitpunkt negativ erhält er den Status *Suspect*.[11]

Suspect Besitzt ein Knoten den Status *Suspect* wird er für einen längeren Zeitabschnitt isoliert. Nach Ablauf dieser Zeit wird er wieder in das Netz eingegliedert und für eine bestimmte Zeit von Watchdog überwacht. Verhält sich der Knoten während der Überwachung korrekt, wird sein Status in *Unstable* umgewandelt. Wenn nicht erhält er den Status *Malicious*.[11]

Malicious Das Verhalten des Knotens ist untolerierbar und er wird permanent aus dem Netz ausgeschlossen. Dabei wird er in einer Dismal-Liste aufgenommen und hat keine Möglichkeit mehr erneut in das Netz eingegliedert zu werden.[11]

In Abbildung 7.3 werden die Status-Übergänge mit den jeweiligen Bedingungen grafisch dargestellt.

Damit ein böswilliger Knoten nicht mit den Ratings spielen kann, sich also z.B. normal Verhalten, dann böswillig (wobei er deklassiert wird) und dann wieder normal, um in den *Member* Status zurückzugelangen, ist ein ”malcounter” implementiert. Jedes mal, wenn ein Knoten von *Member* auf *Unstable* abgewertet wird, erhöht sich der ”malcounter”. Überschreitet dieser Wert eine definierte Schwelle wird der betreffende Knoten in den Status *Suspect* versetzt.[11]

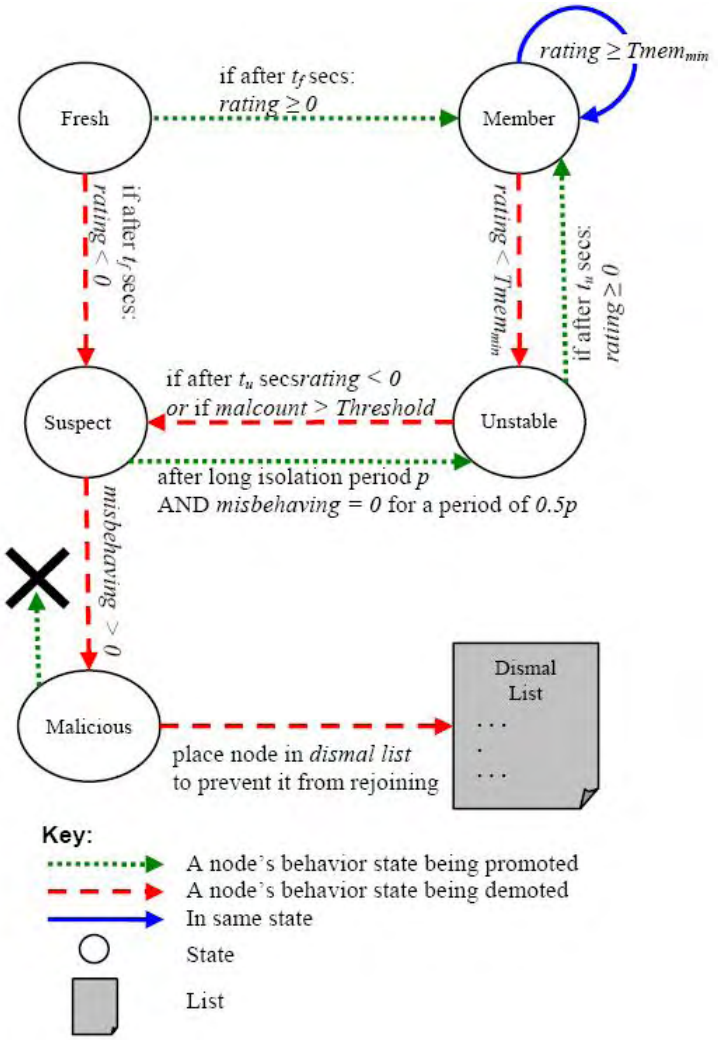


Abbildung 7.3: State chart - Routeguard [11]

Bewertung Routeguard

Obwohl nur Knoten die den Status *Member* haben, voll auf das Netzwerk zugreifen können ist gemäss [11] eine deutliche Steigerung des Paket-Flusses im Netzwerk bemerkbar auch mit einer grossen Anzahl von böswilligen Knoten. Im Gegensatz zu CORE oder CONFIDANT, können böswillige Knoten nicht mehr beliebig mit den Ratings "spielen".

7.4.5 Zone-Based Intrusion Detection System

Werden in einem Intrusion Detection System für mobile ad-hoc Netzwerke beispielsweise in den Hosts nur in lokaler Manier voneinander unabhängige IDS eingesetzt, bleibt die Leistungsfähigkeit eines solchen IDS meist nur schwach, das heisst dass unter anderem die Falschalarmrate zu hoch und die Erkennungsrate zu niedrig ist [25]. Abhilfe schaffen soll der hier vorgestellte Lösungsansatz, welcher die Kollaboration der einzelnen IDS "Agenten" (welche hier noch genauer besprochen werden) zwecks Formung eines

kooperativ und verteilt agierenden MANET IDS bewerkstelligt. Um trotz der dynamischen Topologie eines MANETs (die Topologie eines MANETs ist i.d.R im Vornherein nicht bekannt) eine gewisse hierarchische Struktur[14] einzuführen, kann beispielsweise das Clustering angewandt werden: Es werden Gruppen von Knoten (Cluster) inklusive z.B. periodisch zufällig neu gewählter Hauptknoten (Cluster-Head) gebildet, welche die Ressourcen mobiler Knoten der entsprechenden Clusters in MANETs entlasten können [3]. Intermediäre Gateway-Knoten (eine Art von Schnittpunkten der Cluster) leiten dann allgemein ausgedrückt Datenpakete zwischen Cluster-Head-Knoten weiter. Ferner gruppiert ein Aggregationsalgorithmus in den relativ spontan ausgewählten Gateway-Knoten die registrierten verdächtigen Aktivitäten. Auf diese Weise kann von den Gateway-Knoten entschieden werden, wann Alarme auch wirklich ausgelöst werden sollen.

Idee

Als Erstes muss das strukturlose MANET-Netzwerk strukturiert werden, um eine effiziente Arbeit des Aggregationsalgorithmus zu ermöglichen [25]. Dies geschieht mittels einer Partitionierung des Netzwerkes (vergleiche Abbildung 7.4 unten) in logische, sich nicht überlappende Zonen (eine Art Cluster). Für jede Zone werden spontan einige Knoten als sogenannte Gateway-Knoten (analog zu Cluster-Heads) auserwählt. Auf diesen Gateway-Knoten wird der erwähnte Aggregationsalgorithmus eingesetzt, welcher für die Gruppierung der von den einzelnen IDS registrierten verdächtigen Aktivitäten zuständig ist. Alle Knoten sind bei diesem Lösungsansatz mit IDS Agenten ausgestattet, welche lokal systemspezifische Daten sammeln und auf verdächtige Aktivitäten hin analysieren. Es handelt sich hiermit also um ein spezielles host-basiertes IDS, bei welchem auf jedem zu überwachenden System nicht nur Sensoren, sondern gleich ein ganzes IDS eingerichtet wird. Nur die Gateway-Knoten sind jedoch hier schlussendlich dazu befugt Alarme auszulösen, nachdem sie aus den aggregierten lokal gesammelten Sensordaten die entsprechenden finalen Entscheidungen gefällt haben. Es werden also Alarm-Konzentrationspunkte geschaffen, welche die durch die dynamische Topologie von MANETs verursachte "Alarmüberflutung" (z.B. aufgrund mehrfach generierter Alarme) verhindern sollen. Durch die beschriebene Vorgehensweise wird also ein Typ eines vielschichtigen Netzwerk geschaffen, in welchem hier zwischen Gateway-Knoten und sogenannten Intrazone-Knoten (Nicht-Gateway-Knoten) unterschieden wird. Diese zwei Typen von Knoten werden durch die Schaffung von Zonen also in Gruppen (Cluster) zusammengefasst, welche die Aggregation der lokal gesammelten Daten erleichtern sollen. Eine flache Architektur wäre bezüglich Alarm-Management aufgrund der mangelhaften Skalierbarkeit ungeeignet. Andererseits wäre ein strikt hierarchischer Ansatz mit einem einzigen zentralen Knoten in einem MANET aufgrund der entsprechenden dynamischen Topologie prinzipiell nicht durchführbar.

Partitionierung des Netzwerkes Die Partitionierung des Netzwerkes kann gemäss diverser Kriterien erfolgen [25]. Ein relativ einfacher Ansatz (nebst den vielen möglichen Clustering-Algorithmen) ist die Einteilung von Knoten zu einer Zone gemäss ihres gegenwärtigen geographischen Standortes, wobei die jeweilige Lokalität mit einem GPS

[14] In verkabelten Netzwerken ist diese hierarchische Struktur ja bereits intrinsich gegeben

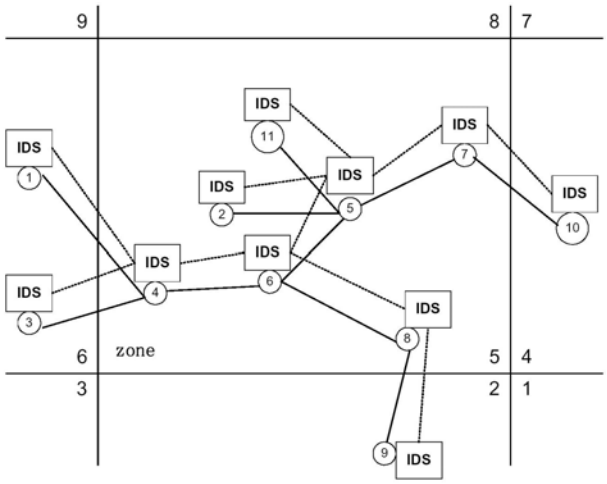


Abbildung 7.4: Partitionierung eines Netzwerkes in sich nichtüberlappende Zonen. In jeder Zone werden Gateway-Knoten ernannt. In Zone 5 wären das beispielsweise die Knoten 4,7 und 8. Diese Gateway-Knoten stehen in Kontakt mit Gateway-Knoten anderer Zonen [25].

(Global Positioning System) bestimmt werden kann. Dies bedeutet, dass die Knoten in einem MANET bedingt durch ihre Mobilität im Laufe der Zeit die Zonenzugehörigkeit bzw. ihre Rolle (Intrazone- oder Gateway-Knoten) wechseln können. Die Zonengrösse wird hierbei gemäss Faktoren wie Mobilität, Dichte und Sendeleistung der Knoten des MANETs gewählt. Die Zonengrösse sollte weder zu gross noch zu klein gewählt werden. Der Nachteil von zu gross gewählten Zonen ist ein hoher Kommunikationsaufwand zwischen den Knoten einer Zone. Zu kleine Zonen bringen den Nachteil, dass dann in einer Zone nicht genügend Informationen für den Aggregationsalgorithmus gesammelt werden können, d.h. eine Gruppierung der lokal festgestellten verdächtigen Aktivitäten wäre nicht mehr sinnvoll.

Gateway-Knoten Die Knoten einer Zone werden nach deren Zuweisung zu einer Zone in zwei Typen eingeteilt [25]: Währenddem Intrazone-Knoten (also Nicht-Gateway-Knoten) dafür zuständig sind, lokal registrierte verdächtige Aktivitäten an Gateway-Knoten derselben Zone zu melden, übernehmen eben diese Gateway-Knoten (auch Interzone-Knoten genannt) mittels Aggregationsalgorithmus die Aggregation und Korrelation der erhaltenen Meldungen der Intrazone-Knoten, wobei Gateway-Knoten benachbarter Zonen kollaborieren können um Intrusion Detection Aufgaben in einem globalen Kontext wahrnehmen zu können. Da lediglich lokal tätige MANET IDS anfällig für eine hohe Falschalarmquote sind (besonders wenn mehrere Angreifer miteinander konsipirieren), sollte eben diese Aggregation integraler Bestandteil eines MANET IDS sein. Um eine gewisse Robustheit des ZBIDS gewährleisten zu können, werden dabei mehrere Gateway-Knoten je Zone ernannt. Wohlgemerkt verfügen nicht nur die Intrazone-Knoten über einen IDS Agenten, sondern auch die Gateway-Knoten. Einerseits beobachten die IDS Agenten dabei lokal und unabhängig voneinander die Systemaktivitäten (Benutzerverhalten, Systemverhalten, Kommunikationsaktivitäten des Systems), andererseits kollaborieren diese Agenten untereinander um den globalen Aspekt der Intrusion Detection wahrzunehmen. Der auf den Gateway-Knoten eingesetzte Aggregationsalgorithmus gruppiert die von den lokal

eingesetzten IDS Agenten registrierten verdächtigen Aktivitäten nach gewissen Kriterien. Schlussendlich fällen aber nur die Gateway-Knoten die finale Entscheidung, wann effektiv ein Alarm ausgelöst werden soll, denn sie vermögen aufgrund ihres besseren Überblickes über die entsprechende Zone die besseren Entscheidungen zu treffen.

Low Level ZBIDS mittels lokaler IDS Agenten

Da es wie erwähnt in MANETs aufgrund des darin vorherrschenden hohen Mobilitätsgrades schwer ist ein IDS zentralisiert zu kontrollieren, wird auf jedem Knoten ein autonom funktionierender lokaler IDS Agent eingesetzt [25]. Um jedoch eine zu hohe Falschalarmrate zu vermeiden, sollen diese IDS Agenten miteinander kollaborieren und ihre registrierten verdächtigen Aktivitäten den Gateway-Knoten derselben Zone melden. Nachfolgend soll nun kurz die Funktionsweise eines solchen lokalen IDS Agenten geschildert werden.

Funktionsweise eine lokalen IDS Agenten Jeder Knoten des mobilen ad-hoc Netzwerkes beteiligt sich an der Intrusion Detection und Response, d.h. individuelle IDS Agenten werden auf jedem Knoten installiert (also nicht nur Sensoren, wie das bei hostbasierten IDS meist der Fall ist) [29]. Jeder IDS Agent läuft hierbei unabhängig und beobachtet lokale Aktivitäten des entsprechenden Systems. Diese individuellen IDS Agenten formen kollektiv das IDS System, welches das MANET vor Angriffen schützen soll. Solche Agenten werden aufgrund ihrer Vorteile (geringe Ressourcenbelastung, flexibel bezüglich des Betriebsystems, updatebar) nicht nur für das ZBIDS, sondern auch für andere Lösungsansätze verwendet.

Es sollen hier kurz die Bestandteile eines IDS Agenten genannt werden (vergleiche auch Abbildung 7.5 unten) [25] [29]:

- **Data Collection Modul:** Diese Komponente ist dafür zuständig, lokale Systemdaten und Aktivitäten (System- und Benutzer-Aktivitäten innerhalb des mobilen Knotens, Kommunikationsaktivitäten dieses Knotens und innerhalb des Empfangsradius) zu sammeln, parsen, filtern und formattieren zwecks nachfolgender Analyse.
- **Detection Engine:** Diese Komponente hat nun die Aufgabe, die gesammelten Daten auf verdächtige Aktivitäten hin zu überprüfen. Bei hohen Sicherheitsanforderungen geschieht dies mit Vorteil mittels verschiedener Erkennungsmethoden (siehe hierzu auch die im Kapitel "Methodik und Architektur von IDS" allgemein besprochenen verschiedenen methodischen Verfahren).
- **Local Aggregation and Correlation Engine (LACE):** Diese Komponente soll die durch die Erkennungsmethoden aufgespürten verdächtigen Aktivitäten lokal aggregieren und korrelieren, d.h. die Erkennungsresultate der eingesetzten Erkennungsverfahren werden kombiniert.
- **Global Aggregation and Correlation Engine (GACE):** Die Funktionalität dieses Moduls hängt hier vom Typ des mobilen Knotens ab: Handels es sich um einen Intrazone-Knoten (also einen Nicht-Gateway-Knoten), ist das GACE-Modul

hauptsächlich für die Informierung der Gateway-Knoten derselben Zone über die lokal aufgespürten verdächtigen Aktivitäten zuständig. Handelt es sich um einen Gateway-Knoten, aggregiert und korreliert das GACE-Modul die Resultate des eigenen und der LACE-Module der Intrazone-Knoten derselben Zone. Das GACE-Modul kooperiert dabei mit GACE-Modulen von Gateway-Knoten benachbarter Zonen.

- **Intrusion Response:** Die Reaktionen auf von Gateway-Knoten generierter Alarme können lokal (z.B. Informierung eines lokalen Nutzers) oder global (z.B. Isolierung eines von mehreren IDS Agenten verdächtigten Knotens) ausfallen (vergleiche hierzu auch die allgemeine Übersicht über Reaktionen eines Intrusion Response Systems am Ende des Kapitels "Methodik und Architektur von IDS").

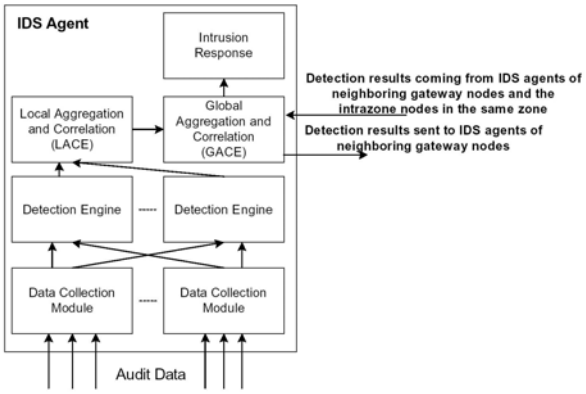


Abbildung 7.5: Schematischer Aufbau eines lokalen Agenten [25].

High Level ZBIDS mittels Aggregationsalgorithmus

Problem Findet keine Kooperation zwischen den Knoten statt (d.h. jeder Knoten versucht unabhängig in lokaler Manier Angriffe aufzuspüren), treten insbesondere in MANETs zu viele Falsch- und Mehrfachalarme auf [25]. Auch bleibt die Erkennungsrate gering, da ausschliesslich lokal agierende IDS sich nur einen unzureichenden Überblick über die globale Angriffssituation verschaffen können. Der in den Gateway-Knoten eingesetzte Aggragationsalgorithmus soll diesem Problem begegnen:

Aggregationsalgorithmus Der in den Gateway-Knoten eingesetzte Aggregationsalgorithmus ist nun dafür zuständig, die eigenen und von den Intrazone-Knoten derselben Zone erhaltenen Informationen über verdächtige Aktivitäten zu aggregieren und korrelieren, das heisst nach gewissen Kriterien zu gruppieren, entsprechende Zusammenhänge festzustellen und dann die finalen Entscheidungen zu fällen, in welchen Fällen effektiv ein Alarm generiert werden soll [25]. Auf diese Weise sollen Falsch- und insbesondere Mehrfachalarme minimiert werden. Auch die Erkennungsrate verbessert sich, denn es lassen sich durch den Aggregationsalgorithmus Angriffe erkennen, die ein einzelnes, unabhängig agierendes IDS womöglich nicht erkennen könnte.

Fazit

Durch die dargestellte Vorgehensweise, das heisst durch die Integration sicherheitsrelevanter Informationen aus einem grösseren Gebiet kann der Aggregationsalgorithmus in diesem Management Framework insbesondere in MANET-Netzwerken die Falschalarmrate erheblich senken und ferner die Erkennungsrate markant steigern [25]. Der in Hinblick auf die Mobilität der Knoten benötigte Kommunikationsaufwand zur Zoneneinrichtung- und erhaltung (es werden hierbei wohlgemerkt keine zusätzlichen Kontrollnachrichten benötigt) wird also durch die viel bessere Leistungsfähigkeit eines solchen IDS mehr als kompensiert. Der Aspekt des Kommunikationsaufwandes ist wichtig, denn das Senden und Empfangen von Nachrichten ist gemessen am entsprechenden Energieverbrauch bei mobilen Knoten sehr teuer. Abschliessend soll bemerkt werden, dass sich die Kollaboration lokal eingesetzter IDS Agenten und der Aggregationsalgorithmus bei der Formung eines kompletten MANET IDS hervorragend komplementieren.

7.4.6 Verteilte Attacken

In den bisher vorgestellten Ansätzen wurde nicht berücksichtigt, dass Attacken von verschiedenen Knoten ausgehen können, wobei die angreifenden Knoten zusammen kooperieren. Ein denkbares Szenario ist, dass mehrere Knoten gemeinsam koordiniert eine Attacke gegen ein Netz starten. Problematisch wird es vorallem dann, wenn einzelne Aktionen eines Knotens in sich legal sind (lokal betrachtet), aber zusammen mit Aktionen von weiteren Knoten eine Attacke darstellen. Eine solche Attacke bringt in Bezug auf IDS mehrere Probleme mit sich.

- Die einzelnen, lokal betrachtet, legalen Aktionen müssen detektiert werden um die Attacke zu erkennen. Dazu müssen die IDS untereinander Informationen über die erhaltenen Pakete austauschen. Denn auch ein Knoten, der bislang noch kein Paket erhalten hat, welches zu der Attacke gehört, kann Angriffspunkt eines weiteren Schrittes sein und muss deshalb informiert werden, dass sich bereits Pakete im Netz befinden, die allenfalls zu einer Attacke gehören. Nur so kann ein Knoten, der als nächstes angegriffen wird (dieser Angriff jedoch lokal als legale Aktion betrachtet wird), erkennen dass er ein weiteres Ziel des Angriffes ist.

- Das wiederum bedeutet, dass jedes Paket an jeden Knoten gesendet werden müsste, sodass jeder Knoten eine globale Sicht auf das Netzwerk hat. Ein Vorteil ist, dass sämtliche Routing Probleme gelöst sind. Praktisch ist dieses Vorgehen in einem grösseren Netz nicht denkbar, da wir uns in einem Umfeld bewegen, dass mit begrenzten Ressourcen auskommen muss (Energie, Rechenzeit, etc.). Weiter könnte eine solche ständige Flutung des Netzes die effektive Kommunikation verunmöglichen, da die meisten Knoten nur noch damit beschäftigt sind Pakete weiterzuleiten und somit nicht mehr die Möglichkeit haben Anfragen zu beantworten oder neue Anfragen zu senden.

- Folglich muss jeder Knoten beim Eintreffen eines Paketes entscheiden ob die Möglichkeit besteht, dass dieses Paket zu einer Attacke gehört. Weiter muss jedes mal

entschieden werden ob es nötig ist die anderen Knoten darüber zu informieren. Es wird also ein Abstraktions-Mechanismus benötigt der entscheiden kann ob das Netz informiert werden soll und wenn dies so ist, eine kurze aber dennoch detaillierte Beschreibung über die mögliche Attacke versendet. Ein Knoten der eine solche Information erhält, sollte danach in der Lage sein gezielt weitere (global gesehen) böswillige Aktionen zu erkennen.

Um eine verteilte Attacke zuverlässig zu erkennen, muss eine globale oder netzweite Sicht auf das Netz von den Knoten zur Verfügung gestellt werden. Um den Workload der einzelnen Knoten zu optimieren, wird eine Abstraktion der Pakete benötigt um die Falschalarmrate niedrig zu halten und trotzdem muss das Netz genügend Informationen erhalten um eine verteilte Attacke zu erkennen.

7.4.7 Zusammenfassung

Das DSR Protokoll oder ein vergleichbares Routing Protokoll dient als Grundlage für die vorgestellten IDS. Mit seiner Hilfe kann in einem unstrukturierten und dynamischen Netz eine Route von einem Quell- zu einem Zielknoten gefunden werden. Dieser Mechanismus kann jedoch leicht missbraucht werden um falsche Routinginformationen zu verbreiten, was für Attacken benützt werden kann. Damit ein Sender sicher sein kann, dass seine Pakete unverfälscht weitergeleitet werden, kann der Watchdog / Pathrater Ansatz verwendet werden. Dabei überprüft die Watchdog-Komponente, dass die Pakete korrekt weiterversendet werden, Pathrater sorgt dafür, dass bei Auffinden eines egoistischen Knotens, dieser nicht mehr für das Routing berücksichtigt wird. Allerdings wird dieser Knoten in seinem Bestreben unterstützt, seine Ressourcen zu schonen. Um Knoten zur Kooperation zu zwingen, werden Bewertungsverfahren verwendet. CORE bewertet die Knoten eines Netzes, ähnlich eines Tauschbörsen-Clients. CONFIDANT kann zusätzlich schlechte Bewertungen abgeben und gegebenenfalls Knoten ausschliessen. Dies kann jedoch auch als Angriff verwendet werden. Routeguard geht noch einen Schritt weiter und führt ein Klassensystem ein. Je nach Klassifizierung haben die Knoten einen unterschiedlichen Handlungsspielraum. ZBIDS partitioniert das Netz, um Aufgaben im Netz zu verteilen. Somit kooperieren Knoten bezüglich der Intrusion Detection und die Knoten entscheiden nicht mehr selber bezüglich den ihnen zugänglichen Informationen. Ein Problem, dass durch die vorgestellten Ansätze nicht gelöst werden kann, sind die beschriebenen verteilten Attacken. Lokal gesehen legale Aktionen können zusammen im netzweiten Kontext eine Attacke sein. Um dieses Problem zu lösen, muss der Workload angemessen auf die Knoten verteilt werden. Auch ein Abstraktionsmechanismus wird benötigt, damit es nicht zu einer Überflutung des Netzes durch Kontrollinformationen kommt.

7.5 Ausblick

Die einzelnen Ansätze sind zum Teil als Prototypen zu Forschungszwecken implementiert und die meisten Systeme wurden auch betreffend Performance mittels Simulationen getestet. Im Einsatz in einer realen Umgebung welche erheblichen Einfluss auf den Empfang

der Knoten hat [16], werden die Resultate der Messungen jedoch anders ausfallen, denn die Bewegungsmuster der Knoten werden komplexer, sowie Hindernisse können die Kommunikation empfindlich stören. Zu bemerken ist ebenfalls, dass sich die vorgestellten Ansätze grösstenteils nur mit Teilproblemen beschäftigen, die sie auch lösen können. Ein IDS für mobile Ad-hoc Netzwerke muss mehrere dieser Lösungsansätze in sich vereinigen. Dabei ist zu beachten, dass sich die verschiedenen Mechanismen nicht gegenseitig behindern.

7.6 Zusammenfassung

Intrusion Detection Systeme (IDS) bilden neben Intrusion Prevention Systemen eine zweite "line of defense". In einem dynamischen und kabellosen Umfeld sind die Anforderungen an ein solches System hoch, da die Ressourcen gering sind und strukturlose Mobilität die Überwachung erschwert. Diverse mögliche Attacken bedrohen Mobile Ad-hoc Netzwerke. Um die Integrität zu erhalten und allfällige bösartige Knoten vom Netz auszuschliessen werden IDS benutzt.

IDS können nach verschiedenen Gesichtspunkten klassifiziert werden. Anhand des Einsatzortes der Sensoren wird zwischen Host-basierten (HIDS) und Netzwerk-basierten IDS (NIDS) unterschieden. Eine weiter Möglichkeit zur Klassifizierung bietet die Methodik, die von den IDS verwendet wird. Signaturbasierte IDS benutzen bereits bekannte Signaturen um Attacken zu erkennen. Sie haben eine niedrige Falschalarm-Rate, können jedoch unbekannte Attacken nicht erkennen. Anomaliebasierte IDS gehen von einem Normalzustand aus und bewerten Abweichungen von diesem als Attacke. Sie sind resistenter gegen unbekannte Attacken, haben jedoch gegenüber von den Signaturbasierten IDS eine deutlich höhere Falschalarm-Rate. Spezifikationsbasierte IDS arbeiten mit Spezifikationen, sie definieren zulässiges Verhalten innerhalb der spezifizierten Grenzen von Programmen oder Protokollen. Werden diese Grenzen überschritten kann eine Attacke angenommen werden. Weiter können IDS nach ihrer Struktur unterschieden werden. Stand-alone, verteilte und kooperative sowie die zentralisierte Architektur wurde kurz behandelt. Wenn eine Attacke erkennt wurde bieten IDS die Möglichkeit sich gegen die Angriffe zur Wehr zu setzen, um weitere Schäden oder ein weiteres Eindringen in das Netzwerk zu verhindern. Die Möglichkeiten der Intrusion Response werden ebenfalls kurz besprochen, sind aber nicht Ziel dieser Arbeit.

Die Lösungsansätze umfassen nicht sämtliche bekannten Ansätze, sondern nur eine Auswahl der bekannteren Systeme. Bei der Diskussion der Ansätze wurde eingehend auf die Routing-Sicherheit eingegangen. Da das Dynamic Source Routing (DSR) Protokoll einfache Möglichkeiten bietet um es für Attacken zu missbrauchen, werden zusätzliche Mechanismen wie Watchdog/Pathrater, Routeguard und andere Bewertungsverfahren (Core, Confidant) eingesetzt um die Routing-Mechanismen zu schützen. Dadurch dass mobile Ad-hoc Netzwerke unstrukturiert sind, erschwert dies ein zuverlässiges Erkennen von Attacken. Der Ansatz des Zone-based Intrusion Detection System, versucht diesen Umstand zu verbessern und dem Netz eine gewisse Struktur zu geben. Die bis dahin besprochenen Systeme und Ansätze sind gut untersucht und es existieren diverse Implementationen, welche zu Forschungszwecken erstellt wurden. Durch von mehreren Knoten ausgehende

Attacken entstehen Probleme wie Alarm-Abstraktion, Organisation und Workload Verteilung welche kurz erläutert wurden.

Die dargestellten IDS-Ansäzte sind wie in der Beurteilung erwähnt wird, nicht gebrauchsfertig implementiert sondern nur zu Testzwecken. Konkrete Erfahrungen in einer realen Umgebung fehlen nach wie vor, da die Versuche in einer simulierten Umgebung stattfinden. Ein IDS muss also mehrere Ansätze in sich vereinigen, ohne dass sich die verschiedenen Mechanismen gegenseitig behindern. Zudem sollte der Ressourcenknappheit entsprechend Rechnung getragen werden.

Literaturverzeichnis

[1] Anjum F., Subhadrabandhu D., Sarkar S.: Signature based Intrusion Detection for Wireless Ad-Hoc Networks: A Comparative study of various routing protocols (IEEE: 01285405.pdf), Februar 2003.

[2] Anderson: Computer security threat, monitoring and surveillance, Technical Report, 1980.

[3] Brutch P., Ko C.: Challenges in Intrusion Detection for Wireless Ad-hoc Networks (IEEE: 01210188.pdf), 2002.

[4] Buchegger S., Le Boudec J.-Y.: Coping with False Accusations in Misbehavior Reputation Szstems for Mobile Ad-hoc Networks., EPFL Technical Report IC/2003/31: `http://icwww.epfl.ch/publications/documents/IC_TECH_REPORT_200331.pdf`

[5] Buchegger S., Le Boudec J.-Y.: Performance Analysis of the CONFIDANT Protocol; Cooperation Of Nodes - Fairness In Dynamic Ad-hoc NeTworks, EPFL Technical Report IC/2002/01; `http://icwww.epfl.ch/publications/documents/IC_TECH_REPORT_200201.pdf`

[6] Cisco Systems, `http://www.cisco.com/en/US/netsol/ns340/ns394/ns171/ns128/networking_solutions_white_paper09186a00801bc111.shtml#wp39639`, November 2006.

[7] Cisco Systems, `http://www.cisco.com/univercd/cc/td/doc/product/iaabu/csids/threat/ctr20/userguid/15289apa.htm`, November 2006.

[8] Computer Base Lexikon, `www.computerbase.de/lexikon/Intrusion_Detection_System`, 16. November 2006.

[9] Da Silva A.P., Martins M., Rocha B., Loureiro A., Ruiz L., Wong H.: Decentralized Intrusion Detection in Wireless Sensor Networks (ACM: p16-da-silva ACM.pdf), Oktober 2005.

[10] Huang Y., Lee W..: A Cooperative Intrusion Detection System for Ad Hoc Networks (ACM: p135-huang ACM.pdf), 2003.

[11] Hasswa A., Zulkernine M., Hassanein H.: Routeguard: An Intrusion Detection and Response System for Mobile Ad Hoc Networks (IEEE: 01512922.pdf), 2005.

[12] Heberlein, Levitt und Mukherjeeh: A method to detect intrusive activity in a networked environment, In Proceedings of the 14th National Computer Security Conference, 1991.

[13] Hijazi A., Nasser N.: Using Mobile Agents for Intrusion Detection in Wireless Ad Hoc Networks (IEEE: 01436049.pdf), Februar 2004.

[14] Institut d'électronique et d'informatique, Université de Marne-la-Vallée: `http://igm.univ-mlv.fr/~dr/XPOSE2004/IDS/IDSSnort.html`, November 2006.

[15] Khoshgoftaar T., Nath S., Zhong S.: Intrusion Detection in Wireless Networks using Clustering Techniques with Expert Analysis (IEEE: 01607440.pdf), 2005.

[16] Klenk A.: MOBILE INTRUSION DETECTION IN MOBILEN AD-HOC NETZWERKEN: `http://net.informatik.uni-tuebingen.de/fileadmin/RI/members/klenk/da_klenk.pdf`

[17] Klenk A.: Mobile Intrusion Detection in mobilen ad-hoc Netzwerken (S.9 bis 23), Universität Ulm, Juni 2003.

[18] Lim Y., Schmoyer T., Levine J., Owen H.: Wireless Intrusion Detection and Response (IEEE: 01232403.pdf), Juni 2003.

[19] Marti S., Giuli T.J., Lai K., Baker M.: Mitigating Routing Misbehavior in Mobile Ad Hoc Networks: `http://www.hpl.hp.com/personal/Kevin_Lai/projects/adhoc/mitigating.pdf`

[20] Mishra A., Nadkarni K., Patcha A., Tech V.: Intrusion Detection in Wireless Ad Hoc Networks (IEEE: 01269717.pdf), Februar 2004.

[21] Noureddine E.: Anwendung für MANETs: `http://www.igd.fhg.de/~pebinger/lectures/adhocnetworks/seminarpapers/01_Anwendungen_fuer_MANETs_Elmarga.pdf`

[22] Schiller J., Mobilkommunikation: Pearson Studium, 2., überarbeitete Auflage.

[23] Search Security Lexikon, `http://searchsecurity.techtarget.com/tip/1,289483,sid14_gci918619,00.html?track=IDSLG`, 16. November 2006.

[24] Stakhanova N., Basu S., Wong J.: A Taxonomy of Intrusion Response Systems, `http://www.cs.iastate.edu/~ndubrov/Response_taxonomy.pdf`, 2006.

[25] Sun B., Wu K., Pooch U.: Alert Aggregation in Mobile Ad Hoc Networks (ACM: p69-sun ACM.pdf), September 2003.

[26] Vigna G., Gwalani S., Srinivasan K., Belding-Royer E., Kemmerer R.: An Intrusion Detection Tool for AODV-based Ad hoc Wireless Networks (IEEE: 01377212.pdf), 2004.

[27] Yang H., Xie L., Sun J.: Intrusion Detection Solution to WLANs (IEEE: 01321948.pdf), Juni 2004.

[28] Yang H., Xie L., Sun J.: Intrusion Detection for Wireless Local Area Network (IEEE: 01347598.pdf), Mai 2004.

[29] Zhang Y., Lee W.: Intrusion Detection in Wireless Ad-Hoc Networks (ACM: p275-zhang ACM.pdf), 2000.

[30] Zhang Y., Lee W., Huang Y.: Intrusion Detection Techniques for Mobile Wireless Networks (ACM: p545-zhang ACM.pdf), 2003.

Kapitel 8

Virus and Spam Threats on Moblie Devices

Patrick Fauquex, Simon Derungs, Martin Schill

In dieser Arbeit werden die Gefahren von Viren, Würmern und Spam, mit speziellem Augenmerk auf mobile Geräte behandelt. Dazu wird erst eine historische Übersicht über die Entstehung und Verbreitung von Viren und Würmern aufgeführt um anschliessend auf die von Ihnen ausgehenden Gefahren detaillierter einzugehen. Dabei wird v.a. auf deren Verbreitungsstrategien sowie den möglichen Schaden den sie anrichten können geachtet. Darauf werden die Auswirkungen von Viren und Würmern auf mobile Geräte beleuchtet. Auch hier wird auf zusätzliche, durch die speziellen Eigenschaften von mobilen Geräten hinzukommende, Verbreitungsmöglichkeiten eingegangen und die Problematik veranschaulicht indem mögliche Attacken auf mobile Geräte aufgelistet und einige Beispiele von Viren und Würmern aufgeführt werden. Auch die Auswirkungen von Viren und Würmern auf die Geschäftsswelt wird genauer betrachtet. Anschliessend werden allgemeine Gegenmassnahmen sowie spezifische Schutzmechanismen für mobile Endgeräte behandelt. Einzelne Massnahmen, welche speziell Unternehmen und somit die Wirtschaft betreffen, werden ebenfalls kurz erwähnt. Bezüglich Spam wird eine Definition gegeben und dessen Ursprung und die Verteilung über die Kontinente im Laufe der letzten Jahre aufgezeigt. Es wird auf die möglichen Distributionskanäle und die verschiedenen Arten von bösartigem Spam eingegangen. Es werden Auswirkungen auf die verschiedenen Akteure erläutert und es wird mit den möglichen Gegenmassnahmen abgeschlossen.

Inhaltsverzeichnis

8.1 Definition Virus und Wurm

Um auf die Gefahren von Viren und Würmern und auf ihre Gegenmassnahmen eingehen zu können, werden diese beiden Begriffe definiert.

8.1.1 Viren

Ein Computervirus ist ein sich selbst reproduzierendes Computerprogramm (eine Sequenz von Instruktionen), welches sich in andere ausführbare Programme einschleust [2]. Viren unterscheiden sich generell durch ihr Existenz-Medium (Dateiviren, Bootfähige Viren, Makroviren, Skriptviren) und durch ihre Infizierungsmethoden.

8.1.2 Würmer

Ein Computerwurm ist ein sich in einem Computernetzwerk selbst verbreitendes Programm [3]. Dies ist dann auch der Hauptunterschied zu einem Virus, welcher für eine Verbreitung auf die "Unterstützung" des Benutzers angewiesen ist. Sie werden aufgrund ihrer Verbreitungsart in Postwürmer (Emailwürmer und Instant Messagewürmer), Internetwürmer und P2P-Würmer kategorisiert.

8.2 Historischer Überblick

Schadprogramme haben sich parallel zu der Informationstechnologie entwickelt. Da sich die Informationstechnologie unglaublich schnell entwickelt hat, haben sich immer wieder neue Möglichkeiten der Schadenszufügung aufgetan. Während früher verschiedene Betriebssysteme und Netzwerke befallen haben, beschränken sie sich heute hauptsächlich auf weit verbreitete Software, wie zum Beispiel Microsoft Windows. Anreize dafür sind vor allem der finanzielle Reiz aber auch die Medienpräsenz und die Reputation.

Am Anfang Die Frage, wann genau der erste Virus aufgetaucht ist, kann man nicht genau beantworten. Fakt ist, dass der allererste Computer, der von Charles Babbadage erfunden wurde, nicht von Viren befallen war. Die ersten theoretischen Ansätze für Computerviren kamen allerdings sehr viel früher auf. So veröffentlichte John von Neumann bereits in den 40er Jahren Arbeiten über selbst reproduzierende mathematische Automaten. Weitere Ansätze, wie denjenigen von Lionel Penrose im Jahre 1959 haben, wenn natürlich auch ungewollt, zur Entstehung der ersten Viren beigetragen.

70er Jahre Anfangs der 70er Jahre tauchte im ARPANET, dem Computernetz der US-Armee und dem Vorgänger des heutigen Internets, der Creeper-Virus auf. Das Programm konnte sich selbständig über ein Modem Zugriff auf einen Computer verschaffen. Auf den Bildschirmen von befallenen Systemen erschien die Nachricht 'I'M

THE CREEPER: CATCH ME IF YOU CAN.' Im Jahr 1975 tauchte ein Computerspiel auf den Markt auf, welches sich mit einer Selbstkorrekturfunktion selber überschreiben konnte. Zusätzlich kopierte das Programm sich selbst in andere Verzeichnisse, sodass nach einer gewissen Zeit sämtliche Verzeichnisse eine Kopie dieses Spiels enthielten. Die endgültige Beseitigung dieses Virus konnte erst mit der Einführung eines neuen Betriebssystems gelöst werden.

80er Jahre Im Zeitalter des Personal-Computer-Boom tauchten auch die ersten Trojaner auf. Sie konnten sich zwar nicht selbst reproduzieren, richteten aber dennoch grossen Schaden an, sobald sie einmal heruntergeladen und installiert wurden. Der Personal-Computer-Boom wurde hauptsächlich von Apple ausgelöst, was sie auch zum Hauptangriffsziel von Virenschreibern machte. Der erste Bootvirus war der Elk Cloner, welcher sich über Floppy-Disks verbreiten konnte und den Bootsektor eines Apple II infizierte. Dieser Virus war nur darum so erfolgreich, weil die meisten Anwender kaum eine Ahnung von Computerviren hatte. Im Jahr 1983 wurde das Computervirus durch Len Adlerman als ein "Programm, welches andere Programme infiziert, indem es sie dahingehend modifiziert, dass sie Kopien seiner selbst installieren" definiert. Die ersten Viren auf IBM Computern tauchten im Jahre 1986 auf. Der Virus Brain wurde von zwei Pakistani geschrieben und verbreitete sich innerhalb von kürzester Zeit weltweit aus. Er infizierte den Bootsektor und das Inhaltsverzeichnis, richtete aber sonst keinen weiteren Schaden an.

1987 Der erste Virus, der Daten beschädigte, war der Lehigh-Virus. Er zerstörte Informationen, die auf Disketten geschrieben waren. Allerdings konnte eine Ausbreitung kurz nach seiner ersten Entdeckung im Jahr 1987 verhindert werden. Ebenfalls im Jahr 1987 gelang es dem Virus Suriv-2, erstmals eine EXE-Datei zu infiltrieren. Der erste Wurm trat ebenfalls im Jahre 1987 in Erscheinung. Der Wurm mit dem Namen Christmas Tree legte das IBM-Vnet-Netzwerk innert wenigen Tagen lahm, indem er unzählige Kopien von sich selbst erzeugte und damit das Netzwerk hoffnungslos überforderte.

1988 Erste Formen von Antivirenprogrammen tauchten im Jahr 1988 auf, allerdings waren diese eher unbekannt und schützen auch nur vor einzelnen Viren. Ende 1988 kam schliesslich die erste bekannte Antivirensoftware "Dr. Solomon's Anti-Virus Toolkit" auf den Markt.

1989 Weitere Viren wie Datacrime, FuManchu und Vacsina lösten immer grössere Epidemien und weltweite Hysterien aus. Die rasante Entwicklung der Virentechnologie und die Nachfrage von diversen Anwendern hatten zur Folge, dass der damalige IT-Marktführer IBM erstmals ein rein kommerzielles Produkt als Schutz gegen Viren anbot.

1990 Auftritt der ersten polymorphen Viren. Die Chamäleon-Viren änderten ihren Quellcode mit jeder neuen Infizierung, was die vorhandenen Antivirenprogramme nutzlos machte. Ebenfalls in diesem Jahr wurde der erste BBS-Server (Bulletin Board System-Server) zum Austausch von Viren eingerichtet, was natürlich die weltweite Produktion von Viren antrieb.

1991 Die Anzahl von Viren stieg nun bereits rasant an. Dies rief auch mehrere Antivirenprogramme auf den Mark, wie zum Beispiel Norton AntiVirus.

1992 Die Anzahl Viren stieg im Jahr 1992 ins astronomische, es kam fast täglich zu Sicherheitsvorfällen. 1992 war auch das Erscheinungsjahr von Win.Vir_1_4, dem ersten Virus für Windows.

1993 Computerviren wurden eine immer ernstere Bedrohung. So traten im Jahr 1993 erstmals Tarnkappen-Viren auf, die sich in dem Code infizierter Dateien verbergen konnten. Microsoft brachte in diesem Jahr erstmals ein eigenes Antivirenprogramm auf den Markt, welches sich durch seine hohe Effizienz auszeichnete.

1994 Im Jahr 1994 traten die komplexen Viren SMEG.Pathogen und SMEG.Queeg auf, welche bis heute noch nicht mit hundertprozentiger Sicherheit aufgespürt werden können.

1996 Anfangs 1996 befiel der Virus Win.Tentacle erstmals das Betriebssystem Windows 3.x. Dieser Windows-Virus ging als erster Windows-Virus in freier Laufbahn in die Geschichte ein. Auch traten in diesem Jahr die ersten Makroviren auf, welche vor allem das MS Excel und das MS Word befielen. Überhaupt wurde Microsoft so langsam immer mehr zum Hauptangriffsziel von Virenschreibern. Es erschienen Dutzende Viren, die die Betriebssysteme Windows 95 und Windows NT sowie das MS Office infizierten.

1997 Jahr 1997 erschienen die ersten Viren für das Betriebssystem Linux. Wäre Linux nur halb so verbreitet wie Windows, so wäre die Anzahl der Linux-Viren vermutlich weitaus größer als es die der Windows-Viren heute tatsächlich ist. Ebenfalls im Jahr 1997 tauchten erstmals Netzwerk-Viren, sowie Viren, die sich via Email und IRC (Internet Relais Chat) auf.

1998 Der Virus Red Team infizierte Windows EXE-Dateien. Eine neue Dimension einer Epidemie löste der Virus Win95.CIH aus, der über populäre Server Spielprogramme infizierte und dadurch auch in der Viren-Hitliste an die Spitze sprang.

1999 Der Wurm Happy99 oder auch Ska war der erste Virus, der sich mit MS Outlook verbreitete, welches sich vor allem in den USA und in Europa bereits zum Standardmailprodukt entwickelt hatte. Neuartige Würmer, welche sich per Email verbreiteten und nur schon durch das lesen des Mails aktiviert wurden, traten ebenfalls im Jahr 1999 auf. Erstmals wurde auch ein Virusautor in Amerika verhaftet und mit einer Gefängnisstrafe von 10 Jahren und einer Geldbusse bestraft. Auch in Taiwan konnte ein weiterer Virenautor identifiziert werden.

2000 Im Jahre 2000 schaffte es der Skriptvirus LoveLetter in das Guinessbuch der Rekorde. Der Virus zerstörte Dateien und verschickte sich selbständig an alle Adressen im MS Outlook. Als erstes Handyvirus ging der Timofonica-Virus in die Geschichte ein, welcher Nachrichten an zufällig ausgesuchte Telefonnummern des Mobilfunknetzes MovieStar verschickte. Das erste Schaden verursachende trojanische Pferd für das PalmOS Betriebssystem von Palm Pilot wurde im August entdeckt. Er zerstörte während der Installation Dateien, konnte sich aber nicht selbständig vervielfältigen.

2001 Mit der steigenden Popularität des Internets vollzog sich auch der Wechsel von den klassischen Viren zu den Würmern. Bekannteste Schadprogramme waren CodeRed, Nimda, Aliz und BadtransII, welche die Ursache für ernsthafte Epidemien waren. Erstmals erschienen auch dateilose Würmer. Diese waren in der Lage, sich selbst zu vervielfältigen und auf infizierten Computern zu laufen, ohne dabei Dateien zu benutzen.

2002 Im Jahr 2002 erschien der Virus Benjamin, welcher vor allem das Dateienaustausch-Netzwerk Kazaa zum Ziel hatte. Mit dem Wurm Klez tauchte ein Schadprogramm auf, welches zwei Jahre lang die Viren-Hitliste anführte. Er war für 3/5 aller registrierten Infizierungen verantwortlich.

2003 Das Jahr 2003 war das Jahr der endlosen Wurmattacken. Anfangs Jahr infizierte der Internetwurm Slammer innerhalb weniger Minuten mehrere Hunderttausend Computer und legte zwischenzeitlich mehrere nationale Internetsegmente lahm. Erstmals erschien auch der Sobig-Wurm, welcher zum Netzwerk-Wurm mit der grössten Ausbreitung seit dem Bestehen des Internets wurde. Weitere Würmer wie Sobig.f, Mimail, I-Worm.Swen und Sobir erschienen ebenfalls im Jahr 2003 und lösten mittlere bis grosse Epidemien aus.

2004 Durch den Wurm Bagle wurden erstmals die Bildung eines Netzes von Zombie-Maschinen bekannt. Die bisher grösste Epidemie in der Geschichte der Viren löste der Wurm Mydoom.a aus. Er enthielt drei Eigenschaften von früheren Viren und infizierte dadurch mehrere Millionen Computer.

Die oben genannten historischen Fakten wurden von [14] entnommen.

Mit dem Boom von mobilen Endgeräten hat sich auch ein neues Angriffsziel für Viren und Würmer aufgetan. In den folgenden Abschnitten werden ausschliesslich Bedrohungen von Viren und Würmern auf mobile Geräte und mögliche Gegenmassnahmen für Benutzer solcher Geräte behandelt.

8.3 Gefahren durch Viren und Würmer

Um die Gefahren, welche von Viren und Würmern ausgehen zu verdeutlichen wird ersteinmal beschrieben, was diese anrichten können. Weiter wird in diesem Teil der Arbeit erläutern mit was für Strategien sie vorgehen und wie sie sich verbreiten. Speziell wird auf die Auswirkungen auf mobile Geräte eingegangen, welche den Viren und Würmer v.a. neue Verbreitungsmöglichkeiten eröffnen. Es werden ebenfalls einige Beispiele aufgeführt. Zudem werden unter Berücksichtigung der immer grösseren Verbreitung dieser Geräte in der Geschäftswelt, die Gefahren von Viren und Würmern innerhalb dieser beleuchtet.

8.3.1 Beschreibung der Gefahren

Viren und Würmer können Computersysteme zum Absturz bringen, einfach “nur” Ressourcen gebrauchen und somit die Performance des ganzen Systems erheblich schmälern,

Daten löschen und somit ganze Computersystem unbrauchbar machen, vertrauliche Dokumente versenden, Systeme ausspionieren um beispielsweise Passwörter zu stehlen oder ganz einfach Computer so zu beeinflussen, dass sie für Angriffe auf Dritte verwendet werden können [1]. Diese Aufzählung, welche keinen Anspruch auf Vollständigkeit erhebt, wird nun durch eine Klassifizierung des schadenanrichtenden Codeteils genauer erläutert wie auch ergänzt.

Beschreibung Payload

Der Teil des Codes eines Viruses oder Wurms, welcher nichts mit der Verbreitung zu tun hat, also der Teil, der die eigentliche "Aufgabe" erfüllt und somit den Schaden anrichtet, wird als Payload bezeichnet. Die verschiedenen Codteile eines Computerwurms respektive -virus sind in Abb. 8.1 ersichtlich.

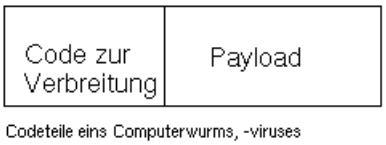


Abbildung 8.1: Codeteile von Computerwürmern, -viren - Quelle: Eigene Darstellung

In [3] werden folgende Arten von Payloads unterschieden.

None / nonfunctional: Es ist kein Payload oder einer, welcher nichts bewirkt, vorhanden. Dieser Typ von Payload kann so gewollt oder durch einen Bug innerhalb des Payloads zustande gekommen sein. Doch auch Viren und Würmer ohne Payload richten Schaden an. Ein Virus kann durch seine Reproduziertätigkeit die Ressourcen des infizierten Systems belasten. Ein Wurm kann durch seine Selbstverbreitung den Datenverkehr massiv belasten.

