University of Zurich
Department of Informatics

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

Internet Economics II

TECHNICAL REPORT – No. ifi-2006.02

February 2006

University of Zurich
Department of Informatics (IFI)
Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

ifi

B. Stiller, T. Bocek, C. Morariu, P. Racz, M. Waldburger (Eds.):
Technical Report No. ifi-2006.02, February 2006
Communication Systems Group
Department of Informatics (IFI)
University of Zurich
Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
URL: http://csg.ifi.unizh.ch

Introduction

The Department of Informatics (IFI) of the University of Zürich, Switzerland works on research and teaching in the area of communication systems. One of the driving topics in applying communications technology is addressing investigations of their use and application under economic constraints and technical optimization measures. Therefore, during the winter term WS 2005/2006 a new instance of the Internet Economics seminar has been prepared and students as well as supervisors worked on this topic.

Even today, Internet Economics are run rarely as a teaching unit. This observation seems to be a little in contrast to the fact that research on Internet Economics has been established as an important area in the center of technology and economics on networked environments. After some careful investigations it can be found that during the last ten years, the underlying communication technology applied for the Internet and the way electronic business transactions are performed on top of the network have changed. Although, a variety of support functionality has been developed for the Internet case, the core functionality of delivering data, bits, and bytes remained unchanged. Nevertheless, changes and updates occur with respect to the use, the application area, and the technology itself. Therefore, another review of a selected number of topics has been undertaken.

Content

This new edition of the seminar entitled "Internet Economics II" discusses a number of selected topics in the area of Internet Economics. The first talk "AAA Support for Multicast Services" discusses authentication, authorization, and accounting services in general and for multicast services. Talk two "Incentive Strategies for Cooperation in Ad-hoc Networks" provides an overview of different mechanism and strategies in order to encourage cooperation in mobile ad-hoc networks. "Losses Resulting from Fraud, Espionage, and Malware" as talk three discusses possible attacks and threats in the area of information technology and their economic aspects. The fourth talk addresses "The Business Model: Open Source" and gives an overview of open source software and its business potential. Talk five "Technical and Economic Aspects of Inter-domain Service Provisioning" presents service provisioning in a multi-domain environment and discusses business models for virtual organizations. The sixth talk "Economy-driven Peering Settlements" presents the current Internet structure and addresses the relationship and interactions between Internet Service Providers (ISP) including peering and transit connections.

Talk seven "Financial Clearing for Roaming Services between Mobile Network Operators" outlines business relations between network operators in mobile telecommunication systems. The talk presents technical and economical aspects in support of roaming services. Talk eight "Software Patents: Innovation Killer or Innovation Supporter" and talk nine "Economic Impacts of Intellectual Property Rights in ICT" discuss copy right and patent issues for software products, analyze their advantages and disadvantages, and presents digital rights management. Talk ten "Charging Models in DiffServ Networks" presents possible charging mechanisms for Quality of Service, provided by the Differentiated Services architecture. Talk eleven continues with "Grid Services and their Market Potentials" and gives and overview of grid services and business models for grid service providers. Finally, talk twelve "New Business Models based on P2P" outlines the business potential of peer-to-peer systems.

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 Thomas Bocek, Cristian Morariu, Peter Racz, 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 Internet Economics, 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, February 2006

Contents

Chapter 1

AAA Suport for Multicast Services

Bernhard Wasser, Edoardo Beutler, Marco Jaggi

This paper deals with the topic of AAA-Support for Multicast services and the challenges of pricing for multicast services. Starting with an introduction to multicast and AAA including functionality, challenges and application examples, the paper focuses more deeply on the economical problems resulting from the characteristics of multicast transmission concerning authentication, authorisation and accounting. Authentication and authorisation are much more complex for multicast transmission because of the large user groups and the resulting key management problems and also because of the size of the transmitted authentication data. Concepts like IGAP servers and AAA-Protocols like RADIUS and DIAMETER help dealing with these issues. The basic accounting and charging problems like calculating the cost and sharing it comprehensibly between all the receivers are far more difficult than in a unicast environment, therefore enhancement of accounting support is a must. The complexity grows when different providers are involved in the delivery, because accounting between those participants is required too. The content owner does not necessarily know how many users are connected to a specific multicast stream; therefore the pricing methods the providing companies may be difficult to choose. There are several ways of cost sharing: splitting the total cost evenly between the users, sharing the cost based on the number of links or a highest bid method. The costs for a specific user may vary massively between those three methods, for every situation the most appropriate method should be chosen. To satisfy the customers, a comprehensible and transparent pricing is necessary. But apart from the accounting accuracy other requirements like security, scalability and robustness are also a must to offer commercial services on a multicast basis. Architectures like IGAP and VIPCAS provide solutions for accounting, charging and pricing; these and other frameworks are compared and presented in the last part of this paper.

Contents

1.1 Multicast

Multicast is a mode of transmission for delivering information to several destinations simultaneously. The message is delivered over each physical link of the network at once. Whenever the links to the destinations split, the data is copied. This strategy will distribute the data efficiently by minimizing the resource use. If unicast - the conventional point-to-point delivery mode - is used to deliver data to several locations, a copy of the data hast to be sent from the source to each recipient for every single user. This procedure results in inefficiency at the sending side due to a lack of scalability. [1]

1.1.1 Application Area

The word "'Multicast"' is typically used to refer to IP Multicast, a protocol for sending to multiple receivers at the same time via TCP/IP networks, using a multicast address. Typically there is one content server, the source where the data is located and multiple users at different locations. A multicast server splits the data stream at every branching point. The resulting behavior is a tree structure as shown in Figure 1.1.

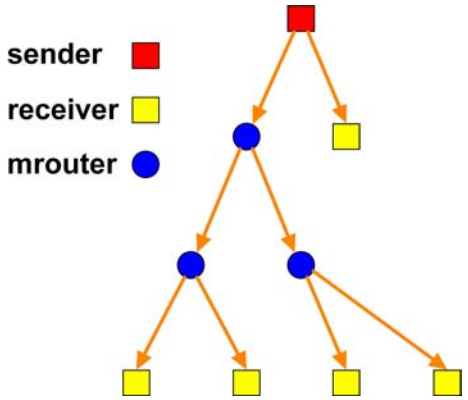


Figure 1.1: Example of a Multicast-Tree, in dependance on [6]

The difference to broadcast is the fact that the data is only delivered to those users, that specifically asked for it.

Examples of multicast technology are web conferencing [2] - used to hold group meetings or presentations over the internet - or the IRC [3], the Internet Relay Chat; a system for multi-user communication and file sharing in different channels. Another application area is video conferencing [4], similar to web conferencing, but using an additional webcam to transmit video streams of the participating user. Videoconferencing is an imperfect replacement to face-to-face meeting in private and business issues. Whenever time or resources for traveling is short, videoconferences may be a viable substitute for a physical meeting. Apart of those mentioned applications, multicast transmission is mostly used for the distribution of large multimedia content to several users at the same time. Due to faster internet connections, multimedia content like video or audio files become more and more common. The actual lack of scalability in traditional transmission mode generates

high bandwidth need; multicast may reduce transmission costs for the provider of large data streams. [5]

1.1.2 Difficulties

Multicast bears a lot of potential in the mentioned application areas, but first, several problems have to be overcome. First of all, scalability is still an issue. Currently there is no efficient way to distribute data between millions of content servers and millions of user groups [1]. The much less complicated scenario of just a small number of content servers is much more practical; to become a real standard for multimedia distribution, much larger multicast delivery-networks need to be created. Another major problem area is security. A certain level of privacy has to be guaranteed, especially for business purpose. During a conference concerning confidential data, every user has to be authenticated, to prevent unwanted participants from joining. Even more devastating is a scenario where an unauthorised person joins the conference disguised as an authorised member. The problem complexity grows with larger user groups. [6]

1.1.3 Requirements

For a commercial service, there needs to be some sort of accounting support. Whenever services are sold to users, accurate cost calculations are needed to deliver viable customer bills. In multicast scenarios, it typically is not known who and how many users are consuming a service, this complicates the creation of correct accounting data massively. So whenever multiple providers are part of the delivery, there need to be means of information sharing for beeing able to calculate the resulting costs. The cost calculation may vary depending on provided quality of service, number of active users and structure of the resulting multicast tree. Furthermore a way to share the costs between the participant providers and between all participant users must be found. A fair and comprehensible method needs to be chosen and communicated to all participant interest groups. Without a viable accounting support handling the pricing and cost allocation problems, there may never be commercial services based on the multicast technology. [7]

1.2 AAA

AAA is an architectural framework for configuring a set security functions in a consistent manner. An AAA protocol is used to control the access to network servers and to define what services users may start. This allows reacting on the security issue in multicast services. Furthermore AAA provides ways for an accurate accounting procedure. [8]

1.2.1 Definition

AAA stands for authentication, authorisation and accounting protocol:

- Authentication: Differentiates between valid and invalid users. Whoever needs access to an AAA-secured network needs to provide his identity and his credentials. Examples for possible credentials are passwords, one-time tokens, digital certificates, and phone numbers [15]. The server compares a user's authentication credentials with other user credentials stored in its database. If they match, the user gains access to the network. If they are different, the authentication fails and access is denied. AAA provides a method to identify users, including login and password dialog, challenge and response, messaging support, and, depending on the security protocol selected, encryption. [16]

- Authorisation: After passing the authentication, the user must gain authorisation for specific services. The authorisation process determines whether or not the user has the rights to perform such commands. AAA authorisation works by assembling a set of attributes that describe the user's authorisation rights. These attributes are compared to the information contained in the server's database for a given user and the result is returned to AAA to determine if the user is authorised, or if access is restricted [8]. Authorisation can be based on different restrictions, for example time-of-day restrictions, physical location restrictions or restrictions against multiple logins by the same user. Usually, authorisation occurs within the context of authentication. [15]

- Accounting: To track the consumption of network resources by users, AAA provides accounting support. The gathered information may be used for planning, billing and reporting purpose. There are two different modes of accounting: real-time accounting and batch accounting. Real-time accounting refers to accounting information that is delivered just in time with the consumption of the resources. In batch accounting the information is saved and delivered later on. Typically information like the identity of the user, the nature of the service delivered, when the service began and when it ended, number of packets, and number of bytes are gathered. [15]

1.2.2 Protocols

There are several protocols that perform the described functionalities. The most widespread AAA protocols are RADIUS, DIAMETER and TACACS.

RADIUS

RADIUS is an abbreviation for "'Remote Authentication Dial In User Service"' and provides the functionality of an AAA protocol [28]. RADIUS is intended to work in both local and roaming situations, so it is usable for mobile systems. RADIUS was originally developed by Steve Willens in 1992 as a commercial application for Network Access Servers, but was published later as RFC by the IETF (Internet Engineering Task Force) [17]. Now, several commercial and open-source RADIUS servers exist [18]. Radius is based on the client server architecture. All user data is stored in a central database on the Radius server. The Radius server itself can be the client for other Radius servers; this

allows creating hierarchical structures with multiple RADIUS servers, to distribute the functionality between them. The transport protocol between client and RADIUS server is UDP [9].

The messages between RADIUS client and server are authenticated with a shared secret that is never transmitted over the network. Additionally encrypted user-passwords are in use. RADIUS supports multiple methods of authentication. Some usable protocols are PAP (Password Authentication Protocol [19]), CHAP (Challenge Handshake Authentication Protocol [20]) or (Extensible Authentication Protocol EAP [21]). Due to the fact that RADIUS exists as an open standard, it is still the most widespread protocol for AAA-functionality.

DIAMETER

In larger and more complexe networks RADIUS loses some of its advantages due to a lack of scalability. To improve the AAA functionality a new protocol was designed. DIAMETER is the successor of RADIUS; the name is supposed to be geometric pun. DIAMETER provides more complexe funtionality as its predecessor and it is backwards compatible. Still there are possibilities to perform an upgrade. DIAMETER is based on a peer-to-peer archticture, not client-server. This may be very usefull in accounting scenarios, because the server has to possebility to request data from the client, without having to wait for the client to send a request. Instead of UDP it's using TCP or SCTP, and it supports transport level security like IPSEC. [22]

TACACS

TACACS stands for "'Terminal Access Controller Access Control System"' and is a remote authentication protocol that is used to communicate with an authentication server commonly used in UNIX networks. The functionality is basically the same as in the other two procols. [23]

As mentionned, server chains can be built to distribute the functionality or to have a second AAA-server in case of problems. Figure 1.2 shows a workstation within a network, sercured by four AAA-Servers, two RADIUS and two TACACS+ servers.

1.2.3 Example

Suppose the system administrator has defined a method list where the RADIUS server on top will be contacted first for authentication information, then the second RADIUS server followed by the two TACACS+ servers. When a user tries to dial in to the network, the network access server (NAS) first queries the RADIUS server on top for authentication information. The Server checks its database and if it can authenticate the user, it issues a PASS response to the network access server and the user is allowed to access the network. If it returns a FAIL response, the user is denied access and the session will be terminated.

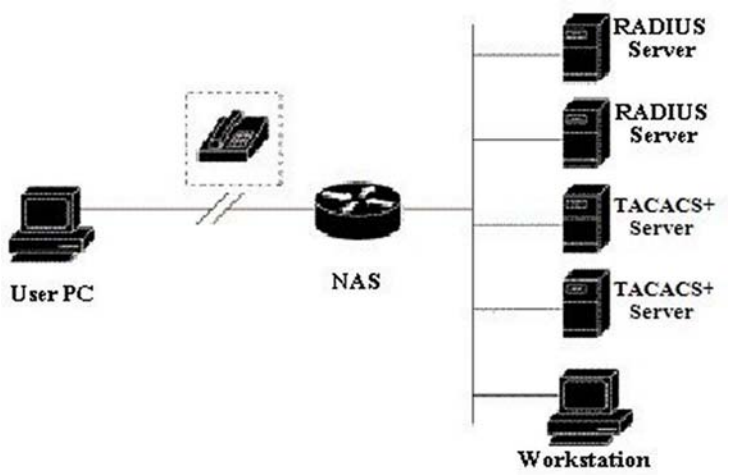


Figure 1.2: Example for an AAA secured network [8]

In both cases the other three servers do not take part in the whole procedure. But the first server does not respond, then the network access server processes that as an ERROR and queries the next server for authentication information. This pattern will continue through the remaining designated methods until the user is either authenticated, rejected, or the session terminated. If all of the authentication methods return errors, which the network access server would process as a failure, the session would be terminated. [8]

1.2.4 Multicast Requirements

All above-mentioned protocols need to provide additional requirements apart of the basic AAA-functionality. Especially in scenarios of multicast transmissions issues like scalability need to be solved to not collide with the aims of providing data to large numbers of users; the AAA-functions must not become the bottleneck in multicast transmission, even if large numbers of groups are requesting the provided data. Security and robustness are always needed in environments where sensible data may be transmitted; all AAA-protocols provide several methods for securing data transmission. [22]

1.3 Authentication and Authorisation

These two expressions are closely related, while the third A, the accounting is a bit separate. Thus authentication and authorisation will be handled in one section, while accounting has its own.

1.3.1 Authentication

In authentication there are two points of view. There is the authentication of the user towards the system (client towards server) to let the server know who he is dealing with

and there is the authentication of the server towards the user (server towards client) to be able to guarantee a trustworthy source to the user.

Client towards server

As mentioned, the first A in Triple A stands for authentication. The authentication has the purpose to identify the connector towards the system (user view: "'Tell the system who I am"', system view: "'Who tries to connect?"'). This authentication is necessary for systems where only a selected group of users should have access (e.g. students of the university Zürich to subscribe for the "'Seminar Internet Economics II"', customers who have payed to watch a certain movie). Furthermore only identified users can be hold responsible for consumed benefits, or also (in a more negative view) for fraud (e.g. in auctions over the internet) or abuse of the provided services, as for example abusing a mail server to send spam-mails. [12]

The authentication is done by asking an identification and credentials (a secret only the specified user can know) from the connector. The most common method for authentication is to ask the users for a user name (identification) and a password (credentials). Other possible credentials are for example credit card informations (especially useful in commercial services where the user pays with his credit card for the provided services), mobile phone SIM card (maybe when downloading images, ring tones, movies to your mobile phone), biometric data (like fingerprints or retina scans) or certificates (a trusted authority guarantees a certain identity). Many other possibilities are imaginable. [10]

Server towards Client

As already mentioned there also is an authentication problem from the other point of view. The user wants to be sure of receiving the demanded data from the chosen source. This requires security to prevent for example someone acting as the sender and transmitting wrong or falsified informations (e.g. a sender acting like a trusted stock market source and writing about a share going up instead of down). An especially difficult situation to guaranteed the authenticity are real time streams. These are often lossy. The bandwidth of the users is different and especially in slow connections many packages get lost. Even in these situations the authenticity of the data needs to be guaranteed. This is an important part of the QoS (Quality of Service) issue. [10], [6] A common solution is to encrypt the multicast packets. There are two possibilities to be considered:

1. In **symmetric key encryption** the same key ist used to encrypt and to decrypt the data. This method is relatively fast, but is only useful to hide the multicast stream from unauthorised users during the data transmission. The authenticity can not be guaranteed. For everybody possesing the key (i.e. for every authorised user) it is not only possible do decrypt the data stream, but also to encrypt his own data using the correct key and this is exactly what we try to prevent. In multicast this danger is even higher because receivers often are also acting as senders.

2. The alternative is encryption with an **asymmetric key**. In asymmetric encryption a so called private key is used to encrypt the data and everybody possessing the public key (public key $\neq$ private key) is able to decrypt the data. This solution works fine for identifying the source. A receiver is only able to decrypt the stream, but can not encrypt his own data with the correct key, faking to be the trusted source. The disadvantage of asymmetric encryption is the much higher system resource needed to handle the data.

1.3.2 Authorisation

This is the second of the three A's. Authorisation refers to the granting of services from a system to a certain (normally already authenticated) user ("'What is a this user allowed to do?"'). In most cases the authorisation process is not only the decision whether access should be allowed or disallowed. Normally there are more restrictions, as for example different users with access only to their account (i.e. a professor would probably not be very happy with all the students having access to his account). Another often used restriction ist the time. Access can be allowed only for one week, or also only on a certain time-of-day (e.g. customers of a phone service provider are allowed to phone for free from 10 pm until 7 am). [9], [13]

1.3.3 The generic architecture

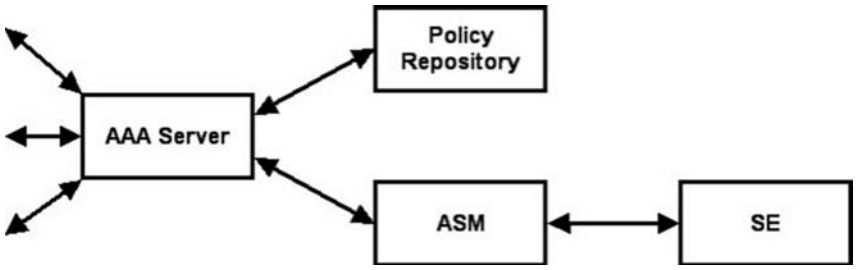


Figure 1.3: Generic AAA Architecture, in dependance on [9]

Often, especially in systems with different access points, the AAA services are spread over several systems. This is (normally) not very efficient, can cause hight administration effort and threaten the systems consistency (It can be efficient though if everything is well planed and organised and not originated in uncontrolled growth). Often a better solution is to centralise the three services on one server, although this poses a new problem. Such a server builds a bottle neck. If this server crashes, the system is no longer reachable. A model for this type of architecture is the generic AAA architecture from the IETF (Internet Engineering Task Force) [17]. This architecture divides the AAA tasks in a generic and an application specific part. [9]

- The **AAA server** builds the heart of the whole system. This server provides the AAA functionality.

- The AAA server is connected to an **ASM (Application Specific Module)**. The ASM is an interface connecting the (generic) AAA server with the systems specific services.

- The ASM is connected to the so called **SE (Service Equipment)**. The Service Equipment is the part of the system where the services the system provides (e.g. providing a multicast stream) are located.

- Further the AAA server is connected to a **Policy Repository**. This repository knows all the registred users, as well as the provided services. Using its decision guidelines the Policy Repository decides (i.e. is a certain user authorised to consume a service?) for the AAA server.

1.3.4 Multicast specific problems

In multicast there are some problems in authentication and authorisation that do not appear (or at leas are not that important) in unicast. Probably the biggest problem is the missing unicast connection (often the receivers are not even known) and that, depending on the net, receivers can even become senders. The obvious solution would be to individually authenticate each user and to work with a session key and encrypted data which can be decrypted on the client by using a MAC (Message Authentication Code) [6]. But there are some problems in this obvious solution:

1. The key management is difficult. For small networks it works pretty well but in a large multicast network the key management causes a lot of problems (scaling problems). The authentication data gets really long and the key management very complex. This is a problem that appears also if a unicast server has many clients connected, but it is much more relevant for multicast since the advantage of using multicast grows with the number of receivers. The whole idea of multicast is to use this technique for high amounts of users.

2. The distributed key has to be refreshed after a certain time. This is the only possibility to get rid of the users who are not anymore authorised to consume the provided services. So the authorised (only the ones who should still have access) receivers have to be informed about the change. This is a huge administrative task in a system with a lot of users.

As described an individual authentication of each user is difficult and poses some hard problems. Also it means you give up a part of the advantages gained through the use of multicast. You do not have to send the whole data to all senders as in unicast, but at least the authentication has to be done as in unicast. Is there no better solution to handle the users than authenticating each of them?

1.3.5 Alternatives

A better solution could be an authentication of, for example, user groups or categories with the same authorisations. For example group *A*, the Swiss people watching a football match and group *B* the Germans attending the match. This would simplify the situation for you have only two groups to handle instead of thousands of individual users. Unfortunately this solution does not go together with the third A, the Accounting. If you want to bill somebody for consumed benefits, you have to be sure that the billed benefits really have been consumed. What if one of the Swiss football watchers is bored by the game after 15 minutes and leaves? If there is only a group of Swiss you can not tell if this customer is still watching or if he really stopped watching. As we can see, for a proper billing (if not using special things like e.g. flat rates to access the services) an individual authentication is necessary.

1.3.6 Example - IGAP

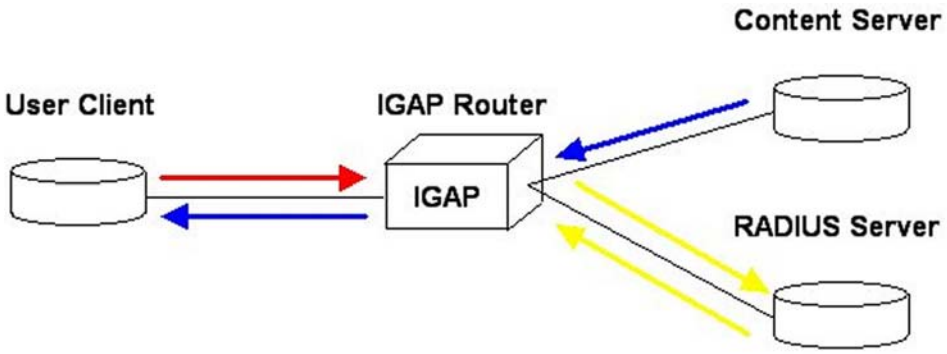


Figure 1.4: Example architecture, in dependance on [11]

An often used solution for Authentication and authorisation (and also accounting, see 1.4.6) is the installation of an IGAP (Internet Group membership Authentication Protocol) router between the content provider and the customer. As you can see the IGAP router is connected to a content-, and a RADIUS server. This architecture supports AAA and works as follows:

1. The client asks on the IGAP router for authentication and authorisation.

2. The IGAP router handles the authentication and authorisation with the RADIUS server (for information on RADIUS see 1.2.2).

3. The IGAP router lets the (now authenticated and authorised) user receive the multicast packages.

1.4 Multicast Accounting

Adequate pricing and charging methods are an issue for all commercial services. Some systems lack requirements as transparency or scalability to a certain level. The main reason for open problems with charging and accounting is the variety of service characteristics. The services differ in requirements concerning Quality-of-Service, network technology and the complexity of their billing system. Multicast services require different accounting systems than their unicast counterparty, due to the nature of multicast transmissions. The question of pricing and accounting is one of the major obstacles for developing commercial applications for a large user groups.

1.4.1 Terminology

Clear boundaries between the terms must be declared to avoid confusion within the terminology:

- Accounting names the process of information collecting concerning all the services a customer consumed. All collected data needs to be saved in a log or database. Typically the log includes information like connection duration, number of connections, delivered packet amounts or data size. The accounting data represents the basis for the charging and billings process. [15]
- Charging is the process of traversing the bridge between technical data and monetary numbers. The accounting data is processed and by measuring the resource use of a specific user the resulting cost is calculated. The whole pricing procedure is build on top of the accounting data. The pricing loses any relevance if accounting data is flawed. [24]
- Pricing is a part of marketing and describes the process of setting a price for specific services. Prices can be cost based, demand based or based on the competition within a market.
- Billing transforms the charging information into actual customer bills. The bill holds information of all consumed services over time along with the total amount of money the customer will have to pay. Main requirements to the customer are transparency and comprehensibility, the pricing needs to be displayed in a way the customer can understand it. [24]

1.4.2 Problems

The complexity of the accounting grows with the number of participant users. This is exactly the case in multicast transmissions due to the intention of this technique. Whenever multiple companies take part in the content delivery the complexity grows. Accounting and billing needs to be performed between those companies, not just between provider

and customer. The point of multicast is to generate efficiency by sharing a delivery tree; this main goal must not be limited by the means of accounting.

Basically we deal with three main problems within multicast accounting:

- Cost of multicast provisioning: The structure of a multicast tree can be highly dynamical, gaining or losing branches with every user fluctuation. With every new branching point the question rises, if this branching only generates additional costs or if it provides its share of the total amount of saved bandwidth. The structure of the tree highly influences the cost generation. Depending on its density and the size of the user groups the costs may differ widely.
- Measuring the resource usage: Again the unpredictability of the tree structure causes the difficulty. A given amount of resources needs to be shared between an unknown numbers of users. This gets even more disturbing if there is no way to trace back in time the locations of the users accurate enough to share the resource in a way to guarantee a certain quality of service.
- Cost sharing: Based on the cost calculations, the resulting amount needs to be shared between all participants. The costs need to be shared in a fair manner between all users. Depending on Quality-of-Service and the chosen price policy, cost between the users may differ; still they need to be comprehensible. With different providers taking part in the delivery another dimension results; the costs need to be shared between them too. One company may sell their content to another, which then sells it to its customers. So the content deliverer can probably share the costs between other companies, without direct contact to the end-users. [26]

1.4.3 Example

The following scenario shall elaborate the difficulties of such a multiple provider delivery. A company specialized on providing video stream decides to offer a service of high quality video streams via multicast. The users may join such a multicast group and accept the delivery of those streams. The company guarantees a certain quality and expects the user to pay for the service according to the chosen pricing. So far the delivery is quite easy to handle, the company has direct customer contact. Now to increase their profit and to make the service much more established the company decides to increase the group of users. Two well known telecommunication companies would like to sell the streaming service to its customers. They pay the owner of the content given price to get access to the content stream and sell the service to their customers with their own pricing policy.

The owner of the content is now in the situation that he can share his expenses between his own customers and the two other companies. The given set of available resources needs to be shared in a way to satisfy all participants. The contract with the direct customer persists; they still have a certain quality assured. Most certainly the deliverer had to set up a service level agreement with the two new customer companies, which have to be provided too. Both companies resell the service to an amount of customers - one company to 200 users, the other one to 600 users. The content owner does not necessarily have

to know how many users are hidden behind those two companies, if they don't display the correct amount of users, the true number will stay a secret. That fact can make the pricing policy between the three rather difficult. Between them some accounting support needs to guarantee that all services can stay comprehensible, even though the whole tree structure may not be known by all parties. Somehow the accounting data needs to be gathered. A resource reservation protocol may help with the resource allocation, to keep up the needed quality level.

1.4.4 Requirements

Apart of the mentioned multicast specific problems some basic requirements need to be fulfilled by the accounting system to run a commercial multicast service:

- Security: All billing information has to be confidential, and the possibility of manipulation or fraud by hacking or other abuse needs to be minimized. The customer needs to be sure that the system can not or can't easily be manipulated to his disadvantage. Many users would lose their trust in a service, if there were growing numbers of manipulation incidents, where some illegal users consume on the cost of viable ones.
- Robustness: If there is a connection interrupt, which can happen for several reasons, it needs to be recognized. If this is not the case, the customer would pay for the service even though he wouldn't really consume it. On the other hand it must not be possible for the user to evade all billing simply by creating an intentional disconnect from the transmission.
- Accuracy: To satisfy the customers, the billing must be absolutely accurate, fair and comprehensible. The pricing needs to be as simple as possible; complicated structures may discourage the casual user. All prices of the services need to be known and communicated to the customer before he starts consuming a service. The provider of the content can't just raise the price spontaneously just because at that moment there were too few customers to cover his costs.
- Scalability: The solution must not become a bottleneck in any way slowing down the transmission. It should have low impact on the network performance both in terms of added traffic and processing overhead by network elements.

1.4.5 Process Diagram

The process diagram is shown in Figure 1.5. Assuming the delivery of video streams once more, the user triggers the whole process. He starts or ends his video stream during the delivery. The service stops or ends for that specific, the streaming server identifies the action and sends the information to the node manager, here in this example an IGAP server. The server saves all user data and timestamps in his log, and searches for the matching access portal, where the date will be collected and evaluated.

The transmission to the access server can be performed in real-time - this means whenever the user triggers an event, the node manager sends the information to the access portal. In that particular case the log in the node manager and the log of the access server are consistent. The other possibility is the batch mode, where the access server has to send a request to every node the get all the needed information fragments to the activities of a certain user. When the user finishes using the service, it's possible to calculate the exact duration and use of the service the customer consumed. The cost can be billed to the customer using the declared pricing policy. If a technical problem would have occurred during the transmission, the accounting would have been paused, and the user wouldn't have to pay for the lost time. The whole process should be as simple as possible from the user's point of view. Authentication needs to be performed before sending any accounting data to other servers, to prevent manipulation or a loss of privacy. [27]

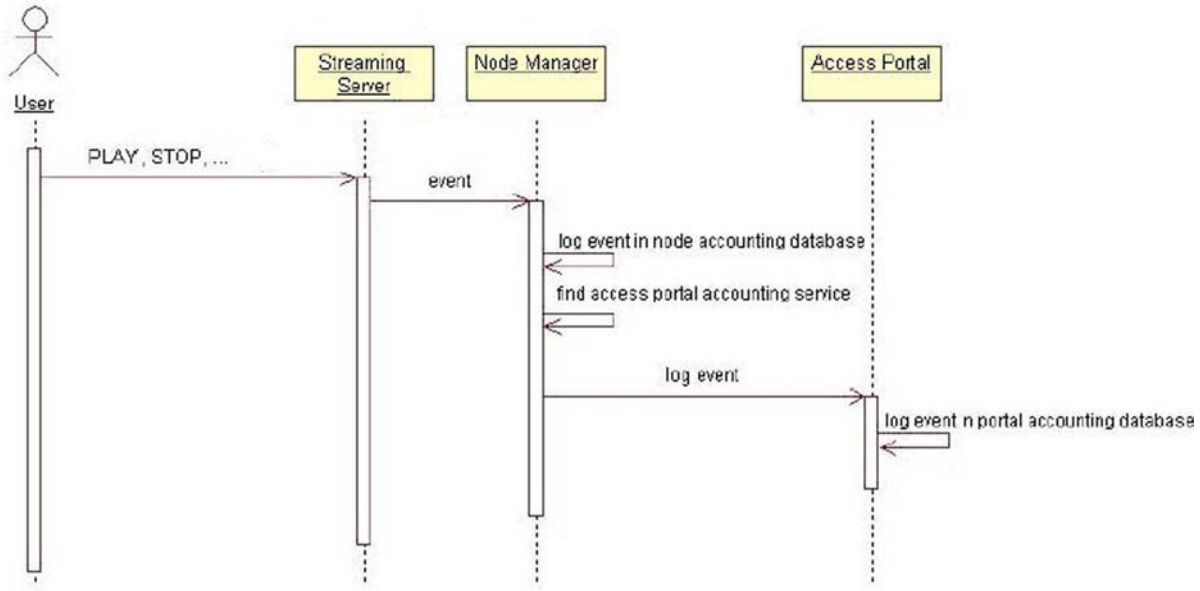


Figure 1.5: Accounting process diagram [27]

1.4.6 Implementation

A possible technical realization could now be reached using an IGAP server and a RADIUS server connected as shown in Figure 1.6.

Performing a join sequence into a multicast group the following steps need to be taken for a successful connection.

1. Send IGAP Join from user client to IGAP router
2. Send RADIUS Access Request from IGAP router to RADIUS server including user ID and password
3. Send RADIUS Access Accept from RADIUS server to IGAP router after user authentication and authorisation

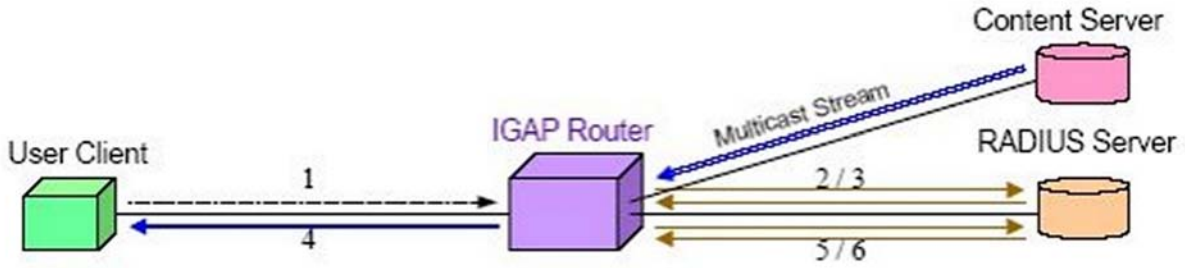


Figure 1.6: Joining an IGAP-Group [11]

4. Start to send multicast packets from IGAP router to user client
5. Send RADIUS Accounting Request from IGAP router to RADIUS server
6. Send RADIUS Accounting Response (start) from RADIUS server to IGAP router, collecting of accounting data starts.

Should in any case the access request not be accepted, the connection would be terminated without any packages of the multicast stream being delivered to the unauthorised user.

1.5 Cost Allocation and Pricing

1.5.1 Introduction

Cost Allocation is the process, where Internet Service Providers, or to be more specific in our topic, the providers of a multicast service determine the total cost their service is generating. Based on those total costs, that are most of the time more assumptions than concrete facts, the real pricing is then done. The pricing of an individual service determines, how much a consumer has to pay. By taking into account the providers different policies (like accounting or billing), the whole charging pocess can be categorized. Also, pricing can be used to influence usage behavior and to measure the policy compliance. In the end, it must pursue a rational cost recovery. To sum it up, cost allocation is affecting the provider, whereas pricing affects the consumer.

1.5.2 Cost Allocation Considerations

The billing model which evolved in the unicast environment is too inequitable when applied in a multicast environment to provide ISP‘s with any incentive. ISP‘s sell access to their combined mix of network nodes (customers, peers, and transit providers) to their customers. In an unicast environment, the sum of the customer‘s use of the ISP‘s network can be measured at the point at which it‘s aggregated, facing the customer. Other expenses can be amortized proportionately across all customers. Multicast traffic is multiplied within the ISP‘s network, such that the sum of the node utilization may be far

greater than what's observed at the point at which it enters from the customer. Customers who send multicast are already paying for a connection, and paying for unicast utilization. So it's needed to identify the difference between what the customer currently pays for, and what the ISP has to provide in a multicast environment. [29]

It is easy to see, that one of the major problems for the cost allocation of a multicast service is, that it is highly dependable from the multicast tree (see 1.7 for an example). For different variations of a multicast tree, the cost structure can change as well. Often the concrete costs for a multicast service cannot be determined before the service starts. So there are many multicast architectures, where the calculation of the costs must be done while the service is running - or even afterwards. Therefore the cost allocation algorithm must be able to perform the computation at the nodes, and to communicate between the nodes in an efficient way.

Many different flavors and generations of multicast routing protocols have been proposed. They can be classified into either dense mode (DVMRP [40], PIM-DM [41]) or sparse mode (CBT [42], PIMSM [43]). In addition to bandwidth usage, there are many other dimensions to the tradeoff between dense mode and sparse mode multicast. As a general rule of thumb, dense mode multicast is perceived to be appropriate for mass-dissemination applications such as webcasting, whereas sparse mode multicast is more suited for teleconferencing and other applications with just a few receivers. The question is: should dense and sparse mode multicast be priced differently?

Dense and sparse mode protocols differ primarily in their tree-construction techniques. Dense mode protocols take a flood-and-prune approach. In this approach data packets are periodically flooded to the entire network, and branches are pruned where there are no downstream receivers. Sparse mode protocols, on the other hand, grow the distribution tree on a branch-by-branch basis as new nodes join the multicast group. Dense mode protocols work well when most nodes in the network are receivers, but are extremely bandwidth inefficient when the group members are few and sparsely located throughout the network. [30]

1.5.3 Pricing in Multicast Services

As in pricing considerations of every service, multicast providers too have to assemble a pricing policy. It determines among other things, if the pricing should reflect the actual costs or not, like the network consumption. Another problem is, that the bandwidth used by a multicast transmission may not be directly attributable, and the sharing of bandwidth complicates the issue of pricing as well. Additionally, it is difficult to track the information on the multicast tree cost, and especially to do it efficiently. In general two approaches can be followed for sharing the costs for multicast communication among the participating nodes: link cost sharing (LCS) and total cost sharing (TCS). Let's have a closer look at how they work, and what the consequences are for the price the different receivers have to pay. Therefore we are looking closely at cost c_1 in Figure 1.7. Receiver R_1 gets the multicast service as the only receiver from the multicast router M_1. On the other hand, the three receivers R_2 through R_4 get the service from router M_2.

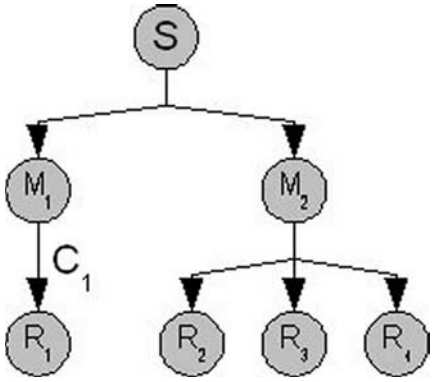


Figure 1.7: Example of a multicast setting with a sender (S), two multicast routers (M) and four receivers (R).

Link cost sharing is sharing the costs of individual links among all connected nodes. LCS schemes split savings from the joint usage of a link among the nodes that are directly connected. In the example of Figure 1.7, this means that receiver R_1 has to pay half of the total costs: $c_1 = C/2$, because there are two routers. The receivers R_2 to R_4 have to pay a sixth of the total costs: $c_{2,3,4} = C/2 * 1/3 = C/6$.

Total cost sharing is sharing the total costs of the multicast tree among all members of the multicast group. TCS can be treated as a simplified, abstract LCS scheme, where the whole network cloud is considered as one link, with all egress routers treated as directly connected nodes. In the example of Figure 1.7, this means that all receivers R_1 through R_4 have to pay the equal amount of a quarter of the total costs: $c_{1,2,3,4} = C/4$.

While LCS schemes usually require recursive calculation (for all nodes/links of the multicast tree) to allocate costs to the receiver, for TCS schemes it can be sufficient (depending on the total cost determination strategy) to consider only the border nodes. Therefore, it is possible to implement a TCS with less overhead than a LCS. [7]

An alternative approach to pricing is to have each receiver place a bid for the content. The network uses these bids to determine the set of receivers that obtain the content, as well as the price these accepted receivers pay. The price charged to an accepted receiver can be no more than its bid, but it is often advantageous to charge receivers a smaller price. [31]

1.5.4 Discussion of different pricing approaches

ELSD [33], standing for Equal Link Split Downstream, is the earliest work on multicast pricing. It splits costs amongst downstream receivers and allocates no costs upstream. This is shown to be an optimal cost allocation for single-source sessions with a source-rooted tree, but requires changes to the IP multicast service. ELSD is modified in [34] for multiple-source sessions, at the expense of some scaleability. It is not clear, however, that ELSD works for dynamic sessions.

EXPRESS [35] modifies IP multicast for large-scale single-source applications. Multiple-source sessions are catered for by a "'session relay"' approach, where all sources' traffic is sent via one node. This master node becomes a single point of failure, and may not be able to cope with delayintolerant applications.

In [36] multicast in an ATM intserv environment is examined. Charges are determined based on requested and received QoS, with the charging protocol encapsulated in RSVP. Since prices are determined a posteriori, they are unpredictable, and the scheme depends on RSVP.

In "'split-edge pricing"' [37], both sender and receiver initially pay a share of the cost of a transmission, and claims over the value of the transmission are settled later. Network providers agree prices for offering each subscription level to their neighbours, and these charges are summed to create the price for a complete transmission. It is shown that for multicast pricing both sender and receiver need to pay, because otherwise an incentive exists for downstream providers to lie about the number of receivers. This problem still exists with split-edge pricing, however, since senders and receivers settle claims after transmission, and no mechanism is provided for verifying the number of downstream receivers. This is the "'collusion prevention"' axiom of Herzog's thesis [34], which also states that this problem cannot be solved through cost allocation alone.

Einsiedler et al. [38] propose assigning weights to each link in a network, to represent the "'cost"' of that link. These weights can be derived from the congestion along the link, the costs of maintenance, or inter-domain costs for links that traverse ISPs. An extra Internet Group Management ProProtocol (IGMP) message or IPv6 header extension is used to store the weight information. Costs are determined, as in ELSD, by splitting the costs at each branching point in the tree. This scheme also violates the collusion prevention axiom, since the charge depends on who is paying, which may create incentives for senders or receivers to always pay, depending on which is cheaper.

In [30] is the cost efficiency of multicast over unicast analyzed. Multicast costs are capped at the unicast level, and it is assumed that new joins to an existing branch in a tree represent a zero marginal cost. A pricing scheme based on this would thus be incentive-incompatible, since it would only charge the first user in a branch.

[32] analyses multicast costs to determine the "'cheapest"' tree topology for a session in terms of network resources. Their method is protocol-independent, although pricing and user incentives are not considered.

The last paper mentioned in this section is [39] that has some good considerations. It discusses specifically the Marginal Cost Mechanism, the Welfare-Maximising Multicast Tree, and the Shapley Value Mechanism.

1.6 Architecture Example

1.6.1 Introduction

As we have already shown in this paper, there are several different architecture approaches to multicast services. They differ mostly in the protocols they are using, as well as the server-client architecture. With the needs of Authentication, Authorisation and Accounting, and therefore the need to calculate costs and charge the users, also different approaches to the client-server architecture have evolved. There is an immense ammount of proposals, some more suited to demonstrate the technical aspects of multicast, others more suited for showing the AAA support.

As we have already shown, some AAA architectures (e.g. IGAP in section 1.3.6 or the accounting solution presented in section 1.4.6), we would like to concentrate in this section here on a proposal for charging and accounting in QoS enhanced IP multicast, as presented in [7]. In this example we will show mainly the technical aspects of the architecture and protocols, and not so much the whole AAA process. The proposed architecture has QoS support and allows an usage based charging. It has a built in protection from abusive resource reservation, and the calculation of multicast tree cost can be done efficiently. Additionally, it has some savings and cost sharing mechanisms, that are especially interesting from an economical point of view. All Information, and especially all figures are taken from [7].

Of course there are other examples, that may illustrate better some different aspects of AAA support in multicast services. But we have chosen this example, because it is easy to explain and very graphic. For other architectures, the reader is referred to the papers mentioned in 1.5.4.

1.6.2 Example Overview

We will discuss the example in five parts. Each of them will be built upon the last, until we have the whole thing together, starting with the framework "'Charging and Accounting Reference Model (ChARM)"'. Then, the architecture called "'Value-Added IP Charging and Accounting Service (VIPCAS)"' is built considering the framework. It uses a data structure called "'Premium IP Network Accounting Record (PIP-NAR)"' whereas for the information exchange, the "'Charging Information Protocol (CIP)"' is used.

1.6.3 Example Details

The Framework

The left part of the figure represents the policy plane, i.e. technical and commercial rules for setting parameters based, e.g., on the network configuration (important for metering

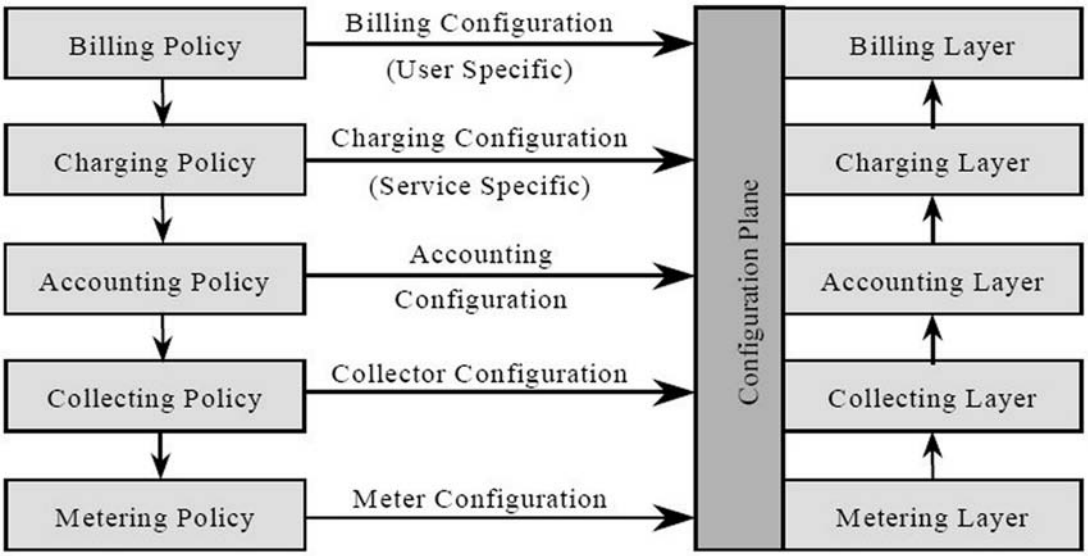


Figure 1.8: Charging and Accounting Reference Model (taken from [7])

policies) or the market situation (important for charging and billing policies). The parameters are injected as temporary data to the layers in the right part of Figure 1.8 through the configuration plane.

Configuration parameters are derived from pricing policy, charging policy, accounting policy and metering policy. Since the metering layer has to provide the data needed for the charging formula, it is useful to derive the basic elements of a lower layer policy (e.g. metering) from higher layer policies (e.g. charging).These policies can be provided by interaction of dedicated policy servers with the corresponding entities of the configuration plane.

The Architecture

Here is the description of the five layers with an outline of their main functions:

- **Metering Layer**: obtain reservation information, meter actual usage of network resources. Placed at the edge routers only or at multiple splitting points.
- **Collecting Layer**: meter data access and forward to accounting layer, select appropriate meters. Transfer of metering data can be initiated explicitly (self-explanatory) or implicitly (after a triggering event such as detection of a new flow).
- **Accounting Layer**: process collected usage data and reservation data, consolidate them based on service parameters, and create accounting data sets, reconstruct the multicast topology including splitting points where required by the cost sharing scheme.
- **Charging Layer**: cost allocation assigns costs to specific endpoints, such as sender(s) and receivers of a multicast group. Derives costs for accounting data sets based on

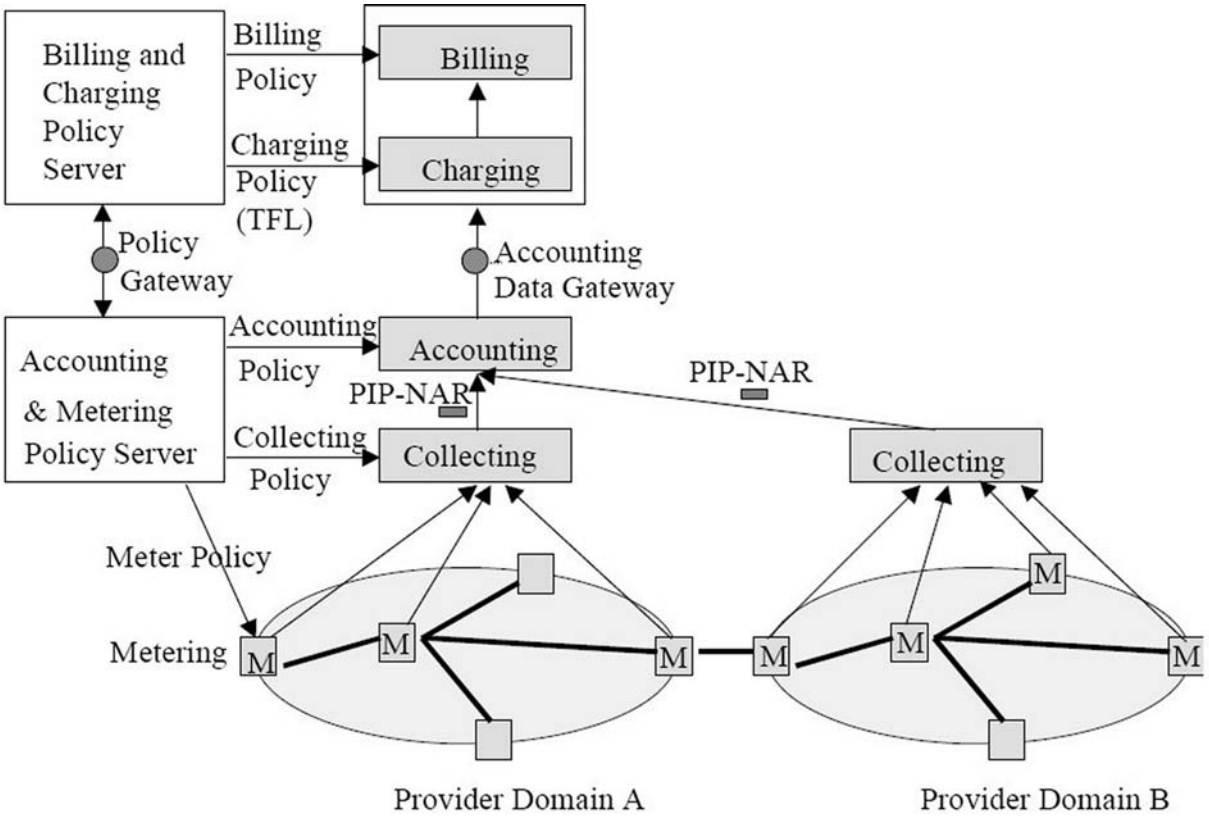


Figure 1.9: Value-Added IP Charging and Accounting Service (taken from [7])

service specific tariff parameters. A simple evaluation of current costs can be used for displaying an estimation of accumulated costs for the service user, or for control purposes by the customer organization or by the provider.

- **Billing Layer**: translates costs calculated by the charging layer into monetary units and generates a bill for a customer. This process may combine technical considerations with economic considerations, such as volume of resources used by the customers, and marketing methods (e.g. offered discounts).

Figure 1.9 is more or less self-explanatory. On the metring points (edge routers and multiple splitting points, designated with "'M"') the data is collected. It is very easy to see the five layers as described above as well as the policies and how they are built upon and interact with each other.

The Data Structure

Figure 1.10 shows a short overview of some of the parameters and flags used for transmitting the collected data to the accounting layer. This overview ist not complete. Only a part of the data record is shown.

Parameter	Type	Length [Bytes]
Record Description		
Version	Char	1
Length of Record	Char	1
Type of Record	Short	2
Measurement start time	Long	4
Measurement stop time	Long	4
Measurement point identification	**(meter IP address)**	**r_id**
Flow Description	**(src,dst IP, ports)**	**f_id**
Reserved Resources	**(e.g. flowspec)**	**rr**
Used Resources	**(#packets, #bytes)**	**umd**
Data Extension	**(e.g. distance, burstiness)**	

Flagname	Set	Not set
Uni-directional	unidirectional record	Bi-directional record
IPv6	IPv6 Flow	IPv4 Flow
Reserved Resources	PIP-NAR contains Reserved Resources	PIP-NAR contains no Reserved Resources
Used Resources	PIP-NAR contains Used Resources	PIP-NAR contains no Used Resources
DiffServ	Differentiated Services	Integrated Services
Extension	Extension Present	No Extension present

Figure 1.10: Selected Parts of Premium IP Network Accounting Record (taken from [7]). (It is not necessary to fully understand the used data record in detail. It is not even showed in full detail.)

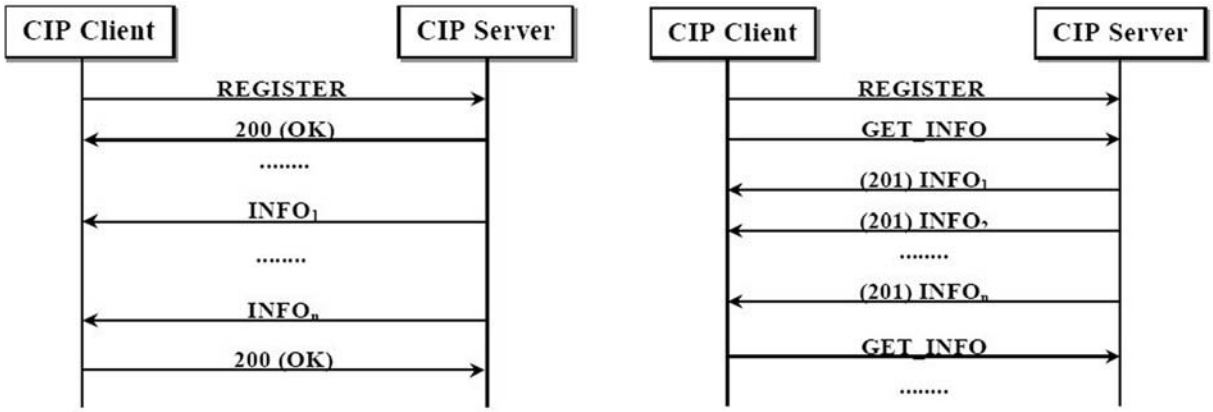


Figure 1.11: Charging Information Protocol (taken from [7])

The Information Exchange

In Figure 1.11 shows the procedure of an information exchange of charging informations. The information messages (INFO) contain the following fields: Identification (name of

the service and the provider), Validity, Tariff, QoS guarantees, Information about the reservation, and Transaction ID.

Distribution of charging information can be done by unicast or multicast transmission. Tariffs for the offered service classes are sent in a sequence of information messages (INFO). In order to allow clients to recognize a loss of a packet, the INFO messages contain sequence numbers. With these numbers it is possible to request a retransmission of lost packets. If a unicast connection is used for the announcement of tariffs, all clients that want to receive information about current tariffs have to register with the CIP sever first. In the registration request clients can choose between two modes to get information from the server. In the push rmode (default setting) information messages are sent periodically to the client. In order to prevent sending to non-existent or non-operational stations, messages are acknowledged by the client. In the pull mode information is only sent on demand. Clients need to send a request (GET_INFO) in order to get the information messages. In the unicast case CIP uses timeout and retransmissions to provide a reliable transport. Besides the reliability and the possibility to use TCP for transport, unicast distribution allows selective individually adapted advertisements. This means that the information can be reduced to tariffs that are new to a particular client. Furthermore, special offers for certain customers can be conveyed individually.

1.7 Summary

Multicast is a transmission mode for the simultaneous delivery of data to several destinations. Typically there are few data sources (content servers) and large numbers of users at different locations. In multicast transmission a tree-like structure is built. This happens through the splitting of the data at every branching point. Multicast transmission is mostly used for conferencing and for the distribution of large multimedia content. One of the basic requirements for a commercial service based on multicast transmission are good mechanisms for the authorization, authentiaction and accounting (called Triple-A or AAA) of the users.

The problems in AAA are due to the architecture of multicast. The service provider does not have a direct connection to the user, worse a user can even become a sender. For a proper accounting and securing of his data the service provider needs to know which user consumes the provided services. These requirements are difficult to fulfill in an efficient way. For a centralised AAA service there is a generic architecture introduced by the IETF.

As in every service, also multicast providers have to assemble a pricing policy. Among other things it has to be determined, if the pricing should reflect the actual costs of the network consumption or not. Another problem is, that the bandwidth used by a multicast transmission is difficult (if not impossible) to directly attribute.

In all commercial services adequate pricing and charging methods are an important issue. Often systems lack important requirements as transparency or scalability to a certain level. The main reason for unresolved problems with charging and accounting is the large

variety of service characteristics. The services differ in requirements concerning Quality-of-Service, network technology and the complexity of their billing system. Due to the nature of multicast transmission, these services require different accounting systems than their unicast counterpart. These questions of pricing and accounting are one of the major obstacles for developing commercial applications for large user groups.

The structure of a multicast tree can be highly dynamical. The tree gains or loses branches with every user fluctuation. The structure of the tree has a high influence on the cost generation. The real costs may differ, depending on the trees density and the size of the user groups, widely. The costs of the transmisson needs to be shared between the customers. There are two general approaches which can be followed for sharing the costs for multicast communication among the nodes participating: link cost sharing and total cost sharing. Link cost sharing is sharing the costs of individual links among all connected nodes on that link. Total cost sharing is sharing the total costs of the multicast tree among all the members of the multicast group. There are also alternative approaches, one of them is to have each receiver place a bid for the content.

There are several different frameworks dealing with these issues. Some examples presented in this paper are Charging and Accounting Reference Model (ChARM), the Value-Added IP Charging and Accounting Service (VIPCAS), Premium IP Network Accounting Record (PIP-NAR) or the Charging Information Protocol (CIP).

Bibliography

[1] Online Wikipedia: Multicast http://en.wikipedia.org/wiki/Multicast, Last accessed: 22.11.2005.

[2] Online Wikipedia: Web Conferencing http://en.wikipedia.org/wiki/Web_conferencing, Last accessed: 22.11.2005.

[3] Website: IRChttp://irc.pages.de/intro.html, Last accessed: 20.11.2005.

[4] Online Wikipedia: Video Conferencing http://en.wikipedia.org/wiki/Video_conferencing, Last accessed: 22.11.2005.

[5] Yoann Hinard and Hatem Bettahar and Abdelmadjid Bouaballah: MCDA: Une Architecture de Sécurité et de Tarification pour la Diffusion de Contenu Multicast. 4ème Conférence sur la Sécurité et Architectures Réseaux, SAR2005

[6] F. Bergadano and D. Cavagnino and B. Crispo: Multicast Security http://security.di.unito.it/research/cambridge.pdf

[7] G. Carle, F. Hartanto, M. Smirnov and T. Zseby. Charging and Accounting for QoS-Enhanced IP Multicast. Proc. of Protocols for High-Speed Networks (PfHSN), Salem, MA, August 1999.

[8] Cisco Systems: AAA Overview http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/12cgcr/secur_c/scprt1/scaaa.htm, Last accessed: 24.11.2005.

[9] B. Stiller and P. Racz and C. Morairu: Praktikum Mobile Systeme http://www.ifi.unizh.ch/csg/docs/WS05/MobSysLab/mobsys_praktikum.pdf, Last accessed: 24.11.2005.

[10] Adrian Perrig, Ran Canetti, J.D. Tygar, and Dawn Xiaodong Song. Efficient authentication and signing of multicast streams over lossy channels. In IEEE Symposium on Security and Privacy, May 2000.

[11] A. Tanabe, D. Andou, K. Izutsu, T. Hayashi, H. Tohjo: IGAP: IP Multicast Management Protocol that can collaborate with User Authentication. 2003 Asia-Pacific Network Operations and Management Symposium, October 1 - 3, 2003

[12] Online Wikipedia: Authentication http://en.wikipedia.org/wiki/Authentication, Last accessed: 22.11.2005.

[13] Online Wikipedia: Authorisation `http://en.wikipedia.org/wiki/Authorisation`, 22.11.2005.

[14] Online Wikipedia: Triple-A System `http://de.wikipedia.org/wiki/Authentication_Authorization_Accounting`, Last accessed: 22.11.2005.

[15] Online Wikipedia: AAA Protocol `http://en.wikipedia.org/wiki/AAA_protocol`, Last accessed: 22.11.2005.

[16] Website: Authentication Authorization and Accounting `http://searchsecurity.techtarget.com/sDefinition/0,,sid14_gci514544,00.html`, Last accessed: 24.11.2005.

[17] IETF Request for Comment: IP Multicast Applications: Challenges and Solutions

[18] Online Wikipedia: RADIUS http://en.wikipedia.org/wiki/RADIUS, Last accessed: 22.11.2005.

[19] Online Wikipedia: Password Authentication Protocol `http://en.wikipedia.org/wiki/Password_authentication_protocol`, Last accessed: 22.11.2005.

[20] Online Wikipedia: CHAP `http://en.wikipedia.org/wiki/Challenge-handshake_authentication_protocol`, Last accessed: 22.11.2005.

[21] Online Wikipedia: Extensible Authentication Protocol `http://en.wikipedia.org/wiki/Extensible_Authentication_Protocol`, Last accessed: 22.11.2005.

[22] Online Wikipedia: DIAMETER `http://en.wikipedia.org/wiki/DIAMETER`, Last accessed: 22.11.2005.

[23] Online Wikipedia: TACACS `http://en.wikipedia.org/wiki/TACACS`, Last accessed: 22.11.2005.

[24] B. Stiller, G. Fankhauser, B. Plattner, and N. Weiler. Charging and accounting for integrated internet services - state of the art, problems, and trends. In Proc. INET '98, Geneva, Switzerland, July 1998.

[25] Online Wikipedia: Pricing `http://en.wikipedia.org/wiki/Pricing`, Last accessed: 22.11.2005.

[26] Tanja Zseby: Policy-based Accounting `http://www.aaaarch.org/merit/azseby/sld023.htm`, Last accessed: 22.11.2005.

[27] M. Czyrnek, M. Lubonski, C. Mazurek: Authentication, authorization and accounting in distributed multimedia content delivery system. TERENA Networking Conference 2003, Zagreb, 19-22.05.2003.

[28] C. de Laat, G. Gross, L. Gommans, J. Vollbrecht, D. Spence: Generic AAA Architecture, RFC 2903.

[29] Bill Woodcock, Zhi-Li Zhang: A Straw-Man Pricing Model Addressing the Multicast Deployment Problem. Packet Clearing House, January 2003

[30] J. Chuang and M. Sirbu. Pricing multicast communications: A cost based approach. In Proc. of the INET'98, 1998.

[31] M. Adler and D. Rubenstein. Pricing Multicasting in More Practical Network Models. Technical report, University of Massachusetts UM-CS- `http://citeseer.ist.psu.edu/adler02pricing.html`, Last accessed: 01.02.2006.

[32] K. Ravindran and Ting-Jian Gong: Cost Analysis of Multicast Transport Architectures in Multiservice Networks. IEEE/ACM Transactions on Networking (TON) Volume 6 , Issue 1 (February 1998).

[33] S.Herzog, S.Shenker: Sharing the Cost of Multicast Trees: An Axiomatic Analysis ,ACM SIGCOMM'95, Aug. '95, Cambridge, MA, USA

[34] S. Herzog. Accounting and Access Control for Multicast Distributions: Models and Mechanisms, Ph.D. thesis, University of Southern California, August 1996.

[35] H. W. Holbrook and D. R. Cheriton: IP multicast channels: EXPRESS support for large-scale single-source applications. ACM SIGCOMM Computer Communication Review, Volume 29 , Issue 4 (October 1999)

[36] G. Carle, M. Smirnov, and T. Zseby: Charging and accounting architecture for IP multicast integrated services over ATM. 1998.

[37] B. Briscoe: The direction of valueflow in connectionless networks. Proceedings of the First International COST264 Workshop on Networked Group Communication, 1999.

[38] P. Hurley, B. Stiller, T. Braun, Charging Multicast Communications Based on a Tree Metric, H. Einsiedler, Proceedings of the GI Multicast Workshop'99, Braunschweig, Germany, 20-21 May, 1999.

[39] JJ. Feigenbaum, C. Papadimitriou, and S. Shenker. Sharing the cost of multicast transmissions. Journal of Computer and System Sciences (Special issue on Internet Algorithms), 63:21-41, 2001.