Internet Remote Control: Dieser Typ von Payload ermöglicht es, an eine infizierte Maschine welche ans Internet angeschlossen ist, über dieses Kommandos zu schicken und so beispielsweise einen Neustart zu veranlassen.

Spam-Relays: Diese Art von Payload wandelt eine infizierte Maschine in ein Spam-Relay, welches von Spammern zur Verbreitung von Spams verwendet werden kann, um.

HTML-Proxies: Web-Anfragen eines Benutzer werden zu einem zufällig ausgewählten Proxy umgeleitet. Auf diese Art können Benutzer auch gezielt zu falschen Seiten umgeleitet werden. So wird auch versucht mit nachgebastelten Bank-Webseiten die Zugangsdaten von Benutzer zu erhalten (sogenanntes Phishing).

Internet DoS: DoS steht für "Denial of Service". Ziel ist es einen oder mehrere angebotene Dienste eines Systems zu überlasten und so unbrauchbar zu machen. Dies wird erreicht indem ein Wurm das mit im infizierte System dazu verwendet, den Dienst eines anderen Systems zu bombardieren. Ist ein Wurm weit verbreitet, kann ein attackierter Dienst eines attackierten Systems schnell überlastet werden.

Data Collection: Wie der Name schon sagt, zielt dieser Payload darauf ab, Daten auf einem infizierten System zu sammeln. Heikle Daten wie Passwörter oder Kreditkarten können hier natürlich von besonderem Interesse sein.

Access for Sale: Dieser Payload ist kurz mit der Frage "Wie mache ich mit Würmern Geld?" zusammenzufassen. Idee ist es, den Zugang zu Computersystemen von bestimmten Opfern, welcher durch Infizierung erlangt wird, zu verkaufen.

Data Damage: Diese Art Payload beabsichtigt schlicht das Löschen von Daten. Sei dies wahllos oder gezielt.

Physical-world Remote Control: Diese kann erlangt werden, indem vernetzte Computer, welche reale Objekte kontrollieren, attackiert werden.

Physical-world DoS: Besitzt ein infizierter Computer ein Modem, können normale Telefonnummern überflutet werden. Beispielsweise Helplines oder Notrufnummern.

Physical-world Reconnaissance: Wieder den Zugriff auf ein Modem der infizierten Maschine vorausgesetzt, können Telefonnummern auf antwortende Geräte gescannt werden und die so erhaltenen Informationen in der wirklichen Welt verwendet werden.

Physical-world Damage: Zielt darauf ab, die Hardware des infizierten Computers zu zerstören. Dies tönt jedoch schlimmer als es ist. Eine Möglichkeit besteht über die Zerstörung des BIOS ein *motherboard* unbrauchbar zu machen und auch dies geht nur wenn der Zugriff darauf nicht geschützt ist.

Worm Maintenance: Die Idee hinter diesem Payload zeigt wie Weit die Entwickler von Würmern bereits denken. Er hat die Aufgabe den Wurmcode auf dem aktuellsten Stand zu halten und sucht zu diesem Zweck laufend nach neueren Versionen des Wurms.

Beschreibung Verbreitungsstrategien

Vor allem die Selbstreproduktion, welche einen Virus per Definition "auszeichnet", sowie die selbständige Verbreitung von Würmern innerhalb von Netzwerken, machen diese zu einer grossen Gefahr. Denn je schneller sich etwas reproduziert und verbreitet, desto schwieriger ist es von diesem ausgehende böswillige Aktionen zu verhindern und den "Übeltäter" zu fangen.

In herkömmlichen Computersystemen werden von Viren und Würmern v.a. folgende Verbreitungsstrategien gewählt:

Verbreitung über E-mail: Würmer nutzen diese Möglichkeit indem sie, sobald sie sich auf einem Computersystem eingenistet haben, die Adressbücher der E-Mail Programme nutzen um ein infiziertes E-mail Attachement an sämtliche im Adressbuch vorhandenen Adressen zu senden. Eine andere Möglichkeit sich via E-mail zu verbreiten ist, sämtliche ausgehenden Nachrichten zu infizieren. Als Beispiel sei hier

der Taripox.B Virus genannt, welcher den Proxy-Server des Client Email-Systems ändert. Der vom Virus kontrollierte neu konfigurierte Proxy hängt nun an jedes ausgehende Mail schädlichen Code an, bevor er das Mail weiterleitet [1].

Verbreitung über Instant Messaging (IM) Programme: Instant Messaging Programme wie ICQ [37] oder MSN Messenger[38], ermöglichen den direkten Dateitransfer zwischen zwei Nutzern dieser Software. Werden nun von Viren infizierte Dateien übertragen, wird der Virus mitübertragen.

Verbreitung via P2P Software: P2P Software wie LimeWire[39] oder Kazaa[40] ist weit verbreitet und bietet sich so geradezu als weitere Verbreitungsmöglichkeit an. Sie ermöglicht den Austausch von Musik- und anderen ausführbaren Dateien. Für Viren und Würmer ist es nun ein leichtes sich als solche in über die P2P Software zugänglichen Ordnern zu tarnen und darauf zu warten, bis sie von einem anderen Benutzer heruntergeladen und ausgeführt werden [1].

Neben diesen Verbreitungsstrategien, welche allesamt auf die Einbettung in Dateien abzielen, können sich Würmer nach [3] noch auf zwei weitere Arten verbreiten. Sie können sich einerseits aktiv selbst verschicken oder über eine zweiten Kommunkationskanal verbreiten. In letzterem Fall, stellt die infizierte Maschine eine Verbindung, beispielsweise über TFTP (Trivial File Transfer Protocol), her, um den Wurm-Body herunterzuladen. Der Hauptteil des Wurms wird somit über einen zweiten Kommunikationskanal übertragen.

8.3.2 Auswirkungen auf mobile Geräte

In diesem Abschnitt wird nun auf die Auswirkungen der beschriebenen Gefahren durch Viren und Würmer auf mobile Geräte eingegangen. Nun, zum einen sind diese, durch dass immer mehr persönliche Daten auf mobilen Geräten gespeichert werden, ein interessantes Ziel für gezielte Attacken. Zum anderen sind sie längst nicht so sicher wie ein fix installierter Desktop Computer. Denn durch limitierte Ressourcen, sei dies Speicherplatz oder Akkuleistung, können Modelle wie im Hintergrund laufende Virenscanner, nicht wie auf einem Desktop Computer implementiert werden. Hinzu kommt, dass sie weitere Verbreitungsmöglichkeiten für Viren und Würmer liefern. Auf diese wird in diesem Abschnitt detaillierter eingegangen. Zudem sollen mögliche Attacken auf mobile Geräte beschrieben werden und einige Beispiele von möglichen Viren für Mobiltelefone die Problematik weiter veranschaulichen.

Zusätzliche Verbreitungsmöglichkeiten durch mobile Geräte

Mobile Geräte bringen auch neue Kommunikationstechnologien mit sich, ohne welche sie gar nicht so mobil sein könnten. Doch diese neuen Technologien haben ihren Preis. Denn je mehr Kommunikationstechnologien ein Gerät unterstützt desto mehr Sicherheitsrisiken besitzt es. Hier werden nun einige dieser Technologien und die damit verbunden Sicherheitsrisiken, welche durch Viren und Würmer missbräuchlich genutzt werden können, aufgezählt.

Wi-Fi: Immer mehr mobile Geräte ermöglichen den Zugang zu WLAN (Wireless Local Area Network). Der gängige Standard 802.11 und das darin verwendete WEP-Protokoll(Wired Equivalent Privacy), offenbaren aber eklatante Sicherheitsmängel, welche es ermöglichen in solche Netzwerke einzusteigen und dort Viren und Würmer zu verbreiten. Diese Sicherheitsmängel sollen hier kurz aufgezeigt werden. Die Mechanismen um Vertraulichkeit, Authentifikation und Zugangskontrolle in Wireless-Netzwerken und somit eine ähnliche Sicherheit wie in verkabelten Netzwerken zu erreichen, sind allesamt übergehbar [11]. Die Vertraulichkeit soll durch eine eindeutige Identität gewährleistet werden. Das einzig Eindeutige eines Teilnehmers innerhalb eines WLAN-Netzes ist jedoch die MAC-Adresse seiner Netzwerkkarte [11]. Diese kann jedoch leicht durch den Gerätetreiber überschrieben werden. Möchte nun also ein Angreifer eine andere Identität vortäuschen, braucht er nur den Netzwerkverkehr nach MAC-Adressen, welche frei versendet werden, zu scannen und schon kann er vorgeben jemand anderer zu sein. Die Zugangskontrolle in WLAN-Netzwerken wird häufig über Listen dieser MAC-Adressen geregelt. Hat ein Geräte eine MAC-Adresse, welche nicht in dieser Liste geführt wird, wird ihm der Zugriff zum Netzwerk verweigert. Diese Art der Zugriffskontrolle kann also leicht Umgangen werden. Eine andere in WLAN-Netzwerken genutzte Art der Zugriffskontrolle, ist der sogenannte *closed network* Ansatz [11]. Dabei geht es darum, dass jeder Teilnehmer im Netzwerk ein Geheimnis kennt und der Zugangspunkt (Access Point) davon ausgeht, dass Teilnehmern welche dieses kennen, der Zugang zum Netzwerk erlaubt sei. Da i.d.R. verwendete Geheimnis ist jedoch der Netzwerkname, welcher die ganze Zeit frei übertragen wird. Diese Art der Zugangskontrolle stellt also ebenfalls keine allzu grosse Hürde für einen Angreifer dar. Die Authentifizierung mit dem WEP-Protokoll in WLAN-Netzwerken funktioniert mit Hilfe eines *shared key* über ein *Challenge-Response* Verfahren. Die hier aufgezeigten Sicherheitsrisiken zeigen, dass es für einen Angreifer ein leichtes wäre, sich Zugang zu einem geschützten Netzwerk zu verschaffen um dort Viren und Würmer zu verbreiten. Es sei jedoch darauf hingewiesen, dass mittlerweile mit WPA eine Alternative zu WEP existiert, die bedeutend sicherer ist.

Bluetooth: Da heutige Bluetooth Geräte nur über eine Reichweite von 10-20 Metern verfügen und sich so beispielsweise ein Wurm nur weiterverbreiten kann, wenn sich ein weiteres Gerät innerhalb dieser Reichweite befindet, liegt der Verdacht nahe, dass sich diese über Bluetooth nicht so schnell ausbreiten können. Das dem aber nicht so ist, soll hier, nach einer kurzen Übersicht über mögliche Bluetooth-Attacken, aufgezeigt werden.

In [12] werden drei Klassen von Bluetooth-Attacken unterschieden.

- Cryptographic Vulnerabilities: Der in [13] erwähnte Versuch, hat gezeigt, dass es in nahezu Echtzeit möglich ist den Bluetooth-PIN zu knacken. Dieser wird von zwei miteinander kommunizierenden Bluetooth Geräten nach dem "handshake" dazu gebraucht ihre Kommunikation zu verschlüsseln. Diese Schwachstelle wird jedoch eher weniger stark zur Verbreitung von Würmern beitragen, offenbart aber eine weitere eklatante Sicherheitslücke von Bluetooth [12].
- Social Engineering-based Attacks: Um diese Art von Attacken zu verstehen sei kurz erwähnt, dass ein Bluetooth Gerät auf zwei Arten eine Verbindung

zu einem anderen herstellen kann. Zum einen durch direkte Adressierung des anderen Gerätes und zum anderen durch das Aussenden einer sogenannten "inquiry" Nachricht, welche bei Empfang durch ein anderes Gerät mit seinem durch den Benutzer konfigurierbaren Gerätenamen und Gerätetyp, beantwortet wird [12]. Damit eine Verbindung zustandekommt, muss der Benutzer eines adressierten Geräts dies zulassen. Als Identifikation des anderen Gerätes dient ihm der eben erwähnte konfigurierbare Gerätename. Wird nun an dieser Stelle ein Name wie "Ein Bewunderer" oder "Ein Freund" gewählt, akzeptiert ein Benutzer eine Verbindung sicher schneller. "Diese Art Attacke ist unter dem Begriff *bluejacking* bekannt" [12].

- Attacks Exploiting Software Vulnerabilities: Schwächen in der Software der Bluetooth Schnittstelle werden ausgenutzt um auf Daten eines Gerätes, welches dies Schwächen offenbart, zuzugreifen.

Damit sich Würmer über Bluetooth nun aber schnell ausbreiten können, müssen aufgrund der limitierten Reichweite einige Voraussetzungen erfüllt werden. In [12] wird gezeigt, dass folgende Voraussetzungen als gegeben betrachtet werden können:

1. Aufspürbare Bluetooth Geräte sind verbreitet
2. Sie sind relativ homogen verteilt
3. Die meisten Geräte bleiben lange genug in der Reichweite eines nach Bluetooth-Geräten scannenden Geräts um infiziert werden zu können
4. Sich mit Schritttempo in Bewegung befindliche Geräte nicht von einer Infizierung geschützt sind

Abbildung 8.2 zeigt, dass sich Würmer auch in Bluetooth-Netzwerken rasch ausbreiten können. Voraussetzung für eine Infizierung ist eine oben als "Attacks Exploiting Software Vulnerabilities" beschrieben Schwachstelle. Die Grafik zeigt die Geschwindigkeit der Verbreitung, wenn 100 Prozent oder nur 25 Prozent der Geräte eine solche aufweisen. Die schwarze Kurve, welche die Verbreitung bei 100 Prozent infizierter Geräten darstellt, bezieht sich auf die linke Achse, die graue Kurve, welche die Verbreitung bei 25 Prozent infizierter Geräten zeigt, bezieht sich auf die rechte Achse. Weiter wurde in [12] bemerkt, dass die Anzahl anfänglich infizierter Geräte zwar die Geschwindigkeit der Ausbreitung beeinflussen, sprich mehr anfänglich infizierte Geräte sorgen auf für eine schnellere Ausbreitung, diese Beeinflussung jedoch sehr gering ist.

SMS und MMS: Aufgrund der limitierten Grösse von SMS (168 Zeichen) sind diese kaum für die Verbreitung von Viren und Würmern zu gebrauchen. Sie treten eher im Zusammenhang mit Spam auf, was weiter unten diskutiert wird. MMS können hingegen bis zu 50 Kbits gross werden und bieten somit genug Platz um Viren und Würmer für Mobiltelefone zu platzieren [7]. Doch obwohl Schwachstellen bei Bildformaten wie JPEG und BMP, welche als MMS übertragen werden können, aufgetaucht sind und diese somit als mögicher Träger von Viren und Würmern hätten genutzt werden können, sind diese nicht genutzt worden [10].

Damit wurde die wichtigste Eigenschaft von mobilen Geräten, welche ebenfalls zur Verbreitung von Viren und Würmern beiträgt, aber noch ausser Acht gelassen: Die Mobilität.

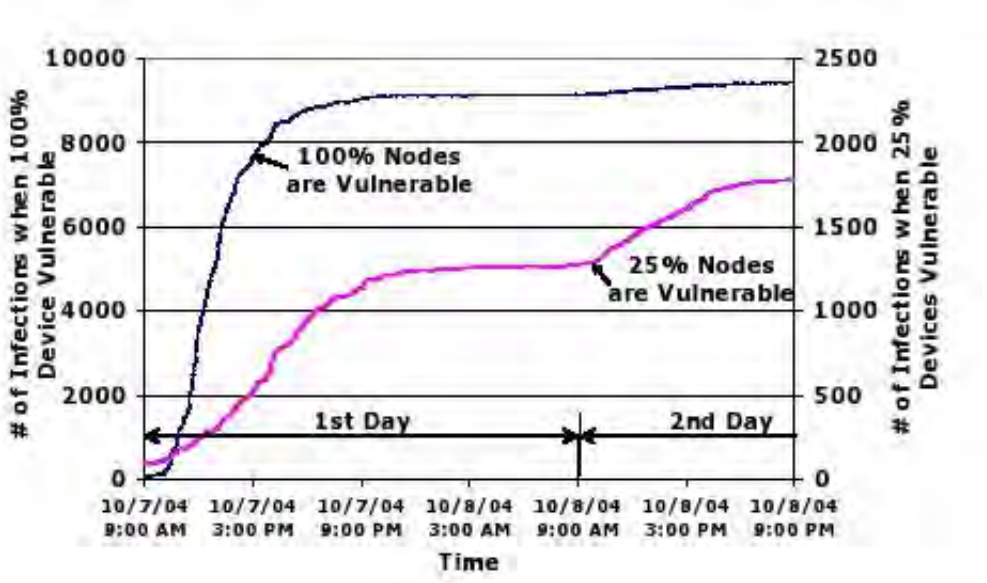


Abbildung 8.2: Würmer Verbreitung in einem Bluetoothnetzwerk - Quelle: [12]

Ein Netzwerk, welches nach aussen geschützt ist, kann von Innen ganz einfach infiziert werden, indem ein infiziertes mobiles Gerät in das Netzwerk und somit hinter die Firewall gebracht wird. So werden Geräte infiziert, welche dann in andere Netzwerke mitgenommen, in welchen die Infizierung weitergeht [4].

Mobile Geräte sind wie bereits erwähnt interessante Objekte für eine gezielte, durch Würmer und Viren unterstützte, Attacke. Dies übrigens nicht zuletzt, da durch das ständige Mittragen dieser Geräte ein subjektives Sicherheitsgefühl entsteht und so viele sensible Daten auf Ihnen gespeichtert werden. Wenige Mobiltelefon-Benutzer dürften wissen, ob ihr Mobiltelefon einen *remote* Zugang erlaubt. Hinzu kommt, dass mobile Geräte in Zukunft Zugang zu immer mehr Objekten ermöglichen werden. So haben Forscher bereits Systeme entwickelt, welche die Haustür öffnen oder den Offen bedienen können [5].

Mögliche Attacken auf mobile Geräte

An dieser Stelle werden nun mögliche Attacken auf mobilen Geräte klassifiziert. Nach [6] können diese wie folgt eingeteilt werden. Im Teil “Auswirkungen auf die Geschäftswelt” werden konkrete Beispiele beschrieben.

Information Theft: Dabei wird zwischen Angriffen auf transiente und statische Information unterschieden. Unter transienter Information ist Information wie der Stromverbrauch oder der Ort an dem sich das mobile Gerät befindet, zu verstehen [6]. Angriffe auf statische Informationen, beabsichtigen an auf dem Gerät gespeicherte Daten heranzukommen. Bei einem Mobiltelefon also beispielsweise die gespeicherten Telefonnummern. Solche Angriffe laufen häufig über Bluetooth oder Wi-Fi. Eine mögliche Attacke ist das sogenannte *bluebugging*. Dabei wird das Gerät des Opfers mit einem Virus infiziert um es so in ein Abhörgerät zu verwandeln [6].

Unsolicited Information: Informationen werden an mobile Geräte geschickt, die diese nicht angefordert haben, beispielsweise SMS-Spam. Auf diese Art von Attacke wird später detailliert eingegangen.

Theft-of-service attacks: Diese zielt darauf ab, die Dienste des infizierten Gerätes für eigene Zwecke zu beanspruchen. Also beispielsweise, teure SMS Nachrichten über ein infiziertes Mobiltelefon zu senden.

Denial-of-service attacks: Was unter "Denial-of-service" zu verstehen ist, wurde weiter oben schon erläutert. In Zusammenhang mit mobilen Geräten kann eine Dienst nicht nur unbrauchbar gemacht werden indem er überlastet wird, sondern auch indem im die Energie genommen wird. Beispielsweise wird darauf abgezielt, den Akku eines mobilen Gerätes möglichst schnell aufzubrauchen.

Beispiele von Viren und Würmern für Mobiltelefone

Um die Problematik von Viren und Würmern im mobilen Umfeld weiter zu veranschaulichen, wird nun auf einige Viren und Würmer, welche für die Infizierung von Mobiltelefonen entwickelt wurden, eingegangen. Viren und Würmer müssen auf ein Betriebssystem ausgerichtet sein. Während bei Computern Windows einen Marktanteil von über 90 Prozent hat, verteilt sich dieser auf dem Markt für Betriebssysteme für Mobiltelefone homogener. Neben Palm und Windows CE ist Symbian das verbreitetste Betriebssystem [7]. Viren und Würmer für Mobiltelefone werden in der folgenden Aufzählung nach Plattform, der Verbreitungsart und den Auswirkungen auf das System nach einer Infektion durch den jeweiligen Wurm oder das jeweilige Virus, charakterisiert. Die hier aufgelisteten Viren und Würmer sind in [7] detailliert beschrieben.

Cabir Plattform: Symbian Series 60, verwendet auf Geräten von Motorola, Nokia, Panasonic und Sony Ericsson
Verbreitungsart: Via Bluetooth oder eine gemeinsam genutzte Applikation. Ist ein Gerät infiziert, nutzt der Wurm dessen Bluetooth Schnittstelle um sich weiter zu verbreiten. Zu diesem Zweck wird die Umgebung ständig nach anderen Bluetooth-fähigen Geräten gescannt. Ist ein Gerät gefunden, versucht sich der Wurm auf dieses zu übertragen. Akzeptiert der Benutzer des attackierten Gerätes den Empfang einer Nachricht über Bluetooth wird die Wurm-Datei übertragen.
Infizierung: Über ein eine .SIS (Symbian installation system) Applikations-Installations Datei.
Auswirkungen: Ist ein Mobiltelefon mit Cabir infiziert, wird auf dem Display ein Text wie z.B. "Caribe VZ/20a" angezeigt. Der genaue Wortlaut des Textes hängt von der Version des Wurms ab. Zudem wird so gleich die Bluetooth-Schnittstelle des Geräts in Beschlag genommen, um sich weiter zu verbreiten.

Skulls Plattform: Symbian Series 60, auf Nokia 7610 ausgerichtet
Verbreitungsart: Skulls ist als nützlich Applikation getarnt und somit ein sogenanntes trojanisches Pferd. Somit wird Skulls vom Benutzer selbst installiert. Die Verbreitung findet also über das Internet statt, von wo aus es heruntergeladen wird und mit den dafür eingerichteten Schnittstellen vom Computer auf das Mobiltelefon

übertragen.
Infizierung: Durch Installation der Applikation die Skulls vortäuscht zu sein.
Auswirkungen: Sämtliche Funktionen mit Ausnahme der Möglichkeit Anrufe zu tätigen und zu empfangen, werden unbrauchbar gemacht. Also das Dateimanagement, Bluetooth, SMS, Web-Funktionen sowie die Möglichkeit Applikationen zu installieren und zu entfernen. Somit kann Skulls auch nicht mehr deinstalliert werden. Es kommt nur noch ein kompletter Reset des Gerätes in Frage, womit sämtliche darauf gespeicherten Daten verloren gehen.

Mquito Plattform:Symbian Series 60
Verbreitungsart:Mquito ist eine modifizierte Version des verbreiteten Mosquito Spiels und kann somit auch als eine Art trojanisches Pferd verstanden werden. Die Verbreitung findet somit wie auf die bei Skulls beschriebene Art über das Internet statt.
Infizierung: Ein Gerät ist infiziert sobald Mquito installiert wurde.
Auswirkungen: Eine ursprünglich zum Copyright-Schutz des Spiels implementierte SMS-Funktion wird genutzt um kostenpflichtige SMS zu verschicken.

WinCE.Duts.A Plattform: Windows CE
Verbreitung: Durch die Übertragung infizierter Dateien.
Infizierung: Hängt sich an alle ausführbaren Dateien die grösser als 4,096 bytes sind an. Dies indem bei Ausführung einer infizierten Datei der Virus ausgeführt wird und somit noch nicht infizierte Dateien infiziert.
Auswirkungen: Belastung des Systems durch Verbreitung innerhalb dessen. WinCe.Duts.A besitzt keinen Payload.

Lasco Plattform: Symbian 60 Series
Verbreitungsart: Via Bluetooth analog Cabir. Zusätzlich infiziert Lasco weitere .SIS Dateien und wird so durch den Austausch dieser weiter verbreitet. Was zu einer erheblich schnelleren Verbreitung führt.
Infizierung: Über ein eine .SIS Applikations-Installations Datei.
Auswirkungen: Die Bluetooth Schnittstelle ist durch die ständigen Weiterverbreitungsbemühungen des Wurms ausgelastet. Dies führt dazu, dass der Akku erheblich schneller aufgebraucht wird.

Diese Beispiele zeigen auf, dass sich Viren und Würmer für Mobiltelefone auf verschiedenste Arten zu verbreiten wissen und sie erheblichen Schaden anrichten können. Aber auch, dass der Faktor Mensch eine entscheidende Rolle bei deren Verbreitung spielt. Akzeptiert ein Benutzer die durch Cabir und Lasco verbreitenden Bluetooth-Nachrichten nicht, ist dieser Verbreitungskanal für diese blockiert. Prüfen Benutzer vom Internet heruntergeladenen Applikationen für ihre Mobiltelefone erst auf Viren und Würmer für Mobiltelefone, können sich diese, sofern sie bereits bekannt sind, auf diesem Weg auch nicht mehr weiter ausbreiten.

8.3.3 Auswirkungen auf die Geschäftswelt

Die Folgen der Verbreitung von Viren und Würmern bleiben natürlich nicht ohne Auswirkungen auf die Geschäftswelt. Zum einen durch ihre blosse Verbreitung und den Schaden

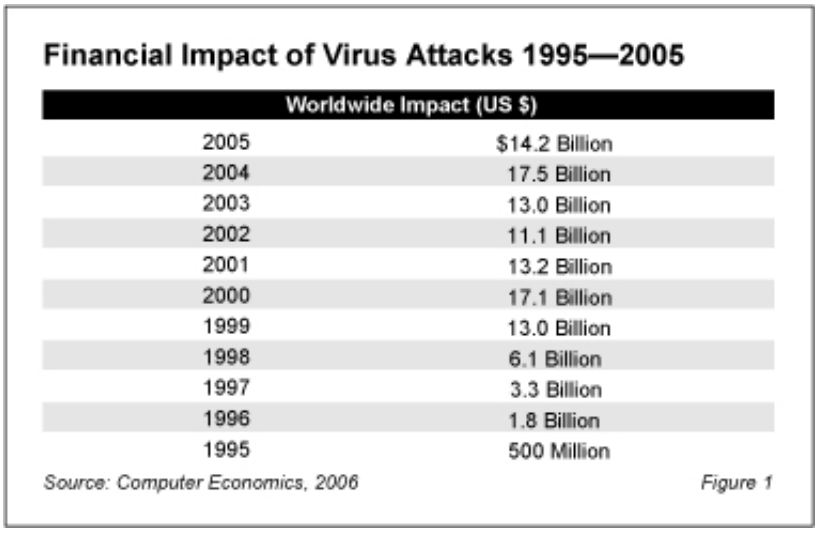
den sie anrichten können - was das sein kann wurde im Abschnitt "Beschreibung der Gefahren" aufgezeigt. Zum anderen können Viren und Würmer durch gezielte Attacken zum kommerziellen Vorteil eingesetzt werden. Firmen die stark von Internet basiertem Verkehr abhängig sind, können von einer gezielten Attacke stark betroffen sein hier. Hinter einer solchen über das Internet geführten Attacke kann der Gedanke stehen, Finanzmärkte zu manipulieren. Dies durch ein, mit einer gezielten Attacke ausgelöstes, wirtschaftliches Desaster. Ein weiterer möglicher Grund für einen Angriff kann sein, dass dadurch der Kundenzugriff auf das Angebot eines Konkurrenten reduziert werden kann. [3] Gezielte Attacken sind aber auch im Hinblick auf mobile Geräte, welche in der Geschäftswelt immer mehr an Bedeutung gelangen, interessant. Diese, gegen mobile Geräte gerichteten, gezielten Attacken bieten einen Möglichkeit an Informationen eines Konkurrenten heranzukommen. In diesem Abschnitt werden nun die wirtschaftlichen Auswirkungen von Viren und Würmern beleuchten und gezeigt wie gezielte Attacken auf mobile Geräte genutzt werden könnten, um sich gegenüber der Konkurrenz einen Vorteil zu verschaffen.

Wirtschaftliche Auswirkungen von Viren und Würmern

Attacken durch Viren und Würmern können erhebliche finanziell Konsequenzen haben. Die weltweite Entwicklung dieser ist in Abbildung 8.3 dargestellt. Diese Kosten setzten sich nicht nur aus den direkten Kosten, welche durch den Ausfall eines Dienstes oder die Behinderung der Mitarbeiter bei der Arbeit durch einen Systemausfall entstehen, sondern auch aus Folge- oder Präventionskosten, zusammen. Dies sind laut [8] sämtliche für die Absicherung der Systeme und Netze gegen solche Attacken entstehende, sowie die zur Säuberung und Wiederaufbereitung von infizierten Systemen anfallende Kosten. Die mit diesen Punkten verbundenen Personal- und Beratungskosten sind eingeschlossen. Wenn man die Zahlen der letzten beiden Jahre vergleicht ist ein Rückgang der Gesamtkosten zu erkennen, was eigentlich als positiv gewertet werden kann. Dies ist jedoch zu relativieren wenn man sich die Gründe für diesen Minimierung vor Augen hält. Dies sind nach [8] auf der einen Seite die verbesserte Netzwerkinfrastruktur der Firmen, sowie grosse Fortschritte in der Sicherheitstechnologie wie Antiviren-Software, so dass traditionelle Attacken nicht mehr so ins Gewicht fallen. Auf der anderen Seite finden weniger generelle Attacken, die Viren und Würmer wahllos im Internet verbreiten statt, dafür immer mehr gezielte Angriffe auf Firmen. So kommt es, dass die Gesamtkosten für die Wirtschaft gesunken, diejenigen für eine attackierte Firma jedoch massiv gestiegen sind [8].

Auf mobilen Geräten basierende gezielte Attacken

Die gezielte Attacken sind wie oben erwähnt, im Vergleich zu den generellen, massiv gestiegen. Hier sollen nun mögliche solcher Attacken, unter Berücksichtigung des Einbezugs mobiler Geräte, betrachtet werden. Unter "Mögliche Attacken auf mobile Geräte" im Abschnitt "Auswirkungen auf mobile Geräte" wurden gezielte Attacken auf mobile Geräte bereits klassifiziert. Nun wird gezeigt wie diese im Geschäftsumfeld eingesetzt werden könnten. Hier sind vor allem als *Information Theft* klassifizierte Attacken interessant. Beispielsweise das bereits erwähnte *bluebugging*, bei dem ein Mobiltelefon in ein Abhörgerät umgewandelt werden kann. Wird das Mobiltelefon auf stumm, der Vibrationsmodus



Financial Impact of Virus Attacks 1995—2005

	Worldwide Impact (US $)
2005	$14.2 Billion
2004	17.5 Billion
2003	13.0 Billion
2002	11.1 Billion
2001	13.2 Billion
2000	17.1 Billion
1999	13.0 Billion
1998	6.1 Billion
1997	3.3 Billion
1996	1.8 Billion
1995	500 Million

Source: Computer Economics, 2006 *Figure 1*

Abbildung 8.3: Wirtschaftliche Kosten von Viren und Würmern - Quelle: [8]

aus und der "Automatisch Antworten" Modus aktiviert - ist das nicht gleichzeitig möglich, kann der Rufton einfach auf sehr leise gestellt werden -, wird das Mobiltelefon einen Anruf ohne Intervention des Benutzers annehmen [6]. So ist es möglich das Mobiltelefon des Managers eines Konkurrenten über Bluetooth mit einem Virus zu infizieren, der diese Einstellungen vornimmt. So kann dieser später belauscht werden. Um an sensible Informationen heranzukommen muss *bluebugging* aber nicht einmal angewendet werden, denn schon auf den mobilen Geräten selbst befinden sich massenhaft sensible Informationen. Es gibt sogar Benutzer die ihre Kreditkartennummer auf ihrem Mobiltelefon speichern [7].

Auf ein weiteres bereits angesprochenen Sicherheitsrisiko ist in der Geschäftswelt ebenfalls zu achten. Die Mobilität von mobilen Geräten. Den diese können auch ganz einfach gestohlen werden. Eine Firma muss sich in diesem Zusammenhang, nicht nur auf gezielte Attacken konzentrieren, sondern auch die auf Vergesslichkeit ihrer Mitarbeiter achten. Den diese lassen ihre Geräte schnell einmal irgendwo liegen. "Reisende in Eile liessen über 6 Monate 62'000 Mobiltelefone, 2900 Laptops und 1300 PDAs in Londons Taxis liegen" [9]. Die Wahrscheinlichkeit, dass auf diesen Geräten wichtige geschäftliche Informationen lagen ist sehr gross. Den 97 Prozent der von Mitarbeitern innerhalb der Organisation genutzten tragbaren Geräte befinden sich in privaten Besitz d.h. gehören den Mitarbeitern und nicht der Organisation[9]. Es ist auch ein leichtes diese Geräte solange sie sich ausserhalb der Firma befinden zu infizieren und so Viren und Würmer, die zur Zerstörung von Firmendaten oder gezielten "Information Theft" Attacken gedacht sind, ins Firmennetzwerk einzuschleusen. Dies kann auf über gestohlene Geräte erreicht werden. Den die meisten Firmen Laptops sind, damit Mitarbeiter auch von zu Hause aus arbeiten können, konfiguriert die Firmen-Firewall zu tunneln [5].

8.4 Überblick über die Gegenmassnahmen

8.4.1 Allgemeine Massnahmen

Eine Absicherung mobiler Geräte kann nur durch eine sinnvolle Kombination von Schutzmassnahmen sichergestellt werden. Die Absicherung kann gemäss [15] in drei verschiedene Phasen unterteilt werden:

1. Präventiver Schutz: Damit sind sämtliche Sicherheitsmassnahmen gemeint, die im Voraus ohne Bezug auf einen konkreten Angriff zur Erhöhung der Sicherheit beitragen.Dazu zählen z. B. Schulung der Mitarbeiter, eine Unternehmenspolicy, Installation von Verschlüsselungssoftware, etc.

2. Angriffserkennung: Damit sind Sicherheitsmaßnahmen gemeint, die Angriffe erkennen und melden. Dazu zählen das Gerätemanagement zur Meldung von Schwachstellen und versuchten Angriffen, die Anzeige verloren gegangener Geräte oder die Analyse von Logfiles.

3. Wiederherstellung und Reparatur eines kompromittierten Systems: Nach einem Angriff müssen ebenfalls Maßnahmen eingeleitet werden. Dies sind meist organisatorischen Maßnahmen, welche die Wirkung des Angriffs und erfolgte Schäden feststellen und beseitigen.

8.4.2 Lösungsansätze für mobile Geräte

Sind die möglichen Bedrohungen gegen mobile Geräte bekannt, so ist es möglich, Informationen besser zu schützen und sich somit besser auf Gefahren vorzubereiten. Dennoch ist es nicht möglich, einen kompletten Schutz zu gewährleisten.

In [6] werden allgemeine Massnahmen aufgezeigt, die dazu beitragen sollen, dass mobile Geräte gegen bösartige Angriffe besser geschützt werden können. Dabei sind vor allem Geräte (Mobiltelefon, Smartphone, PDA,...) gemeint, bei welchen eine starke Interaktion mit Menschen zu Grunde liegt. Diese Massnahmen sollen vor allem zur Minimierung der Risiken beisteuern:

Aufklärung: Eine der wichtigsten Schutzmassnahmen ist die präventive Aufklärung von Kunden, Angestellten und Institutionen. Sie sollten auf die verschiedenen möglichen Attacken gegen mobile Geräte sensibilisiert werden. Es soll ihnen bewusst sein, dass ein mobiles Gerät nicht mit dem Internet verbunden sein muss, um ein potentielles Ziel eines Angriffs zu sein. Nutzer von mobilen Geräten sollten mit gesunder Vorsicht mit den Geräten umgehen und nicht persönlich Information “offen” auf ihren Geräten ablegen.

Visualization: Es sollten Statistiken visuell bereitgestellt werden, welche den Usern kritische Grössen wie Batteriestand, Stromverbrauchsrate, Datenübertragungsrate und Prozessorbelastung aufzeigen. Dadurch können User selbständig auf mögliche Probleme aufmerksam gemacht werden und frühzeitig reagieren.

Profiling: Service Provider können typische Aktivitätprofile von Usern erstellen, um dadurch unbefugten Zugriff auf deren mobile Geräte zu entdecken. Mittels Nachrichten können Service Provider die Benutzer auf solche Zugriffe aufmerksam machen und damit auch den Benutzer schützen.

Hard Switches: Werden Stromversorgungen benützt, welche physikalisch trennbar sind von mobilen Geräten, so kann sichergestellt werden, dass sie auch wirklich abgestellt wurden.

Heterogenität: Eine Plattformverschiedenheit bietet ein inhärentes Schutzlevel vor Viren und anderen Bedrohungen (Bsp.: Betriebssystemmarkt für mobile Geräte: Symbian, Palm OS, ...). Dadurch wird eine mögliche Verbreitung von Schadcode verlangsamt.

Die oben erwähnten Punkte stellen vor allem allgemeine Massnahmen dar, nun sollen auch noch ein paar angewandte Schutzmöglichkeiten erwähnt werden. In [5] werden folgende Massnahmen aufgezeigt:

- Kostensenkung für Zugriff und Unterhalt von Systemen. Sogenannte “key rings” verwalten Benutzerkennworte über verschiedene Webseiten. Den Benutzern reicht ein einziges Passwort, um auf sämtliche Dienste zurückgreifen zu können (z.B. Zugriff auf spezielle Dienste für Kunden, welche nicht für jedermann zugänglich sind).
- Zwei-Faktoren Authentifikation - Passwort plus Sicherheitscode, welcher nach jedem Login wechselt (z.B.: Beim E-Banking wird oft die Zwei-Faktoren Authentifikation und “key rings” verwendet, um die persönlichen Daten der Kunden zu schützen).
- Anstelle von Passwörtern wird eine biometrische Erkennung eingesetzt. (Bsp. PDAs und Laptops mit Fingerabdruckerkennung) Dadurch kann ein feindlicher Zugriff auf mobile Geräte verhindert werden.
- Die Systemkonfiguration sollte einfacher gestaltet werden, sodass jeder Nutzer sie selber und ohne Probleme ausführen kann.
- Die Software aktualisiert sich automatisch und selbständig. Die Aktualisierung findet ohne Einfluss des Nutzers statt. Somit liegt die Verantwortlichkeit für die Aktualisierung nun beim Gerätehersteller sowie beim Service Provider.
- Ein anderer Ansatz wäre, dass mobile Geräte regelmässig an einen physikalischen Ort gebracht werden müssten, um sicher zu gehen, dass nicht unerlaubter Weise auf das Gerät zurückgegriffen wird (Vor allem nützlich bei Diebstahl).
- Es sind persönliche Infos über Interaktionspartner sichtbar, sodass die User sehen, wer mit ihnen verbunden ist. Dadurch können sie dubiose Personen gleich von Anfang meiden und somit ihre Daten besser schützen.

- Es könnte auch ein Trackingsystem (wie z.B. bei Autos) implementiert werden. Würde sich ein gestohlenes Gerät mit einer Telefonleitung oder dem Internet verbinden, könnte es lokalisiert werden.

Lösungsansätze für Bluetooth

Es wurde prophezeit, dass in ein paar Jahren Geräte mit Bluethooth-Funktion die Wi-Fi-Geräte um das fünffache übertreffen werden (77% aller Mobiltelefone, 60% aller PDAs und 67% aller Notebooks werden eingebaute Bluetooth-Radios haben). Es ist klar, dass deshalb Schutzmassnahmen für Bluetooth von grosser Bedeutung sind. Verteidigungslösungen sollten vor allem an hochfrequentierten Lokalitäten angeboten werden, so dass eine erhöhte Ausbreitungsgeschwindigkeit von Viren und Würmern verhindert werden kann. Zum Beispiel an verkehrsintensiven Orten wie Flughäfen können solche Lösungen einen grösseren Wurmausbruch verhindern.

Die heutigen Bluetoothgeräte beinhalten oft eine non-discoverable-Funktion. Hat ein Gerät diese Funktion aktiviert, so reagiert das Gerät nicht auf empfangene Bluetooth-Anfrage-Nachrichten. Allerdings wird dadurch der Gebrauch von Bluetooth eingeschränkt resp. stark unterbunden, was eigentlich nicht Sinn der Sache sein sollte. Lösungen sollten vor allem die Präsenz von Bluetoothwürmer entdecken, den infizierten Code analysieren und einen security Patch innerhalb einer gewissen Zeit erstellen, testen und verteilen. Allerdings sollte die Verteilung solcher Patches während der Nacht oder am Wochenende stattfinden, um am Tag nicht Kommunikationskanäle zu besetzen und unnötige Speicherkapazitäten zu. Diese Randzeiten würden genügen, um solche Updates automatisch zu übertragen, zu installieren und zu testen.

Ebenfalls sinnvoll wäre ein Monitoringsystem, welches an verkehrsintensiven Orten (Flughäfen, Bahnhöfe) frühzeitig eine Wurminfektion erkennen kann. Wird ein Wurm an einem solchen Ort frühzeitig erkannt und unter Quarantäne gestellt, so kann er sich nicht beliebig verbreiten und sich schlussendlich weltweit verbreiten.

Lösungsansätze Mobiltelefone

Nach [7] ist ein Hauptproblem bei Mobiltelefonen, dass sie meistens keine Antivirus Software besitzen. Für Hacker sind sie vor allem deshalb so interessant, weil weltweit mehrere Millionen Personen als potentielle Opfer gelten.

Ein möglicher Vorteil gegenüber Hackern ist allerdings, dass es mehrere verschiedene Betriebssysteme gibt. So müssen für sie für jedes dieser Betriebssysteme einen eignen Code schreiben, um eine sich schnell verbreitende Infektionswelle auslösen zu können. Momentan gilt Symbian hauptsächlich als gefährdet, da dieses Betriebssystem die grössten Marktanteile besitzt.

Von Vorteil wäre es, wenn die Geräteanbieter und Service Providern effiziente Antiviren- und Sicherheitsprogramme anbieten würden. Antispyware und Antivirus Funktionalität soll den Nutzern von mobilen Geräten helfen, resistenter gegenüber potentiellen Angreifern

zu sein. Die Nutzer sollten ein gesundes Sicherheitsgefühl entwickeln, so dass garantiert werden kann, dass Daten und Kommunikationen weiterhin sicher bleiben.

Bereits sind erste Viren auf mobilen Geräten aufgetaucht, was auch die Anti-Virus Unternehmen wie F-Secure dazu gebracht hat, Produkte für Mobiltelefone zu entwickeln und anzubieten. F-Secure benützt SMS und MMS um Updates für Virusdefinitionen zu verteilen. In [7] werden folgende Schutzmassnahmen als zentral betrachtet:

Aufteilung von Sprache und Daten Um einen effektiven Schutz gewährleisten zu können, ist es sinnvoll, die Sprach- und Datenkommunikation aufzuteilen. Dadurch kann vermieden werden, dass Daten in zelluläre Netzwerke gelangen und dadurch die Sprachübertragung beeinträchtigen. Eine solche Aufteilung sollte sowohl in verdrahteten Netzwerken wie auch in drahtlosen Übertragungsschnittstellen stattfinden. Dadurch kann aber nicht garantiert werden, dass Luftschnittstellen überflutet werden können. Um den Verkehr zu gestalten, sollten verschiedene Nachrichten unterschiedliche Prioritäten erhalten. Daher sollten Textnachrichten klassifiziert werden können. Textnachrichten von ausserhalb des Netzwerkes (z.B.: vom Internet ausgehend) sollten eine niedrige Priorität für die Datenkanäle erhalten während Textnachrichten innerhalb eines Telefonnetzwerkes eine höhere Priorität zugewiesen bekommen.

Ressourcen vorbeugend bereitstellen (Resource Provisioning) Bei speziellen Anlässen können zusätzliche Ressourcen bereitgestellt werden, um mögliche DoS-Attacken zu vermeiden. Als Beispiel sind hier die Olympischen Spiele in Griechenland zu nennen, bei der das gesamte Telefonnetz beträchtlich erweitert wurden (Basisstationen und MSC), sodass es möglich war, während 17 Tagen über 100 Mio. Textnachrichten erfolgreich zu versenden. Bei SMS, die aus dem Internet verschickt werden, kann mit einer Kapazitäterweiterung in kritischen Gebieten ähnliches erreicht werden. Allerdings kostet die Bereitstellung zusätzlicher Ressourcen meistens sehr viel, sodass diese Lösung oft nicht in Frage kommt.

Ratenbeschränkung (Rate Limitation) Eine Möglichkeit wäre, drahtlose Übertragungskanäle zu beschränken. Dies ist allerdings keine gute Lösung, da dadurch DoS-Attacken weiterhin möglich sind. Zusätzlich würde auch die Übertragungsgeschwindigkeit legaler Textnachrichten deutlich langsamer werden. Ein weiteres Problem ist die (Un)Zuverlässigkeit von Nachrichten, die aus dem Internet verschickt wurden. Hilfreich wäre es, wenn die Anzahl der Empfänger pro SMS limitieren würden. Damit könnte verhindert werden, dass ein Netzwerk überflutet werden kann.

Aufklärung (Education) Die ersten drei erwähnten Punkte befassen sich hauptsächlich mit Massnahmen für DoS-Attacken. Allerdings nützen diese wenig gegenüber Phishingangriffen. Hierzu genügt oft nur eine Nachricht/ Anfrage ausgehend von einem Angreifer, um an die gewünschten Daten seiner Opfer zu kommen. Ebenso kann ein Virus immer noch mobile Geräte beschädigen, wenn er in ein spezifisches System über verschiedene Benutzerinteraktionen gelangt. Die einzig mögliche Lösung hierfür bietet eine Aufklärung der Nutzer. Dadurch soll diesen bewusst werden, worauf sie achten sollten und wie sie sich schützen können.

Für PCs gibt es bekannterweise mehrere Antivirenlösungen, während diese für Handys noch in ihrer "Kindheit" befinden und somit noch nicht wirklich verbreitet sind. In Ländern

wie Japan hingegen, wo Viren, Würmer und vor allem Spam ein ernsthaftes Problem für Handys sind, ist bereits eine fortgeschrittene Antivirus Software erwerbbar. (Bsp.: McAfee's VirusScan auf Symbian basierende Telefone; bei Nokia mittels Symantec Client Security Software; Trend Micro Mobile Security bietet Antispam- und Antivirenlösungen für SMS Dienste an).

Antivirensoftware für mobile Geräte sind im Prinzip stark mit mit den Antivirensoftware für PCs zu vergleichen, allerdings müssen sie einfacher gestaltet werden, da mobile Geräte mit weniger Speicherplatz und Performance auskommen müssen. Ebenfalls sind die Betriebssysteme und die Viren einfacher, da nur begrenzt Ressourcen zur Verfügung stehen.

8.4.3 Lösungsansätze für die Wirtschaft

Gemäss [9] sind Schutz und Sicherheit von mobilen Geräten ein wichtiger Bestandteil der Sicherheitspolitik von Unternehmen. Sie stellen aufgrund ihrer Grösse und ihrer Speicherkapazität insofern eine Bedrohung dar, als dass Informationen einfach und schnell heruntergeladen und anschliessend ohne Probleme aus der Unternehmung geschafft und weitergegeben oder verkauft werden können.

Generelle Schutzmassnahmen

Alle mobilen Geräte, die sensitive oder proprietäre Daten speichern, ausführen oder übertragen können, sollten Schutzmechanismen haben, welche der Gerätekapazität angepasst sind. Das Gerät selber und die Übertragungswege müssen geschützt werden. Gemäss [9] sind die unten genannten Punkte zentral für jedes Unternehmen:

- Es sind komplizierte Passwörter zu verwenden, die alpha-numerische Zeichen benützten und mindestens 8 Zeichen lang sind.
- Installation von Softwareschutzmechanismen, welche die Inhalte auf mobilen Geräten verschlüsseln, das Gerät nach einer gewissen Zeit in den Standby-Modus überführt (Anmeldung mittels Passwort dann wieder erforderlich) und welche die persönlichen Daten auf der Festplatte löschen, wenn das Passwort mehrmals nacheinander falsch eingegeben wurde.
- Datenübertragung verschlüsseln mittels Zwei-Faktoren Sicherheitsmechanismus (z.B. RSA SecurID und mobile VPN Produkt).
- Installation von Virenschutz auf mobilen Geräten sowie die Durchführung regelmässiger Updates. Wichtig hierbei ist, dass solche Software aktiv das Gerät auf Viren überprüft und nicht erst auf Anfrage.
- Ist das mobile Gerät mit anderen Netzwerken verbunden, so ist eine mobile Firewall wichtig (z.B.: Die für PDAs angebotene Bluefire Mobile Firewall).

- Jedes Unternehmen sollte eine eigene Politik für mobile Geräte besitzen, mit der die Angestellten vertraut sind. Ebenso sollten regelmässig Work Shops durchgeführt werden, die die Angestellten auf den "korrekten Umgang" sowie etwagige Sicherheitslücken sensibilisieren.
- Generiert ein Unternehmen Wettbewerbsvorteile durch Know-How, Goodwill oder spezielle Produktionsvorgänge, so können unter Umständen auch Kamerahandys oder USB Sticks verboten werden, da durch diese ebenfalls geheime Daten nach aussen gelangen können.

8.5 Gefahren durch Spam

8.5.1 Definitionen

Der Begriff Spam ist dem Markennamen für ein Dosenfleisch der amerikanischen Firma Hormel Foods (spiced ham) entliehen und entstand bereits im Jahre 1936. Die momentane Bedeutung des Begriffs als Bezeichnung für Massenmailings erhielt er aber erst durch einen Sketch der englischen Comedyserie Monty Python's Flying Circus, in dem der Begriff mehrere Male erwähnt wird. [16]

Schwartz und Garfinkel [17] unterscheidet vier Sorten von *Email-Spam*:

Unaufgeforderte Werbe-Email (*unsolicited commercial email, UCE*) Dies ist die Bezeichnung für eine Email, "die man nicht angefordert hat und die ein bestimmtes Produkt oder eine Dienstleistung bewirbt." Der Begriff ist auch unter dem Namen *Junk-Mail* bekannt.

Unaufgeforderte Massen-Email (*unsolicited bulk email, UBE*) Dies ist die Bezeichnung für eine Email, "die als Seriennachricht an Tausende (oder Millionen) von Benutzer geschickt" wurde. Der Inhalt entscheidet dabei, ob die Nachricht auch eine Werbe-Email ist.

Kettenbriefe Kettenbriefe versprechen meistens einen Weg zum schnellen Geld und sind häufig mit der Anwerbung weiterer Personen - unter anderem auch per Email - verbunden. Kettenbriefe sind häufig betrügerischer Art und sollten immer mit Vorsicht gelesen werden.

Rufschädigungen (*reputation attacks*) Dies ist die Bezeichnung für eine Email, in der ein Absender (eine Person oder eine Organisation) vorgetäuscht wird. Der Inhalt hat meistens zum Ziel, den Absender in ein schlechtes Licht zu rücken.

Mit dem vermehrten Auftreten mobiler Endgeräte entstand im Laufe der Zeit eine neue Form der Massenwerbung, die mit dem Begriff *mobile spam* bezeichnet wurde.

Dabei kann *mobile spam* definiert werden als Nachrichten, die einer unerwünschten Quelle entstammen und auf mobile Engeräte übertragen werden. Die Nachrichten können dabei unterschiedliche Ziele verfolgen: [18]

- Dem Benutzer etwas zu verkaufen
- Den Benutzer verleiten, eine teure Telefonnummer zu wählen
- Die Einstellungen des Mobilgerätes zu verändern oder gar zu löschen
- Einfache Nachrichten kommerzieller Natur, die in die Privatsphäre des Benutzers eindringen und/oder schädlichen Code beinhalten

Im Sinne dieses Begriffes wird in dieser Arbeit im Speziellen auf die Verbreitung von Spam durch Kurznachrichten in Form von SMS (Short Message Service) auf Mobiltelefone fokussiert. *Mobile Spam* kann aber auch in Form von MMS (Multimedia Messaging Service) oder über Instant Messaging auftreten. [19]

8.5.2 Einleitung

Spam nahm seinen Anfang bereits im Jahre 1975, als individuelle Personen massenweise *Junk-Emails* erhielten. [19] Das Problem des Spams häufte und verbreitete sich im Laufe der Zeit aufgrund einer immer enger vernetzten Gesellschaft. In der heutigen Zeit umfasst der Begriff eine Vielzahl möglicher Arten, u.a. den Begriff des *Email-Spam* und des *Mobile-Spam*, wie bereits im vorhergehenden Abschnitt erwähnt.

Spam in Form von Emails hat die Benutzer während vieler Jahre verärgert. Die Durchsetzung dieser Form ist vor allem den Eigenschaften von Emails zuzuschreiben. Diese sind einfach, schnell, zeitunabhängig, direkt und beinahe kostenlos für den Absender (Spammer), was ihm eine kostengünstige Möglichkeit der massenhaften Direktwerbung ermöglicht. Dabei erzielt der Spammer bereits Profite, wenn bloss schon ein geringer Prozentsatz der Empfänger auf dessen Angebote oder Dienstleistungen eingeht. [20] Solange es also derartige Benutzer geben wird, dürften wir auch noch mit Spam konfrontiert bleiben.

Weil das Internet in den letzten Jahren aber auch Zeichen der Sättigung mit *Email-Spam* zeigten, waren die Vertreiber von Massenwerbung auf neue Wege angewiesen. [21] Da sich die Verbreitung von Mobiltelefonen rasant entwickelt und diese Geräte innert kürzester Zeit immer fortschrittlicher, ausgeklügelter und komplexer werden, dabei stets die Möglichkeit besteht, sich über eine drahtlose Verbindung ins Internet einzuwählen, bot sich den Spammern geradezu dieses neue Medium an. Ausserdem tragen die meisten Personen ihr Mobiltelefon laufend mit sich und kommunizieren fortwährend über SMS, was Massenwerbung noch direkter an den Benutzer bringt. [22]

Mobile-Spam wurde in erster Linie vor allem auf dem asiatischen Kontinent zu einem relevanten Problem, allen voran in Korea, wo das Volumen des *Mobile-Spams* schon Ende des Jahres 2003 das Volumen an *Email-Spam* überstieg (siehe Abbildung 8.4). Aber auch in Japan waren unerwünschte Nachrichten auf mobile Geräte um diese Zeit ein grosses Problem, als 90% aller Spam-Nachrichten auf Mobiltelefonen landeten, während gerade einmal 10% in klassischer Form auf den PC vertrieben wurden. [18] Weil die Kosten für den Versand eines SMS viel höher sind als diejenigen für den Versand eines Emails, wurde *Mobile-Spam* in den westlichen Ländern zunächst nur als geringes Problem erachtet,

obwohl hier das Senden von SMS vor allem bei jungen Leuten schon gang und gäbe war. [23] Inzwischen hat sich Mobile-Spam aber auch in Europa in einem rasanten Tempo entwickelt, wie eine kürzlich veröffentlichte Studie der Universität St. Gallen ergeben hat. [35]

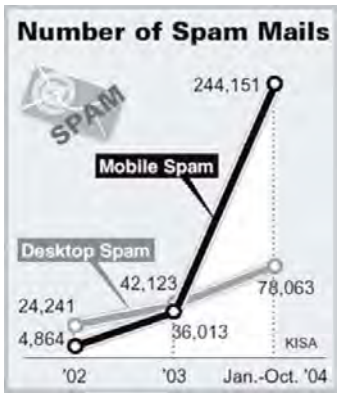


Abbildung 8.4: Anzahl an Spam-Mails in Korea - Quelle: [KISA]

Der Erfolg dieser Werbemethode liegt einerseits darin, dass mit einem relativ geringen finanziellen Aufwand (eine SMS in Korea kostet in etwa 100 KRW, was zum gegenwärtigen Zeitpunkt etwa 0.09 CHF entspricht) eine schier unendliche Zahl von Empfängern in kürzester Zeit erreicht werden kann und andererseits in den zahlreichen Möglichkeiten, wie eine Nachricht am Mobiltelefon ansprechend angezeigt werden kann, aber auch in der Tatsache, dass SMS fast immer gelesen werden müssen. In der Presse wird die SMS-Werbung deshalb vereinzelt auch als "einzig echte zeitaktuelle Kommunikationsmassnahme" bezeichnet. [24]

8.5.3 Verteilkanäle

Spam-Nachrichten, welche beim Benutzer auf einem mobilen Gerät landen, bewirken ein grösseres Ärgernis als Spam-Nachrichten die per Email den PC erreichen. Dies deshalb, weil man enger mit dem mobilen Gerät (z.B. Mobiltelefon) verbunden ist und dies als eine grössere Eindringung in die Privatsphäre empfindet. [25]

Mobile-Spam kann auf unterschiedliche Arten auf ein Mobiltelefon gelangen. Dabei ist wichtig zu unterscheiden, ob die Werbung über eine drahtlose Verbindung direkt aufs Mobiltelefon oder über einen Umweg darauf stösst. Ein Umweg wäre bspw. dann gegeben, wenn Emails zunächst auf einen Heim- oder Arbeitsaccount umgeleitet werden und von dort vom Benutzer abgeholt (pull-Methode) oder ans mobile Gerät weitergeleitet (push-Methode) werden. Wenn Benutzer dabei über einen Umweg auf Spam stossen, sind sie weniger überrascht als bei direkter Art. [25]

In Anlehnung an einen Bericht der "Canadian Wireless Telecommunications Association" [25] seien hier 3 mögliche Formen der drahtlosen Nachrichtenübertragung genannt:

Drahtlose Email-Spam Einige Mobiltelefone können wie bereits erwähnt über einen integrierten Email-Client oder aber über das Wirelss Application Protol (WAP) auf

Emails zugreifen. Entweder wird dabei auf eine vom Mobiltelefon gegebene Emailadresse oder auf eine Emailadresse über einen Heim- oder Arbeitsaccount zugegriffen. Die Menge der Spam-Nachrichten richtet sich dann nach den Filtermethoden des ISP oder des Geschäftes, welche diese Accounts anbieten. Ein Spammer wird eher selten über einen drahtlosen Emailaccount Spam versenden, da es sich aufgrund der relativ hohen Verbindungskosten nicht lohnen wird.

SMS-to-SMS Spam Auch wenn ökonomisch nicht sinnvoll, ist es dennoch möglich, *SMS-Spam* direkt von einem Mobiltelefon aus zu versenden. Ein Mobilfunkanbieter verrechnet üblicherweise einen Tarif für ausgehende Nachrichten. Diese Art des *Spammings* wird weiterhin eingeschränkt, insofern als dass die Mobilfunkanbieter den Zugriff auf ihr Netzwerk kontrollieren können.

Email-to-SMS Spam Hier werden typischerweise Internet-Nachrichten über einen Email-zu-SMS Gateway in Nachrichten für ein Mobiltelefon konvertiert. Die Emailadresse besteht dabei aus zwei Teilen, der Telefonnummer des Mobilteilnehmers und der Internet-Domain (z.B. 0761234567@mail.wirelesscarrier.com) Dieser Weg funktioniert meistens nur in genannter Richtung.

Das bereits erwähnte *Bluejacking* kann im Zusammenhang mit *Mobile-Spam* wie folgt verwendet werden: Wenn ein Mobiltelefon über eine Bluetooth-Schnittstelle verfügt und diese aktiviert ist, kann ein Spammer dem Benutzer eine Visitenkarte zusenden, nur dass in dieser nicht der Name steht, sondern eine beliebige andere Nachricht, so dass also auch Werbung möglich wäre. Voraussetzung dazu ist allerdings, dass beide die Bluetooth-Schnittstelle aktiviert haben und sich in der Nähe befinden, was diese Spam-Methode nicht sonderlich effektiv macht.

Neben den obengenannten Möglichkeiten verfügen einige Mobiltelefone des weiteren über Instant-Messaging Services und Multimedia-Nachrichten Services (MMS), über welche ebenfalls Spam-Nachrichten auf das Mobiltelefon gelangen können. Der Hauptfokus der Spammer richtet sich auf den Email-zu-SMS Weg.

8.5.4 Varianten betrügerischer Mobile-Spams

Mobile-Spam kann zu unterschiedlichen Zwecken benutzt werden. Deren Ziele bzw. Risiken seien hier in Anlehnung an einen Bericht der “Nexus Telecom AG” [26] und der “CiscoSystems” [27] kurz erläutert:

Spamming Beim *Spamming* hat ein Content-Provider einen regulären Servicevertrag mit einem Mobilfunkanbieter. Aus Sicht eines Endbenutzer erscheint eine einzige Spam-SMS unter Umständen als ungewollt und ärgerlich, häufig werden diese aber in grösseren Mengen gleich an mehrere Teilnehmer verschickt. Das Risiko für den Mobilfunkanbieter besteht darin, dass ihn möglicherweise Klagen eines Kunden oder eines Roaming-Partners wegen Weitergabe von Spam erreicht.

Flooding Von *Flooding* ist die Rede, wenn ein Content-Provider mit einem SMS-Center in einem fremden Netzwerk verbunden ist und die Gefahr besteht, dass sich aufgrund massenweiser SMS das Netzwerk überlädt oder im Extremfall gar abstürzt. Dabei spielt es keine Rolle, ob die SMS gültig sind oder nicht. Im Unterschied zu *Spamming* steht also mehr die Gefahr einer Überladung des Netzwerks als die Inhaltsvermittlung der SMS im Vordergrund.

Faking/Spoofing Beim *Faking/Spoofing* simuliert eine Anwendung das Verhalten eines normalen SMS-Centers. Die Quelladresse der SMS gibt dann vor, aus einem anderen Netzwerk zu stammen. Die Absenderadresse einer SMS kann hier also frei gewählt werden. Beim *Spoofing* steht speziell das Abfangen von Informationen des Opfers im Vordergrund.

8.5.5 Auswirkungen von Spam

Obwohl die Auswirkungen von Spam auf mobile Geräte (insbesondere auf Mobiltelefone) im ersten Augenblick ähnlich erscheinen wie die Auswirkungen auf übliche Desktop-Computer, gibt es einige wichtige Unterschiede die an dieser Stelle kurz erläutert werden sollen.

Mobile-Spam ist ein wesentlich ernsteres Problem als bspw. *Email-Spam*, dies deshalb, weil Mobiltelefone als persönlich enger verbunden angesehen werden, schliesslich tragen die meisten Menschen ihr Gerät häufig mit sich herum. Während sich die Lästigkeit beim *Email-Spam* nämlich während dem Durchlesen ergibt, dringen SMS beispielsweise durch ihre zwangsweise Ankunft plötzlich, auf einen Schlag in die Privatsphäre des Benutzers ein, was als unangenehmer empfunden wird. Ein weiterer Unterschied ergibt sich durch die Tatsache, dass Menschen unter Umständen zwar mehrere Emailaccounts, hingegen nur 1 Mobiltelefon besitzen. [28]

SMS-Spam unterscheidet sich aber auch bezüglich bestimmter Eigenschaften von *Email-Spam*, bspw. ist der Absender einer SMS unter Umständen erst ersichtlich, wenn die SMS geöffnet und gelesen wird. Ausserdem können - je nach Einstellung - nervenaufreibende Signaltöne bei mehreren Werbe-SMS eine zusätzliche Belästigung bedeuten. Ein weiterer wichtiger Aspekt ist schliesslich die Speicherkapazität vieler Mobiltelefone bezüglich SMS, so dass bei Überschreiten einer bestimmten, oft kleinen Grenze die Gefahr besteht, dass der Eingang überläuft und keine weiteren Nachrichten mehr empfangen werden können, wesentlich grösser als bei E-Mails. Im Vergleich mit einer E-Mail kann eine SMS also eine weitaus belästigerende Wirkung zeigen. [24]

Einen letzten Punkt könnte man schliesslich noch bezüglich den Kosten hinzufügen. Bei mobilen Geräten kann das Abrufen von *Email-Spam* eine reine Geldverschwendung bedeuten, da sich die Zugriffsgeschwindigkeiten ins Internet typischerweise in einem Bereich von 9600 bit/s bis 19200 bit/s bewegen. Ausserdem belaufen sich in Europa die Kosten für eine Nachricht bis auf 15 Cents, während dessen in den USA eine minim geringere Gebühr verrechnet wird. Der Abruf von *Email-Spam* verursacht durch die zu zahlenden Gebühren also wesentlich höhere Kosten als bei einem Desktop-Computer. [29]

Bei den nun folgend aufgeführten Punkten wird einfachheitshalber angenommen, dass sich die Auswirkungen von *Mobile-Spam* und *Email-Spam* in unterschiedlicher Stärke und Grösse bewegen, vom Prinzip der Problemstellung allerdings gleich sind. Viele hier erwähnte Probleme lassen sich nämlich gut auf *Mobile-Spam* überführen.

Auswirkungen auf Netzwerkbetreiber

Internetressourcen: Im Jahre 2001 wurden über 140 Milliarden Spam Nachrichten versendet, wobei sich Prognosen für das Jahr 2008 auf über 600 Milliarden Nachrichten belaufen. Dies Spam-Massen haben einen wesentlichen Einfluss auf die Bandbreite und Speicherplätze des Internets. [30]

Auswirkungen auf Unternehmen

Hardwareinvestitionen: Für die Weiterleitung und Speicherung der zusätzlichen Massenmails für die Benutzer enstehen bei Unternehmen mit vielen Mailempfängern und ISP's Leistungsengpässe, so dass durch die erhöhte Hardwarebelastung zusätzliche Speicherkapazitäten und Rechnerleistung erforderlich werden. [30]

Personalkosten: Die zusätzlich angeschaffte, leistungsfähigere Hardware macht mehr administratives und technisches Personal notwendig. Ohne personelle Verstärkung wird es für die Administratoren schwieriger, die Angestellten zur Einsparung von Arbeitszeit vor Spam zu schützen. [30]

Opportunitätskosten: Während dem oft stundenlangen Sichten und Aussortieren von Emails, fehlt den Mitarbeitern die Arbeitszeit für ihre eigentlichen Aufgaben. Dabei entgeht dem Unternehmen nicht nur Geld sondern es verringert auch noch seine Produktivität und damit seine Wettbewerbsfähigkeit. [31]

Kommunikationsfehler: Es kann sein, dass wichtige Emails von Kunden oder Geschäftspartnern leicht in Fluten von Spam untergehen, sei es dass sie einfach übersehen oder versehentlich gelöscht werden. Dieser Umstand kann zu unbemerkten Fehlern in der Kommunikation führen und im Extremfall in einem Verlust des Auftrages oder gar eines Kunden enden. [31]

Nichterreichbarkeit: Wenn der Email-Server eines Unternehmens infolge einer in kurzer Zeit eintreffenden grossen Menge von Spam einen Totalausfall erleidet und der Server jegliche weitere Kommunikation verweigert (*Denial of Service*), kommt der Informationstausch nach aussen und innen zum Erliegen, was in manchen Bereichen einen Produktionsstillstand auslösen könnte. [31]

Marketing-Probleme: Eine Behinderung der Marketing-Aktivitäten könnte sich bspw. dadurch ergeben, dass Newsletter-Emails oder Pressemitteilungen, die eigentlich zur Kundenbindung oder -neugewinnung gedacht sind, in einer Vielzahl unerwünschter Werbebotschaften untergehen. Im Extremfall wird das Unternehmen sogar selbst des Spams beschuldigt, was zu einem Imageverlust oder ungerechtfertigten Anschuldigen führen könnte. [31]

Auswirkungen auf Endbenutzer

Übertragungskosten: Das Herunterladen von Spam verursacht bei Verbindungen ohne Flatrate zusätzliche Onlinekosten [30]

Softwarekosten: Für die Einrichtung von Spam-Abwehrsystemen können dem Benutzer erhebliche Kosten entstehen. [31]

Psychische Belastung: Je höher die Anzahl der ungewollten Nachrichten über der Anzahl der gewollten Nachrichten liegt, umso mehr fällt frustrierende Such- und Sortierarbeit für den Benutzer an. Im Falle eines Überlaufs werden neue Nachrichten, gewollt oder nicht gewollt, zurückgewiesen und später zugestellt. Auch inhaltlich kann eine Email besonders für Kinder belästigend wenn nicht sogar gefährdend auswirken. Ein Grossteil der *Email-Spam* umwirbt obszöne Dienstleistungen und Angebote. [30]

Physical damage und Spionage: Spam Nachrichten können auch angehängte Virusdateien, Spionageprogramme und Dialer (Programme, die sich unbemerkt auf teure Telefonnummern einwählen) beinhalten, wodurch dem Benutzer umfangreiche, wirtschaftliche Schäden entstehen können. Eine Einschränkung der persönlichen Freiheit kann neben Schäden -wenn auch nicht direkt physisch- und Datenverlust die Folge sein. Dialer verursachen zudem finanzielle Schäden durch erhöhte Einwahlgebühren. [30]

Einige wenige Probleme, die sich speziell auf Mobiltelefone bzw. SMS beziehen, seien hier noch erwähnt:

Anbieterwechsel bei den Kunden Mobilfunkanbieter sollten ein grosses Interesse daran haben, *Mobile-Spam* weitestgehend einzudämmen. SMS-Nachrichten machen nämlich einen Anteil von ungefähr 10% am Umsatz von Mobilfunkanbieter aus, gemäss einer Studie der Firma IDC. [27] Das zunehmende Volumen an Spam kann genau diese Einnahmequellen gefährden, wenn sich Kunden nämlich zu einem Anbieterwechsel entscheiden. Eine Studie aus Singapore [32] hat nämlich erwiesen, dass ein Grossteil der Kunden den Mobilfunkanbieter für Spam verantwortlich machen.

Betrügerische SMS Einzelne SMS kommen mit einer Aufforderung, eine angegebene Nummer zurückzurufen. Ohne dass es auf den ersten Blick ersichtlich ist, wählt man dann eine teure Mehrwertdienstnummer. Die Telefonnummern werden dabei geschickt verschleiert. Ähnliches kann passieren, wenn man eine SMS beantworten soll, bspw. auf Einladungen von Frauen oder auf die Frage “Wie fandest du das Hockey-Spiel gestern?” Wird eine solche SMS nämlich beantwortet, wissen die Versender, dass die Telefonnummer tatsächlich existiert, und man selbst bezahlt im weiteren Verlauf auch noch jedes empfangene SMS. Dasselbe kann auch bei der Registrierung der Handy-Nummer in einem SMS-Chat vorkommen. Man erhält dann einerseits kostenpflichtige SMS mit diversen Angeboten und zum Anderen kann es sein, dass die Handy-Nummer gar weiterverkauft wird. Selbst mit Antworten wie “STOPP SMS” kann es unter Umständen sein, dass man weiterhin kostenpflichtige Nachrichten erhält. [32]

8.5.6 Gegenmassnahmen

Hinsichtlich der Erkennung von Spam spielt ein wesentlicher Unterschied bezüglich der Eigenschaften zwischen *SMS-Spam* und *Email-Spam* eine grosse Rolle. *Email-Spam* kann dabei über verschiedene Methoden hauptsächlich über die gebrauchten Schlüsselwörter und deren Struktur identifiziert werden. Einige dieser Methoden heissen bspw. "Bayesian approach", "Memory based approach", "Markov Chain approach" oder "Support Vector Machine". Gemeinsam ist all diesen Methoden, dass sie sich auf die in Spam enthaltenen Strukturen von Schlüsselwörtern beziehen. In SMS-Nachrichten wird nun allerdings die verfügbare Information bezüglich Schlüsselwörtern und Korrelationen zwischen Sequenzen von Schlüsselwörtern durch die kleine Grösse der Kurznachrichten beschränkt. Dies führt natürlich zu einer starken Einschränkung der Anwendbarkeit obengenannter Methoden auf das Problem des *SMS-Spams*. [28]

Die relativ kleine Grösse von Kurznachrichten birgt aber auch Vorteile, so können SMS zur Umgehung der Erkennung durch Filter nur begrenzt verändert werden, ohne dass sich dabei ihre Bedeutung verändert. Des weiteren sind SMS keine Echtzeitanwendungen, was einem ein Zeitfenster zur Erkennung von Spams eröffnet. Ein dritter Vorteil besteht darin, dass das SMS-Center in der SMS Netzwerkarchitektur einen zentralen Punkt zur Sammlung von Spam-Nachrichten einnimmt, was ein *Spamming* von unterschiedlichen Quellen aus sinnlos macht. [28]

Zur Identifizierung einer Kurznachricht als Spam reichen die Informationen hinsichtlich des Textes und der Schlüsselwörter nicht aus. Bereits ein 15-Zeichen langes SMS könnte einen Benutzer auf eine teure Mehrwertdienstnummer führen. Daher ist es von Vorteil, wenn Filteralgorithmen für *SMS-Spam* anstatt sich auf Texte und Schlüsselwörter zu beziehen, die Kurznachrichten auf Ähnlichkeiten mit im Netzwerk umlaufenden Spam-Nachrichten vergleicht. [28]

Das Problem des *SMS-Spam* kann gleich wie beim *Email-Spam* über präventive, technische oder gesetzliche Massnahmen verkleinert werden.

Präventive Massnahmen

Präventive Massnahmen bieten Mechanismen, die jedem Benutzer zur Verfügung stehen und die jeder anwenden kann. In Anlehnung an den Bericht "Security Info" [32] seien hier einige davon erwähnt:

- Handynummern sollten nur Personen bekannt gegeben werden, die man kennt und von denen man auch angerufen werden möchte.
- Teilnahmen an SMS-Chats sollten aus Gründen, die bereits weiter oben erwähnt wurden, vermieden werden.
- Auf SMS sollte nicht voreilig geantwortet werden. Es empfiehlt sich zunächst zu prüfen, ob der Absender bekannt ist.

- Bei Wettbewerben sollte vermieden werden, die Handynummer einzugeben. Fallen nämlich persönliche Daten in eine bestimmte Werbezielgruppe, könnten die Daten an SMS-Spammer weiterverkauft werden.
- Es sollten keine Klingeltöne von dubiosen Webseiten heruntergeladen werden, oder jene, die via SMS bestellt werden können.
- Auf *SMS-Spam* sollte nicht geantwortet werden, bspw. um seine Meinung zu äussern.
- Die AGBs sollten bei Dienstleistungen aufmerksam durchgelesen werden. Steht dort, dass der Anbieter die Telefonnummer auch für andere Marketingzwecke verwendet, ist davon abzuraten.

Technische Massnahmen

Im Gegensatz zur Email können Empfänger von *SMS-Spam* keine Gegenmassnahmen auf Benutzerseite durchführen gegen eine zunehmende Anzahl an ungewollten Kurznachrichten. Deshalb obliegt es dem Mobilfunkanbieter, den Benutzern zu helfen, ungewollte SMS zu blockieren. [26]

Technische Massnahmen finden sich bei den Mobilfunkanbietern aber noch relativ selten, die Anbieter verdienen natürlich auch mit an jeder SMS. Es gibt aber dennoch Anbieter, die sich der Sache in ernster Weise annehmen und etwas dagegen tun, z.B. Sunrise, wo *SMS-Spam* gemeldet werden kann. [23]

Eine Vielfalt an technischen Massnahmen gegen Email-Spam existiert bereits. Die meisten davon können effektiv auf das Problem des *SMS-Spam* überführt werden. In Anlehnung an einen Bericht über die Filterung von *SMS-Spam* [23] seien hier einige dieser Methoden kurz erkläutert:

Schwarze und weisse Listen Sender, welche in der schwarzen Liste enthalten sind, gelten als Spammer und deren SMS werden deshalb geblockt. Sender, welche in der weissen Liste enthalten sind, gelten als vertrauenswürdig und deren SMS werden deshalb zugestellt.

Porto Für den Versand einer Nachricht wird ein Porto verlangt. Das Porto kann nach ökonomischen Aspekten (in Cent pro Nachricht), aufgrund einer Berechnung (Kostenfunktion) oder basiert auf einem Turing-Test (computergestützte Unterscheidung von Spam und Non-Spam) berechnet werden.

Adressen-Management Hierbei wird eine Menge an temporären Adressen automatisch generiert. Sobald Spam an eine Adresse gelangt, wird sie gelöscht.

Kollaboratives Filtern Wenn ein Benutzer eine Nachricht als Spam markiert, so gilt dies auch für andere ähnliche Benutzer. Ausserdem kann eine Nachricht an mehrere Empfänger vom Dienstanbieter automatisch als Spam markiert werden.

Digitale Signaturen Nachrichten ohne digitale Signatur gelten als Spam. Signaturen werden von einem vertrauenswürdigen Sender oder Dienstanbieter vergeben.

Inhaltsbezogene Filter Ist die meistverwendete Methode. Jede Nachricht wird auf spezifische Merkmale von Spam und auf Schlüsselwörter hin überprüft. Die bekannteste Methode dazu ist der Bayesian-Filter.

Gesetzliche Massnahmen

- Im Gegensatz zur EU fehlen in der Schweiz ausdrückliche gesetzliche Regelungen für das *Spamming*. Mit der laufenden Revision des Fernmeldegesetzes soll diesem Umstand aber entgegengetreten werden. [33]
- Die USA vertritt als weltgrösster Exporteur eine opt-out-Lösung. Der Empfänger erhält hier beim Empfang von Werbung die Möglichkeit, sich aus der Verteilerliste des Anbieters entfernen zu lassen, wenn dieser keine Werbung mehr wünscht. Das gängige Gesetz dazu heisst "CAN SPAM Act". [17]
- Die EU hingegen vertritt als grösster Block eine opt-in-Lösung. Hier stimmt der Empfänger durch einmaliges Eintragen in eine Abonnentenliste dem Empfang zu. Hier sind zahlreiche gesetzliche Regelungen gängig. [17]
- In Japan wurden zur Regelung des Spam zwei Gesetze erlassen, einerseits das "The Law on Regulation of Transmission of Specified Electronic Mail (Anti-spam Law)" und andererseits das "Specified Commercial Transactions Law". Diese getätigten Massnahmen führten zu einem markanten Rückgang der von einem Mobilgerät gesendeten SMS ab Mitte des Jahres 2003 bis Ende 2004. [34]

Rechtliche Schritte versprechen allerdings wenig Erfolg. Zum einen sind Gerichtsverfahren häufig sehr aufwendig, finanziell risikobehaftet und in ihrem Ergebnis nicht immer vorhersehbar. Zum anderen können allfällige Urteile im Ausland kaum durchgesetzt werden. Und da der Spammer in den meisten Fällen seinen Sitz im Ausland hat, nützen rechtliche Schritte in der Schweiz wenig. Ausserdem lassen sich die Absender häufig gar nicht eindeutig identifizieren. [33]

8.6 Schlussfolgerung

Betrachtet man einzelne Verfahren zur Spambekämpfung für sich, bieten keine Massnahmen einen 100%igen Schutz. Jedes Verfahren bietet für sich betrachtet spezifische Vor- und Nachteile und bekämpft Spam nur unter bestimmten Aspekten. Um einer Verbreitung von Spam entgegenzuhalten, sollte jedoch an möglichst vielen Punkten angesetzt werden können. Einige solcher Punkte sind z.B. die Vermeidung von Adresssammlungen und die Erschwerung des Versands von Spam. Ziel ist es, dies durch eine sinnvolle Kombination der Verfahren sowohl auf Benutzer- wie auch auf Anbieterseite zu gewährleisten. Ergänzend dazu sollten auch umfangreiche gesetzliche Massnahmen ergriffen werden. [30]

Literaturverzeichnis

[1] Lawton G.: Virus Wars: Fewer Attacks, New Threats, Computer, Dezember 2002.

[2] Leitch I.: Computer viruses: a problem of management, ENGINEERING MANAGEMENT JOURNAL, Februar 1994.

[3] Weaver N., Paxson V., Staniford S., Cunningham R.: A Taxonomy of Computer Worms, ACM, Oktober 2003.

[4] Eustice K., Popek Dr. G., Kleinrock Dr. L., Ramakrishna V., Marksturm S., Reiher Dr. P.: Securing Nomands: The Case for Quarantine, Examination, and Decontamination, ACM, 2004.

[5] Hong J.: Minimizing Security Risks in Ubicomp Systems, Computer, Dezember 2005.

[6] Dagon D. Martin T., Starner T.: Mobile Phones as Computing Devices: The Viruses are Coming!, IEEE CS and IEE ComSoc, Oktober-Dezember 2004.

[7] Leavitt N.: Mobile Phones: The Next Frontier for Hackers?, Computer, April 2005.

[8] Computer Economics: 2005 Malware Report: Executive Summary, `http://www.computereconomics.com/article.cfm?id=1090`, Januar 2006.

[9] Halpert B.: Mobile Device Security, ACM, 2004.

[10] Leyden J.: My car has avirus, The register, `http://www.theregister.co.uk/2005/02/09/ibm_security_report/`, Februar 2006.

[11] Housley R., Arbaugh W.: Security problems in 802.11-basesd networks, Communications of the ACM, Mai 2003.

[12] Su J., Chan K., Miklas A., Po K., Akhavan A., Saroiu S.,de Lara E., Goel A.: A Preliminary Investigation of Worm Infection in a Bluetooth Environment, ACM, November 2006.

[13] Tecchannel: Bluetooth-Sicherheit: Binnen Sekunden geknackt, tecchannel, `http://www.tecchannel.de/news/themen/sicherheit/454210/index.html`, Januar 2006.

[14] Viruslist.com: Alles über Internet-Sicherheit, `http://www.viruslist.com/de/viruses`, 2004.

[15] Bundesamt für Sicherheit in der Informationstechnik: Mobile Endgeräte und mobile Applikationen: Sicherheitsgefährdung und Schutzmassnahmen, Bonn, September 2006.

[16] Chastonay M.: Spam: grosses Übel im Internet, WB EXTRA 21.06, Juni 2006.

[17] Schmidt D.: Spam in neuen Informations- und Kommunikationstechnologien, Justus-Liebig-Universität Giessen, http://geb.uni-giessen.de/geb/volltexte/2004/1667/, Februar 2004.

[18] Srivastava L.: spam to your mobile phone, Michigan State University, www.itu.int/osg/spu/presentations/2006/srivastava_mobilespam_2006.pdf, März 2006.

[19] Hameed S.: Mobile Spam in Singapore, Singapore Internet Research Centre, www.ntu.edu.sg/sci/sirc/download/mobile\%20spam\%20report-11\%20may\%202005.pdf, Mai 2005.

[20] Wey C.: Das Spam-Problem im Internet und verbraucherpolitische Antworten, Technische Universität Berlin, www.wm.tu-berlin.de/~wey/seminar/arbeiten_ws0506/arbeit_kalantary_spam.pdf, Oktober 2005.

[21] Enck W., Traynor P., McDaniel P., La Porta T.: Exploiting Open Functionality in SMS-Capable Cellular Networks, ACM, November 2005.

[22] Jamaluddin J., Zotou N., Edwards R.: Mobile Phone Vulnerabilities: A New Generation of Malware, IEEE, August 2004.

[23] Hidalgo J., Bringas G., Sanz E., Garcia F.: Content Based SMS Spam Filtering, ACM, Oktober 2006.

[24] Fellner G.: Spam-SMS, http://www.it-law.at/papers/Fellner_Spam_SMS.pdf, September 2003.

[25] CWTA: Overview of Wireless Spam Issues in Canada, Canadian Wireless Telecommunications Association, www.e-com.ic.gc.ca/epic/internet/inecic-ceac.nsf/vwapj/Wireless_Spam.pdf/\$file/Wireless_Spam.pdf, März 2005.

[26] Buehler W.: Blocking of SMS Spam and Fraud, Nexus Telecom AG, www.nexus-ag.com/spam_fraud.0.html, Mai 2004.

[27] Cisco Systems: SMS Spam and Fraud Prevention, http://www.cisco.com/en/US/netsol/ns341/ns396/ns177/ns278/networking_solutions_white_paper0900aecd80250cb6.shtml, November 2006.

[28] Dixit S., Gupta S., Ravishankar C.: Lohit: an online detection and control system for cellular sms spam, http://www.actapress.com/PaperInfo.aspx?PaperID=22246, November 2005.

[29] Dennis S.: Unimobile Solves Problem Of Mobile Device Spam - Company Business and Marketing, www.findarticles.com/p/articles/mi_m0NEW/is_2001_April_12/ai_73115148, April 2001.

[30] Müller T.: Email-Direktmarketing - Spam, Westfälische Wilhelms-Universität Münster, www-wi.uni-muenster.de/pi/lehre/WS0304/Seminar/06_Spam.pdf, Dezember 2003.

[31] Pfänder D.: Die Problematik vom Spam-Mail für kleine und mittelständische Unternehmen, Fachhochschule Ansbach, `www.danielpfänder.de/t3/fileadmin/user_upload/downloads/spam.pdf`, Oktober 2004.

[32] Werz: SMS SPAM, `http://www.securityinfo.ch/sms.html`, 2003.

[33] Bakom: Spam Spamming Spammer, `www.ai.ch/dl.php/de/20050429102648/spam_d.pdf`, März 2005.

[34] Imaizumi Y.: Japan's Measures against Spam, `www.itu.int/osg/spu/ni/multimobile/presentations/ITUimaizumi.pdf`, Juni 2006.

[35] ITU, Universität St. Gallen, Intrado: Insights into Mobile Spam - World's First Collaborative Empirical Study, `www.mobilespam.org`, Februar 2005.

[36] Borisov N., Goldberg I., Wagner D.: Intercepting Mobile Communications: The Insecurity of 802.11, Proceedings of the Seventh Annual International Conferecne on Mobile Computing And Networking, Juli 2001.

[37] `http://www.icq.com`

[38] `http://get.live.com/messenger/overview`

[39] `http://www.limewire.com/english/content/home.shtml`

[40] `http://www.kazaa.com/us/index.htm`

Kapitel 9

Voice and Video Transmission over Wireless Networks

Fabian Hensel, Andres Petralli, Pascal Suter

Multimedia Applikationen gehören mittlerweile zu unserem Alltag. Vorallem die mobilen Geräte wie Mobiltelefone und PDAs werden immer leistungsstärker und ermöglichen zunehmend multimediale Applikationen. Besonders im Bereich der mobilen Kommunikation zeichnen sich viele mögliche Anwendungen ab; insbesondere wird die multimediale Kommunikation immer populärer. Doch anders als bei fest verkabelten Netzwerken kämpfen mobile Netzwerke mit verschiedensten Herausforderungen. Diese Arbeit gibt einen Einblick in die verschiedenen Anforderungen und Probleme der multimedialen Kommunikation in mobilen Netzwerken. Lösungsansätze und Technologien werden auf den folgenden Seiten diskutiert. Zum Einstieg werden die Eigenschaften von mobilen Netzwerken, in Bezug auf die verschiedenen Parameter untersucht. Weiterhin wird gezeigt wie in modernen Netzwerken die Qualität in Bezug auf Sprachübertragungen erfasst wird sowie die Uebertragungen mit Zuhilfenahme von verschiedenen Codecs gewährleistet wird. Der Vollständigkeit halber werden zu den gezeigten Methoden die möglichen Fehlerkorrekturen und Mechanismen zur zuverlässigen Uebertragung dargestellt. Dies beinhaltet unter Anderem die Diskussion verschiedener Dienstgüteklassen (QoS).

Inhaltsverzeichnis

9.1 Eigenschaften von Wireless Netzwerken

In den folgenden Abschnitten werden die verschiedenen, für Voice- und Videoübertragung in paketvermittelnden Netzen wichtigen Faktoren erläutert und beschrieben.

9.1.1 Soft-Realtime

Vorallem im Zusammenhang mit der Telekommunikation kommt man oft an den Begriff „Realtime“. Doch was genau ist eine realtime Anwendung?

Realtime Systeme sind Systeme, die mit Deadlines arbeiten. Wird eine Nachricht inerhalb der Deadline abegliefert hat sie einen gewissen Wert, kommt sie jedoch nach der Deadline an, ist sie entweder viel weniger wert oder überhaupt nichts mehr. Bei Anwendungsfällen in denen eine verspätete Nachricht überhaupt keinen Wert mehr hat, spricht man von hard-realtime Anwendungen. Wenn die Nachricht ihren Wert nach der Deadline rasch verliert, aber noch nicht sofort wertlos ist, spricht man von soft-realtime Anwendungen. Hard-realtime Anwendungen sind dort anzutreffen, wo das Nichteintreffen einer Nachricht oder das Nichteintreten eines Zustands bis zur Deadline zur Katastrophe führt. Bei soft-realtime Anwendungen führt es lediglich zu Umständlichkeiten oder zur Behinderung, nicht aber zum totalen Kollaps des Systems. Die Abbildung 9.1 zeigt diese beiden Arten von Realtime grafisch auf.

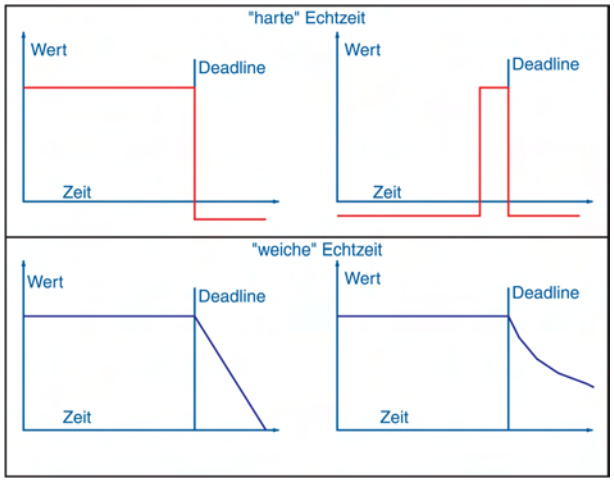


Abbildung 9.1: Bei Hard-Realtime ist nach dem Verstreichen der Deadline die Nachricht nichts mehr wert, während bei Soft-Realtime der Wert lediglich stark abnimmt [12].

Bei hard-realtime Applikationen wir mit Serviceverträgen gearbeitet, die die Einhaltung der Deadline garantieren. Dazu sind entsprechend spezielle dedizierte Systeme und Vorgänge nötig. Es darf nur mit dem Worst-Case gerechnet werden und dementsprechend überdimensioniert müssen auch die Systeme gestaltet werden. Dies gilt nicht nur für die Hardware, sondern auch für die Software.

Bei soft-realtime Systemen hingegen wird mit dem best-effort Ansatz gearbeitet. Es werden soweit möglich Vorkehrungen getroffen um die Wahrscheinlichkeit, dass eine Deadline

verpasst wird zu minimieren. Es wird mit Durchschnittswerten statt mit Extremwerten gerechnet und es werden statistische Werte angegeben anstatt Garantien abzuschliessen. Das Ziel ist es, möglichst viele Deadlines einhalten zu können und nicht keine einzige zu verpassen.

Audiovisuelle Übertragungen sind in der Regel typische soft-realtime Applikationen. Es nervt den Benutzer zwar, wenn seine Telefonverbindung abbricht, doch wenn dies nicht mehr als beispielsweise dreimal jährlich passiert, kann er das akzeptieren.

9.1.2 Latency

Hat ein Netzwerk eine hohe Latency bedeutet das, dass ein Paket, welches von einem Absender zu einem Empfänger über das Netzwerk gesendet wird, relativ lange im Netzwerk unterwegs ist, bis es an seinem Ziel ankommt. Somit entsteht eine Verzögerung, die vorallem dann bemerkbar wird, wenn es sich um bi-direktionale Kommunikation handelt. Der Absender muss dann nämlich nebst der Reaktionszeit des Empfängers auch noch die doppelte Dauer der Latency abwarten, bis er die Antwort erhält.

Im Zusammenhang mit Voice und Videoübertragung spricht man häufig von der einweg end-zu-end Latency. Dies bedeutet beispielsweise bei der Voice übertragung „vom Mund des Sprechers zum Ohr des Zuhörers“. Das heisst also, dass die end-zu-end Latency auch die Zeit beinhaltet, die für das Digitalisieren und das Komprimieren des Tonfragmentes benötigt wird, sowie mögliche Buffer und aufwändige Korrekturmethoden. Dies kann oft zu kontraversen Situationen führen, da zur möglichst problemfreien Übertragung eine möglichst geringe Datenmenge und somit eine möglichst hohe Kompression von Vorteil wäre, jedoch stärkere Kompressionen in der Regel mehr Zeit zum komprimieren brauchen.

Um die entstehende Verzögerung in der Kommunikation etwas zu vertuschen werden gute echocancellation Methoden benötigt, denn wenn die beiden Parteien das Echo ihrer selbst hören, ist die Verzögerung viel stärker wahrnehmbar als sonst.

9.1.3 Jitter

Jitter bezeichnet den Variationsraum der Latency. Das heisst, ein Netzwerk mit stark schwankender Latency hat einen hohen Jitter. Ein häufiger Grund für hohen Jitter sind biespielsweise ständig wechselnde Routen oder eine hohe Last auf dem Netzwerk. Bei der Audio und Videoübertragung hat dies Unterbrüche und Fehler in Ton und Bild zur Folge. Unterbrüche entstehen dann, wenn ein Paket nicht rechtzeitig eintrifft, Fehler können entstehen, wenn ein Paket unterwegs ein anderes überholt, weil es beispielsweise auf der schnelleren Route unterwegs ist.

Um dem entgegen zu wirken kann man einen Jitter-Buffer verwenden. Dies ist ein kleiner Zwischenspeicher, der die ankommenden Pakete sammelt, richtig sortiert und dann in der richtigen Datenrate konstant an die Anwendung abgiebt. Je grösser der Buffer um so höher kann der Jitter sein. Jedoch steigt mit der Buffergrösse wiederum auch die End-zu-End Latency, daher sollte der Jitter-Buffer nur gerade so gross wie nötig sein.

9.1.4 Packet-Loss

Packet-Loss und ganz allgemein Loss bezeichnet den Verlust von Datenpaketen oder Fragmenten davon während der Übertragung über ein Netzwerk. Bei der Audioübertragung hat der Verlust eines Pakets zur Folge, dass ein Teil einer Tonkurve fehlt. Je nach Codec können auch benötigte Parameter zur Wiederherstellung der Tonkurve fehlen. Bis zu einem gewissen Grad sind dadurch entstehende Ungenauigkeiten für das menschliche Gehör nicht hörbar. Ein durchschnittliches Voice-Datenpaket enthält ca. 20ms an Ton. Fehlen einmal 20ms kann beispielsweise das vorherige Paket wiederholt werden und wir werden kaum etwas davon bemerken. Problematisch wird es jedoch bei sehr stark komprimierten codecs, da diese bereits alle für uns nicht hörbaren Daten (und noch etwas mehr) weggelassen haben. Somit ist der nicht wahrnehmbare Audio-Overhead bereits weg und ein Verlust dementsprechend gravierender.

9.1.5 Availability

Availability, also die Verfügarkeit sowie die Reliability, also die Zuverlässigkeit von Voice und Video Verbindungen sind für den Benützer sehr wichtig. Beim Telefon ist man sich als Anwender gewohnt, dass, sobald man den Hörer von der Gabel nimmt, einen Rufton hört. Auch wenn in einem Gewitter die Stromzufuhr zum Gebäude abbricht und alles dunkel wird, das Telefon geht noch! Somit sind wir uns von Telefonen eine extrem hohe Verfügbarkeit gewohnt. Aber auch bei der Zuverlässigkeit sieht es nicht anders aus. Wann wurde das letzte mal eine Telefonverbindung (also eine Festnetz-Verbindung) während eines Gesprächs unterbrochen oder waren Störungen hörbar? Wahrscheinlich ist das schon recht lange her, und dann war es höchst warscheinlich bei einer Ausland-Verbindung, die irgendwo über einen VoIP Gateway geroutet wurde, um Kosten zu sparen. Herkömmliche Telefonnetze bieten eine weitaus höhere Verfügbarkeit und Zuverlässigkeit als moderne Computernetzwerke es bieten können. Dies hängt vorallem damit zusammen, dass bei IP-basierten Netzwerken die ursprünglichen Services alle auf best-effort Leistungen basierten. Wenn zu Stosszeiten jeder seine eMails versenden möchte, bekommt halt jeder etwas weniger Bandbreite. Die Nachricht erreicht deswegen genauso ihren Empfänger, es geht einfach etwas langsamer. Bei Computernetzen sind wir es uns gewohnt, dass es zu kurzen Unterbrüchen kommen kann oder zu kleinen Störungen die aber kaum wahrnehmbar sind. Das Design-Ziel bei IP-basierten Netzen war ein völlig anders, als beim alten Telefonnetz. Bei IP-Netzen war es viel wichtiger, dass auch bei grösseren Ausfällen und Katastrophen irgendwie innert kurzer Zeit wieder eine Verbindung hergestellt werden konnte, dafür ist die Konstanz einer bestehenden Verbindung nicht so wichtig wie beim Telefonnetz.

Es erstaunt auch nicht, dass Computernetzwerke anfälliger sind als dedizierte Telefonleitungen. Alleine schon, weil viel mehr Komponenten nötig sind, um eine Verbindung aufzubauen, sind auch viel mehr Störungen möglich. die Grafik 9.2 zeigt auf, wo bei IP-Netzwerken die meisten Fehlerquellen liegen. Will man nun die Sprach- und Videoübertragung, insbesondere die Telefonie auf ein IP-basiertes Netzwerk verlegen, stellen wir plötzlich ganz neue Anforderungen an das Netzwerk, die zu erfüllen nie geplant war. Bei einer VoIP Verbindung darf nicht plötzlich die geschwindigkeit zusammen fallen, oder ein Computer wegen eines Softwareupdates neu starten und deshalb die Verbindung für

30 Sekunden trennen. Der Endanwender ist schlicht nicht bereit, solche Unterbrüche hin zu nehmen.

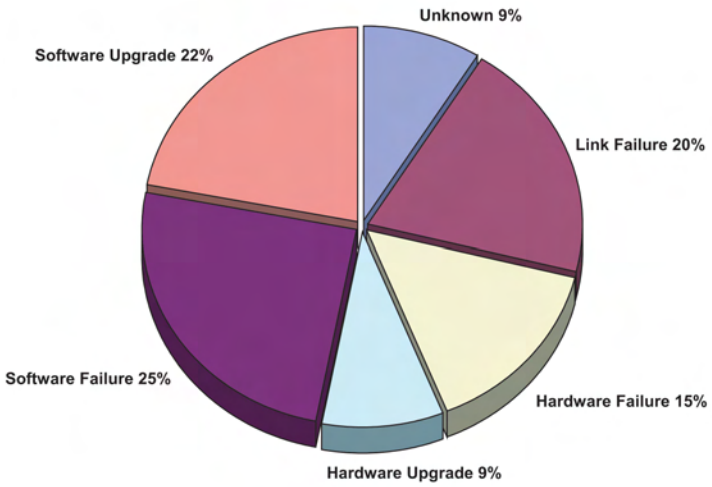


Abbildung 9.2: Ein Grossteil der häufigsten Fehlerquellen in IP-Netzen existieren in Telefonnetzen gar nicht [8].

Etwas anderst sieht es aus, wenn nicht mit dem herkömmlichen Festnetz, sondern mit em Mobiltelefonnetz vergleicht. Dort ist der Konsument bereit, kurze Unterbrüche, Verbindungsabbrüche und Qualitätsschwankungen in Kauf zu nehmen. Somit haben wir zumindest im Bezug auf die Erwartungen an die Dienstleistung einen Vorteil, wenn wir mobile Audio- und Videoübertragung realisieren möchten.

9.1.6 Videospezifische Eigenschaften

Videoübertragung kann in zwei Hauptteile aufgeteilt werden: Interaktives Video (z.B. Videokonferenzen) und Streaming Video. Interkatives Video ist generell gesehen schwieriger zu Übertragen als Streaming Video, da bei Videostreams Buffer verwendet werden können durch die Fehlerkorrekturen und das Ausgleichen von Verzögerungen ermöglicht und vereinfacht werden. Durch den Buffer kann zusätzlich der Datenstrom konstanter gehalten werden (siehe Bandbreitenanforderungen/video). Da die interaktive Videoübertragung ebenfalls zu den soft-realtime Anwendungen gehört, hat man hier praktisch keine Buffer-Möglichkeiten, weswegen die Anforderugen an die Netzwerkverbindung rasant ansteigen. Zudem müssen die Streams in echtzeit kodiert und dekotdiert werden, was vorallem bei stark komprimierenden Codecs sehr rechenintensiv sein kann. Da in der Regel das Codieren wesentlich aufwändiger ist als das decodieren bedeutet dies weitere Nachteile für interaktive Videoübertragung gegenüber dem Streaming.

9.2 Anforderungen an das Netzwerk und die Verbindung

Nachdem nun die grundlegenden Begriffe bekannt sind, gehen wir nun näher auf die Anforderungen an das Netzwerk ein. Die folgenden Abschnitte geben einen Überblick darüber, worauf man achten sollte, was die allgemeinen Richtwerte sind und wie diese voneinander abhängen.

9.2.1 Latency Toleranz

Um eine qualitativ gute Verbindung herstellen zu können, sollte die Latency 150ms nicht überschreiten [11]. Dies wurde durch den ITU Standard G.114 so festgelegt. Dabei meint man mit der Latency nicht die Latency im Netzwerk, sondern die End-zu-End Latency, sprich wie lange ein Laut vom Mund des Sprechers zum Ohr des Zuhörers benötigt. Dies beinhaltet somit beispielsweise auch die Kompressions- und Dekompressionszeit. Die Grafik 9.3 zeigt eine Übersicht wo überall Delays entstehen können. Wie man auf diesem Schema sehr gut sieht, wird alleine bei den beiden Endpunkten für die ganze Codierung und Decodierung sowie für das Buffering etwa die Hälfte der zur Verfügung stehenden Zeit aufgebraucht. Die Netzwerkübertragung sollte also eine möglichst tiefe Latency haben, um den Grenzwert noch einhalten zu können.

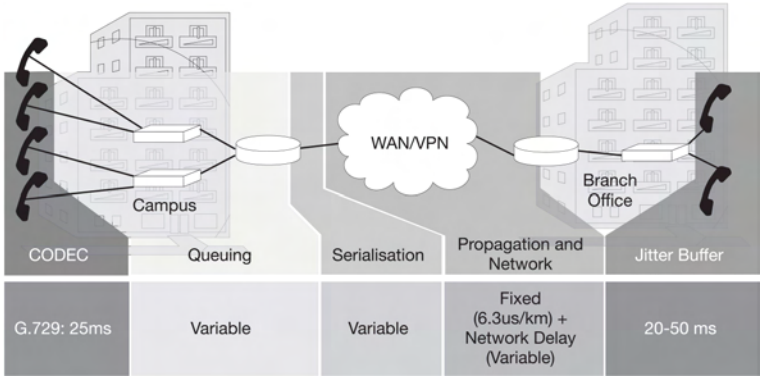


Abbildung 9.3: Etwa die Hälfte der zur Verfügungstehenden Übertragungszeit wird an den Endpunkten gebraucht [11].