[40] D. Waitzman, C. Partridge, S. Deering: Distance Vector Multicast Routing Protocol. RFC 1075, November 1988.

[41] A. Adams, J. Nicholas, W. Siadak : Protocol Independent Multicast - Dense Mode: Protocol Specification. RFC 3973, January 2005.

[42] A. Ballardie: Core Based Trees Multicast Routing - Protocol Specification. RFC 2189, September 1997.

[43] D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. Liu, P. Sharma, L. Wei: Protocol Independent Multicast - Sparse Mode: Protocol Specification. RFC 2362, June 1998.

Kapitel 2

Incentive Strategies for Cooperation in Ad-hoc Networks

Beat Affolter, Simon Bleher, Christian Jaldón

Mobile Ad-hoc Netzwerke werden in Zukunft wahrscheinlich eine wichtige Stellung im täglichen Leben einnehmen. Sie vernetzen verschiedene digitale Geräte auf einfache Art und Weise. Um die Netzwerkfunktionalitäten entscheidend zu verbessern, ist eine gute Kooperation zwischen den einzelnen Teilnehmern nötig. Dank ihr kann die Abdeckung trotz geringerer Strahlung vergrössert und Energie gespart werden. In der Forschung sind verschieden Ansätze zur Verbesserung der Kooperation in Ad-hoc Netzwerken entwickelt worden. Teilweise sind es erst theoretische Konzepte, andere wiederum sind bereits implementiert. In dieser Arbeit werden fünf Konzepte vorgestellt und ihre Stärken und Schwächen beleuchtet. Durch die grossen Unterschiede nur schon in den Voraussetzungen oder dem Ausarbeitungsgrad ist ein Vergleich zwischen ihnen schwierig. Dennoch zeichnen sich gewisse Trends ab die am Ende der Arbeit kurz angeschnitten werden.

Inhaltsverzeichnis

2.1 Einleitung

Sie versuchen mit dem Laptop über ein Mobiltelefon auf das Internet zu gelangen? Wohlbekannt. Das VoIP-Gespräch mit einem Freund wird von ihrem PDA über das Laptop des Sitznachbars zum nächsten Internetzugangspunkt geleitet? Zukunft!

Die mobile Vernetzung von verschiedenen, sich einander sehr wahrscheinlich unbekannten Geräten wird heutzutage immer interessanter. Kaum eine neu auf dem Markt erschienene Anwendung verzichtet heute auf Kommunikationsmittel. Um die Kommunikation zu verbessern oder für kleine und abgelegene Stationen erst zu ermöglichen, sind neue Technologien notwendig. Eine davon sind Ad-hoc Netzwerke.

Diese Arbeit befasst sich mit der Kooperation in Ad-hoc Netzwerken. Da wir gute Vorkenntnisse des Lesers im Bereich Netzwerke und Kommunikation voraussetzen, werden nicht alle technischen Begriffe erläutert. Sie können aber in Standardliteratur wie [1] nachgeschlagen werden.

2.2 Ad-hoc Netzwerk

Dieses Kapitel gibt eine kleine Einführung in Ad-hoc Netzwerke und stellt die Vorteile und Probleme von Kooperation in Ad-hoc Netzwerken vor. Im Englischen ist anstelle von Ad-hoc Netzwerk auch der Ausdruck MANET (Mobile Ad-hoc Network) gebräuchlich. Ausführlichere Informationen findet man auch in der Studie [2], dass für dieses Kapitel als Quelle gedient hat.

2.2.1 Ein Beispielsszenario

Der grosse Konferenzsaal ist voll besetzt. Dringend sollte vor Beginn des nächsten Vortrags noch kurz Kontakt mit einem Geschäftspartner aufgenommen werden. Für das VoIP-Gespräch ist der PDA schnell zur Hand. Er hat jedoch mit seiner schwachen WLAN-Antenne keine Chance sich direkt mit einem Zugangspunkt am Rande des Saals zu verbinden. Zum Glück befindet sich zwei Reihen weiter ein eingeschaltetes Laptop, das mit dem Internet verbunden ist. Über dieses gelingt es dem PDA das gewünschte Gespräch schlussendlich aufzubauen.

2.2.2 Merkmale

Ein Ad-hoc Netzwerk kann sehr unterschiedlich aufgebaut sein. In den allermeisten Fällen sind jedoch sich nicht bekannte, mobile Teilnehmer miteinander verbunden. Teilweise kann ein Zugang zum Internet bestehen und in wenigen Fällen ist auch eine zentrale Administration vorhanden. Die Teilnehmer müssen sich aber generell selbst organisieren und die Pakete gegenseitig austauschen.

2.2.3 Einordung

Die Einordung von Ad-hoc Netzwerken fällt relativ leicht. Der Einfluss einer Basisstation ist in den meisten Fällen äusserst gering. Zudem ist das Routing von Paketen über mehrere Stationen (Multihop) möglich.

Als Gegensatz könnenn GSM-Netzwerke angesehen werden. Heutige Mobiltelefone kommunizieren nur direkt mit der Infrastruktur ihres Anbieters, der gleichzeitig die volle Kontrolle über den Datenverkehr ausübt.

Das bekannte WLAN (IEEE 802.11) ist normalerweise einem GSM-Netzwerk sehr ähnlich. Es kann zwar auch in einem Ad-hoc Modus betrieben werden, jedoch fehlt ihm dann die Möglichkeit Pakete über mehrere Stationen zu senden.

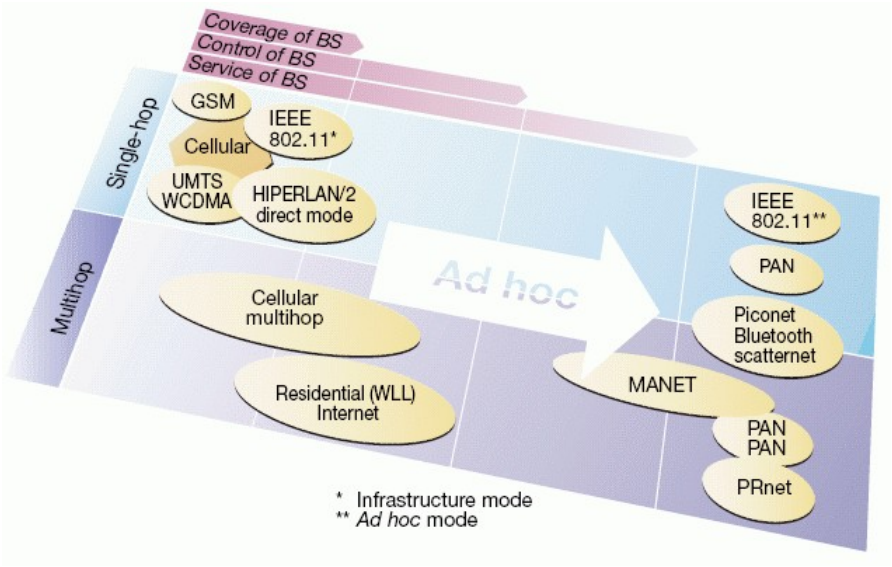


Abbildung 2.1: Einordnung von MANETs [Ericsson]

2.2.4 Einsatzgebiete

Ad-hoc Netzwerke sind heutzutage noch nicht weit verbreitet und dienen hauptsächlich Forschungszwecken. In Zukunft werden jedoch verschiedenen Anwendungsgebiete von den Vorteilen dieser neuen Technik profitieren.

Aus dem obenerwähnten Beispielsszenario ist ersichtlich, dass Ad-hoc Netzwerke auf spontane Weise entstehen können. Meistens dienen sie zur Verbesserung der Kommunikation, die mit den heutigen Mitteln in manchen Situationen nicht ausreichend sichergestellt werden kann. Zum einen kann dies an Orten mit besonders vielen Teilnehmern wie Konferenzen oder öffentliche Räume Vorteile bringen. Ebenfalls ist auch der Einsatz in schlecht erschlossenen, abgelegenen Gebieten denkbar, wie es beim 100$-Laptop des MIT Media Lab der Fall sein wird [3]

2.3 Kooperation in Ad-hoc Netzwerken

2.3.1 Herausforderungen

Es stellen sich verschiedene Herausforderungen an ein Ad-hoc Netzwerk, die nachfolgend kurz erörtert werden.

- **Verteilte Netzwerkfunktionen:** Da die Teilnehmer nicht auf ein bestehendes Netzwerk zurückgreifen können, müssen alle gemeinsam die Funktionen bereitstellen.
- **Dynamische Netzwerktopologie:** Durch die Mobilität der Teilnehmer verändert sich das Netzwerk laufend. Spezielle Beachtung muss deshalb dem Routingprotokoll geschenkt werden.
- **Wechselnde Verbindungsqualität:** Wegen möglicherweise weiten Wegen über mehrere Stationen bis zum gewünschten Teilnehmer, kann die Verbindungsqualität stark schwanken. Geeignete Fehlerkorrekturalgorithmen sind deshalb zur Verbesserung der Verbindung nötig.
- **Begrenzte Energieressourcen:** Auch kleinere Geräte, wie z.B. PDAs, können in einem Ad-hoc Netzwerk präsent sein. Da ihre Energiereserven vergleichbar klein sind, können sie unter Umständen nicht gleich wie die anderen Teilnehmer behandelt werden.

2.3.2 Kooperation in Ad-hoc Netzwerken

Ein Ad-hoc Netzwerk ohne Kooperation funktioniert nur begrenzt. Je besser die Kooperation funktioniert, desto mehr kommen die Vorteile dieses Netzwerktyps zum Tragen. Zum einen kann die Anzahl der Antennen einer vorhandenen Infrastruktur herabgesetzt werden. Durch das Routing der Teilnehmer bleibt die Abdeckung dennoch erhalten oder wird in den meisten Fällen sogar verbessert. Durch die geringeren Funkdistanzen sinkt aussserdem der Energieverbrauch.

Probleme bereiten vor allem die Implementierung eines guten Routingprotokolls und die Sicherheit. Durch die Mobilität ändern sich die Wege der Pakete zum gewünschten Ziel ständig. Ohne ein Routingprotokoll, das sich genügend schnell anpasst, ist ein mobiles Ad-hoc Netzwerk undenkbar. Die Abhängigkeit von den anderen Teilnehmern kann von einzelnen ausgenützt und missbraucht werden. Statt fremde Pakete weiterzuleiten könnten sie zur Optimierung des eigenen Durchflusses abgewiesen werden. Ein gutes Sicherheitskonzept ist deshalb unabdingbar.

Wie bereits unter 2.3.1 beschrieben, müssen gewisse Herausforderungen vorgängig gemeistert werden, damit ein erfolgreiches Routing stattfindet und die Kommunikation zwischen den Teilnehmern funktioniert. Selbst wenn diese Hürden technisch überwunden werden

können stellt sich die Frage, weshalb ein Teilnehmer im Netz überhaupt an der Kommunikation teilnehmen soll. Nebst dem Versenden und Empfangen der eigenen Datenpakete bringt eine Teilnahme wesentlich mehr Nachteile als Vorteile mit sich:

- **Energieverlust:** Das Weiterleiten, Forwarding, von Datenpaketen konsumiert den grössten Teil der Energie. Vor allem bei mobilen Kleingeräten ist Energie einer der kostbarsten Ressourcen. Gemäss [4] wird ca. 80% der Energie für das Forwarding konsumiert.
- **Bandbreite:** Wenn zum eigenen Datenverkehr noch derjenige von fremden Knoten dazukommt, dann resultiert dies in einer geringeren Bandbreite, die für einen Knoten zur Verfügung steht
- **Rechenkapazität:** Nicht jedes mobile Endgerät verfügt über genügend Rechenkapazität um eigene wie auch fremde Jobs zu verarbeiten. Vor allem hinsichtlich der komplizierten Routingalgorithmen haben kleinere Geräte Mühe hinreichende Performance zu leisten.
- **Psychologische Barrieren:** Nebst den oben erwähnten Hindernissen befindet sich hinter jeder Technologie der Faktor Mensch. Die Kenntnis, dass ein *Fremder* seine Daten über den eigenen Computer verschickt hat bei vielen eher eine abscheuchende als motivierende Wirkung.

Die Frage nach der Sicherheit wurde noch nicht diskutiert. Das Sicherheitsrisiko von Funknetzwerken soll hier aber nicht abgehandelt werden. Viel mehr interessiert uns in diesem Kontext das Auftreten von *Falschspielern* oder auch *boshaften Knoten*, wie sie in der Literatur genannt werden. Eine Kommunikation in mobilen Ad-hoc Netzwerken ist solchen Teilnehmern vollkommen ausgeliefert. Was geschieht also, wenn ein Knoten bewusst keine Datenpakete von Dritten weiterleitet, sondern diese einfach verwirt. Was geschieht wenn solche Knoten die Datenpakete Dritter modifiziert?

Dieser Arbeit befasst sich verschiedenen Konzepten, wie die Kooperation in mobilen ad hoc Netzwerken gefördert werden kann und wie *Falschspieler* entdeckt und dementsprechend bestraft werden können. CineMA - Cooperation Enhancement in Manets, eines am Institute of Computer Science der Universität Bonn entworfenes Model, begegnet den erwähnten Herausforderungen, indem es über bestimmte Module, sowohl Störungen im Netz, also unkonventionelles Verhalten der Knoten, proaktiv erkennt und gleichzeitig einen Mechanismus initiert, damit *Falschspieler* eine möglichst kleine Auswirkung auf die Qualität der Kommunikation ausüben können.

2.3.3 Klassifikation der Kooperationsmechanismen

Kooperationsmechanismen in mobilen Ad-hoc Netzwerken kann man anhand von zwei Hauptklassen differenzieren:

- Detection based Approach
- Motivation based Apprach

Detection based Approach

Erkennungsbasierte Ansätze zielen auf proaktives Handeln gegen Störungen, so genannte Incidents, im Netz ab. Dabei horcht ein Mechanismus im Hintergrund die Netzwerkaktivität und versucht zu erkennen, ob die Datenpakete gemäss Routingtabelle tatsächlich weitergeleitet werden.

Die Umsetzung solcher Ansätze findet man in vielen Variationen an. Einige Ansätze werden mit einem Reputationssystem gekoppelt, welches die Teilnehmer anhand von vordefinierten Kriterien bewertet. Die Teilnehmer tauschen gegenseitig ihre Informationen über die Netztopologie und Informationen über die direkten Nachbarn aus. Diese Informationen in kombinierter Form kann dazu verwendet werden, um ein Ranking über die Teilnehmer zu erstellen. Knoten mit negativer Bewertung werden ausgeschlossen oder bestraft, in dem ihre Aktivität im Netz gedrosselt wird.

Motivation based Approach

Motivationsbasierte Ansätze verfolgen das Ziel, Anreize für eine Teilnahme an der Kommunikation zu schaffen.

Vereinfacht dargestellt wird ein Knoten, welcher eigene Datenpakete versendet bestraft, hingegen Knoten, welche fremde Pakete weiterleiten belohnt.

Das Modell *Charging and Rewarding*, siehe Abschnitt 2.6, gehört zu dieser Klasse. Dabei versucht es ökonomische Anreize für das Weiterleiten von Datenpaketen zu schaffen. Dem Sender wird ein gewisser Betrag verrechnet, dem weiterleitenden Knoten wird ein entsprechender Betrag gutgeschrieben. Sofern die Pakete über einen positiven Kontostand verfügen sind sie dazu berechtigt, eigene Informationen über das Netz zu versenden.

CineMA verfolgt einen erkennungsbasierten Ansatz. Über ein integriertes Modul, dem *Watchdog Modul*, versucht CineMA das Netz auf irreguläre Aktivitäten zu untersuchen. Diese Informationen werden an ein weiteres Modul übergeben, dem *Reputation System Modul*, das anhand einer Kooperations-Funktion die Informationen auswertet um diese wieder an ein anderes Modul weiterzuleiten, welches einen Bestrafungsmechanismus zur Verfügung stellt.

Das Exklusive an CineMA ist das Konzept des *gruppenbasierten Ansatzes*. Im Gegensatz zu anderen erkennungsbasierten Ansätzen muss bei CineMA nicht jeder Knoten mit CineMA funktionieren. Eine Teilmenge der Netzwerkteilnehmer genügt, um eine zufrieden stellende Kommunikation, inklusive Erkennung und Bestrafung von Falschspielern, zu gewährleisten. Das ist einer der grossen Vorteile von CineMA gegenüber anderen Kooperationsmechanismen in mobilen Ad-hoc Netzwerken.

2.4 CineMA: Cooperation Enhancement in Manets

2.4.1 Voraussetzungen

Ein CineMA lauffähiges System bedingt nur ein paar weniger Voraussetzungen. Durch den zuvor erwähnten gruppenbasierten Ansatz ist eine Kooperation in einem mobilen Ad-hoc Netz möglich, auch wenn nicht alle Knoten CineMA implementiert haben. Dies impliziert aber auch, dass eine völlige Auschliessung von Falschspielern nicht möglich ist.

An die Infrastruktur wird keine Voraussetzung gestellt. Ob es in einem reinen Ad-hoc Neztwerk oder über ein zentral organisiertes System implementiert sein muss, zum Beispiel über einen Internet Service Provider, wird nicht definiert.

Selbst an die Hardware und Software werden in der jetzigen Version keine Voraussetzungen gestellt. CineMA kann sowohl in Hardware eingebettet sein oder in Software programmiert werden. Auch werden in der Spezifikation keine Anforderungen erwähnt.

Das Routing wurde in der momentanen Version mittels des *Dynamic Routing Protocol* realisiert - einer der wenigen notwendigen Voraussetzungen. Zukünftige Versionen sollen aber unabhängig vom Routing-Protokoll implementiert werden.

Das Sicherheitskonzept wird völlig ausser Betracht gelassen. Die Möglichkeit bestünde, die Sicherheit respektive die Verschlüsselung der Daten in die Hardware zu verlagern. Ähnlich wie beim GSM-Netz könnte man mittels einer SIM-Karte das System sicher gegen Angreifen gestalten [4].

Das DSR-Protokoll macht vom *promiscuous mode* [5] des WLAN's gebrauch. Somit könne alle Teilnehmer jegliche Pakete lesen und sich damit ein Bild über die Topologie des Netzes machen. Gerade solche Informationen könnten missbraucht werden, um eine Kommunikation zu stören oder zu verhindern. Der Austausch der Information über das Netzwerk, oder *Group Communication* wie er in der Literatur genannt wird, muss deshalb geschützt werden. Verschlüsselungsalgorithmen sind dringendst notwendig.

2.4.2 Architektur

Die Architektur von CineMA ist ziemlich simpel ausgelegt: Keine komplizierten Algorithmen, kein komplexes Reputation System, kein komplexer Bestrafungsmechanismus. Das Modell besteht aus drei Modulen, wie in Abbildung 2.2 illustriert.

Im Folgenden wird genauer auf die drei Hauptmodule eingegangen und näher erklärt, wie der Kooperationsmechanismus funktioniert. Der Routing-Algorithmus selbst ist nicht Bestandteil dieser Arbeit.

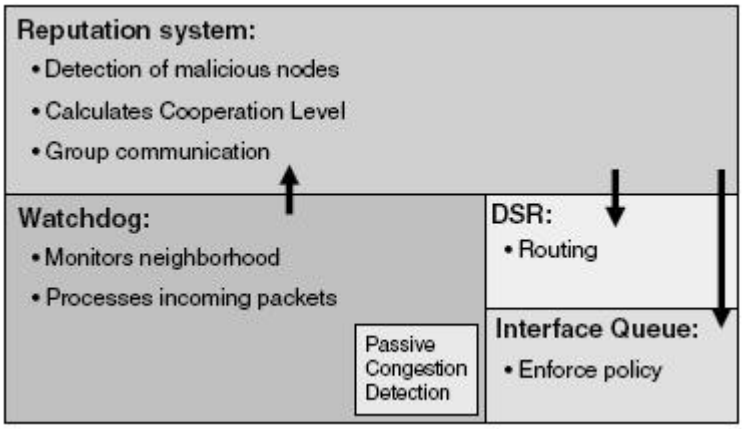


Abbildung 2.2: Architektur CineMA [6]

Das Watchdog Modul

Das Watchdog Modul hat im Wesentlichen folgende Aufgaben:

- Empfangene und gesendete Pakete verfolgen und überprüfen, ob sie ihre Destination erreicht haben
- Monitoring der Netzwerkaktivität: Störungen, Fehler, Performance
- Log-Listen erstellen und Daten an das Reputation System weiterleiten

Dadurch dass das Modul im *promiscous mode* läuft, werden alle gesendeten Datenpakete aufgefangen. Das Watchdog Modul überprüft den DSR-Header und erkennt, ob Knoten die Pakete an ihre Destination weitergeleitet haben.

Das Reputation System

Die Hauptaufgaben des Reputation Systems sind:

- Die vom Watchdog Modul erhaltenen Daten zu analysieren und auf eventuelle Falschspieler zurückzuführen
- Berechnung des *Cooperation Levels*
- Handhabung der *Group Communication*

Ein CineMA-Knoten führt für alle Knoten in seiner Nachbarschaft jeweils eine Liste mit zwei Einträgen:

- `Incoming List`: Hier werden alle Pakete eingetragen, die ein Knoten empfangen hat

- `Forwarding List`: Hier werden alle Pakete eingetragen, die ein Knoten weiterleitet

Der Erkennungsmodus soll illustrativ anhand eines Beispielszenarios erklärt werden, wie Abbildung 2.3 zeigt.

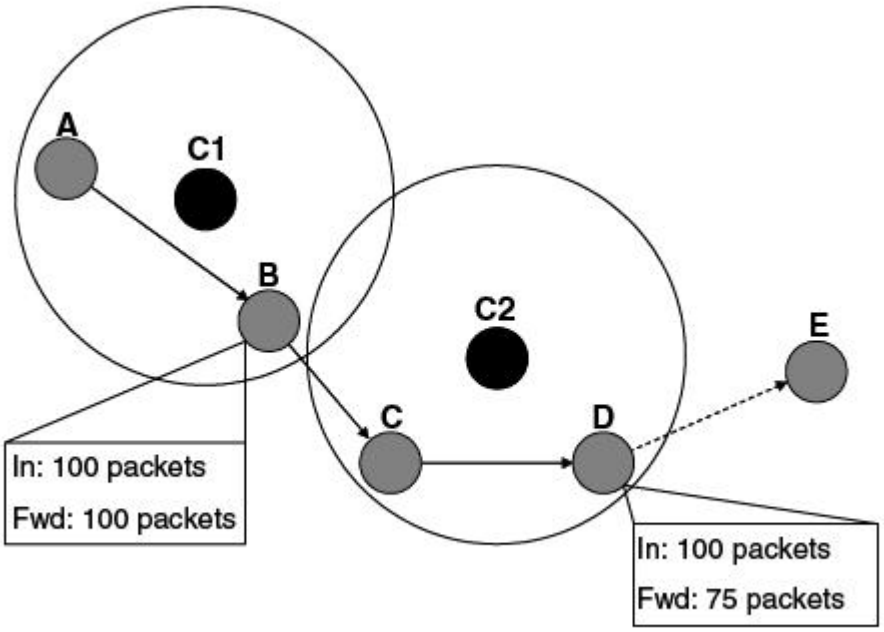


Abbildung 2.3: Beispielszenario CineMA Ad-hoc Netzwerk [6]

Die grau-schraffierten Kreise stellen Knoten ohne CineMA-Implementation. Die schwarzen Kreis solche mit CineMA.

Angenommen Knoten A möchte eine Nachricht an Knoten E senden. Die Route von A nach B wird in den DSR-Header notiert und über das Netz verschickt. CineMA-Knoten C1 erkennt diese Aktivität und fügt in die entsprechende Incoming- wie auch Forwarding List einen Eintrag.

Wenn also das Versenden eines Pakets von A nach E erfolgreich durchgeführt werden konnte, hat jeder Knoten jeweils gleiche viele Einträge in der Incoming wie in der Forwarding List. Nimmt man nun die Anzahl der Einträge der beiden Listen und dividiert man sie miteinander, erhält man einen Quotienten namens *Cooperation Level*. Im optimalen Falle beträge der *Cooperation Level* immer eins.

Nehmen wir nun an, dass Knoten D ein Falschspieler sei und jedes vierte Paket verwerfen würde. Die Kommunikation zwischen A und E würde also verfälscht! (durch die gestrichelte Linie gekennzeichnet). In der Forwarding List des Knotens E hätte es also zu einem Viertel weniger Einträge als in seiner Incoming List. Der Quotient würde als 0.75 betragen.

Eine solche Abweichung vom Optimum wird vom Reputation System als Störung interpretiert. Diese Information verwendet es wiederum um es an die Queue zu senden, welche für die weitere Verarbeitung zuständig ist.

Das illustrierte Szenario macht uns noch auf eine andere Problematik aufmerksam. Im Falle dass Knoten C ein Falschpieler wäre, würde CineMA-Knoten C1 lediglich erkennen, dass Knoten B ein Paket an C sendet, weil es C ausserhalb seiner Reichweite liegt. Und CineMA-Knoten C2 würde nur Pakete welche von C gesendet werden erkennen, wiederrum weil B ausser Reichweite für CineMA-Knoten C2 ist.

Um diesem Problem entgegenzuwirken reicht eine lokale Sicht nicht aus. CineMA-Knoten C1 und C2 haben nur eine beschränkte Sicht des Netzwerkes. Eine globale Sicht des Netzwerkes ist dringend notwendig. Die Informationen über die Netzwerktopologie sowie die beiden Listen müssen unter den Knoten ausgetauscht werden. Die Tatsache, dass diese Informationen von Angreifern zu Nutze gemacht werden könnten erklärt, weshalb die Daten unbedingt verschlüsselt werden müssen.

Die Interface Queue

Die Interface Queue hat folgende Aufgabe:

- Bestrafung von Falschspielern anhand des vom Reputation System gelieferten *Co-operation Levels* (cl)

Für die Umsetzung solcher Bestrafungsmechanismen gibt es verschiedene Ansätze. In [6] werden 2 Modelle vorgestellt:

- `Dropping Packets`
- `FIFO/Leaky Bucket`

Eine vereinfachte Form eines Bestrafungsmechanismus funktioniert, indem Pakete von Falschspielern verworfen werden (Dropping Packets). Dabei macht die Interface Queue vom Cooperation Level gebrauch und berechnet damit die Wahrscheinlichkeit, dass ein Paket eines boshaften Knotens verworfen wird:

- p = 1 - cl

Angenommen der Cooperation Level (cl) eines Knotens sei 0.75, so ist die Wahrscheinlichkeit, dass eine dessen Pakete verworfen werden 0.25

Je kleiner also der Cooperation Level, desto grösser die Wahrscheinlichkeit, dass die Pakete von den CineMA-Knoten verworfen werden.

Eine andere Methode stellt der *FIFO/Leaky Bucket* dar. Das System führt jeweils zwei Datenbehälter:

- In die FIFO-Queue landen alle Pakete von sich korrekt verhaltenden Knoten

- In den Leaky-Bucket (übersetzt *löchriger Behälter*) landen alle Pakete von Falschspielern

Ein `Dispatcher` (eine Analogie zu einer Abfertigungshalle) entscheidet darüber, in welchen Behälter die Pakete jeweils gelangen.

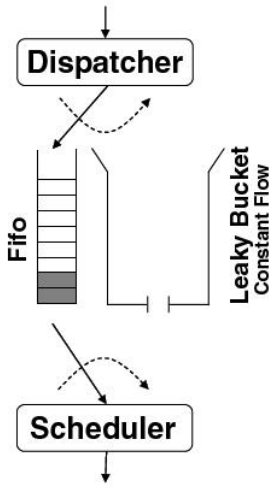


Abbildung 2.4: Interface Queue CineMA

Pakete aus dem FIFO-Behälter werden vom `Scheduler` so rasch wie möglich weiterverarbeitet und verschickt. Pakete des Leaky-Bucket hingegen werden nur mit einer bestimmten Konstante verarbeitet.

Beim Versuch ein Paket des Leaky-Buckets zu versenden wird überprüft, ob der MAC-Layer nicht gerade damit beschäftigt ist, ein Paket des FIFO-Behälters zu verarbeiten. Sollte dies der Fall sein, so wird garantiert, dass das nächste Paket aus dem Leaky-Bucket stammt. Somit kommt eine gewisse Fairness ins Spiel und das System gewährleiset, dass auch Falschspieler Pakete versenden können.

2.4.3 Verwandte Systeme

Andere Kooperationsmechanismen versuchen statt dem Berechnen eines *Cooperation Levels* einen anderen Ansatz zu verfolgen. In [4] ist beispielsweise die Rede eines Zählers, dem *Nuglet Counter*. Der Nuglet Counter gibt stets Auskunft über den *Kontostand* eines Knotens. Sofern der Zähler positiv ist, darf ein Knoten überhaupt eigene Pakete senden. Fürs Weiterleiten von fremden Datenpakete wird ein Knoten belohnt, indem sein Zähler erhöht wird. Will ein Knoten hingegen eigene Pakete versenden, so wird sein Zähler verkleinert.

Im Detail funktioniert der Mechanismus wie folgt:

- Knoten will eigenen Pakete versenden: Zuerst wird die Anzahl der Zwischenknoten berechnet. Wenn Stand des Zählers grösser als n ist, so darf der Knoten senden. Wenn der Wert hingegen kleiner als n ist, so wird ihm das Senden unterbunden.

- Knoten will fremde Pakete weiterleiten: Für das Weiterleiten eines Pakets wird sein Zähler um eins erhöht.

Ein solches System impliziert, dass Knoten zuerst fremde Pakete weiterleiten müssen, um überhaupt eigene Pakete versenden zu dürfen[1].

2.4.4 Beurteilung und Bewertung

CineMA weist gegenüber anderen Kooperationsmechanismen für mobile Ad-hoc Netzwerke eine ganze Anzahl an Vorteilen auf. Es ist nicht an spezielle Hardware oder Endgeräte gebunden und gibt somit enorm viel Handlungsspielraum für viele Anwendungsszenarien.

Da keine Voraussetzungen Punkto Netz-Infrastruktur gelten ist CineMA sowohl für reine Ad-hoc Netzwerke tauglich, wie auch für zentralisierte Dienste über einen Servide Provider.

Durch die simple Architektur lässt sich CineMA auf viele Plattformen implementieren. Der modulartige Aufbau verleiht dem System Flexibilität und kann leicht an spezielle Bedürfnisse angepasst werden.

Ein grosser Vorteil erweist sich durch den gruppenbasierten Ansatz. Da nur eine Teilmenge an CineMA-Knoten in einem Netzwerk vorhanden sein muss, funktioniert eine störungsfreie Kommunikation schon ab einer geringen Anzahl an CineMA-Knoten.

Der Ausdruck des Kooperationsgrades in einem mobilen Ad-hoc Netzwerk durch eine Funktion stellt in diesem Forschungsgebiet eine Innovation dar. Durch die noch so simple Funktion des Cooperation Levels kann die Förderung der Kooperation gefördert werden wie auch ein simpler und effizienter Bestrafungsmechanismus implementiert werden.

Simulationen haben erwiesen, dass das System gut skalierbar ist und ab ein gewisser Menge an CineMA-Knoten die Entdeckungsrate an Falschspielern zufrieden stellend hoch ist [6].

In der momentanen Version von CineMA ist das Vorhandensein des DSR-Routingprokoll eine Voraussetzung. Dies könnte einen Nachteil darstellen, wenn CineMA in Bereichen mit heterogenen Systemen eingesetzt werden soll.

Verschlüsselungsalgorithmen werden nicht spezifiziert. Somit ist es dem Anwender überlassen, welche Art von Verschlüsselung er für den Austausch der Informationen zwischen den Knoten verwenden will.

2.5 CASHNet

CASHNet wurde an der Uni Bern im Jahre 2004 zum ersten Mal publiziert [7]. Als Anreiz zur Förderung der Kooperation in einem Ad-hoc Netzwerk dient reales Geld. Eigene Pakete können nur gegen ein Entgelt versendet werden.

[1] Die Starthöhe des Zählers müsste ebenfalls in Betracht gezogen werden.

2.5.1 Voraussetzungen

Für den Einsatz von CASHNet müssen gewisse Voraussetzungen erfüllt sein. Als Grundbaustein muss in jedem Knoten ein fälschungssicheres System, z.B. eine Smartcard, eingebaut sein. Darauf werden alle benötigten Informationen und Methoden gespeichert. Ausserdem ist ein Routingprotokoll, das die Anzahl der Hops bis zur Basisstation kennt, wie z.B. AODV oder DSR zwingend. Um die ganze Organisation und Abrechnung vornehmen zu können ist auch ein Provider nötig, den der Teilnehmer von Zeit zu Zeit aufsucht.

2.5.2 Mechanismus

Der Mechanismus von CASHNet basiert auf einem Sicherheitssystem mit öffentlichen Schlüsseln die von einem zentralen Provider ausgegeben werden. Um Fälschungen zu erschweren haben sie nur eine beschränkte Laufzeit. Insgesamt können sechs verschiedene Phasen unterschieden werden, die nachfolgend vorgestellt werden.

Installationsphase Der Teilnehmer erhält von seinem Provider seine Schlüssel und ein Zertifikat. Auch kann er Kreditpunkte für das Versenden der eigenen Pakete gegen reales Geld erwerben. Falls er bereits Helferpunkte gesammelt hat, kann er auch diese als Zahlungsmittel einsetzen.

Authentifikationsphase Um im Netzwerk teilnehmen zu können muss sich ein Teilnehmer gegenüber den anderen Knoten zu erst authentifizieren. Dies geschieht anhand des öffentlichen Schlüssels und des Zertifikats. Um die Authentifikation nicht durch die langen Antwortzeiten von weitentfernten Teilnehmern zu verzögern, halten die Knoten die öffentichen Schlüssel ihrer Nachbarknoten in einer Liste gespeichert.

Paketerzeugungsphase Falls ein Teilnehmer ein Paket versenden will, müssen zuerst die Kosten ermittelt werden. Verkehr innerhalb des Netzwerkes also solcher, der nicht über die Basisstation geleitet wird, ist kostenlos. Sobald die Pakete das Netzwerk über den Provider verlassen, werden die Kosten anhand der Anzahl Hops bis zur Basisstation berechnet. Falls der Teilnehmer genügend Kreditpunkte für die Strecke besitzt, wird das Paket abgeschickt.

Paketannahmephase Bei Erhalt eines Pakets von einem benachbarten Knoten wird zu erst überprüft, ob das Paket von einem authentifizierten Teilnehmer stammt. Falls das Paket korrupt oder gefälscht ist, wird es verworfen.

Paketweiterleitungsphase Der Knoten schaut in seiner aktuellen Routingtabelle, welcher Teilnehmer für das gegebene Ziel der beste ist. Er signiert nun das Paket mit seinem Schlüssel und leitet es weiter. Gleichzeitig trägt er die Signatur des Pakets und den gewählten Knoten in eine interne Liste ein.

Belohnungsphase In der Belohnungsphase werden die einzelnen Teilnehmer für ihre geleisteten Weiterleitungen fremder Pakete mit Helferpunkten belohnt. Wenn der Knoten eine Bestätigung für ein erfolgreich versendetes Paket von einem anderem Knoten erhält, vergleicht er sie mit seiner internen Liste. Falls eine Übereinstimmung existiert, wird er mit einem Helferpunkt belohnt.

2.5.3 Untersuchung des Anreizsystems

CASHNet versucht Kooperation durch Einsatz von realem Geld in einem Ad-hoc Netzwerk zu fördern. Jedes selbstgenerierte Paket, das das eigene Netzwerk verlässt, muss mit Kreditpunkten bezahlt werden. Kreditpunkte können entweder beim Provider durch Geld oder durch erworbene Helferpunkte bezogen werden. Helferpunkte gibt es für jedes erfolgreich weitergeleitete Paket eines fremden Teilnehmers.

Da keine Mechanismen zur Benachteiligung von egoistischen Teilnehmern bestehen, ist jedem Knoten selber überlassen, ob er kooperieren will oder tiefer in seine eigene Tasche greift. So ist es auch am Rande eines Netzes liegenden Teilnehmern oder solche mit geringen Energieressourcen möglich ohne Einschränkungen das Netzwerk zu benützen.

2.5.4 Beurteilung und Bewertung

CASHNet ist ein Ansatz zur Förderung der Kooperation der einen Provider für die Organisation des Netzes voraussetzt. Die Teilnehmer erwerben von ihm ihre Hardware, Schlüssel und Kreditpunkte. Der Provider muss vor dem Betrieb nicht zu verachtende Investitionen tätigen, baut aber auch eine enge Beziehung zu seinen Kunden auf.

Die relativ offene Architektur lässt viele Freiheiten in der Wahl des Routingprotokolls oder z.B. der Hardware. Auch deshalb wird CASHNet wohl in verschiedenen Umgebungen eingesetzt werden können.

Der Anreiz zur Kooperation durch reales Geld ist sicherlich eine gut funktionierende Methode. Monetäre Ansätze werden heutzutage ja in fast allen Bereichen des täglichen Lebens erfolgreich eingesetzt. Wichtig ist jedoch ein gutes Sicherheitskonzept, das vor Missbrauch schützt. Durch die öffentliche Verschlüsselung sollte dies sichergestellt sein.

2.6 Charging and Rewarding

Bei *Charging and Rewarding* handelt es sich um eine Entwicklung verschiedener Forscher der ETH Lausanne und der Firma RSA aus den Vereinigten Staaten [9]. Als einziger, der in dieser Ausarbeitung aufgegriffenen Lösungsvorschläge, verzichtet Charging and Rewarding auf jeglichen Zwang zur Mitarbeit. Stattdessen wird versucht, allein mittels ökonomischen Aspekten genügend Anreize zu schaffen, um eine möglichst hohe Zahl an Teilnehmern zur Mitarbeit zu motivieren.

2.6.1 Voraussetzungen

Wie auch bei den anderen vorgestellten Ansätzen sollen an dieser Stelle zuerst die grundsätzlichen Vorraussetzungen aufgezeigt und erläutert werden.

Charging and Rewarding baut auf paketbasierte Kommunikation auf, klammert aber die Problematik des Routing aus. Diese soll speziell betrachtet werden, oder aber ein vorhandenes Protokoll, wie DSR, welches die geforderten Eigenschaften mitbringt, kann angenommen werden.[8]

Weiter wird eine vertrauenswürdige Basisstation angenommen. Dies ist wichtig, da jeglicher Datenverkehr über die Basisstation führen muss und diese für die Verrechnung und die Kontrolle des Verkehrs zuständig ist. Später wird auch aufgezeigt, wie diese Voraussetzung hinreichend erfüllt werden kann.

In diesem System gibt es meist Teilnehmer, welche ausserhalb der Reichweite einer Basisstation liegen, aber welche über andere Teilnehmer eine Verbindung zur Basisstation aufnehmen können. Aus Effizienzgründen wird zur Verschlüsselung ein symmetrisches Verschlüsselungsverfahren eingesetzt.[9]

2.6.2 Architektur

Als System wird eine Menge von Basisstationen angenommen, welche über einen schnellen Backbone miteinander verbunden sind. Ausserdem gibt es eine Menge von mobilen Stationen, wie zum Beispiel Mobiltelefone. Jeglicher Verkehr zwischen diesen Mobilstationen wird über die Basisstation (und ggf. den Backbone) geleitet und erst dann zur Empfängerstation. Ein eigentlicher ad-hoc Verkehr ist somit nicht möglich. Denn selbst zwei benachbarte Stationen können nicht direkt miteinander kommunizieren, sondern müssen den Umweg über mindestens eine Basisstation nehmen. Dies heisst auch, dass das System nicht in jedem Fall einen effizienten Weg wählt. Dafür reduziert es drastisch die Anzahl der Verbindungen, die eine einzelne Mobilstation kennen muss. So muss sie einzig und allein den richtigen Weg zur Basisstation kennen. Im Gegensatz dazu müsste sie im System ohne Basisstation die Erreichbarkeit jedes einzelnen Knotens kennen. Diese grundsätzliche Architekturentscheidung macht diesen Ansatz auch relativ schwer mit den anderen hier vorgestellten Vorschlägen vergleichbar.

Anreizsystem

Wichtigster konzeptueller Teil von *Charging and Rewarding* ist das Verrechnungskonzept, welches die Kernkomponente des Systems darstellt und zugleich sicherstellen soll, dass jederzeit die korrekten Anreize geschaffen werden. Es wird davon ausgegangen, dass es dem Sender einer Nachricht einen Nutzen bringt, diese Nachricht zu versenden. Deshalb ist auch er es, der für die Nachrichtenübermittlung bezahlt. Grundsätzlich kann das Modell aber auch so angepasst werden, dass ein beliebiger Nutzenempfänger des Systems zahlen muss.

Die weiterleitenden Stationen haben hingegen keinen Nutzen davon, wenn sie Nachrichten einer Drittpartei weiterleiten (z.B. zur Basisstation). Im Gegenteil: Sie müssen Energie und Sendeleistung zur Verfügung stellen, was für sie ein negatives Kosten-Nutzen Verhältnis bedeutet. Jeder rationale Agent wird in solch einem Falle die Weiterleitungsfunktion im

Netzwerk nicht wahrnehmen. Er wird dies erst machen, wenn er aus irgendeinem Grunde dazu gezwungen wird, oder er einen Nutzen aus diesem Verhalten ziehen kann, der grösser als sein Ressourcenverbrauch ist. Genau diesen Ansatz verfolgt man bei Charging and Rewarding. Alle weiterleitenden Stationen erhalten nämlich einen bestimmten Betrag für ihre Leistung gutgeschrieben.

Diese Funktion stellt den Ansatz vor eine grosse konzeptuelle Herausforderung. Wie kann die Basisstation überprüfen, ob ein Knoten tatsächlich an der Weiterleitung beteiligt war und dies nicht einfach behauptet, um sich unrechtmässig zu bereichern? *Charging and Rewarding* kann genau dies durch die Verschlüsselung sicherstellen. Indem jede weiterleitende Station das Paket mit ihrem geheimen Schlüssel zusätzlich verschlüsselt, kann die Basisstation bei der Entschlüsselung einwandfrei feststellen, ob das Paket tatsächlich auch durch die Hände dieser Station gegangen ist. Wie diese Authentifizierung der weiterleitenden Stationen genau gelöst ist, wird im nächsten Abschnitt erläutert.[9]

2.6.3 Protokolle und Implementierung

In diesem Abschnitt soll, im Gegensatz zur vorhergehenden konzeptuellen Einführung, auf die Implementierung und die verwendeten Protokolle eingegangen werden.

Allgemeines

Zur einfacheren Darstellung des Mechanismus wird an dieser Stelle ein repräsentatives Beispiel eingeführt und analysiert. Ausgangslage ist ein Netzwerk mit zwei Baisstationen BS 1 und BS 2 (siehe Abbildung 2.5), welche über einen Backbone verbunden sind. Die Mobilstationen MS A und MS D befinden sich jeweils ausserhalb der Reichweite der Basisstationen, erreichen jedoch beide eine Mobilstation (MS B und MS C), welche eine Verbindung mit einer Basisstation aufnehmen kann. Als Szenario nehmen wir an, dass MS A eine Nachricht an MS D senden möchte.

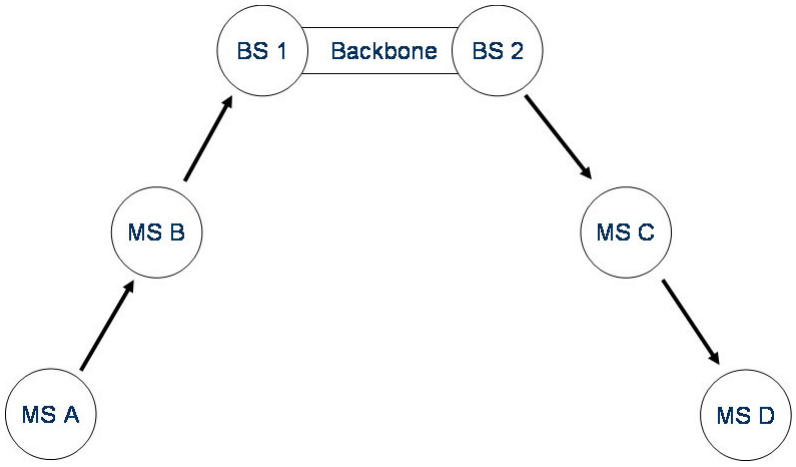


Abbildung 2.5: Setup des Beispiels

Verschlüsselung

Charging and Rewarding setzt vollständig auf symmetrische Verschlüsselung. Hauptgrund ist die eingeschränkte Rechenleistung von mobilen Geräten.

Ein Teilnehmer eines *Charging and Rewarding* Netzes muss sich vorab bei einem Provider registrieren und bekommt von ihm einen geheimen persönlichen Schlüssel, welcher ausser dem Teilnehmer und dem Netzbetreiber niemand kennen darf. Dies ist auch ein Grund, weshalb jeglicher Verkehr über eine Basisstation geleitet werden muss. Die übermittelte Nachricht wird nämlich jeweils mit dem persönlichen Schlüssel codiert. Die Basisstation wird die Nachricht entschlüsseln und für den Empfänger mit seinem persönlichen Schlüssel wieder verschlüsseln.

Verbindungsaufbau

Bevor eine Station mit einer anderen kommunizieren kann, muss eine Ende-zu-Ende Verbindung aufgebaut werden. Dies ist wichtig, um das Routing zu testen und sicherzustellen, dass eine korrekte Route gewählt wird. Ein weiteres Ziel ist es, die dazwischenliegenden Stationen über den bevorstehenden Datenverkehr zu informieren und der Basisstation die Möglichkeit zu geben, alle weiterleitenden Stationen zu authentifizieren. Es ist jeder Station freigestellt, ob sie sich am Datenverkehr beteiligen möchte. Je nach Entscheidung der einzelnen Stationen ist es möglich, dass ein Verbindungsaufbau fehlschlägt. Ein Verbindungsaufbau ist erfolgreich, wenn sowohl eine Verbindung von MS A zu BS 1 zu Stande kommt, wie auch eine Verbindung von BS 2 zu MS D.

Um eine Session zu starten sendet MS A eine initiator session setup request message, welche eine eindeutige ID enthält, die Route von MS A zur ersten Basisstation, und Informationen über den zu erwartenden Verkehr. Über diese Nachricht wird anschliessend ein Message Autenthication Code (MAC) berechnet. Dazu benutzt MS A ihren geheimen Schlüssel Ka. Dieser MAC wird an die Anfragenachricht angehängt.

Jede weiterleitende Station kann nun anhand der Information über den zu erwartenden Verkehr entscheiden, ob sie an diesem Prozess teilnehmen möchte oder nicht. Entscheidet sie sich dafür, berechnet sie (in unserem Beispiel ist das MS B) ebenfalls einen MAC, aber über die gesamte Nachricht, inklusive des MAC von MS A. Schliesslich leitet MS B die Nachricht an die Basisstation BS 1 weiter.

Die Basisstation wiederholt jetzt alle MAC-Berechnungen mit den Schlüsseln von MS A und MS B. Ist das Resultat dasselbe, das sie erhalten hat, leitet sie die Nachricht über den Backbone an die Basisstation weiter, von welcher aus die Zielstation erreichbar ist. Falls das Resultat nicht übereinstimmt, wird die Anfrage verworfen.

Die Zielbasistation schickt anschliessend eine ähnliche Nachricht über verschiedene Stationen zur Zielstation (hier MS D). Wenn diese die Anfrage erhält, berechnet sie darauf einen MAC und sendet die Nachricht wieder zurück. Die Prozedur funktioniert dann äquivalent zum Verbindungsaufbau MS A zu BS 1.

Ist auch dieses Setup erfolgreich verlaufen, versenden BS 1 und BS 2 je eine Bestätigungsnachricht zu MS A und MS B. Der Nachricht werden eine Reihe von MACs angehängt. Für jede Station auf dem Weg zum Ziel wird ein MAC mit dem entsprechenden Schlüssel generiert. So kann zum Beispiel MS B die Echtheit dieser Nachricht anhand des für sie bestimmten MAC überprüfen.[2]

Datenverkehr

Der eigentliche Datenverkehr funktioniert ähnlich wie ein Verbindungsaufbau. Hauptunterschied ist, dass mit dem Schlüssel nicht nur ein MAC berechnet, sondern zusätzlich die Nachricht verschlüsselt wird. Anhand unseres Beispiels kann dies anschaulich aufgezeigt werden:

Wenn MS A über die eingerichtete Ende-zu-Ende Verbindung eine Nachricht zu MS D versenden möchte, wird zuerst wieder ein MAC berechnet und anschliessend die Nachricht von MS A verschlüsselt. Jede weiterleitende Station verschlüsselt die Nachricht wiederum. Die Basisstation entschlüsselt die Nachricht mit den entsprechenden Schlüsseln anschliessend in umgekehrter Reihenfolge und prüft den von MS A generierten MAC. Auf diese Weise ist sowohl die Integrität, als auch die Geheimhaltung des Inhalts gewährleistet. Die Basisstation an der Downstream Verbindung berechnet nun zuerst einen MAC mit dem Schlüssel von MS D und verschlüsselt die gesamte Nachricht mit dem Schlüssel von MS D und von MS C. Die weiterleitende Station entschlüsselt mit dem persönlichen Schlüssel die Nachricht wieder, bevor sie das Ergebnis weiterleitet. Da diese Nachricht aber immer noch mit dem Schlüssel von MS D verschlüsselt ist, können auch die weiterleitenden Stationen auf dem Downstream Ast den Inhalt nicht verstehen. Nachdem MS D die Nachricht entschlüsselt hat, kann sie ebenfalls den erhaltenen MAC wieder mit dem selbst berechneten vergleichen und so sicherstellen, dass die Nachricht korrekt ist und von der Basisstation kommt.

Zur Verschlüsselung von Nachrichten wird neben dem persönlichen Schlüssel auch ein Stromchiffren Generator verwendet. Für die genaue Funktionsweise des Mechanismus wird auf [9] verwiesen.

Für den im nächsten Abschnitt beschriebenen Verrechnungsmechanismus ist es noch notwendig, dass MS D den Empfang der Nachricht bestätigt. Um Ressourcen zu sparen, macht MS D dies nicht nach jedem Paket, sondern am Ende der Session. Die Nachricht enthält folgende zentrale Informationen: Letztes erhaltenes Paket, alle verlorenen Pakete, sowie die ID der Verbindung. Die Nummern aller verlorenen Pakete können angehängt werden, da dies in den meisten Fällen wenige sind, und die Hauptursache für den Verlust ein Verbindungsabbruch ist, in welchem Fall die Information über verlorene Pakete über die Nummer des letzten erhaltenen Paketes transportiert wird.[9]

[2]Bitte beachten: Für die Erläuterung dieses Mechanismus wurden einige für das Verständnis unwichtige Bestandteile der Nachrichten weggelassen. Für genaue Spezifikationen ist [9] empfohlen.

Verrechnung

Der gesamte Verrechnungsmechanismus wird zentral vom Netzbetreiber verwaltet. Die einzelne Mobilstation braucht zu diesem Zweck keine Informationen zu speichern oder zu verarbeiten.

Sobald ein Paket eine Basisstation erreicht und die Entschlüsselung und Überprüfung erfolgreich sind, wird dem Sender der entsprechende Betrag belastet. Dies ist auch der Fall, falls das Paket den Empfänger anschliessend gar nicht erreicht. Gleichzeitig wird auch den Konten der weiterleitenden Stationen ihre Belohnung gutgeschrieben. Die Stationen auf dem Downstream Ast erhalten ihre Betrag erst gutgeschrieben, wenn die Empfängerstation den Erhalt bestätigt hat. Um den Empfänger zu motivieren eine solche Bestätigung zu senden, wird ihm zuerst ein kleiner Betrag belastet, der wieder gelöscht wird, sobald die Bestätigung eingetroffen ist. Falls keine Bestätigung eintrifft, behält der Betreiber diesen Betrag, entschädigt aber auch die weiterleitenden Stationen nicht, da er nicht unterscheiden kann, ob das Paket verloren gegangen ist, oder ob der Empfänger keine Bestätigung senden wollte.

Sicherheitskonzept

Das Unterkapitel der Implementierung abschliessend, soll auch der Bereich Sicherheit untersucht werden. Das ist notwendig um die Robustheit des Systems nachzuweisen, welche unabdingbar für einen praktischen Einsatz ist. Anhand der folgenden Angriffsszenarien soll das Sicherheitskonzept und dessen Implementierung näher betrachtet werden:

- **Refusal to pay**. Jede Station (vor allem der Sender ist in diesem Kontext wichtig) muss der Nachricht einen MAC anhängen, der nur mit einem geheimen Schlüssel berechnet werden kann. Man kann ausserdem davon ausgehen, dass die Basisstation vertrauenswürdig ist und der Schlüssel nicht in falsche Hände gerät. Wenn die Basisstation bei Erhalt einer Nachricht diesen MAC überprüft und zum Schluss kommt, dass dieser tatsächlich mit dem Schlüssel vom Sender berechnet wurde, ist dies ein hinreichender Beweis, dass die Nachricht tatsächlich vom Sender kommt. Dieser kann somit im Nachhinein nicht behaupten, nicht Urheber der Nachricht gewesen zu sein. Weigert sich der Teilnehmer trotzdem die Kosten zu tragen, kann er aus dem Netz ausgeschlossen werden.

- **Incorrect Reward claiming**. Ein Teilnehmer bekommt die Belohnung nur dann gutgeschrieben, falls er beim Setup Teil der Route ist, und während der eigentlichen Übermittlung immer noch daran teilnimmt. Erstere Bedingung wird während des Setup durch die berechnete MAC sichergestellt und letztere Bedingung ergibt sich aus der korrekten Ver- und Entschlüsselung der weitergeleiteten Nachrichten, das entweder von der Basisstation oder vom Empfänger bestätigt wird. Anderweitig kommt ein Teilnehmer zu keiner Gutschrift und kann somit auch keine falschen Ansprüche stellen.

- **Free-riding**. Dies wäre der Fall, wenn zwei Stationen auf einer bestehenden Route miteinander kommunizieren wollen, ohne dabei für die Leistung zu bezahlen. Eine Möglichkeit besteht darin der empfangenen Nachricht zusätzlichen Inhalt aufzuladen, der für den zweiten Teilnehmer gedacht wäre. Da jede Nachricht aber mindestens ein weiteres Mal verschlüsselt wird und der empfangende Teilnehmer den Schlüssel nicht kennt, ist der zusätzliche Inhalt wertlos. Einzig eine Art Morsecode wäre denkbar, bei dem die Nachrichtenlänge manipuliert würde und damit Informationen ausgetauscht werden. Mehr als ein Bit pro Paket ist damit aber auch nicht möglich.

- **Invasive Adversary**. Damit ist gemeint, dass ein Angreifer mehrere Devices beherrscht und nun Pakete mehrmals zwischen diesen Devices weiterleitet, bevor er es wirklich weitergibt. Dies würde ein erhebliches Durcheinander im Belohnungsmechanismus zur Folge haben. Der Angreifer müsste dazu aber auch die Routing-Tabellen von ehrlichen Devices modifizieren können, was in diesem Modell aber nicht angenommen wird.

2.6.4 Beurteilung und Bewertung

Charging and Rewarding bietet einen interessanten Ansatzpunkt. Der Ansatz verzichtet vollständig auf den Zwang zur Mitarbeit. Jedem Teilnehmer ist es somit seiner persönlichen Nutzenrechnung überlassen, ob er für den gebotenen Betrag die Weiterleitungsfunktion wahrnimmt oder nicht. Dieses System ist somit bestimmt fairer, als Zwangssysteme. Jeder Teilnehmer bezahlt nämlich soviel, wie er Datenvolumen verschickt, und jeder weiterleitende Teilnehmer erhält soviel, wie er Aufwand auf sich nehmen musste, um das Netz funktionsfähig zu halten. Bei Zwangsmitarbeit dagegen werden Stationen, welche wenig Verkehr verursachen, übermässig durch Weiterleitung zur Kasse gebeten.

Als Vorteil gilt bestimmt auch die einfach Architektur, die keine speziellen Geräte notwendig macht. Ein Teilnehmer braucht sich nur um den Weg zur nächsten Basisstation zu kümmern, und darum, ob er sich am Datenverkehr beteiligen will oder nicht. Anderweitige Operationen sind nicht nötig.

Dem Ansatz liegt ausserdem ein starkes Sicherheitskonzept zugrunde, was auch eine gewisse Praxistauglichkeit vermuten lässt.

Dadurch, dass jeglicher Verkehr über eine Basisstation läuft, muss der einzelne Teilnehmer sehr wenige Informationen speichern und verarbeiten. Dies wiederum macht das System äusserst skalierbar.

Allerdings bedeutet diese Bindung an Basisstationen auch ein grosser Verlust der Adhoc Eigenschaft. Für jeden Datenverkehr ist ein Provider obligatorisch. Dies macht dieses System auch schwer vergleichbar mit den anderen hier vorgestellten Ansätzen.

2.7 A Robust Reputation System

Dieser Ansatz wurde wie der vorhergehende an der EPFL Lausanne entwickelt [10]. Dabei handelt es sich jedoch nicht um einen Vorschlag, der Anreize schaffen möchte, sondern um einen, welcher Falschspieler entdeckt und gegebenenfalls vom Netz ausschliesst. Es gehört somit zu den Zwang- oder Ausschlusssystemen.

Informationen über das Verhalten anderer Teilnehmer können entweder durch eigene Beobachtung, durch Beobachtung von anderen Mitgliedern des Netzes, oder durch eine zentrale Instanz gewonnen werden. Eine zentrale Instanz gibt es in diesem Netztyp nicht, die eigene Beobachtung reicht oft nicht weit und so liegt der Kernpunkt dieses Reputationssystems in der Verarbeitung von mitgeteilten Beobachtungen der anderen Teilnehmer mittels eines statistischen Ansatzes.

2.7.1 Voraussetzungen

Für dieses Reputationssystem kann ein beliebiges Ad-hoc Netzwerk benutzt werden. Insbesondere ist keine Basisstation und kein zentraler Provider notwendig. Dies erschwert zwar das Routing, erhöht aber in diesem Masse auch die Flexibilität des Netzes. Der zugrunde liegende Kommunikationsmechanismus ist paketbasiert. Die Ausarbeitung dieses Konzeptes konzentriert sich auf den Mechanismus, der, sich falsch verhaltende, Mitglieder entdeckt und sanktioniert. Weitere technische Fragen, wie das Routing und Verschlüsselung werden ausgeklammert und können von anderen Lösungen übernommen werden.

2.7.2 Architektur

Dem Ansatz liegen zwei Werte zugrunde, die jeder Teilnehmer über alle anderen sammelt. Es sind dies Vertrauen und Reputation. Diese Unterscheidung ist für das Verständnis der Implementierung wichtig und deshalb sollen die Begriffe an dieser Stelle bereits eindeutig auseinander gehalten werden.

Reputation beschreibt das Mass, in dem der eine Teilnehmer den anderen als korrekt mitarbeitende Station wahrnimmt und einschätzt. Diese Reputation erhält er aus seinen Beobachtungen und aus den Beobachtungen einiger anderen Teilnehmer. Wie stark die Meinung einer solchen mitteilenden Station in die Bewertung des ersten Teilnehmers einfliesst, entscheidet sich durch das Vertrauen, das dieser Teilnehmer in den anderen hat und wie ähnlich die mitgeteilte Information der eigenen Wahrnehmung kommt. Wie Reputation und Vertrauen aufgebaut werden können, wird im Abschnitt über die Protokolle und die Implementierung erläutert. Konzeptuell funktioniert der Ansatz folgendermassen: Beliebige Teilnehmer sammeln während einer gewissen Zeit Informationen über andere Teilnehmer, welche in ihrem Bereich liegen. Diese Informationen werden mittels eines statistischen Ansatzes zu einem Reputationswert verarbeitet. Diese Werte werden periodisch mit einigen anderen Teilnehmern ausgetauscht. Anschliessend werden diese Informationen bei einem Abweichungstest mit den eigenen Beobachtungen verglichen und je nach

Ergebnis und vorhandenem Vertrauen in den Informationslieferanten in die Berechnung eines neuen Reputationswertes einbezogen oder fallen gelassen. Dies hat aber wiederum Einfluss auf das Vertrauen, das der verarbeitende Knoten in den Informationslieferanten hat. Hat er zum Beispiel selbst sehr viele Beobachtungen gemacht, welche aussagen, dass Teilnehmer C sich immer korrekt verhalten hat, Teilnehmer B ihm aber fälschlicherweise mitteilen möchte, dass sich C falsch verhalten hat, so wird A diese Information nicht verarbeiten, dafür wird aber A's Vertrauen in die Informationen von B sinken.

Sowohl der Reputationswert, wie auch jener für das Vertrauen, werden mit der Zeit aber auslaufen, so dass Vertrauen und Reputation ständig neu erarbeitet werden muss. Dies ermöglicht auch einem Teilnehmer, der sich falsch verhalten hat, später wieder am Datenverkehr teilzunehmen, oder einem Lügner, dass seine Informationen später nicht mehr ignoriert werden. Dieses Auslaufen der Werte hat auch ganz entscheidenden Einfluss auf die Robustheit des Systems, also darauf, wie das System mit Falschspielern umgehen kann (im Sicherheitskonzept beschrieben).

Der Vertrauensindex hat keinen Einfluss darauf, wie der Teilnehmer im Netz von den anderen behandelt wird. Er kann, sofern nicht auch seine Reputation schlecht ist, ohne Einschränkungen am Datenverkehr teilnehmen. Lediglich seine Beobachtungen haben bei den anderen Mitgliedern keinen hohen Stellenwert mehr. Dies ist wichtig, da sich sonst kaum ein Teilnehmer getrauen würde, schlechtes Verhalten eines anderen Teilnehmers zu melden, da die Gefahr besteht, dass, falls dies niemand ausser ihm bemerkt hat, sein Vertrauensindex bei den anderen Teilnehmern sinkt und er so aus dem Verkehr ausgeschlossen würde. Bei der aktuellen Implementierung sinkt zwar sein Vertrauen ebenfalls, die hat jedoch keinen Einfluss auf seine Rolle im Netz.

2.7.3 Protokolle und Implementierung

Um die Funktionsweise und Implementierung dieses Ansatzes darzustellen, geht man am besten wieder von einem Beispielnetz und einem durchzuspielenden Szenario aus.

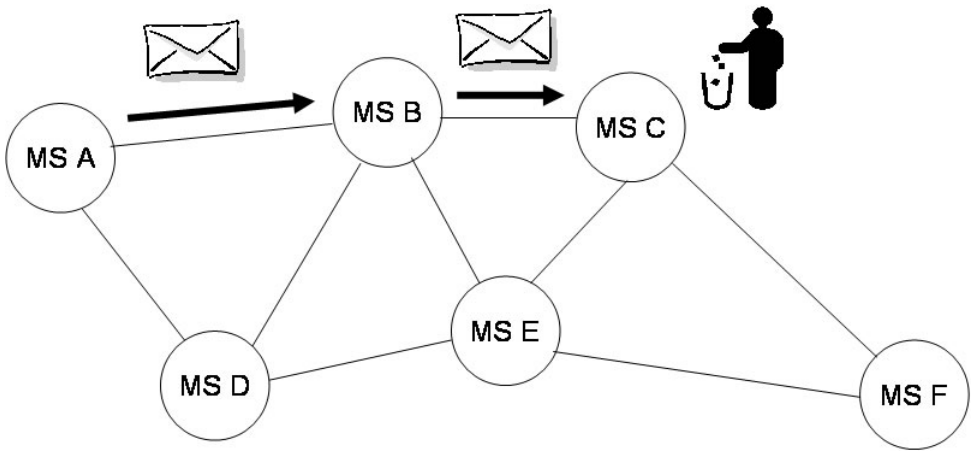


Abbildung 2.6: Das Netz für das Beispielszenario

Die Mobilstation A (MS A) möchte der Mobilstation F (MS F) eine Nachricht senden. MS B leitet die Nachricht korrekt weiter, während MS C die Nachricht einfach verwirft, anstatt

sie dem Empfänger zuzustellen. Die Frage ist nun, wie erhalten die anderen Teilnehmer die Information, dass MS C eigennützig gehandelt hat?