9.2.2 Jitter Toleranz

Jitter entsteht aus zweierlei Gründen. Einerseits wenn die Latency schwankt, andererseits wenn ein Paket ein langsames überspringt. Beide Fälle führen bei Voice und Videoübertragung schnell zu höhrbaren und sichtbaren Störungen. Um dies auszugleichen gibt es Jitter-Buffer. Ein Jitter-Buffer ist ein relativ kleiner Buffer, der die ankommenden Pakete sammelt, sie in einen temporären Buffer ablegt und sie dann in konstantem Abstand an die Voice/Video-Applikation weitergibt. Solange sich mehrere Pakete im Buffer befinden

können auch übersprungene Pakete wieder richtig eingeordnet werden. Somit ist es natürlich gerade in mobilen Netzwerken, wo schnell unterschiedliche Verzögerungen aufgrund wechselnder Routen entstehen können, von Vorteil, einen möglichst grossen Jitter Buffer zu haben. Doch das Problem an diesem Buffer ist, dass er zu einer Erhöhung der Latency führt. Je grösser der Jitter-Buffer umso grösser wird die end-zu-end Verzögerung bei der Übertragung, das heisst, desto geringer darf die Verzögerung im Netzwerk sein. Jedoch sind mobile Netzwerke auch dafür bekannt, dass sie teilweise grosse Verzögerungen haben, also müsste demzufolge der Jitter Buffer möglichst klein angelegt werden. Ein zu klein gewählter Jitter Buffer kann zwei mögliche Folgen mit sich führen: buffer overflow und buffer underrun.

Ein Buffer Overflow tritt dann auf, wenn der Buffer bereits durch ein paar schnell angekommene Pakete gefüllt ist und noch weitere kommen. Die überschüssigen Pakete, für die kein Speicher mehr verfügbar ist, müssen somit weggeworfen werden, was wiederum zu Packetloss führt.

Ein Buffer Underrun tritt dann auf, wenn die Pakete plötzlich so langsam eintreffen, dass zu einem Zeitpunkt keine Pakete mehr im Buffer liegen, die an die Applikation weitergegeben werden könnten. Es entsteht ein Unterbruch in der Übertragung, bis das nächste Paket eintrifft. Ist dies dann auch noch out of order muss das übersprungene Paket weggeworfen werden und es kommt zudem zu weiteren Störungen. Somit gilt es, den optimalen Mittelwert zu finden, bei dem man mit sowenig Latency wie möglich soviel Jitter als möglich ausgleichen kann.

Mittlerweilen gibt es variable Jitter Buffer, die je nach Anforderung ihre Grösse dynamisch Ändern können und somit den oben genannten Herausforderungen besser gewachsen sind.

Wie wir gesehen haben, ist die Grösse des Jitter Buffers eng an die Äusseren Umstände gebunden und somit nicht fix definierbar. Man hat jedoch in Laborversuchen und Tests feststellen können, dass die Qualität der Voice-Übertragung ab einem Jitter von 30ms signifikant abnimmt. Somit kann als Faustregel angenommen werden, dass der Jitter 30ms nicht überschreiten sollte [11].

9.2.3 Packet Loss Toleranz

Methoden wie etwa das Packet Loss Concealment (PLC) dienen zur „Vertuschung“ von Paketverlusten. Sie interpolieren quasi den fehlenden Teil der Information. Im Fall von PLC geschieht dies durch die Repetition der im letzten Paket empfangenen Audio-Fragmente. Damit die Repetition nicht wahrgenommen wird, wird das Signal etwas gedämpft wiedergegeben. Solche Interpolationsverfahren können ungefär eine Lücke von 20ms im Audio-Stream überbrücken. Ist die Lücke Grösser, wird sie für das menschliche Ohr hörbar. 20ms ist gerade etwa der durchschnittliche Inhalt eines Audio-Pakets. Mit 20ms Stücken müssen 50 Pakete pro Sekunde versandt werden, um eine Sekunde Audio komplett zu Übertragen. Wird diese Paketrate erhöht sinkt natürlich automatisch die dauer der darin enthaltenen Audiodaten (vgl Tabelle 9.1). Mit herkömmlichen PLC methoden können also die Pakete maximal 20ms Audiodaten enthalten, da sonst der verlust eines einzelnen Paketes nicht mehr ausreichend kompensiert werden könnte.

Bei erweiterten Formen von PLC können höhere Zeiträume überbrückt werden, sofern alle Voraussetzugen Stimmen. Ist dies der Fall, können längere Audio-Fragmente in ein Paket gepackt werden und somit kann der Overhead Traffic reduziert werden. Somit ist es wünschenswert, gerade für drahtlose Netzwerke, wo in der Regel hohe Paketverlustraten auftreten, eine möglichst gute Packet Loss Toleranz zu schaffen. Als Faustregel gilt, dass ca 1% Packetloss nicht überschritten werden sollte [11].

9.2.4 Bandbreitenanforderungen

Um ein voice Datenpaket über ein IP-Netzwerk zu übertragen wird es in das RTP (Real-Time Transport Protocol) gekapselt, welches wiederum ein UDP basiertes Protokoll ist. Somit haben wir eine dreifache Kapselung der eigentlichen Sprachdaten: der RTP-Header ist 12 Bytes gross, der UDP-Header 8 Bytes und schlussendlich der IP-Header noch 20 Bytes.

Wenn nun ein Codec mit 20ms langen Ton-Paketen arbeitet, muss er innerhalb einer Sekunde 50 Datenpakete versenden, um den Ton rechtzeitig und ohne Verlangsamung zum Empfänger zu bringen. Die gesamthaft anfallenden Header-Daten pro Datenpaket sind etwa 32 Bytes [6] bis 40 Bytes [11] Gross. Das bedeutet, dass die Kapselung in RTP bereits etwa 16 KBytes/s verbraucht. Die Tabelle 9.1 veranschaulicht die Bandbreitenanforderungen verschiedener Voice Codecs. Achtung: diese Angaben sind jeweils ohne den Layer 2 Traffic.

Tabelle 9.1: Bandbreitenanforderungen verschiedener Codecs [6]

Codec	Period (ms)	Payload size (bytes)	Packet size (bytes)	Payload data size (bytes)	Total data rate (kbps)
G.711	20	160	200	64	80
G.729	20	20	60	8	24
G.723.1	30	24	64	6.4	17
GSM FR	20	33	73	13.2	29.2
GSM EFR	20	31	71	12.4	28.4
iLBC 20 ms	20	38	78	15.2	31.2
iLBC 30 ms	30	50	90	13.3	24

Wie schon im vorherigen Kapitel erwähnt kann durch die Vergrösserung der Payload, also der Nutzdaten, innerhalb eines Pakets Bandbreite eingespart werden, da somit weniger Traffic anfällt. Dies bedeutet aber grössere Ton-Fragmente zu versenden, was nur mit besseren Fehlerkorrekturen oder zuverlässigeren Netzen möglich ist.

9.2.5 Video

Bei der Videoübertragung sind die Netzwerkanforderungen bezüglich Jitter, Latency und Loss ungefär die selben wie bei der Voice übertragung, da Video in der Regel zusammen

mit Ton übertragen wird. Jedoch schaut es bei den Bandbreitenanforderungen besonders für die Interaktive Videoübertragung etwas anders aus: Bei QoS Systemen sollte man bei der Festlegung der zu garantierenden Bandbreite noch eine Sicherheitsmarge von 20% zur nominal benötigten Bandbreite hinzurechnen. Wieso so eine Sicherheitsmarge? Nebst den eigentlichen Daten werden wie auch beim Voice-Traffic Headerdaten übertragen. Da jedoch die Paketgrössen, Paketraten und die Grösse der in einem Paket enthaltenen Nutzdaten bei Videotraffic sehr stark schwanken, lässt sich kein genauer Wert für den Overhead berechnen. Es hat sich aber in der Praxis gezeigt, dass man mit 20% Sicherheitsmarge in der Regel diesen Overhead abdecken kann.

Da Videoconferencing nicht ohne Ton auskommt, hat es die selben Anforderungen an die Latency, Jitter und Verlusteigenschaften des Netzwerks, jedoch sind die Datenverkehrs-Muster für Videokonferenzen ganz anders als für reinen Voice-Verkehr.

Während wie vorhin erklärt bei Audiokonferenzen die Paketgrösse nicht allzustark variiert, schwanken bei Videokonferenzen sowohl die Paketgrössen wie auch die Paketraten (Pakete pro Sekunde) sehr stark. Dies ist einerseits bedingt durch die höhere Menge an Daten (eine prozentual gleiche Variation der Pakete hat in den absoluten Werten natürlich grössere Auswirkungen) andererseits auch durch die Funktionsweise moderner Videocodecs. Anstatt jedes Einzelbild zu komprimieren, werden nur einzelne Standbilder komplett übertragen. Es folgt eine Folge von Vektor-Matrizen, die angeben, welches Pixel sich wohin verschiebt, was hinzukommt und was wegfällt. Nach einer gewissen Anzahl solcher Vektorsätzen folgt dann wieder ein komplettes Standbild (Keyframe) von dem aus der ganze Prozess wieder von vorne beginnt. Die Vektor-Matrizen sind einiges kleiner als ein komplettes Standbild. Somit lassen sich Filme durch diese Kompressionsmethode gut komprimieren. Für die Übertragung bedeutet dies jedoch, dass sobald ein Keyframe übertragen werden muss, viel mehr Bandbreite benötigt wird und viel mehr Pakete versendet werden müssen als für die Übertragung einer Vektor-Matrix. Somit ergibt sich ein Übertragungsschema mit hohen, kurzen Spitzen, die das Netzwerk allerdings unbedingt abdecken können muss, denn sonst kommt es bei jedem Keyframe zu Verzögerungen bei der Wiedergabe. Bei Videostreams können diese Spitzen durch ein gutes Buffering geglättet werden, was jedoch bei real-time Anwendungen wie Videokonferenzen nicht möglich ist (es kann weder vorausgeahnt werden, wie das nächste Keyframe aussieht, noch darf es mehr Zeit für die Übertragung in Anspruch nehmen als eine Vektor-Matrix).

Bei Videostreams sehen die Anforderungen auf Grund der Möglichkeiten des Bufferings etwas anders aus. So können höhere Verlustraten von bis zu 5 Prozent toleriert werden. Die Latency sollte für die meisten Anwendungen 4 bis 5 Sekunden nicht Überschreiten, es könnten jedoch je nach Anwendungszweck auch wesentlich höhere Latencies in Kauf genommen werden. Beispielsweise beim Betrachten eines archivierten Filmes spielt die Latency lediglich beim Starten des Films eine Rolle. Ist der Buffer erst mal gefüllt, merkt der Betrachter davon nichts mehr. Jitter spielt nahezu keine Rolle bei Streaming Video Applikationen, da dieser durch den Buffer gut auszugleichen ist.

9.3 Qualitätsmessung in Sprachübertragungen

Selbst in der heutigen multimedialen Zeit ist Sprache immer noch die wichtigste Kommunikationsform. Deshalb ist eine qualitativ hochwertige Übertragung nach wie vor eine der wichtigsten Eigenschaften eines mobilen Kommunikationssystems. Im Gegensatz zur herkömmlichen Sprachübertragung in drahtgebundenen Netzen ist die mobile Sprachübertragung einigen besonderen Eigenschaften ausgesetzt, die es umso schwieriger machen für eine gute Qualität zu sorgen. In den folgenden Abschnitten werden die unterschiedlichen Eigenschaften der verschiedenen Netzen beleuchtet als auch auf verschiedene Methoden zur Bestimmung der Qualitätsmerkmale solcher Netze eingegangen

9.3.1 Unterschiede zwischen drahtgebundenen und mobilen Netzen

Zu den drahtgebundenen Netzen gehören sowohl die klassischen Public Switched Telephone Networks (PSTN) wie das analoge POTS, das digitale ISDN als auch die neuen VoIP Telefonieangebote über IP basierte Netzwerke.

Unter mobilen Netzen versteht man Dienste die über Funktechnologien wie GSM und UMTS angeboten werden. Auch VoIP kann in Spezialfällen über mobile Netze angeboten werden. So wird VoIP heute bereits über 3G oder WLAN Netze verwendet.

Das Medium, welches zur Sprachübertragung verwendet wird, bestimmt massgeblich über die Qualität einer Verbindung. Es sind insbesondere die folgenden vier Merkmale bezüglich der Eigenschaften eines Übertragungsmediums von Bedeutung [18]:

- Die Datenrate (Data/Bit rate)
- Die Verzögerung (Latency)
- Die Verlustrate (Loss rate)
- Die Schwankungen der Laufzeit (Latency jitter)

Dabei ist zu beachten, dass mobile Netze im Bereich der Datenrate, den zu erwartenden Verzögerungen und Verlustraten in den Paketübertragungen als auch im Bereich der Schwankungen der Laufzeiten erheblich höhere Werte aufweisen als drahtgebundene Netze. Besonders die Verlustrate, als auch die Verzögerungen von Paketübertragungen stellen eine besondere Herausforderung dar, welchen mit geeigneten Mitteln entgegen gewirkt werden muss. Zum Thema der Fehlerkorrektur und Übertragungswiederholungen gehen wir in Abschnitt “Retransmissions” weiter ein.

Tabelle 9.2: MOS Referenztabelle nach ITU-T

Beeinträchtigung	Note
Hervorragend	5
Gut	4
Durchschnittlich	3
Ungenügend	2
Schlecht	1

9.3.2 Wie kann man die Qualität von Sprachübertragungen erheben

Die Messung der Sprachqualität ist ein wichtiges Werkzeug in der Planung, dem Bau und der Kontrolle von Sprachübertragungsnetzen. Auch bei der Entwicklung von neuen Endgeräten werden Methoden benötigt um den Erfolg in der Übertragung von Sprache messen zu können. Dabei wurden in der Vergangenheit verschiedenste Methoden gewählt. Ursprünglich wurden vorallem Experimente mit hunderten von Personen durchgeführt, wobei in echten Sprachübertragungsnetzen Verbindungen jeweils einzeln, innerhalb einer Gruppe von Probanden geprüft wurden.

Dieses Vorgehen wird als subjektives Testen bezeichnet. Da ein solches Vorgehen äusserst zeit- und personalintensiv ist und somit hohe Kosten verursacht wurden neue Methoden entwickelt, die diese Faktoren minimieren. Heutzutage werden andauernd neue Geräte oder Chips wie DSPs entwickelt die im Bereich der Sprachübertragung eingesetzt werden. Ein subjektives Testen dieser Neuentwicklungen ist bei der schieren Masse an Entwicklungen einfach nicht mehr möglich. Deshalb wird auf maschinelles Testen verschiedener Parametern in einer Sprachübertragung ausgewichen. Ein solches Vorgehen wird als objektives Testen bezeichnet.

Die verschiedenen Testmethoden wurden insbesondere von der ITU zusammengestellt. Diese sind als ITU-T Recommendations zu finden, stehen aber der nicht-zahlenden Öffentlichkeit leider nicht zur Verfügung.

Mean Opinion Score (MOS)

Der Mean Opinion Score wurde für die subjektiven Tests entwickelt, wird aber heute auch zur Angabe von objektiven Testresultaten verwendet. Der MOS wird statistisch aus den Bewertungen verschiedener Probanden ermittelt und stellt den Mittelwert der Testresultate dar.

Die ITU-T [19] hat für ein subjektiver Testlauf Referenzen zur Verfügung gestellt. Diese Referenzen sind Aufnahmen verschiedener Sprachübertragungen mit verschiedenen Massen an Störungen. Den Referenzübetragungen wurden dabei Noten zwischen 1 (Bad) bis 5 (Excellent) zugewiesen. Siehe Tabelle 9.2.

Ein Proband muss anschliessend den getesteten Verbindungen eine Note nach diesen Massstäben vergeben. Die gesammelten Testergebnisse werden anschliessend als MOS statistisch zusammengefasst.

Perceptual Speech Quality Measure (PSQM)

Die Perceptual Speech Quality Measure (nachfolgend PSQM genannt) ist ein Messwert aus einem Verfahren, das bei KPN Research unter der Leitung von Von Beerends entwickelt und 1993 eingeführt wurde. Es handelt sich dabei um eine angepasste Version von PAQM, (Perceptual Audio Quality Measure), jedoch auf telephonie Anwendungen hin optimiert. Das Verfahren wurde eingeführt, da psychoakkustische Effekte bei Sprache auf den Menschen anders wirken als bei beispielsweise Musikübertragungen. Der Mensch nimmt offensichtlich Sprache anders wahr als gewöhnliche Audioquellen wie Musik. Dies beruht wahrscheinlich auf den Alltagserfahrungen im Umgang mit Sprache, da das menschliche Hirn mit Sprache vertrauter umgehen kann als mit anderen Audioquellen. Dieser Umstand hat dazu geführt, dass ein Testverfahren speziell für Sprache entwickelt wurde, was schliesslich PSQM hervorgebracht hat [19].

PSQM ist ein objektives Testverfahren und wird mit simulierten Umgebungen durchgeführt, welche versuchen die psychoakkustische Darstellung eines Signals im menschlichen Hirn bestmöglich zu repräsentieren. Ausserdem verwendet PSQM auch eine kognitive Modellierung um eine möglichst hohe Korrelation zwischen objektiven und subjektiven Messungen zu erhalten. Die Abweichungen zwischen eines im Voraus kodierten Referenzsignals zu der tatsächlichen Messung wird zur Berechnung des geschätzten Wertes der Qualität einer Sprachübertragung herangezogen. PSQM geht dabei von fixen Verzögerungen der Paketübertragungen auf dem getesteten Netz aus. PSQM wurde 1996 als ITU-T Recommendation P.861 spezifiziert. Bild 9.4 zeigt schematisch das PSQM Blockdiagramm.

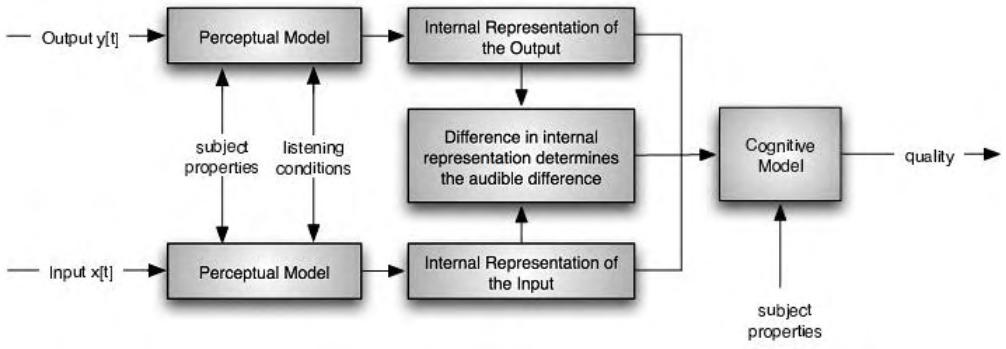


Abbildung 9.4: Opticom's PSQM Blockdiagramm [19]

Perceptual Evaluation of Speech Quality (PESQ)

Perceptual Evaluation of Speech Quality (nachfolgend PESQ genannt) wurde ebenfalls von KPN Research entwickelt. Allerdings hat bei diesem Verfahren nun auch die British Telecom (BT) mitgewirkt [19].

In den letzten Jahren hat die Telekommunikation verschiedenste technologische Wege eingeschlagen. Dabei werden heute Sprachverbindungen über unterschiedlichste Netze aufgebaut und schon lange nicht mehr nur innerhalb eines abgeschlossenen, von der Technologie identisches Netz durchgeführt. Es ist Heute beispielsweise denkbar, dass eine Verbindung über VoIP iniziert wird, diese über ein gewöhnliches ISDN ins PSTN eingeschleust und von dort weiter an ein GSM Netz gereicht wird. Da nun verschiedene Netze in einer einzigen Sprachübertragung involviert sind müssen Testverfahren entwickelt werden, welche die verschiedenen Verzögerungen in der Paketübertragung dieser Netze berücksichtigen. Zusätzlich erschwerend wirkt sich die Tatsache aus, dass in mobilen Netzen jederzeit ein Handover von einer Technologie zur Anderen möglich ist, so wie zum Beispiel UMTS zu GSM oder umgekehrt.

KPN und BT haben mit PSQM(+) als Grundlage ein neues, objektives Testverfahren zusammen gestellt welches diese neuen, hybriden Gegebenheiten berücksichtigt. PESQ ist jedoch wie PSQM eine nicht öffentliche Technologie und muss von KPN und BT lizenziert werden. Das Deutsche Unternehmen Opticom GmbH dient dabei als Vermittler dieser Lizenzen und bietet verschiedene Produkte an, welche die vorgestellten Testverfahren implementieren.

2003 ist bei der IEEE eine Publikation erschienen [21], welche die Korrelation zwischen einer grossen Anzahl von GSM Parametern aus einem echten GSM Netz (GSM-1800) mit den über PESQ ermittelten MOS Werten über zwei verschiedene Verfahren erhebt. Beim einen Verfahren wurde ein SQM (speech quality measure) aus den bekannten GSM Parametern RxQual (geschätzte Qualität eines GSM Kanals), FER (Frame erasure rate for speech frames) und MnMxLFER (Mean of maximum length of erased frames) erstellt, welches mit den aus PESQ ermittelten MOS Werten verglichen wurde. Dabei wurde festgestellt, dass bei 58000 SQM Werten eine "bemerkenswerte" [21] Übereinstimmung der mit PESQ erhobenen Werten und den real gemessenen Werten im GSM Netz vorhanden ist. Es wurde unter Auschluss der Trainingsdaten eine Korrelation von $p = 0.9527$ erzielt. Zu beachten ist, dass unter Anderem Von Beerends als Mitautor von PESQ an dieser Untersuchung beteiligt war. PESQ wurde wiederum von der ITU anerkannt und ist als ITU-T Draft P. 862 zu finden. Bild 9.5 zeigt schematisch das PESQ Blockdiagramm.

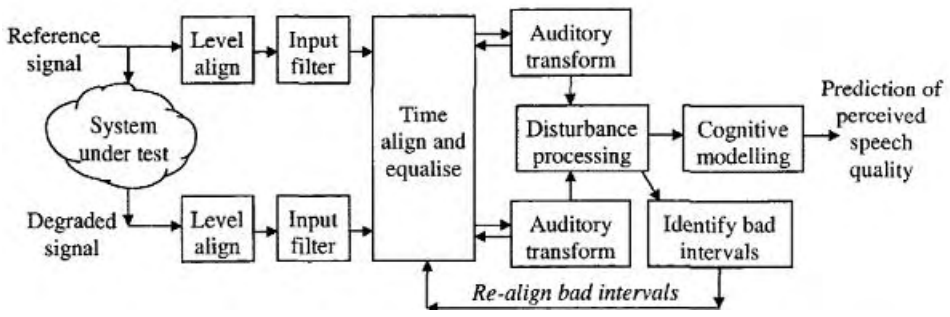


Abbildung 9.5: PESQ Blockdiagramm [20]

9.4 Eigenschaften von Audio- und Video-Codecs

Audio- und Video-Codecs sind die digitale Grundlage der modernen Sprach- und Video-Kommunikation. Wie der Name schon sagt, dienen Codecs der Kodierung und Decodierung von Audio und Video Daten. Dabei werden die Daten den darunter liegenden Übertragungsnetzen entsprechend kodiert. Da diese mittlerweile vollständig digitalisiert sind müssen analoge Daten in digitale Informationen umgewandelt werden. Codecs werden auf der Anwendungsschicht eingesetzt.

Die Entwicklung von Codecs wurde in den letzten Jahren besonders aus Kostenüberlegungen vorangetrieben. Es ist günstiger leistungsfähigere Codecs zu entwickeln als die Netze auszubauen, auf welchen die audio und video Daten übertragen werden sollen. Ebenfalls hat die rasante Steigerung der Rechenkapazität der eingesetzten Geräte dazu beigetragen, dass immer aufwendigere und leistungsintensivere Codecs eingesetzt werden können. Dies hat innerhalb von etwa 10 Jahren dazu geführt, dass die Qualität der übertragenen Daten gegenüber der Steigerung der verfügbaren Bandbreiten im Verhältnis um ein vielfaches zugenommen hat.

9.4.1 Welche Eigenschaften sind für Codecs massgebend

Codecs müssen auf das jeweilige Übertragungsmedium abgestimmt werden. Dabei ist insbesondere die Bandbreite von Bedeutung.

Die verfügbare Bandbreite des Netzes bestimmt mit welcher maximalen Bitrate ein Codec Daten kodieren darf. Je kleiner die Bitrate desto grösser muss die Kompressionsleistung eines Codecs sein. Die verfügbare Bandbreite kann allerdings gewisse Anwendungen wie Videoübertragungen in Echtzeit gänzlich verunmöglichen, falls diese zu klein ist.

Eigenschaften von Audio-Codecs

Audio-Codecs unterscheiden sich in verschiedenen Merkmalen. Es gibt dabei unterschiedliche Codecs die jeweils verschiedene Stärken haben. So gibt es beispielsweise Codecs mit einer sehr hohen Audioqualität wie G.711 wie auch Codecs die speziell für mobile Netze ausgelegt sind wie AMR.

Vorallem unterscheiden sich Audio-Codecs in den folgenden Merkmalen:

- Frequenzgang. Es gibt Codecs die speziell auf die Übertragung von Sprache ausgelegt sind (GSM Codecs), sowie solche die auch einen höheren Frequenzbereich umfassen und somit beispielsweise für die Übertragung von Musik geeignet sind (DAB/DVB Radio). Der Frequenzgang eines Codecs schlägt sich unweigerlich direkt auf die Bitrate eines Codecs nieder.

- Verzögerungen (Latency). In mobilen Anwendungen dürfen nur kurze Verzögerungen auftreten, da wie unter "Anforderungen an das Netzwerk, die Verbindung" beschrieben gewisse Grenzwerte nicht überschritten werden dürfen, bevor diese als störend empfunden werden. Bei der Übertragung von Musik über einen DAB oder DVB Radiosender hingegen können grössere Buffer verwendet werden, wobei das Merkmal der Verzögerung eine kleinere Rolle spielt. Die Verzögerungen treten besonders bei der Kodierung auf. Wenn die Rechenleistung nicht dem Kodieraufwand entspricht können inakzeptable Verzögerungen auftreten.

- Schwankungen in der Laufzeit (Jitter). Wie bei den Verzögerungen hat auch Jitter negative Auswirkungen auf die Qualität von Sprachübertragungen in mobilen Netzen.

Eigenschaften von Video-Codecs

Ähnlich wie bei den Audio-Codecs unterscheiden sich auch Video-Codecs in den Eigenschaften wie der Verzögerung und den Schwankungen der Laufzeit.

Im Gegensatz zum Frequenzgang spricht man aber bei Video-Codecs von Throughput. Dieser bestimmt in der Regel darüber mit welcher Auflösung, Farbtiefe sowie der Anzahl Bilder pro Sekunde ein Video Codec Daten für eine bestimmte, einzuhaltende Bitrate kodieren kann. Der Fortschritt im Bereich der Video-Codecs war in den letzten Jahren enorm und wurde vorallem durch Videotelefonie und multimedia Anwendungen vorangetrieben. Die Kompressionsraten wurden stark angehoben, jedoch ohne dabei die subjektiv wahrgenommene Qualität von Videoübertragungen zu verschlechtern.

9.4.2 Ausgewählte Codecs im Detail

Im folgendnen Abschnitt werden ein paar ausgewählte, bekannte Codecs im GSM und UMTS, DAB und DVB Umfeld portraitiert.

Adaptive Multi Rate (AMR) [13]

AMR wird vorallem in GSM und 3G Netzen (also UMTS) verwendet. Das Codec ermöglicht es alle 20ms die Datenrate anzupassen. Davon stammt auch der Name ab. Es gibt dabei 8 Betriebsmodi. Welcher Modus gewählt wird hängt von der Last auf dem Netz, der Qualität der Funkverbindung sowie der Übertragungsleistung des Gerätes ab. Der Betriebsmodus wird durch das Netz bestimmt und kann mehrmals pro Sekunde angepasst werden. Dieses Codec eignet sich besonders für schmalbandige, in der Qualität unbeständige Funknetze.

Enhanced Variable Rate Coder (EVRC) [13]

EVRC ist ein Codec welches in CDMA Netzen verwendet wird. Es ist somit ebenfalls ein 2,5G/3G Codec. EVRC ist besonders bei tiefen Bitraten sehr effizient und kann bei 9,6, 4.8 oder 1.2kbits eingesetzt werden, wobei 1.2kbits nur für Hintergrundgeräusche, nicht aber für Sprachübertragungen verwendet wird. Durchschnittlich erreicht EVRC eine Datenübertragung von ungefähr 6kbits. EVRC basiert auf RCELP (Relaxed Code Excited Linear Predictive Coding). EVRC kann einen MOS von 3.8 erreichen [17].

G.723.1 [14]

G.723.1 ist ebenfalls ein Codec für kleine Bandbreiten. G.723.1 ist ein ITU Standard und existiert in zwei verschiedenen Betriebsmodi. Einerseits mit 6.3kbits anhand vom MPC-MLQ Algorithmus als auch mit 5.3kbits anhand vom ACELP Algorithmus. Dabei erreicht das Codec gemäss Cisco [16] als MPC-MLQ Version einen Wert von 3.9, als ACELP Version 3.65 MOS (mean oppinion score).

MPEG-1 Audio Layer II

MPEG-1 Audio Layer II ist auch als MP2 bekannt und wird bei 192kbps in DAB Radiosystemen verwendet. MP2 bietet vergleichsweise eine tiefere Kompressionsrate, dafür aber ein viel höherer Frequenzgang als andere Codecs. Das Codec hat eine hohe Originaltreue und wird deshalb vorallem zur Kodierung von Musik eingesetzt. MP2 wird im DAB Radio Bereich demnächst durch AAC+ abgelöst werden [15].

H.263

H.263 ist weit verbreitet und vor allem wegen dem frühen Einsatz als Codec für Videokonferenzen bekannt. Es existiert seit 1995 und wurde von der ITU-T entwickelt. H.263 ist aus MPEG-1/2 entstanden und bietet vergleichsweise eine hohe Bildqualität bei relativ niedrigen Bitraten.

H.264/AVC

H.264 wurde vor kurzem eingeführt und als Standard anerkannt. Es löst dabei das bereits als veraltet betrachtetes H.263 ab. H.264 bietet gegenüber H.263 eine weitaus höhere Bildqualität bei gleichbleibender Bitrate an. Als Nachteil wird jedoch der hohe Rechenaufwand für das Encoding gewertet. Dies dürfte aber durch die rasante Erhöhung der Rechenleistung in mobilen Geräten bald keine Rolle mehr spielen. Trotzdem wird in vielen 3G Anwendungen nach wie vor H.263 verwendet.

Im DVB Bereich spielt der hohe Kodierungsaufwand jedoch keine Rolle, da es sich bei DVB Anwendungen um Broadcast-Systeme wie Fernsehübertragungen handelt. H.264 kommt

dabei in DVB-S2 zum Einsatz für die Übertragung von HDTV Programmen. H.264 ist technisch gesehen mit der MPEG-4 Video-Kodierung identisch.

9.5 Fehlerkorrektur

Drahtlose Netzwerke sind natürlich bedingt viel anfälliger auf Fehler in der Datenübertragung als drahtgebundene. Da die physische Netzwerkschicht hierbei aus dem Luftraum an und für sich als shared medium zur Verfügung steht. Während einige Technologien wie GSM oder UMTS dedizierte Frequenzbereiche nutzen, die von einer Behörde an die Betreiber des Netzwerkes lizenziert werden, nutzen andere Technoligien wie WLAN das für alle frei Nutzbare ISM-Band, was zu zusätzlichen Störungen der Datenübertragung von anderen Geräten auf demselben Frequenzband führen kann.

Als erstes ist zwischen fehlererkennenden und fehlerkorrigierenden Codes zu unterscheiden. Während fehlererkennende Codes nur zur Überprüfung der korrekten Datenübertragung dienen, können die fehlerkorrigierenden Codes zur Bereinigung dieser Fehler verwendet werden. Die in drahtlosnetzwerken Eingesetzen multimedialen Anwendungen verlangen häufig Echtzeitverfügbarkeit, weshalb das erneute Anfordern (Retransmission) zu lange dauert, vor allem über drahtlose Links. Hinzu kommt, dass die Retransmission-Anforderung selbst wieder falsch übertragen werden könnte. Bei unidirektionalen Übertagungen, wie beispielsweise bei DVB, ist es gar unmöglich fehlerhafte Daten erneut anzufordern. Fehlerkorrigierende Codes sind in diesem Anwendungsgebiet deshalb meist sinnvoller, währenddem fehlererkennende Codes meist für zuverlässigere Links eingesetzt werden, da es dort einfacher ist fehlerhafte Daten erneut anzufordern und die Anforderungen auch ohne grossen Zeitverlust übertragen werden können.

In einigen Anwendungen wird auch eine Kombination beider Mechanismen eingesetzt. Die Übertragung wird mit einem bestimmten Mass an Redundanz übertragen. Erst wenn diese nicht mehr ausreicht um den Fehler zu beheben, werden die entsprechenden Daten neu angefordert. Vor allem in Echtzeitmultimediaanwendungen ist es häufig vernachlässigbar, falls ein geringer Teil des Datenstroms fehlerbehaftet ist. Beispielsweise spielt es keine grosse Rolle wenn einige Pixel eines Frames in einem Video mit der falschen Farbwert angezeigt werden, da die eigentlich übertragenen Informationen für den Menschen selbst auch eine gewisse Redundanz enthalten.

Wie in Abbildung 9.6 zu sehen ist, verringert sich der benötigte Signalrauschabstand (x-Achse), um eine gewisse Bit Error Rate (y-Achse) zu erreichen, im Vergleich zu Übertragungen ohne FEC bei der Verwendung von Faltungscodes respektive Turbo-Codes massiv. Dadurch ist es auch möglich Energie für die Übertragung einzusparen, was sich vor allem bei mobilen Geräten durch längere Akkulaufzeiten positiv auswirkt.

Mathematisch gesehen werden bei der Vorwärtsfehlerkorrektur aus k Informationssymbolen n Codesymbole erzeugt. Die Coderate R beschreibt das Verhältnis an übertragenen Informationssymbolen pro Codewort [27]: $R = k/n$ Eine Coderate von 1/2 gibt demnach an, dass aus 1 Datenbit 2 zu übertragende Ausgangsbits entstehen. Die resultierende zu

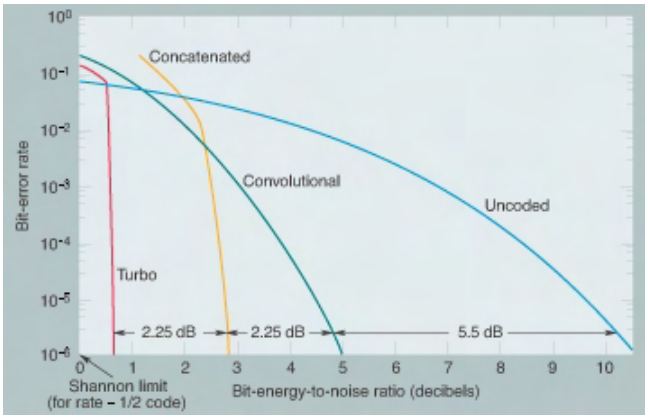


Abbildung 9.6: Vergleich Bit Error Rate bei Verwendung verschiedener Fehlerkorrekturalgorithmen [32]

übertragende Datenmenge ist demnach grösser als die ursprüngliche. Je nach wahl der Coderate geht hierbei mehr oder weniger Übertragungskapazität zu Lasten der Redundanz verloren.

In einem ersten Teil wird anhand von CRC ein fehlererkennender Code kurz erklärt. Im zweiten Teil soll anhand von verschiedenen Implementierungen fehlerkorrigierender Codes gezeigt werden, wobei speziell auf die Forward Error Correction (FEC) eingegangen wird.

9.5.1 Fehlererkennung

In der Praxis wird als fehlererkennender Code meist der sogenannte CRC (Cyclic Redundandy Check) eingesetzt. Dieser interpretiert die Bitstrings der Übertragung als Koeffizienten von Polynomen mit den Werten 0 und 1. Vor dem Sendevorgang müssen sich Sender und Empfänger auf ein Generatorpolynom einigen. In der Praxis werden dabei häufig dieselben Polynome eingesetzt, die sich besonders geeignet für die Erkennung von Fehlern erwiesen haben.

Den zu übertragenen Bitstring, der in jedem Fall länger sein muss als das Generatorpolynom, wird nun eine Prüfsumme angehängt, so dass der resultierende Bitstring durch das Generatorpolynom teilbar ist. Der Empfänger kann durch diese Division nun prüfen, ob sich ein Übertragungsfehler in dem Bitstring ereigenet hat. Mit CRC und der geschickten wahl des Generatorpolynoms können viele, aber nicht alle Übertragungsfehler erkannt werden. Beispielsweise versinken einige Zweibitfehler [22].

CRC wird in vielen Protokollen und Technologien (z.B. IP, Bluetooth Baseband) als Prüfsumme für die Header verwendet, um die korrekte Übertragung der Headerinformationen sicherzustellen. Die auf dem AMR-Codec basierte Sprachübertragung in UMTS verwendet ebenfalls CRC, um die Klasse A Daten sowie die Kontrolldaten zu schützen. Der Prozess der CRC-Prüfsummenberechung werden in den Node B, welche auf dem OSI-Layer 1 (physikalische Schicht) liegen, durchgeführt. Diese Empfangen die Datenströme (Dedicated Channels DCH) in Zeitschlitze unterteilt (Time Transmit Interval TTI) von

den RNCs. Die Daten der Klasse B und C, die für zur Sprachqualität einen geringeren Anteil beisteuern werden hingegen nicht mit CRC gesichert [23]. Abbildung 9.7 zeigt die verschiedenen Dedicated Channels und wie die in den Zeitschlitzen übertragenen Daten aufgebaut sind.

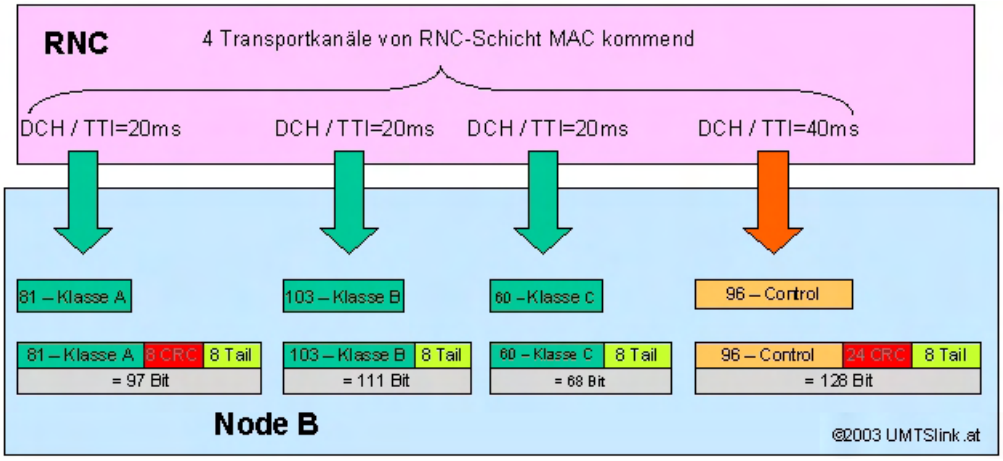


Abbildung 9.7: CRC-Bit Berechung im Node B / UMTS [23]

9.5.2 Interleaving

Interleaving kann als einfaches Durcheinanderwürfeln des Datenstroms verstanden werden. Durch diese Prozedur kann der Datenstrom bzw. die eingesetzten fehlerkorrigierenden Codes robuster vor allem gegenüber Burstfehlern gemacht werden. Burstfehler sind kurzweilige Störungen des Übertragungsmediums. Bei drahtlosen Übertragungen kommen hierbei beispielsweise Störungen von Gewitterblitzen oder anderwertige Entladungen in Frage. Interleaving selbst stellt deshalb bloss eine unterstützende Massnahme zur Fehlerkorrektur, nicht aber selbst ein solcher Mechanismus dar.

Werden Daten in ihrer ursprünglichen Reihenfolge übertragen, sind von einem Burstfehler möglicherweise mehr Bits eines Datenpakets betroffen, als vom fehlerkorrigierenden Code behoben werden können. Durch das Durcheinanderwürfeln verteilt sich in der Folge der Fehler auf mehrere Datenpackete als Einzelbitfehler, die von den fehlerkorrigierenden Codes besser behoben werden können [30]. Interleaving wird bei fast allen Übertragungen eingestzt und geschieht häufig zusammen mit dem Multi- respektive Demultiplexing (bepsielsweise bei UMTS). Ein grosser Nachteil von Interleaving ist die entstehende Zeitverzögerung, da alle Daten die durch den Interleavingprozess laufen sollen, bereits vorhanden sein müssen. Es muss deshalb eine Abwägung zwischen der Anzahl der zu interleavenden Datenpaketen und maximal akzeptierbaren Delay gemacht werden. Abbildung 9.8 stellt den Vorteil des Interleavingvorgangs exemplarisch an einer durch einen Burstfehler beschädigten Übertragung dar.

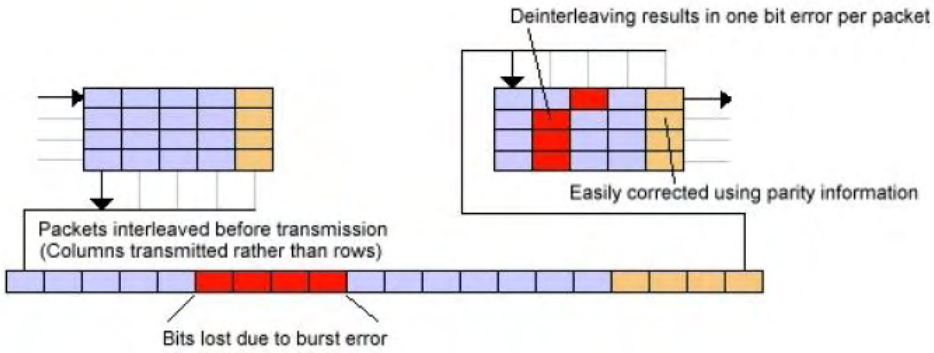


Abbildung 9.8: Interleaving Prinzip [29]

9.5.3 Blockcode: Reed Solomon

Der Reed-Solomon Code ist ein symbolorientierter Blockcode [25]. Als Blockcodes bezeichnet man Algorithmen, die auf Datenblöcke fester Länge angewandt werden [31]. Die Berechungen des Algorithmus werden in einem abgeschlossenen Zahlenraum durchgeführt, der Galois-Feld gennant wird. Ähnlich wie bei CRC geschieht die Encodierung mittels eines Generatorpolynoms. Dabei jeweils doppelt so viele Korrektursymbole den Nutzdaten hinzugefügt werden, wie später beim Decodierprozess behoben werden sollen können. Das Resultat der Division der Nutzdaten durch das Generatorpolynom werden den Nutzdaten angehängt. Anschliessend werden die Daten übertragen.

Beim Decodierungsprozess macht man sich zu Nutze, dass das Generatorpolynom regelmässig auftretende Nullstellen erzeugt. Sind diese nicht vorhanden, wurde das Signal während der Übertragung verfälscht. Der Decoder kann durch algebraische Operationen ein Fehlercodewort erzeugen und durch dieses die ursprünglichen Nutzdaten zurückgewinnen.

Reed-Solomon Fehlercodes werden im GSM Enhanced Full Rate Codec, sowie bei DVB eingesetzt. Abbildung 9.9 zeigt das Ablaufdiagramm für die Reed Solomon Codeberechnung im einem DVB-Encoder und Decoder im Zeit- und Frequenzbereich.

9.5.4 Faltungscodes

Im Gegensatz zu den Blockcodes, werden Faltungscodes auf einen kontinuierlichen Datenstrom angewendet [31]. Faltungscodierung kann als “Verschmieren” der Daten verstanden werden [25]. Es gibt eine Reihe von verschiedenen Arten von Faltungscodes. Unterscheidungsmerkale sind systematisch / nicht-systematisch, rekursiv / nicht-rekursiv, sowie terminert / nicht-terminiert [26].

Der Encoder kann als Zustandsmaschine aufgefasst werden. Im Unterschied zu Blockcodes hat dieser ein Gedächtnis, d.h. die zuvor bearbeiteten Blöcke werden in den Codierprozess miteinbezogen. Die Anzahl miteinzubeziehender Blöcke wird mit dem Parameter m angegeben. Faltungscodes können daher in Hardware sehr einfach durch Schieberegister implementiert werden. Die Payload wird also gewissermassen über mehrere Symbole verteilt.

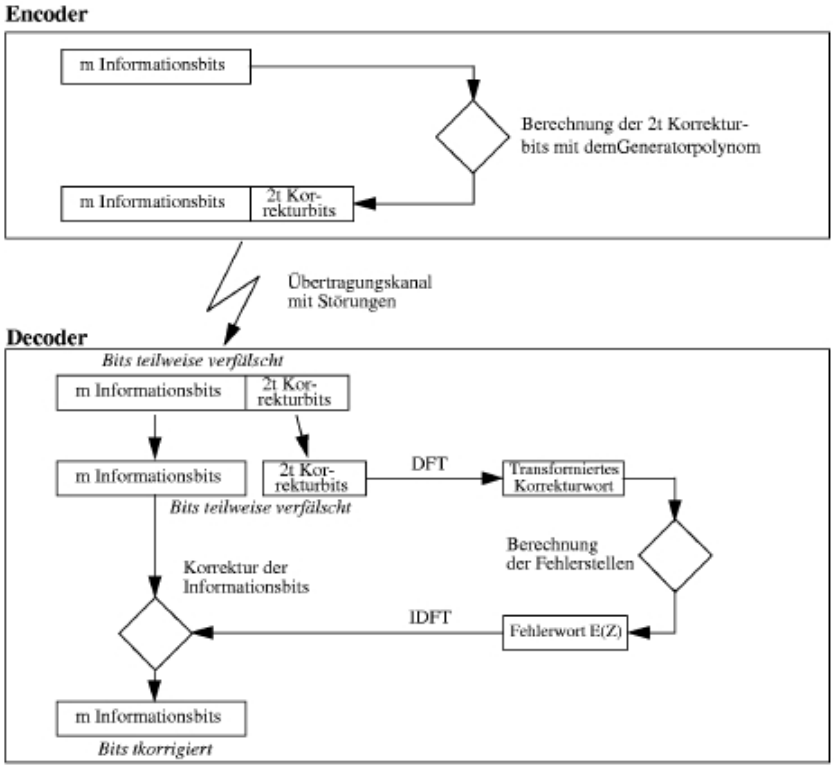


Abbildung 9.9: Reed-Solomon Encoder und Decoder [25]

Im Gegensatz zu den Blockcodes werden auf der Decoderseite keine mathematischen Prüffunktionen eingesetzt, sondern eine Schätzung der Codewörter aus dem empfangenen Datenstrom vorgenommen. Der Decoder schätzt die Ursprungsdaten aus den empfangenen Daten, indem er jene mit der grössten Wahrscheinlichkeit ausgibt. Dies wird als sogenannter Maximum-Likelihood-Decoder bezeichnet. Die bekannteste und verbreitetste Implementierung ist hierbei der Viterbi-Algorithmus [26], [25].

Faltungscodes und der Viterbi-Decoder werden in vielen Systemen verwendet: CDMA und GSM Mobilfunk, Wireless LAN, Raumfahrtmissionen sowie bei DVB. Die Implementierung geschieht dabei jeweils auf der untersten Schicht, der physikalischen Schicht. DVB arbeitet dabei mit einer zweischichten Fehlerkorrektur: der innere Code stellt ein Reed-Solomon Code dar, der äussere Code besteht aus mehreren Faltungscodes. Das Verhältnis zwischen Gesamt- und Nutzdatenrate des Reed-Solomon-Codes beträgt 204 zu 188 [24]. Bei der Kanaleinstellung von DVB-Receivern muss die Coderate des Viterbi-Decoders, welche häufig als FEC-Rate bezeichnet wird, meist manuell miteingegeben werden um Daten von einem Kanal erfolgreich empfangen zu können.

9.5.5 Turbo Codes

Turbo-Codes sind prinzipiell rekursive Faltungscodes [28]. Der Encoder arbeitet auf der Grundlage von zwei Faltungscode-Encoder, wobei die Nutzdaten den ersten Encoder direkt durchlaufen. Bevor die Nutzdaten den zweiten, parallelgeschalteten Encoder durch-

laufen, werden diese mittels eines zuvor definierten Interleavers durcheinandergewürfelt. Die Resultate beider Encoder werden übertragen. Die Decoder arbeiten mit ihren entsprechenden Decodieralgorithmen, wobei das Resultat des einen Decoders in den anderen geleitet wird. Je nach Implementierung geschieht dies mittels Konkatenation oder einem weiteren Interleaver [33].

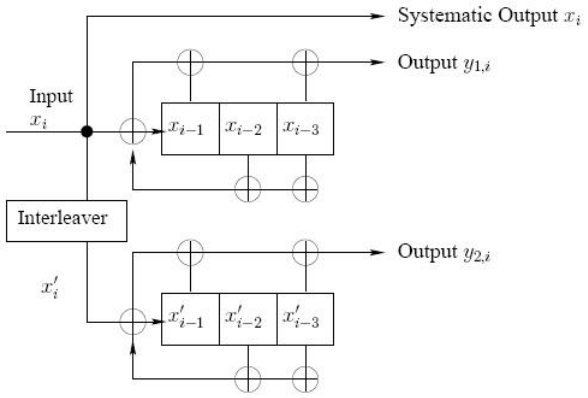


Abbildung 9.10: UMTS Turbo-Encoder [34]

Turbo-Codes werden in CDMA-2000, UMTS und für DVB-RCS eingesetzt. Sie sind im Gegensatz zu den schon vor einigen Jahrzenten entstandenen Block- und Faltungscodes eine relativ neue Entwicklung der 90er-Jahre und bieten eine höhere Verbindungsperformance und Fehlersicherheit. UMTS verwendet für die Kanalcodierung konventionelle 1:3 Faltungcodes für kleinere Rahmengrössen, bei grösseren Rahmen sind diese jedoch ineffizient, weshalb 1:3 Turbo-Codes zum Einsatz kommen [23]. Diese Implementierung ist ist Abbildung 9.10 graphisch dargestellt. DVB-RCS ist ein in der Praxis noch wenig genutztes Verfahren für die Rückkanalkommunikation via Satellit.

9.5.6 Low Density Parity Check Codes (LDPC)

Ähnlich der Hamming-Kodierung werden bei LDPC die Nutzdaten zusammen mit den Redundanzdaten mit einer binären Paritätsmatrix multipliziert. Das Ergebnis dieser Berechnung muss 0 ergeben. Da die Nutzdaten bekannt sind und die Paritätsmatrix zufällig bestimmt wird, können die benötigten Redundanzdaten berechnet werden. Der Decoder verwendet eine ähnliche Formel, um die Nutzdaten zurückzuerhalten. Der Decoder versucht die ursprünglichen Nutzdaten in einem iterativen Verfahren zu schätzen bis ein gesetzter Grenzwert für ein spezifisches Codewort erreicht ist. Ist die maximal zulässige Anzahl Iterationen erreicht, ohne dass der Grenzwert überschritten wird, sind die empfangenen Daten unbrauchbar und der Decoder gibt auf [36].

LDPCs werden heute als eine der Fehlerkorrekturverfahren in DVB-S2 verwendet. Bei der Evaluation verschiedener Algorithmen im Jahre 2003 hat sich dieses Verfahren gegen sechs andere, unter anderem eine Reihe von Turbo Codes, durchsetzen können [35].

9.6 Retransmission

Neben der Methoden der Forward Error Correction (FEC), werden auch die sogennanten Automatic Repeat reQuest (ARQ) zur Kanalkodierung gezählt. Bei reinen ARQ-Verfahren wird die Übertragung nur durch einen fehlererkennenden Code, etwa CRC, geschützt. Es werden keinerlei redundante Daten mitübertragen, so dass im Falle eines Übertragungsfehlers die entsprechenden Daten neu angefordert werden müssen. Dies geschieht mittels NACK-Nachricht (Negative Acknowledgement), die den Sender veranlasst die entsprechenden Daten erneut zu übertragen. Ist die Übertragung erfolgreich wird dies dem Sender mittels einer ACK-Nachricht (Acknowledgement) bestätigt. Es ist offensichtlich, dass für das funktionieren dieses Mechanismus zwingend ein Rückkanal für die Übertragungswiederholungsanforderung existieren muss. Eine Anwendung etwa bei DVB kommt deswegen nicht in Frage. Reine ARQ-Verfahren machen deshalb vor allem bei Übetragungen über zuverlässige Links Sinn, bei welchem Übertragungsfehler selten sind und etwaige ARQs schnell zum Sender übermittelt werden können. Bei schlechten Übertragungsbedingungen führt das ständige Wiederholen meist zu einem fast kompletten Ausfall der Übertragung [28].

9.6.1 Automatic Repeat reQuest (ARQ)

Grundsätzlich werden drei verschiedene ARQ-Verfahren unterschieden: Stop-and Wait, Go-Back-N und Selective Repeat ARQ. Stop-and-Wait ist dabei das einfachste Verfahren, bei dem der Empfang jeder einzelnen Dateneinheit dem Sender bestätigt wird bevor dieser die nächste Einheit senden darf. Wird die Dateneinheit innerhalb eines festgelegten Zeitrahmens nicht bestätigt, so wird die Dateneinheit erneut gesendet. Obwohl diese Methode sehr einfach zu implementieren ist, wird sie kaum eingesetzt, da grosse Leerlaufzeiten beim Warten auf die Bestätigungen auftreten. Das Go-Back-N-Verfahren umgeht dieses Problem, da der Sender nicht auf die Bestätigungen des Empfängers wartete, sondern die Daten kontinuierlich sendet. Stellt der Empfänger ein Fehler in der Übertragung fest, lässt er den Sender wissen welche Dateneinheit N fehlerhaft ist. Der Sender sendet daraufhin diese Dateneinheit sowie alle darauf folgenden noch einmal. Dieses Verfahren wird auf der MAC-Schicht von 802.11 Wireless LAN eingesetzt. Der Nachteil, dass beim Go-Back-N-Verfahren alle Dateneinheiten nach dem Fehler erneut mitübertragen werden, umgeht der Selective Repeat ARQ, indem nur die fehlerhafte Dateneinheit neu übertragen wird, während die folgenden, fehlerfreien Dateneinheiten in einem Puffer gehalten werden, bevor diese der nächsten Schicht übergeben werden. Dieses Verfahren wird beispielsweise von der aktuellen Version von TCP verwendet. Abbildung 9.11 zeigt das senderseitige Zustandsdiagramm der WiMax ARQ-Implementierung.

ARQ-Mechanismen in Wireless Systemen werden im Gegensatz zur Forward Error Correction auf der Bitübertragungsschicht sinnvollerweise in der Sicherungsschicht implementiert. Natürlich können auch weitere auf höheren Protokollschichten angewandt werden, was beispielsweise bei der Verwendung von TCP über wireless Links geschieht. Dies kann jedoch erhebliche Probleme verursachen, da TCP bei Fehlübertragungen von einer Netzüberlastung ausgeht und den Slow Start Algorithmus anwendet. Auf wireless Links hinge-

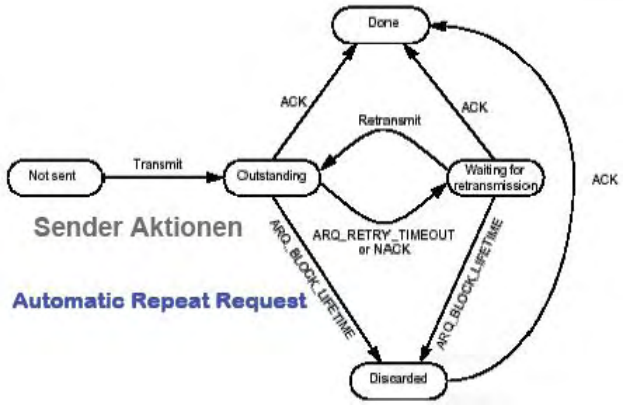


Abbildung 9.11: Zustandsdiagramm eines ARQ-Implementierung auf der Senderseite [37]

gen werden die meisten Übertragungsfehler durch das unzuverlässige Medium verursacht. In der Folge sinkt die Übertragungsgeschwindigkeit unnötig [38].

UMTS verwendet je nach Betriebsmodus ARQs auf der Radio Link Control (RLC) Schicht. Die RLC-Schicht entspricht bei UMTS der LLC-Schicht /cite umtslink. Für Wireless LANs nach IEEE 802.11 werden Stop-and-Wait ARQs auf der MAC-Schicht angewendet. Dies führt bei der Übertragung von Multimediainhalten zu unnötigen Verzögerungen. Die Quality of Service Mechanismen für Wireless LANs in 802.11e sind gegenwärtig in der Endphase der Spezifikation. Diese Erweiterung des WLAN-Standards ermöglicht eine effizientere Übertragung von Multimediainhalten. Durch die sogennante QoSNoACK-Option soll die Acknowledgement-Funktionen der MAC-Schicht komplett deaktiviert werden können. Ausserdem war es ursprünglich geplant Forward Error Correction auf der MAC-Schicht einzuführen, was wegen des grossen Overheads jedoch wieder verworfen wurde. Stattdessen werden nun hybride ARQ-Verfahren vorgeschlagen [39], [40], [41].

9.6.2 Hybrid Automatic Repeat reQuest (HARQ)

Bei HARQ-Verfahren werden FEC- and ARQ-Mechanismen gemeinsam eingesetzt. Während die Forward Error Correction ein gewisses Mindestmass an Fehlertoleranz garantiert, werden ARQs bei grösseren Übertragungsfehlern eingesetzt, bei denen die Fehlerkorrektur allein nicht mehr ausreicht. Dies ermöglich es den FEC-Overhead gering zu halten, da kürzere Codes eingesetzt werden können. Die Latenzzeit ist in der Regel niedriger als bei reinen ARQ-Verfahren. Der Nachteil von HARQ ist die relativ komplexe Implementierung.

802.11e unterscheidet zwei Typen von hybriden ARQ-Verfahren: Beim Type-1 werden auf seiten der Fehlerkorrektur Paritätsbits verwendet. Ist ein Paket fehlerhaft empfangen worden, so wird versucht dieses anhand der angefügten redundanten Daten zu korrigieren. Ist dies nicht möglich, wird das Paket vom Sender erneut angefordert. Dieser Prozess wird wiederholt, bis das Paket erfolgreich empfangen wurde oder die maximal zulässige Anzahl Iterationen erreicht wurde. Während Type-1 Pakete auch dann komplett verwirft, wenn Teile davon korrekt empfangen wurden, werden diese beim Type-2 gepuffert. Ist das bei einer erneuten Übertragung empfangene Paket ebenfalls fehlerhaft, so wird versucht

mittels den korrekten und redundanten Anteilen des gepufferten Pakets zusammen mit jenen des neuen Pakets dieses zu rekonstruieren [40].

UMTS verwendet, wie bereits erwähnt, in seiner ursprünglichen Konfiguration des Radiozugriffsnetz UTRAN ARQ-Verfahren auf der RLC-Schicht. Da die Basisstationen (Node B) von UMTS für die RLC-Schicht komplett transparent sind, müssen die ARQ-Anfragen von den RNCs verarbeitet werden. Diese Radio Network Controller versorgen bis zu mehreren hundert Basisstationen und sind daher oft mit dem Halten von Pufferinformationen aller Datenübertrangssitzungen überfordert. Dies führt dazu, dass die Verarbeitung einer ARQ-Anfrage einige hundert Millisekunden in Anspruch nehmen kann. Da dies der schnellen Datenübertragung von HSDPA nicht genügt, wurde die neue MAC-hs Protokollschicht eingeführt. Diese ist in den Node Bs implementiert, die nun die entsprechenden Puffer verwalten. In Abbildung 9.12 sind die Neuerungen des Protokollstacks graphisch dargestellt. Da die Basisstationen näher bei den Endgeräten gelegen sind und über eine einzelne Basisstation weniger Datenübertragunssitzungen laufen, kann die Reaktionszeit für ARQ-Anfragen auf wenige Millisekunden verkürzt werden. Im Übrigen verwendet HSDPA einen ähnlichen HARQ-Mechanismus wie 802.11e Type-2. Auch teilweise fehlerhafte Daten werden im Endgerätepuffer gehalten, bis genügend Daten gesammelt werden konnten um das Paket rekonstruieren zu können [23], [42].

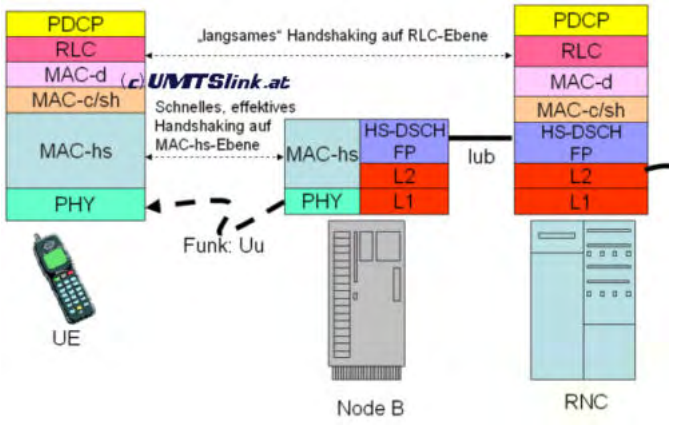


Abbildung 9.12: Unterschiedliche Implementierung der ARQ-Mechanismen in konventionellem UMTS (oben) und HSDPA (unten) [23]

9.7 Quality of Service (QoS)

Netzwerke ohne QoS-Mechanismen folgen dem Best Effort Prinzip. Die zur Verfügung stehenden Ressourcen werden gleichenteils auf die Benutzer aufgeteilt. In drahtlosen Netzwerken ist der limitierende Faktor hierbei die Luftschnittstellt, da diese sehr fehleranfällig ist. Ausserdem sind die Übertragungskapazitäten beschränkt, da nur ein Teil des Frequenzspektrums jeweils einer Technologie zur Verfügung steht, die teilweise auch mit anderen Technologien geteilt werden muss (ISM-Bänder). Netzwerke, die für Multimediaanwendungen verwendet werden, müssen aber gewissen Ansprüchen genügen können, damit die Anwendung einwandfrei funktionieren können.

9.7.1 Dienstgüteklassen

Jede Technologie kann eigene Dienstgüteklassen definieren. Als Beispiel sollen jene von UMTS aufgezeigt werden. UMTS eignet sich besonders gut als beispiel, weil es ein öffentlichen Netz mit vielen Benutzern darstellt, das von Anfang an für die Übertragung von Multimediainhalten konstriert wurde. UMTS unterscheidet zwischen vier verschiedenen Verkehrsklassen, welche den unterschiedlichen Anforderungen verschiedener Anwendungen gerecht werden sollen. Abbildung 9.13 zeigt eine Tabelle, welche die entsprechenden Verkehrklassen mit ihren Eigenschaften und deren Anwendungensbereichen darstellt.

Verkehrsklassen	Konversations-Klasse	Streaming-Klasse	Interaktive-Klasse	Hintergrund Übertragung
Eigenschaften	geringe Zeitverzögerungen keine zeitlichen Schwankungen Unempfindlich gegen kleine Fehler	Konstante Übertragungs-geschwindigkeit Unempfindlich gegen konstante Zeitdelays	Empfindlich gegen Fehler Keine konstante Übertragungs-geschwindigkeit nötig	Unempfindlich gegen Zeitverzögerungen Empfindlich gegen Fehler
Anwendungen	**Sprache** **Videotelefonie**	**Audioübertragung** **Videoübertragung**	**Internetsurfen**	**Email-Download** **Faxübertragung**

Abbildung 9.13: UMTS Verkehrsklassen [23]

Es ist klar, dass bei Telefon- oder Videogesprächen einzelne Fehler in der Übertragung vernachlässigbar sind, da die menschliche Sprache bzw. das Auge selbst redundant ist. Hingegen wird eine Übertragung, die an Jitter leidet von Benutzern schnell als instabil empfunden. Für Streaming-Anwendungen hingegen ist es wichtig, dass für die Übertragung eine konstante Bandbreite zur Verfügung steht, da diese sonst abbrechen könnte. Die Interaktive-Klasse deckt beispielsweise das Surfen im Internet ab. Zeitverzögerungen und Jitter spielen hierbei keine grosse Rolle, allerdings existiert keine Fehlertolernz. Schliesslich gibt es eine Klasse für Hintergrundanwendungen, die Fehlerfreiheit garantiert, allerdings überhaupt keine Rücksicht auf Zeitverzögerungen nimmt.

Die genauen Anforderungen bezüglich Bandbreite, Verzögerungen und Fehlerfreiheit ist individuell von den verwendeten Codecs abhängig. Dazu siehe das Kapitel über Codecs.

9.7.2 Implementierung

UMTS unterscheidet drei verschiedene QoS-Schichten. Der End-to-End Service definiert hierbei QoS zwischen den zwei Endgeräten, also auch über das UMTS Netz hinaus. Beim UMTS Bearer Service beschränkt man sich auf die Anwendung von der QoS-Parameter auf den Teil der Verbindung, der über Komponenten des UMTS-Netzes läuft. Hierzu wird neben dem Mobiltelefon und der Luftschnittstelle auch das Corenetz des Betreibers gezählt. Auf der untersten Ebene wird schliesslich zwischen Radio Access Bearer und Core Network Bearer unterschieden. Die Sicherstellung von QoS wird hierbei jeweils separat auf seiten der Luftschnittstelle respektive dem Corenetz geprüft.

Implementiert werden die QoS-Mechanismen in UMTS in den Schichten 1 bis 3. Die Node B ist bei nicht HSDPA-Betrieb für die Schichten 2 und 3 transparent, weshalb die RNCs für die QoS-Verwaltung auf der Luftschnittstelle auf seiten des Zugangsnetzwerks verantwortlich sind. Die in der Bitübertragungssicht implementierten Turbo-Codes, sowie die ARQ-Verfahren der RLC-Unterschicht dienen der Fehlererkennung- und korrektur. Auf der MAC-Layer wird das Kanalmanagement sowie das Packet Sheduling vorgenommen. Auf der RLC-Schicht sind die ARQ-Mechanismen implementiert. Die entsprechenden Dienstgüteparameter werden ebenfalls auf dieser Schicht festgestellt. Die RRC-Schicht schliesslich sorgt für die Leistungssteuerung der Bearer Services [23], [43]. Abbildung 9.14 gibt eine graphische Übersicht der verwendeten Schichten.

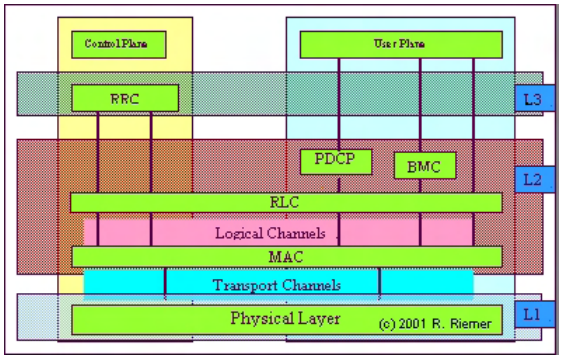


Abbildung 9.14: UMTS Protokollturm [23]

9.8 Zusammenfassung

Wir haben gezeigt, was die speziellen Eigentschaften von mobilen Netzwerken sind und welche besonderen Eigenschaften diese aufweisen. Multimediale Applikationen und Netzwerke müssen aufeinander abgestimmt werden, damit die zur Verfügung gestellten Dienste einwandfrei funktionieren können und von den Benutzern akzeptiert werden. Ein besonderes Merkmal ist vor allem die Qualität der Verbindungen. Am Beispiel der Sprachübertragungen haben wir verschiedene Messmethoden und Möglichkeiten der Qualitätsbestimmung erläutert, sowie deren Kenngrössen eingeführt. Ausserdem haben wir gezeigt, wie Codecs für verschiedene Aufgaben, wie Videotelefonie, Sprachübertragung oder Streaming, eingesetzt werden. Da mobile Netzwerke, die in ihrer Eigenschaft die Luft als Übertragungsmedium nutzen, müssen Methoden der Fehlerkorrektur respektive der Kanalcodierung verwendet werden, um die fehlerfreie Übertragung sicherstellen zu können. Da verschiedene Anwendungen unterschiedliche Anforderungen auf einem gemeinsam genutzten Netz erfordern, haben wir die unterschiedlichen Methoden von Quality of Service dargestellt und deren Implementierungen in UMTS erläutert.

Literaturverzeichnis

[1] Mohamed A. El-Gendy, Abhijit Bose, Kang G. Shin: Evolution of the Internet QoS and Support for Soft Real-Time Applications, IEEE, July 2003.

[2] Gaglan M. Aras, James F. Kurose, Douglas S. Reeves, Henning Schulzrinne: Real-Time Communication in Packet-Switched Networks, IEEE, January 1994.

[3] Racha Ben Ali, Samuel Pierre, Yves Lemieux: UMTS-to-IP QoS Mapping for Voice and Video Telephony Services, IEEE, April 2005.

[4] NENA VoIP Technical Committee: VoIP Characteristics Technical Information Document, NENA, July 2004.

[5] INTEROP LABS: VoIP and 802.11, Interop Labs, May 2006.

[6] Matthew Gast: How Many Voice Callers Fit on the Head of an Access Point?, O'Reilly, `http://www.oreillynet.com/pub/a/etel/2005/12/13/how-many-voice-callers-fit-on-the-head-of-an-access-point.html`, December 2005.

[7] Undisclosed Author: VOIP vs PSTN, Lightreading Reports, `http://www.lightreading.com/document.asp?doc_id=53864`, June 2004.

[8] Network Strategy Partners LLC 2002, `http://www.nspllc.com`, 2002.

[9] Cisco Systems: Quality of Service for VoIP, Cisco, 2001.

[10] Wikipedia, Video, `http://en.wikipedia.org/wiki/Video`

[11] Tim Szigeti, Christina Hattingh: End-to-End QoS Network Design: Quality of Service in LANs, WANs, and VPNs, Cisco Press, Nov 2004.

[12] Dip.-Inf. J Richling: Weiche Echtzeit, Humboldt-Universität zu Berlin, Winter 2003.

[13] Speech Quality Index in CDMA2000, Technical Paper, Ericsson AB 2006, `http://www.ericsson.com/solutions/tems/library/tech_papers/tech_related/speech_quality_index_cdma_2000.pdf`

[14] Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s, ITU-T Recommendation, `http://www.itu.int/rec/T-REC-G.723.1/en`

[15] WorldDMB Press release 3rd of November 2006. `http://www.worlddab.org/upload/uploaddocs/WorldDMBPress20Release_November.pdf`

[16] Understanding Codecs: Complexity, Hardware Support, MOS, and Negotiation, Document ID 14069, http://www.cisco.com/warp/public/788/voip/codec_complexity.html#mos

[17] Atti, V.; Spanias, A., "Embedding Perceptual Metrics in Rate Control Algorithms," Intelligent Control, 2005. Proceedings of the 2005 IEEE International Symposium on, Mediterrean Conference on Control and Automation , vol., no.pp. 861- 865, 2005, http://ieeexplore.ieee.org/iel5/9900/31471/01467127.pdf?isnumber=31471&prod=STD&arnumber=1467127&arnumber=1467127&arSt=+861&ared=+865&arAuthor=Atti%2C+V.%3B+Spanias%2C+A.

[18] Marcus C. Gottwald, Ermittlung, Bewertung und Messung von Kenngrössen zur Bestimmung der Leistungsfähigkeit eines Mobilfunknnetzes in Bezug auf Voice over IP, Freie Universität Berlin, Institut für Informatik, Prof. Dr.-Ing. Jochen Schiller, in Zusammenarbeit mit Qosmotec Software Solutions GmbH Aachen, Juli 2006, S. 50-52

[19] White paper by Opticom GmbH, Germany, State of the art voice Quality testing, http://www.opticominstruments.com/docs/voice_quality_testing.pdf

[20] Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P., "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on , vol.2, no.pp.749-752 vol.2, 2001

[21] Werner, M.; Kamps, K.; Tuisel, U.; Beerends, J.G.; Vary, P., "Parameter-based speech quality measures for GSM," Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. 14th IEEE Proceedings on , vol.3, no.pp. 2611- 2615 vol.3, 7-10 Sept. 2003 14th IEEE Proceedings on, Publication Date: 7-10 Sept. 2003, Volume 3, on pages 2611-2615

[22] Andrew S. Tanenbaum: Computer Networks 4. Edition, Prentice Hall, 2003.

[23] Rudolf Riemer: umtslink.at, http://www.umtslink.at, 04.01.2007.

[24] Kriebel's Sat-Report: http://www.kriebel-sat.de/Technik_digital.htm, 04.01.2007.

[25] Alexander Braun, Markus Hofbauer: Semesterarbeit über digitales Satellitenfernsehen, http://people.ee.ethz.ch/~ambraun/sa1/vorwort.html, 04.01.2007.

[26] Harry Briem, Stefan Gelbke: Faltungscodes, Proseminar Kodierverfahren, Technische Universität Chemnitz, http://www.tu-chemnitz.de/informatik/ThIS/seminare/ss03/kv/Faltungscodes.pdf.

[27] Prof. Dr.-Ing. Jürgen Freudenberger: Skript zur Vorlesung Kommunikationstechnik WS2006/07.

[28] Dr.-Ing. Volker Kühn: Vorlesungsskript Kanalcodierung I WS 2006/07.

[29] Xilinx: Wireless FEC Solutions, http://www.xilinx.com/esp/wireless/collateral/Wireless_FEC_esp.pdf, 04.01.2007.

[30] Wikipedia: Interleaving, http://de.wikipedia.org/wiki/Interleaving, 04.01.2007.

[31] Wikipedia: Vorwärtsfehlerkorrektur, http://de.wikipedia.org/wiki/Vorwärtsfehlerkorrektur, 04.01.2007.

[32] Crosslink - The Aerospace Corporation magazine of advances in aerospace technology. The Aerospace Corporation (Volume 3, Number 1 (Winter 2001/2002)), http://www.aero.org/publications/crosslink/winter2002/04.html, 04.01.2007.

[33] David MacKay: Information Theory, Inference, and Learning Algorithms, http://www.inference.phy.cam.ac.uk/mackay/itila/book.html, 04.01.2007.

[34] Emilia Käsper: Turbo Codes, http://www.hut.fi/~pat/coding/essays/turbo.pdf, 04.01.2007.

[35] Wikipedia: Low-density parity-check code, http://en.wikipedia.org/wiki/Low-density_parity-check_code, 04.01.2007.

[36] Matthew C. Valenti and Shi Cheng and Rohit Iyer Seshadri: Turbo and LDPC Codes for Digital Video Broadcasting.

[37] Hans-Lackner: Next Generation Wireless, 61. K-Stammtisch, http://www.qoscom.de/data/stammtisch/61_NGW_QoSCom.pdf, 04.01.2007.

[38] Dr.-Ing. Volker Kühn: Vorlesungsskript Kanalcodierung II WS 2006/07.

[39] Q. Ni, L. Romdhani, and T. Turletti, A Survey of QoS Enhancements for IEEE 802.11 Wireless LAN, Wiley J. Wireless and Mobile Comp., vol. 4, no. 5, pp. 547-566, Aug. 2004.

[40] Sven Höhne: Dienstgütemechanismen für WLANs nach IEEE 802.11, TU-Braunschweig, http://www.ibr.cs.tu-bs.de/courses/ws0405/skm.

[41] IEEE Std 802.11e-2005, Part 11, Amendment 8.

[42] Juha Korhonen: HSDPA - An Introduction, A TTPCom White Paper, http://www.ttpcom.com/en/downloads/TTPCom_Whitepaper_HSDPA_Introduction.pdf.

[43] Weiwei Liang: Dienstgüte im Zugangsnetz von UMTS, TU-Braunschweig, http://www.ibr.cs.tu-bs.de/courses/ws0405/skm.

Chapter 10

Routing in Multi-hop Mesh Networks

Andreas Bossard, Daniel Dönni, Daniel Rickert

This report gives an overview of current academic and industrial research in the field of Wireless Multi-hop Mesh Networks (WMN). In particular, it focuses on routing protocols. After defining the term WMN, a short review of related work is presented. Since there is a huge number of protocols for WMNs, the report seeks to answer the question why there are so many of them. Various metrics to find the best path between two nodes in a WMN are analyzed in order to show how the performance of routing protocols is measured. The core of this work is made up by a presentation of the most important routing protocols and their classification. Each protocol is described and critically evaluated. The remaining chapters demonstrate what challenges WMNs are confronted with and show why it is difficult to predict their future.

Contents

10.1 Introduction

This paper gives an overview of routing in wireless multi-hop mesh networks. First the term is defined, then we analyze why there are so many protocols and finally we evaluate the most useful protocols.

10.1.1 Definition of a Multi-hop Mesh Network

There are several applications for a Multi-hop Mesh Network. In this paper we consider as the main purpose the wireless connection of clients with the Internet. A WMN consists of three layers (cf. figure 10.1):

- *Access points:* They provide the interface to the Internet for the whole network.
- *Wireless routers:* They form the wireless-backbone which *client nodes* can connect to and perform dedicated routing and configuration. Additionally, they are characterized by minimal mobility [1].
- *Client nodes:* These are the users who connect to the wireless backbone. That is possible directly, by connecting to a router, or indirectly through other client nodes.

For Multi-hop Mesh Networks we consider the hybrid architecture. This architecture is characterized by the fact, that mesh clients don't need a direct connection to a mesh router, but can connect multi-hop-wise over other mesh-clients to a router. The advantage of this architecture is improved connectivity and coverage. The disadvantage is that mesh clients need more ressources because they also need to have routing capabilities [1].

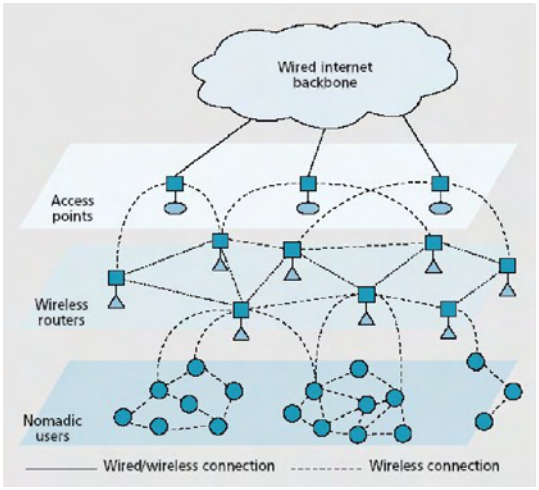


Figure 10.1: A three-layer-architecture for wireless mesh networks (by [2])

10.1.2 Differences between Wireless Mesh Networks and Ad-hoc Networks

The main differences are the following:

- Mesh networks are a flexible and low-cost extension of wired infrastructure networks compared to isolated and manually configured networks [2].
- In WMNs virtually all traffic originates from or is headed to a gateway while in ad-hoc networks the data flows between arbitrary pairs of nodes [3].
- The infrastructure of WMNs is static [4] because of the mesh-routers.
- The network architecture of WMNs is hierarchical (the three layers) in contrast to the flat one of ad-hoc networks [2].
- The mesh routers do most of the configuration and routing, thus they significantly decrease the load of mesh clients.

10.1.3 Advantages of WMNs

Compared to a normal WLAN-Hotspot-Network a WMN has several advantages [2].

- *Low installation costs* For supporting a large area with Internet-access by hotspots, for every hotspot a cabled connection is necessary. Cabling costs a lot. In the wireless mesh network, only a wireless router needs to be added to enlarge the reach of the network.
- *Fast deployment* It takes longer to install a cabled connection, than to install a wireless router.
- *Reliability* There are redundant paths in the wireless backbone between endpoints, so the breakdown of a router should be no problem and there should be no bottleneck-routers.
- *Self-management* The whole network has the advantages of ad-hoc networks: self-configuration, self-organization and self-healing capabilities. For users, the network-setup should be automatic and transparent. For example, when a router is added to the network, it should detect all wireless routers and the optimal path to the wired network. On the other hand, the existing wireless routers should take into account the new routes via the new node.

10.2 Related Work

There are some practical examples of productive WMNs:

- *Strix Systems* Strix Systems [5] has implemented several Wireless Mesh Networks. They are to deploy a country-wide WMN in Macedonia, which means it will be more than 1000 squaremiles big. Furthermore they implemented already a city-wide network in Tempe, Arizona, which is about 40 squaremiles big. Local firefighters and policemen can now get up-to-date information via that network.
- *Wireless St. Gallen* In Switzerland, the town of St. Gallen has initialized a pilot project with three access points and 20 mesh routers to start a town-wide wireless mesh network [6]. The pilot is planned from December 2006 till February 2007. The project hopes for the active help of the citizens to buy routers and install a special firmware. The chosen protocol for the network is OLSR.

Because there is no central coordination entity in the mesh-router-layer WMNs exhibit some similarity to MANETs, Sensor Networks, and Mobile Peer-to-Peer Networks.

MANET is an acronym for Mobile Ad-hoc Network. The term describes a network which is spontaneously established by a number of mobile nodes which happen to be at the same location at the same time. In contrast to a WMN a MANET does not focus on establishing a connection to some infrastructure network but to provide immediate, uncomplicated and reliable connectivity among the participating nodes [7]. These characteristics make MANETs suitable for disaster operations or communication in backcountry [7]. Interested readers are referred to chapter 1.

Sensor networks are networks consisting of a large number of tiny nodes providing only rudimentary functionality. Application scenarios for such networks comprise among others surveillance systems [7] or networks to collect environmental data. Further information can be found in chapter 4.

File-Sharing applications are the most popular ones as far as Peer-to-Peer networks are concerned. The goal of *Mobile Peer-to-Peer Networks* is to run similar and new applications on mobile nodes. This is an interesting and challenging task. A short introduction to Mobile Peer-to-Peer Systems can be found in [8, 9, 10].

10.3 Fundamentals of Routing Protocols

This chapter seeks to answer the question why there are currently so many protocols. Moreover, it gives a short overview of the fundamental paradigms of routing protocols and the underlying mechanisms. This includes the way routing information is retrieved and how it is stored. Common metrics and their suitability to measure various quality aspects concerning routing protocols round off the chapter.

10.3.1 Protocol Variety

Kowalik and Davis [11] estimate that there are currently more than a hundred different routing protocols for WMNs. In their paper [11] they try to answer the question why there are so many protocols and they point out four major reasons that led to current situation. The following argumentation is based on their paper.

1. Lack of Standards The IEEE created a working group named 802.11s in order to standardize WMN protocols. However, such a standard is not expected to be ready before 2008. Therefore, proprietary protocols are springing up like mushrooms, they are usually not compatible with each other and often require users to acquire proprietary hardware, too. This in turn makes customers hesitate to invest in suitable equipment.

2. Range of Application Scenarios There is a tremendous amount of application scenarios for WMNs, such as free public Internet access, wireless VoIP services, highly reliable and fault-tolerant LANs, military communication, or wireless mesh gaming. Their requirements concerning scalability, security, quality-of-service, power efficiency, and other network properties are manyfold which makes it difficult or even impossible to develop one single protocol that meets all demands.

3. Different Design Methods There are always several approaches to solve a problem but some are more suitable than others. Out of this reason, protocols are designed for a specific application. This results in a great variety protocols. Despite this variety, they mainly differ in the strategy they use to find routing information, the way they store this routing information, and the performance metrics that were used to design them.

4. The Bandwidth Problem Although many ideas have been proposed to overcome the bandwidth limitations in wireless networks, the bandwidth problem remains an awkward one. The more stations coexist in the same network, the more they interfere with each other. WMNs are particularly susceptible to suffer from high latency due to the fact that packets are generally routed across several wireless links.

10.3.2 Retrieving Routing Information

There are basically three methods to design algorithms which are supposed to retrieve routing information from the network: The proactive, the reactive, and the mixed design method.

- *Proactive Design Method:* Protocols following the proactive design method actively assemble routing information no matter when or even whether it is used.

- *Reactive Design Method:* Protocols following the reactive design method only collect routing information when it is explicitly requested.

- *Hybrid Design Method:* Protocols following the hybrid design method combine the two methods above.

10.3.3 Storing Routing Information

There are two paradigms to store routing information. Either a link state database or a distance vector is used. The two paradigms can be characterized as follows.

- *Link State Database:* Protocols using a link state database try to map the network or a part of the network onto a database that contains the current state of every link. The state is either *working* or *broken*.

- *Distance Vector:* Protocols using the Distance Vector approach maintain a vector which contains the costs to access other nodes. The vector may not cover the whole network.

10.3.4 Protocol Metrics

The five most common metrics to measure routing performance are: Hop Count, Round Trip Time, Expected Transmission Count, Per-hop Packet Pair Delay, and Weighted Cumulative Expected Transmission Time.

10.3.4.1 Hop Count (HOP)

The Hop Count (HOP) amounts to the number of links a packet passes until it reaches its destination [12, 13].

Advantages

- Implementing the HOP metric is simple.

Disadvantages

- The metric does not take the quality of the path into consideration.

10.3.4.2 Round Trip Time (RTT)

The Round Trip Time (RTT) measures the time it takes for a packet to travel to its neighbor-node and back to the originating host, including the time the destination host needs to process the request [12, 14].

Advantages

- The RTT is an adequate way to measure network latency.

Disadvantages

- The metric includes the time the destination node needs to process the request. However, only the delay between the two nodes is of interest.
- The metric did the poorest job of selecting paths in the experiment of Draves et al [13].

10.3.4.3 Per-hop Packet Pair Delay (PktPair)

To calculate the Per-hop Packet Pair Delay (PktPair) a host sends every other second two packets to its neighbors. A neighbor answers by returning the difference between the arrival times of the two packets. The average of all of these differences of one particular neighbor – which were previously exponentially weighted – constitutes the value which is defined as PktPair for this particular neighbor [12, 13].

Advantages

- PktPair is not affected by queuing delays at the sending node.

Disadvantages

- It is not better at selecting paths than HOP [13].
- The overhead is bigger than the one of RTT because two packets are sent and the second packet is bigger than the first one [13].

10.3.4.4 Expected Transmission Count (ETX)

ETX finds high-throughput paths on multi-hop wireless networks by measuring the loss rate of broadcast packets between pairs of neighboring nodes [13]. ETX improves the throughput of multi-hop routes by up a factor of two over a minimum hop-count metric [15].

An exact definition of Expected Transmission Count (ETX) can be found in [15]: "The ETX of a link is the predicted number of data transmissions required to send a packet over that link, including retransmissions. The ETX of a route is the sum of the ETX for each link in the route.". Other definitions are provided by [12, 13].

Advantages

- It is better at finding multi-hop paths than HOP, RTT or PktPair [13].
- Compared to HOP, ETX selects paths having a more stable throughput [13].

Disadvantages

- The broadcast packets are small and sent at the lowest possible data rate, so they may not experience the same loss rate as data packets sent at higher rates [13].
- The metric does not directly account for link load or data rate [13].

10.3.4.5 Weighted Cumulative Expected Transmission Time (WCETT)

All the information in this section is taken from [16]. WCETT is a metric for routing in multi-radio multi-hop wireless networks. In multi-radio networks each node has multiple radios. For example a node has a 802.11a and a 802.11b radio, or each node has multiple radios that send on different channels.

To find the fastest path it takes into consideration the loss rate and the bandwith of each link, and the interference between links due to the use of the same channel. WCETT builds upon the ETX metric and extends it by adding considerations about bandwidth and channel-diversity of the links.

HOP and ETX don't perform well in multi-radio networks. Consider nodes with two radios: one with 802.11a and one with 802.11b. 802.11b has mostly a longer range and therefore most of the traffic will be carried over the 802.11b links in shortest-path and ETX. Moreover shortest-path and ETX don't take in consideration, that consecutive hops on different channels result in a higher throughput.

Taking channel diversity into account is only useful on short paths or heavy-loaded networks but not on longer paths.

Advantages

- WCETT takes bandwidth and channel-diversity into account.

Disadvantages

- It is the most complicated metric among the presented ones.

10.4 Routing Protocols in Wireless Mesh Networks

As it was shown in the introduction, a WMN consists of traditional ad-hoc networks which are interconnected by access points and wireless mesh routers. Ad-hoc networks consist of lots of nomadic nodes, the participant users with their mobile devices like notebooks. The communication range of these devices is very limited because of the fact that they act with limited energy ressources like batteries or accumulators. So, only a small number of devices is directly reachable (neighborhood) while other devices are out of range.

Because of this situation, wireless mesh routers are responsible for performing routing. Routing denotes the process of moving information, which starts at a source and travels across several networks to a destination while routers decide about the best (shortest, cheapest) path [17]. The most important requirements of routing in WMNs are minimal control overhead, no loops, multi-hop routing, and the management of an extensible dynamic network topology. For performing routing, the forwarding packets are dependent on special routing protocols.

This chapter describes several state-of-the-art routing protocols developed for WMNs. As explained in the previous chapter, there is an overwhelming amount of protocols – each of them mostly developed for a particular task depending on the purpose of communication and the scenario – which makes it impossible to address each and everyone separately. The notion of this chapter is to get the reader familiar with the most important routing protocols as well as their classification schemes in order to enable him to know where to place other protocols that can be found in literature. The most important protocols belong either to the proactive, the reactive or the hybrid protocol family.

10.4.1 Proactive Routing Protocols

As mentioned earlier, routing protocols following the proactive design approach actively assemble information no matter when or even whether it is used. That means that the possible paths for packets are precalculated and stored in routing tables. Proactive routing protocols are built up by conventional Link State or Distance Vector Protocols. Each node manages the route to each other node which allows a constant control overhead. In comparison with an reactive routing protocol, the overhead in proactive routing protocols is mostly higher. There exists a periodically and event-driven exchange of all the routing information between the nodes which means that proactive routing protocols actively pick up and update information about possible routes or link states [7]. Each time a node receives a data packet, the node's task is to route this packet on the cheapest way forward

to the next hop by making use of the underlying routing table. These routing tables are dependent on the used routing processes that we will describe with two proactive routing protocol examples in the sections 10.4.1.1 DSDV and 10.4.1.2 OLSR.

10.4.1.1 DSDV

DSDV is the acronym of Destination-Sequenced Distance Vector, a table-driven, proactive routing protocol for data transmission between different stations in ad-hoc networks. It is based on the Distributed Bellman-Ford algorithm. For each node in a Distance Vector Protocol is determined over which hop it is reachable and what is the total distance to the destination. Therefore, not the complete way to the destination is known, but only the next step. In detail, routing packets across the network need the use of existing, precalculated routing tables at each node of the network which contain available destinations and the number of hops to reach them. The packets will be routed from one hop to the next hop with the best economic aspects with regard to cheapest costs and shortest path. As a conclusion of this behavior, a packet's best path to its destination will be identified during it's travel across the network.

A WMN is dynamically formed, so the arriving of new stations or the leaving of them is very simple. Therefore, each station periodically broadcasts its actual data (distance vectors) to enable its neighbors to adjust their own routing tables. For example, a station leaves the network and a neighbor hop recognizes this, than the neighbor hop sends new information about the distance to all of its neighbor hops, so that they can recalculate their routing tables.

One of the most important problems of Distance Vector Routing is the Count-to-Infinity problem. This means the slowly counting up of costs which occurs by building loops between neighbors if a local link gets broken. For recognizing the actuality of a message, the DSDV contains a sequence number in addition to the hop information. The sequence number is assigned by the station to which the distance has been measured. It will be increased each time update information is broadcasted. This leads us to the goal of this approach: In contrast to traditional Distance Vector Routing, DSDV helps to avoid the looping problem what means that a packet should always reach its destination, and therefore it may not get stuck in a loop between two or more stations because of the sequence numbers. Each station only updates its routing table if the sequence number has been increased or the distance has been decreased. The route with the most actual sequence number has to be chosen, with equal sequence numbers, the shortest route is the best choice.

Advantages

- Very simple and fast realizable topological and routing updates because of the integration of new stations or the leaving of stations.
- Avoidance of the loop problem of Distance Vector Routing.
- No time delay because of the precalculated pathes (distance vectors).

Disadvantages

- Periodic updates of all routing information even if there is no topological changing lead to unnecessary network traffic and high control overhead.
- A large number of participants leads to a fast swelling of the routing tables, but only a few information is needed for routing.

Evaluation Ad-hoc routing protocols can be used for the communication between traffic participants to improve the security of the road traffic. This makes it possible to earlier recognize dangerous situations, speed traps, accidents, and congestion which leads to more comfortable driving. DSDV is a routing protocol to send packets in mobile ad-hoc networks. The elaboration [18] shows that the use of DSDV is possible but limited because there exist a lot of problems to guarantee error-free operation.

10.4.1.2 OLSR

OLSR is the acronym for Optimized Link State Routing Protocol, a table-driven, proactive routing protocol for exchanging topology information between nodes. It is an optimized further development of the traditional Link State Routing Protocol (stability), especially for WMNs, and it takes advantage of the immediate availability of routes because of its proactive nature. Link State means that each node knows the whole network topology to determine the shortest path from a source to a destination. The goal of OLSR is to keep the control messages as small-sized as possible, and to minimize broadcasting.

In [19] a very good explanation on how OLSR works can be found. A short summary will be shown next by using figure 10.2.

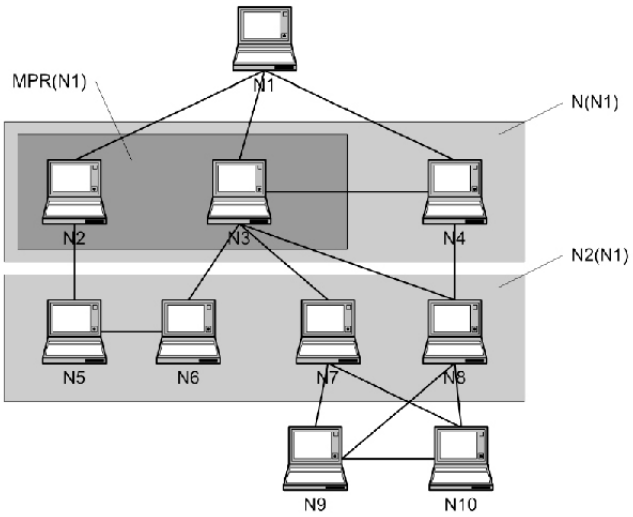


Figure 10.2: OLSR with the amounts N, N2, and MPR [19]

OLSR introduces the idea of multipoint relays (MPRs). These are selected nodes which are responsible for forwarding control messages [19, 20].

Figure 10.2 shows a network with ten nodes, N1 to N10. N(N1) is the amount of the direct neighbor of node N1, N2(N1) is the amount of all nodes which distance to N1 is exactly two hops, and MPR(N1) is the amount of the nodes which are selected as MPRs by N1.

The HELLO messages contain information about the amount of neighbors of a node, and its actual MPRs. In this way, each node is able to detect all nodes with a two hops distance. As shown in figure 10.2, node N1 can detect N2(N1) by using the MPRs N2 and N3. If a node detects a topological change in N(N1) or N2(N1), MPR(N1) has to be refreshed.

Recapitulating can be said, that each node selects an amount of MPRs in that way, that it can reach all nodes with a two hop distance to it over these MPRs. One of the goals of this approach is to reduce the control traffic, and the overhead of flooding messages by reducing redundant retransmissions in the same region [20] by using selected MPRs.

Advantages

- Very short latency because of the fact that routes are precalculated and stored in routing tables.
- The control messages can be kept small-sized by using multipoint relays.
- Broadcasting is minimized because each node selects a set of multipoint relays in a manner that it can reach all nodes with a two hops distance by making use of these multipoint relays.

Disadvantages

- Many redundant routes because of the precalculation of the routes to all other nodes.

Evaluation

- Because of the missing auto assignment of IP addresses, they must be assigned by a central entity to avoid double assigned IP addresses. One of the consequences is the registered access to the Internet [21].
- An end-to-end encryption would avoid the “big brother”-problem what means that without encryption each and every node can see what kind of data other nodes are forwarding [21].

A real-world example Openwireless Schweiz [22] is a Swiss community whose vision is the distribution of free OLSR networks across Switzerland to intensify neighborhood's communication. The goal is to build up free and independent meshed networks of whole districts, i.e. for enabling license free community radio, transmission of local events per stream, VoIP, shared Internet access, and neighborhood WLAN network for gaming, communicating, and data exchange.

Two of the reasons to implement such a OLSR network is the need to get a cheap Internet access, and the possibility of connecting people in free networks without any restriction by commercial providers. Current situation is that the community is growing up, so more and more people provide internet access to expand this network [22].

10.4.2 Reactive Routing Protocols

Routing protocols following the reactive design approach only collect routing information when it is explicitly requested (route discovery), as mentioned earlier. This means that there is no precalculated routing information about shortest and cheapest routes to destinations stored in any routing tables. This results in time delay at the beginning of a request, because the route has to be determined first. On the other hand, reactive routing protocols are bandwidth efficient because of the mentioned on-demand requests what limits the traffic of messages. So, the control overhead of reactive routing protocols is dependent on the number of connections that must be built up. Because of the non-existing precalculated routing tables, the main functions of this kind of protocols are route discovery and route maintenance. On-demand route discovery is looking for the most efficient route across the network with regard to cheapest costs and shortest path. It only maintains active routes. Route maintenance is responsible for keeping the paths across the network alive: it searches for damaged routes and repairs them. Two of these reactive protocols will be described in the chapters 10.4.2.1 AODV and 10.4.2.2 DSR.

10.4.2.1 AODV

AODV is the abbreviation of Ad-hoc On-Demand Distance Vector, an demand-driven, reactive routing protocol which is part of distance vector protocols. The goal of AODV is to avoid system wide broadcasts. An in-depth description of the AODV protocol can be found in an IETF protocol draft [23]. Here, a short summary about the functional aspect is given, it closely follows the very good explanation in [19].

The route discovery process in AODV functions as follows: The sending node X floods a route request (containing the IP address of X, the actual own sequence number, the IP address of the destination, and the newest known sequence number to this node) to all its neighbors. Additionally, the route request gets a broadcast ID. So, the route request is unique identifiable by the IP address of the source and the broadcast ID.

The neighbors receive this route request and check whether they already received a message with identical broadcast ID and IP source address. If so, they ignore the message

in order to avoid loops. Otherwise, they make themselves a reverse route entry which builds up a reverse path to the sender of the route request. This procedure guarantees that the node can forward a possible answer to the route request to the sender. Then the neighbors send the route request again to their neighbors (again by making a reverse route entry to the source) and so on until the destination node is reached or a node with a valid route is found which sequence number is newer than the sequence number in the route request.

The route request contains a time to live (TTL) which will be reduced each time the route request is forwarded one hop to avoid system wide broadcasts.

The answer to a route request is a route reply (containing the sender of the route request, the destination address, the sequence number of the destination, and the TTL). The route reply will be sent along the memorized route (reverse path) back to the source node, while each node between destination and source makes a forward route entry to the destination (forward path: memorizes the path to the destination).

The source node X receives the route reply, makes itself a forward route entry to the destination and sends its data packets via the found route to it. During the whole procedure, the so called route maintenance is responsible for holding up the connections on the found path with the help of hello messages. If a connection breakes, an update of the routing tables will be sent, and if needed, another route discovery will be executed. AODV is based on DSDV. We mentioned the problem in DSDV as the periodically updates of all routing information even if there is no topological changing lead to unnecessary network traffic and high control overhead. In AODV, the drop out of nodes or connections leads to more efficiency (less flooding) and less network traffic. The problem of loopings is solved by sequence numbers because each route request can be identified by the IP address of the starting node and the sequence number of the request. The details can be found in the draft of IETF as described above.

Advantages

- AODV is bandwidth efficient because routes are not precalculated.
- Avoidance of the looping problem by using sequence numbers.
- In comparison with a proactive routing protocol, the overhead in reactive routing protocols is lower because of control functions.
- Better scalability than in proactive routing protocols.

Disadvantages

- Time delay because of non-existing precalculated routing tables (route discovery).
- AODV requires bidirectional links for making route replies.

Application Area AODV is suited well for a network topology which is very dynamically because it recognizes updates very fast and with less additional network load.

10.4.2.2 DSR

In contrast to proactive protocols, reactive protocols do not actively gather and update information about possible routes or link states, unless they are explicitly requested to do so [7]. The goal of this approach is to reduce unnecessary network traffic and the interrelated waste of energy in mobile devices.

An in-depth description of the DSR protocol can be found in an IETF protocol draft [24]. Schiller provides a good overview of the Dynamic Source Routing (DSR) protocol in [7] which explains the algorithm in sufficient detail for our purposes: If a host wants to send data it sends a packet that contains a unique identifier and the destination address to each of its neighbors. All neighbors will react in one of the following ways:

- If the neighbor recognizes that it already received a packet having the same identifier it discards the packet in order to avoid loops.
- If the neighbor recognizes that the destination address matches its own address it knows that the source host would like to send some data to it.
- If neither of the above apply, the host appends its own address to the list of hosts the packet already passed and forwards it to its neighbors.

Given that the topology and link state remain unchanged and a path to the destination actually exists, the algorithm described above will find the path. It is even possible that several (partially) different pathes are found. Nevertheless, because the topology does change time and again, and thus the link state, too, there is no guarantee whatsoever that another packet sent along a route that was discovered this way will still successfully reach its destination [7].

Furthermore, links in wireless networks need not be bidirectional. If they are, the destination host can send a response to the source host by sending a packet simply along the same path by traversing the list of hosts in reverse order. Otherwise, it must apply the same algorithm itself to find a path to the source host and use this one [7].

Advantages

- Less energy is used because routing information is only requested on demand.

Disadvantages

- The algorithm suffers from high latency because every route needs to be established first.

10.4.3 Hybrid Routing Protocols

The discussion of the two previous protocols clearly showed that both proactive and reactive protocols have advantages and drawbacks. One of the most frequently used approaches in science is to combine existing solutions in order to exploit either vantages. This is exactly what hybrid protocols do: They combine the favorable features of proactive protocols with the ones of reactive protocols. We will describe one hybrid routing protocol example in the sections 10.4.3.1 ZRP.

10.4.3.1 ZRP

The Zone Routing Protocol (ZRP) is a typical example of a protocol that combines proactive and reactive design principles. Its working mechanism is explained according the Internet Draft that can be found at Cornell University [25].

The idea of ZRP derives from the observation that proactive protocols flood the network with information about any topology change, even if only local nodes are affected. Reactive protocols only get routing information on request, however, they are often unable to establish connections within narrow time-frames which is especially adverse for real-time applications. Thus, ZRP segments a network into many overlapping parts, termed zones. Routing within such a zone is carried out by a proactive protocol whereas routing between zones is conducted by a reactive protocol [25].

The size denotes the number of nodes that can be reached within a specified number of hops, termed zone radius. The zone radius can be set before the network is created but any host can alter its own zone radius individually at run-time. Consequently, the size of a zone is variable and can be dynamically adjusted to the network characteristics. A proactive protocol, called IARP (IntrAzone Routing Protocol) [26], is deployed to make sure that routing information within a zone is always kept up to date [25].