In dieser Ausarbeitung wird bewusst nicht vertieft auf die statistischen und mathematischen Berechnungen eingegangen, geht es doch in erster Linie darum, Funktionsprinzipien aufzuzeigen und diese auf konzeptueller Ebene miteinander zu vergleichen. An den genauen Berechnungen und Formeln Interessierte seien an dieser Stelle an das Paper der Urheber dieses Ansatzes verwiesen [10].

Reputation und Vertrauen

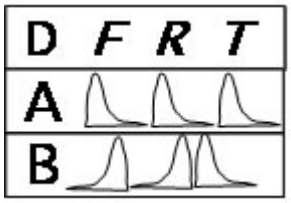


Abbildung 2.7: Verteilung Wahrscheinlichkeit

Um die Berechnung und das Ergebnis zu verstehen, ist es unerlässlich zumindest das Prinzip der Bayesschen Wahrscheinlichkeitsrechnung zu verstehen. Eine Wahrscheinlichkeit p wird dabei nicht als relative Häufigkeit verstanden, sondern als Grad persönlicher Überzeugung.[11]

Der Teilnehmer A denkt also mit einer gewissen Wahrscheinlichkeit p, dass der Teilnehmer B sich unkorrekt verhält. Jeder Teilnehmer hat eine solche Meinung über jeden anderen Teilnehmer im Netz. Für diesen Ansatz wird eine Beta-Verteilung mit den Parametern (a,ß) verwendet (Details zu diesen statistischen Grundlagen entnehmen Sie bitte der [12]). Zu Beginn haben wir eine Verteilung mit den Werten Beta(0,1), welche auf die Absenz jeglicher Information hindeutet. Informationen die in das Modell einfliessen, beeinflussen die Kurve, die das Modell produziert. Eine starke Häufung auf der rechten Seite (Siehe Abbildung 2.7, untere Zeile) bedeutet einen schlechten Wert (hier die Meinung von D über B). In das Modell eingebaut wurde ebenfalls ein Parameter, der den neueren Informationen mehr Gewicht einräumt als älteren, was zum Auslaufen der Reputation und des Vertrauens führt.

Ein neuer Wert für die Reputation wird berechnet, wenn eine neue eigene Beobachtung stattfindet, oder wenn periodisch die Beobachtungen untereinander ausgetauscht werden.

Das Vertrauen nutzt einen ähnlichen Ansatz, wie er für die Reputation verwendet wird. Wenn eine Information sehr nahe an den eigenen Beobachtungen liegt, hat dies eher positive Auswirkungen auf das Vertrauensrating. Eine grössere Abweichung hat entsprechend eine Verschlechterung desselben zur Folge.

Sicherheitskonzept

Damit das System in der Praxis auch wirklich eingesetzt werden könnte ist es unerlässlich es wiederum gegen die wichtigsten Formen der Falschspielerei und die wichtigsten Angriffe zu testen. Erst dann kann ein Urteil darüber abgegeben werden, ob dieses System robust ist, oder nicht.

Die folgenden Angriffszenarien sind die wichtigsten und sollen deshalb einzeln betrachtet werden:

- **Stealthy Lying**. Stealthy Lying bedeutet, dass ein Angreifer jedes Mal ein klein wenig über einen anderen Teilnehmer lügt und so versucht, seine Reputation zu schädigen. Da die Reputation aber ständig wieder ausläuft ist es nicht möglich, die Rufschädigung so zu akkumulieren, dass sie einen Einfluss auf das Verhalten des Systems gegenüber dem Angegriffenen haben kann. Um dies zu erreichen, müsste der Angreifer schon kräftiger lügen. Da dies den Abweichungstest aber nicht überstehen wird, wird das Vertrauen in den Angreifer sinken und seine Information somit nicht mehr verarbeitet.
- **Gain Trust and then Lie**. Eine weitere Mögliche Angriffsstrategie wäre, dass der Angreifer zuerst versucht Vertrauen zu gewinnen, indem er die korrekten Informationen weitergibt. Hat er dann genügend Vertrauen erarbeitet, versucht er wieder mit Falschinformationen einen Teilnehmer zu schädigen. In einem ersten Schritt würde dieser Angriff Erfolg haben, da der falsch informierte Teilnehmer ein leicht schlechteres Bild des angegriffenen Teilnehmers bekommen würde. Aber gleichzeitig nimmt mit dem Abweichungstest auch das Vertrauen in den Angreifer und weitere Falschinformationen werden nicht mehr so stark berücksichtigt und würden das Vertrauen in ihn weiter schmälern.
- **Brainwashing**. Brainwashing ist ein Angriffsszenario bei welchem ein zu täuschender Teilnehmer von lügenden Angreifern umzingelt ist. Sie versorgen ihn gezielt mit Falschinformationen über einen entfernten Teilnehmer. Als Folge davon sinkt die Reputation und der Getäuschte denkt, dass es sich dabei tatsächlich um einen falsch spielenden Teilnehmer handelt. Sobald sich der Getäuschte bewegt, wird er vorerst zwar immer noch ein falsches Bild des anderen Teilnehmers haben und den richtigen Informationen gar nicht glauben. Da die Reputation und das Vertrauen mit der Zeit auslaufen, wird der Getäuschte irgendwann keine Meinung mehr über den anderen Teilnehmer haben und kann an die Wahrheit zu glauben beginnen.

2.7.4 Beurteilung und Bewertung

Im Gegensatz zu einigen anderen Ansätzen kann das Robust Reputation System auf einen zentralen Provider verzichten. Dies macht das Netzwerk und die Zusammenarbeit um einiges flexibler, senkt aber auch die Mächtigkeit des Anwendungsgebietes. Als grosser Negativpunkt stellt sich die Architektur auf dem einzelnen Teilnehmergerät heraus. Jeder Teilnehmer muss über alle anderen Teilnehmer Informationen zur Verfügung haben und

diese mit rechenaufwändigen statistischen Methoden verarbeiten. Es handelt sich dabei immer um subjektive Meinungen der Teilnehmer. Diese individuelle und dadurch eventuell auch redundante Datenhaltung steht im Kontrast zu anderen System, bei welchen eine zentrale Stelle für die Verarbeitung der Informationen zuständig sind.

Da das System auf Zwang basiert und es keine zentrale Koordination gibt, ist ein temporärer Missbrauch auch möglich. Dies ergibt ein etwas getrübtes Bild bei der Sicherheitsbewertung.

Abschliessend kann deshalb festgehalten werden, dass dieser Ansatz für kleinere, überschaubare Ad-hoc Netzwerke durchaus Sinn machen kann. Für grössere, skalierte Netze müsste aber eine andere Lösung vorgezogen werden.

2.8 Fazit und abschliessende Bemerkungen

So unterschiedliche Lösungsansätze vorhanden sind, von so unterschiedlichen Vorbedingungen wird bei den vorgestellten Verfahren ausgegangen. Dies erschwert zwar den allgemeinen Vergleich zwischen diesen Ideen, macht es dafür aber einfacher für eine bestimmte Situation die geeignete Lösung zu finden. So wurden in dieser Ausarbeitung Ansätze präsentiert, welche auf absolute Ad-hoc Fähigkeit ausgelegt sind, aber auch solche, die eine teilweise zentralisierte Struktur oder sogar einen zentralen Provider benötigen. Die Strategien reichen vom kontrollierten Zwang zur Mitarbeit bis zum anreizbasierten Freiwilligkeitsprinzip. Entsprechend den Ansätzen unterscheiden sich auch die Sicherheitsanforderungen und -Lösungen stark. Ein Ziel bleibt aber allen Vorschlägen gleich: Trittbrettfahrer und Falschspieler sollen keine Chance haben.

Wie kann aus dieser Matrix von verschiedenen Ausgangslagen, Architekturen und Sicherheitsprinzipien nun ein Sieger erkoren werden? Die Antwort ist kurz und einfach: gar nicht. Lösungen müssen frei nach dem Motto: *spezielle Situationen erfordern spezielle Lösungen ausgewählt werden.*

Interessant ist die Tatsache, dass Probleme bei der Zusammenarbeit vermehrt mit monetären Anreizsystemen zu lösen versucht werden. Dies können virtuelle oder reale Währungen sein. Dabei wird versucht eine Brücke zwischen Wirtschaftstheorie und Technologie zu schlagen. Nicht die perfekte Lösung (Ziel von vielen Informatikern), sondern die optimalste Lösung soll gesucht werden.

Soweit es den Autoren bekannt ist, gibt es im Moment keine praktischen Anwendungen dieser Ansätze. Die Frage steht im Raum, wie stark sich kabellose Netzwerke weiterentwickeln - sei es in Bandbreite oder Reichweite - und in welche Richtung sich die Anwendung dieser Technologien bewegt. Werden mobile Ad-hoc Netzwerke in Zukunft eine wichtige Rolle einnehmen, oder werden sie im Laufe der Evolution einfach übersprungen? Wie dem auch sei, eines kann den Arbeiten in diesem Gebiet bestimmt entnommen werden. Die Idee dahinter.

Denn Mechanismen zur Förderung der Zusammenarbeit werden mit Bestimmtheit auch in anderen Gebieten zum Einsatz kommen. Eine zunehmend vernetzte Welt bringt dies

praktisch als Anforderung mit sich. Und so werden diese theoretischen Überlegungen aus unserer Sicht auf jeden Fall ihren Platz in einer Anwendung finden.

Literaturverzeichnis

[1] A. S. Tanenbaum: Computer Networks, Prentice Hall, 1985.

[2] M. Frodigh, P. Johansson, P. Larsson: Wireless ad hoc networking - The art of networking without a network, Ericsson Review, 2000.

[3] `http://laptop.media.mit.edu`, 11.12.2005.

[4] L. Buttyán, J.-P. Hubaux: Stimulating Cooperation in Self-Organizing Mobile Ad hoc Networks, Kluwer Academic Publishers, 2003.

[5] `http://searchsecurity.techtarget.com/sDefinition/0,,sid14_gci518283,00.html`, 10.01.2006.

[6] M. Frank, P. Martini, M. Plaggemeier, CineMA: Cooperation Enhancement in Manets, IEEE Computer Society, 2004.

[7] A. Weyland, T. Braun: Cooperation and Accounting Strategy for Multi-hop Cellular Networks, IEEE Workshop on Local and Metorpolitan Area Networks, 2004.

[8] `http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-10.txt`, 10.01.2006.

[9] N. B. Salem, L. Buttyán, J.P. Hubaux, M. Jakobsson. A Charging and Rewarding Scheme for Packet Forwarding in Multi-hop Cellular Networks. International Symposium on Mobile Ad Hoc Networking & Computing (MobiHoc), 2003.

[10] 4 S. Buchegger, J.Y. Le Boudec. A Robust Reputation System for P2P and Mobile Ad-hoc Networks. Second Workshop on the Economics of Peer-to-Peer Systems, Juni 2004.

[11] `http://de.wikipedia.org/wiki/Bayesscher_Wahrscheinlichkeitsbegriff`, 10.01.2006.

[12] James O. Berger. Statistical Decision Theory and Bayesian Analysis. Springer, second edition edition, 1985.

Kapitel 3

Losses Resulting from Fraud, Espionage and Malware

Petra Irène Lustenberger, Daniel Eisenring, Marcel Lanz

Spionage ist vor allem in der Industrie ein wirkungsvolles Mittel dem eigenen Konzern einen Vorteil zu verschaffen oder die Konkurrenz schlechter zu stellen. Die wirtschaftlichen Verluste lassen sich dabei nur schwer beziffern, sind aber immens. Spionage erfolgt durch elektronische wie durch low-tech Aktivitäten. Fraud (Betrug) ist weit verbreitet und hat viele Ausprägungen, seien es Täuschungen an Online-Auktionen oder Kreditkartenbetrüge. Es können dabei Antrags-Fraud, technischer Fraud oder interner Fraud im Sinne von Kategorien unterschieden werden. Malware umfasst als ein abstrakter Oberbegriff jegliche Software die in „böswilliger" Absicht programmiert wurde und besitzt vielfältige Konkretisierungen. Computerviren, seien es Link- oder Bootviren, Computerwürmer, oder Trojaner stellen zusammen mit anderen Erscheinungsformen von Malware für die betreffenden Systeme eine konkrete Bedrohung dar. Neben Malware ist Social Engineering, die elegante Umgehung des aufwändigen „Knackens" bzw. Hackens technisch gut geschützter Ressourcen, bei der auf der Ebene Mensch angesetzt wird, nicht mehr wegzudenken. Die zunehmende Komplexität der Angriffe verlangt eine Anpassung der Unternehmung an die neuen Gefahren aus dem Internet und somit nach einem bewussten Risikomanagement. Die aufgrund der Gefahrenanalyse umgesetzten Massnahmen müssen kontinuierlich auf Wirkung und Schutz geprüft und verbessert werden. Bei der Ermittlung der individuellen Gefährdungslage gelten Technik, höhere Gewalt, Organisation und der Mensch als die vier grundsätzlichen Bedrohungsklassen. Technische Schutzkonzepte sind unter anderem Firewalls, Intrusion Detection Systems und Honeypots. Ausbildung, Training und das Bewusstsein bezüglich Gefahren gelten als die drei essentiellen Ebenen eines Mitarbeitertrainings.

Inhaltsverzeichnis

3.1 Einleitung

Unternehmen sind heute einer Vielzahl von Sicherheitsrisiken ausgesetzt. Die durch Schäden entstandenen Kosten haben sich in den letzten Jahren potenziert. Neben nicht autorisierten Zugriffen auf vertrauliche Daten von aussen durch Hacker oder von unbefugten Mitarbeitern innerhalb des Unternehmens sind die Systeme auch physischen Gefahren wie Feuer, Naturkatastrophen etc. ausgesetzt. Es sind immer weniger technische Fertigkeiten erforderlich, um einen Angriff auf fremde Netze zu starten, wie dies beispielsweise script kiddies[1] tun. Somit kann jeder Ziel eines Angriffs werden. Selbst Unternehmen, die bisher beispielsweise wegen eines gegenüber Grossunternehmen geringeren Bekanntheitsgrades von Angriffen verschont geblieben sind, müssen verstärkt damit rechnen, rein „zufällig" Opfer einer Attacke zu werden. Strategische Vorsorge wird daher immer wichtiger.

Die neuen rechtlichen Entwicklungen haben zur Folge, dass sich mangelnde IT-Sicherheit in den Unternehmen direkt monetär auswirkt. Und zwar sowohl auf die persönliche Haftung der Manager bei Sicherheitsvorfällen als auch auf die Kreditvergabe und die Höhe der Kreditzinsen.

Durch die zunehmende Vernetzung einerseits und das abnehmende Sicherheitsbewusstsein andererseits werden die Bedrohungsformen Spionage (engl. Espionage), Software mit betrügerischer Absicht (engl. Malware) und Betrug (engl. Fraud) begünstigt. Im folgenden Kapitel wird auf besagte Bedrohungsformen eingegangen. Ein weiteres Kapitel widmet sich den konkreten Techniken der Bedrohungsformen. Ein weiteres Kapitel zeigt, wie mit konkreten Schutzkonzepten auf die Bedrohungen reagiert werden kann. Das Ganze wird mit einem Diskurs und einer Schlussfolgerung zum Thema abgeschlossen.

3.2 Wirtschaftliche und statistische Aspekte

Dieses Kapitel soll anhand von Beispielszenarien und Definitionen die Begriffe Espionage, Malware und Fraud erklären, mittels Statistiken die wirtschaftlichen Auswirkungen beleuchten und die Notwendigkeit eines guten Risikomanagements aufzeigen.

3.2.1 Espionage

In folgendem Unterkapitel wird auf die Bedrohungsform Industriespionage eingegangen.

Beispielszenario Espionage [2]

Ein Ehepaar aus Israel fand Auszüge ihres unveröffentlichten Buches auf dem Internet. Die verständigte Polizei stellte darauf einen Trojaner auf dem Computer des Ehepaars

[1] Script kiddies sind Jugendliche, die über wenig technisches Wissen verfügen, sich mit im Internet verfügbaren Programmen und Skripten Viren basteln und damit in fremde Computer eindringen [1]

sicher. Der Trojaner führte zum Exmann der Tochter des Ehepaares, welcher auch der Autor des Trojaners war und die Seiten des Buches im Internet veröffentlichte. Es stellte sich heraus, dass der Exmann ein professioneller Autor von Trojanern war und diese verkaufte. Dessen Kunden waren drei Untersuchungsfirmen, die wiederum im Auftrag von ihren Kunden handelten. Die Trojaner dienten dazu, das Unternehmen gegen das sie eingesetzt wurden, auszuspionieren. Opfer der Angriffe waren unter anderen eine Firma, die Transmitter für unbemannte Flugzeuge herstellt und ein Unternehmen, das Fahrzeuge der Marke Honda importiert. Die Kunden der Untersuchungsfirmen, also die Auftraggeber für die Spionage waren direkte Konkurrenten der ausspionierten Unternehmen. Ein Trojaner kostete 16000 Schekalim, was ungefähr 4100 CHF entspricht. Nach der Bezahlung wurden nur Passwort und Adresse des Trojaners übermittelt. Die Installation auf dem auszuspionierenden Rechner übernahm der Autor. Er benutzte dafür hauptsächlich zwei Methoden. Die Erste war eine mit dem Trojaner infizierte E-Mail. Die Zweite war eine infizierte CD-Rom, getarnt als Geschäftsvorschlag eines angesehenen Unternehmens. Im Zuge der Ermittlungen wurden tausende als geheim deklarierte Dokumente auf FTP-Servern gefunden. Weil der Trojaner spezifisch jeweils für ein Unternehmen programmiert wurde, schlug auch kein Antivirusprogramm Alarm. Laut Medien sei der Verlust noch nicht ermittelt, aber sicher beträchtlich.

Definition Espionage

Man spricht von Spionage, im Beispielszenario von Industriespionage, wenn sich jemand Zugang zu Daten verschafft oder sammelt, die das Unternehmen zu schützen versucht. Die gestohlenen Daten sind für jemanden von wirtschaftlichem Interesse [3]. Es gibt 2 Kategorien von Spionageaktivitäten:

Elektronische Aktivitäten Dazu gehören unter anderen das Abhören von Telefonen, das Ausspionieren von Daten mittels Trojanern, Keyloggern[2], E-Mails abfangen und ähnlichem. Der Schlüsselpunkt ist, dass beim Spionieren elektronische Technologie zum Einsatz kommt.

Low-tech Aktivitäten Im Gegensatz zu den elektronischen Aktivitäten kommen bei den low-tech Aktivitäten wenige bis keine technischen Hilfsmittel zum Einsatz. Beispiele sind: Briefpost lesen, Shoulder-surfing (jemandem bei der Passworteingabe über die Schulter schauen)oder Einbruch.

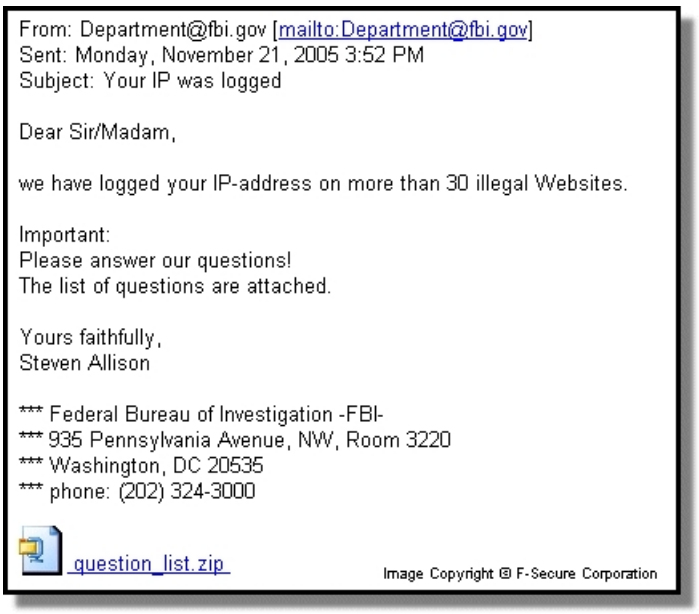
3.2.2 Malware

Im folgenden Unterkapitel wird auf die Bedrohungsform Malware eingegangen.

[2]Ein Keylogger speichert heimlich alle Tastaturanschläge in einer Datei.

Beispielszenario Malware [4]

Im folgenden wird Malware anhand des Virus Sober.Y vorgestellt. Sober.Y wurde am 16 November 2005 entdeckt. Hat der Virus einen Computer erfolgreich infiziert, durchsucht er ihn auf E-Mailadressen, an die er sich weiter versenden kann. Je nachdem wie die Endung der E-Mailadresse ist, generiert er einen Text auf Englisch oder auf Deutsch. Für die Domain @gmx oder Endungen auf .ch und .de wählt er zum Beispiel Deutsch. Der Virus versendet sich dann in einem ZIP Archiv als Anhang in einer E-Mail. Der Inhalt der E-Mail dient dazu, den Empfänger dazu zu bringen, den Anhang zu öffnen und so den Virus zu installieren. In diesem Beispiel gaukelt der Virus vor vom FBI oder CIA zu kommen. Der Empfänger soll sich ertappt fühlen und sich zur Schadensbegrenzung kooperativ verhalten, also das ZIP-Archiv öffnen. Gepackt ist der Virus etwa 55 Kilobyte, entpackt 196 Kilobyte gross. Die Abbildung 3.1 zeigt ein solches E-Mail.


From: Department@fbi.gov [mailto:Department@fbi.gov]
Sent: Monday, November 21, 2005 3:52 PM
Subject: Your IP was logged

Dear Sir/Madam,

we have logged your IP-address on more than 30 illegal Websites.

Important:
Please answer our questions!
The list of questions are attached.

Yours faithfully,
Steven Allison

*** Federal Bureau of Investigation -FBI-
*** 935 Pennsylvania Avenue, NW, Room 3220
*** Washington, DC 20535
*** phone: (202) 324-3000

question_list.zip

Image Copyright © F-Secure Corporation

Abbildung 3.1: E-Mail vesandt vom Virus Sober.Y [4]

Um den Empfänger in Sicherheit zu wiegen, öffnet der Virus ein Dialogfenster, das vom Antivirusprogramm zu kommen scheint mit der Meldung, dass kein Virus gefunden wurde. Beim Öffnen des Anhangs, also der ZIP-Datei mit dem Virus, wird eine Fehlermeldung erzeugt: “Die Datei kann nicht geöffnet werden“.

Definition Malware

Auf die technischen Details wird im zweiten Teil dieser Arbeit eingegangen. Hier soll vollständigkeitshalber nur kurz eine Einleitung und Übersicht gegeben werden.

Malware wird hauptsächlich unterteilt in Viren, Würmer und Trojaner. Für die Unterscheidung stellt sich vor allem die Frage: Ist der ausführbare Code schädlich oder nicht. Für die Entscheidung kann man das Flussdiagramm in Abbildung 3.2 zu Hilfe nehmen.

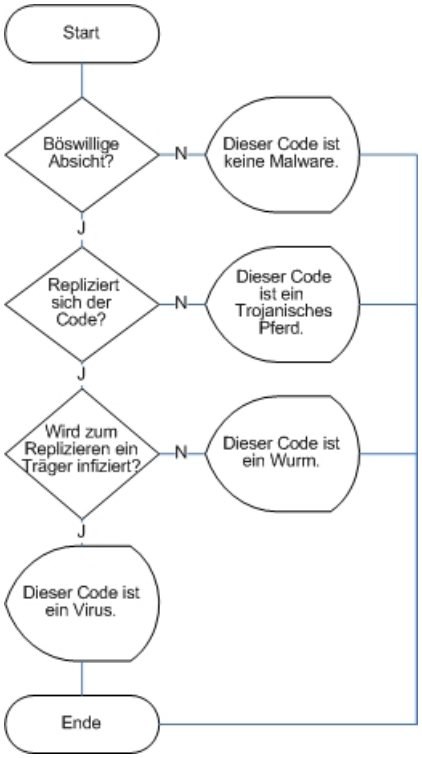


Abbildung 3.2: Flussdiagramm zum Bestimmen des Malwaretyps [5]

Ein Trojanisches Pferd (Trojaner) ist ein nützlich oder harmlos erscheinendes Programm, das zur Ausnutzung oder Beschädigung des jeweiligen Systems konzipierten, verborgenen Code enthält. Im Gegensatz zu Viren und Würmern können sich Trojaner nicht selber verbreiten. Aber es gibt auch Kombinationen von Viren und Trojanern, bei denen der Virus zum Infizieren des Systems dient und dann den Wurm nach lädt. Der Zweck von Trojanern liegt eindeutig im Ausspionieren der Dateien und Passwörter eines fremden Systems. Viren können unterschiedliche Zwecke verfolgen. Es gibt Viren, die einen Keylogger installieren und dann periodisch einen Bericht an den Autor schicken. Eine andere Sorte von Viren und Würmern infiziert Systeme und wartet auf Befehle vom Autor. Solche Malware wird benutzt, um DDoS- Attacken (mehr dazu später) zu starten oder um Spam zu versenden.

3.2.3 Fraud

Im folgenden Unterkapitel wird auf die Bedrohungsform Betrug eingegangen.

Beispielszenario Fraud

Ein klassisches Beispiel für Fraud übers Internet ist das Anbieten von Waren, die nach Bezahlung nicht geliefert werden. Vor allem bei Online-Auktionsplattformen wie eBay[3]

[3] `http://www.ebay.com`

sind solche Betrügereien gängig. Weitere Beispiele für Fraud sind der Handel mit Raubkopien oder der in den USA bekannte Identitätsschwindel. Auch der Kreditkartenbetrug verursacht grossen wirtschaftlichen Schaden.

An dieser Stelle wird weiter auf den Identitätsschwindel und den Kreditkartenbetrug eingegangen.

ID Fraud, also Identiätsschwindel, ist vor allem in den USA geläufig. Um dort als Doppelgänger aufzutreten, ist nur der Name und die dazugehörige 9-stellige Sozialversicherungsnummer einer Person nötig. Mit diesen Angaben kann man zum Beispiel Kreditschecks bis zu $ 44000 ausstellen lassen und damit einkaufen gehen. Es gibt professionelle Unternehmen wie docusearch[4], die zahlenden Kunden für durchschnittlich $ 50 persönliche Daten beliebiger Personen in Erfahrung bringen. Das Angebot

> Search For Social Security Number (Suche nach der Sozialversicherungsnummer einer Person)

für $ 49 ist unterdessen nur noch unter gewissen Bedingungen nutzbar. Um dieses Angebot nutzen zu können, muss man bescheinigen, dass die Anfrage im Zusammenhang mit Betrug steht oder man für den Staat arbeitet und darum die Sozialversicherungsnummer einer bestimmten Person braucht.

Der Kreditkartenbetrug ist bereits so weit fortgeschritten, dass Kreditkartennummern über Botnets getauscht und validiert werden. Im IRC (Internet Relay Chat) existieren Kanäle speziell für Kreditkartenbetrug und bilden das Netz der Botnets. Die Bots sind Programme, die den Chat auf bestimmte Befehle absuchen und automatisierte Antworten geben. So ist es für den Kreditkartenbetrüger möglich schnell die Kreditkartennummern, die er kaufen möchte, mittels spezieller Anfrage an den Bot auf ihre Richtigkeit zu überprüfen. Die Herkunft der Nummern ist sehr unterschiedlich. Sie stammen zum Beispiel aus Einbrüchen auf Webservern oder vom Servierpersonal, das Kreditkartennummern beim Bezahlvorgang abgeschrieben hat [6].

Definition Fraud

Betrug zeichnet sich meistens durch eine Bereicherung desjenigen, der den Betrug begeht, aus. Es handelt sich um arglistige Täuschung und Schwindel. Eine gute Definition ist folgende:

> „[Eine] nicht legitimierte, wirtschaftsschädigende Nutzung der unternehmenseigenen Leistungspotentiale.“ [7].

Katrin Schmitt [7] teilt Fraud in die Kategorien Antrags-Fraud (Missbrauch von Identitäten und Bezahlvorgängen), technischen Fraud (Manipulation von technischer Infrastruktur und Kontrolleinrichtungen im Dienstleistungsbereich) und internen Fraud ein.

[4] `http://www.docusearch.com`

Anrags-Fraud sind zum Beispiel mittels Datendiebstahl ergaunerte Schlüssel und Passwörter, um damit wie im Beispielszenario Waren mit fremden Kreditkartennummern zu bestellen. Unter Fraud fallen auch alle Versuche, per E-Mail ungesetzlich an Geld zu kommen. Verbreitet sind Bettelbriefe oder betrügerische Angebote der so genannten Nigeria-Connection[5]. In den Bereich des technischen Frauds fallen die Beispielszenarien Malware und Espionage (siehe oben).

Laut einer Quelle [7] soll der interne Fraud der am weitesten verbreitete Fall von Betrug sein. Es fällt nicht schwer, das zu glauben, ist es doch für viele Angestellte eines Unternehmens ein Leichtes, an vertrauliche und wichtige Daten zu kommen. Interner Fraud äussert sich dadurch, dass Mitarbeiter durch Datendiebstahl, Datenmanipulation und Datenmissbrauch Dienstleistungen oder Informationen des Unternehmens illegal zur Nutzung anbieten oder diese selbst nutzen.

3.2.4 Risikomanagement

Durch die zunehmende Komplexität der Informationstechnologie werden die Unternehmen mit einer Flut von Daten und Informationen überschwemmt. Das bietet den Angreifern mehr Möglichkeiten und Spielraum für die eigene Bereicherung, und oft bleiben sie dabei auch unentdeckt. Diese zunehmende Unübersichtlichkeit macht eine ständige Überwachung der Prozesse auf Schwachstellen nötig. Das Sicherheitskonzept muss laufend den neuen Gegebenheiten angepasst und wiederum auf Schwachstellen überprüft werden. Es wird empfohlen [7] für dieses sogenannte Fraud Management einen Stab direkt unterhalb der Geschäftsführung einzusetzen. So geniesst er volle Aktionsfreiheit. Seine Aufgabe ist es, den Ist-Zustand zu analysieren und weiter zu verbessern, sowie die Entwicklungen der Technik, Gesetze und Hackerszene zu überwachen und das Sicherheitskonzept daran anzupassen.

3.2.5 Statistiken

In diesem Abschnitt soll das Ausmass des Schadens, das durch Fraud und Malware angerichtet wird anhand von Statistiken und Zahlen aufgezeigt werden.

Die Internet Fraud Watch[6] ist eine Meldestelle für erfolgte Betrügereien im Internet. Aus den eingegangenen Meldungen entsteht halbjährlich eine Statistik. Im Jahre 2003 wird der Link vom Internetauktionhaus eBay auf die Webseite der Fraud Watch entfernt. Seither sind die Meldungen markant eingebrochen und werden für die Statistik hochgerechnet. Dafür sind die Informationen aktuell aus dem ersten Halbjahr 2005. In der Abbildung 3.3, einen Auszug aus der Statistik aller Meldungen über Fraud vom Januar bis Juni 2005, ist die Hauptbetrugsart mit 44% aller Meldungen der Verkauf von Ware, die nicht geliefert wurde. Der durchschnittliche Verlust pro Meldung über diese Art von Fraud, Auctions

[5]Eine Geldsumme von ein paar tausend Dollar soll es einem afrikanischen Geschäftsmann ermöglichen, einen Geldbetrag in Millionenhöhe ausser Landes zu bringen. Versprochen wird ein grosser Anteil am besagten Geldbetrag. Man spricht auch von einem Vorschussbetrug [9]

[6]`http://www.fraud.org`

genannt, beläuft sich auf $ 999. Ein weiterer grosser Posten mit 33% aller Meldungen und einem hohen durchschnittlichen Verlust von $ 4389 pro Meldung sind ebenfalls bezahlte, aber nicht ausgelieferte Güter. Allerdings wurden diese Waren nicht über Auktionen, wie bei eBay möglich, sondern über andere Kanäle wie Webshops bezahlt. Die Nigeria-Connection verursacht auch im Jahr 2005 noch viele Verluste mit ihren Geldversprechen. Mit 7% aller Meldungen und einem durchschnittlichen Verlust von $ 11370 bilden die Nigrian Money Offers die drittgrösste Sparte.

Top Ten Scams

Category	% of All Complaints	Average Loss
Auctions	**44%***	**$999**
Goods never delivered or misrepresented		
General Merchandise	**30%**	**$4,389**
Sales not through auctions, goods never delivered or misrepresented		
Nigerian Money Offers	**7%**	**$11,370**
False promises of riches if consumers pay to transfer money to their bank accounts		
Fake Checks	**5%**	**$4,733**
Consumers paid with phony checks for work or items sold, instructed to wire money back		
Phishing	**4%**	**$298**
Emails pretending to be from well-known source asking to confirm personal information		
Lotteries/Lottery Clubs	**3%**	**$3,953**
Requests for payment to claim lottery winnings or get help to win, often foreign lotteries		

Abbildung 3.3: Statistik Internet Fraud Watch [8]

In 76% aller Fälle wurde der Kontakt zu den Opfern über eine Website, in den restlichen 24% der Fälle per E-Mail hergestellt. Im Jahre 2003 war der Anteil der E-Mails mit 6% gering [8]. Laut einer Statistik von Larry Bridwell und Lawrence M. Walsh [10] wurden im Jahr 2002 pro Monat auf 1000 PC's 105 Virenbefälle registriert.

Die Infektion eines Computers oder Servers kann vielfältige Arten von Kosten verursachen. Dazu gehören vor allem der Produktivitätsverlust, weil der Computer gesäubert werden muss. Das macht meistens eine Arbeitskraft aus der IT-Abteilung. Währenddessen kann der Nutzer seinen Computer aber nicht benutzen, was eine weitere Art von Verlust ist. Unlesbare Daten (corrupted files), Datenverlust und Systemabstürze sind weitere Folgen von Malwarebefällen (siehe Grafik 3.4). Mit dem Entfernen der Malware ist es heutzutage nicht mehr getan. Meistens müssen befallene Server erst vom Netz genommen und gepatched werden, damit sie für die gleiche Art von Malware nicht mehr anfällig ist. Die direkten Wiederinstandstellungskosten waren im Durchschnitt $ 81000 pro Vorfall im Jahr

2002. Im Jahr 2001 beliefen sie sich auf $ 69000. Werden die indirekten Kosten auch mit einberechnet, dann waren die durchschnittlichen Kosten pro Virus-Vorfall $ 500000.

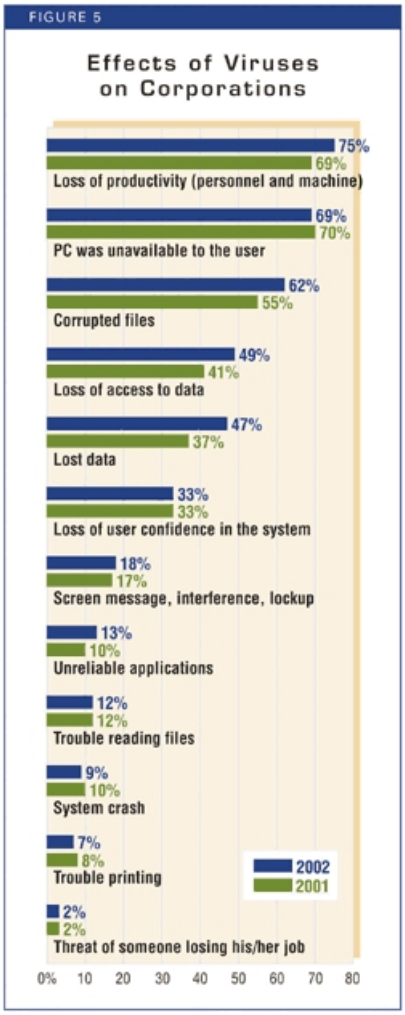


Abbildung 3.4: Statistik Internet Fraud Watch [8]

Es bleiben die wirtschaftlichen Folgen von durch die Konkurrenz gestohlenen, geheimen Daten zu beziffern. Das ist aber aus mehreren Gründen schwierig. Zum einen werden solche Vorfälle nur selten publik, da die betroffenen Unternehmen um einen Imageschaden besorgt sind. Zum weiteren lassen sich die Verluste nicht konkret in Zahlen beziffern. Der Trend bezüglich Malware ist nicht eindeutig. In jüngster Zeit haben sich die Malwareattacken nicht gemäss aller Erfahrungen weiter gehäuft, sondern sind konstant geblieben. Allerdings scheint die Reaktionszeit vom Bekanntwerden einer Sicherheitslücke bis zu deren Ausnutzung mittels Viren und Trojanern gesunken zu sein. Malwareautoren geben sich mehr Mühe, dass ihre Malware unentdeckt bleibt und länger auf dem infizierten System verbleiben kann, um ihren Zweck zu erfüllen.

3.3 Technische Aspekte

Die Begriffe Virus und Trojaner sind unter Internetbenutzern geläufig. Intuitiv wissen viele, worüber man spricht. Wenn es darum geht, die Begriffe näher einzuordnen, wird es meist schwieriger. Den Virus von einem Wurm zu unterscheiden, fällt manchem schwierig. Des weiteren gibt es eine Unzahl weiterer Begriffe, wie Backdoors, Wabbit oder KeyLogger.

Ziel dieses Abschnitts ist es, dem Leser ersichtlich zu machen, welches die Unterschiede zwischen den Begriffen sind, und wie man sie einordnen kann. Damit soll dem Leser das Verständnis über „Betrug“ und „Spionage“, für welche oft „Malware“ als Instrument eingesetzt wird, näher gebracht werden. Zuerst soll der Begriff Malware erläutert werden, dann werden die wichtigsten Konkretisierungen von Malware beschrieben. Darin kommen Formen von Betrug und Spionage vor. Im Anschluss werden zwei Methoden vorgestellt, welche von der Malware Gebrauch machen. Dem sogenannten Phishing und dem Denial of Service. Abgeschlossen wird dieser Abschnitt mit Social Engineering, eine Methode, um an vertrauliche Informationen zu gelangen, die sich gesellschaftlicher Kontakte bedient. Mit dieser Technik wird versucht, technische Barrieren zu umgehen.

3.3.1 Malware

Malware ist ein Akronym, welches zusammengesetzt ist aus „Malicious“ und „Software“, zu deutsch „Böswillige Software“. Das heisst, Malware wohnt die Absicht inne, bösartig zu sein. Auf einem Computersystem irgendeiner Art kann die Bosheit zum Ausdruck kommen, indem beispielsweise Daten gelöscht oder manipuliert werden. Es können auch Informationen gestohlen oder Netzwerkdienste überlastet werden, bis ihre Dienste zum Erliegen kommen. Malware ist ein abstrakter Oberbegriff. Hier wird eine nicht abschlossene Liste mit konkreteren Instanzen dargestellt:

Tabelle 3.1: Liste mit einigen beispielhaften Instanzen von Malware

Virus	Wabbit
Wurm	KeyLogger
Trojaner	Browser Hijacker
Dialer	Exploitr
Backdoor	Rootkit
Spyware	

Dieser Liste könnten zahlreiche weitere Namen von Malwareinstanzen angehängt werden. Doch sie soll nur veranschaulichen, dass der Begriff Malware sehr abstrakt ist und viele auch sehr unterschiedliche Konkretisierungen möglich sind. Es wird im Rahmen dieses Textes nur auf die wichtigsten dieser Begriffe näher eingegangen.

Es gibt noch einen weiteren wichtigen Punkt. Diese Begriffe stellen relativ einfache Definitionen dar. Es ist wichtig, dass man sich das vor Augen führt. Sobald eine Software einer dieser *Definitionen* genügt, zählt sie zu Malware. In der realen Welt entspricht eine konkrete Software -Malware- selten nur einer dieser Definitionen. Meistens sind es mindestens zwei. Der Grund liegt darin, dass es innerhalb der Malware Begriffe gibt, welche

zum Ausdruck bringen, was sie machen, welchen Zweck sie auf fremden Rechnern haben. Andere sagen aus, wie sie den Weg in fremde Systeme finden. Und wieder andere stellen dar, wie sie getarnt sind. Es ist nun offensichtlich, weshalb eine konkrete Malware mehreren Definitionen genügen muss. Denn auf jeden Fall muss sie einen Weg in fremde Systeme finden, und sobald sie ihn gefunden hat, soll eine Funktion ausgeübt werden. Ob getarnt oder nicht hängt jeweils vom Zweck der Malware ab, respektive vom Wunsch des Autors, der sie schreibt.

In der folgenden Tabelle wird versucht, die Begriffe nach den verschiedenen Kriterien einzuteilen:

Tabelle 3.2: Einordnung häufiger Typen von Malware nach nach den Kriterien Verbreitung, Zweck und Tarnung

Verbreitung	*Zweck*	*Tarnung*
Virus	Dialer	Trojaner
Wurm	Backdoor	
	Spyware	
	Key Logger	
	Rootkit	

Im weiteren werden die wichtigsten Malware-Typen näher vorgestellt.

Virus

Computerviren sind definiert als Programmstücke, die sich vervielfachen, in andere Programme hineinkopieren und zugleich (schädliche) Funktionen in einem Rechnersystem ausüben können. Der Virus ist Bestandteil eines anderen Programms (Wirtsprogramm) und führt seine eigenen Anweisungen vor oder während des Ablaufs des Wirtsprogramms aus. [11] Es gibt demnach drei wesentliche Punkte. Ein Virus ist eine Software. Sie kann Kopien (oder modifizierte Kopien) ihrerselbst produzieren. Man spricht auch von replizieren. Auf diesem Weg kann sie von einem System zum anderen gelangen und sich dadurch verbreiten. Der dritte Punkt ist, dass der Virus einen Wirt benötigt. Als Wirt in Frage kommen bestehende, selbstständige Programmstücke. Der Virus allein stellt keine Gefahr dar. Erst durch das Einschleusen in ein anderes lauffähiges Programm kann er etwas ausüben.

Es ist wichtig, zwischen der Infektion und dem Ausbruch eines Viruses zu unterscheiden. Das erste Ziel eines Virus besteht darin, sich zu verbreiten. Er versucht sich in so viele Rechner wie möglich zu schleusen, ohne dass es dem Anwender auffällt. Es kann sein, dass sich über Jahre ein Virus in einem Rechner aufhält, ohne dass der Anwender Notiz davon nimmt. Man spricht davon, dass der Rechner infiziert worden ist, sobald der Virus sich einkapseln konnte, wenn er sich also in eine lauffähige Datei hineinkopieren konnte. In diesem Zustand macht der Virus nichts. Erst der Ausbruch, respektive die Aktivierung veranlasst ihn, seine Funktionen, welche ihm eingebaut sind, wahrzunehmen. Dabei gibt es zwei verschiedene Varianten, welche die Aktivierung auslösen. Erstens die Aktivierung in Abhängigkeit von der Zeit, auch Time bomb genannt. Der Ausbruch findet statt, nach

Erreichung eines Zeitpunktes. Beispielsweise, wird er in Abhängigkeit vom Eintritt einer vorgegebenen Tageszeit oder eines bestimmtem Datums gebracht. Zweitens gibt es eine logische Aktivierung, auch Logic bomb oder trigger genannt. Dieser lässt sich wiederum unterscheiden nach der Art des Auslösers. Auf der einen Seite kann die Ausführung eines bestimmten Vorgangs des Benutzers die Aktivierung auslösen. Dann wird die Aktivierung an eine Funktion eines Anwendungsprogramms gekoppelt. Oft wird hierfür die Exitfunktion gewählt, da sie sonst nicht sehr oft verwendet wird und relativ unauffällig bleibt. Auf der anderen Seite kann die Kopplung an die Durchführung eines Prozesses im System erfolgen.

Eine Virussoftware besteht aus unterschiedlichen Teilen. Um die Funktionsweise von Viren zu durchleuchten, sollen diese hier aufgeführt werden.[14]

Entschlüsselungsroutine Dieser Teil sorgt bei verschlüsselten Viren dafür, dass die verschlüsselten Daten wieder zur Ausführung gebracht werden können. Nicht alle Viren besitzen diesen Teil, da nicht alle verschlüsselt sind. Oft wird die Entschlüsslungsroutine der Viren von Antiviren-Herstellern dazu benützt, den Virus zu identifizieren, da dieser Teil oft klar erkennbar ist.

Vermehrungsteil Dieser Programmteil sorgt für die Vermehrung des Virus. Es gibt keine Viren ohne diesen Teil, weil er Bestandteil der gegebenen Definition eines Viruses ist.

Erkennungsteil (Signatur) Im Erkennungsteil wird geprüft, ob die Infektion eines Programms oder Systembereichs bereits erfolgte. Jedes Wirtsprogramm wird nur einmal infiziert, weil er sonst die Grösse der Wirtsdatei auffallend verändern würde und/ oder durch mehrfache Ausführung auffallen würde, da dies zu einem wesentlich höheren Resourcenverbrauch führen würde. Polymorphe Viren sind in der Lage, mit verschiedenen Signaturen zu arbeiten, die sich verändern können, jedoch stets einer Regel gehorchen. Der Erkennungsteil wird von praktisch allen Computerviren benützt.

Schadensteil (Payload) Im Verhältnis zur Zahl der Computerviren haben nur sehr wenige einen Schadensteil. Der Schadensteil ist der Grund für die Angst vieler Menschen vor Computerviren.

Bedingungsteil Der Bedingungsteil ist dafür verantwortlich, dass der Schadensteil ausgeführt wird. (Aktivierung siehe weiter oben im Text). Er ist in den meisten Computerviren mit einem Schadensteil enthalten.

Tarnungsteil Ein Tarnungsteil ist nur in wenigen, komplexen Viren vorhanden. Er kann den Virus zum Beispiel verschlüsseln, oder ihm eine andere Form geben (Polymorphismus, Metamorphismus). Dieser Teil dient zum Schutz vor Erkennung durch Anti-Viren Herstellern. Es gibt aber nur eine sehr geringe Anzahl von Viren, die nicht vollständig erkannt werden können.

Die Verbreitung eines Viruses geschieht, indem er Dateien infiziert. Er integriert sich in eine ausführbare Datei, in einigen Fällen auch in einen Bootsektor oder als Makro in eine

interpretierbare Datei und wird somit Teil einer schon bestehenden Programmroutine. Die Verbreitung des Virus erfolgt durch Weitergabe dieser infizierten Dateien. Auf welchem Wege sie weitergegeben werden, ob via Datenträger oder via Netzwerke, ist für die Definition „Virus“ unerheblich.

Bekannterweise können Viren schädlich sein. Oft sind in einem Virus Funktionen enthalten, welche Schaden anrichten können. Sie können Dateien manipulieren, löschen oder weitere Dateien hinzufügen. Diese sehr einfache Darstellung spricht nur Daten an, nicht aber die Informationen, welche darin enthalten sind. Das Schadenspotential wird immens grösser, wenn aus den Daten Information gewonnen werden kann. Wie ist die Situation, wenn die Dateien private Informationen enthalten? Was ist, wenn sie dahingehend kritisch sind, dass Anwendungsprogramme nicht oder nicht fehlerfrei funktionieren? Viele Menschen fürchten sich vor den Viren wegen dieser Schadfunktionen und sind sich nicht bewusst, dass die wenigsten Viren eine Schadfunktion haben, welche tatsächlich Daten löscht oder den Rechner zum Erliegen bringt. Das häufigste Ziel von Viren ist, sich zu verbreiten und einen höchstmöglichen Verbreitungsgrad zu erlangen. Um das zu erreichen, ist es notwendig, dass er nicht entdeckt wird. Er muss sich deshalb möglichst unauffällig verhalten.

Exkurs Im Zusammenhang mit der Schädlichkeit ist folgender Punkt interessant: Viren lassen sich auch nach dem Betriebssystem einordnen, für welche sie geschrieben und für welche sie kompatibel sind. So, gibt es Schätzungen, welche besagen, dass es über 60’000 Viren für das Betriebssystem MS Windows geben soll, deren 40 sowohl für Macintosh als auch für Linux und etwa 5 für kommerzielle Unix Versionen. Das Dokument mit diesen Schätzungen geht auf den Oktober 2001 zurück. Dies mindert jedoch nicht die Aussagekraft darüber, dass die Anzahl Windowsviren diejenige der anderen Betriebssysteme bei weitem übersteigt, denn man geht nicht davon aus, dass sich die Verhältnisse inzwischen stark verändert haben. Die Wahl des Betriebssystems hat einen grossen Einfluss auf die Wahrscheinlichkeit einer Infektion. [12]

Linkviren

Diese Viren hängen sich an bereits vorhandene Programme oder Programmbibliotheken und werden immer mit dem infizierten Programm zusammen gestartet. Der Name Linkviren (auch Dateiviren) kommt daher, dass sie immer im Verbund mit einer anderen Datei funktionieren, also einen Link in andere Dateien haben. Die Infektion findet statt, indem sich der Linkvirus in die Wirtsdatei einfügt.[13]

Um auch angesprochen respektive zur Ausführung gebracht werden zu können, muss die Wirtsdatei derart manipuliert werden, dass sie den Virus aufruft. Es gibt unterschiedliche Möglichkeiten, wo in einen Wirt sich ein Virus einfügen kann. Man unterscheidet zwischen Appender und Prepender, wobei erstere Variante häufiger zur Anwendung kommt.[14] Im Falle eines Appenders wird der Virus am Ende des Codes des Wirtsprogramms eingeschleust. Wie bereits erwähnt, muss die Wirtsdatei modifiziert werden, damit der Virus zur Ausführung kommt. Die Modifikation veranlasst das Wirtsprogramm zuerst den Virus aktiv werden zu lassen. Dann führt der Virus das Wirtsprogramm aus, indem er an den ursprünglichen Programmeinstiegspunkt springt.

Bei der Prepender-Variante wird der Virus am Anfang einer Wirtsdatei eingefügt. Dabei wird beim Ausführen der Wirtsdatei zuerst der Virus aktiv, der sich entweder weiterverbreitet oder seine Schadwirkung entfaltet. Erst in zweiter Instanz wird das Wirtsprogramm ausgeführt. Beim Starten des Wirtsprogramms kann ein kleiner Zeitverlust auftreten, der vom Anwender jedoch oft nicht bemerkt wird. Es gibt auch Viren, welche sich weder am Anfang noch am Ende eines Wirtsprogrammes einnisten, sondern irgendwo dazwischen. Der Aufwand der Implementierung ist in dieser Variante erheblich höher, weil es komplizierter ist. Diese Virentypen werden Entry Point Obscuring genannt, was zu deutsch mit „Verschleierung des Einsprungspunkts“ übersetzt werden kann.

Bootviren

Viren dieser Art infizieren den Bootsektor von Wechseldatenträger und Festplatten respektive deren Partitionen. Um Viren dieser Art zu schreiben, erfordert es vom Auter viel mehr technisches Wissen und Fähigkeiten als für andere Viren. Ausserdem ist die Verbreitung langsam und bietet dem Autor wenig Möglichkeiten. In diesem Text soll nicht weiter auf diese Virenart eingegangen werden, da sie aus den genannten Gründen wenig Verbreitung findet.

Wurm

Die Computerwürmer sind verwandt mit den oben beschriebenen Viren. Der wesentlichste Unterschied liegt darin, dass Würmer selbstständige Programme sind. Sie benötigen also keine Wirtsprogramme. Ihr „Lebensbereich“ sind die Rechnernetze. Ein Wurm kann eine Kopie von sich an andere Rechner verschicken. Hierzu muss er das Protokoll des jeweiligen Netzes kennen und er muss die Adressenliste, in welcher die einzelnen Knotenrechner des Netzes verzeichnet sind, inspizieren können.[11] Hilfreich für das Verständnis ist der geschichtliche Hintergrund der Würmer. Ursprünglich wurden Würmer erstellt, um Kontrollfunktionen in Rechnernetzen zu übernehmen. Sie verbreiteten sich willkürlich im Netz und kamen in einem Knoten immer zu dem Zeitpunkt zur Ausführung, da dieser keine anderen Aufgaben zu bewältigen hatte. Die Aufgabe des Wurms bestand darin, diesen Knoten auf die Performance und auf die Funktionsfähigkeit zu inspizieren. Im Falle eines Problems konnte der Wurm eine Meldung auslösen.

Es gibt verschiedene Wege, wie sich ein Wurm tarnen kann. Es kann die Endung der Datei derart gestaltet werden, dass der Anwender glaubt, es handle sich um eine Bilddatei. Er sieht am Ende des Dateinamens „.jpg“ stehen. In Wirklichkeit endet der Name jedoch mit „.exe“. Eine Endung, welche im Betriebssystem Windows für ausführbare Dateien steht. Das Betriebssystem verschleiert die Endungen aller Dateien, und das nutzt der Autor des Wurmes aus. (siehe auch weiter unten, Trojaner). Ein weiterer Weg um einen Wurm zu tarnen, besteht darin, eine Endung für den Dateinamen zu wählen, welche einerseits ausführbar ist, welche jedoch auf der anderen Seite den meisten Windowsbenutzern unbekannt ist. Beispiele hierfür sind „.pif“ (steht für DOS-Datei-Verknüpfungen), „.scr“ (wird von Bildschirmschonern verwendet), „.vbs“ (Visual Basic Script Dateien), und schliesslich „.bat“ (für DOS-Batch-Dateien).

Aus Sicht des Wurms ist es am einfachsten, wenn er ohne Zutun des Anwenders aufgerufen werden kann. Prinzipiell dürfte das vom Betriebssystem jedoch nicht zugelassen werden. Deshalb ist bei dieser Variante der Tarnung der Wurm auf das Vorhandensein von Sicherheitslücken angewiesen. Des weiteren gibt es auch Möglichkeiten, sich Sicherheitlücken von Anwendungsprogrammen zu bedienen. Typischerweise werden hierfür E-Mailprogramme herangezogen. Die Anlage von E-Mails können als HTML geschrieben werden. Dieser HTML-Code enthält Skripte, wie beispielsweise Javascript und über diese Skripte vermag sich der Wurm Eingang in den Rechner zu verschaffen. Denn die Sicherheitslücken in der Implementierung des Anwendungsprogramms lassen es zu, dass sich der Wurm - nur aufgrund des Öffnens einer E-Mail - starten lässt.

Für Würmer stehen mehrere Wege zur Verfügung, um sich verbreiten zu können. Grundsätzlich gibt es keinen grossen Unterschied, welcher Weg gewählt wird. Das Prinzip ist stets ähnlich. Eine erste Möglichkeit wurde bereits erläutert: Die E-Mail. Es können auch E-Mails verschickt werden, welche den Wurm im Anhang haben. Der Empfänger muss dann dazu verführt werden, den Anhang zu öffnen. Alternativ können Würmer sich über Instant-Messaging-Programme verbreiten. Das sind Programme, welche zum Chatten grosse Verbreitung gefunden haben. Bekanntestes Beispiel ist der MSN Messenger. Es wird versucht, allen Kontakten einen Link zu der Seite zu senden, die den Wurm enthält. Im Grunde genommen können Würmer immer dann zum Einsatz gebracht werden, wenn Nachrichten mit Anhang oder Dateien zwischeneinander ausgetauscht werden. So werden in ähnlicher Funktionsweise Würmer auch über File Sharing-Programme und Peer-to-Peer-Netzwerke eingesetzt.

Trojaner

Trojanische Pferde haben ihren Namen aus den Überlieferungen rund um Homers Epos Ilias, in dem von einem grossen Pferd die Rede ist, in dessen Bauch sich die Griechen versteckt hatten, um in die Stadt Troja zu gelangen. Mittels List, hier mit dem Tarnen als harmloses Geschenk, wurde für den Empfänger unerkannt etwas Gefährliches, Zerstörerisches in Gang gesetzt.[15] Ähnlich verhält es sich mit den Computer-Trojanern. Hinter einer anscheinend harmlosen Datei befindet sich ein weiteres Programm versteckt, das sich unbemerkt selbst installiert und unterschiedliche Funktionen ausführen kann.[17] Haben solche Programmstücke den Weg in einen Rechner gefunden, sind sie nur sehr schwierig zu entdecken, da sie über einen langen Zeitraum hin „schlafen“ und erst aufgrund eines Ereignisses ihre oft negative Wirkung entfalten.

Am Anfang waren die Trojaner gezielt auf Opfer zugeschnitten und eingesetzt worden. Das Ziel hatte darin bestanden, bestimmte Informationen respektive Daten zu erlangen (Spionage). Heute werden sie in Formen realisiert, in welchen sie sich weniger gezielt ausbreiten. Hierfür werden sie beispielsweise als Würmer implementiert. Für Trojaner steht die Tarnung im Zentrum. Trojaner sind ausführbare Dateien. Sie sind demnach davon abhängig, dass jemand sie aufruft. Der Anwender verlässt sich normalerweise auf das Erscheinungsbild des Icons, welches der Datei zugeordnet ist, um zu wissen, von welcher Art die Datei ist. Da die Zuordnung der Icons auf Dateien jedoch manuell verstellbar ist, kann der Autor eines Trojaners diese derart manipulieren, dass der Anwender getäuscht wird. So kann es sein, dass im guten Glauben eine Bilddatei geöffnet wird. Später wird

sich jedoch herausstellen, dass es sich um eine ausführbare Datei gehandelt hat und mit dem Anklicken die unkontrollierte Ausführung des Schadprogramms in die Wege geleitet worden ist.

Die Grösse solcher Programme ist je nach Typ unterschiedlich. In Maschinensprache umfassen Würmer meist über 1'000, Viren zwischen 100 und 300 Befehlen und Trojanische Pferde deren mehr als 50.[11]

Weitere Ausprägungen von Malware

Nachdem die wichtigsten drei Formen von Malware dargestellt sind, sollen weitere Typen von Malware beschrieben werden. Typisch ist, dass die folgenden Namen der Begriffe alle auf das Ziel der Attacke hinweisen und dass sie alle in der einen oder anderen Form mit den drei besagten Typen installiert werden.[7]

Backdoor Das Ziel eines Backdoorsbesteht darin, den Authentifikationsprozess, welcher normalerweise durch den Anwender durchlaufen werden muss, zu umgehen. Der Autor ist bestrebt, Zugang zu einem fremden Rechner zu erlangen, ohne dass es der Anwender bemerkt. Der Backdoor kann als eigenständiges Programm implementiert werden, welches im Hintergrund läuft. Oder er kann auch als Virus, das heisst in einem legitimen Anwendungsprogramm, eingekapselt sein.

Spyware Es handelt sich hier um eine Software, welche die Aufgabe hat, Informationen von dem Rechner, auf dem sie sich befindet, ausfindig zu machen und an den Autor über das Netzwerk weiterzuleiten. Der Anwender soll auch hier nicht im Bilde der Existenz der Malware sein. Gemäss Earthlink sollen ungefähr 90% aller am Internet angeschlossenen Rechner von Spyware infiziert sein.[16] Meistens sind Spywareprogramme geschrieben, um das Verhalten des Anwenders im Internet insbesondere das Kaufverhalten zu erkunden. Es gibt jedoch auch Varianten, welche weiterreichende Ziele verfolgen. Beispielsweise sollen persönliche Informationen für Diebstähle, oder andere unlautere Vorhaben mit Spyware ergattert werden.

Wabbit Wabbits sind selbstreplizierende eigenständige Programme. Sie infizieren dabei keine anderen Programme oder Dateien, sondern sind selbstständig. Was sie von Würmern unterscheidet, ist ihre Verbreitung. Während Würmer sich auf einem Rechnernetz verbreiten, besteht die Absicht des Wabbits darin, sich lokal auf der Maschine zu verbreiten. Das Resultat ist offensichtlich. Der Wabbit bindet derart viele Resourcen an sich, dass der Rechner bald nicht mehr funktionsfähig ist. Man spricht bei solchen Attacken von Denial of Service (DoS), siehe Abschnitt weiter unten.

KeyLogger KeyLogger werden verwendet, um die Tastatur eines Rechners zu beobachten und zu protokollieren, was der Benutzer eingibt. Mit KeyLogger können Passwörter, PINs und dergleichen ausfindig gemacht werden. Es gibt verschiedene Varianten. Einige speichern die Eingaben lokal auf der Festplatte. Das setzt voraus,

[7] Quelle: Wo nicht anders vermerkt, gilt [13]

dass der Angreifer sich zu dem Rechner Zugriff verschaffen kann. Andere speichern die gewonnenen Daten nicht lokal, sondern sie senden sie über das Netzwerk an einen bestimmten Empfänger.

Zwei Methoden von Attacken

Phishing Das Wort ist eine Zusammenstetzung aus *P*assword, *H*arvesting und F*ishing*. Es lässt sich daraus ableiten, was mit Phishing erreicht werden soll. Mittels Phishing versuchen Betrüger, an vertrauliche Daten, wie Login-Daten von Internetbenutzern zu gelangen. Dabei kann es sich beispielsweise um Kontoinformationen von Online-Auktionsanbietern oder Zugangsdaten für das Internet Banking handeln.

Die Betrüger nutzen die Gutgläubigkeit ihrer Opfer aus, indem sie ihnen E-Mails mit gefälschten Absenderadressen zustellen. Diese E-Mails sind in HTML-Code geschrieben und haben ein Layout mit Bildern und Logos, welche demjenigen des Instituts, welches sie zu sein vorgeben, identisch ist. In den E-Mails wird das Opfer beispielsweise darauf hingewiesen, dass seine Kontoinformationen und Zugangsdaten nicht mehr sicher oder aktuell sind und es diese unter dem im E-Mail aufgeführten Link ändern soll. Der Link führt dann allerdings nicht auf die Originalseite des jeweiligen Dienstanbieters, sondern auf eine vom Betrüger identisch aufgesetzte Webseite.

Denial of Services (Dos) Diese Attacken zielen darauf ab, einen Serverdienst unzugänglich zu machen oder außer Betrieb zu setzen. Bei DoS-Attacken wird ein Server gezielt mit so vielen Anfragen konfrontiert, dass das System die Aufgaben nicht mehr bewältigen kann und im schlimmsten Fall zusammenbricht. Auf diese Weise wurden schon Web-Server bekannter Unternehmen wie zum Beispiel Amazon, Yahoo, eBay, mit bis zur vierfachen Menge des normalen Datenverkehrs massiv attackiert und für eine bestimmte Zeit für normale Anfragen außer Gefecht gesetzt. Solche Angriffe werden mit Backdoor oder ähnlichen Programmen umgesetzt, die sich selbstständig auf dem Rechnernetz verbreiten und dadurch weitere Wirte zum Ausführen von Angriffen bringen. Eine andere Variante von DoS, welche nicht auf Servern, sondern auf jeder Art von Rechnern durchführbar ist, wird - wie oben beschrieben - von Wabbits verursacht.

Schlusswort

Der Tatbestand, dass sehr oft nicht zwischen einem Virus und einem Wurm differenziert wird, sollte man nicht stehen lassen. Oft werden in Literatur und im Internet noch viel mehr, Würmer als Viren gleichgestellt. Dabei heisst der Oberbegriff Malware, und dieser wäre meist zutreffender. Auch der Duden Informatik beispielsweise führt den Begriff Wurm im Text über Viren auf. Man muss den Begriff Virus nachschlagen, um etwas über den Wurm in Erfahrung zu bringen.

Der Hintergrund dieser unachtsamen Verwendung der Begriffe liegt in der Geschichte. Den Virus gibt es schon länger. In Zeiten, da das Internet noch keine grosse Verbreitung genossen hatte, konnten die Dateien und Programme nur über Disketten von einem System

zum anderen gelangen. In dieser Zeit waren die Viren die dominierende Variante. In dieser Zeit ist es dem Begriff Virus gelungen, sich derart zu etablieren, dass heute zu oft von Viren gesprochen wird, wo Malware gemeint ist.