If the destination node is located outside the zone, the source host sends a route query packet to a selected number of hosts within its zone. If the destination node is located inside their zone or they already have a cache entry pointing to it, they send a route reply message to the source host whereupon a connection can be established. Otherwise, they forward the route query packet to some of their neighbors. This discovery process is conducted by IERP (IntErzone Routing Protocol) [27] which makes use of the BGR (Bordercast Resolution Protocol) [28] to determine the subset of nodes that the route query packet is forwarded to [25].

Advantages

- ZRP combines the advantages of proactive and reactive routing protocols.
- ZRP increases the scalability of the system.
- ZRP is a modular protocol which allows IARP and IERP to be replaced by suitable other protocols [25].

Disadvantages

- Although there is an IETF protocol draft, ZRP was never standardized.
- Combining pro- and reactive protocol elements results in increased complexity.
- Scalability is still limited.

Evaluation ZRP serves as a good example to describe characteristics of a hybrid routing protocol. It addresses both scalability and performance issues and can be adapted to the requirements of the system and its users. However, an increased protocol complexity requires more powerful hardware.

Furthermore, there are still scalability problems. For example, if the size of a system using ZRP is large with respect to the average zone radius, the system will be almost as slow as a purely reactive system. While increasing the zone radius will mitigate the problem, too large a radius will lead to excessive traffic within the zone and result in poor performance as well.

10.4.4 Other Protocols

10.4.4.1 Hierarchical Routing Protocol

A major disadvantage of AODV and DSR is that they do not scale [7]. Hierarchical routing protocols try to overcome this disadvantage by building node clusters. They exploit the fact that under normal conditions only a few nodes join or leave the cluster they currently belong to. Therefore, it is neither necessary nor efficient to propagate such minor topology variations throughout the whole network, it is perfectly acceptable to inform only the cluster members. As long as a cluster can be reached, the nodes within this cluster can be reached, too [29, 7].

One node within such a cluster is chosen as clusterhead [29]. A clusterhead acts as gateway and router for all cluster members. At the same time it constitutes the interface to other clusters and their nodes. An aggregation of two or more clusters make up a super-cluster. Super-clusters can be combined likewise. In this manner, a scalable, hierarchical network is formed [7].

Advantages

- In small or heterogeneous networks the hierarchical routing protocol might be more efficient than AODV and DSR.

Disadvantages

- The load on the clusterhead is significantly higher than the one on the other nodes within the same cluster.
- The load on the clusterhead in a super-cluster is even higher.
- Clusterheads are likely to become resource bottlenecks.

Evaluation While the hierarchical routing protocol might be useful in wired networks to build a scalable network it is quite unlikely that this is applies to WMNs, too. The reason is that the clusterhead establishes connections to other clusterheads but is punished for it. There is no incentive whatsoever for a node to become clusterhead, it is even likely that nodes would actively try to avoid to be selected for this job.

Even in the unlikely case that nodes would accept and carry out the job altruistically the approach would not scale. This is due to the fact that the resources of mobile devices are limited and it is improbable to find any nodes which could sustain to load of "backbone routers".

The only way this approach might work is if there is a certain amount of much more powerful nodes in the network which are capable of dealing with these issues. In a network of PDAs or cell phones several laptops would be required. However, it cannot be expected to this is always the case in a WMN.

10.5 Challenges

10.5.1 Scalability

Scalability is probably the most critical issue in WMNs [1]. It was shown that increasing the size of a WMN sharply diminishes its performance (cf. section 10.5.2) [30, 31]. One reason is that well-known medium access schemes such as CDMA, TDMA and CSMA/CA cannot simply be ported to WMNs:

- WMNs are decentralized systems. Thus, it is much more difficult to realize code management in CDMA [1].
- TDMA is not unproblematic either because time synchronization in large systems is hard to achieve [30].
- CSMA/CA does not have the above problems but its frequency spatial-reuse efficiency is poor [32].

Obviously, the mentioned medium access schemes have their limitations. Therefore, Akyildiz et al. [1] suggest to create new schemes by either merging CSMA/CA with TDMA or CDMA respectively.

10.5.2 Performance

Performance is a critical factor for any novel technology. If the performance cannot compete with related technologies it will usually not be successful, unless it makes specific application scenarios possible that could not have been realized with existing products. However, dealing with performance issues in WMNs is difficult. One reason is that many attempts to increase performance result in negative feedbacks (NFBs). A typical example is the one described by Iannone et al. [33] "[On] the one hand, adding power increases the rate and reliability of transmission, and creates longer hops; on the other hand, more power means more interference, reducing the global throughput of the network". Since there are many NFBs in WMNs the solution space is reduced significantly.

It is possible to provide some numbers about WMN performance but due to the vast amount of different technologies and research approaches they cannot easily be compared. The lack of reference values which could help to put the measured values into context makes it even more difficult to draw the right conclusions. As long as a widely accepted standard does not exist (cf. section 10.5.4) or more WMNs are deployed, the present situation is unlikely to change. Nevertheless, we try to show some representative data.

Current vendors indicate very different numbers concerning their system's performance. In fact, most of them do not provide any at all. Some exceptions are Strix Systems [5] which promises a throughput of 108 MBit/s [34] and Firetide [35] which claims to achieve 25MBit/s [36, 37, 38, 39]. However, neither of them provides any useful details about the conditions these values were measured under and how the performance decreases if these conditions are not fulfilled any longer. Since either company sells proprietary products it is more difficult to measure or verify the indicated numbers.

The numbers indicated by academic research do not look a quarter as good as the ones indicated by the industry. Measurements conducted in Roofnet [40, 41] in April 2004 showed that in a network of 18 nodes an average data rate of 2857.6 KBit/s and latency of 9.7 ms can be achieved given that no more than 1 hop must be carried out [1]. However, the average data rate drops to 378.4 KBit/s while the latency increases to 43 ms if 4 hops in a network of only 7 nodes must be completed [1]. This impressively demonstrates the performance and scalability behavior of WMNs. Obviously, in spite of the considerable research effort, WMNs still cannot nearly compete with traditional WLAN and considerable efforts are necessary to achieve a performance level that make it a real alternative to IEEE 802.11a/b/g.

10.5.3 Power Efficiency

Power efficiency is a hot topic whenever it comes to dealing with wireless devices. The basic problem is that the energy per volume fraction of current energy sources lies below the desired threshold. Research is conducted in the development of efficient power saving schemes that keep mobile devices running as long as possible as well as in the exploration of alternative energy sources. However, despite long lasting searches for new energy sources, there have not been many noteworthy results. In contrast, the development of energy preserving mechanisms proved to be far more fruitful.

The development of energy preserving mechanisms has been quite successful, nevertheless, solving the power problem remains a complex task. Most power saving mechanisms suffer from NFBs. Additional mechanisms attempting to reduce power consumption will consume power themselves and might consume the energy they just generated themselves.

10.5.3.1 Affected Devices

The power efficiency requirements in WMNs depend on the node's role in the WMN (cf. figure 10.1). Access points are not commonly mobile and connected to the power plugs and the Internet by wires [1]. Wireless routers are also primarily immobile but in contrast to access points they are not directly connected to the Internet but only forward packets to nomadic nodes and access points. Since they carry quite a large amount of traffic it is best to power them by wires. Even though mobility is less of a concern, they may also be equipped with batteries. These can be rather bulky compared to the ones which are used in nomadic devices. As a consequence, it is first of all the nomadic devices which require energy preserving mechanisms.

Still, it important to note that the energy problem does not solely concern nomadic devices. This is due to the fact that many technical features will only work properly if they are implemented in routers and access points, too. This means that eventually *all* devices in WMNs are affected by the power problem and therefore required to support such mechanisms, even if they gain little or nothing at all themselves.

10.5.3.2 Importance of Power Saving Mechanisms

A first aspect is convenience. A major advantage of wireless over wired devices is enhanced convenience. Though, the battery of any mobile device must be recharged after a while, rendering it dependent on external power sources. The longer a device runs without external power sources, the less dependent and the more convenient it is to use.

Another aspect is the fact that performance and scalability critically depend on the amount of available ressources. The need for performance (cf. section 10.5.1) and scalability (cf. section 10.5.2) makes power saving a fundamental requirement for any wireless device.

10.5.4 Standardization

10.5.4.1 Standards: Failure and Success

Standards are developed in collaboration with many industry partners, all of which made considerable research investments and are therefore primarily interested that their technology becomes the new standard. Due to the individual interests, defining standards is usually a tedious and long-lasting task. But in many cases, standards are a requirement

for a technology to become successful. This is because they prevent customers and companies to a certain degree from investing into technologies and equipment that might soon vanish from the market because a standard is an incentive for other companies to invest into the standard, too.

However, it is also possible that a standard fails. This might happen if the standard comes out too late and proprietary technologies have already become a de-facto standard. Another reason is that a standard contains too many compromises and therefore is, from a technological point of view, considerably worse than existing, proprietary solutions. Finally, a standard may also fail if the manufacturers simply do not care about it but add incompatible extensions or prefer to sell their own solutions in order to strengthen customer retention.

10.5.4.2 Trends

Despite the fact that there have been so many different ideas during the last couple of years, a standard is in the offing that many important companies are committed to. The ESS Mesh Networking Task Group [42] reduced the the original 35 parties which made a proposal for 802.11s to a single one that consists of the former SEEMesh[1] and the Wi-Mesh Alliance[2] [46, 47]. Still, the standard is not expected to be ready before 2008 [47].

Besides the companies working towards a common standard, there are also many that do not cooperate. First of all the ones that already have successfully been selling products for a couple of years prefer not to take part in the standardization process [47]: Among them are Strix Systems [5], BelAir Networks [48] and RoamAD [49]. It can be expected that they will defend, possibly even extend their current influence on the market.

10.5.4.3 The Future of WMNs

Many companies are willing to cooperate. One reason might be that large-scale WMNs covering hundreds or even thousands of km^2 are possibly rather the result of several small WMNs growing into medium and eventually large ones. The lack of standards would cause such processes to fail miserably. On the other hand, according to Wi-Fi Planet [50], Strix Systems [5] started in 1995 with the deployment of a nationwide WMN in Macedonia in close cooperation with On.net [51]. The WMN is supposed to cover an area of over 1000 mi^2 and to provide data, voice and video services to more than 2 million people.

Based on current academic and industrial research it impossible to predict the future of WMNs, there are simply too many unknowns. Only time will tell whether the technology will be successful.

[1] Some important members include: Intel, Nokia, Motorola, Firetide as well as NTT DoCoMo (June 20, 2005) [43, 44].

[2] The alliance currently consists of Nortel, InterDigital, MITRE, Philips, Accton, Thomson, ComNets, nexthop, and Swisscom (February 8, 2007) [45].

10.6 Summary and Conclusion

In the previous sections, the advantages and drawbacks of WMNs compared to ad-hoc networks and traditional hotspot-architectures were shown. The primary advantages are low installation cost, fast deployment, reliability, and smart self-configuration mechanisms [2].

There is a great variety of different protocols. Reasons which lead to the current situation comprise the lack of standards, the vast range of application scenarios, different design methods as well as a rich set of attempts to increase the poor throughput [11].

The most important routing protocols and their classification were presented. Proactive routing protocols actively gather routing information. They are characterized by low latency and high performance at the cost of a high amount of routing overhead. Reactive protocols only look for routing information when they are explicitly requested to so. The routing overhead is reduced significantly, however, the network latency is comparably high due to the fact that a route must first be built in order to establish a connection. Hybrid routing protocols try to combine the advantages of the previous approaches at the cost of increased protocol complexity.

There are still many problems which must be solved in order for WMNs to be successful. Scalability and performance issues are the thoughest challenges but the power problem which occurs in any wireless networks is not solved yet either. A large amount of renowned companies participate in standardization activities. The process is on the right track companies are obviously quite optimistic. Reliable data is rare and difficult to evaluate because it cannot easily be put into context. It is remarkable that the numbers indicated by the industry are significantly higher than the ones measures by academic research institutions. Whatever the reason might be, we have not been able to figure it out.

Certainly, there are many potential application scenarios for WMNs. The technology is not market ready though. Another unknown is the impact of related technologies such as WiMAX or high-speed private area networks. Thus, the future of WMNs remains unclear.

Bibliography

[1] Ian F. Akyildiz, Xudong Wang, and Weilin Wang. Wireless mesh networks: a survey. *Computer Networks*, 47(4):445–487, March 2005.

[2] R. Bruno, M. Conti, and E. Gregori. Mesh networks: commodity multihop ad hoc networks. *Communications Magazine, IEEE*, 43(3):123–131, 2005.

[3] Jangeun Jun and M. L. Sichitiu. The nominal capacity of wireless mesh networks. *Wireless Communications, IEEE [see also IEEE Personal Communications]*, 10(5):8–14, 2003.

[4] Thomas Scherer and Thomas Engel. Bandwidth overhead in wifi mesh networks for providing fair internet access. In *PM2HW2N '06: Proceedings of the ACM international workshop on Performance monitoring, measurement, and evaluation of heterogeneous wireless and wired networks*, pages 40–47, New York, NY, USA, 2006. ACM Press.

[5] Strix Systems. Access/One® Network OWS. `http://www.strixsystems.com`, Accessed: 01-10-2007.

[6] Openwireless St. Gallen. `http://sg.openwireless.ch`, Accessed: 01-21-2007.

[7] Jochen Schiller. *Mobile Communications*. Addison-Wesley, 2nd edition, August 2003.

[8] Wolfgang Kellerer, Rüdiger Schollmeier, and Klaus Wehrle. Peer-to-peer in mobile environments. In Steinmetz and Wehrle [52], pages 401–417.

[9] Andreas Heinemann and Max Mühlhäuser. Spontaneous collaboration in mobile peer-to-peer networks. In Steinmetz and Wehrle [52], pages 419–433.

[10] Christoph Lindemann and Oliver P. Waldhorst. Epidemic data dissemination for mobile peer-to-peer lookup services. In Steinmetz and Wehrle [52], pages 435–455.

[11] Karol Kowalik, Mark Davis. Why Are There So Many Routing Protocols for Wireless Mesh Networks? `http://www.cnri.dit.ie/rpublications/Kowalik-ISSC.pdf`.

[12] Jakob Eriksson, Sharad Agarwal, Paramvir Bahl, and Jitendra Padhye. Feasibility study of mesh networks for all-wireless offices. In *MobiSys 2006: Proceedings of the 4th international conference on Mobile systems, applications and services*, pages 69–82, New York, NY, USA, 2006. ACM Press.

[13] Richard Draves, Jitendra Padhye, and Brian Zill. Comparison of routing metrics for static multi-hop wireless networks. In *SIGCOMM '04: Proceedings of the 2004 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 133–144, New York, NY, USA, 2004. ACM Press.

[14] Atul Adya, Paramvir Bahl, Jitendra Padhye, Alec Wolman, and Lidong Zhou. A multi-radio unification protocol for ieee 802.11 wireless networks. In *BROADNETS '04: Proceedings of the First International Conference on Broadband Networks (BROADNETS'04)*, pages 344–354, Washington, DC, USA, 2004. IEEE Computer Society.

[15] Douglas S. J. De Couto, Daniel Aguayo, John Bicket, and Robert Morris. A high-throughput path metric for multi-hop wireless routing. In *MobiCom '03: Proceedings of the 9th annual international conference on Mobile computing and networking*, pages 134–146, New York, NY, USA, 2003. ACM Press.

[16] Richard Draves, Jitendra Padhye, and Brian Zill. Routing in multi-radio, multi-hop wireless mesh networks. In *MobiCom '04: Proceedings of the 10th annual international conference on Mobile computing and networking*, pages 114–128, New York, NY, USA, 2004. ACM Press.

[17] Routing Basics. `http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/routing.htm`, Accessed: 01-21-2007.

[18] Destination-Sequenced Distance Vector (DSDV). `http://www.ihp-ffo.de/systems/lv/ws0405/SP_Text.pdf`, Accessed: 02-05-2007.

[19] Andreas Gohr. Spontane Wireless LANs. Master's thesis, Fachhochschule für Technik und Wirtschaft Berlin, May 2004. Language: German.

[20] T. Clausen, P. Jacquet. Optimized Link State Routing Protocol (OLSR). `http://www.ietf.org/rfc/rfc3626.txt`, Accessed: 11-23-2006.

[21] Wikipedia. Optimized Link State Routing protocol. `http://en.wikipedia.org/wiki/OLSR`, Accessed: 11-19-2006.

[22] Openwireless Schweiz. `http://www.openwireless.ch`, Accessed: 11-24-2006.

[23] C. Perkins, E. Belding-Royer, S. Das. Ad hoc On-Demand Distance Vector (AODV) Routing. `http://www.ietf.org/rfc/rfc3561.txt`, Accessed: 11-23-2006.

[24] IETF Internet-Draft: The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR). `http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-10.txt`, Accessed: 11-20-2006.

[25] Zygmunt J. Haas, Marc R. Pearlman, Prince Samar. IETF Internet-Draft: The Zone Routing Protocol (ZRP) for Ad Hoc Networks. `http://people.ece.cornell.edu/~haas/wnl/Publications/draft-ietf-manet-zone-zrp-04.txt`, Accessed: 11-20-2006.

[26] Zygmunt J. Haas, Marc R. Pearlman, Prince Samar. The Intrazone Routing Protocol (IARP) for Ad Hoc Networks. `http://www3.ietf.org/proceedings/02jul/I-D/draft-ietf-manet-zone-iarp-02.txt`, Accessed: 11-20-2006.

[27] Zygmunt J. Haas, Marc R. Pearlman, Prince Samar. The Interzone Routing Protocol (IERP) for Ad Hoc Networks. `http://www3.ietf.org/proceedings/02jul/I-D/draft-ietf-manet-zone-ierp-02.txt`, Accessed: 11-20-2006.

[28] Zygmunt J. Haas, Marc R. Pearlman, Prince Samar. The Bordercast Resolution Protocol (BRP) for Ad Hoc Networks. `http://www3.ietf.org/proceedings/02jul/I-D/draft-ietf-manet-zone-brp-02.txt`, Accessed: 11-20-2006.

[29] Wikipedia. Hierarchical state routing. `http://en.wikipedia.org/wiki/Hierarchical_state_routing`, Accessed: 11-20-2006.

[30] Lifei Huang and Ten-Hwang Lai. On the scalability of ieee 802.11 ad hoc networks. In *MobiHoc '02: Proceedings of the 3rd ACM international symposium on Mobile ad hoc networking & computing*, pages 173–182, New York, NY, USA, 2002. ACM Press.

[31] Kamal Jain, Jitendra Padhye, Venkata N. Padmanabhan, and Lili Qiu. Impact of interference on multi-hop wireless network performance. In *MobiCom '03: Proceedings of the 9th annual international conference on Mobile computing and networking*, pages 66–80, New York, NY, USA, 2003. ACM Press.

[32] A. Acharya, A. Misra, and S. Bansal. High-performance architectures for ip-based multihop 802.11 networks. *Wireless Communications, IEEE [see also IEEE Personal Communications]*, 10(5):22–28, 2003.

[33] L. Iannone, R. Khalili, K. Salamatian, and S. Fdida. Cross-layer routing in wireless mesh networks. In *Wireless Communication Systems, 2004. 1st International Symposium on*, pages 319–323, 2004.

[34] Strix Systems. Wireless Mesh Networks for Metropolitan, City And Country-wide deployments – Strix Systems is the worldwide leader in Mesh Networks. `http://www.strixsystems.com/products/datasheets/strix_ows_system_description.pdf`, Accessed: 01-10-2007.

[35] Firetide. Firetide: Instant Mesh Networks. `http://www.firetide.com`, Accessed: 01-10-2007.

[36] Firetide. HotPort® Indoor Wireless Mesh Nodes. `http://www.firetide.com/hotfusion/documents/data_sheets/HotPort3100.pdf`, Accessed: 01-10-2007.

[37] Firetide. HotPort® Outdoor Wireless Mesh Nodes. `http://www.firetide.com/hotfusion/documents/data_sheets/HotPort3200.pdf`, Accessed: 01-10-2007.

[38] Firetide. HotPort® Indoor Wireless Mesh Nodes. `http://www.firetide.com/hotfusion/documents/data_sheets/HotPort3500.pdf`, Accessed: 01-10-2007.

[39] Firetide. HotPort® Outdoor Wireless Mesh Nodes. `http://www.firetide.com/hotfusion/documents/data_sheets/HotPort3600.pdf`, Accessed: 01-10-2007.

[40] Daniel Aguayo, John Bicket, Sanjit Biswas, Douglas S. J. De Couto, Robert Morris. MIT Roofnet Implementation, August 2003. `http://pdos.csail.mit.edu/roofnet/design`, Accessed: 01-22-2007.

[41] Daniel Aguayo and John Bicket and Sanjit Biswas and Glenn Judd and Robert Morris. Link-level measurements from an 802.11b mesh network. *SIGCOMM Comput. Commun. Rev.*, 34(4):121–132, 2004.

[42] Stuart J. Kerry, Donald Eastlake 3rd. IEEE P802.11 TGs. `http://grouper.ieee.org/groups/802/11/Reports/tgs_update.htm`, Accessed: 01-09-2007.

[43] See Mesh? SEEMesh Proposed. `http://www.wi-fiplanet.com/news/article.php/3522041`, Accessed: 02-08-2007.

[44] IEEE starts hammering out mesh network standard. `http://www.computerworld.com/networkingtopics/networking/lanwan/story/0,10801,103353,00.html`, Accessed: 02-08-2007.

[45] Members of the Alliance. `http://www.wi-mesh.org/index.php?option=com_content&task=view&id=12&Itemid=26`, Accessed: 02-08-2007.

[46] Guido Hiertz, Spiro Trikaliotis. Funknetze stricken – Gemeinsamkeiten und Unterschiede von WLAN und Mesh-Netzen. `http://www.heise.de/mobil/artikel/68923`, Accessed: 01-09-2007, Language: German.

[47] Wikipedia. IEEE 802.11s – Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/802.11s`, Accessed: 01-09-2007.

[48] Muni WiFi Mesh Networks. `http://www.belairnetworks.com`, Accessed: 01-21-2007.

[49] RoamAD. `http://www.roamad.com`, Accessed: 01-21-2007.

[50] Wi-Fi Planet – The Source for Wi-Fi Business and Technology. `http://www.wi-fiplanet.com`, Accessed: 01-21-2007.

[51] On.net. `http://www.on.net.mk/default-MK.asp`, Accessed: 01-21-2007.

[52] Ralf Steinmetz and Klaus Wehrle, editors. *Peer-to-Peer Systems and Applications*, volume 3485 of *Lecture Notes in Computer Science*. Springer, 2005.

Kapitel 11

QoS-enabled MAC Schemes for Wireless Networks

Aggeler Mattias, Hochstrasser Martin, Ma Seung Hee

QoS in Netzwerkapplikationen gewinnt mit den stetig steigenden Anforderungen multimedialer Anwendungen an Bedeutung. Speziell in drahtlosen Netzwerken, wo die Frequenzbänder beschränkt sind und alle Netzwerkknoten dasselbe Medium teilen müssen, ist die Bereitstellung von QoS eine besondere Herausforderung. Um der anwachsenden Nachfrage nach QoS gerecht zu werden, entstehen zahlreiche verschiedene Lösungsansätze für verschiedene drahtlose Netzwerktechnologien. In dieser Arbeit werden zuerst typische Anforderungen für QoS vorgestellt. Anschliessend werden Probleme erörtert, welche sich für die MAC-Schicht (Media Access Control) von drahtlosen Netzwerken ergeben und die für QoS hinderlich sind. Weiter werden grundlegende Eigenschaften von MAC-Verfahren aufgezeigt, die aufgrund von typischen QoS Anforderungen aus höheren Schichten notwendig sind. Danach wird ein Einblick gegeben in einige Ansätze für verschiedene drahtlose Übertragungstechnologien, welche versuchen QoS auf der MAC-Schicht zu implementieren. Dabei werden ausgehend von der verwendeten Technologie jeweils zuerst die klassischen und danach die neuen Ansätze präsentiert. Abschliessend gibt eine Diskussion der vorgestellten MAC-Verfahren im Bezug auf die im Kap. 11.2 vorgestellten QoS Anforderungen einen Überblick und die Grundlage für die abschliessende Bewertung.

Inhaltsverzeichnis

11.1 Einleitung

Der Begriff des QoS (Quality of Service) lässt sich nach den Unterlagen zur Vorlesung "Protocols for Multimedia Communications" [6] wie folgt definieren:

> Quality of Service: the grade, excellence, or goodness of a service; in the considered case, communication services.
>
> A concept for naturally describing communication-relevant features, formally specifying communication requirements, and "protocolly" defining rules for acquiring requested QoS.

Vor allem in drahtlosen Netzwerken gewinnt QoS immer mehr an Bedeutung. Dies zum einen wegen der Beschränkung des Mediums und zum anderen wegen den steigenden Qualitätsanforderungen von Multimedia-Applikationen. Aufgrund der Beschränkung der Frequenzbänder und der Nutzung eines gemeinsamen Mediums durch alle Benutzer, ist das Anbieten von QoS eine grosse Herausforderung. Die MAC-Schicht stellt die Grundlage für ein wirkungsvolles QoS dar. Nur wenn bereits in dieser Schicht die benötigten Funktionen für QoS eingebaut wurden, können sie auch von höheren Schichten benutzt werden. In dieser Arbeit wird der Begriff "MAC-Schema" im abstrakten mathematischen Sinne als MAC-Algorithmus verwendet. In der praktischen Anwendung jedoch als MAC-Protokoll.

In den weiteren Kapiteln werden zuerst die QoS Anforderungen allgemein erläutert und nachfolgend spezifisch aus Sicht des Medienzugriffs. Abschliessend wird dies an zwei Anwendungsbeispielen erläutert. Im Hauptteil der Arbeit werden die MAC-Schemata von drei drahtlosen Übertragungstechnologien präsentiert und hinsichtlich dieser Anforderungen betrachtet; 802.11, ATM, GPRS. Abschliessend werden diese im Bezug auf ihre QoS-Möglichkeiten beurteilt.

11.2 QoS Anforderungen

Die Zahl von verteilten Applikationen, welche mit Hilfe von Computernetzwerken ihre Dienste erbringen, steigt stetig. Gleichzeitig werden immer mehr drahtlose Netzwerke als Kommunikationsinfrastruktur verwendet. Immer mehr Applikationen benötigen QoS-Leistungen dieser Kommunikationsinfrastruktur, da sie nur damit sinnvolle Leistungen erbringen können. Bekannte Vertreter solcher Applikationen sind Multimedia-Anwendungen.

Die typischen QoS Anforderungen an Netzwerke lassen sich wie folgt zusammenfassen [3]:

- Durchsatz
- Transmission Verzögerung
- limitierter Jitter
- limitierte Fehlerrate

Probleme von geteilten drahtlosen Medien

FDM (Frequency Division Multiplexing) ist mit den zur Verfügung stehenden schmalen Frequenzbändern nur bedingt möglich. Ein Frequenzkanal muss bei der zu erwartenden grossen Anzahl an Benutzern von mehreren Stationen gleichzeitig verwendet werden (auch wenn diese nicht unbedingt miteinander kommunizieren).

Da die betrachteten Funknetzwerke die Mobilität der Benutzer ermöglichen müssen, ist die Signalemission omnidirektional. Dadurch schwächt sich das Signal mit wachsender Distanz stark ab. Weiter hinzu kommt das bekannte Problem der "hidden endnodes", welches einen zusätzlichen Koordinationsaufwand mit sich bringt [2]. Die bereits zahlreichen vorhandenen Funkquellen verursachen je nach Standort mehr oder minder starke Interferenzen, welche die Performance des betrachteten Mediums reduzieren können.

Zusammenfassend lassen sich daraus die folgenden Komplikationen ableiten.

Probleme für Datenübertragung (Folien M1-11 [2])

- Erhöhte Fehlerraten, ausgelöst durch Interferenzen
- Beschränkte Verfügbarkeit von Frequenzbändern
- Tiefere Datenraten
- Höhere Latenz
- Varianz der Latenz (erhöhter Jitter)

Allgemein führt die gleichzeitige Benutzung desselben Mediums zu grösseren Latenzen und erhöhten Fehlerraten. Durch die sich ändernde Auslastung entsteht auch Jitter. Diese Probleme der drahtlosen Übertragung vergrössern die Schwierigkeiten zur Erbringung von QoS.

11.2.1 QoS Anforderungen aus Sicht des Medienzugriffs

In einem Schichten-Modell wie zum Beispiel dem ISO OSI (International Standards Organisation / Open Systems Interconnection) Referenzmodell [1], erbringen höhere Schichten ihre Leistungen mit Hilfe der darunter liegenden Schichten. Um QoS in höheren Schichten anbieten zu können, müssen die tieferen Schichten ebenfalls QoS Leistungen erbringen. Bei Kommunikationsnetzen mit mehreren Teilnehmern, welche ein Medium gemeinsam benutzen müssen, ist eine Koordination des Medienzugriffs erforderlich. Im ISO OSI Referenzmodell ist MAC-Schicht dafür verantwortlich. Die Unterstützung von QoS in der MAC-Schicht ist erforderlich, damit höhere Schichten wiederum QoS implementieren und anbieten können. Die Eigenschaften der MAC-Schicht mit dem darin implementierten MAC-Protokoll hat Einfluss darauf, inwiefern QoS angeboten werden kann.

Um beispielsweise einen Datenstrom mit konstanter Bitrate senden zu können, müssen die Daten in Frames zerteilt und zu fixen Zeitpunkten gesendet werden. Das heisst, dass eine Garantie für die Zugriffszeit gegeben werden muss. Kommt es zu Kollisionen mit anderen sendenden Stationen oder wurde das Paket aufgrund von Störungen fehlerhaft übertragen oder empfangen, ergibt sich dadurch eine Unterbrechnung im Strom von Paketen und die effektive Durchsatzrate vermindert sich.

Die Latenz wird sowohl von der Dauer der Signalausbreitung als auch massgeblich von der Wartezeit vor dem Senden eines Frames beeinflusst. Wenn ein Datenframe an der MAC-Schicht ankommt und gesandt werden soll, muss die Station zuerst abwarten bis zum erlaubten Medienzugriff. Je nach MAC-Protokoll ist dies zum Beispiel Warten auf das freie Medium, Abwarten von inter frame spaces, Warten auf einen reserviertes Zeitfenster für die Übertragung oder Abwarten eines Pollings durch einen zentralen Koordinator. Je kürzer diese Wartezeit ist, desto kürzer wird auch die gesammte Übertragungslatenz. Um eine Garantie für eine maximale oder konstante Latenz leisten zu können, muss das MAC-Protokoll eine maximale Wartezeit für den Medienzugriff garantieren können. Ein weiterer Faktor für die Latenz ist auch die Fehlerrate durch verlorene Frames, wenn diese wiederholt übermittelt werden müssen.

Jitter verstärkt sich auf der MAC-Schicht, wenn die Zugriffszeiten für das Medium sehr unregelmässig sind, obwohl eine Station eigentlich in regelmässigen Abständen senden will. Eine Garantie auf regelmässigen Medienzugriff ist erforderlich um Jitter zu minimieren. Übertragungswiederholungen aufgrund von verlorenen Frames verstärkt Jitter zusätzlich. Je weniger Fehler bei der Übertragung von Frames eintreten, desto geringer kann der Jitter gehalten werden.

Um QoS an höhere Schichten anbieten zu können, müssen auf der MAC-Schicht vor allem Zeitgarantien oder zu mindest Zeitoptimierungen für den Medienzugriff einzelner Stationen angeboten werden können. Eine geringe Fehlerrate bei der Übertragung ist für die Erbringung von QoS von grossem Vorteil, manchmal unbedingt erforderlich. MAC-Schemata können durch die Vermeidung von Kollisionen einen wichtigen Beitrag zur Reduktion unbrauchbar übertragener Frames leisten.

11.2.2 Beispiele von Anwendungen für QoS

In den folgenden zwei Abschnitten werden anhand zweier populärer Applikationen - welche auf QoS angewiesen sind - Beispiele für QoS Anforderungen in Datenübertragungen vorgestellt.

Video-Streaming

Videoanwendungen zeichnen sich primär durch die höhere benötigte Bandbreite im vergleich zu anderen Kommunikationsanwendungen aus. Je nach Codierung der Videodaten haben diese eine konstante oder variable Bitrate. Um die Videodaten ohne Unterbrüche für den Betrachter übertragen zu können, ist eine minimale oder sogar konstante garantierte Datenrate im Netzwerk erforderlich.

Die Latenz zwischen Sender und Empfänger spielt nur eine untergeordnete Rolle, da auch ein Eintreffen des Datenstroms mit einigen Sekunden Verzögerung den Nutzen der Anwendung (im Gegensatz zu interaktiver Kommunikation) nicht spürbar mindert.

Varianzen in der Verzögerung (Jitter), können zwar mit entsprechenden Buffern ausgeglichen werden, und haben somit auch keinen grossen Einfluss auf die vom Endnutzer wahrgenommene Qualität. Die wichtige zu erfüllende Anforderung an ein Kommunikationsnetz für eine Video-Streaming Applikation ist somit die zu Verfügung gestellte und optimalerweise garantierte Bandbreite.

VoIP (Voice over Internet Protocol)

Telefon- oder Konferenzapplikationen wie VoIP benötigen im Vergleich zu Video - Applikationen weniger Bandbreite. Die benötigte Datenrate ist auch häufig nicht konstant, da Daten in der Regel nur übertragen werden müssen, wenn ein Teilnehmer gerade spricht. Die Anforderung an ein Datennetzwerk ist daher eine minimale garantierte Bandbreite. Der Kommunikationskanal wird zwar nicht immer genutzt, aber wenn er gebraucht wird, soll eine mindest Datenrate bereitstehen.

Eine sehr wichtige Eigenschaft, welche sich direkt auf die wahrgenommene Qualität der Anwendung und deren Nutzen auswirkt ist, die Übertragungslatenz. Hohe Latenzzeiten erschweren die Kommunikation, da diese interaktiv erfolgen soll. Ein Teilnehmer will nicht lange auf die Antwort seines Gesprächspartners warten müssen. Latenzzeiten müssen sich auf einige Zehntelssekunden beschränken und müssen daher auch von der Kommunikationsinfrastruktur ermöglicht werden. Dabei spielt neben der Latenz der Signalübertragung auch die Wartezeit für den Zugriff auf das Netzwerk eine grosse Rolle.

Jitter ist ein weiteres grosses Problem für die interakive Sprachübermittlung. Grosse Varianzen in der Übertragungs- und Ankunftszeit von Sprachdaten beeinträchtigen die wahrgenommene Qualität massiv und sind vor allem durch unerwünschte Unterbrechungen wahrnehmbar. Im Gegensatz zu nicht interaktiven, gestreamten Daten können durch Jitter verursachte Probleme nur sehr begrenzt durch Puffer gelöst werden, da diese die Latenz erhöhen würden. Demzufolge ist neben einer Mindestbandbreite und einer begrenzten Latenz ein möglichst geringer Jitter eine QoS Anforderung and das Kommunikationsnetzwerk.

11.3 802.11

11.3.1 Traditionelle Verfahren DCF / PCF

Distributed Coordination Function (DCF)

Das grundlegende MAC-Schema für WLAN (802.11) ist die DCF (Distributed Coordination Function). DCF regelt den Medienzugriff ohne zentrale Koordinationsstelle. Deshalb kann DCF auch in Ad-Hoc-Netzwerken verwendet werden.

Die DCF benutzt CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance) für den Medienzugriff. Damit eine Station zu Senden beginnen kann, darf das Medium nicht gerade durch den Sendevorgang einer anderen Station belegt sein, da andernfalls eine Kollision entsteht. Um festzustellen, ob das Medium gerade zur Übertragung frei ist, horcht die Station in das Medium hinein. Wenn das Medium unbenutzt ist, muss die Station zuerst noch eine vordefinierte Zeit abwarten, den IFS (Inter Frame Space), und eine zusätzliche zufällig gewürfelte Zeit, bis sie schliesslich zu Senden beginnen kann. Die minimale und maximale Länge dieser zusätzlichen gewürfelten Zeitspanne wird durch eine Zeitfenster definiert, welches Contention Window (CW) genannt wird. Kommt eine andere Station der wartenden Station zuvor und beginnt zu senden, muss die wartende Station wieder auf ein freies Medium warten. Die schon verstrichene Wartezeit wird beim nächsten Zugriffsversuch angerechnet. Somit erhalten Stationen, welche bereits lange auf den Medienzugriff gewartet haben eine höhere Zugriffswahrscheinlichkeit.

Die zufällige Zeitspanne aus dem CW soll dazu dienen, Kollisionen zu vermeiden. Wenn aber zwei Stationen zufällig die gleiche Zeitspanne gewürfelt haben, kann es vorkommen, dass beide gleichzeitig zu senden beginnen und somit eine Kollision im Medium entsteht. Dies wird von den sendenden Stationen durch Ausbleiben einer Empfangsbestätigung erkannt und die Stationen versuchen einen erneuten Sendeversuch. Nach jeder Kollision verdoppeln die Stationen die Zeitspanne, die abgewartet werden muss. Die Verdopplung kann solange wiederholt werden, bis die maximale Länge des CW erreicht wird.

Der IFS ermöglicht eine Priorisierung von verschiedenen Stationen. Zum Beispiel muss eine Station, welche eine Empfangsbestätigung zurücksenden will, eine kürzere Zeit (SIFS, Short Inter Frame Space) abwarten als andere wartende Stationen, welche eine längere Zeitspanne warten müssen (DIFS, DCF Inter Frame Space). Auch der zentrale Steuerungsknoten im PCF Zugriffsverfahren (siehe nächstes Kapitel) hat durch eine kürzere Wartezeit (PIFS, Point Inter Frame Space) den Vortritt vor anderen Stationen [10].

DCF bietet keinerlei QoS Garantien. Es ist nicht möglich einen Medienzugriff zu reservieren, oder eine Zugriffszeit zu garantieren. Durch die unterschiedlich langen IFS besteht zwar eine Priorisierung, diese ist jedoch sehr einfach und statisch, und bietet keine Möglichkeit, eine Kategorisierung in Übermittlungsprioritäten an höhere Schichten anzubieten [8].

Point Coordination Function (PCF)

Die PCF basiert auf der DCF und erweitert diese. Dabei übernimmt der PC (Point Coordinator) teilweise die Kontrolle über den Medienzugriff. Der PC, der meist auch gleichzeitig der Access Point eines fixen Wireless-LANs ist, sendet Beacon Frames aus, welche helfen, die Uhren der verschiedenen Stationen synchron zu halten, sowie die Zeit in eine Wettbewerbsphase und eine wettbewerbsfreie Zeitspanne einzuteilen.

Während der Wettbewerbsphase wird die herkömmliche DCF für den Medienzugriff verwendet. Der PC hat dabei jedoch immer den Vortritt vor den Mobilen Stationen, da er nur die kurze PIFS abwarten muss, bis er senden darf. Während der wettbewerbsfreien Periode dürfen Stationen ihrerseits keinen Sendeversuch starten. Nur der PC darf dann

eine Kommunikation initiieren. Der PC pollt während dieser Phase alle ihm bekannten mobilen Stationen, welche ein Frame zurücksenden, wenn sie etwas zu übermitteln haben. Ansonsten senden sie ein Null-Frame zurück, welches keine Payload enthält. Während dieser Zeit entstehen keine Kollisionen, da durch das Polling der PC das Übertragungsmedium für einen kurzen Zeitraum exklusiv für eine mobile Station reserviert [10].

Obwohl einige Vorteile der PCF gegenüber der DCF bestehen, können trotz Priorität des PC, des Pollingmechanismus und der Vermeidung von Kollisionen keine QoS Garantien an höhere Schichten angeboten werden. Dies hat nach [12] folgende Ursachen:

Das erste Problem besteht darin, dass kurz vor dem geplanten Ende der Wettbewerbsphase eine Station zu senden beginnen kann, obwohl die Übermittlung und das Zurücksenden einer allfälligen Empfangsbestätigung länger als die geplante Dauer des Wettbewerbsphase dauern kann. Der geplante Beginn der Wettbewerbsfreien Phase wird verzögert, da der PC zuerst auf ein freies Medium warten muss, bis er mit dem Beacon Frame den wettbewerbsfreien Zeitraum initialisieren kann. Durch diese mögliche Verzögerung kann keine Zeitgarantie für den Medienzugriff für Stationen geboten werden, da das Polling der einzelnen Stationen verzögert werden kann.

Ein weiteres Problem entsteht dadurch, dass - nachdem eine Station gepollt wurde - kein Zeitlimit für die Übertragung existiert. Die Übertragung des Frames hängt von der Grösse in Bytes, sowie der Übertragungsgeschwindigkeit ab, welche bei den 802.11 Standards bei jedem Frame wechseln kann. Diese Nichtabsehbarkeit der Sendedauer macht eine genaue Zeitgarantie für die Datenübertragung der nachfolgenden gepollten Stationen unmöglich.

Ein drittes Problem ist, dass der PC während der wettbewerbsfreien Zeitspanne alle Stationen genau einmal pollen muss, ungeachtet der Prioritäten der verschiedenen Stationen oder des zu übertragenden Datenvolumens.

Da keinerlei Zeitgarantien für die Übertragung angeboten werden, können trotz der besseren Auslastung des Mediums keine QoS Garantien, über zum Beispiel eine Mindestdatenrate oder eine maximale Übertragungsverzögerung, gegeben werden [8].

Ein weiteres erschwerendes Element für QoS besteht darin, dass es keine Beschränkung der Anzahl der teilnehmenden Stationen im Netzwerk gibt. Alle Stationen, unabhängig der Anzahl, haben bei der DCF die gleiche Priorität und Sendewahrscheinlichkeit und müssen auch bei der PCF vom PC beim Polling berücksichtigt werden. Durch das hinzukommen weiterer mobiler Stationen verringert sich die Verfügbarkeit des Mediums pro Station entsprechend. Bei vielen teilnehmenden Stationen kann es aufgrund der steigenden Kollisionswahrscheinlichkeit zu sehr hohen Verzögerungen in der Übermittlung von Datenframes kommen. Dies erhöht nicht nur die Wahrscheinlichkeit von verlorenen Frames wegen Überschreitung der maximalen Sendeversuche, sondern auch die Wahrscheinlichkeit von verlorenen Frames durch Pufferüberlauf aufgrund der langen Wartezeiten [12].

11.3.2 Neue Verfahren: 802.11e Hybrid Coordination Function (HCF)

Die HCF (Hybrid Coordination Function) ist das Kernstück des neuen 802.11e Standards. HCF baut auf der DCF Grundlogik auf, beinhaltet aber neue Zugriffsregeln für die einzelnen Stationen. Grundsätzlich besteht HCF ähnlich wie das herkömmlichen Medienzugriffsverfahren aus zwei Teilen: Einerseits gibt es neu die EDCF (Enhanced Distributed Coordination Function), welches im Grunde eine überarbeitete Version der DCF ist. Das zweite Zugriffsverfahren ist HCCA (HCF Controlled Channel Access), welches ein überarbeitetes Polling Verfahren - ähnlich wie PCF - darstellt.

Ein weiteres wichtiges neues Element, welches mit HCF eingeführt wird, ist die so genannte TXOP (Transmission Opportunity). Eine TXOP definiert eine Zeitspanne, während der eine Station senden darf. Stationen dürfen diese Zeitspanne im Sendevorgang nicht überschreiten [10].

Enhanced Distributed Coordination Function (EDCF)

Bei der EDCF werden an jeder Station die zu sendenden Daten in vier verschiedene Zugriffskategorien (AC, Access Categories) mit verschiedenen Prioritäten eingeteilt.

Die verschiedenen Prioritäten der ACs werden durch verschieden lange Wartezeiten auf den Medienzugriff durch die Parameter der IFS und des CW erreicht. Jede AC hat dabei eine eigene Warteschlange auf den Medienzugriff und steht mit den anderen ACs im Wettbewerb. Eine grafische Darstellung der Warteschlangen ist in Abbildung 11.1 dargestellt. Hat eine AC zu sendende Frames in der Warteschlange muss, nachdem das Medium frei geworden ist, zuerst die Dauer des AIFS[AC](Arbitration Interframe Space) abgewartet werden, welcher den DIFS von DCF ersetzt. Danach muss das CW abgewartet werden. Dessen Länge wird durch Maxima und Minima (CWmin[AC], CWmax[AC]) beeinflusst. Hinzu kommt noch für jede AC der Wert der TXOPlimit[AC], welche die maximale Sendedauer für die Frameübertragung bestimmt.

Eine AC mit kurzem AIFS[AC] und niedrigen Werten für CWmin[AC] und CWmax[AC] hat durch die kürzere obligatorische Wartezeit eine höhere Sendewahrscheinlichkeit als eine AC mit höheren Werten.

Die Abbildung 11.2 zeigt eine Übersicht über die verschiedenen Interframe Spaces und CWs. Falls mehrere ACs gleichzeitig sendeberechtigt werden, bekommt diejenige mit der höheren Priorität den Vorzug.

Damit die gleichen Zugriffskategorien der verschiedenen Stationen die gleiche Fairness zum Medienzugriff haben, werden die Parameter für die AIFS[AC], der CWmin[AC] und CWmax[AC] sowie die TXOPlimit[AC] der einzelnen Zugriffskategorien in den Beacon Frames des Access Points bekannt gegeben. Ist die Auslastung des Mediums gross, erhöht sich durch kleine Werte für CWmin[AC] und CWmax[AC] die Kollisionswahrscheinlichkeit. Wenn jedoch die Werte für CWmin[AC] und CWmax[AC] bei geringer Auslastung

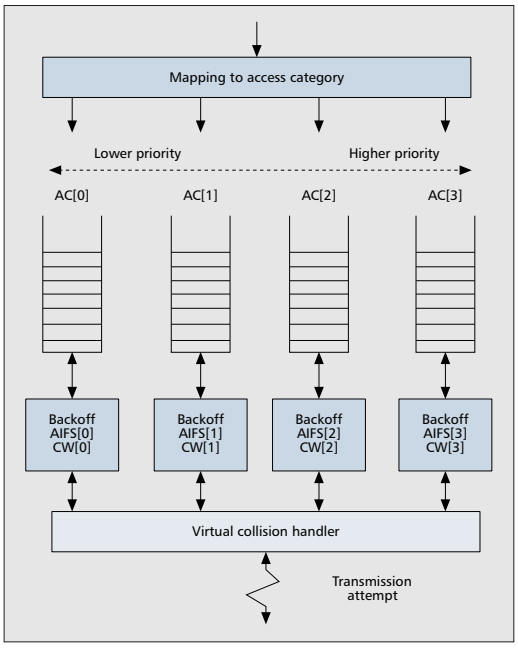


Abbildung 11.1: Die verschiedenen Warteschlangen auf den Medienzugriff von EDCF [10]

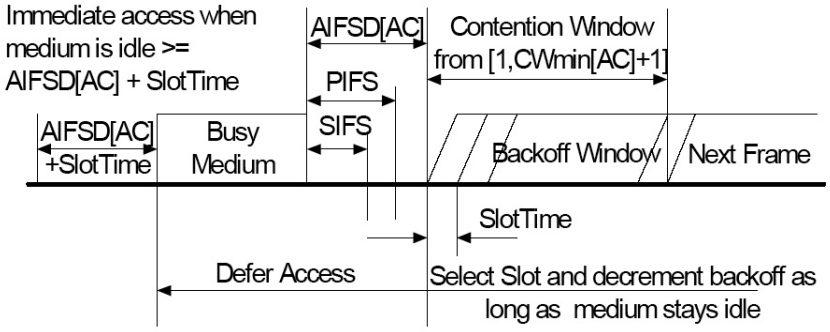


Abbildung 11.2: Übersicht der Interframe-Spaces [11]

des Mediums relativ hoch gesetzt sind, wird Bandbreite durch unnötige Wartezeiten verschwendet.

Eine Möglichkeit, die den einzelnen Stationen erlaubt die MAC-Parameter selbst anzupassen, aber dabei immer noch eine statistische Fairness auf den Medienzugriff der einzelnen Zugriffskategorien gewährleistet, wird in [9] vorgestellt. Dabei werden in regelmässigen Zeitabständen die Anzahl der übermittelten Frames, die Anzahl der fehlgeschlagenen Sendeversuche, sowie die Anzahl der Übertragungen, welche mehrere Sendeversuche benötigten ermittelt. Danach wird aus diesen Werten die Fehlerwahrscheinlichkeit für die Übertragung und die Kollisionswahrscheinlichkeit berechnet. Diese Wahrscheinlichkeiten werden durch eine speziell Funktion auf Werte für CWmin und CWmax abgebildet. Mit dieser Methode können die CWmin und CWmax Werte ohne die zentrale Kontrolle eines AP dynamisch an die Auslastung des Mediums angepasst werden. Die benötigten Daten für die Berechnung können von jeder Station autonom ermittelt werden.

Auch in diesem Zugriffsverfahren ist es wie in DCF möglich, dass Kollisionen vorkommen. Wenn zwei Zugriffskategorien zufällig ein gleich lange Wartezeit aus dem CW gewürfelt haben, beginnen beide gleichzeitig zu senden, und es kommt zu einer Kollision. Diese wird gleich wie beim herkömmlichen Verfahren durch die Verdopplung der Wartezeit bis hin zu CWmax[AC] und erneutem Sendeversuch behandelt [10].

Aufgrund der Differenzierung von vier verschiedenen Zugriffskategorien auf der MAC-Schicht können bei EDCF Prioritäten für den Datentransfer unterschieden, und höher liegenden Schichten angeboten werden. Datenströme aus höheren Schichten müssen somit auf die vier verschiedenen Kategorien abgebildet werden. Diese Abbildung erfolgt nach den acht verschiedenen Nutzerprioritäten aus dem 802.1D Standard und ist in Tabelle 11.1 ersichtlich.

Tabelle 11.1: Mapping der Prioritäten auf die Zugriffskategorien

Priorität	User Priorität in 802.1D	Zugriffskategorie (AC)	Verwendungszweck (informativ)
niedrigste	1	AC[0]	Hintergrund
	2	AC[0]	Hintergrund
	0	AC[1]	Best effort
	3	AC[1]	Video
	4	AC[2]	Video
	5	AC[2]	Video
	6	AC[3]	Voice
höchste	7	AC[3]	Voice

Gewisse Applikationen (wie zum Beispiel eine Videoübertragung) auf einzelnen Stationen können somit den Vorrang vor andern Applikationen mit tieferer Priorität anderer Stationen (zum Beispiel ein Filetransfer) für den Medienzugriff erhalten. Es ist jedoch zu beachten, dass trotz der bevorzugten Behandlung der höheren Zugriffsklassen innerhalb der Klassen immer noch das Prinzip des Best-Effort gilt. Somit können keine strikten QoS Garantien wie Mindestdatenraten oder maximale Verzögerungszeiten angeboten werden.

HCF Controlled Channel Access (HCCA)

HCCA ist ein polling basiertes MAC-Schema, bei welchem ein zentraler Koordinator - in der Regel ein AP (Access Point) - den Medienzugriff aller Stationen im Netzwerk kontrolliert. HCCA stellt im Grunde eine abgeänderte Version des PCF Verfahrens dar, basiert aber nicht darauf, sondern stellt für sich einen eigenständigen Teil der HCF dar. Wie beim alten MAC-Standard wechseln sich wettbewerbsbasierte und wettbewerbsfreie Phasen für den Medienzugriff über die Zeit ab. Während der wettbewerbsfreien Zeitperiode pollt der AP Stationen nach seiner Pollingliste ab. Aber auch während der wettbewerbsbasierten Phase kann der AP eine Station pollen. Wie bei PCF hat der AP beim neuen HCCA den Sendevorrang vor allen anderen Stationen. Die Mindestwartedauer ist wie beim PCF für den AP lediglich die Dauer des PIFS [10].