3.3.2 Social Engineering

Social Engineering (auch Social Hacking) ist das Erlangen vertraulicher Informationen durch Annäherung an Geheimnisträger mittels gesellschaftlicher Kontakte. Dieses Vorgehen wird von Geheimdiensten und Privatdetektiven seit langem praktiziert, der Begriff wird jedoch meist im Zusammenhang mit Computerkriminalität verwendet, da er hier das Gegenstück zum rein technischen Vorgehen beim Eindringen in fremde Systeme bildet.[18] Durch Täuschung oder Beeinflussung von Mitarbeitern können technische Barrieren umgangen oder ausgehebelt und effizient überwunden werden. Solche Szenarien drohen insbesondere dann, wenn die Informationen oder Systeme des Unternehmens gezielt und in krimineller Absicht angegriffen werden, beispielsweise im Bereich der Industriespionage. Dabei nutzen die Angreifer die Unkenntnis oder Arglosigkeit der betroffenen Mitarbeiter aus, in dem sie diese - oft unter Vortäuschung einer falschen Identität - dazu veranlassen, wichtige Informationen wie Passwörter oder Netzwerkadressen preiszugeben oder gar selbst Einstellungen an IT-Systeme zu manipulieren.

Wenn die Sicherheit der Technischen Systeme höher ist als die organisatorische und mitarbeiterbezogene Sicherheit, dann wird ein geplanter Angriff auch eher auf die Mitarbeiter gerichtet sein als auf die technischen Systeme. Um z.B. an das Passwort für eine Web-Applikation mit vertraulichen Daten zu gelangen, ist es häufig einfacher, das Passwort unter falschen Vorwänden von einem Nutzer zu erfragen, als auf technischem Wege die Verschlüsselung zu brechen, die das Passwort bei der Übertragung und Speicherung schützt. Solche Angriffe werden leichter möglich, wenn sich die Mitarbeiter in Sicherheit wiegen, weil sie sich durch technische Sicherheitssysteme wie Firewalls und Virenschutzsoftware ausser Gefahr glauben.

Angriffstechniken, bei denen - wie oben dargestellt - der Angreifer Mitarbeiter der angegriffenen Organisation einbezieht, indem er sie durch eine Täuschung dazu bringt, ihm Informationen preiszugeben, Sicherheitsmassnahmen ausser Kraft zu setzen oder in anderer Weise das Gelingen des Angriffs zu unterstützen, bezeichnet man als Social Engineering. Angriffe durch Social Engineering werden zumeist kaum in der Öffentlichkeit bekannt. Zum einen ist es für viele Firmen peinlich, derartige Angriffe zuzugeben, zum anderen geschehen viele Angriffe so geschickt, dass diese nicht oder erst viel später aufgedeckt werden.[19]

3.4 Erkennung, Schutz, Abwehr - Prävention

Ergänzend zu den vorangegangenen Kapiteln, die sich mit den unternehmerischen und ökonomischen Aspekten von Verlusten, Betrug und Wirtschaftsspionage und den technisch-konzeptionellen Aspekten auseinander setzten, wird in diesem Kapitel auf die Erkennung

(Detection) und den Schutz (Defense) näher eingegangen. Fragen wie: „Welche Möglichkeiten stehen einer Unternehmung zum Schutz zur Verfügung?“ „Was kann getan werden, um einen Angriff rechtzeitig zu erkennen und abzuwehren?“ und „Wie ist die unternehmerische Prävention sinnvoll zu gestalten?“ werden dabei behandelt.

Zuerst wird das Thema in Punkt 3.4.1 konzeptuell angegangen. Unter 3.4.2 werden zunächst technische, danach unter 3.4.3 organisatorische und mitarbeiterbezogene Schutzkonzepte aufgeführt. Es werden pro Konzept einige konkrete Erscheinungs-Formen exemplarisch beschrieben. 3.4.4 fasst die Erkenntnisse und die behandelten Schutzmassnahmen dieses Kapitels abschliessend zusammen.

3.4.1 Sicherheits-Konzept

Eine IT-Sicherheitsstudie von security.de 2004 [20] bestätigt die Relevanz von Sicherheitsbedrohungen, wie sie im vorangegangenen Kapitel unter „technische Aspekte“ beschrieben wurden.

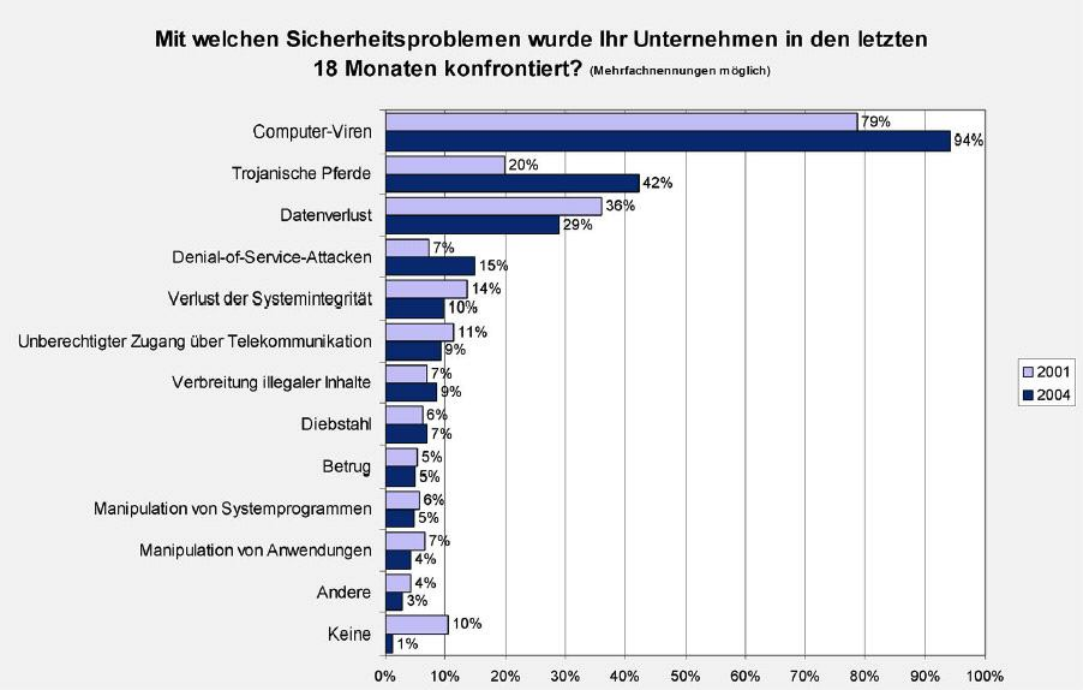


Abbildung 3.5: Mit welchen Sicherheitsproblemen wurde Ihr Unternehmen in den letzten 18 Monaten konfrontiert? [20]

Ein Sicherheitskonzept zeigt auf, wie den herrschenden Sicherheitsproblemen von Seiten der Unternehmung am besten entgegengewirkt wird. Die unter Abbildung 3.5 dargestellten komplexen Gefahren erfordern ganzheitlich integrierte Sicherheitslösungen in den Unternehmungen, die sowohl organisatorisch als auch technisch sind. Ein individuell auf ein Unternehmen zugeschnittenes Sicherheitsmanagementsystem ist eine unabdingbare Voraussetzung für sowohl dauerhafte wie konstante Sicherheit.[19]

Ein umfassendes Sicherheitskonzept hat folgende strategische Ziele zu erfüllen: Integrität (Unversehrtheit und Korrektheit der Daten), Vertraulichkeit (Schutz vor unautorisierter Kenntnisnahme), Verfügbarkeit (Erreichbarkeit, Zuverlässigkeit, Wartbarkeit) und Authentizität (Echtheit, Glaubwürdigkeit).[19]

Aus unternehmerischer Sicht bilden sich als weitere wichtige Zielsetzungen: Die Sicherung der Geschäftskontinuität, der Schutz der Unternehmensreputation bzw. ein Imagegewinn, die Förderung des Vertrauensverhältnisses mit den Geschäftspartnern, die Minimierung finanzieller Risiken und die Einhaltung gesetzlicher Vorschriften sowie die Vermeidung der Kosten inklusive Folgekosten aus Sicherheitsvorfällen. [19]

„IT-Sicherheit kann nicht durch unkoordinierte Einzelmassnahmen erreicht werden, sondern nur durch die Anwendung von aufeinander abgestimmten technischen, organisatorischen und personellen Massnahmen.“ [19]

Die folgenden Unterkapitel behandeln die Gestaltung eines Sicherheits-Konzeptes sowie die Ermittlung der Gefährdungslage und die vier grundsätzlichen Bedrohungsklassen zur Ermittlung der Gefährdungslage (für eine Unternehmung).

Gestaltung eines Sicherheitskonzeptes [19]

Die wichtigsten drei Bausteine eines ganzheitlichen Sicherheitskonzepts sehen wie folgt aus:

1. Eine Analyse der spezifischen Risiken des jeweiligen Unternehmens steht an erster Stelle. Dabei gilt es zu untersuchen, was die schützenswerten, materiellen und immateriellen Güter sind, welche Gefahren ihnen drohen und wie hoch sich das jeweilige Risiko beläuft.

2. Basierend auf den Ergebnissen der Risikoanalyse ist anschliessend ein Massnahmenkatalog zu erarbeiten und umzusetzen. Die Kapazitäten (Kosten, Zeit, Personal, etc.) werden in der Regel nicht ausreichen, um sofort das gesamte System durch Massnahmen wie Backup, Virenschutz, Firewall, Festlegung der Zugangs- und Zugriffsrechte, Schaffung eines Sicherheitsbewusstseins in der Belegschaft etc. vollkommen zu schützen.

3. Alle Sicherheitsmassnahmen sind als ein Grundprinzip stetig fortzuschreiben. „Wenn man eine Firewall hat, die nicht kontinuierlich gepflegt wird, sollte man diese lieber abschalten. Dann hat man wenigstens nicht mehr das trügerische Gefühl, man sei sicher.“(Frank Rustemeyer, secunet Security Networks AG)

Unternehmen können bei der Erstellung eines umfassenden Sicherheitskonzeptes auf verschiedene Standards zurückgreifen.

Die vier grundsätzlichen Bedrohungsklassen zur Ermittlung der Gefährdungslage [19]

„IT-Sicherheit bedeutet die Gewährleistung eines angemessenen Schutzes der Vertraulichkeit, Verfügbarkeit, Integrität und Nachvollziehbarkeit der elektronischen Datenverarbeitung." Um allen diesen Schutzzielen gerecht zu werden, erfordert sie die Betrachtung eines breiten Spektrums von Bedrohungen, die sich in folgende vier Klassen zusammenfassen lassen:

Technik: Technisches Versagen oder Fehlfunktion von Hardware oder Software sind eine häufige Quelle von Schäden, die insbesondere die Verfügbarkeit betreffen. Oft jedoch werden durch ihre mittelbaren oder unmittelbaren Auswirkungen auch die anderen Schutzziele tangiert. In besonderem Mass gilt dies für Fehler in der Software.

Höhere Gewalt: Übergreifende Schadenereignisse wie Natur-Katastrophen oder Terror-Aktionen können die IT-Systeme in ihrer Funktionsfähigkeit beeinträchtigen. Dabei sind insbesondere die indirekten Auswirkungen wie z.B. der Personalausfall nach einer Grippe-Epedemie zu beachten. Von Bedrohungen durch höhere Gewalt wird in erster Linie das Schutzziel der Verfügbarkeit betroffen.

Organisation: Organisatorische Mängel umfassen z.B. fehlende Regelungen und Anweisungen, mangelnde Kontrollen und Tests sowohl beim Personal als auch bei der eingesetzten Software oder die unzureichende Dokumentation von den vorhandenen IT-Systemen. Mängel dieser Art führen zwar häufig nicht unmittelbar zu Schäden, sie bergen jedoch grosse Risiken in Form von latenten Bedrohungen, welche insbesondere in der Kombination mit anderen Bedrohungen wie unerkannte technische Mängel, fehlende Verhaltensregeln für Personal etc. wirksam werden.

Mensch: Menschliches Fehlverhalten verursacht die am weitaus umfassendste Gruppe von Bedrohungen. Dabei gilt es zu unterscheiden zwischen vorsätzlichen Handlungen, bei denen Mitarbeiter oder externe Angreifer bewusst versuchen dem Unternehmen Schaden zuzufügen und fahrlässigem oder irrtümlichem Handeln von Mitarbeitern, was ohne Vorsatz des Mitarbeiters zu Schaden führt. Während das Schadenspotential bei vorsätzlichen Handlungen durch das gezielte Vorgehen des Angreifers besonders hoch ist, bildet gerade das nicht vorsätzliche Fehlverhalten von Mitarbeitern in der Praxis eine der Hauptursachen für die auftretenden Schäden durch Malware, Daten- oder Informations-Verluste oder Spionage. Im dem Sinne sind auch die immer wieder zu ähnlichen Ergebnissen führenden Studien zu verstehen, welche besagen, dass ein wesentlicher Anteil (70-80%) der IT-Vorfälle von internen Mitarbeitern verursacht werden.

Während die Bedrohungen der ersten beiden Kategorien, Technik und höhere Gewalt, sich vor allem auf dingliche Gegebenheiten in Unternehmen beziehen (Technik, Gebäude, Brandschutz etc.), betreffen die Bedrohungen der beiden letztgenannten Kategorien, Organisation und Mensch, direkt die Mitarbeiter.

In den folgenden Kapiteln werden technische Massnahmen vorgestellt, welche potentiellen Bedrohungen entgegen wirken. Darin wird auf die Schutzkonzepte Firewalls, Intrusion

Detection Systems und Honeypots eingegangen. Im Anschluss werden Inhalte und Aufbau eines Security Awareness Konzeptes als Massnahme und Schutzkonzept der organisatorischen Ebene und der Ebene Mensch ergänzend aufgeführt. Abschliessend folgt eine Zusammenfassung.

3.4.2 Technische Schutzkonzepte

Zum Schutze der eigenen Computersysteme und Netzwerke existieren zahlreiche technische Tools und Schutzkonzepte, welche, wie die Graphik aus der IT-Sicherheitsstudie siehe Abbildung 3.6 belegt, zum Einsatz kommen.

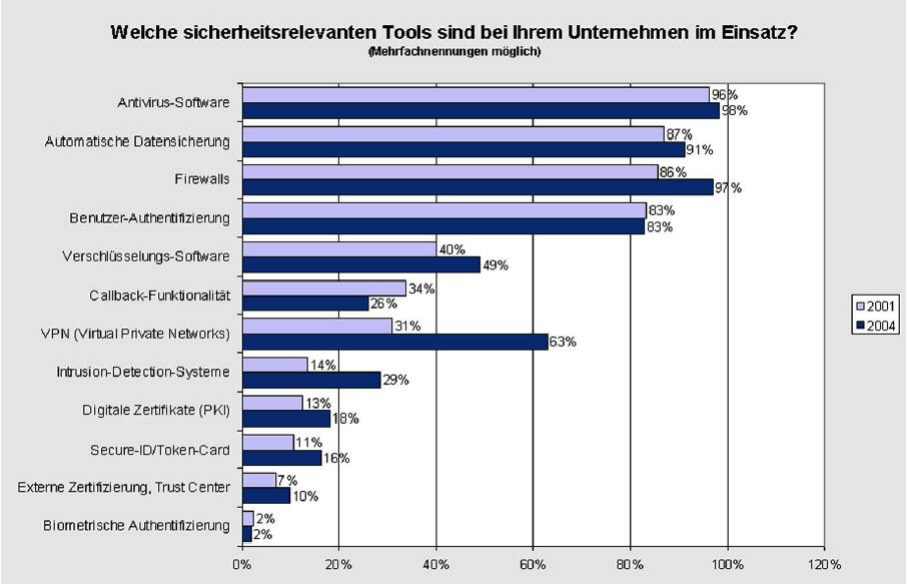


Abbildung 3.6: Welche sicherheitsrelevanten Tools sind in Ihrem Unternehmen im Einsatz? (Vgl. 2001 zu 2004) [20]

Im Folgenden werden exemplarisch drei mehr oder weniger bekannte solche technische Konzepte vorgestellt, die für die Sicherheit innerhalb und zwischen Netzwerken zum Einsatz kommen, ausgeführt. Computer, Informationssysteme, Infrastruktur und Daten sind deutlich weniger verwundbar, wenn die Netzewerke, die sie umgeben, optimal überwacht und gesichert sind.

Zunächst wird das Prinzip einer Netzwerk-Firewall und dann ein Intrusion Detection System erklärt. Anschliessend wird noch ein weiteres in der Allgemeinheit noch eher weniger bekanntes Schutz-Konzept mit der Bezeichung Honeypot, welches vorwiegend Erkennung von Angriffen dient, festgehalten. Auf die Wirksamkeit technischer Schutzkonzepte wird in einem letzten Unterkapitel noch kurz eingegangen.

Firewalls [21]

Eine Firewall (vom engl. Begriff „Brandwand“ abgeleitet) ist ein System aus Software- und Hardwarekomponenten, welches den Zugriff zwischen verschiedenen Rechnernetzen

beschränkt, um ein Sicherheitskonzept umzusetzen. Hardwarekomponenten einer Firewall sind beispielsweise Rechner mit Netzwerkschnittstellen wie Router oder Hosts. Softwarekomponenten sind Paketfilter oder Proxyserver. Der häufigste Einsatzzweck einer Firewall besteht darin, den Datenverkehr zwischen einem zu schützenden lokalen Netzwerk (LAN) und dem Internet zu protokollieren.[22]

Firewalls werden an den Schnittstellen zwischen einzelnen Netzen oder Computersystemen gesetzt, damit sie den Datenverkehr zwischen den Teilbereichen kontrollieren, um in ihrer Funktion ungewünschten Verkehr beispielsweise ein Angriff von aussen zu verhindern und nur einen gewünschten Verkehr passieren zu lassen. Die zwei wesentlichen Aufgaben einer Firewall liegen also im Unterbinden von einerseits ungewolltem Datenverkehr von externen Computersystemen zum geschützten Bereich und vom geschützten Bereich zu externen Systemen andererseits.[23]

Umgangssprachlich meint man mit einer Firewall häufig jene Software, welche den Datenverkehr zwischen getrennten Bereichen kontrolliert und regelt. Weil das durch die landläufige Vorstellung assoziierte Verständnis des Begriffs Firewall zwar teilweise stimmt, jedoch auch gewisse Mythen und Halbwahrheiten beherbergt, sollte der Begriff differenziert betrachtet werden.[22]

Zunächst einmal sollte zwischen dem (Sicherheits-)Konzept Firewall und der konkreten Firewall-Realisierung unterschieden werden. Das Sicherheitskonzept beschreibt Regeln, um zu bestimmen, welche Informationen die Firewall passieren dürfen und welche nicht. Realisiert wird das Konzept durch eine Software, welche auf einer (oftmals speziellen) Hardware läuft.[22]

Die umgangssprachliche Unterscheidung zwischen Hardware- oder Software-Firewalls macht eigentlich wenig Sinn, weil es sich bei dieser Unterscheidung in erster Linie um eine nichttechnische Definition handelt. Zu jeder Firewall gehört Software, welche auf einer Hardware ausgeführt werden muss. Aufgrund von diesem Missverständnis werden z.B. dedizierte Router, auf denen die Firewallsoftware ausgeführt wird, fälschlicherweise als Hardware-Firewall bezeichnet oder Personal-Firewalls werden als Software-Firewall bezeichnet, die aber auch auf einem PC ausgeführt werden müssen, also Hardware benötigen.[22]

Üblicherweise wird ein Gerät erst dann eine Hardware-Firewall genannt, wenn es sich um ein spezifisches Produkt für genau diesen Zweck handelt - also ein Gerät, welches mit mehreren Netzwerk-Schnittstellen und einer darauf laufenden Software zusammen, als eine Firewall dient.[22]

Die Softwarekomponente der Firewall arbeitet auf den Schichten 2 bis 7 des OSI Referenzmodells. Das Implementationsniveau kann demzufolge sehr unterschiedlich ausfallen, was begründet, warum eine Firewall oft nicht nur aus einer sondern eben aus verschiedenen Softwarekomponenten besteht.[22]

Zum Schluss soll die trockene Materie an einem repräsentativen Beispiel kurz verdeutlicht werden: Eine Firma, die ihre Arbeitsplatzrechner ins Internet bringen will, entscheidet sich für das Sicherheits-Konzept Firewall. Eine Firewall soll als Netzwerkinterface dienen, um die Gefahr durch Würmer und/oder Viren abzuwenden, und dafür sorgen, dass nur Verbindungen zu einem Mail-Server aufgebaut werden können. Damit für die Mitarbeiter

auch eine Recherche im Internet möglich ist, soll ein PC den Zugriff auf Webseiten über einen Proxyseiten erhalten. Der Surf-Rechner wird zusätzlich noch dadurch geschützt, dass die potentiell gefährlichen Elemente, wie beispielsweise Active X, aus den angeforderten HTML-Seiten einfach herausgefiltert werden. Sonstige Zugriffe von außen auf das Firmennetz werden abschliessend aus Sicherheitsgründen konsequent abgeblockt.[23]

Das Sicherheitskonzept Firewall kann also, wie es als ein erstes technisches Schutzkonzept in den Abschnitten dieses Kapitels beschrieben wurde, in seiner konkreten Realisierung den Datenverkehr zwischen einem lokalen Netzwerk einer Unternehmung und dem Internet zu protokollieren und damit kontrollieren. Wie eben im Beispiel beschrieben können unter anderem gefährliche Elemente wie Aktive X via Firewall den angeforderten Webseiten herausgefiltert werden. Somit kann sich eine Unternehmung durch die Umsetzung des technischen Schutzkonzeptes Firewall vor Bedrohungen in Form von Malware, Betrügereien und/oder Spionage proaktiv schützen. Im anschliessenden Kapitel wird noch ein weiteres technisches Schutzkonzept, das von Intrusion Detection Systemen, dargestellt.

Intrusion Detection Systeme [24]

„Ein Intrusion Detection System (IDS) ist ein Programm, das der Erkennung von Angriffen auf ein Computersystem oder Computernetz dient.“ [24] Wie der Name bereits vermuten lässt, richtet sich ein IDS auf die Erkennung von „Einbrüchen“ (nicht physischen sondern eben auf Computern oder in Netzwerken). Richtig eingesetzt, ergänzen sich die beiden Schutzkonzepte Firewall und IDS und erhöhen die Sicherheit von Netzwerken gemeinsam. Es können grundsätzlich drei IDS-Arten unterschieden werden: erstens hostbasierte IDS (HIDS), zweitens netzwerkbasierte IDS (NIDS) und drittens hybride IDS (HIDS).

Ein IDS benutzt für die Einbruchserkennung grundsätzlich zwei Verfahren: das erste beinhaltet den Signatur-Vergleich mit bekannten Angriffsmustern. Das zweite Verfahren wendet zur Erkennung eine statistische Analyse an. Die meisten IDSs arbeiten jedoch mit Filtern und Signaturen, welche spezifische Angriffsmuster beschreiben. Der Nachteil dieses Vorgehens ist, dass dabei nur bereits bekannte Angriffe erkannt werden können.[25]

Der IDS-Erkennungs-Prozess wird in drei Schritte unterteilt: Zunächst erfolgt die Wahrnehmung eines IDS durch Sensoren, die entweder (bei einem HIDS) Logdaten oder Daten des Netzwerkverkehrs (bei einem NIDS) sammeln. Während der Mustererkennung überprüft und verarbeitet das IDs die gesammelten Daten und vergleicht sie mit Signaturen aus der Musterdatenbank. Falls Ereignisse auf eines der bekannten Muster zutreffen, wird ein sogenannter „Intrusion Alert“ (Einbruchs-Alarm) ausgelöst. Dieser kann je nach Konfiguration vielfältiger Natur sein: Dabei kann es sich beim Alarm lediglich um eine E-Mail oder SMS handeln, welche dem Administrator zugestellt wird oder, je nach Funktionsumfang, kann sogar eine Sperrung oder Isolierung des vermeintlichen Eindringlings innerhalb des Netzwerkes erfolgen.[25]

Die HIDS, als hostbasierte IDS, stellen die älteste Art von IDS dar. Sie wurden ursprünglich vom Militär entwickelt, um die Sicherheit von Großrechnern zu garantieren. Ein HIDS muss im Gegensatz zum NIDS auf jedem zu überwachenden System installiert werden.

Im HIDS-Kontext ist mit „Host“ jedes System, auf dem ein IDS installiert wurde, gemeint. Ein HIDS sollte also als Vorrausetzung das Betriebssystem der zu überwachenden Systeme unterstützen. Seine Informationen erhält es aus Log-Dateien, Kernel-Daten und anderen Systemdaten wie etwa der Registry. Alarm wird geschlagen, sobald in den überwachten Daten ein vermeintlicher Angriff erkannt wurde. Bei einer HIDS-Unterart, den sogenannten „System Integrity Verifiers“, beispielsweise wird mit Hilfe von Prüfsummen bestimmen, ob System-Veränderungen vorgenommen wurden.

Die NIDS, Netzwerkbasierten Intrustion Detection Systeme, versuchen insbesondere, alle Pakete in einem Netzwerk aufzuzeichnen, diese zu analysieren und verdächtige Aktivitäten zu melden. Ausserdem versuchen diese Systeme Angriffsmuster aus dem Netzwerkverkehr zu erkennen. Weil in der heutigen Zeit überwiegend das TCP/IP-Protokoll eingesetzt wird, muss auch ein potentieller Angriff mit grösster Wahrscheinlichkeit über dieses Protokoll erfolgen. Für ein HIDS bedeutet das insbesondere, dass mit nur einem Sensor ein ganzes Netzsegment effizient und effektiv überwacht werden kann. Als einziges Handycap bildet jedoch die Datenmenge eines modernen 1 GBit-LANs, welche die Bandbreite des Sensors übersteigen kann. Falls dies auftritt, müssen vom einem NIDS Pakete verworfen werden. Eine lückenlose Überwachung ist bei einem solchen Vorkommnis nicht mehr garantiert.[26]

Als letzte sind hier noch die hybriden IDS, HIDS, zu erwähnen. Sie verbinden, um eine höhere Abdeckung bei der Erkennung von aufgetretenen Angriffen gewährleisten zu können, beide möglichen Prinzipien. In diesem Zusammenhang spricht man von netz- und hostbasierten Sensortypen. Sie sind an ein zentrales Managementsystem angeschlossen, und viele der heute eingesetzten IDS verfügen über eine solche, hybride Funktionsweise.

Zusammenfassend dienen IDS in ihrer vielfältigen Natur der Einbruchserkennung in Systemen und/oder Netzwerken. Ein IDS kann richtig eingesetzt eine Firewall als weiteres technisches Schutzkonzepte ergänzen und die Sicherheit von Netzwerken, was Verluste, Betrügereien, Malware und/oder Industriespionage angeht, für ein Unternehmen noch zusätzlich erhöhen.

Im Anschluss wird ergänzend zu Firewalls und den Intrusion Detection Systemen, technisches Schutzkonzept für Netzwerke eingegangen.

Honeypots [27]

„Als ein Honeypot (zu Deutsch: Honigtopf) wird ein Dienst bezeichnet, der die Aufgabe hat, Angriffe auf ein Netzwerk zu protokollieren. Dieser Dienst kann ein Programm sein, das einen oder mehrere Dienste zur Verfügung stellt, oder ein Server.“ Das Schutzkonzept Honeypots dient der Überwachung eines Netzwerkes. Dabei ist die Grundidee, in einem Netzwerk als „Köder“ für potentielle Angreifer einen oder mehrere Honeypots zu installieren, die einem legitimen Netznutzer unbekannt sind, und von einem solchen daher auch niemals angesprochen werden. Ein Hacker und Angreifer, der nicht zwischen echten Servern/Programmen und Honeypots unterscheiden kann und routinemässig alle Netzkomponenten auf Schwachstellen untersucht, wird früher oder später die von einem Honeypot angebotenen Dienste in Anspruch nehmen. Dabei wird er aber vom Honeypot protokolliert - das heisst „er bleibt am Topf kleben“ indem er eine Spur hinterlässt. Die

bloße Tatsache, dass irgendjemand versucht, mit einem Honeypot zu kommunizieren, wird dabei bereits als ein potentieller Angriff betrachtet.

Mittels eines Honeypot-Programmes werden übliche Netzwerkdienste wie Mail-, Datei-Server u.a. eines einzelnen Rechners oder sogar ein vollständiges Netzwerk simuliert. Falls je ein unberechtigter Zugriff auf einen derartigen virtuellen Dienst erfolgt, werden alle ausgeführten Aktionen umgehend protokolliert und gegebenenfalls wird ein Alarm ausgelöst.

Als unterschiedliche Honeypot-Typen gilt es generell zwischen low interaction und high interaction Honeypots zu unterscheiden. Ebenfalls gehören noch in die Kategorie der Honeypots sogenannte Tarpits (engl. Teergruben).

Ein Low-Interaction Honeypot ist meist ein Programm, welches einen oder mehrere Dienste emuliert. Weil die Fähigkeiten eines Low-Interaction-Honeypots beschränkter sind als die eines High-Interaction Honeypots, hat ein versierter Angreifer weniger Probleme, ihn zu erkennen. Um automatisierte Angriffe von Würmern wie zum Beispiel Sasser zu protokollieren, reicht ein Low-Interaction Honeypot jedoch vollständig aus.[28]

High-Interaction Honeypots bieten zumeist als vollständige Server ihre Dienste an. Der Fokus bei einem High-Interaction Honeypot liegt weniger auf der Beobachtung und Protokollierung von automatisierten Angriffen als auf manuell ausgeführten Angriffen. Dadurch sollen neue Methoden der Angreifer rechtzeitig erkannt werden. Optimalerweise stellt der „Köder“ eines High-Interaction Honeypots ein high value target dar, also einen Server, dem von potentiellen Angreifern ein hoher Wert nachgesagt wird. Zur Überwachung eines High-Interaction Honeypots wird eine spezielle Software eingesetzt. Meistens ist dies das frei verfügbare Sebek, das vom Kernelspace aus alle Programme des Userspace überwacht und die anfallenden Daten vom Kernelspace aus an einen loggenden Server sendet. Grösstes Ziel von Sebek ist es, unerkannt zu bleiben. Ein Angreifer soll davon nichts wissen oder ahnen. Er soll nichts ändern können, ohne dass er dabei überwacht wird.[28]

„Tarpits“, dienen dazu die Verbreitungsgeschwindigkeit zum Beispiel von Würmern zu verringern. Das Verfahren ist auch unter dem Namen LaBrea bekannt. LaBrea ist im IT-Bereich eine „Teergrube“, d.h. ein Programm, mit dem man virtuelle Netzwerke vortäuschen und so z.B. Internetwürmer festhalten und/oder Netzwerkscans blockieren kann. Ebenso gibt es aber auch Teergruben die offene Proxy Server emulieren, und falls jemand versucht Spam über diesen Dienst zu verschicken, den Sender dadurch ausbremst, dass es nur sehr langsam die Daten überträgt.“ [27]

Abschliessend sind also Honeypots und Tarpits geeignete technische Schutzkonzepte zur Früherkennung von Angriffen, je nach Motiv verursacht durch Datenverluste, Betrügereien, Malware oder Industriespionage. Im kommenden Abschnitt gilt es noch, die drei hier vorgestellten Schutzkonzepte zu diskutieren.

Wirksamkeit Technischer Schutzkonzepte [19]

Bezüglich der Wirksamkeit sowohl von Firewall, Intrusion Detection Systems wie Honeypots sollte abschliessend und als Übergang zum nächsten Kapitel „Organisatorische und

Mitarbeiter bezogene Schutzkonzepte“ noch folgendes erwähnt werden: Die Wirksamkeit solcher Systeme ist in grossem Masse abhängig von der richtigen Konfiguration, einer ständigen Pflege wie Aktualisierung und der laufenden Überwachung der Systeme. Selbst gut gewartete Sicherheitssysteme der hier vorgestellten Art schützen jeweils nur gegen die Bedrohungen, gegen die sie entwickelt wurden. Ein grosser Anteil eines Spektrums potentieller Bedrohungen durch Verluste, Betrügereien, Malware oder Industriespionage bleibt jedoch beim Einsatz solcher Systeme erhalten, wenn nicht zusätzlich zum technischen Schutz organisatorische und mitarbeiterbezogene Schutzkonzepte umgesetzt werden. Um über diese eine bessere Vorstellung zu erhalten werden im anschliessenden Kapitel auch die Schutzkonzepte solcher Art noch vorgestellt.

3.4.3 Organisatorische und mitarbeiterbezogene Schutzkonzepte

In der Diskussion um die Sicherheit von Informationen und IT-Systemen stehen häufig Angriffe über Computernetze und die hierfür geeigneten Schutzmassnahmen Firewalls, Intrusion Detection Systeme oder Honeypots wie oben aufgeführt im Mittelpunkt. Die Aufmerksamkeit gilt dabei vor allem Hackerangriffen, Verunstaltungen von Webseiten und verteilte Denial-of-Service-Angriffen. Inbesondere jedoch bei eingeschleppter Malware, Viren, Würmern oder Trojaner als Systemplagen spielt nebst der technischen die menschliche Komponente doch eine entscheidende Rolle.

Der Mensch stellt für die IT-Sicherheit ein zentrales Risiko dar. Bei der Betrachtung der IT-Sicherheit wird dieses Element der Gefährdungslage oft nicht ausreichend berücksichtigt. Technische Schutz-Systeme erfüllen zwar eine wichtige Funktion, schützen aber jeweils nur vor ganz bestimmten, definierten Bedrohungen. Das in diesem Kapitel vorgestellte Security Awareness Konzept ergänzt sinnvollerweise die technischen Massnahmen.[29]

Weil die Kenntnis über Security Awareness im Gegensatz den z.B. unter bereits aufgeführten Technischen Schutzkonzepten oft weniger verbreitet zu sein scheint, wird dieses Konzept im kommenden Abschnitt ausführlich dargestellt. Zuerst wird auf den Sinngehalt und Zweck eines Security-Awareness-Programms eingegangen. Das darauf folgende Kapitel erklärt dessen wesentliche Inhalte. Ein Security-Programm hat verschieden Komponenten, worauf näher eingegangen wird. Abschliessend wird noch ein mögliches Vorgehensmodell für die Implementierung aufgezeigt.

Security Awareness

Mit dem Aufsetzen und der Einführung eines Security-Awareness-Programms können Unternehmen den Gefahren der Unkenntnis, des fehlenden Bewusstsein für Risiken und der Überrumpelung auf Mitarbeiterseite insbesondere dem Social Engineering entgegenwirken, indem sie

- allen betroffenen Mitarbeitern die erforderlichen Grundlagen und Kenntnisse der Informationssicherheit vermitteln,

- Mitarbeiter für Risiken bei der Informationsverarbeitung sensibilisieren, so dass sie Gefahrensituationen erkennen und vermeiden können,
- wichtige Verhaltensweisen trainieren, die die Wirksamkeit von Sicherheitsmassnahmen verbessern und dadurch Schadensfälle verhindern oder die Auswirkungen begrenzen

Dazu greift das Security-Awareness-Programm im Unternehmen vorhandene Regelungen und Richtlinien zur IT-Sicherheit auf und vermittelt die erforderlichen Inhalte durch ein Massnahmenbündel aus Schulungen, praktischen Demonstrationen und Übungen, verschiedenen Informations- , und Sensibilisierungsmaterialien (Flyer, Plakate, Rundmails, Bildschirmschoner, Intranet...) und Tests.[30]

Inhalte eines Security-Awareness-Programms [31]

Die Inhalte eines Security-Awareness-Programms lassen sich in Anlehnung an den NIST-Standard 800-50[8] in drei Punkten zusammenfassen:

- Die Vermittlung der erforderlichen Grundkenntnisse (Education) zur IT-Sicherheit bei allen Mitarbeitern, die bei ihrer Arbeit IT-Systeme einsetzen und damit sensible Daten verarbeiten oder bei der Erfüllung kritische Geschäftsprozesse mitwirken. Massnahmen aus dem Bereich Education zielen auf den Erwerb von Wissen.
- Das Einüben von Verhaltensweisen (Training) und damit verbunden das Frischhalten bzw. Auffrischen des erworbenen Wissens. Massnahmen aus dem Bereich Training zielen auf den Erwerb von Fertigkeiten.
- Die Sensibilisierung (Awareness) für den Wert von Daten und Informationen und die Ausbildung eines „offenen Auges“ für verdächtige Vorfälle und Anzeichen von Gefährdungen der IT-Sicherheit. Massnahmen aus dem Bereich Awareness zielen auf die Förderung der Wahrnehmung der IT-Sicherheit durch den Mitarbeiter.

Die konkrete Ausgestaltung der im Rahmen des Programmes vermittelten Inhalte muss sich individuell nach der Organisation, der Risikosituation und den damit verbundenen Sicherheitsanforderungen richten. Neben der Vermittlung der Inhalte der Sicherheitsrichtlinien des Unternehmens müssen bei der Programmgestaltung auch aktuelle Themenstellungen berücksichtigt werden, damit auch den sich weiterentwickelnden Gefährdungen und Angriffsszenarien Rechnung getragen wird. Mögliche Themenfelder für ein Security-Awareness-Programm umfassen:

- Die Vermittlung der Sicherheitsziele des Unternehmens und die Konsequenzen bei einer Kompromittierung der Vertraulichkeit Verfügbarkeit, Integrität oder Nachvollziehbarkeit der Datenverarbeitung

[8]Das National Institute of Standards and Technology (NIST) publiziert regelmässige aktuelle Standards.

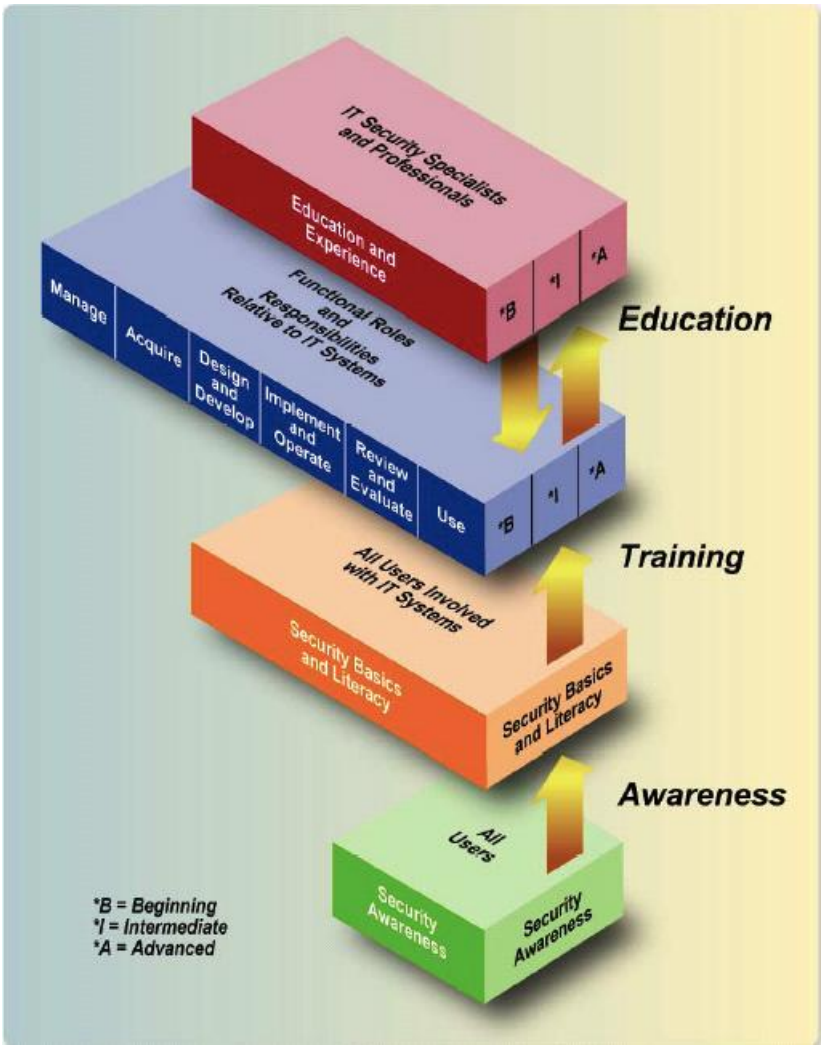


Abbildung 3.7: The IT Security Learning Continuum [31]

- die Bedeutung von Daten und IT-Systemen für den Geschäftszweck und den Fortbestand des Unternehmens und die Regelungen zur Klassifikation von Informationen und Dokumenten im Unternehmen;
- Regelungen zur Passwortsicherheit
- Viren, Würmer und andere Gefahren bei der Internet-Nutzung und Verhalten bei einem Virenbefall
- Risiken und Gefahren bei der Internet-Nutzung und Verhalten bei einem Virenbefall
- Risiken und Gefahren beim Einsatz von E-Mail und betriebsinterne Regelungen (z.B. zur privaten Nutzung von E-Mail, zum Weiterleiten, zur E-Mail-Archivierung etc.)
- Regeln zur Nutzung bzw. zum Verbot der Nutzung dienstlicher IT-Komponenten und -Infrastrukturen für private Zwecke, bzw. des Einsatzes privater Komponenten und Software für dienstliche Zwecke

Dabei sollte nicht versucht werden, diese Vielzahl von Themenstellungen auf einmal und alle auf demselben Wege zu vermitteln. Die Zielgruppe, die genauen Inhalte, die geeignete Art der Vermittlung und ggf. der Aktualisierungszyklus müssen für jeden einzelnen Themenblock vor dem Hintergrund des betrachteten Unternehmens und der Gefährdungslage bestimmt werden und führen so zusammen zu einem Bündel von Schulungs-, Trainings-, und Sensibilisierungsmassnahmen.[29]

Komponenten eines Security-Awareness-Programms [31]

Um die verschiedenen Inhalte des Programms richtig und nachhaltig zu vermitteln, empfiehlt sich die Gestaltung eines abgestimmten Massnahmenbündels. Das zentrale Element hierbei bilden kurze Schulungsveranstaltungen, die das jeweilige Thema möglichst anschaulich und praxisnah vermitteln und die vor allem auch Gelegenheit zu Fragen, Diskussion und eigener Beschäftigung der Teilnehmer mit der jeweiligen Themenstellung geben.

Diese Schulungsveranstaltungen sollten durch weitere Massnahmen flankiert werden, die das Vermittelte „wach halten“ bzw. aktualisieren. Solche zusätzlichen Massnahmen können z.B. umfassen:

- Definition eines „Einarbeitungspakets“ für neue Mitarbeiter, damit auch diese mit ihren IT-Sicherheitskenntnissen auf den Stand ihrer bereits geschulten Kollegen gebracht werden. Dabei sollte auch der neue Mitarbeiter bei der Einarbeitung die Möglichkeit zu Fragen und die Anregung zur eigenen Beschäftigung mit dem Thema erhalten. Dies kann z.B. erreicht werden, indem in den einzelnen Unternehmensbereichen einzelne Mitarbeiter als „Multiplikatoren“ weiter qualifiziert werden, damit diese mit erweiterten Kenntnissen als Ansprechspartner für neue und alte Kollegen fungieren können.

- Gestaltung eines Intranet-Bereichs zur IT-Sicherheit. Dieser Intranet-Bereich dient einerseits als Nachschlagemöglichkeit zu IT-Sicherheitsfragen (Hintergrundwissen, Ansprechpartner, Telefonnummern...), er kann andererseits auch genutzt werden, um neue und alte Informationen, die die Mitarbeiter des Unternehmens betreffen, zu veröffentlichen. Wichtig ist hierbei die ansprechende Gestaltung, die eine möglichst einfache Nutzung erlaubt und gleichzeitig die Lerninhalte möglichst ansprechend vermittelt. Hierfür können z.B. kurze Videosequenzen mit praktischen Demonstrationen oder animierte Darstellungen der Funktionsweise von IT-Sicherheitskomponenten eingesetzt werden.

- Bei aktuellen Themen oder Neuerungen und Veränderungen der Sicherheitskomponenten kann ein Artikel in der Mitarbeiterzeitschrift aufmerksam machen und die Hintergründe erklären. Hier können auch Berichte über die erfolgreiche Abwehr von Angriffen oder richtiges Verhalten von Mitarbeitern platziert werden. Neuigkeiten können auch über Rundmails oder elektronische Newsletter im Unternehmen verbreitet werden.

- Durch „Werbeträger“ wie Screensaver oder Poster können eingängige Verhaltensregeln und Sicherheitsgrundsätze im Unternehmen kommuniziert werden.

Die wirksamste Methode zur Ausprägung des Sicherheitsbewusstseins bleibt dabei die praktische Übung. Durch „gestellte“ Angriffsszenarien mit unbefugten Mitarbeitern durchgesprochen und ausgewertet werden, können Mitarbeiter auch praktisch so auf den Ernstfall vorbereitet werden, dass der Faktor Überrumpelung für den Angreifer entfällt. Solche Massnahmen müssen natürlich sorgfältig vorbereitet und mit den relevanten Stellen (Betriebsrat, Wachschutz etc.) abgestimmt werden.

Bei der Gestaltung des richtigen Massnahmenbündels sind neben didaktischen Erwägungen natürlich auch Kostenabwägungen durchzuführen, so dass das für die Gesamtmassnahme verfügbare Budget möglichst zielgerichtet eingesetzt wird. Auch in dieser Hinsicht kann das oben angesprochene Multiplikatorenmodell dazu beitragen, dass Massnahmen in kleinerem Rahmen durchgeführt und die vermittelten Kenntnisse dann von den Multiplikatoren im Unternehmen weitergetragen werden.

Security Awareness Vorgehensmodell [31]

Wie die vorgehenden Abschnitte gezeigt haben, bestehen für ein Security-Awareness-Programm vielfältige Gestaltungsmöglichkeiten hinsichtlich der Inhalte, der Art der Vermittlung und des jeweils angesprochenen Personenkreises. Für die erfolgreiche Planung und Durchführung eines solchen Programms ist daher ein methodisches Vorgehen erforderlich, damit bei jedem Projektschritt die erforderlichen Einflussfaktoren adäquat berücksichtigt werden.

Ein solches methodisches Vorgehen lässt sich am besten in vier einzelne Projektphasen gliedern, die jeweils mit einem Meilenstein abgeschlossen werden sollten. Die erste Phase dient der Ausbildung eines Security-Awareness-Prozesses, damit der erreichte Stand langfristig gehalten werden kann. Hierfür werden Abläufe definiert, die einerseits die Auffrischung der Kenntnisse und Fähigkeiten innerhalb der Zielgruppe zum Ziel haben (Etablierung regelmässiger Sicherheitskolumnen in Mitarbeiter-Zeitung oder Intranet, Gestaltung einer zyklisch wechselnden Posterreiche etc.), und die andererseits dafür sorgen, dass Fluktuationen in der Zielgruppe nicht zu einem Verfall des erreichten Ausbildungsstands führen (Einarbeitung und Ausbildung neuer Mitarbeiter).

Im Rahmen der vierten Phase sollte auch eine Erfolgskontrolle durchgeführt werden, in der Feedback aus der Zielgruppe gesammelt wird (Selbsteinschätzung der Teilnehmer) und Anregungen für die Fortführung des Security-Awareness-Prozesses aufgenommen und bewertet werden.[29]

3.4.4 Zusammenfassung Schutz, Prävention und Abwehr

Ein Risikokonzept als ganzheitliches Konzept basiert auf einer soliden Gefahrenanalyse, welches hinübergeleitet wird in Massnahmen, deren Schutz und Wirkung kontinuierlich

neu-überprüft und verbessert wird. Bei der Ermittlung der individuellen Gefährdungslage gelten Technik, höhere Gewalt, Organisation und der Mensch als die Bedrohungsklassen. Technische Schutzkonzepte sind Firewalls, die im Optimalfall den Datenverkehr zwischen einem lokalen Netzewerk (LAN oder Intranet) und dem Internet kontrollieren, Intrusion Detection Systems zur Alarmierungs-Zwecken über Angriffe von aussen wie von innen und Honeypots, welche Fremdlinge protokollieren und festzukleben versuchen. Education, Training und Awareness gelten als die drei essentiellen Ebenen eines Security Awareness Programms, worin Mitarbeiter für einen alltäglichen Umgang im Hinblick auf Sicherheit kontinuierlich trainiert und begleitet werden.

3.5 Diskurs und Schlussfolgerung

Bei einer ganzheitlichen Betrachtung des Themas IT-Sicherheit im Unternehmen wird deutlich, dass nicht nur technische Einflussfaktoren und Sicherheitssysteme für das insgesamt erreichte Sicherheitsniveau bestimmend sind. Vielmehr stellen auch Faktoren wie Unkenntnis und fehlende Sensibilität der Mitarbeiter ein erhebliches Risikopotenzial dar, das in die Betrachtung mit einbezogen werden muss.

Ein umfassendes Management der Informationssicherheit im Unternehmen darf sich nicht ausschliesslich auf technische Angriffsszenarien und Sicherheitssysteme beschränken, sondern muss auch die Mitarbeiter als Risiko mit in die Betrachtung einbeziehen. Viele Angriffswege beruhen auf der Manipulation von Mitarbeitern oder beziehen diese als Element eines komplexen Angriffs mit ein. Dabei bilden die Mitarbeiter aber nicht nur ein Risiko, sondern auch eine Chance für das Sicherheitsmanagement: Technische Sicherheitssystem bieten meist nur Schutz gegen bestimmte, bekannte Arten von Angriffen. Ein aufmerksamer, sensibilisierter Mitarbeiter hingegen kann auch bei völlig neuen Angriffszenarien Verdacht schöpfen und durch die Benachrichtigung der Sicherheitsverantwortlichen im Unternehmen helfen, Schaden abzuwenden und Beweismaterial zu sichern.

Durch die aktive Schulung und Vorbereitung von Mitarbeitern auf Sicherheitsvorfälle kann nicht nur dazu beigetragen werden, Fehlverhalten zu vermeiden und dadurch Angriffe zu erschweren. Sensibilisierte und aufmerksame Mitarbeiter stellen hingegen selbst einen aktiven Bestandteil der Sicherheitsinfrastruktur dar, weil sie - anders als die meisten technischen Schutzvorkehrungen - auch völlig neuartige Angriffsszenarien erkennen können und durch Benachrichtigung der verantwortlichen Stellen eine frühzeitige Reaktion von seiten des Unternehmens fördern können.

Um die von den Mitarbeitern ausgehenden Risiken zu minimieren und die mit den Mitarbeitern verbundenen Chancen optimal zu nutzen, empfiehlt sich die systematische Planung und Durchführung eines unternehmensweiten Security-Awareness-Programms, das methodisch die erforderlichen Kenntnisse vermittelt, kritische Abläufe einübt und vor allem den Mitarbeitern ihre eigene Verantwortlichkeit für die Sicherheit des Gesamtunternehmens bewusst macht und so dazu beiträgt, eine gelebte Sicherheitskultur im Unternehmen auf- bzw. auszubauen.

Literaturverzeichnis

[1] Wikipedia: Script Kiddie, http://de.wikipedia.org/wiki/Script_Kiddie, 17.Oktober 2005.

[2] Financial Cryptography: Industrial Espionage using Trojan horses, http://financialcryptography.com/mt/archives/000487.html, 31. Mai 2005.

[3] Wikipedia: Industriespionage, http://de.wikipedia.org/wiki/Industriespionage, 29. Januar 2006.

[4] F-Secure Virus Descriptions: Sober.Y, http://www.f-secure.com/v-descs/sober_y.shtml, 29. Januar 2005.

[5] Microsoft Technet: Leitfaden zur erfolgreichen Virenabwehr, http://www.microsoft.com/germany/technet/datenbank/articles/900002.mspx, 29. Januar 2006.

[6] Bill McCarty: Automated Identity Theft, http://www.honeynet.org, 29. Januar 2006..

[7] Kathrin Schmitt: Fraud Management - vom Umgang mit IT-Betrug http://www.silicon.de/cpo/ts-cio/detail.php?nr=15783, August 2004.

[8] Statistik Internet Fraud Watch http://www.fraud.org/internet/internet_scams_halfyear_2005.pdf,31. Januar 2006.

[9] Nigeria-Connection: Was ist die Nigeria-Connection?, http://www.nigeria-connection.de/, 20. Februar 2005.

[10] Larry Bridwell und Lawrence M. Walsh: Inundated by Infections http://infosecuritymag.techtarget.com/2003/apr/virussurvey.shtml, April 2003.

[11] Duden Informatik, Dudenverlag, 2001, 3. Auflage

[12] Quinetiq: Analysis of the Impact of Open Source Software, http://www.govtalk.gov.uk/documents/QinetiQ_OSS_rep.pdf, Oktober 2001.

[13] Stefan Ernst, Hacker, Cracker und Computerviren, Schmidt (Otto), Köln

[14] Kefk Network: Computerviren, http://www.kefk.net/Security/Malware/Viren, 14. Januar 2006.

[15] Schwab, Die schönste: Sagen des klassischen Altertums, 2001.

[16] Earthlink, `http://www.earthlink.net/software/nmfree/spyaudit/about/`, 29. Januar 2006

[17] Pfitzmann/Federrath/Kuhn, Technischer Teil des DMMV-Gutachtens.

[18] Wikipedia: Social Engineering, `http://de.wikipedia.org/wiki/Social_Engineering`, 11. Januar 2006

[19] J. von Knop, H. Frank. Netzwerk- und Computersicherheit. W. Bertelsmann Verlag, 2004.

[20] Studie IT-Security 2004: Im Wettlauf mit der Lernfähigkeit der Hacker, `http://www.silicon.de/cpo/studien/detail.php?nr=14998`, 31. Januar 2006.

[21] Wikipedia: Firewall, `http://de.wikipedia.org/wiki/Firewall`, 31. Januar 2006.

[22] W.Barth, Das Firewall-Buch. Grundlagen, Aufbau und Betrieb sicherer Netzwerke mit Linux. 2., überarbeitete Auflage, SuSE Press, 2003.

[23] P. Wouters, Designing a Safe Network Using Firewalls. Linux Journal, August 1997.

[24] Wikipedia: Intrusion Detection Systeme, `http://de.wikipedia.org/wiki/Intrusion_Detection_System`, 19.Oktober 2005.

[25] L. Zhuowei, A. Das, and Z. Jianying, Theoretical basis for intrusion detection Proceedings from the Sixth Annual IEEE, 15-17 Juni 2005.

[26] S. Northcutt, J. Novak, Network Intrusion Detection. 1.Auflage, Hüthig, 2004.

[27] Wikipedia: Honeypot, `http://de.wikipedia.org/wiki/Honeypot`, 25.Oktober 2005.

[28] The Honeynet Project: Know Your Enemy Whitepapers, `http://www.honeynet.org/papers/kye.html`, 29. Januar 2005.

[29] M. T. Siponen, Five Dimensions of Information Security Awareness. Computers and Society, Juni 2001.

[30] C. McCoy, and R. Thurmond Fowler, You Are the Key to Security: Establishing a Successful Security Awareness Program. ACM, Oktober 2004.

[31] M. Wilson, and J. Hash, Building an Information Technology Security Awareness and Training Program. Special Publication 800-50 of National Institute of Standards and Technologies, Oktober 2003.

Kapitel 4

The Business Model: Open Source

Christine Richartz, Bettina Koch, Roman Wieser

Open Source Software, heute sicherlich schon jedem einen Begriff, schaffte in den letzten Jahren den Sprung vom Computerspezialist ins Wohnzimmer und in viele Bereiche der Industrie, sprich in viele Firmen und Unternehmen. Wo liegen aber die Vorteile nicht proprietäre Software mit Lizenzen und zuverlässigem Support herzustellen, sondern hingegen auf gratis Software umzusteigen und diese dann zum Downloaden zur Verfügung zu stellen. Heute gibt es schon zahlreiche Unternehmen, die auf Open Source Software setzen und damit auch sehr erfolgreich sind. Gibt es dabei nicht eine Reihe von Problemen, da jeder ein wenig daran programmieren kann? Ist diese Software sicher, da der Code frei erhältlich und für jeden zugänglich ist? Welche Geschäftsmodelle werden praktiziert um erfolgreich im Geschäft zu bleiben, auch ohne Lizenzeinnahmen? Welche Strategien werden dabei verfolgt und angewendet?
In dieser Arbeit wird auf die Vor- und Nachteile von Open Source Software eingegangen und mit Closed Source Software verglichen. Es wird abgewogen, wann welches Modell besser geeignet ist. Schliesslich werden einige heute praktizierte Geschäftsmodelle vorgestellt und anhand von Beispielen erläutert.

Inhaltsverzeichnis

4.1 Einleitung

4.1.1 Ziel und Gliederung der Arbeit

Die vorliegende Arbeit stellt verschiedene Geschäftsmodelle im Rahmen von Open Source Software vor. Sie soll ausserdem verschiedene Parallelen und Unterschiede zwischen diesen Modellen aufzeigen. Um diese Thematik richtig behandeln zu können, empfiehlt es sich vorab einige Begriffsdefinitionen zu klären und auch eine klare Abgrenzung zwischen dem Open Source und dem Closed Source Business vorzunehmen.

Es ist nicht Ziel dieser Arbeit die Eigenschaften wie Vor- und Nachteile, die Anwender beim Benutzen einer Software erfahren zu identifizieren und genauer zu beschreiben. Die Ausarbeitung orientiert sich thematisch zum grössten Teil an der Sicht der Anbieter von Closed Source und Open Source Software. Sie soll deren Geschäftsideen, die Art und Weise wie ihr Produkt entsteht und ihre Bestrebungen auf unterschiedliche Art und Weise Gewinne zu erstreben, darlegen.

Das methodische Vorgehen, das dieser Arbeit zu Grunde liegt, basiert auf Recherchen von Fachliteratur, Onlinequellen und Zeitungsartikel, sowie der kritischen Auseinandersetzung mit den jeweiligen Quellen.

Wie bereits erwähnt, beginnt diese Arbeit im Kapitel eins mit den wichtigsten Begriffsdefinitionen, konzentriert sich dann in Kapitel zwei auf die beiden sehr unterschiedlichen Geschäftsmodelle von Closed und Open Source, betrachtet darin auch deren wichtigste Vor- und Nachteile und erklärt eine Methode zu herausfinden, wann das Closed oder Open Source Geschäftsmodell eingesetzt werden soll. Im Kapitel drei folgen die sieben wichtigsten Geschäftsstrategien für Unternehmen, die auf freien und Open Source Software aufbauen. Diese verschiedenen Strategien werden anhand von Beispielen genauer erklärt und es wird ein Überblick darüber geschaffen, wie auch mit Open Source Geld verdient werden kann. In Kapitel vier folgen darauf praktische Beispiele von erfolgreichen Projekten aus der berühmten LAMP-Architektur, welche Linux, Apache, MySQL und PHP beinhalten. Abschliessend werden im Kapitel 5 die Schlussfolgerungen der Arbeit zusammengefasst. Im Anhang der Arbeit werden noch die wichtigsten Lizenzen des Open Source bzw. der freien Software miteinander verglichen.

4.1.2 Wichtige Definitionen

Um richtig in dieses Thema einsteigen zu können, müssen vorab einige Begriffe genau abgegrenzt werden. In diesem Teil folgen deshalb die Definitionen von „Closed Source Software“, „Freie Software“ und „Open Source Software“.

Closed Source Software

Closed Source Software sind Programme, bei denen der Quellcode nicht zur Verfügung steht. Der Programmcode wird vor der Herausgabe kompiliert, damit der Einblick durch

Dritte möglichst erschwert wird. Die Programme werden meist in binären Formaten verteilt. Wie schon erwähnt, bliebt der Quellcode im Besitz der Firma bzw. der Entwickler. Diese verkaufen dann meist äusserst restriktive Lizenzen. Beispiele solcher Lizenzenbestimmungen sind:

- Die Software darf nicht ausgeliehen / weiterverkauft werden.
- Die Software darf nicht analysiert und erweitert bzw. mutiert werden.
- Es dürfen keine Fehler behoben und Korrekturen weitergegeben werden.
- Die Software ist an bestimmte Rechner oder Rechnerkonfigurationen gebunden. [1]

Free Software

„Freie Software“ ist Software, die frei für viele Zwecke verwendet werden darf. Es ist darauf zu achten, dass man Freie Software nicht mit Freeware verwechselt, das „Frei“ im Namen bezieht sich auf die gegebenen Freiheiten und nicht auf die Tatsache, dass die Software kostenlos ist. Der Quellcode bei freier Software ist frei zugänglich und kann benutzt, studiert, den eigenen Bedürfnissen angepasst, verändert, kopiert und verteilt werden. Falls Verbesserungen vorgenommen werden, müssen diese unter den gleichen Bedingungen zur Verfügung gestellt werden. Der Begriff Freie Software bezieht sich laut „gnu“ auf genau 4 Freiheiten [2]:

- Die Freiheit das Programm für jeden Zweck zu benutzen.
- Die Freiheit, zu verstehen, wie das Programm funktioniert und wie man es für seine Ansprüche anpassen kann.
- Die Freiheit Kopien weiterzuverbreiten, so dass man seinem nächsten weiterhelfen kann.
- Die Freiheit, das Programm zu verbessern und die Verbesserungen der Öffentlichkeit zur Verfügung zu stellen, damit die ganze Gemeinschaft davon profitieren kann.

Open Source Software

„Quelloffen“ („open source“) bedeutet nicht nur freien Zugang zum Quellcode. Bei quelloffener Software müssen die Lizenzbestimmungen in Bezug auf die Weitergabe der Software noch mehreren Kriterien entsprechen. Die Organisation „Open Source Initiative“ hat folgende Definition des Begriffes „Open Source“ erstellt [3]:

1. **Freie Weitergabe:** Die Lizenz darf niemanden in seinem Recht einschränken, die Software als Teil eines Software-Paketes, das Programme unterschiedlichen Ursprungs enthält, zu verschenken oder zu verkaufen. Die Lizenz darf für den Fall eines solchen Verkaufs keine Lizenz- oder sonstigen Gebühren festschreiben.

2. **Quellcode:** Das Programm muss den Quellcode beinhalten. Die Weitergabe muss sowohl für den Quellcode als auch für die kompilierte Form zulässig sein. Wenn das Programm in irgendeiner Form ohne Quellcode weitergegeben wird, so muss es eine allgemein bekannte Möglichkeit geben, den Quellcode zum Selbstkostenpreis zu bekommen, vorzugsweise als gebührenfreien Download aus dem Internet. Der Quellcode soll die Form eines Programms sein, die ein Programmierer vorzugsweise bearbeitet. Absichtlich unverständlich geschriebener Quellcode ist daher nicht zulässig. Zwischenformen des Codes, so wie sie etwa ein Präprozessor oder ein Konverter („Translator“) erzeugt, sind unzulässig.

3. **Abgeleitete Software:** Die Lizenz muss Veränderungen und Devirate zulassen. Ausserdem muss sie es zulassen, dass solcherart entstandene Programme unter denselben Lizenzbestimmungen weitervertrieben werden können wie die Ausgangssoftware.

4. **Unversehrtheit des Quellcodes des Autors:** Die Lizenz darf die Möglichkeit, den Quellcode in veränderter Form weiterzugeben, nur dann einschränken, wenn sie vorsieht, dass zusammen mit dem Quellcode so genannte „Patch files“ weitergegeben werden dürfen, die den Programmcode bei der Kompilierung verändern. Die Lizenz muss die Weitergabe von Software, die aus verändertem Quellcode entstanden ist, ausdrücklich erlauben. Die Lizenz kann verlangen, dass die abgeleiteten Programme einen anderen Namen oder eine andere Versionsnummer als die Ausgangssoftware tragen.

5. **Keine Diskriminierung:** Die Lizenz darf niemanden benachteiligen.

6. **Keine Einschränkung bezüglich des Einsatzfeldes:** Die Lizenz darf niemanden daran hindern das Programm in einem bestimmten Bereich einzusetzen. Beispielsweise darf sie den Einsatz des Programms in einem Geschäft oder in der Genforschung nicht ausschliessen.

7. **Weitergabe der Lizenz:** Die Rechte an einem Programm müssen auf alle Personen übergehen, die diese Software erhalten, ohne dass für diese die Notwendigkeit bestünde, eine eigene, zusätzliche Lizenz zu erwerben.

8. **Die Lizenz darf nicht auf ein bestimmtes Produktpaket beschränkt sein:** Die Rechte an dem Programm dürfen nicht davon abhängig sein, ob das Programm Teil eines bestimmten Software-Paketes ist. Wenn das Programm aus dem Paket herausgenommen und im Rahmen der zu diesem Programm gehörenden Lizenz benutzt oder weitergegeben wird, so sollen alle Personen, die dieses Programm dann erhalten, alle Rechte daran haben, die auch in Verbindung mit dem ursprünglichen Software-Paket gewährt wurden.

9. **Die Lizenz darf die Weitergabe zusammen mit anderer Software nicht einschränken:** Die Lizenz darf keine Einschränkungen enthalten bezüglich anderer Software, die zusammen mit der lizenzierten Software weitergegeben wird. So darf die Lizenz z. B. nicht verlangen, dass alle anderen Programme, die auf dem gleichen Medium weitergegeben werden, auch quelloffen sein müssen.

10. **Die Lizenz muss technologie-neutral sein:** Keine Bestimmung der Lizenz oder deren Erfüllung darf an eine bestimmte Technologie oder die Verwendung von bestimmten Schnittstellen oder Oberflächen gebunden sein.

4.2 Closed Source Geschäftsmodelle vs. Open Source

Warum gibt es eigentlich diese zwei völlig unterschiedlichen Geschäftsmodelle? In diesem Kapitel werden das Open Source bzw. Closed Source Geschäftsmodell und ihre Vor- und Nachteile verglichen. Die Betrachtung bezieht sich auf die Seite der Hersteller und nicht auf die Benutzer der Programme.

4.2.1 Das Closed Source Geschäftsmodell

Der Grundgedanke des Closed Source Geschäftsmodells ist mit den meisten anderen Branchen zu vergleichen: ein Produkt wird entwickelt, hergestellt und verkauft. Bei diesem Geschäftsmodell wird der Softwareerstellungsprozess als herkömmlicher Warenerzeugungsprozess angesehen. Wie diese Softwareentwicklung genau aussieht und welche Vor- bzw. Nachteile das Modell mit sich zieht, wird nun genauer erläutert.

Die Softwareentwicklung im Closed Source Geschäftsmodell

Die kommerzielle Softwareentwicklung ist sehr marktorientiert. Vor der Erstellung eines neuen Programms wird meist eine Marktanalyse erstellt, um sicher zu gehen, dass das Produkt, nach der Realisierung auch genügend Abnehmer findet. In dieser Marktanalyse werden Fragen bezüglich der Produkteigenschaften, der voraussichtlichen Konkurrenten oder auch Kundenwünsche analysiert. Ziel ist es, möglichst eine Marktlücke zu finden und sich in diesem Gebiet dann erfolgreich positionieren zu können. [4]

Als zweiter Schritt nach dieser Analyse folgt das eigentliche Projekt. Der Projektplan wird erstellt, danach kommt der erste Entwurf zur Ausführung. Als Kern des Softwareentwicklungsprozesses wird darauf die Implementierung vorgenommen und das ganze in einer Systemumgebung getestet. Wenn alle Tests zufrieden stellend verlaufen sind, kann das Produkt auf den Markt gebracht werden und es ist nur noch für angemessenen Support zu sorgen.

Der ganze eben erwähnte Ablauf des Projektes läuft sehr isoliert ab und erfolgt in einer hierarchisch strukturierten Organisation. Raymond, ein Autor und Programmierer in der Open Source Szene, nennt dieses Geschäftsmodell auch das „Kathedralen-Modell“ [5], weil ein kleines Grüppchen von Leuten in Abgeschiedenheit etwas vollständig erstellt, wie beim Bau einer Kathedrale. Konkret auf die Softwareentwicklung bezogen gibt es hier ein Projektleiter und je nach Umfang des Projektes ihm unterstellte Hauptverantwortliche für gewisse Gebiete, sowie deren Mitarbeiter. Die Mitarbeiter werden alle für ihre Arbeit bezahlt, haben aber wenig Einfluss darauf, was die Software alles beinhaltet. Sie erstellen

den von ihnen verlangte Code für die Software, testen diesen oder leisten Support-Dienste. Der Support fällt je nach Software mehr oder weniger gut aus. Bei einer Standardsoftware wie beispielsweise Windows Office beschränkt sich der Support auf Hilfefunktionen und Updates. Bei einer Individualsoftware beinhaltet er je nach Supportvertrag mehrere Hilfeleistungen. [4]

Vor- und Nachteile das Closed Source Geschäftsmodells

Wie schon bei der Definition von Closed Source Software erwähnt, wird der Quellcode von entwickelten Programmen beim Closed Source Geschäftsmodell geheim gehalten. Die Entwickler sehen ihr Produkt, als eine Art geistiges Eigentum, das sie für sich behalten möchten. Was gibt es eigentlich für Vorteile um dies zu tun? Der Quellcode wird deshalb nicht besser. Streng genommen gibt es für die Unternehmen zwei Gründe weshalb eine solche Strategie angewendet wird:

- Profitmaximierung
- Schutz des Wettbewerbvorteils

Durch den Verkauf von Lizenzen für die entwickelte Software lassen sich nicht nur die Entwicklungskosten abdecken, sondern auch zusätzliche Profite erwirtschaften. Dies macht natürlich den Hauptvorteil für die Unternehmen, die nach diesem Geschäftsmodell arbeiten aus.

Der zweite Grund lässt sich am besten anhand eines Beispieles erklären: Ein Unternehmen veröffentlicht ein Softwarepaket unter Open Source. Diese Software wird nach einer gewissen Zeit sehr beliebt, da sie von den Benutzern verbessert wird. Daraufhin beginnen auch die Mitbewerber dieses Unternehmens diese Software zu nutzen, diese profitieren alle von diesen Vorteilen ohne vorher aber etwas investiert zu haben. Dies spürt das investierende Unternehmen durch einen Geschäftsrückgang. Das Risiko den Mitstreitern Vorteile zu verschaffen wollen natürlich viele Unternehmen nicht eingehen und entscheiden sich deshalb für eine Closed Source Software. [6]

Die Nachteile vom Closed Source Geschäftsmodell liegen sowohl in der Entwicklungsphase als auch in der Unterhaltungsphase, eventuell sogar im Vertrieb. Konkret sind dies folgende:

- Entwicklungskosten müssen alleine getragen werden
- Begrenzte Anzahl Mitwirkende
- Gleichwertige kostenlose Programme

In der Entwicklung gibt es zwei Nachteile für die Unternehmen. Erstens muss es alle anfallenden Kosten selbst tragen und zweitens hat es nur die eingestellten Mitarbeiter zur Verfügung. Diese erstellen die Anforderungen, die Software und erledigen das Testen.

Auch wenn die genannten Arbeiten von verschiedenen Personen erledigt werden, ist ein breiteres Spektrum an Mitwirkenden, wie es bei einem erfolgreichen Open Source Projekt der Fall ist, besser. Ebenso gilt dieses Argument für die Unterhaltungsphase, mehr Menschen finden mehr Fehler und haben auch mehr Verbesserungsvorschläge.

Die grössten Nachteile, die im Verkauf entstehen können, sind gleichwertige Produkte, die für den Anwender kostenlos erhältlich sind. Momentan ist es zwar noch nicht so, dass die breite Masse sich mit Open Source Software eindeckt, aber für die Zukunft ist dies ein mögliches Problem.

Zusätzlich zu diesen direkten Vor- bzw. Nachteilen für die Hersteller gibt es auch noch soziale und ökologische Sichtweisen, beispielsweise schafft das Closed Source Geschäftsmodell viele Arbeitsplätze und es stärkt die Wirtschaft. Wenn ein Unternehmen jedoch einen starken Aufschwung erlebt und eine dominierende Marktposition einnimmt, kann es andere, konkurrierende Firmen schwächen und möglicherweise Innovationen verhindern und somit sowohl in wirtschaftlicher als auch in technischer Hinsicht destruktiv wirken. [4]

4.2.2 Das Open Source Geschäftsmodell

Die Idee des Open Source Modells ist eine vollkommen andere, als diejenige des Closed Source Geschäftsmodells. Der entwickelte Quellcode wird hier nicht als die handelbare Ware angesehen, sondern das Geschäft wird mit den ganzen Notwendigkeiten, die nebenbei bei der Benutzung entstehen, gemacht. Wie diese Art der Software Entwicklung, die Linus Torvalds so populär gemacht hat [7], funktioniert und welche Vor- und Nachteile dieses Modell bringt, wird nun genauer untersucht.

Die Softwareentwicklung beim Open Source Geschäftsmodell

Die Softwareentwicklung bei Open Source Projekten verläuft anders, als bei der oben genannten Variante. Sie findet in einer offenen und locker organisierten Gemeinschaft statt. Für jedes Softwareprojekt schreiben viele Programmierer einen Anteil. Die neuen Erkenntnisse werden weitergegeben, übernommen, wieder verändert, usw. Dieses Nehmen und Geben wird auch der „Basar-Stil“ [5] genannt.

Der Auslöser für ein neues Open Source Projekt ist laut Eric Raymond „die persönliche Sehnsucht eines Entwicklers“ [5] Dies kann beispielsweise ein Problem sein, für die er nicht die passende Software hat, ein Mangel an einer bestehenden Software oder ein bereits gesehenes Produkt, dessen Lizenz er nicht kaufen möchte. Nachdem die Idee für ein Projekt entstanden ist, wird von diesem Entwickler (bzw. einem Open Source Unternehmen) zuerst ein grobes Design erstellt. Er entscheidet, ob er auf einer bestehenden Software aufbauen will, welche Teile er davon übernehmen bzw. weglassen möchte oder ob er bevorzugt ganz von vorn zu beginnen. Mit Hilfe von diesen Entscheidungen erstellt er dann eine Vorabversion seiner gewünschten Software. Sobald die erste Version entstanden ist und diese genügend Interessen von anderen Entwicklern und Anwendern weckt, kann

mit dem eigentlichen Projekt begonnen werden. An dem Projekt arbeiten dann verschiedene Personen an unterschiedlichen Orten auf der Welt. Dies erfordert natürlich eine gute Organisation bzw. Koordination, damit alles funktioniert, deshalb steht an der Spitze eines solchen Projekts meistens auch eine oder mehrere Führungsperson(en), welche aber nicht alleine den Werdegang einer Software bestimmt. Wie man sieht, ist in dieser Art der Softwareentwicklung nicht alles so voll durchgeplant, wie beim oben genannten Modell, es wird spontaner entwickelt und erstmal das gemacht, was gerade benötigt wird.

Neben der guten Organisation müssen noch weitere Vorraussetzungen für das gelingen des Open Source Projektes erfüllt sein. Die wichtigsten sind zum einen motivierte Mitglieder, denn eine noch so gute Idee kann nicht im grossen Ausmass ausgeführt werden, wenn niemand daran arbeitet und zum anderen eine gute Infrastruktur im Internet, die der Kommunikation und Präsentation dient.

Ein Projekt ist mit der ersten, vollständigen Version der Software nicht beendet, es steht danach unter ständiger Evaluation, die die Software an der vorhandenen Umgebung anpasst, somit hat die Software eine Chance anwendernah, innovativ und stabil zu bleiben. Diese Anpassungen werden wiederum von verschiedenen Entwicklern vorgenommen. [4]

Vorteile und Nachteile des Open Source Geschäftsmodells

Die zentrale Frage, die sich wahrscheinlich jeder anfangs im Zusammenhang mit diesem Geschäftsmodell stellt ist: Wie kann man mit einem freien Gut Geld verdienen? Es ist offensichtlich, dass mit Open Source nicht das schnelle Geld verdient wird, sondern die Taktik vielmehr auf ein mittle- bis langfristiges Geschäft ausgelagert werden muss. Die Möglichkeiten, wie mit Open Source Software Geld verdient werden kann, verteilen sich in der Softwareentstehung nach der Entwicklung. Es können Beratungen, Installation und Konfiguration, Wartung und Support, spezielle Lösungen und Anpassungen, sowie auch Schulungen angeboten werden, die sich sehr gut verkaufen lassen, da sie jedem Kunden individuell weiterhelfen. Diese Varianten werden aber im dritten Kapitel anhand der verschiedenen Open Source Geschäftsmodelle genauer erklärt.

Nach dem die wichtige Frage des Geldverdienens geklärt ist, folgen nun die Vorteile, die ein Unternehmen erfährt, wenn es dieses Geschäftsmodell verwendet. Diese sin im wesentlichen:

- geteilte Entwicklungskosten
- ev. Minimierung des Weiterentwicklungsrisikos
- Steigerung des Bekanntheitsgrades möglich

Der erste liegt in der Entwicklung. Bei der Entwicklung entstehen grosse Kosten, die ein Unternehmen, welches mit dem Open Source Geschäftsmodell arbeitet, nicht alleine tragen muss. Bei Open Source Projekten arbeiten sehr viele Entwickler mit, die ihre Zeit unentgeltlich zur Verfügung stellen bzw. von grossen Software Firmen angestellt sind um an Open Source Software Projekten mitzuwirken, somit fallen weniger Entwicklungskosten

für Programmiertätigkeiten und das Testen an. Diese Mitarbeit von vielen Programmieren trägt zudem auch dazu bei, dass ein sicheres ausgereiftes Produkt entsteht, da es eben von mehreren auf Fehler und Sicherheitslücken hin geprüft und darauf auch verbessert wird. Wie das Linus' Gesetz besagt: „Having enough eyes, all bugs seem shallow". Diese Tatsache führt zu vielen zufriedenen Anwendern, was zu einem grösseren Kundenpotential führt, dies wiederum schafft mehr Nachfrage für Dienstleistungen und den damit verbundenen Einnahmen. [1]

Ein anderer Vorteil den Quellcode freizugeben ist auch, dass das Weiterentwicklungsrisiko minimiert wird. Dieser Vorteil kann natürlich nur genutzt werden, wenn es auch Entwickler gibt, die sich für das Projekt interessieren. Vor allem bei Softwareprojekten, bei denen nur ein oder wenig Mitarbeiter mitgewirkt haben, besteht für ein Unternehmen das Risiko, dass sie die Kontrolle über ihre Software verlieren, wenn die Entwickler aus dem Unternehmen ausscheiden. Aufgrund schlechter Dokumentation oder komplexer Algorithmen ist es dann eventuell nicht mehr möglich, an einem Projekt eines anderen Verbesserungen und Ergänzungen vorzunehmen. Durch das „Freigeben" des Quellcodes an die Opensourcegemeinde lässt sich dieses Risiko minimieren, da viel mehr Programmierer daran arbeiten und somit das Wissen über die Software auf eine grössere Anzahl von Personen verteilt ist. [6]

Open Source Software kann einem Unternehmen helfen den Bekanntheitsgrad zu stärken und damit immense Werbeerfolge zu erzielen. Ein bekanntes Produkt wird immer mit einem bestimmten Logo assoziiert. Als Beispiel lässt sich hier Sun erwähnen, dass mit Java kein Geld verdient, aber jeder Mensch weiss trotzdem aus welchem Hause Java kommt. Dieser Effekt kann auch als Markterschliessungshilfe oder als Lockmittel eingesetzt werden um Kunden an die gewinnbringende proprietäre Produktpalette heranzuführen, die dann den Einkommensverlust wieder ausgleicht. [8]

Neben den genannten Vorteilen gibt es natürlich auch bei diesem Geschäftsmodell Nachteile. Diese sind hauptsächlich:

- keine vollständige Kontrolle über Produktentwicklung
- Abhängig von interessierten Programmierern und Entwicklern
- Entwickler von Closed Source Programmen können ebenfalls von Open Source Software lernen

Den wichtigsten Nachteil findet man in der Entwicklung des Produktes. Ein Unternehmen, das seine Software der Open Source Gemeinde zugänglich macht, hat nicht die hundertprozentige Kontrolle über die Produktentwicklung, es kann sich nicht alleine entscheiden, welcher Weg eingeschlagen wird. Die Mitarbeitenden Programmierer und Anwender haben ebenso Einfluss auf die Zukunft eines Projektes. Zudem steht es in einer Abhängigkeit von motivierten, befähigten Mitwirkenden. Falls diese Helfer nicht über die benötigten Kenntnisse verfügen, kann dies auch negative Auswirkungen haben. Es könnten schlechtere Versionen publiziert werden, was den Ruf eines Unternehmens nicht unbedingt positiv beeinflusst.

Ein weiterer negativer Effekt, der aus Open Source Software entstehen könnte, ist die Tatsache, dass auch Entwickler, die mit dem Closed Source Business Modell arbeiten, viel aus den freigegebenen Quellcodes lernen können und dieses Wissen verwenden um eine Closed Source Software zu entwickeln und vermarkten. Zwar ist es nicht erlaubt Teile von Open Source Software für diesen Zweck weiter zu verwenden, aber die Ideen oder Wissen daraus mitzunehmen, lässt sich wohl kaum verbieten.

Als eher nebensächlicher Nachteil lässt sich noch erwähnen, dass es in den Open Source Gemeinden Entwickler gibt, die sehr stark antikommerziell eingestellt sind und daher wirtschaftliche Erfolge von Open Source Software Unternehmen nicht akzeptieren wollen. Diese könnten bei einem zu erfolgreichen, vermarktbarem Projekt abspringen, weil sie die finanziellen Erfolge einer Firma nicht unterstützen wollen. [9]

Wann ist welches Geschäftsmodell zu empfehlen?

Wie im vorherigen Teil bei den Erläuterungen zum Open bzw. Closed Source Geschäftsmodell schon angedeutet wurde gibt es zwei wirtschaftliche Werte von Software: Zum einen den Warenwert und zum anderen den Gebrauchswert. Der Warenwert ist der Wert der Software als handelbare Ware und der Gebrauchswert ist definiert als Wert von Software im Sinne eines Werkzeugs.

Bevor ein Unternehmen die Auswahl des geeigneten Geschäftsmodells bei einem neuen Projekt trifft, muss es klar festgelegen durch welchen der beiden Werte es hauptsächlich sein Geld verdienen möchte, also welche Variante wirtschaftlich und persönlich am sinnvollsten ist.

Dabei gilt es natürlich nicht nur die gegenwärtige Situation zu prüfen, sondern auch Annahmen zu treffen, welche Vorteile die Zukunft bringen wird. Rein ökonomisch für die Gegenwart gesehen, wäre die Entscheidung leicht zu treffen, denn mit einem Closed Source Projekt lässt sich kurzfristig sicherlich den grösseren Gewinn erwirtschaften, da man das Produkt direkt verkaufen kann und somit gleich Geld einnimmt. Bei Open Source Software hingegen dauert es eine gewisse Zeit, bis ein Produkt sich etabliert und auch Dienstleistungen dazu verkauft werden können. Die Vorzüge der Open Source Methode sind deshalb schwieriger zu messen bzw. hervorzusehen.

Eine Methode von Eric Raymond um zu entscheiden ob ein Open Source Projekt begonnen werden soll, geht von den Fällen aus, in denen Open Source Software erfolg hatte, oder eben versagte. Durch diese Methode filtert er verschiedene Merkmale heraus, die einen Evolutionsdruck in Richtung Open Source erzeugen. Aufgrund dieser Merkmale müsste man laut Raymond jedem Unternehmen, dass die gestellten Fragen eher verneint zu einem geheim halten seines Quellcodes raten und denjenigen, die fast alle bis alle Punkte bejahen die Freigabe des Quellcodes empfehlen. Diese fünf zentralen Fragen, die bei einer Entscheidung genauer betrachtet werden müssen, sind [6]:

- Sind Zuverlässigkeit, Stabilität und Skalierbarkeit von ausschlaggebender Bedeutung?

- Kann die Korrektheit des Designs und der Implementation nicht leicht auf anderem Weg als dem der Peer Review überprüft werden?
- Ist die Software ist für den Benutzer geschäftskritisch?
- Ist die Software etabliert oder ermöglicht eine gemeinsame Computer- oder Kommunikationsinfrastruktur?
- Sind die Schlüsselmethoden (oder deren funktionale Äquivalente) Teil wohlbekannter Ingenieurskunst?

Raymond ist der Meinung, dass hohe Qualität und Zuverlässigkeit am besten durch die Bewertung des Produkts durch unabhängige Gutachter entsteht und er schwört deshalb auf die beim Open Source stattfindenden Peer Reviews. Durch diese werden Korrektheit von Design und Implementation am erfolgreichsten geprüft. Weiters überlegt er, dass ein Unternehmen (der Anwenderseite) Open Source bei einer zentralen Anwendung interessanter finden sollte um nicht in die Falle eines Herstellermonopols zu geraten. Diese betrachteten Merkmale führen zwar alle zu grösserem Vertrauen der Anwender, somit eventuell auch zu einem Zuwachs an Anwendern und mehreren Möglichkeiten um mit Open Source Software Geld zu verdienen. Ob dies aber die einzig richtigen Möglichkeiten sind, lässt sich jedoch bestreiten.

Die beiden letzten Kriterien, die bei Open Source zum grösseren Erfolg führen sollen, sind das Vorhandensein von Netzwerkeffekten und die Tatsache, dass Schlüsselmethoden oder funktionale Besonderheiten allgemein unter Programmierern bekannt sein müssen. Diese Merkmale sind die Grundlagen aller Open Source Projekte, ohne dessen Erfüllung muss gar nicht erst in Erwägung gezogen werden mit dem Open Source Geschäftsmodell zu arbeiten. Denn wenn niemand die Software benutzt, gibt es auch kein Bedarf an Dienstleistungen und somit keine Möglichkeit damit Geld zu verdienen.

Raymond erwähnt auch noch ein Beispiel, bei dem selbst er ein Closed Source Projekt empfehlen würde, dies ist jedoch ein absoluter Extremfall. Eine Firma entwickelt Software, die den Arbeitern eines Sägewerks ermöglichen die Schnittbilder für ihre Sägen computerisiert berechnen zu lassen, um die Länge der Bretter, die man aus einem Baumstamm gewinnen kann zu maximieren. Wenn man die Erfüllung der oben genannten Fragestellungen betrachtet, wird hauptsächlich die dritte Frage mit ja beantwortet. Diese Berechnungen lassen sich von einem erfahrenen Mitarbeiter aber auch händisch berechnen, deshalb rät hier selbst der Open Source Anhänger, das Closed Source Geschäftsmodell anzuwenden.

Wie schon mit etwas Kritik an einzelnen Punkten erwähnt, muss man sich bei dieser Methode bewusst sein, dass sie von einem Mitglied der Open Source Gemeinschaft entstanden ist. Es lassen sich auch auf anderen Wegen Stabilität und Qualität erzeugen, dies ist nur eine Frage des Aufwandes, den man dafür betreibt. Natürlich spielen bei einem Entscheidungsprozess ob ein geschlossenes oder offenes Geschäftsmodell verwendet wir auch die persönlichen Ziele, Merkmale und Philosophien eines Unternehmens eine grosse Rolle.

4.3 Ökonomie und Open Source Geschäftsmodelle

4.3.1 Die ökonomische Sicht von Open Source

> „Für mich ist es ein Rätsel, wie eine Software-Branche existieren kann, wenn Anwender Software per Open Source kostenlos erhalten“
> Microsoft-Manager Jim Gray [10]

Das Copyleft, was heisst ohne kommerzielle/kostenpflichtige Lizenzen anzubieten, selbst ist es, über das nachgedacht werden muss. Ohne einen Anreiz zur Produktion von freier Software, fällt das ganze System in sich zusammen (externe Effekte). Wo bleibt die extrinische und intrinische Motivation, der Reputationsgewinn und die kommerzielle Nutzbarmachung von Erfahrung durch Mitarbeiter in Open Source Projekten? Software ist kein gewöhnliches Gut mit gewöhnlichen Eigenschaften, dessen Nutzen kann erst nach dem Gerbrauch festgestellt werden. Der Markt läuft normalerweise aber genau umgekehrt: Geld gegen Ware. Da aber der Nutzen erst später einsehbar ist, entstehen Informationsasymmetrien, die zu einer falschen Selektion führen kann. [12]

Eigentlich ist der Open Source Software-Entwicklungsprozess frei von monetären Transaktionen und es können keine Exklusivrechte an den Bestandteilen geltend gemacht werden. Trotzdem spielt Open Source Software eine immer wichtigere Rolle im kommerziellen Umfeld. Ein Mehrwert entsteht durch den Einsatz im kommerziellen Umfeld.

Im Folgenden werden ein paar der wichtigsten Open Source Geschäftsmodelle vorgestellt und diskutiert. Grundsätzlich gibt es drei grundlegende Unterscheidungen bezüglich der Geschäftmodelle:

- Ware
- Marke
- Dienstleistung

Von diesen drei Kriterien geht schliesslich der Gewinn aus. Es gibt sehr viele Strategien im Open Source Bereich, zum Teil mit sehr kreativen und hochgradig konkurrierenden Marketing- und Dienstleistungsansätzen. So kann ein Open Source Projekt zum Beispiel einen neuen Industriestandard setzen. John Koenig schreibt im IT Manager's Journal von sieben grundlegenden Geschäftsstrategien auf die hier tiefer eingegangen wird. [14]

- **Die Optimierungsstrategie:** Modularität erlaubt Optimierung einzelner Schichten
- **Die Doppellizenz Strategie:** Freier Gebrauch von Software mit Einschränkungen
- **Die Consulting Strategie:** Lizenzkosten vermeiden und Service verkaufen
- **Die Abonnementen Strategie:** Open Source Wartung auf jährlicher Basis verkaufen

- **Die Gönnerstrategie:** Original Gerätehersteller unterstützen - OSS um eine Umgebung für andere Produkte und Dienste zu schaffen
- **Die Benutzer Strategie:** OSS benutzen um eigene Dienste zu verkaufen
- **Die Eingebettete Strategie:** OSS in Hardware benutzen um die Akzeptanz zu erhöhen

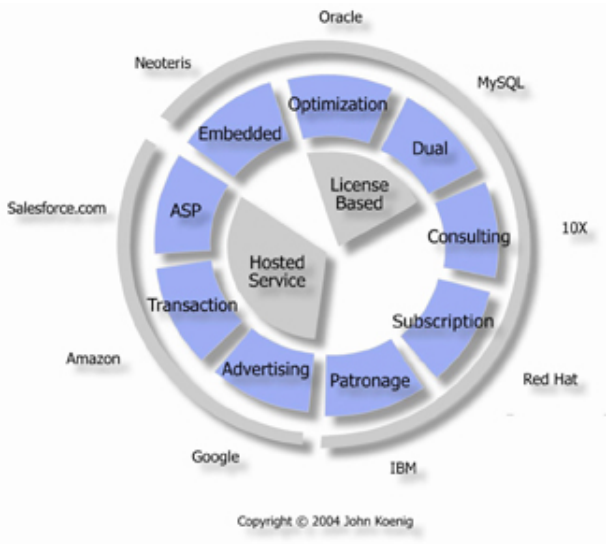


Abbildung 4.1: Strategien und Beispiele [13]

In Abbildung 4.1 sieht man diese Strategien und Firmen, die diese schon mit Erfolg anwenden. Bei der Optimierungs- un der Doppellizenz Strategien werden die Kosten mit Lizenzen, die weggelassen um dadurch Kosten zu sparen oder auch durch eine eingebaute kommerzielle Lizenz um die Vermarktung von Drittanbieter zu verhindern oder dadurch Gewinn zu erwirtschaften, gedeckt. In der Abbildung 4.1 tun dies zum Beispiel Oracle und MySQL. Weiter folgen die Consulting und die Abonnementen Strategie, in welchen der Gewinn hauptsächlich durch Dienstleistungen gemacht werden. Im linken unteren Drittel folgen dann eine Reihe von Firmen die mit der Benutzerstrategie arbeiten. Zum Schluss dieses Abschnittes wird noch auf die Eingebettete Strategie eingegangen, wie sie hier in der Abbildung 4.1 von Neoteris benuzt wird. [14]

4.3.2 Die Optimierungsstrategie

Die Optimierungsstrategie ist eine Abhandlung von Clayton Christensen's „Gesetz der Erhaltung der Modularität". Für den Open Source Ansatz bedeutet dies, dass die Software „modular und passend" sein sollte, so dass angrenzende Softwareschichten optimiert werden können. Dabei sind die einzelnen Schichten nur begrenzt profitabel, jedoch erzielen die daran anknüpfenden Komponenten eine bessere Performance. Linux ist dafür ein sehr gutes Beispiel. Die durch die Modularität eines Betriebssystems gewonnenen Schnittstellen dienen zur Verminderung der Gewinne anderer Hersteller wie Sun und Microsoft, die durch ihr komplexes System kaum Möglichkeiten zur Anpassung bieten. Die Gewinner

dieser Strategie sind dabei die unabhängigen Softwareschichten, da Anwendungen optimiert werden können und dabei ein grösserer Warenwert erzielt werden kann. Der Vorteil in diesem System ist, dass jede Unternehmung nun genau diese Softwareschicht optimieren kann, die für sie am wichtigsten ist. So entsteht ein wichtiges Produkt das höhere Gewinne erzielen kann als Komplettlösungen anderer Anbieter.

Ein gutes Beispiel dafür ist Electronic Arts, die schnelle und zuverlässige Server für ihr Spiel „Sims“ benötigte. Oracle hatte dabei die Linuxversion vom Oracle9i Real Application Server (RAC) vorgeschlagen. Portabilität war schon lange eine wichtig Voraussetzung, da es den Konsumenten nicht an eine Hardware bzw. an einen Betriebssystem Hersteller bindet. Für dieses Projekt baute Oracle eine Datenbanklösung für Standard-Linux und optimierte den Oracle RAC für Linux Clusters. Die Hardware für die Oracle UNIX (non-RAC) Lösung hätte Oracle 2 Millionen Dollar mehr gekostet ohne allerdings zu einer besseren Leistung führen zu können. Schliesslich lieferte Oracle die Linux RAC Lösung, die EA mehr als 1.3 Millionen Dollar sparen liess. [14]

4.3.3 Die Doppellizenz Strategie

Bei dieser Strategie bietet der Software Hersteller einerseits freie Benutzung der Software mit Einschränkungen, andererseits kommerzielle Lizenzen mit mehr Funktionalitäten gegen Gebühr. Freie Benutzung beinhaltet meist die Einschränkung, dass alle Veränderungen, die veröffentlicht werden, zusammen mit dem Source Code veröffentlicht werden müssen. Um Rivalen des Open Source Projektes zu unterbinden, dürfen somit keine Unternehmen, die frei erhältliche Version als Teil ihres Produktes oder Lösung vertreiben. Die Doppellizenz Strategie ist eine Art Business Politik, die dem Benutzer erlaubt zwischen der kommerziellen oder der General Public License (GPL) zu wählen. Der Vorteil für den Vertreiber ist eine höhere Kundenaufmerksamkeit und schnellere Anpassung, eine stärkere strategische Positionierung und eine grosse Anzahl von Kunden, die Fehler melden und Verbesserungen empfehlen. Des Weiteren ermöglicht dieser Ansatz eine Art Versuchsprojekt, in dem unabhängige Programmierer die Software testen und verbessern können und das Projekt dadurch weiter bringen. Das Recht die Software frei zu benutzen, ist besser als jede Geld zurück Garantie und kreiert einen mächtigen Vorteil gegenüber einschränkenden Lizenzen. Dieser Ansatz ergibt komplementäre Gewinnströme des traditionellen Softwaremodells, nämlich durch Instandhaltungsangebote oder durch Service wie Beratung und Training. Dabei wird eine starke defensive Marktposition erreicht. Beispielsweise können Unternehmen den Datenbank-Server MySQL für Open Source Projekte bzw. für die innerbetriebliche Anwendung gemäß der GPL kostenlos einsetzen. Wenn Sie hingegen ein kommerzielles Produkt auf der Basis von MySQL entwickeln und samt MySQL verkaufen möchten (ohne Ihren Quellcode zur Verfügung zu stellen), kommt die ebenfalls verfügbare kommerzielle Lizenz zum Einsatz. (Das bedeutet, dass die Weitergabe von MySQL in diesem Fall kostenpflichtig wird.) [14]

4.3.4 Die Consulting Strategie

Unter der Consulting Strategie wird grundsätzlich verstanden, die Lizenzkosten zu reduzieren oder wegzulassen und eine Serviceleistung zu verkaufen. Das heutige Open Source Modell wird eher als ein Service Modell angesehen, indem die Grundfunktionalitäten nichts kosten und die Gewinne durch Anpassungen gewonnen werden (Clay Shirky, 1999). Dies hat auch McKinsey Consulting in einer Studie von 1999 herausgefunden. Lizenzen sind nur noch ein kleiner Teil der Investitionen in die Informationstechnologie, jedoch Beratung und Service beginnen stark zu steigen. 10X ist eine Firma, die sich auf solche Beratung spezialisiert hat. Sie unterstützt Firmen bei der Integration von populärer Open Source Software wie Apache oder MySQL. Gemäss Red Hat sind die Einnahmen von Betriebssystemen nur rund 4% des gesamten Umsätzen einer Linuxbasierten Lösung. Die hohen Gewinne sind durch Integration von Hardware, Software und Instandhaltung zu erwirtschaften, dies entspricht der Integration von Middleware. Alles in allem entsteht durch Open Source die Möglichkeit für Verkäufer fast alle Lizenzkosten zu vermeiden und für den Konsumenten eine Lösung zu guten Preisen anzubieten und trotzdem hohe Gewinne zu erzielen. [14]

4.3.5 Die Abonnementen Strategie

Nach Clupepper, „Gewinne durch Serviceleistungen, sei es Unterhalt oder Beratung, steigen relativ zu den Gewinnen aus Lizenzen. Nach 20 Jahren hat eine typische Software Firma rund 2 Dollar für jeden Dollar, der beim Erwerb einer Lizenz gezahlt wurde, verdient.“ [15] Die Benutzer zahlen periodisch einen kleinen Beitrag um sich für einen Service anzumelden. Es kann zwischen Gratisinhalten und „Premium“-Inhalten ausgewählt werden. Die Abbildung 4.2 zeigt den Trend, wie die Gewinne von Dienstleistungen von Novell und Red Hat stark am steigen sind, wie sie auch viele andere Open Source Software Firmen erfahren, jedoch Lizenzeinnahmen an sinken sind. Bei Red Hat sieht man die steigenden Unterhaltserträge, die jedoch schneller wachsen als bei Novell. Red Hat sagt zu dem ersten Quartal 2004: „Das Grundabonnement berechtigt den Endverbraucher zu einem Jahr von Wartung, die diesen zu Konfigurationssupport und Updates berechtigt.“ Wobei Red Hat Dienstleistungen in Form von Unternehmensberatung, Ingenieurservice, Konsumententraining und Schulung anbietet.

	Redhat		Novell	
	2003	Q104 RPT	2003	Q104 RPT
Revenue	100%	100%	100%	+3%
License			20%	-10%
Maintenance	64%	+70%	79%	+7%
Services	34%	+37%		
Gross	70%		60%	
SGA	56%	+8%	44%	-8%
R&D	24%	+29%	16%	+2%
Operating	(1)%		-	
Net Income	(7)%		(14)%	

Abbildung 4.2: Vergleich Red Hat - Novell [17]

Neben Red Hat und Novell gibt es viele Marktsegmente, die dieses Modell benutzen. Zu den bekanntesten Open Source Projekten zählt LAMP (Linux, Apache, MySQL, PHP), bei der die Firma Covalent Support bietet. [14]

4.3.6 Die Gönnerstrategie

Gönner eines Open Source Projektes zu werden, hat viele strategische Gründe. IBM tut dies z.B. um Standards durchzusetzen und starke Märkte zu öffnen. Wenn eine Firma dazu beiträgt, wird erwartet, dass dabei ein De-Facto-Standard und eine Unterstützungsgemeinschaft darum herum entsteht. Ein anderer Grund um diese Strategie anzuwenden, kann in der Tatsache liegen, dass man eine Firma, die mit einer Software Gewinne erzielt, zu eliminieren möchte. IBM, einer der Hauptanbieter von Linux, versucht x86 Betriebsysteme zu veräussern, um Servergebühren von Microsoft Windows und Sun Solaris zu vermindern. Dabei bietet sich die Gelegenheit einen höheren Wert von Verfügbarkeit und Sicherheit, wie es zum Beispiel mit FireFox und den IExplorer [16] der Fall ist. [14]

4.3.7 Die Benutzer Strategie

Es ist wohl offensichtlich, dass wenn in einer Firma anstelle von teurer, lizenzkostenbehafteter Software Open Source Software benutzt wird, viel gespart werden kann. Die GPL Lizenz erlaubt Software unternehmensintern und privat zu benutzen, ohne die Codeveränderungen und Modifikationen preiszugeben, so lange man die Software, die man kreiert nicht weiter verkauft oder ohne Sourcecode veröffentlicht. Tiefere Kosten und trotzdem einen extrem zuverlässigen Unternehmensservice zu liefern ist äusserst interessant für viele. So gestalten heute viele Unternehmen ihr Geschäftsmodell. Das beste Beispiel dafür ist sicherlich Google oder Amazon. Google verkauft keine Software, bietet lediglich einen Service an oder vermieten eine Service. Die Kosten, die durch Linux auf den 10'000 (2002) Servern gespart werden sind immens. Heute gibt es Gerüchte die 100'000 Google Server schätzen und Dienste die weit über das Suchen hinaus reichen. Amazon verkauft Online Bücher, und es gehen Millionen über den Tisch. E-Trade, ein erfolgreiches internetbasiertes Bankingsystem mit sicherem Tauschservice, benutzt neuerdings auch Linux Server. Was haben alle diese Firmen gemeinsam? Alles sind Dienstleistungsfirmen, die als Eckstein ihre IT Open Source Software benutzen. Der Erfolg spricht für sich. [14]

4.3.8 Die Eingebettete Strategie

Linux wird heute in mehr als der Hälfte der eingebetteten Systeme die auf dem Markt sind benutzt. Von grossen Servern bis zu Mobiltelefonen ist Linux überall zu finden. In der ganzen Welt ist Linux in vielen Tiefpreis- Kommunikationsprodukten mit dabei. Hardwareverkäufer benutzen Linux, da es sehr ausbaufähig und schnell mit minimalem Kapitalaufwand zu implementieren ist. Zudem gibt es dabei nur wenige Komplikationen um mit dem Design zu starten und es ausgiebig zu testen. Dies kommt dem Kunden natürlich auch zu gute. Es werden keine Gelder für Funktionalitäten vergeudet, die gemäss der General Public Lizenz auch gratis zu haben sind. [14]

4.4 Erfolg durch Open Source am Beispiel von LAMP

Mittlerweile gibt es unzählige Beispiele von Open Source Software, die sehr erfolgreich ist. Die LAMP Plattform ist eine der großen Erfolgsgeschichten der Open Source Welt. Dass Akronym LAMP beinhaltet die Kombination aus dem Betriebssystem Linux, dem Web-Server Apache, der MySQL-Datenbank und den Sprachen PHP, Perl und Python und gehören zu den meistgenutzten Systemen für Web-Anwendungen und dienen der dynamischen Webseitenerzeugung. Alle diese Komponenten sind als freie Software erhältlich. Durch die hohe Anzahl von Installationen hat sich eine sehr grosse Community gebildet, was zu einem grossen Netzwerk und einer schnellen und aktiven Weiterentwicklung führt. Kompetente Hilfe ist dabei auch schnell verfügbar und Bugs in der Software werden ebenfalls schnell behoben. Das Grundgerüst das LAMP bereitstellt, wird von zahlreichen Unternehmen genutzt, um ihre Applikationen kostengünstig und effektiv zu entwickeln und zu betreiben. [19]

4.4.1 LINUX

> „I knew I was the best programmer in the world. Every 21-year-old programmer knows that. How hard can it be, it's just an operating system.“ Linus Torvalds, Entwickler des Linux-Kernels [4]

Der Name Linux ist abgeleitet von dessen Initiator, Linus Torvalds. Linux fand seinen Ursprung 1991 an der Universität von Helsinki. Der dort studierende Linus hatte die Idee, ein richtiges „Unix“ für den damals aufkommenden 80386 Prozessor, also ein Unix für einen PC zu schreiben. Als Grundlage nahm er das von Professor Andrew Tanenbaum entwickelte Minix. Er orientierte sich während seiner Programmiertätigkeit an bereits vorhandenen, ausgereiften Standards wie z.B. dem System V- und dem POSIX-Standard. Linus machte den Quellcode frei verfügbar. Er unterstellte ihn der GPL und veröffentlichte ihn im Internet. Heute konzentriert er sich eher auf die Koordinierung der Entwicklung. Linux wird von Software Entwicklern der ganzen Welt weiterentwickelt und gewartet.

Wenn man es genau nimmt, bezeichnet Linux nur den freien Kernel des Betriebssystems. Dazu gehören auch Dateiverwaltung, Speicherverwaltung und Low-Level Funktionen. Für den Einsatz eines Linux-Systems ist aber weitere Software notwendig. Diese darüberliegenden Softwarekomponenten wie zum Beispiel die mächtige Shell als Komandointerpreter, das X-Windowsystem, welches die graphische Benutzeroberfläche darstellt und die vielen weiteren Systemtools, sind nicht an Linux gebunden. Man kann sie austauschen. Dies zeigt einmal mehr den modularen Aufbau des Linux Systems. Der Benutzer kann sein System nach eigenem Ermessen anpassen und zusammenstellen. Kaum ein anderes Betriebssystem kann hier mithalten.

Viele dieser Softwarekomponenten wie zum Beispiel der Compiler oder der Debugger, wurden bereits vor der Entwicklung von Linux als freie Software für Unix Systeme unter dem GNU Projekt entwickelt. Die Softwarekomponenten werden dann mit dem Kernel zu einem Gesamtpaket vereint. Diese Linux-Distribution genannten Systeme greifen insbesondere auf das GNU System des GNU Projektes zurück. Dies ist ein Grund dafür, weshalb

einige Entwickler diese Software-Bündel auch als GNU/Linux und nur den Kernel als Linux bezeichnen. Software für Linux wird neben GNU von etlich weiteren Arbeitsgruppen, Firmen und auch von interessierten Privatleuten geschrieben.

Die so gennanten Distributoren fassen sämtliche Programme zu einem Paket zusammen, und bieten dieses als Ganzes an. Oft setzen sich die Distributoren auch politisch stark für die Verbreitung von freier Software ein, denn sie sind die nächstliegende Verbindung zwischen den neuen Anwendern und der Software. Um die erweiterten Linux-Systeme unter einen Hut zu bringen und ein zu weites Auseinanderdriften zu vermeiden, wurde der Linux Standard Base LSB ins Leben gerufen. Dies ist ein Standard, der die Lauffähigkeit von Anwendungen auf allen Distributionen garantieren soll. [20] Als namhafte Distributoren können folgende genannt werden [4]:

- **Debian:** Nichtkommerzielle Distribution, die auf Basis der GPL entwickelt wird.
- **Mandriva:** Eine RPM-basierte Distribution aus Frankreich, auch für Anfänger geeignet
- **Red Hat / Fedora:** Fedora ist die Community Edition von Red Hat. Im amerikanisch-englischen Raum meistgenutzte Distribution, ebenfalls auch für Anfänger geeignet.
- **SuSE:**Im deutschen Raum meistgenutzte Distribution, für Anfänger geeignet
- **Turbolinux:** Weitverbreitete Distribution in China und Japan
- **Slackware:** Älteste Distribution, minimalistisch, Unixähnlich, Linux Grundwissen muss vorhanden sein

In der Abbildung 4.3 werden die wichtigsten Linux-Distributionen miteinander verglichen.

Distribution	Active sites Sep '04	Active sites Mar '05	6-month Growth Rate
RedHat	1630382	1610427	-1.2%
Debian	693941	791086	14.0%
Cobalt	619960	516963	-16.6%
SuSE	399031	442908	11.0%
Fedora	182421	405682	122.4%
Mandrake	62972	73459	16.7%
Gentoo	43525	63160	45.1%

Abbildung 4.3: Anteile von Betriebssystemen auf Sites [21]

Diese Distributionen unterscheiden sich hauptsächlich in folgenden Punkten [22]:

- Welche Software-Pakete werden verwendet (welches GUI, Vorkonfiguration)

- Welche Verwaltungswerkzeuge werden beigelegt
- Ist eine gute Dokumentation vorhanden
- Gibt es offiziellen Support
- Disitributionspolitik
- Wird die Distribution offiziell von kommerziellen Software-Anbietern unterstützt

Die meisten Installationen von Linux sind auf Servern zu finden. Dies lässt sich durch die Stabilität, Flexibilität und die Geschwindigkeit von Linux erklären. Die Auswahl eines Betriebssystems für den privaten Gebrauch, geschieht nach anderen Gesichtspunkten. Es muss einfach zu bedienen sein und viele Anwendungen enthalten. Software-Projekte wie zum Beispiel KDE und GNOME versuchen dies zu realisieren. Linux ist immer mehr als Desktop System für den privaten Gebrauch als ernstzunehmende Alternative zu betrachten. Den Ruf eines Betriebssystems für Freaks hat Linux mittlerweile abgelegt. [4]

4.4.2 Apache

> „When you talk to someone who's not technical, they need to realize that there's an option, that there's a Pepsi to Microsoft's Coke.“ Brian Behlendorf, Apache Software Foundation [23]

Brian Behlendorf ist der bekannteste Kopf der Apache Gemeinde. Seiner Meinung nach hängt der Erfolg von freier Software entscheidend davon ab, dass der Massenmarkt gleich viel Vertrauen in Open Source steckt, wie zum Beispiel in Microsoft. Es muss verdeutlicht werden, dass freie Software eine Alternative ist, so wie „Pepsi eine Alternative zur Microsofts Coke“ ist, um es mit den Worten von Behlendorf zu formulieren.

Der Apache Webserver entstand als Erweiterung des alten NCSA-HTTP-Servers. Es ist eines der erfolgreichsten Open-Source-Projekte überhaupt. Das Apache HTTP Server Projekt ist ein gelungener Versuch einen freien, robusten, erweiterbaren und skalierbaren Webserver zu kreieren.

Das Projekt wird gemeinsam von einer Gruppe freiwilligen aus der ganzen Welt betreut. Zusätzlich haben hunderte von Benutzern Ideen geliefert und zur Entstehung des Codes bzw. der Dokumentation beigetragen.

Die Entwickler arbeiten ununterbrochen an der Qualität des Webservers. Der Quellcode wird also ständig auf Fehler untersucht und das Risiko für Fehler im Code wird dadurch verhindert. Bei einem proprietären Server hingegen, dauert es relativ lange bis zum Beispiel ein Sicherheitspatch zur Verfügung gestellt wird. Somit muss der zahlende Kunde auf die nächste grössere Version warten. Ein weiterer Vorteil von Apache ist der modulare Aufbau. Durch entsprechende Module kann er beispielsweise die Kommunikation zwischen Browser und Webserver verschlüsseln oder als Proxy Server eingesetzt werden. Ein wichtiger Punkt der Apache zum Durchbruch verholfen hat, ist seine Fähigkeit mehrere

verschiedene Webseiten unter einem Apache zu betreiben. Bis anhin galt die Regel, pro Webserver nur eine Domain zu betreiben.

Vorteile des Apache [25]:

- modularer Aufbau
- Virtual Hosts
- für verschiedene Betriebssysteme
- extrem leistungsfähig und stabil
- wird laufend weiterentwickelt
- breite Konfigurationsmöglichkeiten
- gratis

Der Open Source Webserver Apache ist unumstrittener Marktführer auf der ganzen Welt (siehe Abb. 4.5). Zwei Drittel der Webserver sind vom Typ Apache. Alle acht Sekunden kommt eine neue Apache-gestützte Web-Site hinzu (siehe Abb. 4.4).

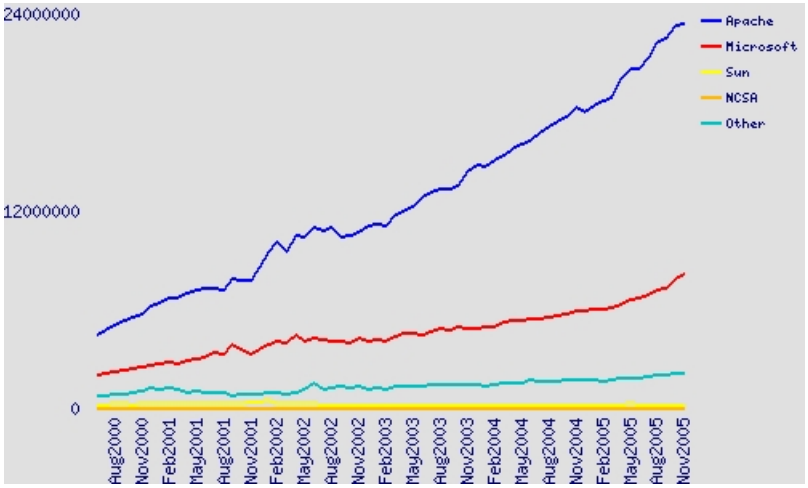


Abbildung 4.4: Aktive Server [17]

Das Projekt ist ein Teil der Apache Software Foundation, welche 1999 gegründet wurde. Die Apache Software Foundation (ASF) ist eine ehrenamtlich arbeitende Organisation zur Förderung der Apache-Softwareprojekte. Obwohl die ASF als Non-Profit Organisation gilt, weist sie jedoch durchaus Charakteristika eines Unternehmens auf. So gibt es zum Beispiel einen Aufsichtsrat, der von den Mitgliedern der ASF gewählt wird. Er trägt Verantwortung und fällt Entscheidungen. Die Aufgaben der ASF sind der rechtliche Schutz aller Projekt-Mitarbeiter und der Schutz der Marke Apache. Zudem ist die Apache Lizenz dort entsprungen. Vor allem kommerzielle Firmen, wie IBM stellen ihr privates Entwicklerteam zur Verfügung, weil sie Apache Webserver als Teil ihres Web-Sphere Produktes verkaufen. Von solchen Partnerschaften innerhalb einer Organisation wie ASF profitiert

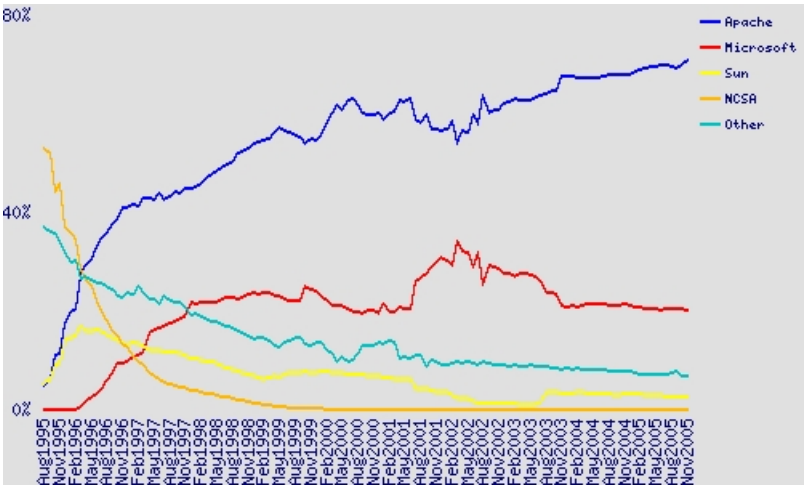


Abbildung 4.5: Marktanteile von Servern [17]

nicht nur IBM, sondern auch die Open Source Gemeinde. Die ASF glänzt durch ihre professionelle Struktur und kann daher eine sehr hohe Akzeptanz im IT Markt sicherstellen. Dadurch können auch finanzielle und organisatorische Unterstützungsleistungen für ihre Open Source Projekte erwirtschaftet werden. [4]

4.4.3 MySQL

MySQL ist ein leistungsfähiges, relationales Datenbank Management System (RDBMS). Es ist freie Software, die zu den weitverbreitesten Open Source Programmen überhaupt gehört. [24] Zu ihren Besonderheiten gehören der grosse Funktionsumfang, Flexibilität, seine hohe Geschwindigkeit und die Stabilität selbst unter Last. Weltweit schätzt man über 6 Millionen Installationen und zählt über 30.000 Downloads täglich. Die führende Open Source Datenbank MySQL wird schnell zum Kern von vielen großen, geschäftskritischen Anwendungen.

MySQL ist ein Open Source Unternehmen der zweiten Generation mit einem dualen Lizenzmodell, das Open Source Werte und Methoden in einem gewinnbringenden und tragfähigen Geschäftsmodell unterstützt. Das Produkt ist unter der Open Source Lizenz GPL erhältlich, es werden aber auch kommerzielle Lizenzen verkauft. [26]

Zu den wichtigsten Kunden gehören Yahoo!, DaimlerChrysler, Cox Communications, die NASA und die Nürnberger Versicherungsgruppe. Die Kunden erzielen wesentliche Kosteneinsparungen, indem sie MySQL für ihre Webseiten, für geschäftskritische Unternehmensanwendungen, und für ihre Hochlastsysteme einsetzen. Vor allem in den Sektoren Lizenz-, Hardware-, Verwaltungs-, Entwicklungs- und Supportkosten können Einsparungen von 50-90 Prozenz verzeichnet werden. Relevante Vorteile sind die um 60 Prozent verringerten Ausfallzeiten, die praxisgeprüften Produkte und die schnelle Weiterentwicklung. [24]

Bei genauem Betrachten des Entwicklungszyklus [24] (siehe Abb. 4.6), ist die schnelle Entwicklung von MySQL klar ersichtlich. Alle 4-6 Wochen wird ein neuer Release entwickelt, der anschliessend von einer grossen Community eingesetzt und getestet wird. Parallel dazu

findet ein intensives Debugging statt. Sobald der Release fehlerfrei läuft, beginnt die Auslieferung an die kommerziellen Lizenznehmer. Die Erlöse dieser kommerziellen Lizenzen fliessen wiederum zurück in die Entwicklung.

MySQL wird für alle gängigen Linux-Distributionen sowie für Unix, Mac OSX, Windows und viele weitere Betriebssysteme als ausführbares Paket zum freien Download zur Verfügung gestellt.

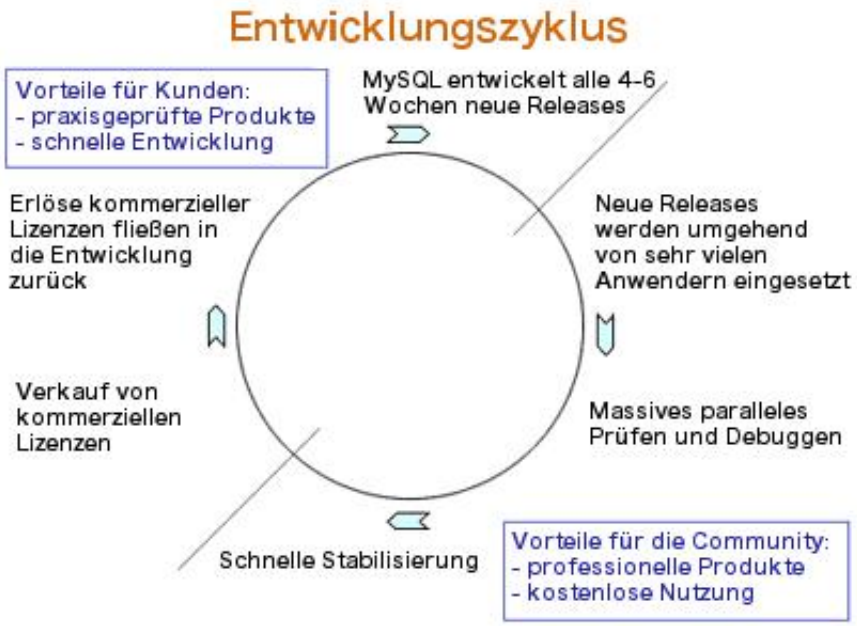


Abbildung 4.6: Entwicklingszyklus von MySQL [24]

4.4.4 PHP

PHP (Hypertext-PreProcessor, einstmals belächelt als 'Personal-Homepage') hat sich im Laufe der letzen Jahren zu der weitverbreitetsten und meistverwendesten Scripting-Sprache im Web entwickelt. Laut den Oktober-Statistiken von Netcraft (siehe Abb. 4.7) hat es PHP4 geschafft auf über 20 Millionen Domains und mit rund 1 Million IP's im Einsatz zu sein.

Der Erfolg war absehbar, da Funktionsumfang und Portabilität schon sehr früh zu überzeugen wussten. Die neulich erschienene Version 5 der Sprache (PHP5) brachte letztendlich die langerwartete 'Reife' und erweiterte OOP-Fähigkeiten. Das heutige Einsatzgebiet von PHP reicht vom simplen Gästebuch-Skript bis hin zu komplexem E-Commerce, Groupwares, Blogs, Wikis oder Content-Management-Systemen. Seine Beliebtheit verdankt PHP seiner leichten Erlernbarkeit, der engen und mühelosen Anbindung an performante Datenbanken wie MySQL oder PostgreSQL.

Die Geschichte von PHP begann als „Personal Homepage". Das waren ein paar, ursprünglich in Perl, später in C programmierte Funktionen, um Zugriffe auf einen Webserver zu

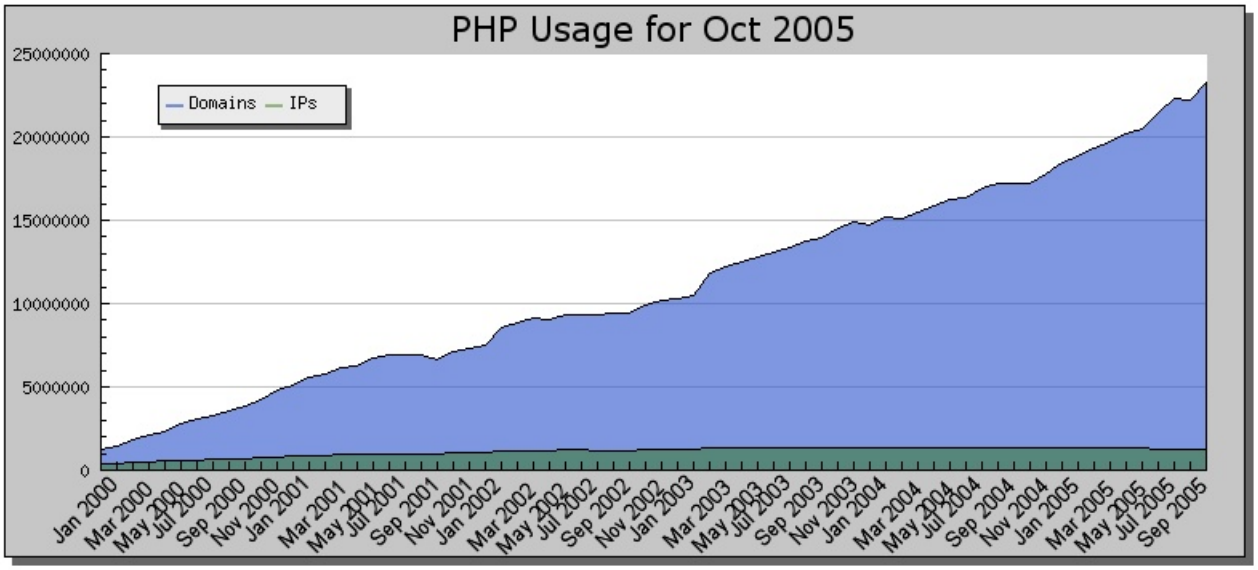


Abbildung 4.7: Usage PHP [27]

zählen (Counter). Der Autor, Rasmus Lerdorf, programmierte zudem ein paar Funktionen, um SQL Abfragen vom Webserver aus auszuführen. Aus diesen beiden anfänglichen Funktionalitäten folgten 1995 PHP/FI. Dieses Kürzel stand für „Personal Homepage / Forms Interpreter“, welches 1997 komplett neu geschrieben und als PHP/FI2.0 veröffentlicht wurde.

Im Jahre 1997 schrieben Andi Gutmans und Zeev Suraski PHP komplett neu. Sie taten es um PHP für ein eigenes Projekt zu verwenden. Mit diesem ReWrite kam der Durchbruch. In Zusammenarbeit mit Rasmus Lerdorf entstand so PHP3. Es verbreitete sich schnell weiter, dank der offenen Spracharchitektur und der leichten Erlernbarkeit. PHP4 lässt sich einfach in HTML einbinden und hat eine eingebaute Schnittstelle zur Datenbank MySQL (was allerdings in der neuen Version 5 der Sprache nicht mehr der Fall ist). Dank all seiner Vorteile erfreute sich PHP in Kreisen von Webmastern und Providern schon sehr bald immer grösser Beliebtheit und Akzeptanz. Diese trug letztendlich wesentlich zum raschen Zuwachs an Funktionalität und Modulen (Extensions) bei. [30]

Der Erfolg von MySQL ist eng mit dem Erfolg von PHP verknüpft. Das Gespann ist der Unterbau vieler Content Management Systemen (CMS) und E-Commerce-Anwendungen. Im Jahre 2000 entstand PHP 4. Zu den erweiterten Funktionalitäten gehörten die verbesserte Modularität und die Pufferung von Ausgaben. Der Kern wurde neu entwickelt (ZendEngine) und somit konnten auch komplexe Anwendungen effizient umgesetzt werden. Bislang war PHP nur teils objektorientiert (pseudo-OOP). Es war zwar möglich Klassen mit Methoden zu schaffen, diese konnte man aber nicht überladen oder auch Zugriffsmodifizierer wie „private“ oder „protected“ kannte PHP4 noch nicht. Mit PHP 5 steigt die Skript-Sprache nun in die Liga der objektorientierten Sprachen auf [28]. Der überarbeitete Kern (ZendEngine2) bietet nun umfänglichere OOP-Fähigkeiten an. Dank PEAR [29] und phpclasses.org hat PHP öffentliche Code-Repositorien, welche für jeden zugänglich sind und bereits getestete und stabile Werkzeuge (Klassen) für den täglichen Gebrauch zur Verfügung gestellt.

4.5 Fazit

Bei allen Vorteilen, die hier mit Open Source Software genannt wurden, muss man sicherlich sagen, dass es eine gute Sache ist. Nicht nur die vielen Vorteile, sondern auch die vielen erfolgreichen Firmen, die sich mit Open Source Software eine gute Existenz aufgebaut haben und hier genannt wurden, sprechen nur für die Sache Open Source. Wir wollen aber dabei immer realistisch bleiben. Im Vergleich von Kassensturz [31] haben Markforscher errechnet, dass Unternehmen, unter bestimmten Bedingungen, durchaus erhebliche Einsparungen in Server-Bereich erreichen können. Man muss sich immer im Klaren darüber sein, dass die meisten Benutzer nicht über ein grosses Wissen an Open Source Produkten verfügen und so erhebliche Support und Schulungskosten entstehen können. In diesem Zusammenhang meinten Markforscher, dass für kleine und mittlere Unternehmen sich der Umstieg auf Open Source Software finanziell kaum lohnt. Vor allem die jährlich zu zahlenden Maintenance-Kosten und die hohen Schulungkosten belasten KMU stärker als der Einsatz von kommerzieller Software. Für grosse, multinationale Unternehmungen ist es jedoch durchaus lohnend. Bei grossem Einsatz von Open Source Software kann viel gespart werden. Dies gilt für alle Unternehmen. Ob nun kommerziell Software oder Open Source Software benutzt werden soll ist jedem selbst überlassen. Diese Arbeit hat durchaus die Vorteile einer Open Source Lösung gezeigt.

4.6 Anhang

4.6.1 Die wichtigsten Lizenzen

Im folgenden Teil werden die wichtigsten Lizenzarten der Open Source Bereichs erläutert. Die Tabelle 4.8 zeigt eine Übersicht über die unterschiedlichen Lizenzarten und Eigenschaften.