Bei der Spezifikation des HCCA wurde darauf geachtet, die oben genannten Probleme des PCF zu lösen: Im Gegensatz zum Zusammenspiel von DCF und PCF ist es bei HCF nicht möglich, dass eine Station, welche gegen Ende der wettbewerbsbasierten EDCF-Phase sendet, den Anfang der HCCA-Phase verschiebt. Eine Station beginnt nicht zu senden, wenn sie nicht vor dem im Beacon Frame des AC angekündigten Beginn der HCCA-Phase fertig wird. Damit ist es garantiert, dass die Polling-Phase des HCCA immer zum geplanten Zeitpunkt anfangen kann. Des weiteren ist es nicht mehr möglich, dass eine Station während des Pollings die Polling-Zeitpunkte der anderen Stationen durch unabsehbare Sendedauer verzögern kann. Die im Polling-Frame vergebene TXOP Länge ist für die Station verbindlich. Dadurch sind die Pollingzeitpunkte voraussehbar und können garantiert werden. Ein weiterer Vorteil gegenüber der PCF besteht darin, dass der AP nur diejenigen Stationen pollen muss, welche auch Daten zu senden haben. Der AP kann die Pollingreihenfolge und die TXOPs für jede Station anpassen. Damit der AP weiss, welche Stationen wie oft gepollt werden müssen, senden Stationen ein spezielles TSPEC-Frame (Traffic Specification Frame), welches Parameter für das Polling enthält.

Durch die Neuerungen in HCCA ist es dem AP nun möglich neben der kurzzeitigen exklusiven Reservation des Mediums für eine Station auch genaue Zeitpunkte für den Medienzugriff zu garantieren und einzuhalten. Die Stationen können dem AP durch die Übermittlung des TSPEC-Frames benötigte QoS Garantien für den Medienzugriff mitteilen. Die wichtigsten TSPEC Parameter sind nach [10]:

- Durchschnittliche Datenrate
- Maximale Verzögerung
- Maximaler Serviceintervall
- Nominelle Paketgrösse
- Minimale physische Übertragungsrate

Der AP entscheidet, ob er die QoS-Garantien gewährleisten kann und teilt dies der Station mit. Der AP ist dafür verantwortlich, durch die Zeitplanung des Pollings und der TXOPs die im TSPEC-Frame beschriebenen Parameter einzuhalten. Der AP kann mit HCCA somit einen garantierten Medienzugriff für verschiedene Stationen anbieten [8].

Durch die neu ermöglichte, strikt zeitgebundene Reservation für den Zugriff des Mediums und der QoS-Parameter Anfrage über TSPEC ist es nun möglich, höheren Schichten parametrisierte QoS-Dienstleistungen, wie zum Beispiel garantierter Übertragungsraten oder maximale Verzögerungszeiten, anzubieten.

11.3.3 Abschliessende Beurteilung und Zusammenfassung

Der WLAN Standard (802.11) kann aufgrund seiner MAC-Schicht kein QoS für die Datenübertragung anbieten. Auch die polling-basierte PCF kann aufgrund von möglichen Zeitverzögerungen keine QoS Garantien bieten. Um QoS in WLAN zu ermöglichen wurde

mit 802.11e ein neuer Standard für die MAC-Schicht ausgearbeitet. Er erlaubt mit EDCF die Priorisierung von Datenströmen nach deren Wichtigkeit und Anforderungen für die Übertragung und ermöglicht dadurch statistische QoS. HCCA erlaubt durch ein neues Polling-Verfahren das Medium für einzelne Stationen explizit zu reservieren. Dies erlaubt genau durch TSPEC spezifizierte QoS Garantien anzubieten, welche speziell auf einzelne Applikationsanforderungen zugeschnitten sind.

11.4 ATM

ATM als Übertragungsstandard bietet viele technische Möglichkeiten, die von modernen Netzwerken verlangt werden. ATM kann mit einer fast unbegrenzten Kapazität aufwarten und besitzt sehr tiefe Fehlerraten [4]. Da ATM im Gegensatz zu typischen drahtlosen Netzwerken verbindungsorientiert arbeitet, ist es möglich eine breite Unterstützung für QoS anzubieten. Die typischen Service-Kategorien, welche von ATM angeboten werden sind in Tabelle 11.2 aufgeführt.

Tabelle 11.2: Service Typen in ATM [4]

Service	Anwendungsfelder
Constante Bit Rate (CBR)	Digital voice und video
Real-Time Variable Bit Rate (rt-VBR)	Komprimiertes voice und video
Non-Real-Time Variable Bit Rate (nrt-VBR)	Bursty traffic, keine Verzögerungsgarantien
Available Bit Rate (ABR)	Nur eine minimale Bandbreite wird garantiert, ohne andere Zusagen
Unspecified Bit Rate (UBR)	Best effort service, keine Garantien

Um kabellos überhaupt Hochgeschwindigkeits-Netzwerke wie ATM zu ermöglichen, ist es notwendig, das limitierende Medium auf eine andere Art und Weise zu benutzen, als man sich dies von drahtlosen Netzwerken gewohnt ist. Eine wichtige Anforderung betrifft die Mobilität: Diese muss zwischen verschiedenen Funkzellen gewährleistet sein. Ebenfalls wird ein funktionierendes QoS verlangt.

Die Kombination von ATM und Wireless beinhaltet noch einige andere Schwierigkeiten wegen der Kompatibilität, welche hier kurz dargestellt werden:

- ATM ist ursprünglich für ein Medium mit grosser Bandbreite konzipiert worden. Wireless bietet jedoch eine relativ geringe Bandbreite.

- ATM fordert eine sehr niedrigere BER (Bit Fehler Rate).
  Wireless hingegen ist ein Medium, dessen Qualität stark variiert und welches viel Hintergrundrauschen beinhaltet.

11.4.1 Traditionelle Verfahren: DAMA

DAMA steht für Demand Assignment Multiple Access. Es ist ein Verfahren, welches Time-Slots verwendet, welche in Rahmen aufgeteilt sind. Jeder Rahmen ist dabei aufgeteilt in einen Up- und einen Downstream-Kanal. Diese wiederum sind aufgeteilt in zwei Untereinheiten (Subslots). Der Upstream-Kanal ist in den RA (Request Access) und den TA (Transmission Access) Subslot aufgeteilt. Der Downstream-Kanal ist aufgeteilt in die Subslots ACK (Acknowledgement) und DD (Downstream) [4]. Um dies zu veranschaulichen sei auf Abbildung 11.3 verwiesen.

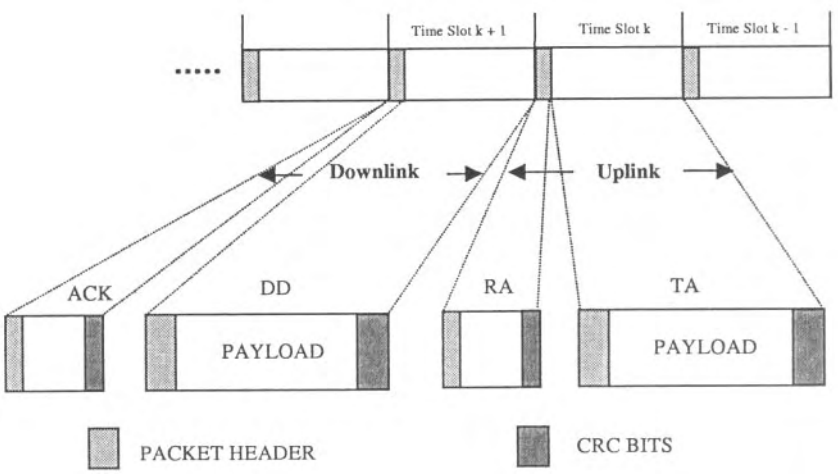


Abbildung 11.3: Radio channel classification: slot-by-slot [4]

Die Übertragung des Downstreams wird jeweils von der Basisstation durchgeführt und kontrolliert. Dies geschieht konfliktfrei unter der Verwendung des TDM (Time Division Multiplexing). Diese Übertragungen resultieren in einer geringen Verzögerung. Beim drahtlosen Service ist die Bandbreite, welche von einer Basisstation benötigt würde, um alle mobilen Stationen gleichzeitig zu versorgen, viel grösser als die zur Verfügung stehende. Aber die mobilen Geräte verwenden ein Random Access Schema in RA Untereinheiten, da sie nicht alle gleichzeitig aktiviert sind [4]. Der Sender reserviert die zukünftigen Zeitperioden. Dadurch werden die Daten ohne Kollision gesendet. Zum Beispiel wird bei DAMA eine explizite Reservationsmethode und bei PRMA (siehe Variationen des DAMAs) eine implizite eingesetzt. Folien M3-9 [2].

Variationen des DAMAs

RAMA (Resource Auction Multiple Access) wurde zur schnellen Zuteilung von Ressourcen und als Handoff-Mechmismus vorgeschlagen. In diesem deterministischen Protokoll stellen die mobilen Benutzer Zugriffsanfragen, indem sie ihre Access ID übermitteln.

PRMA (Paket Reservation Multiple Access) ist konzipiert um die Effizienz der Bandbreite beim TDMA zu verbessern. Wenn ein mobiler Benutzer mit einem regelmässigen Datenaufkommen erfolgreich ein Paket in einem freien Slot übermittelt (unter

Verwendung von Slotted Aloha), werden ebenfalls zukünftige Slots in regelmässigen Abständen reserviert. Dies soll einen kontinuierlichen Durchsatz ohne Unterbruch und ohne zusätzlichen Overhead ermöglichen.

Das Aloha Protokoll ist ein OSI Layer 2-Protokoll. Wenn der Sender Daten hat, dann schickt er sie, ohne im Voraus das Medium auf Kollisionen geprüft zu haben. Deshalb ist dieses Protokoll ein Random Access Protokoll. Wenn die Daten durch Kollisionen nicht richtig ankommen, schickt der Sender die Daten noch mals. Bei Slotted Aloha werden die Daten nur gesendet, wenn der Sender die Erlaubnis durch einen Start-Slot erhalten hat.

DQRUMA (Distributed Queuing Request Update Multiple Access) wurde entworfen, um eine effiziente Auslastung der Bandbreite zu ermöglichen und QoS-Parameter zu unterstützen. Diese sind mit den QoS-Service-Typen von ATM (siehe Tabelle 11.2) kompatibel. DQRUMA verwendet das Konzept von dynamischen Upstream Untereinheiten. Die Upstream Untereinheit kann bei Bedarf gänzlich mit RA Minislots gefüllt werden (bei grossem Aufkommen von Requests). Dies reduziert den Request-Konflikt enorm. Requests in RA-Kanal verwenden ein Random Access Protokoll wie Slotted ALOHA. Pakete sind auf der Warteliste des Buffers bis die BS (Basestation) diese nach ihrer Scheduling Policy abarbeitet [4]. DQRUMA benutzt ein zusätzliches Bit, das Piggyback-Bit, in Upstream- Kanal. Es informiert die BS, ob die betroffene Übertragung aus mehreren Paketen besteht. Dadurch kann die BS direkt einen längeren Kanal-Zugriff erlauben.

11.4.2 Neue Verfahren: ARCMA

ARCMA steht für "Adaptive Request Channel Multiple Access" und ist ein nach Bedarf zuteilendes Protokoll (Demand Assignment Multiple Access Protocol) mit dynamischer Bandbreiten-Zuweisung. Der RA-Kanal in ARCMA ist fähig zusätzliche Information für verschiedene ATM Service-Typen zu übermitteln, z.B. CBR, VBR, usw. Dank diesen Zusatzinformationen kann die Basisstation bessere QoS Unterstützung für die verschiedenen Traffic-Klassen (Service-Typen) anbieten. ARCMA benutzt ein Piggyback-Bit ähnlich wie DQRUMA um den Konflikt im RA-Kanal zu reduzieren.

ARCMA verwendet jedoch im RA-Kanal einen komplexeren Algorithmus für das Random Access Schema: Der Slotted Aloha mit BEB (Binary Exponential Backoff) wird hier als Protokoll für ARCMA eingesetzt. Dies hilft vor allem gezielt Kollisionen im RA-Kanal zu reduzieren und die Unterstützung der verschiedenen Klassen von ATM zu verbessern [4].

Beschreibung

1. Request / Acknowledgement Phase

   Fast alles in dieser Phase funktioniert wie bei DQRUMA. Der Request wurde im RA-Kanal (RA Minislots) geschickt. Der Request beinhaltet die mobile Access-ID und er wird in der Setup-Phase durchgeführt. Bei ARCMA umfasst das Request-Paket neben dem Request auch den TOS (Type of Service)-Wert des gewünschten Services (siehe auch Abbildung 11.4).

2. Permission / Transmission Phase

   Standard Traffic: Die BS ist für die Allokation von Bandbreite (time slots) zuständig. Dabei wird die "Packet Transmission Policy" verwendet. FIFO (First In First Out)-Policy wird häufig dafür eingesetzt, d.h. das Element, das als Erstes ankommt, wird zuerst abgearbeitet.

   CBR Traffic: Eine spezielle CBR-Request-Queue wird verwendet, welche als erste abgearbeitet werden muss vor den restlichen Request-Tabellen. Da dieser CBR Traffic immer Priorität im Vergleich zu anderen Traffic besitzt, muss die Anzahl von MSs mit CBR Traffic in einer Zelle beschränkt werden.

3. Adaptive Request Channel Strategy

   Durch die Anpassung der Traffic-Umgebung werden die Kollisionen auf dem RA-Kanal reduziert. Wenn die BS entdeckt, dass die Request-Liste leer ist, werden die nächsten Upstream TA-Slots zu RA-Minislots umgewandelt und eingesetzt. Im nächsten Time-Slot werden diese RA-Minislots der MS für die Request-Transmission zur Verfügung gestellt (siehe auch Abbildung 11.4).

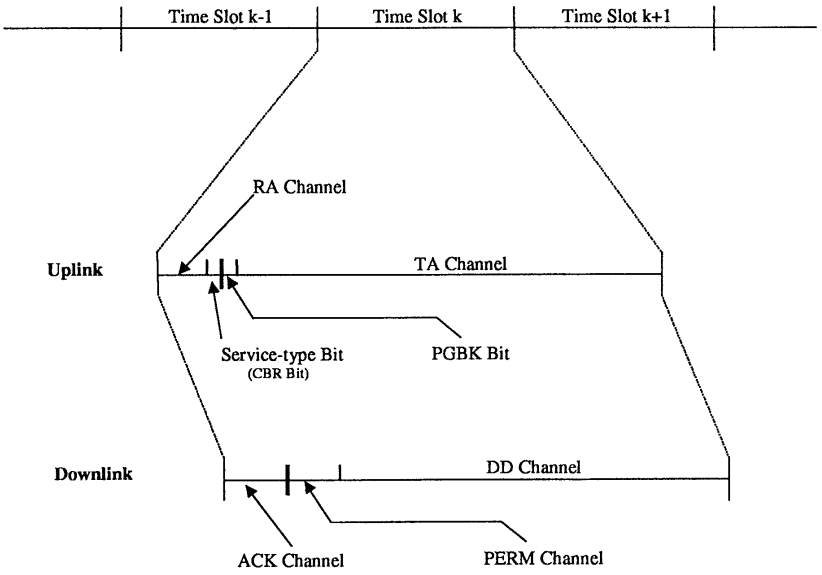


Abbildung 11.4: Timing Diagram ARCMA [4]

Performance

1. Adaptiver RA-Kanal: Die mögliche Auslastung des Kanals wird durch das Ausnutzen von freien TA Slots verbessert. Die Belastung des RA-Kanals wird dadurch verringert. Alle Anfragen werden geschickt, nachdem der Kanal zum Multiple RA-Modus umgewandelt wurde. Dadurch wird die Anzahl der mobilen Anfragen pro RA-Kanal und die Wahrscheinlichkeit von Kollisionen weiter reduziert.

2. CBR Behandlung: CBR Traffic wird mit einer minimalen Verzögerung übermittelt, da der Zugriff frei von Anfragen erfolgt und CBR eine höhere Priorität geniesst.

Diese Eigenschaft ist bei CBR Traffic sehr wichtig, da dieser Traffic verzögerungsempfindlich ist. Ausserdem reduziert der folgende, von Anfragen freie Zugriff, den Wettbewerb im RA-Kanal. Aus diesen Gründen wird die Performance enorm verbessert.

3. Slotted Aloha unter Verwendung des "Exponential Backoff Algorithmus": Slotted Aloha wird für das "Random Access Protokoll" im RA-Kanal eingesetzt. Die Verwendung des BEB Algorithmus erhöht die Stabilität des Protokolls. Durch diese Kombination soll die Zugriffsverzögerung verringert werden, indem aufeinanderfolgende Kollisionen im RA-Kanal vermieden werden. Wird für den RA-Kanal zusätzlich "Multiple RA Modus" eingesetzt, steht eine zusätzliche Kontroll-Schicht zur Verfügung um Kollisionen zu verringern.

Merkmale [4]

1. Effiziente Nutzung des Kanals: Adaptive RA-Kanal Strategie, besondere Behandlung des CBR-Traffics und Piggyback-Strategie werden dafür eingesetzt.
2. Übertragung erfolgt "slot by slot": Falls eine Kollision entsteht, werden die MSs dies sofort feststellen und das Paket wird in der nächsten Zeitperiode wieder geschickt.
3. Es besteht Transparenz gegenüber dem AAL. Dadurch wird die Komplexität zwischen kabelgebundenen und kabellosen Netzwerken reduziert.
4. Die verwendeten RA Pakete sind klein. Bei der Kollision entsteht wenig Verlust.
5. Die Reihenfolge der Pakete bleibt erhalten. Alle Pakete werden im Buffer der MSs zwischengespeichert und in ihrer Reihenfolge abgearbeitet. Dies garantiert die ursprüngliche Reihenfolge.
6. Es bestehen mehrere Upstream und Downstream Kanäle simultan.

11.4.3 Vergleich ARCMA mit DQRUMA

Die drei grossen Unterschiede: Besondere CBR-Traffic Behandlung, Priorität der Anfrage-Liste, Adaptiver RA-Kanal

ARCMA benutzt den Kanal effizienter als DQRUMA bei den verschiedenen Traffic-Bedingungen. Je intensiver der Traffic wird, desto grösser wird der Unterschied zwischen den beiden Protokollen. ARCMA stellt für verzögerungs-empfindlichen Daten-Traffic (CBR Traffic) die bessere Wahl dar, da hier eine spezielle Behandlung für CBR Traffic angeboten wird. Grund hierfür ist v.a. die Priorisierung der "Transmission Scheduling Policy".

Die Performance von ARCMA liegt laut Testresultaten bei allen Simulationsbedingungen über jenen von DQRUMA, unabhängig von der Traffic-Intensität [4]. Dabei werden Kanal-Durchsatz, durchschnittliche Transmissions-Verzögerung und die durchschnittliche Länge der Warteliste als Perfomance-Massstab verwendet. Der Kanal bei ARCMA besitzt einen höheren Durchsatz und eine niedrigere durchschnittliche Verzögerung. Vor allem durch die Vermeidung von Konflikten im RA-Kanal wird der Kanal viel effektiver verwendet.

11.4.4 Abschliessende Beurteilung und Zusammenfassung

Es wird ein neues Protokoll gefordert, welches beschränkte Medien wie drahtlose Netzwerke effektiver ausnutzen kann. Ausserdem wird die Unterstützung vom Realtime- und verzögerungsempfindlichen Traffic erwartet.

ARCMA ist konzipiert um diese Anforderungen im MAC-Sublayer zu erfüllen. Es existieren mehrere Protokolle, die hohe Bandbreiten und tiefe BER bei kabellosen ATM-Netzwerken gewährleisten können. Es fehlt jedoch die Unterstützung der verschiedenen ATM Service-Typen. ARCMA hingegen bietet einen besseren Support für den Verzögerung-Empfindlichen CBR Traffic durch die Priorisierung der "Transmission Scheduling Policy". Ausserdem verbessert ARCMA die Kanal-Ausnutzung durch Kollisions-Reduktion [4].

11.5 GPRS

Die heute in Verwendung befindlichen GPRS-Netzwerke verfügen für die Übertragung von Daten und insbesondere Multimediadaten nicht über optimale QoS-Möglichkeiten. Ein Ansatz, um dessen QoS mit mehr Funktionalität auszustatten, wird von OLIVER T.W. YU vorgeschlagen [5]:

> [...] a dynamic adaptive guaranteed QoS provisioning scheme over GPRS wireless mobile links by proposing a guaranteed QoS media access control (GQ-MAC) protocol and an accompanying adaptive prioritized-Handoff call admission control (AP-CAC) protocol to maintain GPRS QoS guarantees under the effect of mobile Handoffs.

Zuerst wird das klassische QoS beschrieben und anschliessend der neue Ansatz vorgestellt, unterteilt in die beiden Protokolle GQ-MAC und AP-CAC. Dabei werden den Erweiterungen "Mobile Label Switched Tree" und "Dynamic guard bandwidth scheme" separat betrachtet.

11.5.1 Traditionelle Verfahren: PRMA

Die Zugriffsmethoden in einem GPRS Netzwerk basieren auf FDD und TDMA (Frequency Division Duplex / Time Division Multiple Access). Für die Zeit einer aktiven Verbindung werden dem Benutzer jeweils zwei Frequenzen für Uplink und den Downlink zugeteilt. Kombiniert mit dem erwähnten Multiplexing, ist es möglich dieselben Frequenzen mehreren Benutzern zuzuteilen. Im Downlink-Kanal wird Bandbreite nach nach dem first-come, first-served Verfahren zugeteilt. Im Uplink-Kanal kommt Slotted ALOHA zum Zug. Die heute verwendeten Profile unter FDD/TDMA in GPRS-Netzen bieten folgende QoS-Klassen an:

- precedence

- delay
- reliability
- mean and peak throughput

In GPRS-Netzwerken wird PRMA als Protokoll verwendet. Dabei wird von MSs Slotted ALOHA für die Reservation verwendet. Bei Sprach-Quellen wird zusätzlich dieselbe Slot-Position in den nächsten Frames reserviert. Damit kann eine unterbrechungsfreie Unterhaltung (QoS Profil: conversational) bestmöglich garantiert werden (dabei handelt es sich nicht um ein klassisches Gespräch über GSM, sondern um paketorientierte Voice-Daten). Daten-Quellen jedoch müssen jedes mal, wenn sie Pakete senden wollen, einen neuen Slot im Wettbewerb mit anderen Daten-Quellen beantragen. Ein detaillierter Beschrieb aller anderen DAMA Varianten wird in Kapitel 11.4.1 gegeben. Durch gewisse Erweiterungen wie Centralized-PRAMA oder Integrated-PRAMA kann die Effizienz zwar gesteigert werden und unter den Daten-Quellen eine optimierte Service-Fairness erreicht werden. Jedoch bietet keiner der Ansätze die Möglichkeit eine Delay Garantie (Bounded Delay) zu geben, denn alle weisen den Nachteil eines variablen access delays auf [5].

Die Abbildung 11.5 soll einen Überblick ermöglichen über das vorgesehene Mapping zwischen den neu definierten QoS Service Typen **Conversational, Streaming, Interactive, Background** und die klassischen QoS-Möglichkeiten in einer GPRS Umgebung.

GPRS QoS class	QoS attributes	Implemented QoS profiles			
		Streaming	Conversational	Interactive	Background
Precedence	Congestion packet discard probability	Tolerable ($<10^{-2}$)	Tolerable ($<10^{-2}$)	Loss sensitive ($<10^{-5}$)	N/A
Delay	Latency	Bounded (<500 ms)	Bounded (<80 ms)	Less stringent than that of conversational & streaming	N/A
	Jitter	Stringent	Stringent	N/A	N/A
Reliability	Packet loss probability	Tolerable ($<10^{-2}$)	Tolerable ($<10^{-2}$)	Loss sensitive ($<10^{-5}$)	N/A
Mean and peak throughputs	Throughput	Guaranteed	N/A	Guaranteed	N/A
	Burstiness	Low	High	Higher than conversational	N/A

Abbildung 11.5: Mapping der neuen Profildefinitionen auf GPRS QoS-Klassen [5]

conversational meint hier die paketorientierte Übertragung von Gesprächen. **interactive** bezeichnet Kommunikation, welche auf eine garantierte Bandbreite angewiesen ist.

11.5.2 Neue Verfahren: GQ-MAC / AP-CAC

GQ-MAC

Das Protokoll GQ-MAC (Guaranteed QoS Media Access Control) ermöglicht die Reservation für Übertragungen der Klasse "streaming traffic" und eine priorisierte Reservation für die Klassen "conversational" und "interactive". Burst Traffic wird durch eine dynamisch adaptive Zuteilung von Ressourcen (Peak Bandwidth Allocation) kompensiert [5].

Prioritätsregelung beim Verbindungsaufbau

Die Datenkanäle in GPRS (Paket Data Channels / PDCH) sind folgendermassen unterteilt:

- Paket Random Access Channel (**PRACH**)
  Kanal für die Zugriffsanfrage im Uplink
  - Signalisierungskanal (**S-PRACH**)
  - Datenkanal (**U-PRACH**)
- Paket Access Grant Channel (**PAGCH**)
  Kanal für die Bestätigung der Zugriffsbewilligung im Downlink
- Paket Data Traffic Channels (**PDTCH**)
  restliche PDCH Kanäle im Up- und Downlink für Nutzdaten

Die Prioritäten beim Verbindungsaufbau über den PRACH Kanal sind in Tabelle 11.3 ersichtlich. Dabei wird sowohl für die Signalisierung auf dem S-PRACH Kanal als auch für den Datentransfer auf dem U-PRACH Kanal eine Unterscheidung getroffen. Der S-

Tabelle 11.3: Zugriffs Prioritäten unter GQ-MAC

Request access type		Access priority	PRACH used
Signaling	New call	Low	S-PRACH
	Handoff	High	
User data	Conversational traffic	High	U-PRACH
	Interactive traffic	Low	

PRACH Kanal wird über Slotted ALOHA sowohl von neuen Verbindungsanfragen als auch von Handoff Anfragen gleichzeitig benutzt um Zugriff auf die PDCHs zu erhalten. Die Handoff-Anfragen erhalten dabei eine höhere Priorität. Im U-PRACH Kanal wird unterschieden zwischen conversational und interactive Traffic. In einem Wettbewerbs Zeitfenster (Contention Cycle) ist nur conversational Traffic zugelassen. Durch diese Restriktion ist es möglich für conversational Traffic ein Bounded Delay für den Zugriff zu garantieren. (Dies ist notwendig, um unterbrechungsfreie Gespräche zu garantieren.)

AP-CAC

Das AP-CAC Protokoll (Adaptive Prioritized Handoff CAC) bietet eine sich dynamisch anpassende Zulassungskontrolle an, welche "Handover" über Anfragen für einen Neuaufbau priorisiert. Dies wird v.a. deshalb so gehandhabt, da ein Unterbruch einer schon bestehenden Verbindung für den Benutzer als störender empfunden wird, als die temporäre Verweigerung eines Neuaufbaus. Die Kapazität, welche dabei für Handovers bereitsteht

wird dynamisch angepasst an die jeweiligen Belegungszustände der benachbarten Zellen. Je höher belegt die benachbarte Zelle ist, desto höher fällt die Reservation für Handovers aus. Der hierzu verwendete Vorschlag wurde in [7] präsentiert. Es handelt sich dabei um ein "single guard channel scheme", in welchem mehrere verschiedene Prioritätsstufen verwendet werden. Dabei sind für Handoff Requests eine fixe Anzahl an "guard channels" (N_G) reserviert. Ist also die Anzahl verfügbarer freier Kanäle kleiner als (N_G), sind neue Verbindungen nicht zugelassen (Handoff jedoch schon). Damit kann die Wahrscheinlichkeit von nicht erfolgreichen Handoff Anfragen drastisch reduziert werden, ohne dass die Anzahl an abgewiesenen neuen Verbindungen wesentlich ansteigt. In AP-CAC wird ein verfeinertes Schema verwendet, welches Handoff Anfragen noch unterteilt, abhängig von den verschiedenen QoS-Klassen. Dabei sind drei Prioritätsstufen vorhanden. (N_{G1}) und (N_{G2}) sind für zwei Kategorien von Handoff Anfragen zuständig und haben ihrerseits Vorrang vor der dritten Kategorie, welche für neue Verbindungsanfragen zuständig ist. Die Behandlung von Anfragen läuft dabei folgender Weise ab [5]:

1. Klasse N ist für neue Verbindungen zuständig
   Solche werden nur zugelassen, wenn die Anzahl freier Kanäle grösser als (N_{G1}) ist.

2. Klasse H1ist für Handoff Anfragen zuständig
   Beinhaltet die Kategorien **interactive, conversational, background**
   Solche werden nur zugelassen, wenn die Anzahl freier Kanäle grösser als (N_{G2}) ist.

3. Klasse H2 ist für Handoff Anfragen zuständig
   Beinhaltet die Kategorie **streaming traffic**
   Solche werden dann zugelassen, wenn irgendein freier Kanal zur Verfügung steht

Die Anzahl freier Kanäle wird laufend dynamisch an die Netzsituation angepasst. Im nächsten Abschnitt wird im Detail darauf eingegangen.

Um das QoS durchgehend über das ganze drahtlose GPRS-Netzwerk garantieren zu können, muss ein Mapping zwischen den IntServ-Parametern des Backbone-Netzwerks und den QoS Bedürfnissen des GPRS-Netzwerks vorgenommen werden Dazu muss auch das AP-CAC Protokoll vom drahtlosen Netzwerk auf das Backbone-Netzwerk erweitert werden um eine einheitliche, durchgehende Zulassungskontrolle mit sich dynamisch anpassenden Prioritätsstufen für die Zulassung zu ermöglichen [5].

11.5.3 Erweiterung von AP-CAC

Das vorgestellte Protokoll AP-CAC unterstützt die adaptive Zulassung von Traffic auf Basis der Prioritätsklassen. Durch die Verwendung des "dynamic guard channel scheme" ist es möglich die Zugriffs-Kapazitäten auch für Handoff Anfragen verschiedener "traffic classes" dynamisch anzupassen. Dies geschieht abhängig von den momentan aktiven Benutzer-Verbindungen in den benachbarten Funk-Zellen als auch abhängig von den beobachteten Bewegungsmustern (mobility pattern). So können potentielle "Handovers" schon frühzeitig mit in Betracht gezogen werden.

Um nun das AP-CAC erfolgreich auf das Kernnetz auszudehnen, müssen die kalkulierten Handoffs auf alle Verbindungen im Netzwerk einen Einfluss haben. D.h., dass die Zulassung eines neuen Verbindungsaufbaus nicht nur vom Zustand einer Zelle abhängig ist, sondern abhängig vom Zustand des gesamten Netzwerkes [5].

MLST (Mobile Label Switched Tree)

Die in den vorangegangenen Kapiteln vorausgesetzte Durchgängigkeit des QoS-Steuerflusses bedingt auf der Kontrollebene des Netzwerks eine gewisse Einheitlichkeit. Die verwendeten Technologien GPRS und ATM verfügen auf dieser Schicht jedoch über verschiedene Technologien, IP im GPRS-Teil und ATM / frame relay im Backbone-Netzwerk. Als verbindende Lösung wird MPLS eingesetzt inklusive der zur Verfügung stehenden Kontrollfunktionen für traffic. Das bedeutet, dass Pakete, welche in das Netzwerk eintreten, von einem PE Router (Provider Edge Router) mit einem Label versehen werden. Verlassen die Pakete das Netzwerks, wird das Label durch einen solchen auch wieder entfernt. Hiermit können in einem paketorientierten Netzwerk gewisse Eigenschaften eines "circuit-switched" Netzwerks emuliert werden, welche für das vorgeschlagene QoS benötigt werden. Für eine detailliertere Beschreibung von MPLS sei auf die Vorlesungsfolien zu "Protocols for Multimedia Communications" verwiesen [6].

MLST ist im wesentlichen dafür zuständig, die Mobilität von Endgeräten zu ermöglichen. Mithilfe dieses Baumes, welcher sich kontinuierlich anpasst, wird die Mobilität von Benutzern innerhalb verschiedener Zellen gewährleistet unter gleichzeitiger Optimierung der Kapazitätsauslastung. Abbildung 11.6 stellt einen Ausschnitt des Netzwerks dar. Dabei sind sowohl die Funkzellen, als auch das Backbonenetzwerk eingezeichnet. In Zelle 0, 7 und 10 befinden sich zur aktuellen Zeit jeweils ein aktiver Benutzer. Für die Benutzers in Zellen 0 und 10 besteht die Voraussage eines möglichen Zellwechsels (Handoff). Dieser löst eine präventive Reservation von Ressourcen aus (gestrichelte Linien). Die ganze Gruppe der momentan durch einen Benutzer verwendeten und für mögliche Handoffs reservierten BSs (Basestations) inklusive der beteiligten Gegenstelle, bilden den erweiterten MLST (multipoint-to-point). Dabei sind die aktive und die reservierte BS über den MMP (mobile merged point) verbunden. Bei einem eintretenden Handoff kann somit die Verbindung dort "übergeben" werden.

MLST verspricht eine Optimierung in zweierlei Hinsicht. Die Verzögerung bei Handoffs wird minimiert, ohne dabei jedoch "unnötig" Ressourcen zu reservieren. Dies wird dadurch bewerkstelligt, dass bei der Initialisierung der Verbindung nur die möglichen Wege für Handoffs ermittelt werden. Bandbreite wird dabei noch keine reserviert, jedoch wird die "guard bandwith" erhöht. Die effektive Reservation von Bandbreite wird erst kurz vor dem tatsächlichen Eintreten des Handoffs vorgenommen. Bei einer Änderung der Verbindung (Initialisieren, Trennen, usw.) werden die Werte für potentielle Handoffs angepasst und über den ganzen betroffenen Link propagiert. Somit passt jeder betroffene Knoten den Wert für die zu erwartenden Handoffs an [5].

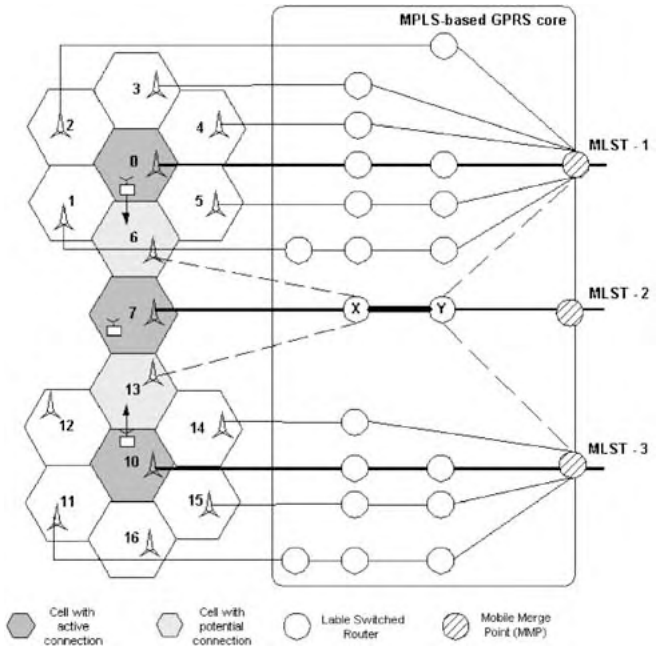


Abbildung 11.6: Schematische Darstellung von MLST [5]

Dynamic Guard Bandwith Scheme and Controller

Ein aktiver mobiler Benutzer wie zum Beispiel derjenige in Zelle 0 aus Abbildung 11.6 kann eine Anfrage für einen Handoff an eine benachbarte Zelle stellen, in welche er wahrscheinlich wechseln wird. Dabei ist die Wahrscheinlichkeit eines Requests an eine angrenzende Zellen von verschiedenen Faktoren abhängig: Mobilitätsverhalten, Zellengrösse und geschätzte verbleibende Verbindungsdauer [5].

11.5.4 Abschliessende Beurteilung und Zusammenfassung

Die Analyse des Protokolls AP-CAC zeigt, dass damit QoS Vereinbarungen auch unter der zusätzlichen Erschwernis von "mobile Handoffs" garantiert werden kann. Dies geschieht durch die Verwendung einer "adaptive priorized admission control for multiple traffic classes" mit Hilfe des beschriebenen "dynamic guard channel scheme", welches die Bandbreite für Handoffs anpasst, je nachdem wie die Belastung von benachbarten Zellen und allfällige Bewegungsmuster von Benutzern aussehen. Bezüglich der Problematik von Handoffs ist es mit dem gewählten Ansatz in den Versuchen gelungen nicht erfolgreiche Zellenwechsel (Handover) zu minimieren, und zwar in Abhängigkeit von der Priorität des betroffenen DatenTraffics. Es ist also durch die Kombination dieser Ansätze möglich ein sich dynamisch anpassendes Ende zu Ende QoS zu realisieren. Dazu wird das AP-CAC Protokoll auf das gesamte Netzwerk ausgedehnt, um das QoS kontinuierlich garantieren zu können. Ausserdem muss für ein durchgängiges QoS im ganzen Netzwerk ein Mapping der Parameter, auf das im Backbone Netzwerk verwendete IntServ, stattfinden. Die Mobilität von Benutzern auf der anderen Seite wird durch die Verwendung des beschriebenen MLST

Verfahrens hinsichtlich Handoff Verzögerung und optimierter Ressourcenverwendung verbessert [5].

11.6 Diskussion und Fazit

Alle vorgestellten Protokolle verbessern nachweislich die Unterstützung von QoS durch Veränderungen der MAC Schichten. Dabei werden je nach Ansatz verschiedene Massnahmen vorgeschlagen um das Ziel zu erreichen. Das Protokoll soll drahtlose Netzwerke effektiver ausnutzen und Unterstützung vom Realtime- und verzögerungsempfindlichen Traffic bieten.

Tabelle 11.4: Protokolle, QoS Anforderungen und MAC Parameter im Überblick

Protokoll	Garantie für Mindest/ konstante Datenrate	Garantie für limitierte Latenz	Garantie für begrenzten Jitter
802.11 DCF/PCF	Keine	Keine	Keine
802.11e EDCF	Keine, aber höhere Datenrate für höhere Prioritäten	Keine, aber geringere Latenz für höhere Prioritäten	Keine, aber geringerer Jitter für höhere Prioritäten
802.11e HCCA	Ja	Ja	Ja
DQRUMA	Kein	Keine	Keine
ARCMA	Keine, aber höhere Datenrate für höhere Prioritäten	Keine, aber geringere Latenz für höhere Prioritäten	Keine, aber geringerer Jitter für höhere Prioritäten
GPRS Standard auf PRMA	Keine	Keine	Nur für Gesprächsquellen
GPRS GQ-MAC und AP-CAC	Ja	Ja	Ja

Für den WLAN-Standard ermöglicht EDCF die Priorisierung nach Wichtigkeit und Anforderungsprofil und erlaubt statistisches QoS. In HCCA erlaubt ein neues Polling-Verfahren, die explizite Reservation des Mediums. Durch diesen Entwurf wird QoS nach TSPEC möglich, speziell zugeschnitten auf einzelne Applikationsanforderungen.

ARCMA ist konzipiert, um QoS Anforderungen auf der MAC Schicht anzubieten und die verschiedenen ATM Service-Typen zu unterstützen. ARCMA bietet Unterstützung für den verzögerungsempfindlichen CBR Traffic durch Priorisierung und erhöht die Kanal-Ausnutzung durch eine Reduktion von Kollisionen.

Bei GPRS wird ein kombinierter Lösungsansatz präsentiert, welcher neben der Priorisierung von verschiedenen Traffic-Arten auch eine Begrenzung des Zugriffs durch ein verbessertes Handoffverfahren vornimmt, um eine Überbelegung des Mediums zu verhindern.

Damit wird das effektive Funktionieren der Priorisierung soweit verbessert, dass auch harte Zusagen für die QoS Parameter gegeben werden können.

Allgemein tendieren alle Verbesserungsansätze für QoS in drahtlosen Netzwerken zur Verwendung von Priorisierung und Reservation.

11.7 Glossar

AAL	ATM Adaptation Layer
ABR	Available Bit Rate
AC	Access Categories
ACK	Acknowledgement
AIFS[AC]	Arbitration Interframe Space
AP	Access Point
AP-CAC	Adaptive Prioritized Handoff CAC
ARCMA	Adaptive Request Channel Multiple Access
ATM	Asynchronous Transfer Mode
BEB	Binary Exponential Backoff
BER	Bit Error Rate
BS	Base Station
CBR	Constant Bit Rate
CSMA/CA	Carrier Sense Multiple Access / Collision Avoidance
CSMA/CD	Carrier Sense Multiple Access / Collision Detection
CW	Contention Window
DAMA	Demand Assignment Multiple Access
DCF	Distributed Coordination Function
DD	Downstream
DIFS	DCF Inter Frame Space
DQRUMA	Distributed Queuing Request Update Multiple Access
EDCF	Enhanced Distributed Coordination Function
FDD	Frequency Division Duplex
FDM	Frequency Division Multiplexing
GPRS	General Packet Radio Service
GQ-MAC	Guaranteed QoS Media Access Control
HCCA	HCF Controlled Channel Access
HCF	Hybrid Coordination Function
IFS	Inter Frame Space
MAC	Media Access Control
MLST	Mobile Label Switched Tree
MMP	Mobile Merged Point
MS	Mobile Station
PAGCH	Paket Access Grant Channel
PC	Point Coordinator
PCF	Point Coordination Funktion
PDTCH	Paket Data Traffic Channels
PIFS	Point Interf Frame Space

PRACH	Paket Random Access Channel
PRMA	Paket Reservation Multiple Access
QoS	Quality Of Service
RA	Request Access
RAMA	Resourcce Auction Multiple Access
S-ALOHA	Slotted ALOHA
SDM	Space Division Multiplexing
SIFS	Short Inter Frame Space
TA	Transmission Access
TDM	Time Division Multiplexing
TDMA	Time Division Multiple Access
TSPEC	Traffic Specification Frames
TXOP	Transmission Opportunity
UBR	Unspecified Bit Rate
VBR	Variable Bit Rate
VOIP	Voice Over Internet Protocol

Literaturverzeichnis

[1] IEEE: 802® IEEE Standard for Local and Metropolitan Area Networks: Overview and Architecture, The Institute of Electrical and Electronics Engineers, Inc., 2002.

[2] Burkhard Stiller: Vorlesungsunterlagen Mobile Communication Systems, Universität Zürich, 2006.

[3] Xinping Guo and Colin Pattinson: Quality of Service Requirements for Multimedia Communications, School of Computing, Leeds Metropolitan University.

[4] Anna Hać and Boon Ling Chew: ARCMA?adaptive request channel multiple access protocol for wireless ATM networks, INTERNATIONAL JOURNAL OF NETWORK MANAGEMENT, 2001.

[5] OLIVER T.W. YU: End-to-End Adaptive QoS Provisioning over GPRS Wireless Mobile Network, Kluwer Academic Publishers, 2003.

[6] Burkhard Stiller: Vorlesungsunterlagen Protocols for Multimedia Communications, Universität Zürich, 2005.

[7] O.T.W. Yu and V.C.M. Leung, Adaptive resource allocation for prioritized call admission over an ATM-based wireless PCN, IEEE Journal on Selected Areas in Communications 15 (September 1997) 1208?1224.

[8] Shugong Xu: Advances in WLAN QoS for 802.1 1 : an Overview, Sharp Labs of America, 2000.

[9] Jun Zhao, Zihua Guo, Qian Zhang and Wenwu Zhu: Distributed MAC Adaptation for WLAN QoS Differentiation, Microsoft Research, Asia, 2003.

[10] Deyun Gao, Jianfei Cai, King Ngi Ngan: Admission control in IEEE 802.11e wireless LANs. IEEE Network 19(4): 6-13, 2005.

[11] Javier del Prado Pavón and Sai Shankar N: Impact of Frame Size, Number of Stations and Mobility on the Throughput Performance of IEEE 802.11e, Philips Research, USA, 2004.

[12] Hoiigqiang Zliai and Yugoang Fang: Performance of Wireless LANs Based on IEEE 802.1 1 MAC Protocols, University of Florida. Gainesville,Florida, 2003.

Kapitel 12

Mobile-Inhalteanbieter und Netzneutralität

Matthias Alder, Stéphanie Eugster, Philipp Kräutli

Netzneutralität steht für eine nicht-diskriminierende Übermittlung von Datenverkehr. Dies meint, dass unabhängig von Ausgangspunkt oder Inhalt der Datenpakete, im Internet alle Datenübermittlungen gleichbehandelt werden sollen: Gleichbehandlung aller Datenpakte im offenen Internet - unabhängig von deren Ausgangspunkt oder Inhalt. Das Thema Netzneutralität wird seit Jahren kontrovers diskutiert. Diverse Experten und Lobbyisten haben Position gegen oder für die Prinzipien eines nicht-diskriminierenden Internetzugangs eingenommen. Generell können zwei Lager identifiziert werde: Telekommunikationsoperatoren und Internet Service Provider (ISP) versus die Endkunden und Content Provider. Im Rahmen dieser Arbeit wird der Begriff Netzneutralität erklärt und mit Fallbeispielen genauer erläutert. Anschliessend wird der aktuelle Stand der Netzneutralität auf nationaler, europäischer und internationaler Ebene untersucht. Danach werden Interessenskonflikte, Auswirkungen und mögliche regulatorische Massnahmen von Netzneutralität, sowie potentielle Abweichungen vom status quo auf Mobile Content Provider und deren Geschäft untersucht. Da der Fokus auf Mobile Content Provider gesetzt wird, liegt folglich der zentrale Punkt bei den ISPs und Telekommunikationsoperatoren, welche Mobile- oder Nomadic-Kunden Internetzugang anbieten.

Inhaltsverzeichnis

12.1 Einführung und Motivation

Dieser Abschnitt führt in das Thema Netzneutralität ein und es werden die wichtigsten Begriffe erklärt. Anhand von Fallbeispielen wird die Problematik vertieft behandelt.

12.1.1 Einleitung

Der Begriff Netzneutralität steht für eine nicht-diskriminierende Übermittlung von Datenverkehr. So sollen unabhängig von Ausgangspunkt oder Inhalt der Datenpakete, im Internet alle Datenübermittlungen gleichbehandelt werden; Gleichbehandlung aller Datenpakete im offenen Internet, unabhängig von Ausgangspunkt oder Inhalt.

Netzneutralität wird seit Jahren kontrovers diskutiert. Diverse Experten und Lobbyisten haben Position gegen oder für die Prinzipien eines nicht-diskriminierenden Internetzugangs eingenommen. Generell sind zwei sich konkurrierende Lager identifizierbar: Telekommunikationsoperatoren und Internet Service Provider (ISP) versus die Endkunden und Content Provider.

In dieser Arbeit wird grossen Wert auf einen kritischen und objektiven Ansatz gelegt, da das Thema Netzneutralität stark von subjektiven und emotionalen Äusserungen und Stellungnahmen geprägt ist.

12.1.2 Der Begriff Netzneutralität

Bis anhin galt der Grundsatz, dass alle Betreiber von Telefonnetzen (Carrier) alle Datenpakete gleichmässig weiterzuleiten und deren Inhalt nicht zu öffnen und anzusehen haben. Nun wird unter dem Begriff Netzneutralität die Frage diskutiert, ob die grossen Carrier beim Weiterleiten von Datenpaketen nach deren Herkunft oder Inhalt unterscheiden und anhand davon verschiedene Gebühren erheben dürfen. So könnten beispielsweise die Gebühren für Voice over IP oder für Datenverkehr von Suchmaschinen anders ausfallen als die für File Transfer Protocol (ftp) oder für Universitäten [1].

Die Aufhebung der Netzneutralität wird durch unzufriedene Carrier forciert. Diese stellen eine aufwendige Infrastruktur bereit, die Content Providern, wie beispielsweise Google, sehr grosse Umsätze ermöglichen [1].

Über die letzten Jahre hinweg, wurde der Sachverhalt von Netzneutralität-Regulationen ein heiss diskutiertes Thema in telekommunikationspolitischen Debatten. Befürworter der Netzneutralität haben wiederholt die Federal Communications Commission angehalten, Regeln auszuarbeiten, die den Operatoren von Breitband-Netzzugängen verbieten, Drittpartei-Applikationen, -Inhalt oder -Portale (“unabhängige Applikationen”) zu diskriminieren und sie von ihrem Netzwerk auszuschliessen. Dieses Anliegen basiert auf den Befürchtungen, dass durch Fehlen einer solchen Regulation, Netzwerk-Operatoren solche

Produkte diskriminieren können und dadurch die Provider den Anreiz verlieren, die Innovation der betroffenen Produkte weiter zu fördern, was wiederum sich schädlich auf die Gesellschaft auswirken kann.

Regulations-Gegner dementieren die Notwendigkeit für eine Netzneutralität-Regulation. Sie argumentieren, dass eine Regulation nicht von Nöten sei, da Netzwerk-Operatoren keinen Antrieb hätten, die unabhängigen Applikationen zu diskriminieren. Im Gegenteil, Regulationen können sogar schädlich sein, da sie die Anreize der Netzwerk-Operatoren reduzieren, ihre Netzwerke in Zukunft weiter auszubauen oder zu verbessern [2].

Über den exakten Umfang von Netzneutralität-Regeln wird noch immer verhandelt, das allgemeine Grundprinzip hinter den verschiedenen Vorschlägen liegt jedoch darin, ein Regelwerk auszuarbeiten, dass es Netzwerk-Operatoren und ISPs verbietet, ihre Macht über die Übertragungstechnologie so auszunutzen, dass es dem Wettbewerb im komplementären Markt von Applikationen, Inhalten und Portalen schadet [2]. So ist es beispielsweise noch immer eine offene Frage, ob Netzneutralität-Regeln Preisdiskriminierungen verhindern sollen oder nicht [3]. Die Regulation von Netzneutralität beabsichtigt es aber auch nicht, vertikale Integrationen zwischen Netzwerkprovidern und Applikationsprovidern zu verhindern, so ist es z.B. Netzwerkprovider möglich, auch Applikationen anzubieten [4].

Veranschaulichung der Problematik [5]

Netzneutralität ist ein Prinzip, das besagt, dass diejenigen die ein Netzwerk unterhalten, welches der breiten Öffentlichkeit einen umfassenden Nutzen stiftet und auf öffentlichem Eigentum basiert, sollten ihre Besitzermacht gegenüber ihren Kunden nicht diskriminierend nutzen können.

Zur Veranschaulichung ein Vergleich mit Schnellstrassen: Es existiert für Strassenfahrzeuge, welche keine Offroad-Vehikel sind, von Ort A nach Ort B nur eine einzige Schnellstrasse. Es ist das Recht des Besitzers dieser Strasse, Mautgebühr für die Nutzung der Strasse verlangen zu können. Auch ist es dessen Recht, die Mautgebühr nutzerabhängig oder fahrspurabhängig zu gestalten. So kann beispielsweise für die Nutzung der gut unterhaltenen Hochgeschwindigkeitsfahrspur mehr Mautgebühr verlangt werden, als für die weniger gut unterhaltene Normalgeschwindigkeitsfahrspur. Nach diesem Prinzip ist gegenwärtig das US-amerikanische Breitbandnetzwerk gestaltet. Das Szenario könnte nun soweit gehen, dass der Besitzer der Schnellstrasse von einem Fahrzeughersteller gegen eine Gebühr nur noch Fahrzeug desselben die Schnellstrasse benützen lässt. Die anderen Kunden werden durch diesen ungleichen Zugang zu dem Schnellstrassen-Service diskriminiert. Das Neutralitätsprinzip ist nicht mehr gewährleistet. Dies ist die momentane Situation in den Vereinigten Staaten, so argumentieren Netzneutralität-Befürworter, dies sei was die US-Telekommunikations- und Cable-Unternehmungen wollen und was durch gewisse Politiker unterstützt wird.