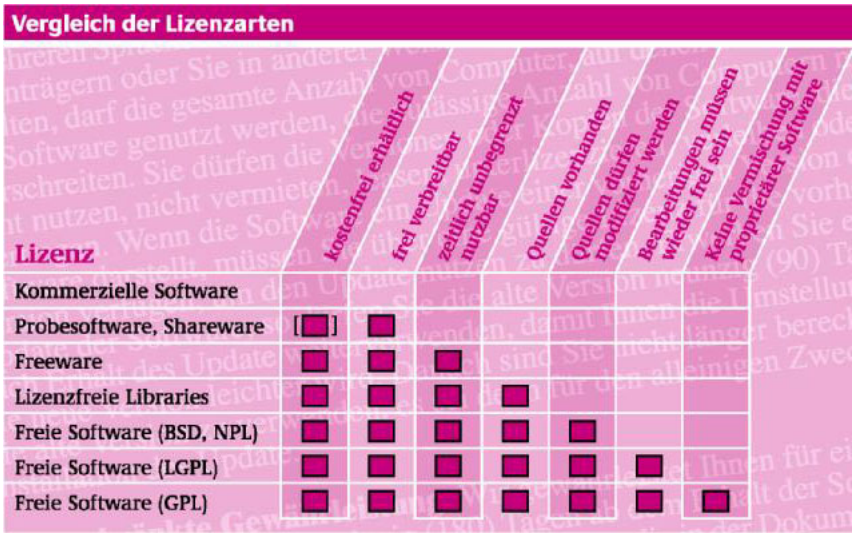


Abbildung 4.8: Vergleich von Lizenzarten [32]

GPL: GNU General Public License

Die GNU General Public License umfasst die Nutzungsbedingungen der freien Software, wie sie in Teil eins der Begriffsdefinition anhand der vier Freiheiten erklärt wurden. Diese sind kurz wiederholt: Ausführen, kopieren, verändern und verbreiten des Quellcodes mit bestimmten Einschränkungen wie bspw. der Pflicht Veränderungen ersichtlich zu machen, die Lizenzbedingungen weiterzugeben und auf den Ursprung zu verweisen. Diese Freiheiten verfallen, wenn gegen die Bestimmungen der GPL verstossen wird. Das Zusammenschliessen von GPL-Software und proprietärer Software ist nur möglich, wenn die daraus resultierende Software unter die GPL gestellt wird. Dies verhindert, dass Teile von freier Software in proprietärer Software verwendet werden kann. Die GPL schliesst wie die meisten Lizenzen, Haftung und Gewährleistung aus, sofern dies mit den länderspezifischen Rechtsgrundlagen vereinbar ist. [33]

LGPL: Lesser General Public License

Die Lesser General Public License ist weniger restriktiv als die eben genannte Variante. Sie übernimmt die gleichen Freiheiten, wie die GPL, aber sie gestattet ausdrücklich, dass alle unter dieser Lizenz stehenden Bibliotheken und Programme in proprietärer Software eingebunden werden können. Dieser Zustand kann zur Verbreitung freier Bibliotheken beitragen. Ein Wechsel von der LGPL zur GPL ist jederzeit möglich, umgekehrt kann nachträglich jedoch nicht gewechselt werden. [34]

BSD-License: Berkeley Software Distribution License

Die BSD-Lizenz enthält im Wesentlichen die gleichen Freiheiten wie die GPL. Sie verlangt jedoch bei weitem nicht so viele Einschränkungen. Software, die unter BSD-Lizenz steht, darf mit oder ohne Veränderungen als Quellcode oder Binary verbreitet werden. Zudem kann sie auch in kommerzielle Systeme eingebunden werden und dabei sogar unter konventielle Lizenzen gestellt werden. Die BSD-Lizenz verlangt aber, dass der Copyright Vermerk bei weiterentwickelten Programmen auch den ursprünglichen Autor beinhaltet. [35]

MPL (Mozilla Public License)

Die Mozilla Public License gilt als eine schwache Copyleft Lizenz und ist vereinfacht gesagt ein Kompromiss zwischen der GPL und der BSD-Lizenz. Sie garantiert gebührenfreien Zugang zur Software. Falls an dieser Veränderungen vorgenommenen werden, müssen sie im Quellcode verfügbar sein. Die veränderte Software sollte grundsätzlich auch unter der MPL stehen, es dürfen jedoch Verbindungen mit proprietären Software gemacht und unter einer anderen Lizenz als MPL veröffentlicht werden. [36]

Literaturverzeichnis

[1] Was ist Open Source und Free Software, Open Source Training and Consulting GmbH, http://www.ostc.de/opensource.html#FreeSoftware, rechechiert 12.11.2005

[2] Philosophy of the GNU Project, FSF (Free Software Foundation), (updated) 22.07.2005, http://www.gnu.org/philosophy/free-sw.de.html, November 2005

[3] The Open Source Definition, deutsche Übersetzung, November 2005, http://www.opensource.org/docs/definition.php, recherchiert am 8.11.2005

[4] Bravehack, Jens Sieckmann, März 2001, http://www.bravehack.de/html/node40.html, recherchiert am 8.11.2005

[5] Die Kathedrale und der Basar, Eric S. Raymond, 8. August 1999, deutsch Übersetzung, http://gnuwin.epfl.ch/articles/de/Kathedrale/catb_g.1.html, recherchiert am 8.11.2005

[6] Der Verzauberte Kessel, Eric S. Raymond, Juni 1999, deutsche Übersetzung, http://www.oreilly.de/opensource/magic-cauldron/cauldron.g.01.html, recherchiert am 8.11.2005

[7] Open Sources - Voices from the Open Source Revolution, Chris Di Bona, Sam Ockman, Mark Stone, Januar 1999, http://www.bravehack.de/html/node41.html, rechechiert am 12.11.2005

[8] HP sounds Net ralling cry, Dan Briody and Ephraim Schwartz , 7.5.1999, http://www.infoworld.com/cgi-bin/displayStory.pl?99057.hnfremont.htm, recherchiert am 12.11.2005

[9] Open Source Business Models, Rapberger, Sommer 2000, http://wwwai.wu-wien.ac.at/~koch/lehre/inf-sem-ss-00/rapberger/main.htm, recherchiert am 14.11.2005

[10] Open Source gefährdet die Software-Wirtschaft, Microsoft, 17.03.2004, http://www.tecchannel.de/news/themen/business/417433/index.html, recherchiert am 15.11.2005

[11] Open Source wird sterben, wenn Softwarepatente kommen, Ökonomin, 18.9.2004, http://www.heise.de/newsticker/meldung/51217, recherchiert am 15.11.2005

[12] Open Source: Dimensionen eines Phänomens, Mathias Bärwolff, Oktober 2004, http://ig.cs.tu-berlin.de/lehre/w2004/ir1/ablauf/date-13/Baerwolff-OpenSource-DimensionenEinePhaenomens-2004-12-01.pdf, recherchiert am 15.11.2005

[13] Setting Up Shop: The Business of Open-Source Software, Frank Hecker, 20 June 2000, http://www.hecker.org/writings/setting-up-shop, recherchiert am 15.11.2005

[14] Seven open source business strategies for competitive advantage, John Koenig, 14.5.2004, http://manage-ment.itmanagersjournal.com/management/04/05/10/2052216.shtml?tid=85, recherchiert am 15.11.2005

[15] Open Source Paradigm Shift, Tim O'Reilly, Juni 2004, http://tim.oreilly.com/articles/paradigmshift_0504.html, recherchiert am 15.11.2005

[16] Firefox und IE: Die Sache mit der Sicherheit, Tobias Röhrig, 28.12.2004, http://www.netzwelt.de/news/69036-firefox-und-ie-die-sache.html, recherchiert am 15.11.2005

[17] November 2005 Web Server Survey, wss, November 2005, http://news.netcraft.com/archives/2005/11/07/november_2005_web_server_survey.html, recherchiert am 15.11.2005

[18] Open Source Software Business Models, 9.11.2004, http://www.jarche.com/node/356, rechechiert am 15.11.2005

[19] MySQL in der LAMP-Area/LinuxTag, Juni 2005, http://www.mysql.de/news-and-events/press-release/release_2005_19.html, recherchiert am 10.11.2005

[20] Open Source Tutorial, 3.2 Linux, http://www.dbus.de/eip/kapitel03b.html, recherchiert am 12.11.2005

[21] Colin Phipps, März 2005 http://news.netcraft.com/archives/around_the_net.html, recherchiert am 15.11.2005

[22] http://de.wikipedia.org/wiki/Linux-Distribution, recherchiert am 12.11.2005

[23] Brian Behlendorf, Apache Software Foundation, Juni 1999, http://www.my-opensource.org/lists/myoss/1999-06/msg00135.html, recherchiert am 13.11.2005

[24] MySQL AB, Über MySQL, 2005, http://www.mysql.de/company, recherchiert am 10.12.2005

[25] Webserver Apache, http://www.ancoso.com/opensource/Apache/document_view, recherchiert am 13.11.2005

[26] HTTP Server Project - The Number One HTTP Server on the Internet, http://httpd.apache.org/, recherchiert am 15.11.2005

[27] Usage Stats for October 2005, Information kindly provided by netcraft.com, `http://www.php.net/usage.php`, recherchiert am 15.11.2005

[28] PHP 5 - Die Neuerungen, 1. Auflage 2004, Martin Goldmann, Markus Schraudolph, Galileo Computing

[29] Stig Bakken, Pear - PHP Extension and application repository, 2005, `http://pear.php.net`, recherchiert am 10.12.2005

[30] Anhang A. Die Geschichte von PHP und verwandten Projekten, `http://php.mirrors.ilisys.com.au/manual/de/history.php`, recherchiert am 15.11.2005

[31] Open-Source-Studie: Microsoft mitunter im Vorteil, 11.10.2004, `http://www.heise.de/newsticker/meldung/52027`, recherchiert am 20.11.2005

[32] Freie Software - Rechtsfreier Raum?, Jürgen Siepmann, München 2000, LinuxLand International

[33] GNU General Public License, Free Software Foundation, `http://www.gnu.org/copyleft/gpl.html`, recherchiert am 22.12.2005

[34] GNU Lesser General Public License, Free Software Foundation, `http://www.gnu.org/copyleft/lesser.html`, recherchiert am 22.12.2005

[35] The BSD License, Open Source Initative, `http://opensource.org/licenses/bsd-license.php`, recherchiert am 22.12.2005

[36] Mozilla Public License Version 1.1, Mozilla Foundation, `http://www.mozilla.org/MPL/MPL-1.1.html`, recherchiert am 22.12.2005

Kapitel 5

Technical and Economic Aspects of Inter-domain Service Provisioning

Bas Krist, Markus Sonderegger, Roland Haas

Die vorliegende Arbeit beschäftigt sich mit dem Einsatz verteilter Systeme in virtuellen Organisationen und fokussiert dabei auf die ökonomischen und technischen Aspekte von domänenübergreifenden Infrastrukturen. Zusätzlich zur evolutionären Entwicklung der Wertschöpfungskette und der Definition daraus resultierender virtuellen Organisationen stehen die technischen und ökonomischen Anforderungen im Vordergrund. Weiter werden ausgewählte Aspekte eines Frameworks für das Service Provisioning detaillierter betrachtet.

Inhaltsverzeichnis

5.1 Aufbau der Arbeit

Folgend wird kurz beschrieben, wie die Arbeit aufgebaut ist. Zu Beginn wir die Problemstellung erklärt und anhand des Begriffes der Wertschöpfungskette nach Porter und der Virtuellen Organisation ins Thema eingeführt. Es werden dann Anforderungen an ein verteiltes System, das durch eine Virtuelle Organisation genutzt wird, beschrieben und die Möglichkeit eines generischen Frameworks (Open Grid Services Architecture) zur Umsetzung der Anforderungen vorgestellt. Zudem werden die Problemstellungen in den Bereichen A4C und Regulations skizziert. Zum Schluss der Arbeit folgt Zusammenfassung und Ausblick.

5.2 Einführung und Problemstellung

Im folgenden Kapitel wird anhand der Werteschöpfungskette beschrieben, wie Unternehmen früher, heute und in der Zukunft, ökonomischen Mehrwert schaffen. Angefangen wird mit Porters Wertschöpfungskette, bis zur Virtuellen Organisation der Zukunft.

5.2.1 Evolution der Wertschöpfungskette

Porters Wertschöpfungskette

Die traditionelle Wertekette [1] zeigt das Unternehmen als Aneinanderreihung von verschiedenen Wertschöpfungsstufen. Auf der linken Seite steht zum Beispiel die Einkaufslogistik, und ganz rechts könnte die Kundenbetreuung stehen. Die Unternehmung verbindet verschiedene Stufen der Wertschöpfung zu einer Wertschöpfungskette und der Kunde interagiert mit dem Unternehmen. In Abbildung 5.1 sieht man, dass das Unternehmen auf einer Stufe der Wertschöpfungskette mit einem anderen Unternehmen zusammenarbeitet. Dies ist ersichtlich durch den vertikalen Pfeil in der Mitte der Abbildung.

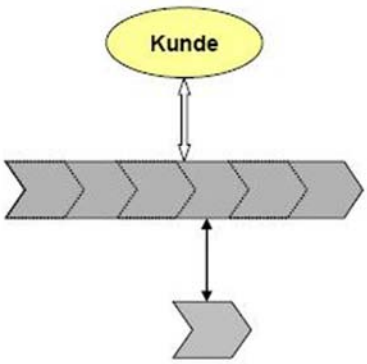


Abbildung 5.1: Traditionelle Wertschöpfunskette [2]

Fokus auf Kernkompetenzen

Durch den Einfluss von Economies of Scale und Scope in Zusammenhang mit technologischen Fortschritten auf dem Gebiet der Informations- und Kommunikationstechnologie, verschiebt sich der Fokus für einzelne Unternehmen auf ihre Kernkompetenzen. Diese Entwicklung führt zu typischem Outsourcing, wie zum Beispiel in der Automobilindustrie. BMW kauft für ihre neue Mini Cooper Motoren von Peugeot [3], damit sie nicht selber grosse Summen in die Entwicklung eines neuen Motors investieren muss. Ein weiterer Vorteil ist die Kostentransparenz für das outsourcende Unternehmen, sie wissen genau, wie viel eine einzelner Schritt in der Wertschöpfungskette kosten wird. Das Unternehmen hat jetzt die einzelnen Wertschöpfungsschritte nicht mehr unter dem eigenem Dach, daher wird jetzt das Qualitätsmanagement sehr wichtig. Die beiden Unternehmen, Lieferant und Outsourcer, müssen eng zusammenarbeiten um dafür zu sorgen, dass die Wertschöpfungsstufen weiterhin reibungslos zueinander passen.

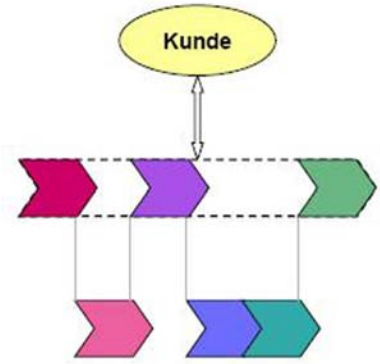


Abbildung 5.2: Zeitgemässe Wertschöpfunskette [2]

5.2.2 Virtuelle Organisation

Der nächste Schritt in Richtung 'Fokus auf Kernkompetenzen' ist die Virtuelle Unternehmung. Durch den geschickten Einsatz von Informations- und Kommunikationstechnologie wird dieses neue Businessmodell ermöglicht. Die Virtuelle Unternehmung, in Abbildung 5.3 'Agent' genannt, wird aus verschiedenen Unternehmen zusammengefasst. All diese Unternehmen konzentrieren sich vollumfänglich auf ihre Kernkompetenzen, und die Virtuelle Organisation sorgt dafür, dass sie alle integriert werden. In Abbildung 5.3 wird nur eine flache Unternehmenshierarchie gezeigt, es ist aber durchaus möglich, dass jedes Unternehmen für sich seine Aufgaben weiter auf andere Unternehmen verteilt. Die Virtuelle Unternehmung behält die Übersicht über ihre Wertekette, ihr Kunde nimmt sie als eine Unternehmung wahr. Die Ressourcen der einzelnen Unternehmung werden zusammengefügt, die Kernkomptenzen komplementieren einander, und die ganze Integration zu einer Virtuellen Unternehmung wird durch den Einsatz geeigneter Informations- und Kommunikationstechnologie ermöglicht. Beim traditionellen Outsourcing wird das Management und die Kommunikation an den Unternehmensschnittstellen zum wichtigen Erfolgsfaktor, bei der Virtuelle Organisation sind diese beide Kompetenzen Kern der Unternehmung.

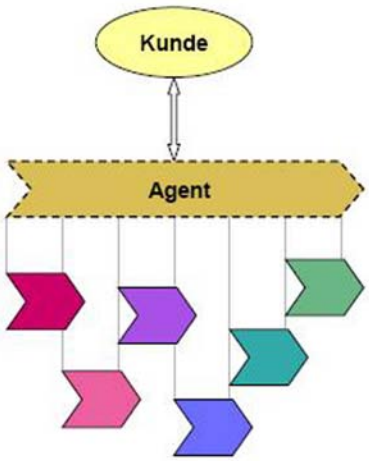


Abbildung 5.3: Zukünftige Wertschöpfunskette [2]

Um den Begriff der Virtuellen Unternehmung zu konkretisieren, wird hier ein Beispiel aus [5] beschrieben. Es handelt sich hierbei um eine Autovermietung, welche sich aber vom traditionellen Vermieter unterscheidet. In Abbildung 5.4 wird die Virtuelle Unternehmung dargestellt.

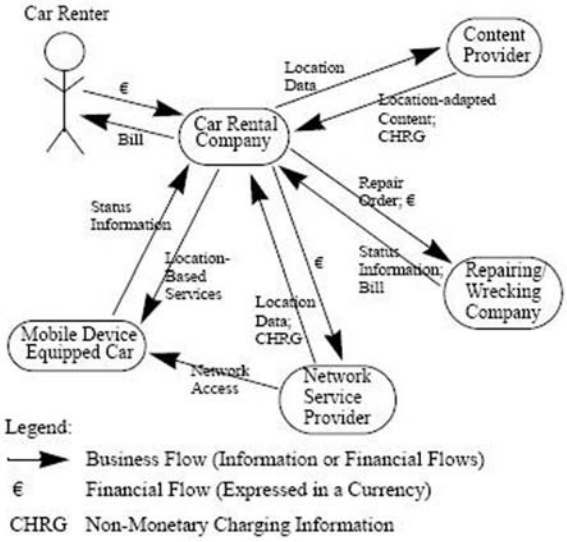


Abbildung 5.4: Virtuelle Organisation [5]

Das Auto ist ausgerüstet mit einem mobilen Gerät, welches es ermöglicht, das Auto mit Hilfe eines Networkproviders zu lokalisieren. Wenn das Auto zum Beispiel verunfallt, kann automatisch eine Nachricht an eine Garage geschickt werden, um das Auto abzuholen und zu reparieren. Für den Mieter kann das mobilen Gerät mit Hilfe eines Contentproviders ortsabhängige Informationen liefern.

Die Beteiligten sind die folgenden: der Kunde, der Autovermieter, der Contentprovider, der Netzwerkprovider und die Garage. Alle diese Marktteilnehmer sind Teil der Virtuellen Organisation, der Kunde aber interagiert nur mit dem Autovermieter und erhält auch nur

eine Rechnung. In dieser Abbildung werden auch noch Daten- und Geldflüsse dargestellt, um ersichtlich zu machen, wer was macht und wer wo Geld verdient.

5.3 Aufbau einer virtuellen Organisation

Nachdem beschrieben wurde, wie eine Virtuelle Organisation entsteht, wird jetzt tiefer auf die Eigenschaften eingegangen, in denen sich eine Virtuelle Organisation von anderen Unternehmensformen abgrenzt. Als Abschluss des Kapitels 5.3 wird einen ersten Blick auf die Infrastruktur der Virtuellen Organisation, die Services, geworfen.

5.3.1 Eigenschaften einer VO

Bis jetzt wurde die Virtuelle Organisation nur grob beschrieben, aber was genau macht eine Virtuelle Organisation aus? In der Theorie existieren viele Definitionen einer Virtuellen Organisation, wobei die meisten mindestens einen Absatz lang sind. Um die Virtuelle Organisation besser zu charakterisieren, orientiert sich die Arbeit an der Definition von [4]. Eine virtuelle Organisation hat eine Reihe bestimmter Eigenschaften [4], welche sie von anderen Unternehmensformen abgrenzt.

- Unternehmensgrenzen überschreitend
- Geographisch Verteilte Teilnehmer
- Zusammenfügen von Ressourcen
- Komplementierung von Kernkompetenzen
- Einsatz von IKT

Unternehmensgrenzen überschreitend: Verschiedene Unternehmen sind an der Virtuellen Organisation beteiligt. Jedes Unternehmen bringt seine Kernkompetenzen mit ein, damit das gewünschte Preis- und Qualitätsniveau erreicht wird. Für die Virtuelle Organisation ist die Kommunikation und das Management der Unternehmensschnittstellen überlebenswichtig.

Geographisch verteilte Teilnehmer: Durch den Einsatz von Informations- und Kommunikationstechnologie verschwinden geographische Distanzen. Die Relevanz des Unternehmensstandortes wird immer kleiner, und ermöglicht damit die Zusammenarbeit mehrerer geographisch verteilter Unternehmungen. Zum Beispiel gibt es Firmen, die ihre Produktion am Abend ins Ausland verlegen, um somit die verschiedenen Zeitzonen auszunutzen und den Arbeitstag zu verlängern. Wenn ein Arbeitstag in den USA zu Ende geht, fängt er in Indien gerade erst an. Für die Virtuelle Organisation ist es wichtig, dass die lokalen Gesetze eingehalten werden. Auf diese Thematik wird in Abschnitt 5.5.3 eingegangen.

Zusammenfügen von Ressourcen: In der virtuellen Organisation arbeiten unterschiedliche Unternehmungen eng zusammen, daher treten Problemen durch unterschiedliche Systeme auf. Heterogene Ressourcen müssen integriert werden, damit eine Zusammenarbeit ermöglicht wird. Diese Integration stellt eine grosse Herausforderung für die Virtuelle Organisation dar. Teilkapitel 5.4.1 geht näher auf diese Problematik ein.

Wechselnde Teilnehmer: Durch den dynamischen Charakter der Virtuellen Organisation ist deren Aufbau nicht immer gleich. Sich ändernde Anforderungen und Umfeldbedingungen führen dazu, dass neue Teilnehmer einer Virtuellen Organisation hinzukommen, andere wiederum aus ihr verschwinden. Klar definierte Schnittstellen sind hier die grösste Herausforderungen, um diese Dynamik unterstützen zu können.

Komplementäre Kernkompetenzen: Die komplementären Kernkompetenzen rechtfertigen die Virtuelle Organisation. Erst der Fokus auf die Kernkompetenzen führt, wie in Abschnitt 5.2.1 aufgezeigt, zur neuen Unternehmensform der Virtuellen Organisation.

Einsatz von IKT: Den Technologieeinsatz könnte man als Haupteigenschaft bezeichnen, ohne den Einsatz von IKT wäre die Virtuelle Organistion gar nicht erst möglich. Die Informations- und Telekommunikationstechnologie ist in diesem Zusammenhang als Enabler zu betrachten.

Die Eigenschaften einer virtuellen Organisation führen zu einer Vielzahl von Schwierigkeiten, die sich im Vergleich zu einer traditionellen Wertschöpfungskette ergeben. Möchte die Car Rental Firma die Services anderen Anbietern im Sinne einer Portalintegration zur Verfügung stellen, so kommt es hierbei zur Schnittstellenproblematik. Die Services müssen in einer Art und Weise zur Verfügung gestellt werden, dass sie von einer, im Voraus unbekannter Anzahl Nachfrager, ausfindig gemacht und verwendet werden können. Eine weitere Schwierigkeit folgt aus der Problematik der Rechnungsstellung. Ein Car Rental-Kunde möchte in der Regel durch eine Sammelrechnung in periodischen Abständen verrechnet werden. Dabei stellt sich in einem verteilten System die Problematik der globalen Sichtweise. Gibt es eine zentrale Verrechnungsstelle oder werden die Services auf der Basis eines kollaborativen P2P-Netzes verrechnet? In der Praxis ergeben sich weitere Probleme, die in dieser Arbeit aus Platzgründen nicht behandelt werden können. Der nachfolgende Abschnitt 5.4 geht mit dem Ziel einer generischen Architektur auf eine Vielzahl auftretender Probleme ein.

5.3.2 Services

Nach den allgemeinen Schilderungen der obigen Teilkapitel zur Notwendigkeit und den Eigenschaften Virtueller Organisationen wird der Fokus in diesem letzten Abschnitt der Einführung auf die Bestandteile Virtueller Organisationen gelegt. Der wichtigste Baustein einer Virtuellen Unternehmung sind Services. Die diversen Teilnehmer der Virtuellen Organisation sind untereinander durch Services 'verbunden'. Man könnte die Services als Infrastruktur oder 'Zement' der Virtuelle Organisation bezeichnen. Sie halten die Organisation zusammen. An den verschiedenen Schnittstellen zwischen den beteiligten Unternehmen müssen Informationen ausgetauscht werden. Damit diese Informationen überall

von allen Teilnehmern verstanden werden, werden Services eingesetzt, welche die Informationen durch automatisierte Verfahren im Notfall in ein unternehmensinternes Format umwandeln. Services können auch aus mehreren anderen Services bestehen (Verschachtelte Services, auch „Service Orchestration“ genannt), und werden so von den Teilnehmern der Virtuellen Organisation angeboten oder durch diese benutzt.

Die folgende Auflistung enthält die Grundvoraussetzungen der Services, welche diese erfüllen müssen, damit eine Virtuelle Organisation entstehen kann. Später werden diese Eigenschaften detaillierter erklärt. Hier geht es in erster Linie um den Überblick.

- Services müssen standardisiert sein: Heterogene Systeme von verschiedenen Unternehmen müssen integriert werden.
- Die QoS (Quality of Services, siehe Abschnitt 5.4.4) muss garantiert werden, sonst kommen keine Virtuelle Organisationen zu Stande. Die Services müssen bestimmte Qualitätsansprüche betreffend Verfügbarkeit und anderer Faktoren gewährleisten, sonst kann die Wertschöpfungskette der Virtuellen Organisation nicht funktionieren.
- Authentizierung und Autorisierung für die Benutzung der Services: Nur berechtigte Teilnehmer dürfen einen Services in Anspruch nehmen, daher kommt der Authentifizierung und Autorisierung eine zentrale Bedeutung zu. Die Thematik wird im Rahmen des Teilabschnitts 5.4.7 wieder aufgenommen.
- Accounting und Billing: Wer hat welche Ansprüche, wer hat was geleistet? Dies sind zentrale Fragestellungen im Rahmen des Accounting und Billing. Die Unternehmen innerhalb der Virtuellen Organisation müssen nach Einsatz nach Gebrauch der Ressourcen bezahlt werden, wobei die Verrechnung für den Kunden nachvollziehbar, transparent sein muss.
- Rechtliche Bedingungen: Die Virtuelle Organisation überschreitet sowohl Unternehmens-, als auch geographische Grenzen. Um die rechtlichen Bedingungen einzuhalten, müssen diese stets verfolgt und deren Einhaltung kontinuierlich überprüft werden. Die Regulations sind eine zentrale Anspruchsgruppe bei der Realisierbarkeitsanalyse zu Beginn des Lebenszyklus einer Virtuellen Organisation. Teilkapitel 5.5.3 greift diesen Aspekt erneut auf.

5.4 Anforderungen an eine virtuelle Organisation

In diesem Abschnitt werden die in den obigen Kapiteln bereits angesprochenen Schwierigkeiten mittels Anforderungen an eine virtuelle Organisation zusammengefasst. Zusätzlich werden neue, bisher noch nicht erwähnte Anforderungen betrachtet und beschrieben. Gemäss [5] muss für die Unterstützung neuer oder veränderter Businessmodelle die Architektur eines verteilten, servicebasierten Systems im Sinne virtueller Organisationen verschiedenen Anforderungen hinsichtlich Sicherheits-, Verrechnungs- und weiterer Mechanismen Rechnung tragen.

Diese Anforderungen bilden die Grundlage für die generische Architektur eines Frameworks einer virtuellen Organisation, wie es in Kapitel 5.5.1 näher beschrieben wird. Die nachfolgenden Anforderungen werden dabei losgelöst von möglichen technischen, konkreten Implementierungen betrachtet.

5.4.1 Integration von heterogenen und dynamischen Umgebungen (Interoperabilität)

Bei einem verteilten System handelt es sich in den meisten Fällen um eine Ansammlung von heterogenen und (geographisch) verteilten Komponenten, basierend auf unterschiedlichen Technologien. Betrachtet man beispielsweise die Server-Komponenten eines verteilten Systems, so können diese mit unterschiedlichen Betriebssystemen betrieben werden (bspw. Unix, Linux, Windows o.ä.), eine Vielzahl von Services mittels unterschiedlichsten Technologien implementieren (bspw. J2EE oder .Net) oder weitere Services von diversen Drittanbietern verwenden. „In addition, Grid environments are frequently intended to be long-lived and dynamic, and may therefore evolve in ways not initially anticipated“ [6]. Das bedeutet, dass der Einsatz eines verteilten Systems über einen längeren Zeitraum hinweg geplant ist und dieses im Falle von Veränderungen angepasst werden muss. Damit sind nicht nur technologische Anpassungen wie z.B. die Aktualisierung von Laufzeitumgebungen sondern auch der Einbezug von bisher unbekannten Neuentwicklungen gemeint.

Die Architektur für ein verteiltes System muss also zwingend die Interoperabilität zwischen solchen unterschiedlichen, heterogenen und verteilten Ressourcen und Services ermöglichen und ebenfalls die Komplexität der Administration solcher heterogener Systeme minimieren. „Interoperabilität ist die Fähigkeit unabhängiger, heterogener Systeme, möglichst nahtlos zusammen zu arbeiten, um Informationen auf effiziente und verwertbare Art und Weise auszutauschen bzw. dem Benutzer zur Verfügung zu stellen, ohne dass dazu gesonderte Absprachen zwischen den Systemen notwendig sind.“ [7]. Schwierigkeiten bei der Sicherstellung der Interoperabilität zwischen verschiedenen Komponenten in einem verteilten System ergeben sich vor allem bei Legacy-Systemen, die bereits wichtige Funktionen bereitstellen. Für die Zusammenarbeit zwischen solchen Systemen ist ein gemeinsamer Standard unumgänglich. „Moreover, many functions required in distributed environments, such as security and resource management, may already be implemented by stable and reliable legacy systems. It will rarely be feasible to replace such legacy systems; instead,we must be able ot integrate them into the Grid“ [6].

Die folgenden Anforderungen für die Unterstützung homogener und dynamischer Systeme lassen sich in Anlehnung an [6] festhalten:

- *Resource virtualisation*: Um die Komplexität des Managementvorgangs heterogener Systeme zu minimieren, ist es wichtig, dass viele sich unterscheidende Ressourcen einheitlich behandelt werden können.
- *Common management capabilities*: Um die Administration von heterogenen Systemen zu vereinfachen, werden Mechanismen für ein einheitliches und konsistentes Management der Ressourcen benötigt. Dazu wird eine gemeinsame Sammlung von Managementfähigkeiten der Ressourcen verlangt.

- *Resource discovery and query*: Um einen Service mit bestimmten und gewünschten Attributen in einem verteilten und heterogenen System ausfindig zu machen, sind Mechanismen und Protokolle notwendig, die das Beschreiben von Services und Ressourcen, sowie die Abfrage dieser Beschreibungen ermöglichen.

- *Standard protocol and schemas*: Um Interoperabilität zu ermöglichen, sind so genannte *high-level* (Standard-)Protokolle nötig, die von den verteilten und heterogenen Ressourcen für die Service- und Ressourcenbeschreibung implementiert und verwendet werden. Die konkrete Implementierung der Protokolle ist dabei für das Gesamtsystem unwichtig, lediglich das gemeinsame Protokoll ist von Bedeutung.

5.4.2 Resource Sharing

In der Praxis gibt es diverse Arten von verteilten Systemen. So sind Systeme innerhalb verschiedener administrativen Einheiten einer grossen, möglicherweise internationalen Unternehmung denkbar. Ebenfalls möglich ist der Zusammenschluss verschiedenster Ressourcen und Services unterschiedlichster Organisationen zu einem vernetzten System. Um den Zusammenschluss der unterschiedlichen Domänen hinsichtlich eines gemeinsamen Zugriffs und Verwendung von Ressourcen zu ermöglichen, sind Mechanismen nötig, die einen Kontext bereitstellen, in welchem Benutzer, Anfragen, Ressourcen etc. auch ausserhalb unternehmerischer Grenzen zusammenkommen und -arbeiten können.

In diesem Zusammenhang kommt den Sicherheitsaspekten und den länderspezifischen Regulierungen grosse Bedeutung zu. Mehr dazu in den Abschnitten 5.4.7 (Sicherheit) und 5.5.3 (Regulations). Die notwendigen Anforderungen an die Architektur eines verteilten Systems hinsichtlich gemeinsamer Nutzung von Ressourcen über Organisationsgrenzen hinweg lassen sich wie folgt zusammenfassen [6]:

- *Global name space*: Globale Namen vereinfachen den Zugriff auf verteilte und potenziell replizierte Ressourcen in einem verteilten System auf einfache und transparente Art und Weise. Globale Namen erleichtern gleichzeitig das Aufsetzen, Überwachen und Einhalten von Sicherheitsrichtlinien.

- *Metadata services*: Metadaten werden benötigt, um verteilte Ressourcen innerhalb des Systems zu finden, aufzurufen und zu verfolgen, ohne auf die Ressourcen selbst zuzugreifen. Die Metadaten enthalten dabei mehr als nur die eigentlichen Informationen über die Ressource selbst, sondern es müssen darin auch weiterführende Informationen über die Sicherheitsbeschränkungen, Service Level Agreements, Verrechnung usw. enthalten sein.

- *Site autonomy*: Jede Domäne soll in der Lage sein, eigene Sicherheitsbestimmungen zu erlassen, die bei einem Zugriff auf diese eingehalten werden müssen. Hierbei wird auch von der *„local control and policy“* gesprochen (siehe auch Kapitel 5.4.7).

- *Resource usage data*: Das *Monitoring, Accounting, Charging und Billing*, ein zentraler ökonomischer Anforderungspunkt eines verteilten Systems, muss ebenfalls in den Anforderungskatalog mit aufgenommen werden. Effizientes Monitoring muss

die Grundlage für die anschliessende Verrechnung der verursachten Kosten an den Nachfrager bilden. Dies ist unter Umständen schwierig, da ein einzelner Service aus mehreren Teilservices zusammengesetzt sein kann. Diese Informationen müssen über die gesamte Hierarchiestruktur eines Services für eine leistungs- und ressourcengerechte Verrechnung gesammelt werden. Schliesslich soll jeder Teilnehmer eines Systems nur für diejenigen Kosten belangt werden, die er auch wirklich verursacht hat. Die Accounting-Thematik des A4C (Authentication, Authorization, Accounting, Auditing and Charging) wird in Teilabschnitt 5.4.11 detaillierter betrachtet, jedoch kann eine ausführliche Sichtweise der Problematik aus Platzgründen nicht in die Arbeit mit aufgenommen werden. Weiterführenden Informationen finden sich in [8].

5.4.3 Optimierungen

Im technisch motivierten Anspruch der Optimierungen geht es um die effiziente und optimale Anordnung der Ressourcen sowohl aus Kunden- als auch aus Anbietersicht.

Aus Anbietersicht beinhaltet dies die Optimierung der Ressourcen. Oftmals wird vom *worst case*-Szenario ausgegangen, in welchem die Ressourcen an die maximale Nachfrage angepasst werden. Dies führt unweigerlich zu Überversorgung während der meisten Zeit und verursacht durch Nichtbenutzung dieser Ressourcen unnötige Fixkosten. Dies gilt es aus Sicht der Anbieter zu optimieren. Hierbei handelt es sich um einen Trade-off, da die Anbieter auch genügend Ressourcenreserve besitzen müssen, um die vereinbarten SLA's einzuhalten. Eine mögliche Strategie ist hierbei die flexible Anordnung von zusätzlichen Ressourcen zu Spitzenzeiten (so genannte *peaks*).

Aus Kundensicht ist die Optimierung verschiedener Typen der Lastnachfrage, insbesondere die aggregierte Workloadnachfrage entscheidend, wenn ein Service durch mehrere verschachtelte Services repräsentiert wird. Die Bestimmung der gesamten nachgefragten Ressourcen ist dabei aber generell schwierig. „An important requirement in this area is the ability to dynamically adjust workload priorities in order to meet the overall service level objectives. Mechanisms for tracking resource utilization, including metering, monitoring and logging; for changing resource allocation; and for provisioning resources on-demand are the required foundation of demand-side optimization."[6]. Dieses Zitat beschreibt die Notwendigkeit, in Abhängigkeit der aktuellen Lastverteilung die Priorisierung einzelner Teilservices zu verändern mit dem Ziel der kleinstmöglichen Bearbeitungszeit aller Teilservices eines zusammengesetzten Services. Dafür und zu Verrechnungszwecken müssen die Teilservices effizient überwacht (Tracking) und geloggt werden.

5.4.4 Quality of Service (QoS)

Services wie die Bereitstellung und das Erbringen einer Dienstleistung müssen immer einen vor der Leistungserbringung getroffenen Qualitätslevel erreichen und garantieren. Die Hauptbestandteile eines solchen SLA (Service Level Agreement) sind Verfügbarkeit,

Sicherheit und Durchführung (*performance*). „Das SLA beschreibt die IT-Services in nichttechnischen Begriffen und ist so auch für technisch ungeschulte Kunden verständlich. Für die Dauer der Vereinbarung gilt es als Vertrag in Bezug auf die Leistungserbringung und Steuerung der IT-Services. SLAs können servicebasiert (ein SLA für einen Service) oder kundenspezifisch (ein SLA für alle Services eines Kunden) vereinbart werden.“ [9]. SLAs spielen eine wichtige Rolle, weiss doch ein Nachfrager vor dem Bezug eines Services nichts über dessen Güte. Hierbei wird auch von *asymmetrischer Information* gesprochen. In der Organisationstheorie wird die Thematik der asymmetrischen Informationsverteilung im Rahmen der Principal-Agent-Theorie behandelt. „In Principal-Agent-Beziehungen eröffnen sich diskretionäre Verhaltensspielräume für nicht vollständig kontrollierte Agents, die diese - ungleiche Interessen zwischen Principal und Agent vorausgesetzt - zu ihrem eigenen Vorteil und zum Schaden der Principals ausnützen können bzw. werden.“ [10]. Aus den verschiedenen Principal-Agent-Beziehungen lassen sich drei Organisationsprobleme ableiten. Bei der *Adverse Selection* sind die Qualitätseigenschaften der Leistung des Vertragspartners unbekannt, beim *Moral Hazard* sind die Anstrengungen des Vertragspartners nicht beobachtbar bzw. nicht beurteilbar und beim *Hold up* ist die Vollständigkeit von Verträgen nicht realisierbar. Um diese Informationsasymmetrien und die damit verbundenen Unsicherheiten des Nachfragers (Agent) zu mindern, verpflichtet sich der Anbieter (Principal) über die Definition eines SLA zur Erbringung eines Services mit festgesetzter Güte. Bei Nichteinhalten des Qualitätsgüte sind Strafen - meist monetär - fällig. Die Grundlage für ein funktionierendes System aus SLA bilden ein lückenloses Monitoring (Metering) und Accounting der Daten, deren Korrektheit mittels Auditing-Verfahren überprüft und sichergestellt werden kann. Eine ausführliche Behandlung des Principal-Agent-Problems, mögliche Einflussgrössen und Problembegrenzungen finden sich in [10].

Die Anforderungen hinsichtlich Qualitätsgüte in homogenen und dynamischen Systemen lassen sich wie folgt summieren [6]:

- *Service level agreement (SLA)*: Qualitätsgüte sollte durch Vereinbarungen festgehalten werden, die vor der eigentlichen Leistungserbringung zwischen dem Service Provider und dem Service Requestor (Nachfrager) ausgehandelt und getroffen werden. Dazu werden Standardmechanismen benötigt, welche das Aushandeln, Festsetzen und die Kontrolle solcher Vereinbarungen vereinfachen.

- *Service level attainment*: Falls das zwischen den beiden Parteien getroffene Arrangement eine *erwartete* Servicegüte beinhaltet, dann sollten die für den Service benötigten Ressourcen in der Art angeordnet, reserviert etc. werden, damit die erwartete Qualitätsgüte erreicht wird. Dazu sind Mechanismen zur Überwachung der Qualitätsgüte, zur Schätzung des Ressourcenbedarfs und zur Planung des Ressourceneinsatzes nötig.

- *Migration*: Die beiden obigen Punkte erfordern, dass ein sich in Ausführung befindlicher Service ausgelagert, sprich verschoben (*migrate*) werden kann, um die vereinbarte Servicegüte zu garantieren.

Die Definition eines SLA folgt dem SMART-Prinzip. „SMART ist ein Akronym für Specific Measureable Achievable Relevant Timely und dient im Projektmanagement zur eindeutigen Definition von Zielen.“ [11]. Die Begriffe lassen sich wie folgt kurz umschreiben:

- *Specific*: unmissverständliche und eindeutige Zieldefinition.
- *Measureable*: Definieren von Kriterien um sicherzustellen, dass das Erreichen des Ziels messbar ist.
- *Achievable*: Die Ziele sollen erreichbar sein.
- *Relevant*: Nur Ziele von hoher Bedeutung werden gesetzt.
- *Timely*: Zu jedem Ziel gehört eine klare Terminvorgabe.

5.4.5 Aufgabenmanagement

Das Aufgabenmanagement soll das Management der Services während der gesamten Lebensdauer eines Services, von dessen Initialisierung über die Durchführung bis hin zur Verrechnung, übernehmen. Zum Aufgabenmanagement gehören daher Funktionen wie Warteschlangenmanagement, Verrechnung, Fehlerbehandlung und weitere. Besondere Wichtigkeit erlangen diese Funktionen, falls sich eine Aufgabe aus vielen Teilaufgaben, die sich über mehrere Domänen erstrecken, zusammensetzt. Die Komplexität des gesamten Aufgabenmanagements nimmt dabei im Vergleich zu einem „normalen“ Auftrag, wo nur eine Domäne für die Abwicklung des Auftrags beteiligt ist, markant zu.

Gemäss [6] lassen sich die Anforderungen an das Aufgabenmanagement, besonders im Hinblick auf verschachtelte Services, wie folgt zusammenfassen:

- *Support for various job types*: Neben einfachen Aufgaben, die nur in einer Domäne ausgeführt werden, müssen auch komplexe und zusammengesetzte Jobs unterstützt werden, welche in mehrere Teilaufgaben zerlegt und auf unterschiedlichen und unabhängigen Domänen zur Ausführung verteilt werden. Des Weiteren sollen die Aufgaben ebenfalls im Sinne eines *Workflows* geplant und zusammengesetzt werden können.
- *Job management*: Dem Management eines Jobs kommt eine besondere Bedeutung zu. So ist die Arbeit des Job Managements nicht damit erledigt, einen Service zu starten. Vielmehr müssen generische Schnittstellen vorhanden sein, über welche während der Ausführung eines Jobs beeinflussend auf diesen zugegriffen werden kann. Die Anforderungen an die Schnittstelle können an dieser Stelle nicht weiter verfeinert werden, jedoch muss diese den Zugriff auf komplexe, (mehrfach) verschachtelte Aufträge erlauben (Workflows, Jobmatrix). Neben dem reinen Management während der Durchführung eines Services müssen auch Strukturen für die Überwachung der Aufträge, Möglichkeiten der individuellen Auftragszusammensetzung und dessen (sequentielle) Ablaufplanung vorhanden sein.
- *Scheduling*: Das *Job Management* (siehe Punkt oberhalb) erlaubt eine Aufsplittung eines Auftrags auf mehrere Teilaufträge und dessen sequentielle Ausführung. Daneben soll ein Auftrag zusätzlich mit einer Priorität versehen werden können, die es erlaubt, (Teil-)Aufträge nach deren Dringlichkeit zu bearbeiten. Besonderes Augenmerk gilt den zusammengesetzten Aufträgen, die mit hoher Priorität in einer

Domäne gestartet werden. Es muss nach dem Start über alle weiteren betroffenen Domänen sichergestellt werden, dass die Teilaufgaben in diesen Domänen ebenfalls mit der höchsten oder zumindest mit erhöhter Priorität durchgeführt werden. Dies führt zu komplexen Situationen mit mehreren betroffenen Priorisierungsinstanzen in mehreren administrativen Domänen.

- *Resource provisioning*: „To automate the complicated process of *resource allocation, deployment, and configuration*, it must be possible to deploy the required applications and data to resources and configure them automatically...“ [6]. Das Zitat umschreibt die Notwendigkeit eines automatisierten Verfahrens für das Anpassen der Laufzeitumgebung einer Domäne. Dazu gehören unter anderem Anpassungen des Betriebssystems der Hosting-Umgebung (z.B. Installation neuer Patches oder Zusatzmodulen), die für die korrekte Ausführung der Aufgabe benötigt werden. Damit sind nicht nur Anpassungen zur Durchführung von Rechenaufgaben gemeint, sondern auch solche für die Bereitstellung von Daten- und Netzwerkdiensten.

5.4.6 Datenorganisation und -Services

Trotz immer weiter steigenden Bandbreiten für den Datentransfer bei gleichzeitig sinkenden Preisen, müssen immer grössere Datenmengen bewegt werden. Auch verlangen immer mehr technologische Felder schnellen und effizienten Zugriff auf grosse Datenmengen. Zusätzlich wird der Zugriff auf Informationen, die an unterschiedlichen Orten, sprich in unterschiedlichen administrativen Domänen, gespeichert sind, wichtiger. Gemäss [6] sind vor allem für die *Business areas* das Management und die Archivierung von grossen Datenmengen essentielle Anforderungen. So sind beispielsweise die für einen auf mehrere Domänen verteilten Gesamtprozess Daten notwendig, die an völlig anderen Orten innerhalb des verteilten Systems gespeichert sind. Es sind also nicht nur Verfahren zum Auffinden von verfügbaren Services nötig, sondern auch solche für die Lokalisierung der benötigten Daten.

In [6] sind die Anforderungen an Datenorganisation und -Services folgendermassen zusammengefasst:

- *Datenzugriff*: Für den effizienten Zugriff auf eine Vielzahl von unterschiedlichen Datenbeständen (bspw. relationale Datenbanken, Dateien, Streams etc.) wird eine Sicht auf die Daten benötigt, die von der Art der darunter liegenden physischen Speicherung und vom Ort der Speicherung abstrahiert.

- *Datenkonsistenz*: Die Zugriffsschicht auf die Datenbestände muss sicherstellen, dass die Datenkonsistenz auch dann sichergestellt ist, wenn Daten zur Aufgabenausführung aus dem Datenspeicher ausgelesen und auf eine andere Domäne transferiert werden. Dabei spielt es keine Rolle, ob die Originaldaten oder die replizierten Daten modifiziert werden.

- *Datenpersistenz*: Daten und die dazugehörigen Metadaten müssen über längere Zeit gepflegt und aufbewahrt werden. Die Dauer der Aufbewahrung wird meist durch regulatorische Vorschriften bestimmt.

- *Datenintegration*: Wie bereits angesprochen, sind die Daten in einem verteilten System meist über mehrere Domänen verteilt und in unterschiedlichsten Datenformaten verfügbar. Es werden daher effiziente Mechanismen benötigt, welche die Suche in solchen verteilten, heterogenen und zusammengeschalteten Datenbeständen erstens ermöglichen und zweitens auf einfache und einheitliche Art und Weise erlauben.

- *Datenlokalisierung*: Neben dem ausfindig machen von gesuchten Datenbeständen in einem System kommt mit der Lokalisierung der Daten eine weitere Anforderung hinzu. Nachdem die gesuchten Daten und deren Speicherort bestimmt ist, müssen sie eventuell aus ihrem aktuellen Speicherort heraus kopiert und lokal in die nachfragende Domäne kopiert werden.

5.4.7 Sicherheit

Die Bereitstellung von Services in einer virtuellen Organisation verlangt strenge Authentifikations- und Autorisationsmechanismen für den kontrollierten Zugriff auf die Services. Es muss sichergestellt werden, dass nur berechtigte Benutzer auf Services zugreifen können. Der kontrollierte Zugriff bildet die Grundlage für die leistungsgerechte Verrechnung der Services. Nur falls zu jedem Zeitpunkt die Identifikation eines Benutzers eindeutig bestimmt werden kann, ist es möglich, ihm die verursachten Kosten zu belasten.

Daneben spielen weitere Faktoren, wie sie in *„Talk1: AAA Support for Multicast Services“* aufgezeigt wurden, eine gewichtige Rolle. Nach [6] lassen sich die folgenden Punkte festhalten:

- *Authentication and authorization*: Authentifikationsmechanismen werden benötigt, um die Identität von Benutzern und Services zu garantieren. Weiter müssen Service Provider Autorisierungsfunktionen implementieren, welche dafür sorgen, dass die Richtlinien zur Benutzung von Informationen und Services überwacht und eingehalten werden.

- *Multiple security infrastructures*: Eine über mehrere Domänen verteilte (Gesamt-) Operation verlangt den Umgang mit mehreren Sicherheitsumgebungen. Eine möglichst breite Unterstützung der gängigen Sicherheitskonzepten ist nötig. „The Grid system should follow each domain's security policies and also may have to identify users' security policies. Authorization should accomodate various access control models and implementations.“ [6]. Schwierigkeiten treten hierbei vor allem in denjenigen Fällen auf, in welchen sich ein Service eines weiteren Services bedient. Dabei müssen die security policies des Kunden weitergereicht und überprüft werden, was die Aufgabe der Delegationsfunktion ist (siehe Punkt weiter unten).

- *Perimeter security solutions*: Mit grosser Wahrscheinlichkeit müssen bei der Ausführung eines Task in einem verteilten Systems verschiedene administrative Domänen, auch über organisationale Grenzen hinweg, zusammenarbeiten. Dazu müssen Sicherheitsmechanismen und -Standards implementiert sein, die gleichzeitig sowohl die einzelne Organisation schützen als auch die interorganisatorische Kommunikation

und Interaktion erlauben, bei gleichzeitigem Aufrechterhalten der lokalen Sicherheitsrichtlinien (z.B. Firewall- und Intrusion Detection).

- *Delegation*: Zusätzlich zu den bisherigen Anforderungen werden Funktionen verlangt, welche den Transfer von Sicherheitsrichtlinien zwischen Domänen ermöglichen. Die Rede ist hierbei vom Transfer der Berechtigungen zwischen dem Nachfrager und dem Anbieter eines Services. Ist der Service aus mehreren weiteren Services zusammengesetzt, so benötigt der für die Zusammensetzung der Teilservices verantwortliche Service Provider die Berechtigungen des Nachfragers. Daraus ergeben sich natürlich potentielle Sicherheitsrisiken, die es zu minimieren gilt. Eine mögliche Restriktion ist die zeitliche Beschränkung von transferierten Berechtigungen.
- *Security policy exchange*: Um die komplette und korrekte Ausführung eines Jobs sowohl aus der Sicht des Anbieters als auch aus der Sicht des Nachfragers zu gewährleisten, müssen die beiden beteiligten Parteien währen der „Verhandlungsphase“ vor dem Ausführen eines Task die Sicherheitsinformationen austauschen. Auf der einen Seite muss der Service Provider dem Nachfrager *beweisen*, dass er die vom ihm geforderten Sicherheitsrichtlinien einhält. Auf der anderen Seite muss sich der Nachfrager eindeutig gegenüber dem Service Provider identifizieren können. Diese Mechanismen gelten nicht nur für die erste Stufe eines Services, wo ein Nachfrager einen Service nachfragt, sondern ebenfalls auf den darunter liegenden Hierarchien, wo ein Service Provider als Manager diverser Teilprozesse einen Service bei einem anderen Service Provider nachfragt.
- *Intrusion detection, protection and secure logging*: Effizientes und sicheres Logging muss die Grundlage der Intrusion Detection und der Missbrauchsentdeckung bilden. Neben den Gefahren des Missbrauchs muss das gesamte System und jede einzelne Domäne ebenfalls gegen die „üblichen“ Gefahren wie Viren und Würmer geschützt werden.

5.4.8 Skalierbarkeit

Neben vielen Vorteilen, die ein verteiltes System eröffnet und dadurch neue Arten von Dienstleistungen ermöglicht, desto schwieriger ist die Skalierung tausender unterschiedlichster Ressourcen in einem solchen System. [6] empfiehlt ein Management dieser Ressourcen auf einer gemeinsamen Peer-to-Peer-Basis. Das Einrichten einer zentralen Stelle mit einer globalen Sicht auf das System zu Managementzwecken vernichtet viele Vorteile der Idee verteilter Ressourcen, da eine zentrale Stelle die Struktur des Systems gefährdet. Im Falle eines Ausfalls dieser zentralen Stelle ist das Gesamtsystem nicht mehr lauffähig. Also muss das Management auf eine gemeinsame und kollaborative Art und Weise durch Selbstorganisation der Ressourcen geschehen.

Weiter rät [6] zum Einsatz so genannter *High-throughput computing mechanisms* [12] für die Optimierung paralleler Aufträge in einem verteilten System mit dem Ziel der Steigerung der globalen Durchsatzzahl. An dieser Stelle kann nicht weiter auf die technischen Aspekte verteilter Systeme eingegangen werden, aber das folgende Statement verdeutlicht die Problematik: „Ein System ist nur immer so leistungsfähig wie seine schwächste Komponente...“.

5.4.9 Verfügbarkeit

Üblicherweise wird hohe Verfügbarkeit durch teure, fehlertolerante Hardware oder komplexe Clustersysteme erreicht. In einem verteilten System ist der transparente Zugang zu einem breiten Ressourcenpool, sowohl in als auch zwischen Organisationen, möglich. Dadurch können stabile und zuverlässige Systeme aufgebaut werden. Fällt eine spezifische Ressource innerhalb eines Systemverbundes aus, so stehen weitere Ressourcen innerhalb des Systems zur Verfügung, die nach potentiellen, automatisierten Umgebungsanpassungen die gleiche Aufgabe wie das ausgefallene System übernehmen können. Es spielt in diesem Zusammenhang keine Rolle, welche Infrastruktur dieses System besitzt, da die Architektur des Frameworks keine Aussagen über die zu Grunde liegende Infrastruktur macht.

Das *Disaster Recovery* muss Verfahren bereit stellen, die einen Vorgang in einem verteilten System rückgängig machen können, ohne dabei das System über längere Zeit hinweg ausser Betrieb zu setzen. Unabdingbares Hilfsmittel hierfür sind automatisierte und einfache Backups und Wiederherstellungsprozeduren.

Das *Fault management* muss seinerseits Methoden anbieten, die sicherstellen, dass aktuell aktive Prozesse nicht aufgrund mangelnder oder fehlerhafter Ressourcen verloren gehen, wie in [6] wie folgt aufgeführt wird: „Mechanisms are required for monitoring, fault detection, and diagnosis of causes or impacts on running jobs. In addition, automation of fault-handling, using techniques such as checkpoint recovery, is desirable“.

5.4.10 Benutz- und Erweiterbarkeit

Für den Benutzer der Services ist es entscheidend, dass ihm die Komplexität des zugrunde liegenden Systems verborgen bleibt. „As much as possible, tools, acting in convert with run-time facilities, must manage the environment for the user and provide useful abstractions at the desired level“ [6]. In diesem Zusammenhang darf man jedoch die so genannten „power user“ nicht vergessen, die sich durch ein vertieftes technisches Know-how auszeichnen und meist den vollen Zugriff auf das System, also ohne Maskierung oder Abstraktion, wünschen. Daher soll es dem Benutzer überlassen bleiben, mit welcher Schicht des Systems er kommunizieren möchte.

Es ist unmöglich, alle Bedürfnisse und Anforderungen der Benutzer an ein System und im Speziellen an die Services im voraus zu erkennen und bestimmen. Aus diesem Grund soll das System aus modularen Komponenten gebildet werden, die jederzeit erweitert, ausgetauscht oder entfernt werden können, um so veränderten Anforderungen gerecht zu werden. Neben der Ersetzbarkeit der Komponenten durch die modulare Bauweise erlaubt das Bausteinprinzip die Implementierung zusätzlicher Module durch einzelne Domänen, durch so genannte Drittanbieter.

5.4.11 Accounting

Die Anforderungen an das Accounting in einem verteilen System bezüglich der Verrechnung von Services muss vereinfacht zwei Aufgaben erfüllen.

Erstens muss das Accounting alle Daten über die gebrauchten Ressourcen sammeln und gegebenfalls aufbereiten. Dafür kann in den meisten Fällen auf die Monitoring-Daten zurückgegriffen werden, die im Rahmen der gemeinsamen Nutzung von Ressourcen anfallen (siehe Abschnitt 5.4.2). Es gibt zwei Möglichkeiten, wie die Monitoring-Daten für das Accounting verarbeitet werden können. Denkbar ist eine zentrale Stelle, wo alle Daten zusammenlaufen und die für die Verrechnung aller benützter Services in einem System verantwortlich ist. Die andere (und bevorzugte) Variante ist die Weiterleitung der Monitoring-Daten an den Aufrufer des Services, der selbst wiederum diese Daten an seinen Aufrufer weiterleitet oder aber für die Verrechnung verwendet. Zusätzlich sind weitere Entscheidungen zu treffen, in welchem Rhythmus die Verrechnung erfolgen soll. Hier wird zwischen periodischer und just-in-time Verrechnung unterschieden.

Zweitens ist das Accounting für die persistente Speicherung der Accounting-Daten zu Archivierungszwecken verantwortlich. Dem Benutzer soll die Möglichkeit geboten werden, jederzeit Informationen über den bisherigen Verbrauch beim Anbieter anzufordern. In vielen Geschäftsbereichen ist die Dauer und der Detailgrad der Datenaufbewahrung durch gesetzliche Regulatorien festgelegt (z.B. Verbindungsdaten in der Telekommunikationsbranche).

5.5 Anforderungsumsetzung mittels einer Grid-Architektur

Nachdem der Fokus des letzten Kapitels auf der Beschreibung der Anforderungen an eine Virtuelle Organisation lag, so wird in diesem Kapitel auf die Umsetzung der gestellten Anforderungen anhand eines Frameworks fokussiert. Die Anforderungsumsetzung erfolgt dabei anhand des OGSA (Open Grid Services Architecture). Die Strukturierung des Kapitels ist dabei [6] entnommen. Es ist festzuhalten, dass die Anforderungen aus dem vorangehenden Kapitel in der OGSA nicht 1:1 umgesetzt werden. Es existiert also nicht für jedes Unterkapitel aus 5.4 ein Unterkapitel.

5.5.1 OGSA (Open Grid Services Architecture)

OGSA ist eine Spezifikation für eine serviceorientierte Grid-Computing-Umgebung für kommerzielle und wissenschaftliche Zwecke. Sie soll den Gebrauch und das Management von verteilten und heterogenen Ressourcen vereinfachen [6]. Sie wurde innerhalb des Global Grid Forum entwickelt. Das Global Grid Forum will die Standardisierung im Grid Computing vorantreiben. OGSA basiert auf verschiedenen Webservice-Technologien, die Infrastruktur wird im Abschnitt „Infrastructure Services“ ausführlicher behandelt. Eine

Schwachstelle der OGSA ist das Accounting, Pricing and Charging, welches nicht behandelt wird. Im Verlaufe der Arbeit wurde kein anderes Framework gefunden, das die drei Punkte ausführlich behandelt.

Übersicht

Abbildung 5.5 zeigt die logische und abstrakte Repräsentation der in Kapitel 5.4 beschriebenen Anforderungen und bringt diese mit der Grid-Infrastruktur in Zusammenhang.

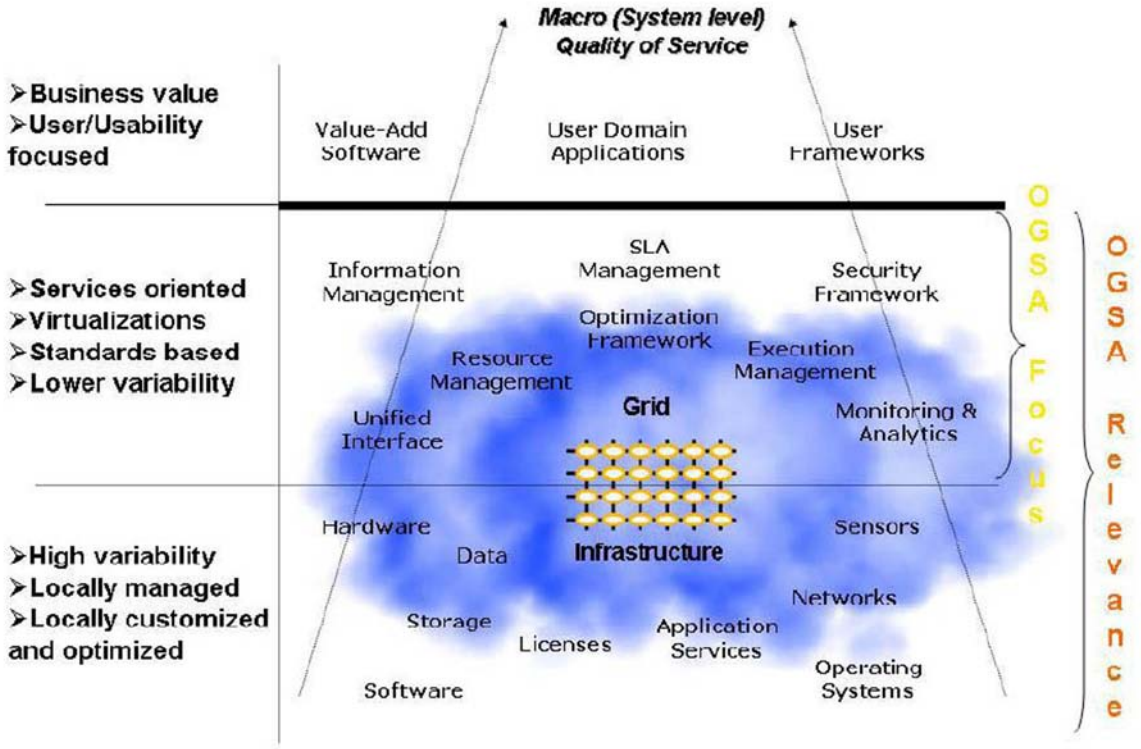


Abbildung 5.5: Konzeptionelle Grid-Infrastruktur [6]

Der untere Teil in der Abbildung 5.5 beschreibt die *Basisressourcen (engl. "base resources")*. "Base resources are those resources that are supported by some underlying entitites or artefacts that may be physical or logical, and that have relevance outside the OGSA context" [6]. Beispiele für Basisressourcen sind diesem Zitat zufolge beispielsweise Arbeitsspeicher, CPU oder Festplatten, aber auch einzelne Prozesse, Lizenzen etc..

Im mittleren Teil der Abbildung 5.5 sind die Fähigkeiten, wie sie im letzten Kapitel behandelt wurden ersichtlich und verwenden Basisressourcen der darunter liegenden Schicht, um ihre Funktionalität bereitzustellen. An dieser Stelle muss angemerkt werden, dass die die Basisressourcen von der mittleren Schicht als so genannte *peers* wahrgenommen werden, und nicht als Ressourcen im eigentlichen Sinn, sondern als abstrakt als eine Art Ressourcenservice.

Auf der obersten Schicht in Abbildung 5.5 befindet sich die logischer Darstellung der Anwendungen oder anderen Entitäten, welche die Fähigkeiten des OGSA-Frameworks in

Form von Services in Anspruch nehmen. Obwohl diese Schicht nicht mehr direkt zum OGSA-Framework gehört, ist sie trotzdem relevant, da sich die Anforderungen an das OGSA-Framework (siehe Kapitel 5.4) primär von diesen Anwendungen ableiten. Das OGSA-Framework richtet sich nach den Benutzeranforderungen und ist nicht an der Technik orientiert.

Diese drei kurz umrissenen Schichten müssen in der Praxis zusammenarbeiten um die Services mit einer bestimmten und verlangen Qualitätsgüte anbieten zu können. Dieser Zusammenhang der Interoperabilität der diversen Basisressourcen ist durch die vertikal verlaufenden Pfeile symbolisiert und wird als *Macro Quality of Service* bezeichnet.

Framework

Die in Abbildung 5.5 gezeigte mittlere Schicht wird vom OGSA-Framework durch Services realisiert, die durch Schnittstellen beschrieben werden. Für die individuellen und durch mehrere Services zusammengesetzten Zustände dieser Services und zu Grunde liegenden Basisressourcen wird das Konzept *SOA (Service-oriented Architecture)* verwendet. "Eine serviceorientierte Architektur ist ein Konzept für eine Systemarchitektur, in welchem Funktionen in Form von wieder verwendbaren, voneinander unabhängigen und lose gekoppelten Services implementiert werden. Services können unabhängig von zugrunde liegenden Implementierungen über Schnittstellen aufgerufen werden, deren Spezifikationen öffentlich und damit vertrauenswürdig sind. Serviceinteraktion findet über eine dafür vorgesehene Kommunikationsinfrastruktur statt. Mit einer serviceorientierten Architektur werden i. d. R. die Gestaltungsziele der Geschäftsprozessorientierung, der Wandlungsfähigkeit (Flexibilität), der Wiederverwendbarkeit und der Unterstützung verteilter Softwaresysteme verbunden."

Infrastructure Services

Wie oben bereits erwähnt, basiert OGSA auf verschiedenen Webservice-Technologien. Es sind dies vor allem WSDL (Web Services Description Language), XML (Extensible Markup Language) und SOAP (Simple Object Access Protocol). WSDL dient der Schnittstellenbeschreibung, XML der Beschreibung und Repräsentation und SOAP als primäres Format für den Austausch von Nachrichten. Nicht alle Anforderungen, die an ein verteiltes System gestellt werden, können mit den existierenden Standards erreicht werden. Modifikationen sind deshalb nötig, weshalb bei der Definition von WSDL 2.0 zum Beispiel OGSA-Architekten mitwirkten. Die beiden wichtigsten Bereiche, in denen Modifikationen nötig sind, sind Sicherheit und Zustandsrepräsentation und -manipulation [14].

Execution Management Services

Die Execution Management Services (EMS) beschäftigen sich mit dem Problem der Instanziierung und dem Management von Arbeitseinheiten. Es werden hier die Anforderungen aus Kapitel 5.4.5 umgesetzt. Die EMS müssen nach [6] folgende Aufgaben erfüllen:

- *Finden der Execution Candidate Locations*: Hier wird das Problem gelöst, wo die Speicherstellen sind, wo eine Arbeitseinheit, unter Berücksichtigung der Restriktionen betreffend der Ressourcen, beispielsweise CPU oder verfügbare Lizenzen, ausgeführt werden kann. Weiter wird geprüft, ob weitere einschränkende Richtlinien vorhanden sind.
- *Auswahl Execution Location*: Aus den möglichen Locations für die Ausführung muss die optimale ausgewählt werden. Dies geschieht mittels verschiedener Auswahl-Algorithmen, die auf die Optimierung verschiedener Zielfunktionen ausgelegt sind.
- *Vorbereitung für die Ausführung*: Ohne Vorbereitung kann eine Arbeitseinheit nicht ausgeführt werden. Es braucht beispielsweise die Bereitstellung und Konfiguration von Binaries und Bibliotheken.
- *Initiierung der Ausführung*: Wenn alles wie oben beschrieben bereit ist, muss die Ausführung gestartet werden.
- *Management der Ausführung*: Wenn die Ausführung gestartet ist, muss sie bis zur Beendigung gemanaged und überwacht werden. Hier wird festgelegt wie mit Fehler umgegangen wird. Oder was zu tun ist, wenn SLAs nicht erreicht werden, etc..

Die EMS sind äusserst wichtig, da in einem verteilten System sehr viele verschiedene Zustände auftreten können und die Ressourcen beziehungsweise deren Auslastung grossen Schwankungen unterworfen sind, aber trotzdem sicher verlässlich verfügbar sein müssen.

Zur Erledigung der oben beschriebenen Aufgaben wird folgendermassen vorgegangen: Verschiedene Dienste zerlegen das Problem der Instanziierung und des Managements von Arbeitseinheiten in mehrere, austauschbare Komponenten. Zur Veranschaulichung können die Execution Management Services in Angebots- und Nachfrageseite aufgeteilt werden. Die Angebotsseite stellt Ressourcen zur Verfügung, die Nachfrageseite verbraucht sie. EMS-Dienste ermöglichen den Applikationen, eben der Nachfrageseite, den aufeinander abgestimmten Zugang zu Ressourcen, egal wo sich diese physisch befinden und wie die Zugangsmechanismen ausgestaltet sind. Daraus folgt, dass Protokollunabhängigkeit und Schnittstellenkonformität notwendig sind [15]. Die EMS-Dienste können in drei Klassen eingeteilt werden. Es sind dies:

- Ressourcen
- Aufgabenmanagement- und -überwachungsdienste
- Ressourcenauswahldienste

Data Services

Data Services werden gebraucht, um Daten dorthin zu verschieben, wo sie benötigt werden, um replizierte Kopien zu managen, Abfragen und Aktualisierungen auszuführen und Daten in neue Datenformate umzuformen. Sie haben weiter die Möglichkeit, Metadaten, die die OGSA Data Services [6] beschreiben, zu managen.

Aufgrund der heterogenen Gegebenheiten eines verteilten Systems müssen viele verschiedene Datenformate unterstützt werden. Die wichtigsten sind:

- *Flat Files*: Flat Files sind die einfachste Form von Daten und haben eine applikationsspezifische Struktur. Diese Dateien können mit normalen Schreib- und Leseoperationen verändert werden.
- *Streams*: Streams sind potenziell unendliche Sequenzen von Datenwerten. Die Datendienste unterstützen Abfragen und Transformationen, die auf Streams angewendet werden.
- *DBMS*: Verschiedene Arten von Datenbank Management Systemen können Teil eines verteilten Systems sein. Unter anderen sind relationale, XML oder objektorientierte Datenbanken möglich.
- *Catalogues*: Ein Katalog strukturiert und findet andere Datendienste. Ein einfaches Beispiel eines Katalogs ist ein Verzeichnis, in dem Dateien aufgeführt sind.
- *Derivations*: Daten können das Resultat von asynchronen Abfragen oder Transformationen auf anderen Daten sein. Diese Derivations werden häufig als Streams behandelt.
- *Data Services*: Können selbst Datenressourcen für andere Dienste sein.

Folgende funktionellen Fähigkeiten, von denen nicht alle zwingend implementiert werden müssen, stellen die OGSA Data Services zur Verfügung:

- *Transparency und Virtualization*: Die OGSA Data Services können Virtualisierungen der, aufgrund der oben erwähnten Heterogenität eines verteilten Systems, vielen verschiedenen Datenformaten definieren. Virtualisierungen sind abstrakte Sichten auf die Daten, die es erlauben, die Daten zu bearbeiten, ohne dass die Unterschiede betreffend den Datenformaten dem Bearbeiter ersichtlich sind. Weiter muss es aber möglich sein, dass Clients die Virtualisierungen umgehen und direkt auf die ursprünglichen Schnittstellen zugreifen können. Mit dieser funktionelle Fähigkeit werden die Anforderungen aus Kapiel 5.4.10 umgesetzt.
- *Client APIs*: Es muss möglich sein, die OGSA Data Services mit den APIs bereits bestehender Applikationen zu benutzen, da es in vielen Fällen zu teuer oder unmöglich ist, die APIs neu zu schreiben.
- *Extensible data type support and operation*: Da in Zukunft neue, zur Zeit noch unbekannte Datenressourcen, mit denen die OGSA Data Services verwendet werden, hinzukommen werden, ist es zwingend notwendig, dass zusätzliche Data Services hinzugefügt werden können, um die neuen Ressourcen zu benutzen. Das Gleiche gilt für neue Operationen, die auf Ressourcen ausgeführt werden.
- *Data Location Management*: Die OGSA Data Services stellen den verlässlichen Transfer von Daten von einem Ort zum anderen sicher.

- *Queries (Structured Access)*: Von den OGSA Data Services müssen Mechanismen bereitgestellt werden, damit Abfragen auf strukturierten Datenressourcen möglich sind. Beispielsweise eine SQL-Abfrage über eine relationale Datenbank.

- *Transformation*: Die Data Services müssen Daten von einem Format in ein anderes umwandeln können.

- *Data Update*: Durch die OGSA Data Services soll für die verschiedenen Datentypen eine Vielzahl von Mechanismen bereithalten, um Updates der Datenressourcen durchführen zu können.

- *Security Mapping Extensions*: DBMS verfügen meist über ausgeklügelte Sicherheitsmechanismen, die Data Services stellen sicher, dass die OGSA-Sicherheitsinfrastruktur mit den Sicherheitsmechanismen der DBMS erweitert werden kann.

- *Data Resource Configuration*: Die OGSA Data Services sollen die den Datenressourcen eigenen Optionen zur Konfiguration zugänglich machen.

- *Provenance*: Zusammen mit den Metadaten werden Informationen betreffend des Ursprungs und der Qualität von Datenressourcen zur Verfügung gestellt. Informationen über den Ursprung von Daten können deren Wiederherstellung ermöglichen. Damit werden die Anforderungen aus Kapitel 5.4.9 umgesetzt.

Für die OGSA-Data Services sind die folgenden Eigenschaften wichtig und definiert: Scalability, Quality of Service, Coherency, Performance, Availability und Legal and Ethical Restrictions.

Resource Management Services

In einem OGSA-Grid gibt es drei Arten des Management von Ressourcen:

- Management der Ressourcen selbst, zum Beispiel der Neustart eines Rechners. Dies geschieht auf dem „Resource level“ und dem „Infrastructure level“.

- Management der Ressourcen in einem verteilten System, zum Beispiel die Reservation einer Ressource oder deren Überwachung. Dies geschieht auf dem „OGSA functions level“.

- Management der OGSA-Infrastruktur, die selbst aus Ressourcen besteht. Zum Beispiel die Überwachung eines Services. Diese Aufgaben sind ebenfalls auf dem „OGSA functions level“ angesiedelt.

Für die OGSA-Resource Management Services sind die folgenden Eigenschaften wichtig und definiert: Scalability, Interoperability, Security und Reliability.

Security Services

Die Security Services sollen nach [6] die Durchsetzung von Richtlinien einer virtuellen Organisation betreffend der Sicherheit, wie in Kapitel 5.4.7 beschrieben, erleichtern. Es ist zu beachten, dass die richtige Balance zwischen striktem Durchsetzen der Richtlinien, was höhere Kosten verursacht, und dem weniger strikten Durchsetzen, was ein erhöhtes Risiko für Verluste birgt, gefunden wird. Weiter gibt es natürlich betreffend der Sicherheit gesetzgebende Richtlinien, die erfüllt werden müssen. Es gilt zu berücksichtigen, dass bei einem verteilten System die Grenzen von administrativen Domains überschritten werden, das heisst, dass jede administrative Domain ihre eigenen Sicherheitsrichtlinien hat und diesen Rechnung getragen werden muss. Hinzu kommt, dass zusätzlich die Sicherheitsanforderungen einer virtuellen Organisation als Ganzes berücksichtigt werden müssen.

Folgende funktionelle Fahigkeiten müssen von den OGSA-Security Services zur Verfügung gestellt werden:

- *Authentication*: Eine Person oder Instanz muss die eigene Identität nachweisen. Zum Beispiel mittels Benutzername und zugehörigem Kennwort.
- *Identity Mapping*: Dieser Dienst ermöglicht es, eine Identität aus einer Domäne in einer anderen Domäne weiterzuverwenden.
- *Authorization*: Es wird geprüft, ob ein bereits identifizierter Nachfrager eines Dienstes oder einer Ressource berechtigt ist, die Ressource oder den Dienst zu nutzen.
- *Credential Conversion*: Dieser Dienst kann einen Typ eines Berechtigungsnachweises, wann und wo für einen anderen Autorisierungs-Dienst nötig, in einen anderen umwandeln.
- *Audit and Secure Logging*: Der Dienst ist verantwortlich, dass sicherheitsrelevante Vorkommnisse aufgezeichnet werden.
- *Privacy*: Daten sollen nur für berechtigte lesbar sein.

Self-Management Services

Die Self-Management Services sollen die Kosten der Infrastruktur eines verteilten Systems senken, sowohl auf der Soft- als auch auf der Hardware-Seite. Dies soll anhand von drei funktionellen Fähigkeiten der OGSA Self-Management Services erreicht werden. Die entsprechenden Services werden nachfolgend aufgelistet und erläutert:

- *Self-configuring Mechanisms*: Diese Mechanismen sollen es beispielsweise erlauben, je nach Last, basierend auf Richtlinien, Komponenten hinzuzufügen oder überflüssige zu entfernen.

- *Self-healing Mechanisms*: Das System soll nicht richtig funktionierende Ressourcen und Dienste selbständig erkennen können und gemäss vorgegebenen Richtlinien korrigierende Massnahmen ergreifen. Dies ohne die Umgebung zu stören. Weiter sollen mit diesen Mechanismen ebenfalls Attacken wie zum Beispiel Verseuchung mit Viren oder eine Denial-of-Service-Attacke gegen das verteilte Systems erkannt werden und Massnahmen ergriffen werden, die das System sicherer machen.

- *Self-optimizing Mechanisms*: Dieser Mechanismus soll helfen, höchste Effizienz zu erreichen.

Die drei Mechanismen können ihre Aufgabe nur durch Zusammenarbeit erledigen. Für die OGSA Self-Management Services sind die folgenden Eigenschaften wichtig und definiert: Availability, Security und Performance.

Information Services

Für andere OGSA-Dienste ist es von grosser Wichtigkeit, dass sie einfachen Zugang zu Informationen über Applikationen, Ressourcen und andere Dienste im verteilten System haben. Ohne diese Informationen können sie ihre Dienste nicht vollbringen. Die Information Services in OGSA bieten anderen Services diesen Dienst an. Die Information Services halten die entsprechenden Informationen für die Phase von der Publikation eines Dienste bis zu dessen Verbrauch bereit. Informationen werden vom Produzenten der Information oder einem Intermediär zugänglich gemacht. Die Information wird dann von einem oder mehreren Konsumenten nachgefragt. Oder ein oder mehrere Produzenten wollen ihre Information einem oder mehreren Konsumenten übermitteln. Es soll nicht nötig sein, dass sich Produzent und Konsument vorgängig kennen. Konsumenten sollen via Push- oder Pull-Mechanismus an die benötigten Informationen kommen. OGSA schreibt nicht vor, in welchem Format die Informationen zugänglich sein müssen. Die folgenden funktionellen Fähigkeiten sind für die OGSA-Information Services definiert:

- Naming scheme
- Discovery
- Message delivery
- Logging capabilities
- Monitoring capabilities

Für die OGSA-Information Services sind die folgenden Eigenschaften wichtig und definiert: Security, Quality of Service, Availability, Performance and Scalability.

5.5.2 Accounting, Charging, Pricing

Nach [16] geht es beim Accounting darum, über eine Metering-Komponente gemessene Daten zu speichern. Es besteht die Schwierigkeit, Metriken zu finden, die im Hinblick auf das Charging und das Pricing flexible Tarifstrukturen erlauben. Beim Charging werden dann die gesammelten Accounting-Daten in einen nicht monetären Wert umzuwandeln. Zum Beispiel User X hat Service Y zu 10 Einheiten benutzt. Beim Pricing werden dann diese nicht-monetären 10 Einheiten in einen monetären Wert gewandelt. Zum Beispiel 1 Einheit Y zu 10 CHF. Natürlich gibt es beim Pricing dann auch kompliziertere Funktionen mit denen beispielsweise ein Flatrate-Angebot abgerechnet werden kann. Beim Charging kommt beispielsweise das Problem auf, wie die Daten zwischen den entsprechenden Stellen transferiert werden. Oder wer die Beziehung zum Endkunden unterhält.

5.5.3 Regulations

Das Thema der Regulations wird nur im Sinne einer Beschreibung der Problemstellung beschrieben, Lösungsansätze können im Rahmen dieser Arbeit nicht entwickelt werden. Betreffend der Regulations ist die Problemstellung die Folgende: Werden in der Schweiz Daten, die unter das Bankgeheimnis fallen, innerhalb eines verteilten Systems benutzt, muss darauf geachtet werden, dass die entsprechenden Daten in Teilen des verteilten Systems, die sich in anderen Ländern befinden, nicht zugänglich sind und schon gar nicht dort gespeichert werden. Das Gleiche ist beim Datenfluss innerhalb des verteilten Systems der Fall, es müssen Mechanismen eingebaut werden, um zu verhindern, dass Daten durch gewisse Länder beziehungsweise Regionen fliessen. Wenn auf solche Regulatorien Rücksicht genommen werden muss, ist es möglich, dass diese ein verteiltes System verunmöglichen. Die Auswirkungen der regulatorischen Rahmenbedingungen müssen also unbedingt zu Beginn der Planung des verteilten Systems berücksichtigt werden.

5.6 Schlussfolgerung und Zusammenfassung

Das Studium der Literatur hat gezeigt, dass die theoretischen Entwicklungen von Frameworks und Architekturen voranschreitet. Auf der anderen Seite müssen durch zukünftige Studien die möglichen Einsatzfelder verteilter Systeme in der Praxis identifiziert und eingegrenzt werden. Weiter müssen methodische Ansätze für den strukturierten Einsatz der neuen Wertschöpfungskette entwickelt und anhand praxisnaher Versuchsprojekte überprüft, weiterentwickelt und konsolidiert werden.

Neben den organisationalen Aspekten des Projektmanagements und den kontinuierlichen technischen Weiterentwicklungen verteilter Systeme stellt das A4C ein kritischer Erfolgsfaktor dar. Die in der Arbeit geschilderten Schwierigkeiten und noch ungenügend definierten Anforderungen an das A4C müssen analog dem Akogrimo-Projekt [17] durch praxisnahe Projekte verifiziert und weiterentwickelt werden. Die zukünftigen Entwicklungen der Regulations auf gesamteuropäischer Ebene können zu zusätzlichen positiven Effekte

auf die Weiterentwicklung verteilter Systeme bewirken, vereinfacht doch eine länderübergreifende Regelung gesetzlicher Vorschriften den Aufbau globaler, verteilter Systeme im Sinne virtueller Organisationen.

Ein erster Ansatz in dieser Richtung stellt das in dieser Arbeit vorgestellte Vorgehen des OGSA-Frameworks dar, welches versucht, die Anforderungen an virtuelle Organisationen zu konkretisieren und mittels eines generischen Frameworks zu implementieren. Die Grid-Computing-Community beschäftigt sich seit mehreren Jahren mit der Definition der Anforderungen (Kapitel 5.4), welche beim Aufbau einer virtuellen Organisation mit dem Ziel einen ökonomischen Mehrwert zu realisieren, beachtetet werden müssen. Das OGSA-Framework realisiert diese Anforderungen durch ein Set einer generischen Architektur, welches Konzepte unabhängig von der physischen Realisierung beschreibt (Kapitel 5.5). Ausgehend von dieser Architektur ist es das Ziel zukünftiger Projekte, technische Realisierungen zu entwerfen. Die aktuellen Bemühungen basieren auf den bekannten Technologien der Web Services. In wie weit sich diese Technologien eignen, oder ob völlig neue Konzepte herangezogen werden müssen, ist offen.

Literaturverzeichnis

[1] M. Porter: Competitive Strategy, The Free Press, 1980

[2] H. Häuschen: Zwischenbetriebliche Integration, Vorlesung eBusiness, 2004/2005.

[3] John Neff: Next MINI engines to be built by BMW and Peugeot, 22.06.2005, http://www.autoblog.com/2005/06/22/next-mini-engines-to-be-built-by-bmw-and-peugeot/, 08.01.2006

[4] Hans Jägers, Wendy Jansen, Wilchard Steenbakkers: Characteristics of Virtual Organizations, http://www.ve-forum.org/apps/pub.asp?Q=932&T=BookofKnowledge&B=1, 20.11.2005.

[5] M. Waldburger & B. Stiller, Toward the Mobile Grid: Service Provisioning in a Mobile Dynamic Virtual Organisation, August 2005.

[6] I. Foster, H. Kishimoto, A. Savvaa et al., Gridforum.org, http://www.gridforum.org/documents/GWD-I-E/GFD-I.030.pdf, 18.11.2005.

[7] Autoren: Wikipedia.de, Interoperabilität, Wikipedia.de http://de.wikipedia.org/wiki/Interoperabilit%C3%A4t, 07.01.2006

[8] D. Haage, V. Olmedo et al.: Akogrimo - Overall Network Middleware Requirements Report.

[9] Wikipedia.de, Service Level Agreement, Wikipedia.de http://de.wikipedia.org/wiki/Service_Level_Agreement, 07.01.2006

[10] A. Picot, H. Dietl, E. Frank, Ogranisation - Eine ökonomische Perspektive, 3. Auflage, Schäfer-Poeschel Verlag Stuttgart, 2002.

[11] Wikipedia.de, SMART (Projektmanagement), Wikipedia.de http://de.wikipedia.org/wiki/SMART_%28Projektmanagement%29, 07.01.2006

[12] Andrew Grimshaw, High Throughput Computing, University of Virginia http://www.cs.virginia.edu/~grimshaw/CS851-2004/HTC-links.htm, 07.01.2006

[13] Wikipedia.de, Service Oriented Architecture, Wikipedia.de http://de.wikipedia.org/wiki/Service_Oriented_Architecture, 25.11.2005

[14] Wikipedia.org, Open Grid Services Architecture, Wikipedia.org http://en.wikipedia.org/wiki/OGSA, 9.1.2006

[15] T. Dimitrakos, D. M. Randal, F. Yuan et al., eu-grasp.net, `http://eu-grasp.net/english/dissemination/articles/EDOCconference03.pdf`, 9.1.2006.

[16] Global Grid Forum, Grid Economic Services, Juni 2003.

[17] Communication Systems Group, University of Zurich, Access to Knowledge through the Grid in a Mobile World (Akogrimo), `http://csg.ifi.unizh.ch/research/akogrimo/`, 26.01.2006.

Chapter 6

Economy Driven Peering Settlements

Barbara Schwarz, Gian Marco Laube, Sinja Helfenstein

In the Internet's structure, peering is one possibility for Internet service providers to interconnect. It enables to avoid a solely hierarchical architecture. Even though it is based on collaboration, it is still driven by strategic and economical goals. Our work does not focus on the technical aspects, but concentrates on the economical consequences when ISPs decide to peer. We start by looking at the general motivation factors and continue with a more detailed analysis of a business case. Having shown the situation from the ISP's perspectives, we proceed to investigate the implications for the entire industry. To illustrate these theoretical concepts we look at the current situation and the main players in Switzerland. While transit agreements lead to deadweight loss, pure peering strategies are not stable due to market dynamics. As a consequence new settlement models are required not only for a fair cost allocation, but also to introduce guaranteed quality of service for a connection. The discussed issues are still hidden by today's market growth. It can be expected though, that a more competitive environment will lead to increased awareness of exactly these topics in the near future.

Contents

6.1 Introduction

How much does it cost to send a long-distance e-mail? This question might sound silly at first sight. Asking it to an arbitrary number of people would presumably lead to answers such as *'It's free - of course!'* or *'It is included in the monthly flat rate'*. These statements are certainly true - but what do the economics beneath look like? How can the ISP ensure that the long-distance e-mail actually reaches its destination, and what are the cost for this service?

Once we realise that Internet service providers (ISP) do not access the Internet for free either, but are charged for all data that they send to and receive from destinations „outside their own network" it becomes obvious that there must be strategies to keep costs low. This need is even strengthened due to an environment with continuously declining end customer fees for (unlimited volume) Internet access and fast increasing performance requirements.

Hence to survive in their uttermost competitive market, ISPs must continuously focus on keeping cost low and performance high. Peering settlements are a very effective method to achieve exactly these two goals and have therefore recently gained importance. Although it might sound like a very easy decision *'To Peer or not to Peer'*, the story is much more complex and includes further factors such as market dynamics, politics, technological (dis)advantages and many more. The matter of fact that peering always includes two parties, increases the complexity and brings us to a point where we have to study and evaluate the various settlement models that are viable.

6.1.1 Related Work

A good introduction to the economical basics of peering is the business case analysis of W. Norton [18]. The assets and drawbacks of peering are easier to understand if looked at from a provider's perspective than from a solely theoretical view. For our work we use this business case as a framework and update the numbers to recent Swiss prices. The appendix of Norton's work also provides a brief but accurate description of reasons not to peer. S. Gibbard [8], [9] described the deadweight losses of a pure hierarchical Internet architecture and underlines the positive aspects of peering for the society. Aiming to approach the topic with a more neutral view, we extend this view by also showing the disadvantages of peering. A detailed paper from G. Huston [10] gives a broad overview over the economics and few some technical aspects of peering. Comparing the Internet with the Telephony market (see Section 6.5.1) he concludes that up to now only zero-settlement peering and transit agreements are stable settlement models in the Internet architecture and that the most natural business outcome of today's Internet is one of aggregation of providers. He believes that the exponential growth of the market has masked the problems of unfair cost distribution under providers up to now (see Section 6.2.3). At the end of our work, we try to expand these thoughts and include recent work about new settlement models. The papers [5] and [1] show the economical factors that can hinder ISPs of different dimensions from peering. A more technical report of L. Gao and J. Rexford [7] concentrates on the technical aspects and problems of the BPG

(Boarder Gateway Protocol [26]) routing in relation to peering. As we found out in our interview with Fredy Küenzler from Init 7 AG [12], strategic and political problems are by far the bigger handicap for peering arrangements than conflicting BGP policies, so we concentrated more on the economical factors.

In the third part of this paper we concentrate on new settlement models for the Internet. The CATI project [23] categorized between the years 1998 to 2000 different possible settlement models and discussed their utilisability. Important for our discussion was the categorization of the different models. The project showed also that most peering agreements in practice are not perfectly balanced and that in the long term better settlement structures are needed. A recent work of T. Nguyen and J. Armitage [17] gives a very compact overview of this field. For our discussion of proposed models we used also the evaluating scheme of this report that makes a distinction between three different dimensions (technical, economical efficiency and social impact).

6.2 Peering Fundamentals

Before we can start discussing the peering settlement models and their economic relevance, we need to establish a basic understanding of the Internet's structure. Having introduced this technical and economical background, we can then narrow our focus and concentrate on peering. What it is in general, how it works from a technical perspective and how the underlying economics look like.

6.2.1 The Internet

The Internet can be partitioned into autonomously administrated domains, which vary in size, function and geographical extent. These domains are called *autonomous systems (AS)* [25] and either consist of at least one ISP, a large number of web servers or a company's network. While we are not further interested in the internal structure and administration of such ASs, we will have a closer look at their interaction. In other words we will investigate the Internet's structure on an AS-level and discuss the different interconnection strategies that ASs can follow. When using the term *interconnection* we generally speak of the actual physical connection between two independent networks.

We start with a brief introduction to an AS in general, its routing protocol and further proceed to looking at the Internet's structure with regard to the technical setup and the related economics.

Autonomous Systems and BGP

From a pure networking perspective AS are understood as domains that are centrally administrated and have a unified internal routing policy. In order to be acknowledged as an AS in the Internet domain, one further condition applies: The property of an

official AS-Number (ASN). These ASNs are globally administrated by the *RIPE Network Coordination Centre* [24] and in order to be eligible for one an ISP must be a member of RIPE NCC and fulfil a set of conditions. More information about these conditions and RIPE NCC itself can be found on the organisation's website at http://www.ripe.net.

The protocol used for routing between AS is called BGP - Broader Gateway Protocol [26]. BGP is working on the TCP/IP's Application Layer and is responsible for controlling the handling of IP-packages. Its main function is therefore to define the import and export rules for packages for inter-domain routing: Which incoming packages to accept and where to send outgoing packages, depending on their destination. BGP is also used for AS-internal routing, then called iBGP, as shown in Figure 6.1. [7]

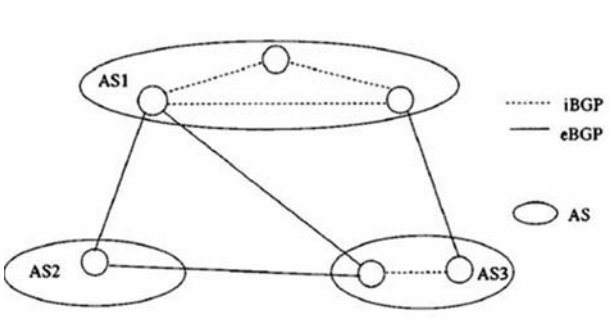


Figure 6.1: AS Interaction [7]

Hierarchical Structures

Despite the absence of a central point of control there are hierarchical relationships existing in today's Internet. With regard to these relationships a pyramid-structure can be identified, with a small number of so-called *Tier 1 Internet service providers (ISPs)* on top. Each Tier 1 ISP has its own (physical and logical) Network and offers transit-services to lower-level (*Tier 2*) ISPs: Access to the backbone network. Tier 2 ISPs can resell this service to their customers. As reselling is a transitive property, we can reach arbitrary deep tree-structures, where ISPs are related to each other by customer-provider-relationships [19]. Wherever a depth of more than two levels is reached some ISPs are in the role of customer and provider simultaneously, depending on the perspective. In fact, the only exception of this would be the tier 1 ISPs that do not have to buy Internet access from anywhere, because they actually build the backbone themselves and the end-customers that do not resell any traffic. This transitivity property of reselling and the role of customer-provider-relationships for the resulting tree-structure is shown in Figure 6.2.

Whenever we are talking about such hierarchical relationships where one ISP is buying Internet access from another, we can also use the terms „transit“ and „upstream“ traffic for the actual data-flow and „upstream carrier“ respectively for the ISP in the role of the provider.

As these customer-provider relationships are very dynamic and by far not the only relationship character existing, there is no stable structure of the Internet and it can rather

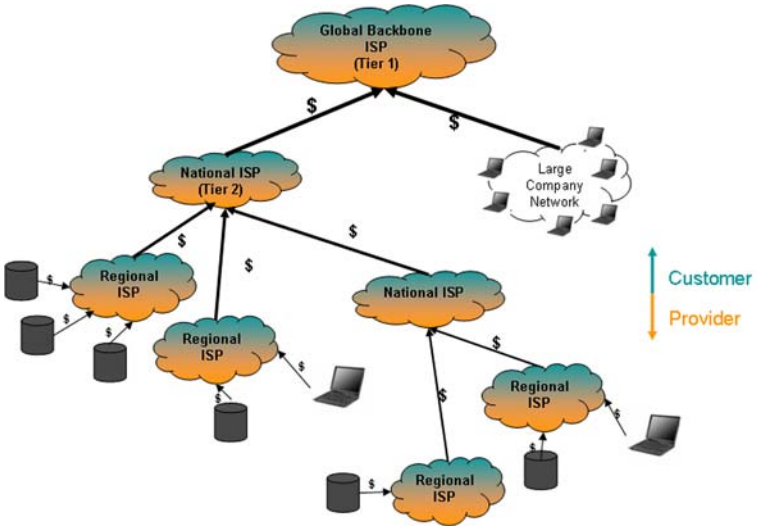


Figure 6.2: The Internet's hierarchical customer-provider relationships

be viewed as a complete mashed network than a strict hierarchical one. Recalling the Internet's structure as a set of autonomous systems, the term '*autonomous* systems' already indicates that there is no central point of control in the Internet due to the mutual interaction of the ASs, as shown in Figure 6.1. But in order to further investigate these ISP relationship-characters that go beyond customer-provider dependency, a closer look at the basic economics of an ISP is needed.

Economics of Internet Service Provisioning

As mentioned in the introduction of this paper, the actual economics of Internet-service provisioning are not yet a matter of common knowledge, it therefore makes sense to have a closer look at it. In the previous writing the term „customer" has been used repeatedly. As ISPs are profit-oriented companies, they usually do not provide their data-transportation services for free. But how exactly is the term „customer" to be defined in the ISP-world and how do the ISP's economics look like?

The first thing you encounter when looking at the Internet from a bottom-up perspective are the end-customers that buy Internet access from their regional ISP. The physical link (last mile) from the customer's location to the ISP's POP (point of presence) is not necessarily part of the ISP's service, as it is usually depending on the existing infrastructure. Therefore a telephone circuit from an incumbent operator can be used to access the POP of a competing start-up ISP.

From such an end-customer's point of view the cost of Internet access are built up as following:

- Initial investment in hardware and software
- Connection to ISP's POP

- Monthly access fee for ISP

Today's broadband access offerings usually do not have connection cost, as they do not use the conventional telephone circuits (dial-up). It is usual to charge a flat rate only, which allows to send and receive a specified or unlimited data volume. [29]

From an ISP's point of view the cost of Internet access are built differently:

- Initial investment in hardware and software
- Service assurance: Upstream package forwarding, usually volume-based charges

ISPs are therefore aiming to keep the upstream data-volume as low as possible and prefer internal traffic.

6.2.2 Peering Basics

As already mentioned in Section 6.2.1 there are further relationship types besides the hierarchical one. W. Norton [18], [19] defined the two main types of business relationships between two physically connected ISPs as following:

- **Transit:** The business relationship whereby one ISP provides (usually sells) access to all destinations in its routing table.
- **Peering:** The business relationship whereby IPSs reciprocally provide access to each other's customers.

Recalling the hierarchical relationships described in Section 6.2.1, they are exactly what Norton refers to when describing the transit business relationship. Transit is thus simply another terminology for the previously discussed customer-provider-relationships.

Figure 6.3 shows the difference between these relationships and illustrates how transit can be avoided by appropriate peering partnerships.

G. Huston [10] mentioned as a third alternative the extension of a peer-to-peer relationship to act as backup link for upstream traffic, meaning that the peering partners provide each other with the option to use the peer's upstream capacity in case of failure of one's own upstream carrier. This will not be discussed further in this paper, as it does not have an impact on the settlement models. Anyhow it should be kept in mind as one further positive aspect of peering.

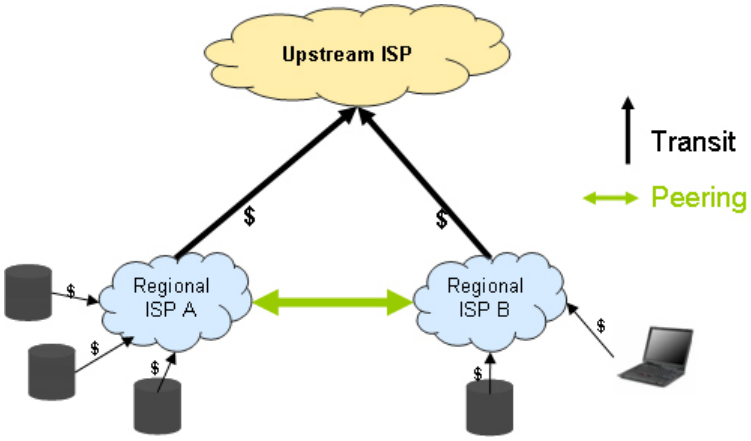


Figure 6.3: Avoiding Upstream Traffic by Peering

Non-Transitivity of Peering

As the name already indicates, the transit relationship is transitive, meaning that the ISP provides universal access to all destinations in its routing table. Peering in contrast is non-transitive and the ISP's provide only access to destinations within their own networks [18], [19].

The following example shall demonstrate this property of (non-) transitivity:

We assume the following situation:

- ISP A is interconnected with ISP B
- ISP B is interconnected with ISP C

In case of these connections being customer-provider ones, this implies that ISP A is automatically a „sub-customer" of ISP C, simply by being a customer of B. This would for example be the case with A being the end-customer, B being a local ISP and C being an international backbone ISP such as MCI or AT&T.

In a second scenario, where the connections are peer-to-peer, A's peering relationship with B does not include access to B's peering partner C. To get a better picture of this example as well, we apply it to the Swiss market and assume A to be sunrise, B cablecom and C swisscom. In that case sunrise and swisscom would still use their upstream carrier's traffic exchange, as they do not have each other's customers in their routing tables. Figure 6.4 shows the resulting routing table for each of these ISPs after peering as described in the second scenario.

6.2.3 The Cost of Peering

As peering-partners do not charge each other for the traffic volume, peering appears to be free at first sight. However this is not the case as there are some more factors to be

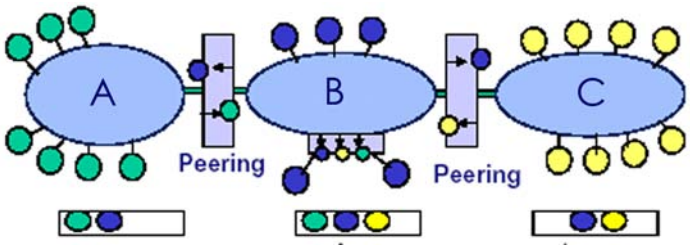


Figure 6.4: Non-transitivity of peering relationships

considered, such as setup cost occurring on both sides: Investment in new infrastructure and expenditures of human labour for the technical setup. After that the continuous maintenance and monitoring expenditures should not be underestimated either.

A second point is the fact, that actually only at so-called *'zero settlement peering'* agreements the partners do not charge each other for any traffic. In fact we often encounter situations where one partner is more powerful and can therefore demand certain advantageous conditions. As soon as this is the case, we do not speak of actual zero settlement peering anymore, as the two partners are not equal.

Recalling the two business relationships described in Section 6.2.2 it makes sense not to take them as exclusive states, but much more as two extremes, leaving much space for further models in between, as shown in Figure 6.5.

Hence whenever two parties decide to peer, negotiations arise about who will bear the related expenses. Market power plays a very important role in cost-allocation and peering-decisions. In fact, high differences in market power can even lead to one party not wanting to peer, or applying certain additional cost to the smaller partner. This leads to a situation where we do not have a digital decision of peer - no peer, but we have many more options in between. The stronger the difference in market power, the more we have a customer-provider situation, and the more equal the two parties are, the higher the probability that they will achieve a so-called *'Zero Settlement Peering'* where the cost are shared and both parties bear the same rights and duties. Later in this paper we will have a closer look at the possible agreements and charging models for all the in-between solutions.

Where on the graph in Figure 6.5 a specific business relationship is actually to be allocated, depends on the following factors:

- **Data-Volume Allowance and Charges:** No limit, free of charge for both sides in a zero-settlement partnership vs. one-side volume-based restrictions and charges in customer-provider situations.

- **Bearing of Setup Cost:** Partners equally share the peering cost, while the further we move towards a customer-relationship, the more has to be defrayed by the weaker party.

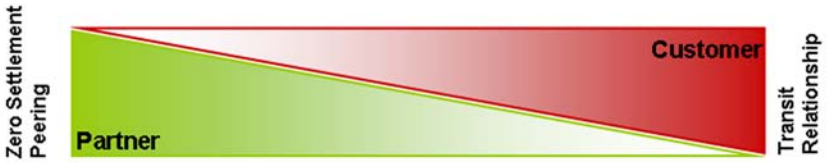


Figure 6.5: Possible Business Models from Customer to Partner

6.2.4 Implementation

On an implementation level two different approaches for interconnection are possible:

- Direct-Circuit Interconnection: Bilateral Point-to-Point connection.
- Exchange-Based Interconnection: Participation at a public Internet Exchange.

The latter one involves a meeting point at which ISPs exchange traffic whereas the direct-circuit model is based on a direct and private point-to-point connection between the two exchange parties only [20]. While the direct-connection is principally about setting up a physical and exclusive connection between the two POP's, the exchange-based model involves many more aspects and is thus worth a closer examination.

Internet Exchanges

An Internet Exchange (IX) is a facility where multiple ISPs can interconnect. It is generally an ethernet switch that all participants plug into and use to establish BGP sessions between their networks and update their routing tables accordingly.

Many IXs also offer private cross-connects, cables going directly between two participating networks. Private cross-connects are useful when two networks have a large amount of traffic going between them, and do not want to fill up the capacity of their exchange point switch ports or whenever parties do not want to interconnect with all the participants at the IX.

The occurring infrastructure cost are usually shared between the participants of an exchange and it is in every participant's own responsibility to ensure its access to the IX's POP.

When participating at an IX, an ISP has two options:

- Follow an open peering policy and interconnect with all willing ISPs at the IX.
- Keep some restrictions and only do private interconnections with selected partners.

In both cases the main advantage compared to the direct-connection is the fact, that by establishing the physical connection to the IX once, there are suddenly many more potential interconnection partners, with low marginal cost.

Comparison of the Implementation Models

As just mentioned there is an apparent difference on the cost-side when comparing the two interconnection models. The direct-circuit interconnection obviously leads to significantly higher setup and maintenance cost and does not profit from economies of scale at all.

W. Norton has compared the two models more precisely from a technical and financial standpoint and drew the following further conclusions:

> „The direct circuit interconnection model suffers from the scaling over both bandwidth and number of participants. As the bandwidth and number of participants grow, the number and size of circuits must increase in a way that disallows economies of scale through traffic aggregation."
>
> „The cost savings function is an incremental function leading to potentially significant savings."
>
> „There is a potential for substantial savings for ISP interconnection under the exchange-based model."
>
> „The first-to-market in a neutral Internet Business exchange will realize the greater revenue by being able to participate in a greater number of business opportunities."[20]

Peering Policies

Based on the above discussed considerations an ISP has to decide about its peering strategy. Does it want to peer at all and if so, with whom, to what conditions and how (technically). The results of these considerations are usually written down in a document called *Peering Policy*. Peering policies define the prerequisites for interconnection - if those defined rules are not met by a potential partner there is no need to further evaluate or negotiate.

Peering policies include the following criteria

- **Technical Aspects:** POP at specific IX, Protocol Version (e.g. BGP-4), Membership of RIPE NCC (Having an ASN), etc.
- **Business Aspects:** Minimum size of customer base, Restriction on further peering partners, etc.
- **Legal Aspects:** Non Disclosure Agreements, Security Standards, Legal Binding, etc.

Some ISPs publish their peering policy publicly on their website, others keep it private and only apply it when they encounter a potential partnership.

The often used term *Open Peering Policy* refers to ISPs that do not apply economically driven exclusions of potential peering partners. As long as a potential partner fulfils the technical requirements, they are willing to engage in a zero settlement peering with anyone.

6.2.5 Development and Motivation

In the Internet's beginning the discussed financial aspects where much lower prioritized than they are today. Anyhow peering was already relevant, but dominated by other reasons.

The discussed more-or-less hierarchical structure of the Internet has emerged historically and lead to today's existing stable and well-connected core. But this core only covers a small part of the world's population, where high redundancy and low communication cost can be observed. This is mainly the case in the more developed world. On the contrary to this phenomenon, many peripheral regions and less-developed countries still have a comparatively low-speed and high-cost connection to the Internet backbone with low or even no redundancy.

A strong hierarchical structure would imply that all the local traffic traversed those weak and expensive upstream connections to the backbone, even if it was for a computer only a few miles away. This would lead to enormous cost and high latency. Steve Gibbard [9] uses Nepal's situation to illustrate this fact. In Nepal, the international transit was received for $5,000 per Mb/s via a satellite-connection with high latency. This lead to high cost and low quality, which in turn had a negative impact on the Internet's popularity and growth:

> 'ISPs in Kathmandu, Nepal, are buying their international transit over high-latency satellite connections. The high costs and poor performance are keeping traffic volumes pretty low in those areas, and thus seem to be a significant barrier to Internet development.' [9]

Such local problems can be solved by introducing national *Internet Exchange Points*, which interconnect the local ISPs. By routing national traffic directly to the destination ISP, the expensive and slow detour via backbone can be avoided. In addition to lower cost and higher quality of service, the resilience can be improved by keeping the local traffic working even at a breakdown of the upstream link or carrier. Taking up Steve Gibbart's example of Nepal again, the use of such a national exchange can be shown by the significantly lowered cost:

> 'Since the Nepal Internet Exchange is run by its member ISPs, switch ports are currently free. They are talking about increasing the switch port price to $800 per year to cover some spare equipment. 2 Mb/s circuits to get to the exchange cost around $13 per month. These circuits (dry copper) require about $1,000 worth of equipment, but when spread out over three years that is $27 per month. The total cost of connecting to the exchange ends up being $107 per month for the first two Mb/s, and $40 per month for each additional Mb/s. So, in the worst case scenario, the cost for local peering ends up being $53.50 per month per Mb/s, as compared to $5,000 per month per Mb/s for transit.' [9]

While resilience and dependency are primarily problems for less developed countries with little redundancy, the so-called core-regions focus more on aspects such as performance, cost and market dynamics. Of course there is no sense for an ISP in Zurich to send data for an ISP in Basel via Boston, as this would lead to higher cost and latency.

In the following two sections we will examine the financial reasons for and against peering and later look closer at the further factors impacting the peering decisions.

6.3 Peering vs. Transit

As described in Section 6.2 an ISP has normally two choices: Either he pays a transit provider to interconnect with the rest of the Internet or he agrees to a peering partnership with few specific partners.

The main factor for this decision is the direct cost. The following chapter will also cover many indirect cost, such as additional technical know-how that is needed for peering agreements.

6.3.1 A Business Case

The best way to illustrate the economics of peering decisions is to calculate a concrete example of a cost situation for an ISP. The following example has been created by G. Huston [10], and updated with current data: approximated costs for a Swiss ISP in the year 2005. [12]

First we analyze the cost of a customer ISP, if he has as transit provider. In the past a transit provider was mostly paid by a measured amount of traffic volume that is used by the customer ISP during a certain time period. The measurement unit is normally Megabits-per-second (Mbps). The volume exchanged is normally measured only a couple of times a month. The peak 5% (the time where traffic volume exchanged is highest) is then ignored, and the 95th percentile is charged to the customer ISP. Today it is more and more common to charge a certain capacity instead of the real volume used. That way a critical success factor for an ISP is to „keep the backbone full“. A transit provider in Switzerland charges between CHF 20 - CHF 100 / Mbps by a minimum capacity of 100 Mbps. That means that this ISP pays between CHF 2000 and CHF 10’000 per month for a capacity of 100 Mbps. The more traffic the customer sends / receive (what exactly is to be measured can be different too) over the transit provider, the less he has to pay for each Mbps. This way the pricing structure has different price levels for different volumes exchanged. This way the customer ISP profits indirectly from economies of scale, as shown in graph 6.6.

During the last years of the previous century, traffic increased dramatically. Transit costs went also down, but not proportional to the increased traffic volume. This leads to a cost growth especially for local ISPs. As a lot of traffic went to neighbouring nets, peering became an issue. Naturally the ISP still needed a transit provider in order to access the

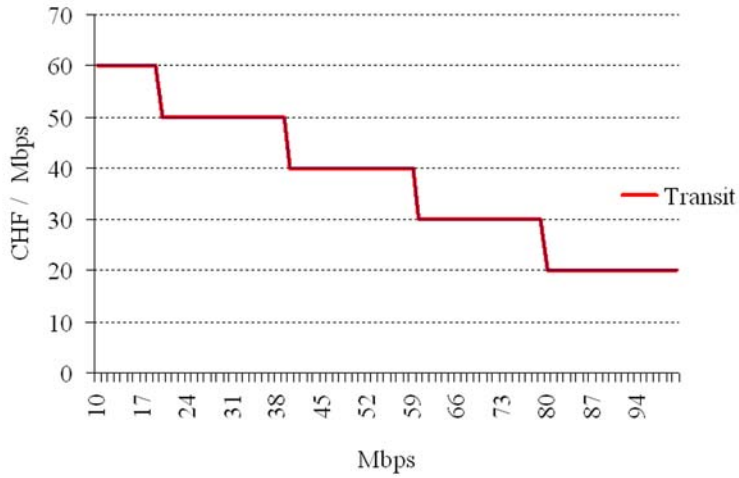


Figure 6.6: Peering vs. Transit: Example transit costs for a local ISP

global Internet, but with peering they found an inexpensive possibility to interconnect each other's customers.

Peering is often considered to be „free". As already stated earlier in this work, this is not correct. Even peering partners in a zero settlement peering relationship that do not charge each other for the traffic exchanged, still have cost factors to consider. As already shown there are different peering possibilities, which bring also different cost factors with them.

By peering at a local Internet exchange point, an ISP has to consider the transport of the traffic to the IX, a fee for the rack space in the IX and for the switch port.

The TIX in Zurich charges (2.12.2005) the following costs for hardware and service elements:

- CHF 500 for a 10/100BaseTX Port
- CHF 2000 for a 1000BaseLX Port
- CHF 950 for half a rack

In addition to these cost an ISP would also have to include the cost for the physical connection to the point of presence of the exchange point. We ignore this cost here as we assume that the cost for the link to a peering point are about the same as the ones for a link to the next transit provider access point.

The main economic difference between a peering and a transit relationship is that in the first case the cost factors are all fixed cost per month, while in the second case the customer pays a variable price depending on how much Mbps he uses. While a customer ISP pays more for higher traffic volume, the peering partners can exchange as much traffic as the underlying ports / bandwidths can handle. While the transit provider only returns

Table 6.1: Prices of a transit provider

Mbps	CHF / Mbps
0 - 20	60
20 - 40	50
40 - 60	40
60 - 80	30
80 -	20

part of his economies of scale to his customer, the peering partners profit from the full cost distribution when more traffic is routed over their connection.

Back to our example: Consider the Swiss ISP A sends and receives all his traffic over one transit provider. A finds out that it regularly sends about 35 Mbps of traffic to the other Swiss ISP B, and receives about 10 Mbps of traffic from B. This is quite a share of A's total traffic volume. A pays for the transit service between 50 - 80 CHF / Mbps depending on the used volume (see price Table 6.1).

Because the exchanged traffic volume is under 100 Mbps (congestion avoidance would limit the maximum average load to about 70 - 80 Mbps), A could peer with B at the Internet exchange point TIX nearby using a 10/100Base-Port. We assume here, that ISP A is not yet present at this IX.

Half of a rack and the 100 Mbps port sum up to a total cost of CHF 1450.

As in the transit case the distribution of the total cost for peering over the volume used can be shown as graph 6.7.

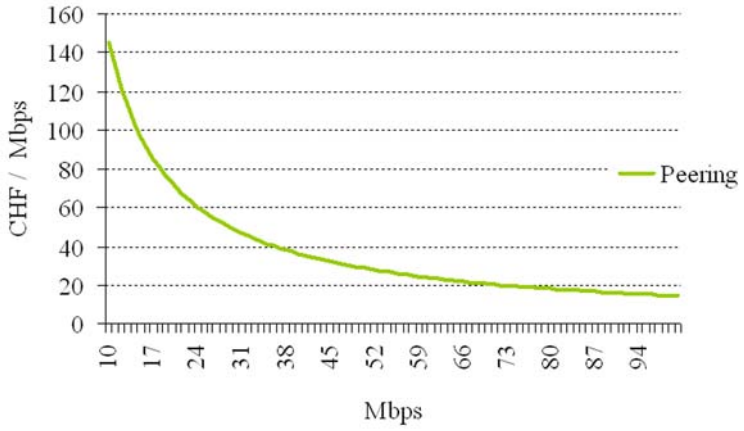


Figure 6.7: Peering vs. Transit – Example peering costs for a local ISP

All this information about peering or transit interconnection can be aggregated to one break even analysis graph 6.8.

The graph 6.8 shows that there is a peering breakeven point at a traffic volume of 29 Mbps. Left of that point the ISP does not generate enough traffic to make peering profitable (peering risk). If ISP A only has to pay for the traffic it sends, it has to

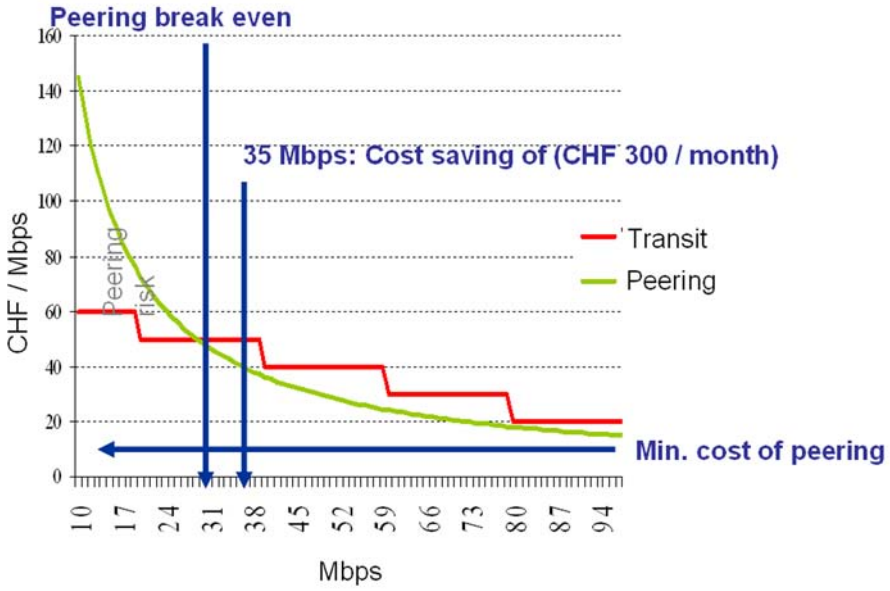


Figure 6.8: Peering vs. Transit – Breakeven analysis

compare peering and transit at the point of 35 Mbps per month, else with 45 Mbps per month. In both cases peering is profitable. ISP A is able to save CHF 300 (35 Mbps) or CHF 350 (45 Mbps) per month. It is also possible to define a minimum cost of bandwidth, which depends on the effective peering bandwidth. Either the ports or the connection to the peering point limit the possible bandwidth to a certain amount. The fixed costs of peering divided trough the maximum of this bandwidth gives us the minimum cost per Mbps trough peering (about 15 - 20 CHF / Mbps in this case).

Up to now this was the perspective of ISP A, but it is also necessary to decide if peering is the right choice for ISP B. If A and B pay their transit provider only for sending data, the situation could look different for ISP B. If B only sends 10 Mbps to A, peering costs 145 CHF / Mbps compared to 50 CHF / Mbps over the transit provider. It is not even possible for A to pay B additionally to agree to peering, because there is no way that it might be profitable for A to cover the peering hardware and service costs for both at the IX.

If both have to pay their transit provider for up- and downloaded traffic (as it is normally the case) peering is the best choice also for B.

A situation similar to the example described above happened between two big Swiss ISPs and a global transit provider. Peering was in that case the better choice, as the two providers had an alternative peering point, at which the costs for the equipment was by far lower. We also ignored here the fact, that the Swiss ISPs are already present at the big Swiss IX and that they can use one peering port multiple times. So as long as A and B have open ports at the IX, A and B have not to take the full fixed cost into account for peering.

Many more factors can influence a peering decision (as discussed in the next section). Because of all these side effects that one has to include in a peering or transit decision,

common statements from peering responsibilities address topics as „the art of peering“ or the fact that „peering is all politics“.

6.3.2 Further Decision Factors for Peering

Why else to Peer

Besides the fact that peering enables lower transit costs, there exist further decision-influencing parameters. Subsequently some additional and important topics why else peering is interesting are pointed out.

Peering facilitates some technical benefits. First, it provides an improved quality of service (QoS). Due to the redundancy of the highly intermeshed infrastructure connections, the reliability increases. The traffic passes shorter and more efficient paths between request sender and response receiver. This results in lower latency and a smaller probability of packet losses. These two aspects are fundamental in order to offer high quality products to end customers.

An ISP with peering agreements and a widely expanded network has the possibility of configuring the load distribution on the different connection links as desired. This is a very interesting option in case of different relationships concerning the traffic costs between business partners. Consider an ISP A that has to send traffic to a requester whereas this is a customer of an ISP B and the two ISPs are connected with each other at IXP X and ISP C, whereas A peers with C, the transit provider of B. In this case, ISP A can choose to route the response traffic over the link at the IX where no fees arise for B or to send the response via C where B has to pay for the downstream traffic. This possible routing control is not to brand market power as shown in the example, but can help in case of bad performance situations where the configuration is altered suitably.

However peering has not only benefits. The required technical competences should not be forgotten or underestimated. If no BPG protocol specialist is available in-house the service for the routing setup and maintenance has to be purchased and this is aligned with costs to be considered. Another issue faced is the support problem. Among peering partners, there normally exist no service level agreements as between customers an transit providers. Thus in case of connection breakdowns or other technical routing problems there are no service efforts to be expected.

How and with Whom to Peer

StrategicDecisions

Strategic considerations are fundamental to the decision on the kind of peering to be implemented. Peering can be used to enlarge the own network, help to increase the company's attractiveness and improve the corporate identity at the same time due to increased publicity. One possibility is to publish the peering policy publicly, meaning to disclose the

own peering strategy with the requirements to be fulfilled by the potential peering partner. It is not worth to start a peering negotiation, if an interested ISP does not meet the requirements. Therefore, it helps not to waste time on finding favourable peering partners. Nevertheless not all ISP's publish their peering policy, neither is it mandatory. According documentations can often be found on a provider's homepage. Today, disclosing one's peering policy is very popular and enables a fundament for successful communication.

In order to attempt peering negotiations an ISP has to be aware of the risk never to gain a peering partner as customer. The slogan „once a customer never a peer" is often heard in this context. Thus an ISP should avoid peering with ISP's to be considered customers.

The market information asymmetry in the Internet sector is another obstacle to overcome. Information about competitors' peering strategies is hard to obtain in this fast moving domain, facts about effective traffic volume and characteristics of existing peering relationships in particular. An ISP can only monitor its own crosslinks but has no information about the load of other network connections it does not share.

Because not every peering negotiation effort avails as desired, peering coordinators have used different tactics to obtain peering notwithstanding. W. Norton's [21] observed tactic analysis demonstrates that the procedures are not only amiable. Instead of peering with some coequal ISP, one tactic seeks to peer with the others most important customer ISPs so that no further peering agreement with the larger ISP is required. Another approach is to manipulate the traffic by configuring the traffic routing temporary so that the other ISP has to send a lot of traffic back to the requester ISP. Even faked requests are sent to activate an increased traffic flow. If the peering requester achieves an agreement, the traffic load is processed again as before. Deferent tactics are mentioned fortunately as well and break new applied tactics can be tracked with interest.

Political Aspects

In addition to the strategic considerations to be done, amazing political moves can be noted. To avoid any contact or communication exchange with competitors possible effort are completely prevented and thus the peering offer refused. Another observable behaviour is to offer peering agreements only to customer-contract-like conditions. This attitude can be observed by large ISPs to maintain their market power. Thereby the peering for the interested party is however coupled with high costs and shows again the difficulty to distinguish between customer and peering relationships if the peering agreement is realized nevertheless.

Concluding this section the one of the possibly most important factors should be stressed out. Sometimes interpersonal differences prevent ISP's from peering, also because negotiations costs endeavours and not everything seems clear from the beginning. These matters have led to a surprising number of failed peering negotiations [18] and that the phrase „peering is all politics" is not completely devious.

Considerations

Recapitulating the impacts just discussed the following characteristics affect the decision whether, how and with whom to peer:

- The Company's Market Power
- Composition and Size of Customer Base (Content Providers vs. Content Recipients)
- Backbone Scope
- Existing Peering Agreements
- Business Strategy in General
- Market Know-how and Information about Competitors

Reasons not to Peer

Despite the advantages, there exist some reasons against peering agreements. Large ISPs do not have the same benefits by peering as smaller providers have. One disadvantage is referred as „backbone free-riding" and occurs if a nationwide operating provider A has to maintain a larger backbone than a regional provider B and is therefore not interested to provide B traffic exchange with A's customers in distant regions for free [1].

The second argument against peering agreements is mentioned by providers with a large content server base and illustrated in Figure 6.9 and 6.10. It is called the „business-stealing effect". Consider a customer of a large ISP A, whose main sites of interest are also customers of A. But ISP B offers a financially more interesting product but is not directly connected to network A and the customer would (by switching to ISP B) have to accept a higher latency when accessing his favorite sites on A's network. With a peering agreement between A and B this barrier to change the provider falls and A might lose customers to B. The quality of service for the customer increases, but from the point of view of ISP A it is not profitable and might thus be reluctant to peer.

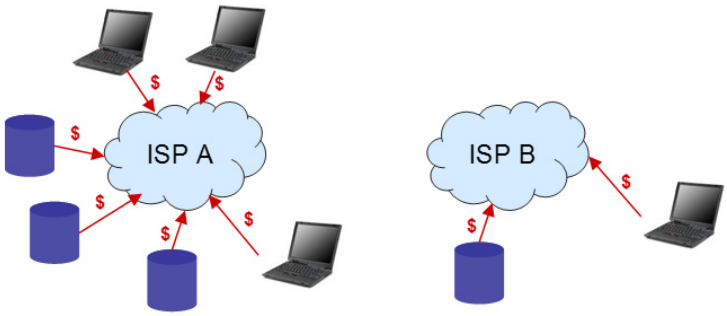


Figure 6.9: Business-Stealing Effect: Customer Allocation without Peering

To come to an optimal peering decision all these possible consequences have to take into account and reveal the complexity of the topic.

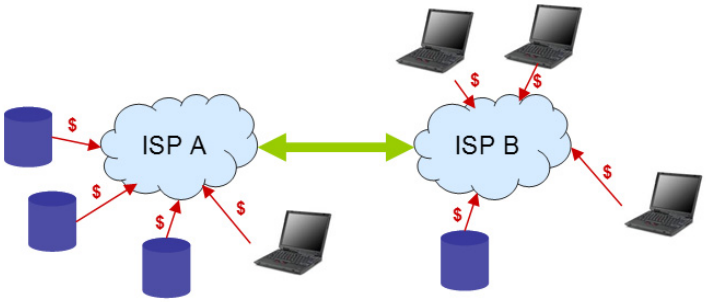


Figure 6.10: Business-Stealing Effect: Customer Allocation with Peering

6.4 Peering in Switzerland

6.4.1 Development

The CERN (European Organization for Nuclear Research, with its headquarter in Geneva) Internet Exchange Point CIXP, is a historical European Internet landmark, through which the first pan-European Internet backbone and the first transatlantic T1 connection to NSFnet were established in 1989 and 1990 [4].

Ten years later in 1999, the Telehouse Internet Exchange TIX of IXEurope (former Telehouse) started operating [33]. The set of the three Swiss IXP players has been completed by swissix in 2001 when building up their geographically distributed Internet exchanges [32].

6.4.2 Current Architecture

The following Table 6.2 shows the IXP's sites and the number of active ports. The total of connected ISPs differs from the figures shown below due to the fact that few providers are connected at several ports like in case of large traffic volume.

Table 6.2: Swiss Internet Exchange Points

IXP	Location	# of Active Ports (new)
CIXP	Geneva	29 (n/a)
TIX	Zurich	60 (4)
swissix	Glattbrugg	14 (2)
	Zurich	33 (4)
	Basel	6 (1)
	Bern	7 (3)

Topology

Today's Swiss interconnection points are distributed along the main transportation axis N1 with a concentrated area around Zurich. The following Figure 6.11 shows the geo-

graphical distribution of the exchanges. The allocation correlates with the demographic population density.

Different traceroute analysis confirm a high interconnectivity rate because cases where request and response originates, respectively ends in Switzerland and the data path does leave the country are very rare. The absence of a large tier1 ISP located in Switzerland concludes peering activity.