12.1.3 Fallbeispiele

Nachfolgend werden einige konkrete Fallbeispiele betrachtet, die sich besonders im Thema Netzneutralität engagieren oder besonders davon betroffen sind.

eBay

In diversen Blogs wird das Thema Netzneutralität heiss diskutiert. So wird beispielsweise auch auf twoday.net darauf hingewiesen, dass die weltweite Marktplatz-Plattform eBay an seine sechs Millionen US-Mitglieder E-Mails gesandt hat, in denen die Mitglieder um Unterstützung im Kampf für die Netzneutralität gebeten wurden. Gemäss twoday.net argumentierte eBay CEO Meg Whitman: "Konsumenten, Organisationen und Firmen bezahlen ja bereits für den Zugang ins Internet. Den Breitband-Anbietern sollte deshalb nicht erlaubt werden, die Kunden doppelt abzukassieren." [6] Im folgenden wird auf klare Stellungsnahmen auf den Websites von eBay zum Thema Netzneutralität eingegangen.

Newsletter eBays.com's Stellungnahme zur Netzneutralität in ihrem Newsletter an die eBay Mitglieder vom Sommer 2006 [7]: Net Neutrality:

> What is it and why does the Community care? From Meg Whitman's eBay Live! keynote address to Jon Stewart's Daily Show, it seems like everyone is talking about network neutrality. If you think the issue sounds confusing, don't worry, you're not alone. In short, the phone and cable companies are using their political muscle to promote legislation in Washington that would allow them to divide the Internet into a two-tiered system, a 'Pay to Play' tier for large companies that can afford the fees and a slow lane for everyone else. Congress needs to act to preserve the Internet as we know it during the next generation of faster and better broadband. Congress has to re-establish basic safeguards that require broadband providers to treat all Internet traffic in a non-discriminatory manner; and by prohibiting tiering schemes that impose additional fees upon website owners to 'deliver' their broadband content to you on top of the fees you already pay to connect to the Internet. Your collective voice as part of the eBay community of millions of average citizens is making a difference. If the giant telephone and cable companies get their way, it will affect your freedom on the Internet. They will use their power to dictate your content, steering you only to those services they own or have exclusive deals with. You could end up paying more or not be able to use non-preferred sites and services. And for sellers, you might have to pay premium prices to make your content and products available on the preferred network. Meg has reached out to millions of eBay members, encouraging them to write their Members of Congress in support of net neutrality. If you have not yet weighed-in with your Members of Congress, please click here to join us today.

eBay Member [8] eBay's Government Relations Team lädt dazu ein, dem eBay Main Street Member Programm beizutreten. Die Mitglieder dieses Programms werden von eBay in Sachen Regierungsregulierungen, welche das Nutzen und Benutzen von eBay-Handel berühren, auf dem Laufenden gehalten. eBay hat eine Kampagne ins

Leben gerufen, die Mitglieder und noch Nicht-Mitglieder dieses Programms auffordert, sich im Kampf für die Netzneutralität zu beteiligen und sich an ihren Abgeordneten zu wenden und gegen die Pläne der US-Telekommunikationsanbieter zu protestieren:

Protect the internet by making your voice heard. It is hard to believe, but lawmakers in Washington are debating whether consumers should be free to use the Internet as they want in the future. The phone and cable companies are using their political muscle to promote legislation that would divide the Internet into a two-tiered system, a Pay to Play tier for large companies that can afford the fees and a slow lane for everyone else. Join eBay CEO Meg Whitman and millions of other Internet users and write your U.S. Senators and U.S. Representative today! You may receive an email from Meg Whitman, depending on your geographic location, asking you to contact Congress in support of network neutrality. If you received the email directly from Meg, you should also see a copy in your My Messages account on eBay.com. If you were directed to this campaign from a forwarded email or another website, we welcome your participation as well! Please log in with the following if you are a Main Street Member or sign up now to become a Main Street Member.

Net Neutrality [9] Auf dieser speziell für die Kampagne für den Kampf für die Erhaltung der Netzneutralität erstellten Website, erklärt eBay ihren Standpunkt. Es wird auf den Hintergrund des Internets eingegangen, und der Grundgedanke der Netzneutralität erläutert: The Internet has always been governed by a regulatory regime based on principles of openness and non-discrimination. This approach has been integral to making the Internet the home to the most innovative and exciting new businesses and ideas. Internet companies have spent billions on new content and services that have transformed American life. This investment has fueled the American economy.

Es wird des Weiteren auf die Folgen von einem Einbruch der Netzneutralität vor Allem aus Kunden- und Inhalteanbieter-Sicht eingegangen. Es wird darauf hingewiesen, dass Konsumenten bereits für den Internetzugang bezahlen, und nun Netzwerkoperatoren den von Konsumenten besuchten Inhalt ebenfalls verrechnen wollen. Ebenso wird auf ökonomischen Folgen hingewiesen: [9]

Fragmenting the Market - The Internet is a global network based on the principle of openness, potentially connecting everyone with everyone. As we have seen with eBay, PayPal, and Skype, the Internet has the power to create communities on a scale never seen before. Replacing the Internet with technologically advanced but closed private networks will end the Internet as we know it and reduce the ability of Internet users to reach a global market. Small business sellers rely on that global community and could be hardest hit by new fees and tiered services that impede existing and potential customers from accessing their sites.

CEO Meg Whitman und eBay, sonst kaum dafür bekannt, sich öffentlich in aktuelle politische Diskussionen einzumischen, schalten sich mit ihrem Newsletter (Brandmail) massiv in den gegenwärtig vor allem in den USA geführten Kampf für ein netzneutrales World Wide Web ein.[10] Dass sich eBay öffentlich per Lobbying in die Diskussion um Netzneutralität einmischt, zeigt, wie virulent dieses Problem derzeit ist. “Meg meinte, dass

es an der Zeit wäre, die ebay-Community einzuschalten, um sicher zu gehen, dass der Kongress hört, was die Community von den Gesetzen zur Reform der Telekommunikation hält", begründete Tod Cohen, eBay Vice President für Regierungsangelegenheiten, die ungewöhnliche Aktion seiner Chefin.[10]

Google

Bei Google hat die Netzneutralität ähnlich wie bei eBay oberste Priorität. Verschiedentlich ist zu erfahren, dass TCP/IP-Miterfinder Vint Cerf, Vize-Präsident und Chief Internet-Evangelist bei Google sich auf einer Konferenz zu dem Thema Netzneutralität geäussert hatte. Sollte die Netzneutralität nicht mehr gegeben sein, so würde sein Unternehmen es "notfalls einklagen", so Cerf [11].

Auch Google hat eine Website eingerichtet, auf der sie die Kunden über das Thema Netzneutralität informiert. Der Begriff wird erklärt und der aktuelle Status wird ebenfalls erläutert. Es wird auch auf verschiedene fremde Websites verwiesen, welche sich diesem Thema widmen. Die Besucher von dieser Website werden dazu aufgefordert, sich ebenfalls bei ihren Repräsentanten zu melden und ihre Stimme für die Beibehaltung der Netzneutralität zu geben [12].

Am 7. Februar 2006 hat Vinton G. Cerf, Vice President und Chief Internet Evangelist von Google Inc., einen 8 seitigen, offenen Brief an das U.S. Senate Committe on Commerce geschrieben [13]. Zusammenfassend das Statement von Vinton G. Cerf: "Allowing broadband carriers to control what people see and do online would fundamentally undermine the principles that have made the Internet such a success... A number of justifications have been created to support carrier control over consumer choices online; none stand up to scrutiny."

Auch von Tim Berners-Lee, dem Erfinder des World Wide Webs, ist ein Statement auf der Google Website publiziert: "The neutral communications medium is essential to our society. It is the basis of a fair competitive market economy. It is the basis of democracy, by which a community should decide what to do. It is the basis of science, by which humankind should decide what is true. Let us protect the neutrality of the net."

SaveTheInternet [14]

Im November 2006 demonstrierte die Koalition `SaveTheInternet` [15] gegen ein Gesetz, welches Carrier nicht auf strikte Netzneutralität verpflichtet. Hinter der Vereinigung Save the Internet steht eine bunte Gruppe zivilgesellschaftlicher Organisationen aller politischen Richtungen. Diese Organisation macht sich gemeinsam mit Internetgrössen wie eBay, Google, Microsoft, Amazon.com und Yahoo für eine strenge gesetzliche Netzneutralitätsregeln stark. Google war auch an dieser Demonstration präsent: So Andrew McLaughlin, Senior Policy Counsel: "Man hätte es nie geschafft, mehr als eine Garagenfirma zu werden, hätten die Internetanbieter den individuellen Zugang zu Google blockieren oder langsamer machen können. Die SavetheInternet-Initiative überreichte den Senatoren insgesamt 18.000 Unterschriften für die Einfügung einer Netzneutralitätsregel."

Die Allianz zur "Rettung des Internet" hat derweil 1,1 Millionen Unterschriften für ihr Anliegen gesammelt und wirbt auch in fast 200 Videos auf YouTube für ein offenes Breitbandnetz [16]. In der Abbildung 12.1 ist die Entstehung, der Verlauf und die Zielsetzung der Save the Internet Koallition dargestellt.

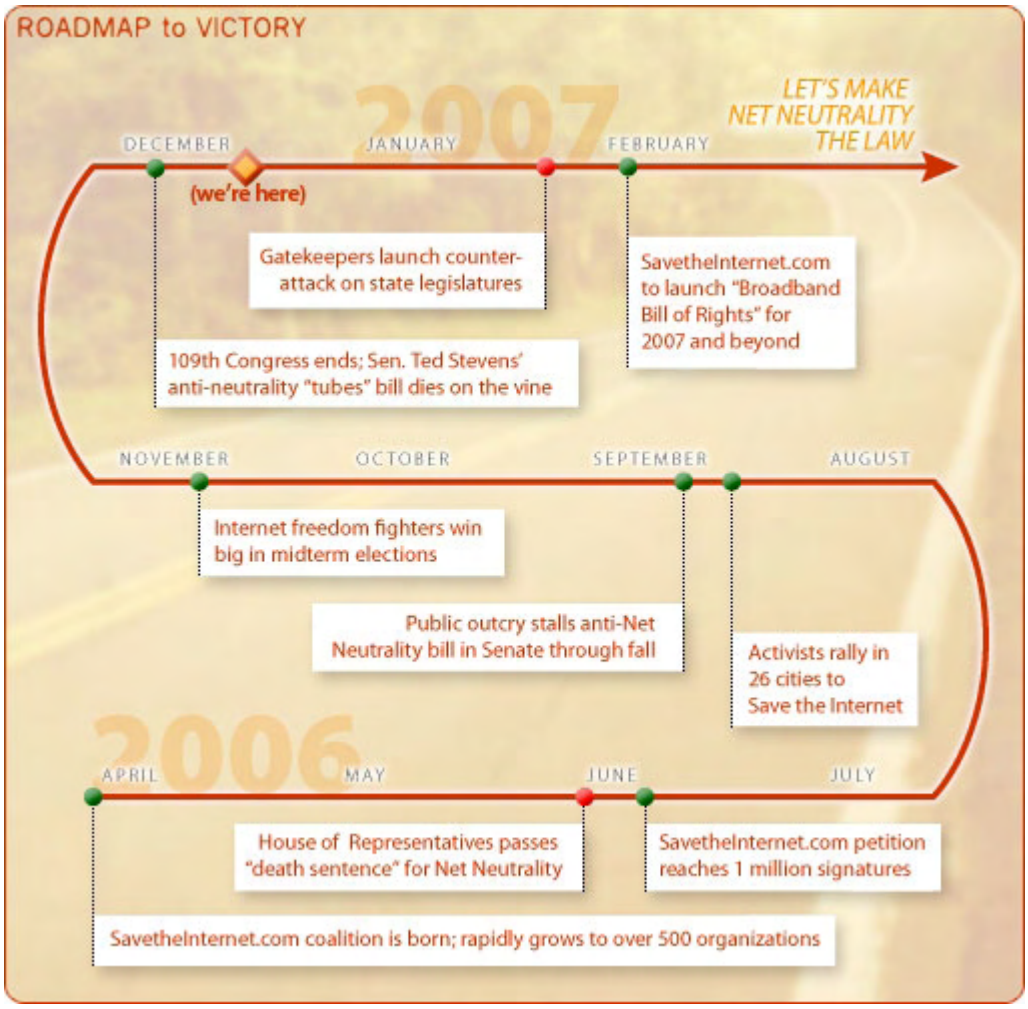


Abbildung 12.1: Roadmap ab Start der Koallition SavetheInternet.com [15]

Objektive Stellungsnahmen von Gegnern der Netzneutralität wie beispielsweise Telekommunikationsunternehmungen, wie AT&T, sind schwierig auszumachen. Einzig die Stellungsnahmen von AT&T Chairman und CEO Edward Whitacre sowie von AT&T Vicepresident Jim Cicconi, in einem Interview vom 21. März 2006 zeigen die Haltung von AT&T gegenüber der Netzneutralität. Edward Whitacre meint, dass jeder Provider, der den Internetzugang blockiert, seine Kunden so direkt in die Arme von der Konkurrenz führt, und dass AT&T Niemandem den öffentlichen Zugang zum Internet blockieren noch einschränken wird [27].

In diesen Abschnitten wurde der Begriff Netzneutralität erläutert und umschrieben und es wurde auf aktuelle Fallbeispiele eingegangen. Im anschliessenden Kapitel wird weiter die Netzneutralität im nationalen und internationalen Kontext, speziell der Stand in der EU, behandelt.

12.2 Netzneutralität im nationalen und internationalen Kontext

Die Debatten und Diskussionen rund um das Thema der Netzneutralität haben ihren Höhepunkt noch längst nicht erreicht. Ganz im Gegenteil: wärend sich der allergrösste Teil der Internetbenutzer überhaupt nicht bewusst ist, welche ökonomischen Anreize die neuesten technologischen Errungenschaften und die Entwicklungen des Marktes und der Benutzerverhalten für die Netzwerkbetreiber geschaffen haben, wurden auf politischer Ebene erst vor wenigen Monaten die ersten Entwürfe für die Sicherstellung der Netzneutraliät in Form von Gesetzen und Reglementen erstellt. Es scheint also nur eine Frage der Zeit, bis auch die ganz grosse Masse der Endbenutzer mit diesem Thema konfrontiert wird. Wobei man gespannt sein darf, ob die Diskussion in der breiten Öffentlichkeit durch eine vermehrte Berichterstattung in den Medien oder durch direkte Auswirkungen für die Benutzer initiiert wird.

12.2.1 Situation in der Schweiz

Erste Anzeichen dafür, dass auch Internetanwender in der Schweiz in naher Zukunft mit Einschränkungen rechnen müssen gibt es bereits heute. Beispielsweise hat der Mobilfunkanbieter Vodafone angekündigt, VOIP-Telefonie über ihr Datennetz in Zukunft mit technischen Mitteln zu unterbinden, um mit VOIP-Produkten von Fremdanbietern nicht ihr eigenes Telefonnetz zu konkurrenzieren [17]. Auch der schweizer Datennetzbetreiber Cablecom sucht nach neuen Möglichkeiten, um möglichst alle Profiteure des Datennetzes zur Kasse zu bitten [29]. Nebst den Endanwenderen sollen vermehrt auch fremde Internet Service Provider für den Datenverkehr bezahlen, welchen sie verursachen. Geht ein Provider nicht auf diese Forderung ein, werden die direkten Verbindungen gekappt. Die leidtragenden sind einmal mehr die Endkunden, da der Datenverkehr in diesen Fällen meist über eine alternative Verbindung mit sehr vielen Zwischenstationen umgeleitet werden muss, was die Datenübertragungsraten und Antwortzeiten negativ beeinflusst.

12.2.2 Auswirkungen auf die Content Provider

Nicht nur die Endanwender, insbesondere auch die kleineren Anbieter von Diensten und Inhalten werden unter dieser Entwicklung in Zukunft leiden. Gewisse Szenarien prognostizieren die Zahlungspflicht für sämtliche Inhalte- und Dienstanbieter für den Datenverkehr, welchen sie im gesamten Netzwerk verursachen. Während diese Forderungen Kleinstanbieter in den Ruin treiben könnten, schaffen sie gleichzeitig Anreize für grössere Anbieter wie Google, Yahoo etc. eigene Netze aufzubauen oder mit nur vereinzelten Internet Service Providern Verträge einzugehen. Dies könnte im Extremfall zu einer Aufsplittung des gesamten Internets führen [30].

12.2.3 Das Internet: Ein neutrales Netzwerk?

Das Internet gilt als offenes, unreguliertes und globales Netzwerk. Damit sind aber längst nicht die Voraussetzungen für ein neutrales Netzwerk gegeben, in welchem sämtliche Teilnehmer gleichberechtigt sind. Die Tatsache, dass zur Zeit die meisten Internet Service Provider die Enkunden kaum einschränken und den Datenverkehr nicht in Abhängigkeit von dessen Zweck klassifizieren und unterschiedlich verarbeiten, könnte sich also schon bald ändern. Zum einen sind die Kosten für eine derartige Auswertung des Datenverkehrs enorm gesunken, zum anderen wird die Abhängigkeit der Endbenutzer von internetbasierten Diensten immer grösser, was deren Zahlungsbereitschaft ansteigen lässt. Da ein grosser Teil der Infrastruktur des Internets in den Händen von privaten Unternehmungen ist, kann davon ausgegangen werden, dass die Entscheidungen betreffend der Einführung von Mechanismen, welche die Netzneutralität beeinflussen, aufgrund rein ökonomischer Überlegungen gefällt werden [37].

12.2.4 Öffentliche Netzwerke

Nebst den Netzwerkinfrastrukturen welche von privatwirtschaftlichen Unternehmungen unterhalten werden, gibt es auch Universitäten und andere öffentliche Stellen, welche frei zugängliche Netzwerke betreiben, und somit einen Teil des Internets bilden. Die Interessen und Ziele dieser Institutionen unterscheiden sich meist grundlegend von deren der Privatwirtschaft. Unternehmungen wie Internet Service Provider versuchen hauptsächlich den Gewinn, welchen sie mit ihren Dienstleisunten erwirtschaften, zu maximieren, während bei universitären Netzwerken die Zurverfügungstellung von wissenschaftlichen Daten und Informationen im Mittelpunkt steht. Aufgrund der Privatisierung der meisten universitären Netzwerke anfang der 90er Jahre sind aber heute die Hauptschlagadern des Internets in privaten Händen [35][30].

So viele, hauptsächlich finanziellen, Vorteile die Privatisierung dieser Dienstleister auch mitsichbringen mag, bezüglich der Netzneutralität war diese mit Bestimmtheit ein Rückschritt. So bleibt zu hoffen, dass der Trend zum Aufbau von staatlichen oder städtischen Netzwerken weiter anhält, um den gewinnmaximierenden Unternehmungen einen Gegenpol zu setzen. Auch wenn beispielsweise im Rahmen der aktuellen Diskussionen um ein Stadtzürcher Hochleistungsdatennetz das Stichwort Netzneutralität noch kaum genannt wurde, würden die Benutzer in Zukunft diesbezüglich wohl stark profitieren können [38].

12.2.5 Drahtlose Netzwerke

Deutlich stärker als bei den drahtgebundenen Netzwerken scheint die Netzneutralität aktuell bei Mobilfunknetzen gefährdet zu sein. Während die Preise, und damit auch die Margen der Anbieter, für die Benutzung von drahtgebundenen Netzwerken und entsprechenden Diensten bereits auf ein sehr tiefes Niveau gesunken sind, versuchen die Betreiber von drahtlosen Netzwerken die Preise so lange wie möglich auf einem hohen Niveau halten zu können. Zu diesem Zweck werden nun vermehrt spezifische Dienste mit technischen Mitteln unterbunden [17].

12.2.6 Stand in Europa

Mit etwas Verzögerung zu den USA wird seit einigen Monaten auch innerhalb der EU-Kommission zum Thema Netzneutralität debattiert. Rechte und Pflichten betreffend der Netzneutralität im weiteren Sinne werden innerhalb der EU in der Rahmenrichtlinie zu elektronischen Kommunikationsnetzen und -diensten geregelt. Dieser Rechtsrahmen, welcher im Frühling 2002 als Teil des EU-Richtlinienpaketes, in Kraft getreten ist betrifft nicht nur die Telekom-Netze im engeren Sinn, sondern insbesondere auch Rundfunknetze und Rundfunkdienste. Die Richtlinie ist darauf ausgelegt, die Entwicklung des für den Wettbewerb geöffneten Telekommunikationsmarktes zu regulieren. Dabei wird die Stärkung des Wettbewerbs durch die Förderung von Investitionen sowie der Erleichterung der Firmengründungen in diesem Sektor als Hauptziel genannt [35].

EU-Richtlinien [32][33][36]

Aufgrund der rasanten Entwicklung dieser Branche sah sich die EU-Kommission veranlasst, die Rahmenrichtlinie zu elektronischen Kommunikationsnetzen und -diensten zu überprüfen. Zu diesem Zweck wurde im Juni 2006 ein Arbeitspapier veröffentlicht, welches unter anderem auch die zweifelhafte Sicherstellung der Netzneutralität analysiert. Es wird festgehalten, dass die aktuelle Entwicklung kein Anlass zur Veränderung der gesetzlichen Rahmenbedingungen gebe. Angeblich werden sich auf einem funktionierenden Markt immer Anbieter finden, welche die Benutzung der Netzwerkdienste möglichst ohne Einschränkungen anbieten. Es wird also davon ausgegangen, dass die Kunden mit ihrem Recht, den passenden Anbieter für den Netzwerkzugang selbst wählen zu können, selbst für ein entsprechendes Angebot bei den Dienstleistern sorgen. Es kann davon ausgegangen werden, dass sich diese prognostizierte Entwicklung zumindest mittelfristig bei den drahtgebundenen Netzwerken bewahrheiten wird.

Es stellt sich aber die Frage, ob dieser Mechanismus auch bei mobilen Netzwerksystemen die Netzneutralität zufriedenstellend wahren wird. Entsprechende Dienste sind noch längst nicht auf einem Entwicklungsstand, welcher sich mit jenem Dienste drahtgebundener Netze messen könnte. Dementsprechend klein ist das Bewusstsein seitens der Endbenutzer bezüglich der zu erwartenden Dienstgüte, was den Anbietern die Möglichkeit einräumt, den Spielraum, welchen die technischen Möglichkeiten zur nichtneutralen Datenverarbeitung bieten, grösstenteils auszuschöpfen. Insbesondere die klassische Mobiltelefonie generiert bei den Anbietern noch immer einen enorm grossen Umsatz. Die Mobilfunkanbieter begründen die hohen Verbindungspreise mit dem enormen Investionsvolumen, welches für die Erstellung der Infrastruktur notwendig war, sowie mit den hohen Unterhaltskosten. Am Beispiel gewisser anderer Länder zeigt sich aber, dass Mobilfunknetzwerke durachaus auch mit tieferen Verbindungskosten erfolgreich betrieben werden können. Aufgrund dieser Tatsache stellt sich also die Frage, ob die Privatisierung der Mobilfunkunternehmen tatsächlich über jeden Zweifel erhaben ist.

Triple Play [34][31][28]

Der neuste Trend zeigt, dass die Netzwerkbetreiber immer mehr auch die Rolle eines Anbieters für Dienste, welche auf diesen Netzen angeboten werden, übernehmen. In gewissen Bereichen kann diese Entwicklung als äusserst bedenkenswert im Bezug auf die Netzneutralität betrachtet werden. Die Anbieter sind versucht, ihre eigenen Dienste gegenüber jenen der Fremdanbieter netzwerktechnisch zu priorisieren. Dadurch kann die Dienstqualität der eigenen Angebote gegenüber den Fremddiensten massiv gesteigert werden, was den Netzwerkbetreibern einen entscheidenden Vorteil verschafft. Das Begriff "Triple Play" macht zur Zeit in fast allen europäischen Ländern Schlagzeilen. Die Dienste Telefonie, Fernsehen und Internet sollen den Endanwendern über nur eine einzige Breitbandleitung zur Verfügung gestellt werden. Auch hier stellt sich die Frage, ob es auch in einigen Jahren noch möglich sein wird, nur den Internetanschluss ansich von einem Anbieter zu beziehen, um die Dienste Telefonie und Fernsehen von anderen Anbietern über dieses Netzwerk in Anspruch zu nehmen. Es ist anzunehmen, dass die Netzwerkbetreiber solche Konstellationen so weit wie möglich zur verhindern versuchen werden.

Die europäischen Behörden halten ganz klar fest, dass es den Unternehmungen bereits heute erlaubt ist, unterschiedlichen Kundengruppen unterschiedliche Dienste anzubieten. Kunden in vergleichbaren Situationen hingegen müssen in Anlehnung an die aktuelle Gesetzgebung gleich behandelt werden. Das unmittelbare Blockieren von Diensten kann auf der Grundlage der Zugangs- und Verbindungsregeln verhindert werden. Schwierig zu beurteilen sind jedoch Verfahren, bei welchen Datenpakete aufgrund bestimmter Kriterien mit einer höheren oder tieferen Priorität behandelt werden. Die EU-Kommission sieht daher für die Zukunft vor, Mindestanforderungen an die Dienstgüte zu stellen, welche von den Anbietern eingehalten werden müssen. Man darf gespannt sein, ob sich diese Forderungen in der Praxis tatsächlich durchsetzen lassen. Insbesondere die Überwachung der zahlreichen Angebote der Service Provider dürfte sich als äusserst schwierig herausstellen.

12.2.7 Stand im Internationalen Umfeld

Die amerikanische Federal Communications Commission schlägt folgende vier Prinzipien als Leitsätze vor [24]:

1. *To encourage broadband deployment and preserve and promote the open and interconnected nature of the public Internet*, consumers are entitled to access the lawful Internet content of their choice.

2. *To encourage broadband deployment and preserve and promote the open and interconnected nature of the public Internet*, consumers are entitled to run applications and use services of their choice, subject to the needs of law enforcement.

3. *To encourage broadband deployment and preserve and promote the open and interconnected nature of the public Internet*, consumers are entitled to connect their choice of legal devices that do not harm the network.

4. *To encourage broadband deployment and preserve and promote the open and interconnected nature of the public Internet*, consumers are entitled to competition among network providers, application and service providers, and content providers.

Nationalen Regulierer die Netzbetreiber allenfalls gestützt auf Artikel 5 Absatz 1 der Zugangsrichtlinie zur Netzneutralität verpflichten. Auf diesem Weg wäre aber wohl nur diskriminierungsfreier Zugang in dem Sinne zu erhalten, dass Exklusivzugang unzulässig wäre, nicht aber ein kostenloser Zugang verlangt werden könnte [22].

Einige Fälle aus den USA, wo der Wettbewerb die Netzneutralität nicht gewährleisten konnte und die Federal Communications Commission eingegriffen hat, finden sich u.a. auf der Webseite savetheinternet.com [25]:

1. In 2004, North Carolina ISP Madison River blocked their DSL customers from using any rival Web-based phone service.
2. In 2005, Canada's telephone giant Telus blocked customers from visiting a Web site sympathetic to the Telecommunications Workers Union during a contentious labor dispute.
3. Shaw, a major Canadian cable, internet, and telephone service company, intentionally downgrades the "quality and reliability" of competing Internet-phone services that their customers might choose -- driving customers to their own phone services not through better services, but by rigging the marketplace.
4. In April, Time Warner's AOL blocked all emails that mentioned `http://www.dearaol.com` – an advocacy campaign opposing the company's pay-to-send e-mail scheme.

Eine weitere Möglichkeit welche von der EU-Kommission in einem Arbeitspapier beschrieben wird [26] wäre, dass die nationalen Regulierungsbehörden Mindestanforderungen an die Dienstgüte festlegen sollen, damit aus Kundensicht ein Mindestmass an Übertragungskapazität eingefordert werden könnte. Die EU-Kommission geht davon aus, dass ausreichender Wettbewerb zwischen den Netzwerkbetreibern die Netzneutralität von allein gewährleisten wird. Sollte dies wider Erwarten nicht der Fall sein, könnten die nationalen Regulierer die Netzbetreiber allenfalls gestützt auf Artikel 5 Absatz 1 der Zugangsrichtlinie zur Netzneutralität verpflichten. Auf diesem Weg wäre aber wohl nur diskriminierungsfreier Zugang in dem Sinne zu erhalten, dass Exklusivzugang unzulässig wäre, nicht aber ein kostenloser Zugang verlangt werden könnte [22].

12.2.8 Diskurs der Verschiedenen Ansätze zur Sicherstellung der Netzneutralität

So vielfältig wie die unterschiedlichen Interessen der an den Netzwerken beteiligten Parteien, so vielfältig sind auch die Anforderungen an die Mechanismen, welche die Netzneutralität sicherstellen sollen. Je nach Sichtweise unterscheiden sich die Prioritäten der

verschiedenen Dienste. Einen Ansatz für eine neutrale und gerechte Verteilung der zur verfügungstehenden Bandbreiten, welchen sämtlichen Anforderungen gerecht wird, gibt es wohl nicht [37], [2].

Vom ursprünglichen Internet, welches als reine Austauschplattform für wissenschaftliche Dokumente gedacht war ist bis heute nicht mehr viel übrig geblieben. Das heutige Internet fungiert als Plattform für den Austausch von riesigen Datenmengen, als Schnittstelle für den Austausch von hochvertraulichen Daten wie beisielsweise beim E-Banking, als Netzwerk für die Verbreitung von Radio- und Fernsehkanälen oder sogar als Basis für Dienstleistungen wie Applications Service Providing. Diese Vielfalt von Diensten erfordert eine Klassifizierung und Priorisierung der entsprechenden Datenpakete. Wo genau die Grenze zwischen Mechanismen zum Zweck des QoS und Mechanismen zu Zwecken, welche die Netzneutralität in Fragestellen, liegt, kann nicht eindeutig beantwortet werden.

Anspruchsvolle Datenströme limitieren

Die grosse Masse der Internet Service Provider scheint zur Zeit zwei Methoden zu kennen um die zur Verfügungstehende Bandbreite unter den Benutzer aufzuteilen. Grösstenteils werden überhaupt keine Vorkehrungen getroffen. Das heisst jeder Benutzer darf sich innerhalb gewisser Grenzen frei bewegen. Dabei kann es durchaus vorkommen, dass eine kleine Anzahl von sogenannten Powerusern einen sehr grossen Anteil der Gesamtkapazität für sich beanspruchen, und somit für schlechte Antwortzeiten und geringe Bandbreiten der anderen Benutzer verantwortlicht sind. Diese Problematik besteht allerdings erst seit der Etablierung von Internetangeboten im Bereich des Filesharings oder Plattformen wie youtube. Aus diesem Grund sahen sich gewisse Anbieter gezwungen, gewisse Einschränkungen vorzunehmen. Der am häufigsten zum Einsatz kommende Ansatz ist die Limitierung der Bandbreiten, welche gewisse Applikationen beanspruchen. Dabei handelt sich in erster Linie um Filesharing-Applikationen. Bereits gibt es jedoch diverse Methoden, um diese Beschränkungen zu umgehen. Dies veranlasst die Netzwerkbetreiber, nach neuen Ansätzen zu suchen, um den Datenverkehr zwecks Bandbreitenbeschränkung und Priorisierung zu klassifizieren.

Fair Use Networking: Netzneutralität dank verbessertem QoS

Das Ziel dieses Verfahrens ist, die Bedürfnisse möglichst aller beteiligten Parteien am internetbasierten Datenaustausch zu berücksichtigen. Auf umstrittene Verfahren wie dem direkten Port-Blocking, welches die Benutzer sehr stark einschränkt, soll verzichtet werden. Weiter soll dem Datenschutz eine sehr hohe Priorität eingeräumt werden. Datenpakete sollen nur anhand der standardisierten Header der Protokolle klassifiziert werden. Auf die sogenannte Deep Packet Inspection soll verzichtet werden. Die so gewonnenen Informationen bezüglich der Klassifizierung der Datenströme werden für die Anwendung der folgenden Policies verwendet:

1. Proportionale Verteilung der verfügbaren Kapazität auf alle aktiven Benutzer

2. Priorisierung der Datenströme, welche für kritische Dienste verwendet werden
3. Anwendung von QoS Policies

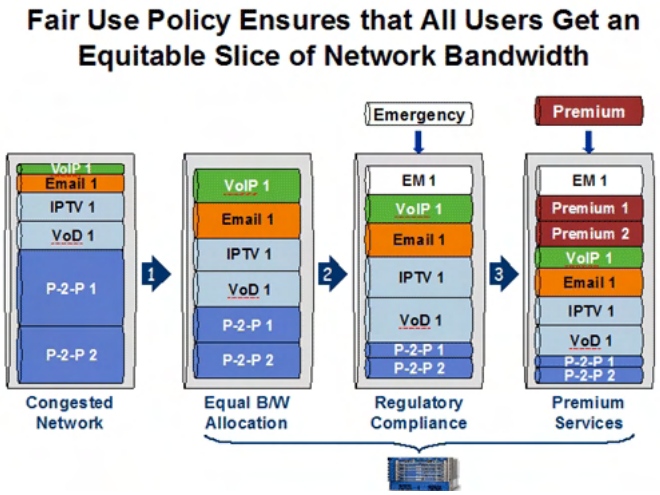


Abbildung 12.2: Fair Use Networking

12.3 Mobile Inhaltsanbieter und Netzneutralität

Da der sogenannte "Mobile Content" nur durch den Zugriff durch ein mobiles Empfangsgerät zu einem solchen wird, kann man den "Mobile" Aspekt in dieser Betrachtung weitgehend ausser Acht lassen und sich auf Content Provider als solches beziehen.

12.3.1 Geschäfts- und Kooperationsmodelle im Markt Mobiler Inhalte

Mobile Provider versuchen hauptsächlich, ihre Kunden an die Services zu binden, welche sie selber anbieten. Ein Telekommunikationsunternehmen welches Mobiltelefonie und zugleich drahtlose Datenverbindungen per UMTS anbietet, möchte verhindern, dass Kunden zum Beispiel Voice-over-IP Dienste über UMTS bei einem anderen Anbieter nutzen, anstatt über das Mobilfunknetz zu telefonieren.

So unterbindet zum Beispiel Vodafone ab August 2007 VoIP über UMTS: "Ab 08.07.2007 wird eine Voice- over-IP-Nutzung mit den Tarifoptionen technisch unterbunden." [17]

T-Mobile andererseits verbietet die Nutzung von VoIP und Instant Messaging bei einigen Wireless Angeboten: "Die Nutzung von VoIP und Instant Messaging ist ausgeschlossen." [18]

T-Mobile Österreich behält sich vor, die Nutzungsmöglichkeit von VoIP und Instant Messaging Diensten nicht zu garantieren: "Wir können nicht garantieren, dass Sie Voice over IP und Instant Messenger Produkte von Dritten nutzen können." [19]

Mobile Inhaltsanbieter versuchen derweil, Kooperationen mit Mobile ISPs oder Herstellern von Mobile Devices einzugehen, um den Zugang zu ihren Inhalten zu gewährleisten. So geschehen beispielsweise zwischen Google und Sony Ericsson. Der Mobiltelefonhersteller hat in den Modellen K800i und K610i Funktionen von Google Web Search integriert [20].

Google gehört brisanterweise zusammen mit Yahoo! und Microsoft zu den prominentesten Befürwortern der Netzneutralität (s. Interessenskonflikte der Martkteilnehmer).

12.3.2 Bedeutung und Auswirkungen von Netzneutralität auf Mobile Inhaltsanbieter

Für Inhaltsanbieter bedeutet Netzneutralität hauptsächlich, dass die von ihnen angebotenen Inhalte allen Netzbenutzern zur Verfügunge stehen, egal über welchen ISP sie mit dem Netz verbunden sind. Dies bedeutet weiter, dass auch kleine und/oder neue Unternehmen die Möglichkeit haben neue innovative Technologien an den Markt bzw. ins Netz zu bringen, ohne sich zuerst Zugang erkaufen zu müssen.

Eines der Hauptargumente für Netzwerkneutralität ist, dass ein diskriminatorisches Netzwerk den Märkten, die von dem Internent abhängig sind, stören, und damit letztendlich das Wirtschaftswachstum nachhaltig hemmen würde. Dadurch wäre es möglich, dass Unternehmen A einen Markt dominiert, obwohl die Technologie von Unternehmen B besser ist. Ermöglicht wird das durch die stärkere Finanzkraft von A, das sich den schnellen Zugang zu Netzwerken erkaufen kann. Auf die gleiche Weise kann auch der Einzug einer neuen Technologie erschwert oder gar unmöglich gemacht werden, weil das diskriminatorische Netzwerk eine andere zeitgenössische Technologie bevorzugt (behandelt). Was das mit ökonomischen Wachstum zu tun hat, lässt sich folgendermassen umreissen: die Innovationskraft eines Landes ist ein wichtiger Motor ökonomischen Wachstums. Führt nun die Abschaffung der Netzwerkneutralität zu einer geringeren Innovationskraft, wirkt sich das unmittelbar auf das wirtschaftliche Wachstum eines Landes aus. Demgegenüber stehen Kritiker, die eine Vernachlässigung grundlegender Anforderungen an Netzwerktechnologien befürchten, denn ein neutrales Netzwerk wäre gegenüber zeit- und bandbreitenkritischer Anwendungen blind, was einen ähnlichen Effekt auf die wirtschaftliche Entwicklung haben würde [21].

Die Inhaltsanbieter versuchen nun entweder durch Kooperation mit Netzbetreibern und/oder Hardwareherstellern oder durch Engagement im Kampf für die Netzneutralität vorteilhafte Wettbewerbspositionen zu erlangen. Wie man am Beispiel von Google und Sony Ericsson sieht, ist auch eine zweigleisige Strategie möglich.

12.3.3 Interessenskonflikte der Martkteilnehmer

Während Netzbetreiber auf ihr Recht pochen, selbst zu bestimmen, wer ihre Netze benutzt, und für die Benutzung auch Geld zu verlangen, plädieren Serviceanbieter und Konsumentenorganisationen für Netzneutralität. Auch Wissenschafter schließen sich den Forderungen nach Netzneutralität an. Sie argumentieren, dass die neue Entwicklung den

bislang stürmischen technischen Fortschritt im Bereich der Services (Inhalte und Anwendungen) massiv behindern könne und dass dadurch volkswirtschaftliche Schäden zu befürchten seien [22].

Interessenskonflikte beim Thema Netzneutralität bestehen hauptsächlich zwischen Netzbetreibern und Inhaltsanbietern. Netzbetreiber möchten selber entscheiden wie sie ihre Netze einsetzen, um beispielsweise bestimmte Protokolle oder Pakete bevorzugt behandeln zu können. Es bestehen auch Pläne, welche die Erhebung von Gebühren für die Durchleitung und Bereitstellung von Infrastrukturen qualitativ höherwertigen Inhalte vorsehen. So haben Netzbetreiber wie T-Online angekündigt, von Serviceanbietern wie Google, eBay, Skype oder dergleichen für den Zugang zu ihrer Kundschaft dereinst Geld zu verlangen.

Auf `http://www.savetheinternet.com` befindet sich eine Liste von Befürwortern der Netzneutralität. Zu den prominentesten Vertretern gehören Professor Lawrence Lessing von der Stanford und Professor Timmothy Wu von der Columbia Universität. Sie schreiben in einem gemeinsamen Brief an die FCC:

"Fundamentally, should the Commission care if the Internet remains a "neutral" network—more precisely, one that does not favor one application (e.g., the World Wide Web), over others (e.g., mass online gaming)? Is it of any concern to the public if the Internet is biased to favor some things over others?

The answer is yes. There are two reasons the Commission should care about maintaining a neutral network, both reflecting the Commission's interest in "stimulating investment and innovation in broadband technology and services." First, guaranteeing a neutral network eliminates the risk of future discrimination, providing greater incentives to invest in broadband application development today. Second, a neutral network facilitates fair competition among applications, ensuring the survival of the fittest, rather than that favored by network bias." [23]

Die prominentesten Befürworter aus der Wirtschaft sind Google, Yahoo!, Microsoft, Ebay und Amazon. Vor allem Google, Ebay und Amazon sehen ihre Quasi-Monopolstellungen in Gefahr wenn beispielsweise Netzbetreiber, welche gleichzeitig ähnliche Services anbieten, den Kunden den Zugang zu Google, Ebay oder Amazon sperren dürften.

Auf Seiten der gemeinnützigen Organisationen haben Moveon.org, Consumer Federation of America, AARP, American Library Association, Gun Owners of America, Public Knowledge, the Media Access Project, Free Press, the Christian Coalition und TechNet eine Allianz für Netzwerkneutralität gebildet und schliesslich hat sich auch Tim Berners-Lee zu Wort gemeldet.

Die Seite der Gegner besteht hauptsächlich aus den grossen Netzbetreibern sowie einigen gemeinnützigen Organisationen, darunter Freedom Works Foundation, National Black Chamber of Commerce, Progress and Freedom Foundation und der einflussreiche Think Tank New American Century (PNAC).

12.4 Zusammenfassung und Schlussfolgerungen

Wie Schewick [2] in ihrer Arbeit analysiert hat, sind die Rufe nach Netzneutralität-Regulationen gerechtfertigt. Bei Fehlen von solchen Regulationen, besteht eine reelle Gefahr, dass Netzwerk Provider unabhängige Applikations-, Inhalte- oder Portalproduzenten diskriminieren oder von ihrem Netzwerk ausschliessen werden. Diese Gefahr reduziert die Anzahl der Innovationen im Markt für Applikationen, Inhalte und Portale mit signifikanten Kosten für die Gesellschaft. Während Netzneutralität-Regulationen diese Gefährdung beseitigen, sind sie nach Schewick jedoch nicht ohne Kosten. Neben den Kosten der Regulation selbst, reduzieren Netzneutralität-Regulationen die Anreize der Netzwerk Provider auf dem Netzwerk-Bereich zu Innovationen vorzunehmen und Netzwerk Infrastrukturen weiter zu entwickeln. Folglich besteht in der Regulation ein Zielkonflikt.

Wie Schewick aufgezeigt hat, ist der potenzielle Nutzen von Innovationen im Applikationsbereich für den ökonomischen Wachstum von enormer Bedeutung. Ein Anstieg solcher Innovationen durch Netzneutralität-Regulationen ist um einiges bedeutungsvoller als die damit verbundenen Kosten. Bevor Netzneutralität-Regulationen entworfen werden können, ist jedoch auch nach Schewick noch mehr Forschung notwendig. So muss beispielsweise die Frage nach dem Ausmass des Regelungskreises der Regulationen gelöst werden. So muss auch der bestmögliche Weg der Implementationen noch identifiziert werden [2].

Seit einigen Monaten wird innerhalb der EU-Kommission das Thema Netzneutralität, mit etwas Verzögerung zu den USA, nun auch debattiert. Rechte und Pflichten betreffend der Netzneutralität im weiteren Sinne werden innerhalb der EU in der Rahmenrichtlinie zu elektronischen Kommunikationsnetzen und -diensten geregelt. Dieser Rechtsrahmen, welcher im Frühling 2002 in Kraft getreten ist, betrifft insbesondere auch Rundfunknetze und Rundfunkdienste. Die Richtlinie ist darauf ausgelegt, die Entwicklung des für den Wettbewerb geöffneten Telekommunikationsmarktes zu regulieren. Aufgrund der rasanten Entwicklung in dieser Branche sah sich die EU-Kommission veranlasst, diese Rahmenrichtlinie zu überprüfen. Dafür wurde im Juni 2006 ein Arbeitspapier veröffentlicht, welches unter anderem auch die zweifelhafte Sicherstellung der Netzneutralität analysiert. Es wird festgehalten, dass die aktuelle Entwicklung kein Anlass zur Veränderung der gesetzlichen Rahmenbedingungen gebe. Angeblich werden sich auf einem funktionierenden Markt immer Anbieter finden, welche die Benutzung der Netzwerkdienste möglichst ohne Einschränkungen anbieten. Es stellt sich also die Frage, ob die Privatisierung der Mobilfunkunternehmen tatsächlich über jeden Zweifel erhaben ist.

Mobile Provider versuchen vorwiegend, ihre Kunden an die Services zu binden, welche sie selber anbieten. Ein Telekommunikationsunternehmen welches Mobiltelefonie und zugleich drahtlose Datenverbindungen per UMTS anbietet, möchte verhindern, dass Kunden zum Beispiel Voice-over-IP Dienste über UMTS bei einem anderen Anbieter nutzen, anstatt über das Mobilfunknetz zu telefonieren. Mobile Inhaltsanbieter versuchen derweil, Kooperationen mit Mobile ISPs oder Herstellern von Mobile Devices einzugehen, um den Zugang zu ihren Inhalten zu gewährleisten. Die Inhaltsanbieter versuchen somit entweder durch Kooperation mit Netzbetreibern und/oder Hardwareherstellern oder durch Engagement im Kampf für die Netzneutralität vorteilhafte Wettbewerbspositionen zu erlangen.

Literaturverzeichnis

[1] Lexexakt: Netzneutralität http://www.lexexakt.de/, 26.10.2006.

[2] Dr.-Ing. Barbara van Schewick: Towards an Economic Framework for Network Neutrality Regulation, http://www.lessig.org/blog/archives/b_paper.pdf, Version: September 20, 2005.

[3] T. Wu: Network Neutrality and Broadband Discrimination, Journal on Telecommunications & High Technology Law. 2: 141., 2003.

[4] T. Wu: The Broadband Debate: A User's Guide, Journal on Telecommunications & High Technology Law. 3: 69., 2004.

[5] Aron Weiss: Net neutrality?: there's nothing neutral about it, ACM Press, http://delivery.acm.org/10.1145/1140000/1138097/p18-weiss.pdf?key1=1138097&key2=1972781611&coll=GUIDE&dl=GUIDE&CFID=2067615&CFTOKEN=29012352, 06.2006.

[6] Side Effects, http://litart.twoday.net/stories/2489238/, 26.12.2006.

[7] eBay, The Main Street Crier - Summer 2006, http://www.ebaymainstreet.com/newsletter/?id=000091, 26.12.2006.

[8] eBay, eBay Main Street Member Program, http://www.ebaymainstreet.com/mainstreet/?campaign_id=neutrality1, 26.12.2006.

[9] eBay, Net Neutrality, http://www.ebaymainstreet.com/federal/net-neutrality/, 26.12.2006.

[10] Alfred Krüger: eBay kämpft für ein neutrales Internet, IT-News World, 02.06.2006 http://www.it-news-world.de/news_862/eBay+k\%E4mpft+f\%FCr+ein+neutrales+Internet, 26.12.2006.

[11] Stefan Krempl: Google will Netzneutralitätnotfalls einklagen, Heise.de, 5.7.2006, http://www.heise.de/newsticker/meldung/75084, 26.12.2006.

[12] Eric Schmidt: A Note to Google Users on Net Neutrality, http://www.google.com/help/netneutrality_letter.html, 26.12.2006.

[13] Vinton G. Cerf: Network Neutrality, Prepared Statement, http://commerce.senate.gov/pdf/cerf-020706.pdf, 26.12.2006.

[14] Monika Ermert: Demonstration für Netzneutralität in den USA, 29.11.2006, http://www.heise.de/newsticker/meldung/81709, 26.12.2006.

[15] Save the internet.com, http://www.savetheinternet.com/, 26.12.2006.

[16] Stefan Krempl: Demonstranten in 25 US-Städten fordern Netzneutralität ein, 3.9.2006, http://www.heise.de/newsticker/meldung/77634, 26.12.2006.

[17] Vodafone, http://www.vodafone.de/unternehmen/presse/6613_74139.html.

[18] T-Mobile, http://www.t-mobile.de/business/email/0,9929,14533-_,00.html.

[19] T-Mobile AT, http://www.t-mobile.at/_PDF/AGB/AGB_25092006.pdf.

[20] heise online: 3-Megapixel-Handy mit Google- und Blog-Unterstützung, http://www.heise.de/newsticker/meldung/70182, 28.02.2006.

[21] Steve Graegert: Was bedeutet Network Neutrality (Netzwerkneutralität)? http://eth0.graegert.com/index.php?section=docsys&cmd=details&id=19.

[22] Wikipedia: Netzneutralität http://de.wikipedia.org/wiki/Netzneutralit\"at.

[23] Tim Wu & Lawrence Lessig: Letter to the FCC, http://faculty.virginia.edu/timwu/wu_lessig_fcc.pdf, 22.08.2003.

[24] FCC: Policy Statement, http://www.lasarletter.net/docs/2005/fcc_netneutrality.pdf, 25.09.2005.

[25] Save the internet.com http://www.savetheinternet.com/=threat.

[26] COMMISSION OF THE EUROPEAN COMMUNITIES: Review of the EU Regulatory Framework for electronic communications networks and services, http://europa.eu.int/information_society/policy/ecomm/doc/info_centre/public_consult/review/staffworkingdocument_final.pdf, 28.06.2006.

[27] WHAT DOES AT&T HAVE TO SAY ABOUT NET NEUTRALITY? http://www.netcaucus.org/events/2006/netneutrality/one-pagers/nn-atandt.pdf, 24.01.2007.

[28] Kristina Sam: US-Repräsentantenhaus lehnt Netzneutralität ab: Republikaner setzen auf Selbstregulierung http://www.pressetext.ch/pte.mc?pte=060609034, 09.06.2006.

[29] Cablecom will mehr Geld und ruiniert Backbone http://www.infoweek.ch/news/NW_single.cfm?news_ID=14536&sid=0, 12.10.2006.

[30] Ben Schwan: Netzneutralität in Gefahr http://www.heise.de/tr/artikel/69272, 07.02.2006.

[31] Monika Ermert: Netzneutralität: USA debattieren, EU wartet ab http://www.heise.de/newsticker/meldung/75524, 16.07.2006.

[32] Europäisches Parlament: Richtlinie 2002/19/EG http://www.bmvit.gv.at/telekommunikation/recht/downloads/rl2002de019.pdf, 07.03.2002.

[33] Commission of the European Communities: Review of the EU Regulatory Framework for electronic communications networks and services http://europa.eu.int/information_society/policy/ecomm/doc/info_centre/public_consult/review/staffworkingdocument_final.pdf, 28.06.2006.

[34] Fiete Stegers: Das ist das Tony-Soprano-Geschäftsmodell: Interview mit Barbara van Schewick http://www.tagesschau.de/aktuell/meldungen/0,1185,OID5692984,00.html, 11.07.2006.

[35] Wikipedia: Netzneutralität http://de.wikipedia.org/wiki/Netzneutralit\%C3\%A4t, 16.12.2006.

[36] Hans Peter Lehofer: Zur Umsetzung des neuen EU-Retsrahmens für elektronische Kommunikationsnetze und -dienste http://www.rtr.at/web.nsf/deutsch/Portfolio_Fachpublikationen\%20RTR/\$file/UmsetzungEURechtsrahmen.pdf, 07.05.2002.

[37] Junaid Islam: Fair Use Networking: Preserving Net Neutrality with enhanced QoS http://www.convergedigest.com/bp-ttp/bp1.asp?ID=354&ctgy=, 05.04.2006.

[38] Benno Gasser: Für ein schnelles Glasfasernetz http://tages-anzeiger.ch/dyn/news/zuerich/700946.html, 21.12.2006.