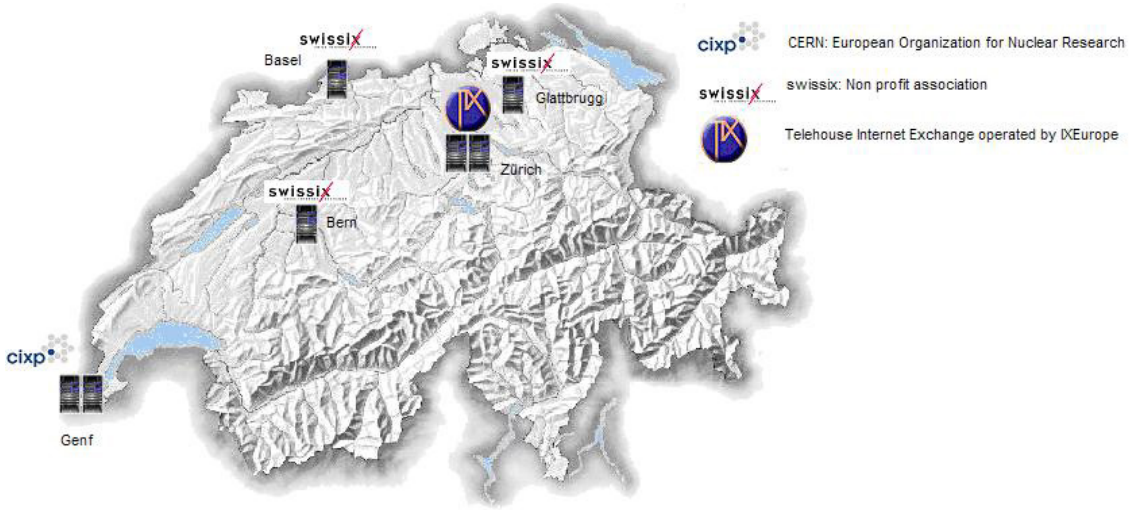


Figure 6.11: Swiss IX Topology

Special Role swissix

Besides the big IXPs in Zurich and Geneva there exists a distributed Internet exchange built and run by swissix, an association formed out of Swiss regional ISPs and carriers. Originated by an international operating company called Interxion, the unsuccessful attempt to establish a commercial Internet exchange has resulted in an association that runs different Internet exchanges. All members have the possibility to connect at any of the listed sites above (see 6.2) and to peer with present associated members. Except the sites at IXEurope, the infrastructure is sponsored by the provider of the hosting locations and the resulting accompaniment are neither costs for setup nor recurring cabling costs that arise.

Swiss Peering Conduct

Swiss ISPs (since only the Swiss market has been analysed, no conclusion are made about foreign handlings) provide peering agreements at the known IXPs. During the enquiries the already mentioned political moves could be reported and the various number of open communities affirm the shown importance of interpersonal contacts. Regular meetings serve to exchange experience, discuss ongoing or future issues and provide a platform for new business contacts. Since the topics are not about information management's

business development but more technical level, common participants are some engineering personnel. Significant differences to abroad have not been found, the situations seem to look roughly similar.

To find information about other ISPs' peering policy the import and export statements of the according AS definition can be queried in the Ripe database [24]. Still these information have to be handled with care as the data sets are updated by the owner of the respective AS and you only find information he is willing to make public or have been kept up-to-date.

An addition to be made is the observation that besides ISPs also enterprises agree to peer. At CIXP customers are allowed to have private peering, that means that two members can directly connect their router without using the CIXP switch. This usually happen between big companies that exchange large amount of traffic whereas those private peerings are seldom publicized.

6.4.3 Trends and Future

Since operating an Internet exchange does not gain economical revenue there is no trend to be identified. The Internet market is a global construct and the Swiss players' attitude does not vary from other foreign.

6.5 Settlement Models Today

As we have already showed, peering is mostly directly connected to large cost advantages for both partners. As long as the providers are about equal in terms of size and market shares and cost structures, a peering agreement is mostly a good choice. Lower costs for the traffic to and from upstream providers combined with better quality for the own customers weight normally more than the (mostly relatively small) costs for additional hardware and sometimes technical skilled (BGP skills) personnel. For this reason, peering was very attractive for ISPs until recent years. Real obstacles for a zero settlement peering agreement are mostly of strategic nature. Direct backbone freeriding can be prevented by BGP-policies, but if a large national ISP accepts a small regional ISP as a peering partner, the smaller one profits from a kind of a free „national backbone“. There is also the risk of the business stealing effect. The main problem is always an „unfair“ cost distribution.

UUNET - a major US backbone provider - was one of the first ISP in 1997 to declare that they will only continue to peer with other big backbone ISPs that fulfil certain criteria. Since then, there exists a hierarchical situation where in most cases ISPs can only peer for free with other ISPs of equal dimension/magnitude. A small local ISP generally has no possibility to peer with a big backbone ISP. Also does such a small provider not receive any access to the important big internet exchanges.

When there are no peering agreements, the ISP has to find a transit provider. As already shown inter-regional traffic over a global transit provider is connected with deadweight

losses (lower quality of service and higher costs for customer). So if an ISP declines a peering agreement with an other provider, this is connected with a loss for the customer of both parties involved.

Still, there can be found nearly only these two types of settlement models: peering or buying traffic. There are many possibilities for peering agreements, which would give the possibility for fairer cost distribution:

- Balancing advertised routes area
- Balancing traffic volume
- Compensation payments

But most ISPs choose - when they agree to peer - a zero settlement policy. Any further possibilities are probably often ignored in favour of simplicity of the negotiations.

6.5.1 Comparison to the telephony market

There are fundamental differences in the accounting of connection services between today's ISPs and the market of the classic telephony companies. In that market there is one dominating concept for the accounting and charging of interconnection services. First of all, the sender pays for the whole end-to-end connection (except for special R-Calls in which the receiver pays). Then the sender pays for the call time, and not for the actual traffic volume. The price per minute is determined through the bilateral pricing agreements between the involved telephony companies [10]. The telephony provider pays the next provider for the transit service of the call. In the Internet market there is normally a 'sender keeps all' situation. The prices in the telephony market are much more transparent, as a global call normally only involves 2-4 telephony companies and the route of a phone call is fixed. This cannot be assured in the Internet. Although settlement models similar to zero settlement agreements do exist in the telephony market they are extremely rare.

Compared to the telephony, one could ask if the Internet market could also use a more direct accounting for the interconnection services. Today's Internet raises the problem, that it is based on a best effort service, which means that packets can get lost and that the connection speed can fluctuate on daily or hourly rates (depending on the load of the respective links). Even the TCP protocol gives a possibility to determine if a packet has reached his end destination or not. Packet accounting on a best effort architecture seems to be far away from practical use by today [10]. For that, many issues as an effective accounting model or how to prevent abuse have to be considered first. The dynamic routes of the packets also provides another big difference to the telephony market and is a significant handicap for more transparent prices between ISPs, as a sender of a packet does not know his packet's exact route in the moment he releases it. All this factors hinder the adaption of the charging concepts of the telephony market.

But the question how long the Internet is based on best effort remains. N. Semret [28] believes that Internet has reached a period of transition. He states that today's Internet with the complex peering-hierarchy can be compared to the situation that the European postal companies faced from the 16th to the 18th century. At that the time, the delivering service between the postal providers was also dominated by complex bilateral agreements. Typical for these situations are that the receiver pays for the service as well. When QoS was introduced in the postal market (e.g. in the form of express or air mail), only the sender paid and new settlement models with a more direct payment for specific services were introduced.

Certainly, we cannot directly compare a 16th century market with today's complex and global ISP market, but the question who pays for the delivery of a packet and if there are different quality service levels is tightly coupled.

6.6 New Settlement Models

Between 1995 and 2000 the research has provided many new settlement models for the interconnection service, but only very few of them have yet found practical use. At that time the Internet was at the beginning of his boom and many new application with a need for high quality links became popular. So congestion of the given links was big concern. Today, these issues seem not to be of the same importance anymore, as the transfer rates increased up to now enough to fulfil most of the needs. But considering the fact that most important telephony providers will in the next couple years also switch their networks on the common IP-technology, these issues will come up again. Today, the motivation of this research lies maybe more on better economical models for the ISPs, since the restrictions on peering or transit models are suboptimal.

As already mentioned, many different settlement models have been suggested. They can be categorized by different criteria [23]. The most important one is the architecture of the Internet they build on. Today there are mainly three different network architectures:

- Best-effort
- DiffServ
- IntServ

It is also possible to categorize the settlement models on their charging units. Theoretically, it is possible to charge per packet, per flow, per reservation or as today per contract.

Then there are different pricing strategies, meaning the ISP's price-setting model for a specific transmission service. Today, there is mostly a direct cost sharing between connected ISPs. An auction of their services (among the participating ISPs) to a potential customer could gain popularity in the future.

The chosen factors have an impact on the following three dimensions, as defined by Nguyen [17]:

- Economical efficiency
- Technological efficiency
- Social impact

As an example: Today's Internet architecture is based on a best effort service with no QoS SLAs. Prices are set by contract on a bilateral peering or transit basis (cost distribution or flat rate for a certain capacity) between the different providers. The low technical overhead for charging and accounting leads to a high technological efficiency. As discussed broadly in this work, the economical efficiency is far away from optimal (deadweight losses) and there is also no social impact on the use of congested links through any price differentiations.

6.6.1 Settlement Models in a best-effort Service

Best-effort is the state of the Internet today. It is packet based (connectionless) and has no defined QoS-levels. The Smart Market Model and the Paris-Metro Pricing are two suggested settlement models which try to bring a certain degree of quality distinction into the best effort network. But QoS constraints are not possible without a DiffServ (soft) or a IntServ (hard) architecture.

Smart Market

The Smart Market model as proposed by K. MacKie-Mason [14] uses a sender-based auction to determine which package are forwarded or not when a link is congested. The sender of a packet has to indicate what is his willingness to pay for its transmission. This value is recorded by a „bid"-field in the packet-header. As long as there are no congestions, the router simply forwards the packets received as common today. When the link reaches a state where the attached router is not able to forward all packets anymore, the respective bid-fields come into consideration. If the value of the bid-field is higher than the market-clearing price, the packet is forwarded, else not. The market-clearing price is determined in a simple way. It is the minimum bid of all the packets that are forwarded (Vickrey Auction), or in other words, the price of the package that „ just made the cut". The sender of the forwarded packets is then charged with the market clearing price for these packets. All the other packets are dropped by a boarding router of the congested link. It is important that the market clearing price is used instead of the actual bid, because else there would be the risk that the sender do not bid their true values, but try to manipulate the competition.

The advantages of this model are that the sender is priced a „fair value" for its packets and that it takes the social costs for delivering the packet into account (sender is only charged extra if other packets have to be dropped out). In the terms that were introduced above, the Smart Market has a high social impact. But it also has low technological efficiency, because it brings a big accounting overhead. There has to be a complex auction on every

router that is anyway already congested, and then there are many small amounts to be charged globally. Additionally the problem that already charged packet might get lost later on their transmission path, is not considered.

Paris Metro Model

The Paris Metro Model of A. M. Odlyzko [22] is by far simpler to implement as the Smart Market, but gives also less pricing flexibility to the providers. The main idea can be derived from most public transport train systems, which have different classes. As fewer people are willing to pay extra for the transportation service, those means of transport have normally at all times open seats, while the customer with second class tickets are used to ride to work or home, standing in the hallway during rush hours. While a second class passenger normally always can squeeze himself into a full car, IP packets are often just dropped by the routers because of a buffer overflow. In the case of the Internet, it would be easily possible to divide a physical link into two logical links and charge the „first" class with a price to provide two different service levels. The „first class" channel would be more attractive for customers who are willing to pay extra for a better link quality. When there are too many packets through the first class link, and this link also gets congested, some customer automatically stop to pay extra, because their willingness to pay sink under the price for that service, and the first class link quality gets better again.

The self regulation of this system is its greatest advantage. The most important disadvantage is based on the problem of the best effort architecture itself: The sender is always restricted to the weakest link. If one provider dumps his first class net compared to the rest too much, than all first class packets that cannot find an alternative route are restricted to the quality of that link.

6.6.2 Settlement models and the DiffServ Architecture

DiffServ and IntServ are considered shortly here and only as comparison to the new settlement models for the best effort network architecture, because they are very complex and therefore build a separate area of studies.

The DiffServ architecture tries to introduce QoS guarantees into the Internet. For our purpose every ISP's autonomous system (AS) is presented in this system as one „DiffServ cloud". Between these clouds service level agreements about the QoS of a bulk of data flows are negotiated. A settlement model can be built when border routers of the ISP's AS serve as brokers for the DiffServ architecture [6]. In that way from the sending point A to the receiving B, there are many explicit SLAs between each neighbouring ISP involved, which results in an implicit guarantee of QoS between A and B [31]. This guarantee is in no case a „hard" service level restriction. Such QoS guarantees can only be provided with the IntServ architecture.

Before the data is sent, the SLA have to be defined and priced. One of the parameters that the neighboring brokers have to negotiate is the volume of the data that is going

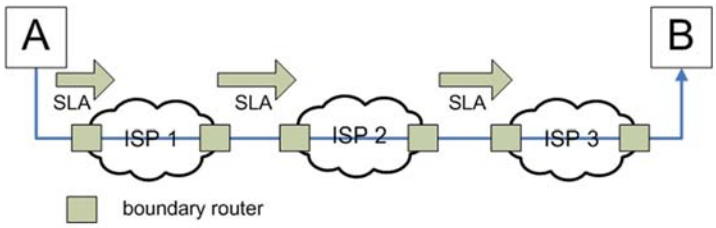


Figure 6.12: The DiffServ architecture with AS boundary routers as brokers

to be sent / received. The problem of the DiffServ architecture is then, that there is no way to finally predict in advance how much volume has to be reserved. So either a broker overbuys traffic (reserve more than it actually will need) or then there is a signalling overhead, because for every little data flow there has to exist the whole signalling and negotiating process. Even there are SLAs, the main problem of the Paris Metro model exist within the DiffServ settlement model, too. The end-to-end connection depends always on the weakest link within the path of the IP packets. But at least there is a soft guarantee (SLA) for the end-to-end quality of the connection. It will be also difficult to realize a price transparency for the customer of a DiffServ connection, since all prices / SLAs exist only for specific data flows and time and can change dynamically.

6.6.3 Settlement Models and the IntServ Architecture

The IntServ architecture [3] delivers compared to DiffServ an end-to-end QoS guarantee. B. Stiller [30] and M. Karsten [11] suggest to use the RSVP protocol to exchange accounting and charging information while setting up an IntServ connection. Each ISP can add the cost for the requested service according to its private price table to the PATH messages. As this information is sent back with the RESV message before any other data is transmitted, the receiver could also use an auction to choose the route for its packets (or the first outgoing ISP) based on the costs and QoS levels offered by the different providers.

IntServ with the RSVP protocol is up to now the most named future possibility to guarantee quality levels while having a dynamic market for these services (e.g. compared to setting up every time a VPN connection for having hard QoS constraints). Several problems are to be solved for using IntServ as base for charging and accounting of ISP services before. Examples are that every router involved has to be capable to handle RSVP. In the DiffServ case, only the AS boundary routers of the ISP involved have to handle the SLA, while internally the router can work as usual. Then there is also the problem to balance the signalling and accounting overhead (as well in the DiffServ case). And as it applies for nearly all settlement models: a lot of research has to be done about how security and efficient payment methods should be achieved.

6.7 Summary

Today's Internet structure basically offers two different possibilities for providers to interconnect. Either one ISP is the other's customer and pays for all traffic that is sent through and into the network of that provider, or they agree to peer and allow bidirectional free traffic. Peering interconnections are mostly located at an Internet exchange, either over the distributed network at the exchange or over a private cross-connection. The cost savings through peering is potentially high for both providers, as the traffic that flows through a peering connection would otherwise go through the net of a charging transit provider. It is important to understand though, that peering does not come for „free". The traffic load that is needed to exceed the peering breakeven point depends on the fixed cost of infrastructure for the peering compared to the cost for the same traffic handled by a transit provider. As we have shown there are also many further to consider, such as the cost factor to acquire technical know-how or political or business strategic backgrounds. Peering normally improves the reliability of a network and reduces the delay times. A peering point can also create a new backup-route in case of a failure of the upstream link.

When looking at today's peering situation in Switzerland we soon realised that the market is too small to have a significant different situation than the rest of Europe. We have a total of three IX organisations where most of the large ISP's are connected. Interesting is the fact that not only ISP's are peering, but large enterprise networks as well. Another point to be mentioned is the existence of a small non-profit ISP that is aiming to offer peering-opportunities for small enterprises and ISP's.

When the Internet started as a non-commercial network that was mostly used by universities, peering became a very popular subject. Since 1997 big backbone providers started to realise that there can be strategic disadvantages related to peering (freeriding and business stealing effect). As a result it is difficult today for a small ISP to find willing peering partners, which again leads to deadweight losses for the whole economy. Put it in other words we are today still operating an inefficient market. This inescapably leads to the question about how the market will develop.

To answer this question there are comparisons with the telephony market, where it was discussed whether the Internet market could also use a more direct accounting for the interconnection services. The main factor hindering the Internet from adapting these charging concepts is the fact that it is still a best-effort service and cannot guarantee timely package delivery or route-pre-selection as we have it in the Telephony market. Another comparison has been drawn with the European postal market, where in the introduction of QoS in the 18th century lead to new settlement models where the receiver was released from charges for the first time. Together these comparisons lead us to believe, that the Internet actually is in a transition period and that there is potential for new settlement models to emerge in the near future.

We mentioned that settlement models can be categorised after the three criteria Architecture, Charging Unit and pricing strategy. We decided to put the focus on the architecture criteria and distinguished between the three architectures Best-effort, DiffServ, IntServ when studying the models. Starting with best-effort-models we explained the *Smart Market Model* that is built like a sender based auction and the *Paris-Metro-Pricing* model

that offers one equal service for two different prices in order to achieve a self-regulating market. The main disadvantage of these best-effort models is that they still result in best-effort service and can thus not guarantee a high QoS.

The DiffServ architecture is a first trial in introducing QoS guarantees into the Internet, but can only offer „soft “ service level guarantees, whereas the IntServ model is much stronger and shows high potential in QoS. Today IntServ is not known good-enough and still needs a lot of development.

Bibliography

[1] P. Baake and T. Wichmann: On the economics of Internet peering, Netonomics vol. 1 no. 1, 1999.

[2] Joseph P. Bailey: Economics and Internet Interconnection Agreements, Presented at MIT Workshop onInternet Economics, March 1995.

[3] R. Braden, D. Clark, S. Shenker: Integrated Services in the Internet Architecture: An Overview, IETF RFC 1633, June 1994.

[4] CERN's Internet Exchange Point CIXP, `http://www.cixp.ch`.

[5] R. Dewan, M. Freimer und P. Gundepudi: Interconnection Agreements between Competing Internet Service Providers, Proceedings of. the 33rd Hawaii International Conference of System Sciences, 2000.

[6] G. Fankhauser, B. Plattner: Diffserv Bandwidth Brokers as Mini-Markets, Workshop on Internet Service Quality Economics, MIT, 1999.

[7] Lixin Gao und Jennifer Rexford: Stable Internet Routing Without Global Coordination, IEEE/ACM Transactions on Networking, Vol. 9, No. 6, Dezember 2001.

[8] Steve Gibbard: Internet Mini-Cores, pch.net, 13/02/2005.

[9] Steve Gibbard: Economics of peering, Packet Clearing House/Gibbard Consulting, Oktober 2004.

[10] Geoff Huston: Interconnection, Peering, and Settlements, Proceedings of INET99, 1999.

[11] M. Karsten, J. Schmitt, L. Woff, and R. Steinmetz: An Embedded Charging Approach for RSVP, International Workshop on Quality of Service in Napa, California USA, May 1998.

[12] Fredy Künzler, Init 7, Zürich - Interview about Peering in General and the Swiss particularities, December 1st, 2005.

[13] N. Luethi: Knotenpunkt im Zivilschutzkeller, Der Bund, 8/07/2005.

[14] K. MacKie-Mason, R. Varian: Pricing the Internet, Proc. of Public Access to the Internet Conference at JFK School of Government, April 1993.

[15] S. Marble: The Impacts of Settlement Issues on Business Evolution in the Internet, 26th Telecommunications Policy Research Conference, 1999.

[16] T. McGarty: Peering, Transit, Interconnection: Internet Access In Central Europe, Presened at the MIT Internet & Telephony Consortium Conference, 2002.

[17] T. Nguyen, G. Armitage: Evaluating Internet Pricing Schemes A Three-Dimensional Visual Model, ETRI Journal, vol. 27, no. 1, 2005.

[18] William B. Norton: A Business Case for ISP Peering, Equinix.com, 19/02/2002.

[19] William B. Norton: Internet Service Providers and Peering, Equinix.com, 30/05/2001.

[20] William B. Norton: Interconnection Stategies for ISPs, Equinix.com, 21/04/1999.

[21] William B. Norton: The Art of Peering: The Peering Playbook, Equinix.com, 25/06/2002.

[22] A. M. Odlyzko: Paris Metro Pricing for the Internet, Proc. of the 2nd International conference on Information and Computation Economies, November 1999.

[23] P. Reichl, B. Stiller, S. Leinen: Pricing Models for Internet Services, Netnomics - Economic Research and Electronic Networking, Vol. 2, No. 3, 2000.

[24] RIPE Network Coordination Centre, `http://www.ripe.net`.

[25] RFC1930: Guidelines for creation, selection, and registration of an Autonomous System (AS), Internet RFC/STD/FYI/BCP Archives, `http://www.faqs.org/rfcs`.

[26] RFC1771: A Border Gateway Protocol 4 (BGP-4), Internet RFC/STD/FYI/BCP Archives, `http://www.faqs.org/rfcs`.

[27] N. Semret: Peering and Provisioning of Differentiated Internet Services, UCLA and Invisible Hand Networks, Inc., 2000

[28] N. Semret, R. Liao, A. Campell, A. Lazar: Pricing, Provisioning and Peering: Dynamic Markets for Differentiated Internet Services and Implications for Network Interconnections, IEEE Journal on Selected Areas in Communication, Vol. 18, No. 12, 2000.

[29] Padmanabhan Srinagesh: Internet Cost Structures and Interconnection Agreements, Presented at MIT Workshop on Internet Economics, March 1995.

[30] P. Reichl, S. Leinine, B. Stiller: A practical Review of Pricing and Cost Recovery for Internet Services, Proc. of the 2nd Internet Economics Workshop Berlin, 1999.

[31] G. Dermler, M. Günter, T. Braun, B. Stiller: Towards a scalable system for per-flow charging in the Internet, Applied Telecommunication Symposium, Washington D.C., U.S.A., 2000.

[32] Swiss ISP and Carrier Association running several Internet Exchanges, `http://www.swissix.ch`.

[33] IXEurope's Internet Exchange Point TIX, `http://www.tix.ch`.

Chapter 7

Financial Clearing for Roaming Services between Mobile Network Operators

Tobias Schlaginhaufen, Martina Vazquez, Pascal Wild

In today's society mobile communication is a rising necessity. People are used to use their cell phone around the globe and demand an equivalent availability from their mobile network operator. Due to the economical fact, that a worldwide network infrastructure is far too expensive for one operator, they have to work together for offering connectivity in other countries. This kind of cooperation between operators is known as roaming and this paper shall give an overview about the financial clearing processes for services of such an interworking.

The documentation builds on an economical and a technical part. The economical part is focused on the different kinds of possible roaming agreements between operators and the pricing models which are used in praxis. The required technical infrastructure, incompatibilities and the used call records for the clearing and settling process are the main topics in the technical part.

Contents

7.1 Introduction

International roaming is defined as the service being provided from providers to costumers in order that they can call, be called, send and receive data outside and from outside of their home country without „interruption in service and loss in connectivity“[9]. After the introduction of the cordless telephone in the US in the 1970ies [10] the importance of the mobile phone and with this also the international roaming grew incredibly and today, thinking of a live without the mobile phone and its service is practicably impossible. Worldwide, nearly everybody has nowadays a mobile phone, even children. There are several aspects why the mobile phone has so much success today. One of them is certainly fast help. If something happens, one is able to dial immediately the emergency numbers and it does not take hours to look for a phone cabin. This can be quite important if people are in deserted places or if they drive in case of an accident. The main aspect of roaming however, is to call people and exchange data with the same mobile phone from home and abroad. Whether the caller or the call receiver is in his home country or in a foreign country he can call and receive calls on his mobile phone. The latest generation of mobile phones (3G) offers another dimension of calling, of connecting people. With this mobile phone the call receiver can see the caller and the caller can see the call receiver. The costumer can use the 3G mobile phones nearly for everything as it is multi functional. For example he can watch films, buy things, book holidays, do transaction of money, consult the information of his bank account and search for data in the Internet.

The international roaming, however, provides also a lot of jobs as telemarketing, help desk, vendor of mobile phones etc. worldwide. Providers of different countries work together. The help desk is maintained by people all around the world. In France for example when a customer calls the help desk of a telecommunication company, he sometimes gets connected to one of the same firm situated in Morocco or in Canada. In development countries such as Mozambique the international roaming can even be a significant economical resource in providing jobs and salaries [11].

Looking on the mobile phone from another view of scientific researches such as psychology, the mobile phone is not only favorable for the development of our society. Certain persons are addicted to the mobile phone. Calling here and calling there, they spend hundreds of Swiss francs a month. For various reasons the importance of the mobile phone increased immensely in the last few years. Today, in a world where communication and connection is one of the qualities everyone should have, the mobile phone is needed and appreciated.

Why offer international roaming? Today, the globalization is indispensably. Worldwide people work together, people travel and to use their own mobile phone or computer without having to change SIM card or computer while being abroad is a need, which is very high not only for private use but also for enterprises. Looking at statistics for what people spend money for, the communication has one of the key positions. In [24] it is stated that international roaming revenues make on average over 10 percent of the whole revenue of a provider and offers higher margin contributions. Moreover, the latest Informa Global Mobile Roaming report mentions that international roamers will increase and reach 850 million by 2010, which means four times the number of today. The revenue increase expected is about US Dollars 211.8 billion [24].

7.2 Economical aspects

7.2.1 Benefits for providers

Increasing income, being better as other providers and attracting year by year more customers is the provider's aim by offering the international roaming.

To succeed in their goals, providers have to attract costumers in order that they call in their roaming area. Therefore, their costumers have to make calls when they are abroad and foreign costumers have to call when they are in their country. Focused are costumers calling a lot such as enterprises. Giving their costumers the possibility to be reachable for everyone at each time of the day wherever they are, always on the same phone number, costumers will call and will be called more often. The bill of the costumers will be higher and the provider's aim will be reached by increasing their revenue.

If there was only one provider offering the international roaming and people would use this service, he would have huge financial advantages, but also a lot of work establishing roaming in each country of the world. Costumers needing the international roaming would change to this provider. However, if this same provider was the only one not to have the international roaming than his costumers may change to other providers.

Another aspect which can be taken into consideration is that costumers, who call from abroad, pay more and will not change operator if they are satisfied with the one they already have. This as well will increase the revenue of the provider.

When a provider is new on the market, he can have a lot of advantages, too. Usually, in the beginning of his existence, he has no problems with getting a good market position if he introduces well-appreciated ideas. The already existing providers, however, have then to rethink their strategies and come out with for instance a new product or service or to renew the existing product in order to survive.

Today, in a world in which globalization is unalterable, offering international roaming is a must for providers who want to be successful. To communicate even though being in a foreign country, has become one of the important needs for costumers. In fact, costumers do invest a lot of money into international roaming.

7.2.2 Benefits for customers

Customers want to be able to make and receive calls from wherever they are. International roaming is their solution. Making calls with the international roaming is simple and comfortable. Taking the phone, dialling the phone number and within seconds being speaking with the person, whom one wants to speak to. Another advantage of the international roaming is the billing. The customer receives one single bill from his home operator listing all the calls made from the home country and from abroad, in other words 'one phone, one number, one bill' [25].

7.2.3 What information needs to be exchanged between operators?

For establishing international roaming [3] operators need to know from each other which clearinghouse they use, the number of customer they have and other economical and technical aspects. Such elements could be for instance:

- Air interface: The air interface is used to communicate between the handset and the base station what means the frequency of the communication has to be decided on and the way the two stations interact. There are analogue and digital air interfaces. Different air interfaces are incompatible. Economically, it depends on the frequencies one wants to use and how well the two stations collaborate. Only by deciding on these two aspects the provider knows how much air interfaces will cost.

- Network requirements: Several network technologies are available to exchange information between providers but as air interfaces networks are incompatible. For knowing how much money it has to be spend on networks, there are two options a provider has. On one hand he can build a network by his own and spend a lot of money in doing so, or on the other hand he can use a network which is already done by another provider and save money. Which option the provider takes is a question of his budget.

- Switching systems: A Mobile Switching Centre treats digit strings and transmits them to the other provider's roaming area. In the Visitor Location Register data about the costumer such as the costumer has the right to call, where the caller is etc. is stored and transmitted to other providers. Here as well the different switching systems are incompatible. Some provider delivering SMS expect that they will be paid for this service but as other services this has to be mentioned in the roaming agreement. And till it is not mentioned, this service will not be factored.

- Signalling and numbering: Signalling is when control information is exchanged between network components. It is done to establish connections to control connections of the international roaming. Standards have to be followed exactly when signalling. Today, there are two different standards which are used: GSM MAP and TIA/EIA ANSI-41. They are also incompatible. When signalling and numbering costs can be saved by using the same standard.

- Data clearing and settlement: As already mentioned above, the data clearing is done by a clearing house. For further information see under how is financial clearing done between operators.

- Fraud management: Fraud management are concepts how to combat the wireless fraud which arise from stolen mobile phones, simple counterfeiting and tumbling to cloning fraud. Fraud protection is a concern which has to be exactly thought of, before signing a roaming agreement. Some providers do not sign agreements until they have the guaranty that their future collaborator has a fraud management system. Fraud is a key problem in the international roaming. Thousands of mobile phones are stolen and the thieves use the phones to call people. The problem is that

the costumer whom belonged the mobile phone gets factored for the calls made by the thief. The providers invest a lot of money to make mobile phones and their usage more secure. In our seminar Internet Economics II another Talk will be discussing this topic in more details.

7.2.4 Roaming Agreements

For being able to offer customers an international roaming the provider has to sign with providers all over the world roaming agreements. Problematic is that some of the countries as the United States, Russia, India and Brazil have regional coverage and not nationwide [3]. So the provider has for instance 20 different agreements with providers in India. Other problems are the time zones and the different billing record formats and settlement cycles. For charging costumers the provider needs a process for the exchange of roamer billing records with his roaming partners.

Usually, before operators decide to negotiate, they are searching providers with the same technical conditions, if possible, to establish the international roaming. Because the network technology used between operators is not always compatible, it is hard to find a solution if the agreement was already signed. Sometimes however, the operator does not have the choice and has to sign roaming agreements with operators not being compatible. Syniverse Technologies Inc. [3] and Comfone [14], which is from Switzerland, are enterprises which help operators in this case, but also in the case of billing to find an adequate solution. Generally, roaming agreements do have a standard part where it should be written how the exchange of protocol-files is defined and how the interconnection prices are determined. In the annex of the roaming agreement the operators are free to write their own arrangements about their country (Common Annex) and their own arrangements between themselves (Individual Annex) such as the name and positions of the people signing the roaming agreement, the location of their company, the people to contact when there are questions about the TAP-files or the billing, the persons who will organize the technical preparation of the international roaming and some more specified aspects [15]. According to Syniverse Technologies Inc. [3] there are some possibilities how the international roaming agreements can be done:

- Direct: This roaming agreement is done between providers. Each time the provider has a new roaming partner he signs a new roaming agreement with him. At the end a provider has for instance 100 different roaming agreements with 100 different roaming partners.
- Piggyback: Here as well the roaming agreement is signed between two providers but with the difference that both providers accept that the other provider expands his roaming area not only in his area but also in the roaming area of all his roaming partners with whom he signed agreements before.
- Consolidator: All providers, who sign this main agreement, which is done by one single operator, do have the same conditions of roaming. Each provider can use the roaming area of those providers who signed this master agreement.

- Alliance (consortium): Each single intermediary manages and makes signatures for multiple roaming agreements available. For each new provider of the alliance a new roaming agreement is established with each provider being already in the alliance. These kind of agreements can correspond to standard conditions, to conditions and to fees of the alliance or even make changes.
- Roamers Broker: Today, this agreement, which is based on the idea of GSM Working Groups, is one of the best one's. For the signing of this agreement there is one Roaming Broker who acts for some Roaming Beakers. This agreement supersedes the multiple bilateral roaming agreements in order that providers do not have hundred's of different roaming agreements, what means that expansion of the roaming area can be established much more simple.

After having decided on a form of international roaming agreement, there are two tests which according to Comfone and Weissbuch Mobilkommunikation have to be passed successfully before operators can open their roaming area for each other. The first is the IREG (International Roaming Experts Group) test where the connection to the fix net and the mobile net, the emergency calls, the locking of the calls and the sending of SMS are verified. Are all these aspects in order, the second test can be done. It is the TADIG (Transferred Account Data Interchange Group) test done to know if calls made in the other operator's roaming area are written down correctly. For this, TAP (Transfer Account Procedure, see page 212) files are needed. In the TAP-files there is written which costumer called in the foreign roaming area at which time and for how long. Only now, when this test is well done, too, the operators decide upon a time when they will open each other their roaming area. [14], [15]

Problem of the direct roaming agreement

Here, an introduction into the problem why providers do not always publish their roaming fees and why there are firms like Comfone. For being able to offer customers international roaming, the provider has to sign roaming agreements with providers all over the world. Problematic is that some of the countries as the United States, Russia, India and Brazil have regional coverage and not nationwide [3]. If a provider signs the Direct Agreement, than he has to sign each time he wants to use a new roaming area, an new agreement. Here an example, which shows well this problematic:

For example for n = 500 operators, the number of bilateral roaming agreements would be 125 000.

$$\binom{n}{2} = \frac{n!}{(n-2)! \cdot 2!} = 125000.$$

Swisscom for example has 400 roaming partners in over 170 countries [31].

Other problems are the time zones and the different billing record formats and settlements cycles. For charging costumers the provider needs a process for the exchange of roamer billing records with his roaming partners.

7.2.5 Problem of the fees

For being able to understand how the fees for international roaming are being calculated, there has to be differentiated between the fee the costumer has to pay and the fee providers negotiated. The agreements are based on it. First, there are the fees for roaming of the home provider which appear while making calls in a foreign country. To receive SMS is out of this charge. Second, there are the IOT's, the inter operator tariffs, which are not used from all providers. This fee is only used by GSM providers, providers in the US who operate with CDMA do not have this fee. Introduced in the 1999 by the GSM operator trade association for decreasing the fees and establishing fairness and nondiscrimination, the IOT's are more or less equally offered by all GSM providers. One GSM operator gives all the other GSM operators the same wholesale roaming fees except the Irish provider Meteor which has three different kinds of IOT's. Before 1999 the operators had a normal network tariff, the NNT. Introducing the IOT's, the fees increased [26], [30]. IOT is the fee which an operator of a foreign country charges an operator of the home country, when costumers of the home operator are roaming in the foreign country. The IOT is the fee one provider has to pay to another provider. At the end the costumer pays a combination of the IOT and the given surcharge from the operator [30], [26], [24]. High IOTs charge the costumer because he has to pay them. The surcharge of the home operator has increased since 1999 from 15-25 percent to 35 percent [26], [30]. In Europe there are strategic alliances: FREEMOVE: Telefonica, Telecom Italia, T-Mobile International, Orange or EUROPE STARMAP: Amenia, Eurotel, 02, One, Sonofon, PannonGSM, Sunrise, Telenor Mobil, Wind. Although in this strategic alliances reductions are done on fees, the IOTs are still high [27], [28].

7.2.6 Disadvantages of international roaming

For providers it is a great challenge to offer international roaming because there are many economical and technical aspects which have to be thought about. This process takes a lot of time and is a complicated issue. The Syniverse Technologies Inc. [3] is an enterprise which helps providers to realise fast their own roaming area nationwide and worldwide so that in realizing the international roaming they do not lose time and all their costumers. For costumers however the disadvantages of international roaming are important due to the lack of financial information of providers. Costumers never do know how much they have to pay for making a call or receiving one. The IOTs can be changed, cannot be available everywhere to costumers and sometimes they are too old. Here, the wholesale tariff is spoken about, which the costumer has to pay.

In the following two figures the above mentioned problem of the pricing is very well shown. In Figure 2.1 prices are shown of different kinds of services English costumers can use while being in Spain. According to a survey done in March 2005 [32] the parliament came to the conclusion that there were huge differences in pricing prepay costumers. The light blue bars in the figure are sometimes very high. Prepay costumers of O2 have so high prices that even the wholesale tariffs of the Spanish provider are lower. What providers sometimes tried to do is to make the prices equal in order that the prices were higher, the prices of the local call were increased to the level of the call made to the home country. In

the second figure however, char bars show the prices for Irish costumers with Vodaphone , who are roaming in France or who are calling to France. Surprising is that in the third bar, the received call with a prepaid SIM card, the prices differ a lot between the providers. The costumer surely asks himself what is the price he will be paying for the roaming by receiving a call.

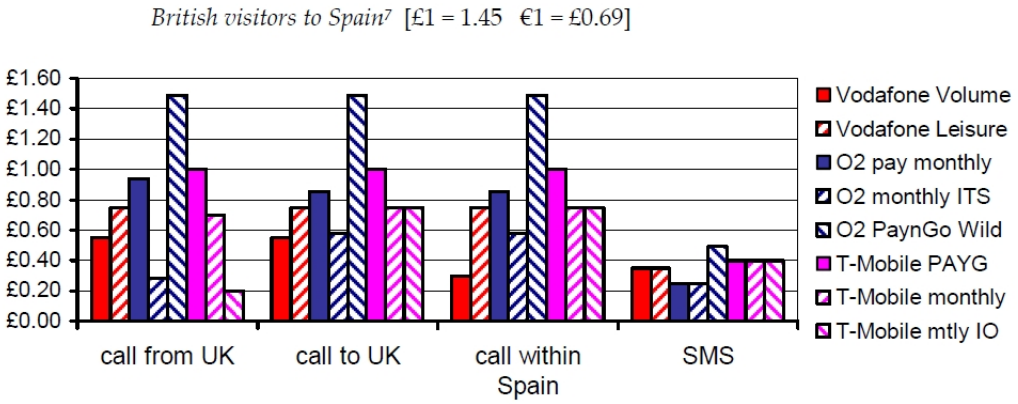


Figure 7.1: Source: [32]

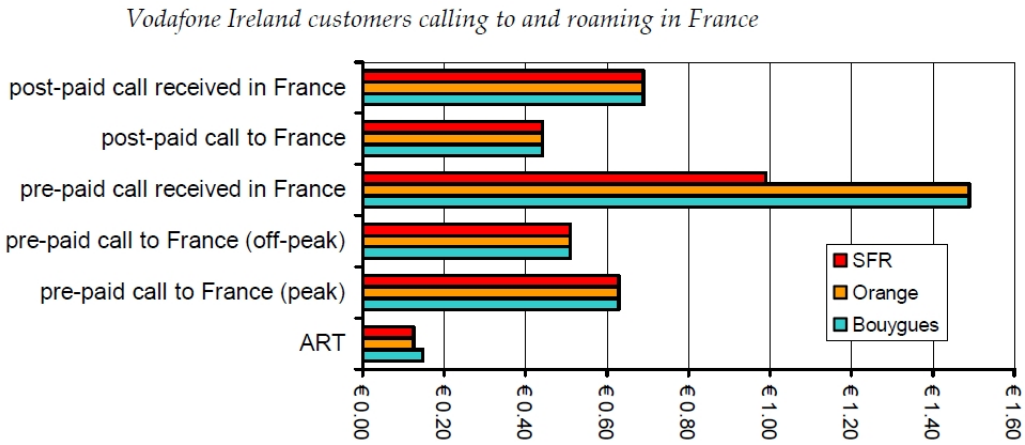


Figure 7.2: Source: [32]

IOTs have to be used for at least 6 months before they can be changed for a network inter-operator fee or for an introduction of a new product. The IOTs can be changed a lot of times. Swisscom for instance publishes the retail fees on the net (because the fees can be changed easily) or in brochures. The retail fees are divided in outgoing connection countries from A to D with the prices the costumer pays for in being in country D calling to country A and in incoming connection countries, which are also divided in five groups [31]. As mentioned above the roaming fees are the combination of the IOTs and the surcharge of the provider.

Usually, the costumer's bill is high because of the IOTs and not because of the surcharge of the provider. The provider tries to attract more and more costumers and offers the same fee and optional programs how a costumer can use the international roaming much easier. The reason in doing so is, that the more roaming minutes one provider can give the

other provider, the other provider would give reductions. For instance Orange is trading the fees all the time new.

According to the Telecommunications Policy Research Conference hold in Alexandria, USA, in the year 2000 [7] most of the providers did not let the costumers know their international roaming prices. On the web pages there was no information and if there was some, so not enough. Mails were written, calls done and faxes sent to the costumer care numbers but no relevant information was given either. On a web page of Proximus, a Belgian provider, on the Telecommunications Policy Research Conference the following text was found out: ' The prices quoted in the brochure are merely given as information and depend on the foreign networks as well as on the current exchange rates. Belgacom Mobile N.V./S.A. can in no way be held accountable for any changes or discrepancies. Belgacom Mobile N.V./S.A. can not guarantee the availability of services offered abroad. Foreign operators may decide to change them without prior notice.' The question, which was asked then, was what costumers could do, because this procedure was not fair at all, but done often also by other providers not only by Proximus. In Greece for example it was Panafon and TeleSTET, in Italy Telecom Italia Mobile and in Portugal Telecel. No of the mentioned providers gave information about pricing. Asking the provider of the foreign roaming area could not help either. The provider of the home roaming area could increase the charges of the provider of foreign roaming. Another problem of international roaming pointed out by the Telecommunication Policy Research Conference was the overcharging of calls.

Today, nearly 6 years after the above mentioned Conference, on the web page of the providor Proximus [12] the international roaming prices are determinate. TeleSTET [13] however does only mention that the price for international roaming is the international call rate. Costumers need two kind of information: 1) they need to know how much the call will cost at the precise moment of the call. In the best case they would have all the different prices of all the different networks, in order to choose which provider is the cheapest one. And 2) they need the above mentioned information with the bill for having the possibility to check the due price. But what does the amount of the international call rate consists of, is the next question to ask, but no information is available.

Each week or month the interconnection prices between operators change. This is another disadvantage for costumers because information about pricing is not updated for them fast enough. Time zones and different network partners make the pricing difficult and sometimes wrong. The charges for international roaming are counted twice or triple. Providers often do not know what price they should charge when during a call the caller has peak time charges and the call receiver has not. In Table 7.1 and 7.2 the differences between pricing can be seen. It depends from where the costumer calls to know how much he has to pay.

7.2.7 How are the interconnection prices determined?

For costumers

Firstly, there has to be differentiated between the different kind of calls that can be made:

Table 7.1: Denmark and Ireland prices in Euros (1999 in brackets)

Calling to Denmark from Ireland		EirCell	Esat Digifone
Danish subscriber	Sonofon	2.14 (1.85)	2.21 (2.11)
	TeleDanmark Mobil	2.14 (2.05)	2.21 (2.12)
Irish subscriber	non-roaming	1.25 (1.74)	1.24 (1.57)

Source: INTUG Europe data [7].

Table 7.2: Denmark and Ireland prices in Euros (1999 in brackets)

Calling to Ireland from Denmark		Sonofon	TeleDanmark Mobil
Irish subscriber	EirCell	1.95 (2.80)	1.36 (2.90)
	Esat Digifone	1.02 (2.20)	0.87 (1.96)
Danish subscriber	non-roaming	n/a (1.42)	0.93 (1.37)

Source: INTUG Europe data [7].

- home to home (same provider)
- home to home (different provider)
- home to abroad
- abroad to home
- abroad to abroad (caller is from the country, where the call receiver is)
- abroad to abroad (caller and call receiver are both abroad)

International Roaming is defined as calling from home to abroad, from abroad to home and from abroad to abroad. Calling from home to home with the same or a different provider is national roaming and consequently will not be explained.

- Home to abroad: The caller, staying in his country, has to pay for the roaming until the frontier and the call receiver, who is abroad, has to pay for the international roaming from the frontier to the place he is staying at. This is the only case that the receiver of a call has to pay for it. Receiving SMS is out of charge.
- Abroad to home: In this case it is the caller who pays for the whole connection, which is very expensive. To write SMS is expensive, too. It is highly recommended that costumers call with a local phone when they want to call home. This sort of calling does not increase the revenue of the home providor significantly.
- Abroad to abroad (caller is from the country, where the call receiver is): When both, the caller and the call receiver, are in the same county, the call receiver does not pay for international roaming and receives SMS for free.
- Abroad to abroad (caller and call receiver are both abroad): Both, the caller and the call receiver have to pay in this case. Usually, the caller has to pay for the international roaming to the home country and the call receiver has to cover the transmission fee of the call from the boarder of his home country to the country he is staying at.

For providers

The prices to use the international roaming vary from provider to provider. In the Telecommunications Policy Research Conference [7] it was said that it can depend on marketing strategies of operators, costs of access to international circuits, volumes of traffic, the ease of negotiating roaming contracts, comparative economic costs (labor, capital, etc), international exchange rates and license fees (especially initial charges). Exactly how interconnection prices between the operators are determined, which percentage of the price of one call which provider receives, is a professional secret and only defined in the agreements between the operators. Here, interconnection fees and not IOTs are considered. They are published and therefore there is more transparent, because of the Roaming Brokers. Also defined in the roaming agreement is how the charging is done [7] if it is done on price per second, per pulse or per unit. As mentioned, the interconnection prices are based on roaming agreements. Here are some options how they can be done [8]:

- Accounting Rate System: The Accounting Rate System wants to obtain that revenue between providers is equally distributed and not based on principles of commerce. This method is the oldest in pricing. Providers sign contracts where international calls being made in their countries should go through and decide which percentage of the charge each provider receives whose roaming area will be touched by a call. The distance and meters from providers are calculated. Today, this system is facing problems. This system is too old, the resellers do not use the settlements used by the accounting rate system and put the connection outside of the interconnection route, already agreed on. One provider takes the connection of another provider and gives it to a third provider.

- Sender Keeps All or Bill And Keep Arrangements: This system is very simple, providers invoice all calls made in their roaming area but do not pay international roaming of other providers. The idea is based on the thought that a provider having this agreement need other's roaming area equally as the other provider need his. The disadvantage of the Sender Keeps All or Bill And Keep is that the caller can phone only to a certain destination. Calls cannot get through providers roaming area.

- Interconnect Agreements: Interconnect Agreements are defined as a roaming service between two providers. These agreements are based on the Cascade Billing which means that the provider B is factored by a provider A for transmitting the call. Even if the call does not end by the provider B, A gets a terminating charge to the next operator of the chain. This has some advantages as the provider of the calling costumer does not have to have agreements with all providers the call is going through. Needed is only one agreement with one roaming partner and to have the shortest route from the caller to the call receiver. In the interconnect arrangements there are three different kinds of charging:

  - Distance Based: Here the interconnecting charges are factored by distance. The provider looks if it is a local, national or worldwide call and whether it is short or long. According to Intec Telecom Systems PLC the costs are calculated

on connection and not on the place the call came from. Problematic, are the charges which the costumer should pay. They do not necessarily correspond to fair charging of the calls done because the provider transmits the call in roaming areas which are not on the way from caller to call receiver and therefore the costumer has to pay much more. These agreements sooner or later will disappear.

- Revenue Sharing: Paying to the other provider a fixed percentage of the revenue for roaming in their area, new providers have to adapt themselves to the already fixed prices. These percentages are defined in the agreements. Innovation is not possible. This kind of agreement is used for new markets as it was done in Malaysia.
- Cost Based: Before signing the cost based agreements providers have to discuss how they want to fix the interconnection fee. Usually they agree that costs are calculated on the charges a call really costs and not on historical aspects of providers.

7.3 Technical Aspects

7.3.1 Functional principle of mobile communication

This section shall give an overview about the infrastructure of a mobile communication network. The focus is on the Global System for Mobile Communication (GSM) technology and the involved components.

GSM

The main components, which are involved in mobile communication, are shown in Figure 7.3.

Mobile Station (MS): The mobile station is your personal mobile phone. It consists of the mobile equipment (ME) and the subscriber identity module, known as SIM card. The SIM contains all required information for subscriber administration, access control and encryption.

Base Transceiver Station (BTS) and base station controller (BSC): The base transceiver station is responsible for the transformation from the air interface to the Abis-interface, which it is connected to over a base station controller. The base station controller controls several base transceiver stations and manages the radio link (e.g. power regulation) and the local handover. It is connected with a mobile switching center (MSC) over the A-interface.

The A- and Abis-interfaces are mostly PCM30 connections with thirty data channels, a signalling and synchronisation channel. The difference between the two channels is in

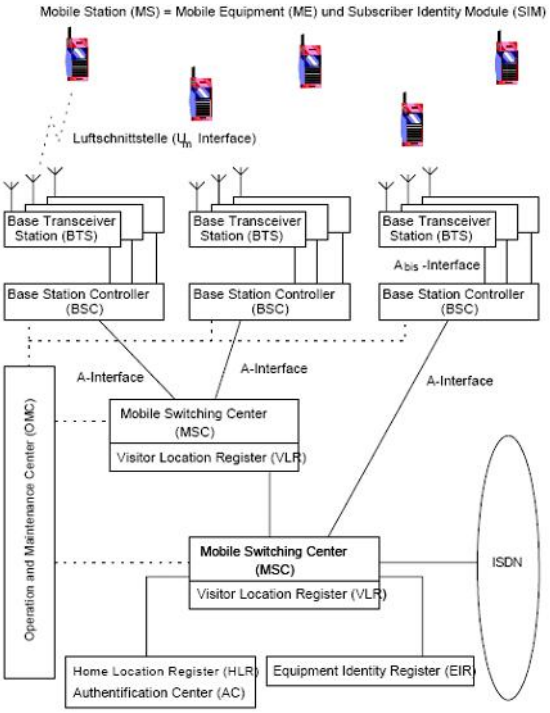


Figure 7.3: Architecture of a mobile network (GSM) [17]

signalling. The Abis uses GSM signalling and the A interface ”signaling system number 7” (SS7) which is used for conventional telephone networks.

Mobile Switching Center (MSC): The MSC, which is the GSM-switching center, is connected with several base station controllers and other MSCs. The MSC has to fulfill the following tasks:

- Switching of internal GSM connections
- Interworking with telephone network and ISDN
- Mobility management (roaming)

Operation and management center (OMC): The operation and management center supports the following functions:

- subscriber administration
- rates administration
- maintenance and net optimisation

A typical net for Switzerland builds on ca. nine MSC where each manages ten BSC and each BSC is connected to ten BTS. These are more or less 900 cells for the whole country (statement Sunrise).

The following part considers each component of the architecture and their functionality within the network.

Each mobile equipment which leaves the factory of a mobile manufacturer, gets a unique 15 digit identity number. With this so called international mobile equipment identity (IMEI), every mobile equipment can be identified around the globe. All mobile phones are registered in the equipment identity register (EIR). The EIR is connected with a mobile switching center of the operators network. The EIR administrates a white list with all phones which are permitted cell phones, a grey list with all supervised cell phones and a black list with the cell phones which are blocked.

When you buy a subscriber identity module (SIM), the card comes with a 15 digit international mobile subscriber identity (IMSI), which is a worldwide unique identifier for this card. The IMSI number uses 3 digits for the country identifier, 2 digits for the used network and 10 digits for the telephone number.

One mobile switching center of the network is connected with the home location register (HLR). The HLR stores for each user of this network the following data:

- International mobile subscriber identity (IMSI)
- Mobile subscriber ISDN number (MSISDN): ISDN number of the cell phone user
- Authentication key
- Information about the current location of the user: The last MSC where the user has been registered.

Each mobile switching center runs a visitor location register (VLR), which stores all information about the users, which are currently in the reception area of the MSC.

The first time the subscriber turns the cell phone with a new SIM card on, the mobile phone sends the international mobile subscriber identity (IMSI) to the mobile switching center (MSC). The MSC uses the IMSI for getting the subscriber information from the home location register and stores them in the visitor location register (VLR). Now the MSC generates a 32bit temporary mobile subscriber identity (TMSI) and a 40bit location area identifier (LAI) and sends them back to the mobile equipment. The location area identifier covers the current location with 3 digits for the country, 2 digits for the network and 16bit for the identification of the cell.

If the subscriber is moving within a network the location has to be updated continuously. For this the mobile equipment measures the signal quality regularly. If the quality gets to weak the cell phone automatically changes to another cell by sending the LAI and TMSI to another BTS. In case of the subscriber is not using the mobile phone at the moment there are two possibilities:

- If the new cell belongs to the same mobile switching center as the weak cell, there is only an update of the LAI in the visitor location register and mobile equipment required.

- If the new cell belongs to another mobile switching center, the subscriber information has to be copied from the previous VLR to the current and the TMSI and LAI have to be updated as well.

If the signal quality gets weaker during the phone is in use, a Handover will be initiated. In this case the following options are possible:

- Intracell handover
- Intercell handover
- Inter MSC handover

These handover functions are complex and time critical, as the subscriber shouldn't recognise the change.

International roaming is based on contracts between network operators. Each country and network operator has his own unique identification number, which is stored in the IMSI and LAI. The first three digits indicate the country and the following two digits the network operator (e.g. Swisscom = 22801 / Orange = 22803). During the sign on process the MSC recognizes, if the mobile equipment is from a foreign network. If this is the case the MSC demands the IMSI from the mobile equipment to find the home network operator. This is the place, where the roaming agreements come in. If an agreement between the two operators exist, there is a link to the HLR of the home operator and the subscriber can sign on at the local MSC.

GPRS

General packet radio service (GPRS) is a further development of GSM. GPRS is a packet switched data service, with the basic idea to use a physical channel for several users. With a packet oriented data service the efficiency of one channel is much higher.

For realising the general packet radio service the GSM architecture has to be extended with the following components:

- The serving GPRS support node (SGSN) is responsible for the packet oriented data traffic with the mobile equipment. The MSC is just used for signalling.
- The gateway GPRS support node (GGSN) establishes connections to other networks.
- The GPRS register (GR) stores all GPRS based data.

Before the mobile station is able to send any data it has to sign on to the SGSN. The SGSN checks in the user records of the GPRS register, if the subscriber is permitted to use this service. As result the MS receives a temporary logical link identity TLLI, which is required for the communication.

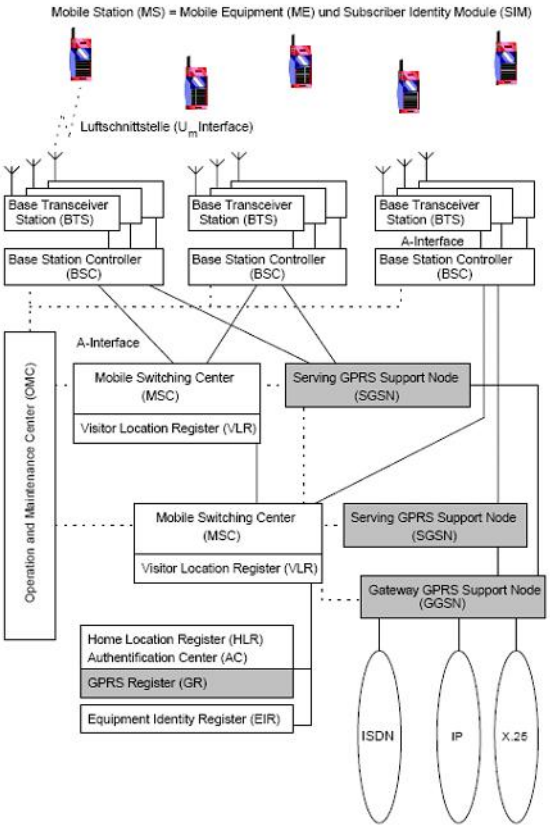


Figure 7.4: Architecture of a mobile network (GPRS) [17]

UMTS (3G)

Universal mobile communication system (UMTS) is the European standard for the mobile communication system of the 3rd generation.

UMTS builds on two parts, the radio access network and the core network. The radio access network contains the mobile station, the base station and the air interface, which is between the other two components. The core network connects the base stations among each other and with other networks like ISDN, Internet and others.

The core network of UMTS is a further development of the existing GSM core network. The radio access network, especially the air interface is a new development. This means GSM and UMTS are using the same core network, but the radio access networks are separated from each other [17].

7.3.2 Important technical issues for offering roaming services

This section is addressed to the technical requirements for offering international roaming successful. The first part focuses on the air interface and the second part on the core network specific technologies.

Before two operators can offer roaming within each others network they have to check the compatibility of their network technology used. The most important parts they have to pay attention to are listed below:

- Signalling protocol (TIA/EIA-41, GSM MAP)
- Air interface (CDMA, TDMA, AMPS, GSM)
- Frequency (800, 900, 1800, 1900 MHz)
- Switching system type
- Data clearing and settlement procedures
- Fraud prevention and management systems

Air interface

As discussed in the previous section the air interface is used for the communication between the handset and the base station. For this cooperation the compatibility, on the physical layer, of mobile equipment and BTS is vital. Incompatibility between the handset and the base station, because of differences in frequency band or interface technology used, create barriers to international roaming.

There exist many different technologies or frequency bands, as shown in Table 7.3.

Table 7.3: Common Technologies and Frequencies by Country [3]

Country	Technologies	Frequencies
China	GSM, CDMA	800MHz and 900MHz
United Kingdom	GSM	900MHz and 1800MHz
USA (C,D,E,...)	AMPS, CDMA, TDMA, GSM	1.8 GHz to 2.0GHz

To enable international roaming, where different technologies are involved the mobile equipment manufacturer had to develop equipment, which accommodates more than one of the air interface technologies. These kinds of mobile phones are known as dual, tri or quad band cell phones.

Core network

Due to the evolutionary process of the development of mobile communication, various infrastructure technologies to transport signalling, validation and other data exist. The following list is just a selection of the commonly used network technologies for wireless communication:

- ANSI SS7/ and ITU-C7

- X.25
- Frame relay
- ATM
- Transmission Control Protocol/Internet Protocol (TCP/IP)
- General Packet Radio Service (GPRS)

To enable international roaming the switches and data formats used by two roaming partners, have to be compatible with each other. Only then, they are able to exchange signalling messages, billing records and other information. At the moment this is just within countries or regions the case. But world wide compatibility is far away. Different approaches to cope with the incompatibility exist. One is to look for roaming partners, which are using the same technology. This approach reduces the choice of possible roaming partners in a massive way and an operator might even not find a compatible partner in some regions. Another and maybe cheaper solution is to connect to a network backbone provider which is already connected to other wireless operators. This kind of centralized point is responsible for the conversion of the different data formats and assures a smooth communication between different operators.

Switching systems

Other barriers for international roaming are the different mobile switching systems which exist on the market. The switch manufacturer extended the existing standards for the signalling interfaces with proprietary features. These extensions and the various interpretations of standards are responsible for the interoperability issues between networks using switches from different manufacturer.

Numbering

As mentioned in the section before the IMSI (International mobile subscriber identity) is used in the GSM world to uniquely identify a subscriber world wide. The first three digits of the fifteen digit long number are used to identify the country of the operator and the following two digits to identify the operator itself. Beside the IMSI there exists another identifier called Mobile Identification Number (MIN) which has been developed by the American National Standard Institute (ANSI). The MIN identifier is a ten digits long number and builds on the North American Dialling Plan. It was originally planned for the area of the United States and Canada. No one thought about international roaming at this time. With the number schema of the MIN it is not possible to distinguish between countries and it doesn't associate any international numbering plans. To ensure international roaming the unique identification of each subscriber world wide is an essential condition. The International Forum on ANSI-41 Standards Technology (IFAST) recommends for enabling international roaming switching from the MIN standard to the ITU standard IMSI. To assure backward compatibility from IMSI to MIN a special IMSI format for the United States has been defined.

[310 + 00 + MIN]

310 is the Mobile Country Code (MCC) and 00 is the Mobile Network Code. To support this MIN based IMSI universally a list of them should be shared among roaming partners. Each partner has to implement this list in their Mobile Switching Centre to ensure, that the MSC can recognise the identification and the registration is successful. The analog technology Advanced Mobile Phone Service (AMPS) is still used in North America for wireless communication. This kind of technology bases on frequency modulation FM and operates in the 800MHZ band. Because the IMSI approach is not supported by analog systems the world wide implementation of IMSI will not be realised as long as analog technology is in use.

Signalling

For establishing and control wireless communication signalling information has to be sent between the involved operators. This process of metacommunication is called signalling. Two main standards for signalling exist:

- GSM MAP - used by GSM operators
- TIA/EIA ANSI-41 - used by AMPS, N-AMPS, TDMA and CDMA operators

Without the use of a conversion point the standards GSM MAP and TIA/EIA ANSI-41 are incompatible with each other. If an operator is not interested in operating or using such a conversion gateway he has to look for a roaming partner who is using the same signalling standard. The cooperation between GSM operators shouldn't be a problem, but the operators which are using TIA/EIA ANSI-41 might have difficulties to roam within each others network. This is because of the many variations within ANSI-41 (Rev. C, Rev. D and Rev. E).

To cope with the interoperability between GSM MAP and ANSI-41 as well as the different versions of ANSI-41 it is recommended to use a centralized call processor-based Service Control Point (SCP). The SCP will then convert the signalling messages to the specific signalling protocol used by your roaming partner [3].

7.3.3 GRX - GPRS Roaming Exchange

Interconnection of phone calls, no matter if it's a mobile or a fixed line call, is (still) done using circuit switched connections. Because of historical reasons and also because there is a permanent communication channel open in voice calls, setting up circuit switched connection absolutely makes sense. Interconnection in packet switched data communication especially GPRS will be introduced in this section. The General Packet Switched Radio Service exists since '2.5G', that is, a technology between the second (2G) and third (3G) generations of mobile telephony. GPRS supersedes technologies like HSCSD.

When using GPRS in a foreign network, all data traffic is transmitted to the home provider through a GPRS Roaming Exchange network (GRX). If for example a mobile phone accesses an Internet server, which is located in the same town somewhere in a foreign country, the data travels all the way to the home provider, where the access to the public Internet takes place. One might think that it would be much easier to just access the Internet at the foreign mobile network provider. The reason that GPRS traffic is handled this way is the fact that many data services are subject to fees. Accessing a WAP site of a newspaper for example, is often liable to charges. These fees are levied by the home provider.

Because GPRS is packet oriented on the air interface, the underlying (fixed) networks understandably are packet oriented as well. With GRX an international packet data network was introduced designed to interconnect mobile network providers in order to provide GPRS roaming to their customers. A GRX is an IP based packet switched network. One might think of it as a worldwide private Internet. Compared to the public Internet, a GRX network offers quality of service guarantees and it is supposed to be secure, since the participating nodes are well known and small in number.

To offer GPRS roaming to its customers, the home operator and the foreign operator need to take part in such a GRX, not necessarily in the same one. The GRX networks are run by companies such as Comfone (Switzerland). For global roaming these GRX networks have to be connected together as illustrated in Figure 7.5. Such a connection point, called Global Peering Point, is AMS-IX, in Amsterdam. In 2004 Asia's first Peering Point for GPRS roaming was launched: Peering Singapore [6].

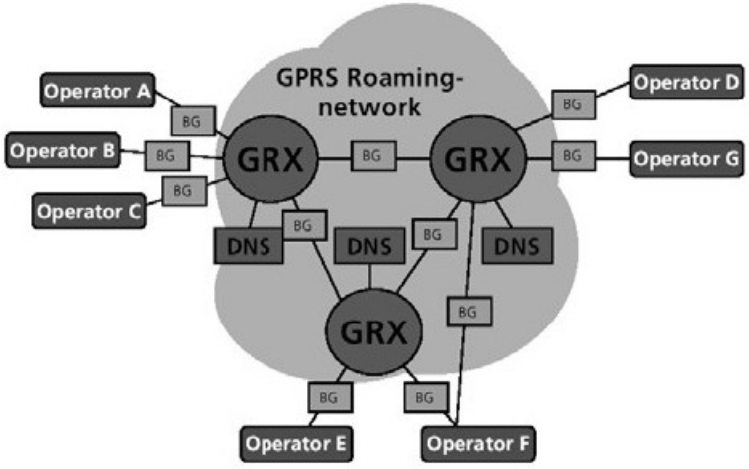


Figure 7.5: GRX networks. Source: [23]

GRX is standardized under the leadership of the International Roaming Expert Group (IREG). Mobile operators together with the GRX providers form the GRX Working Party, a sub-organisation of IREG. The GRX Working Party issues guidelines about different areas like end-to-end Quality of Service, Peering, Security and special services, like MMS relay services and DNS architecture [6].

7.4 Settlement and Data Clearing

This section outlines the settlement of roaming charges between mobile network operators, how information is exchanged and where data clearing is needed.

TAP - Transferred Account Procedure

An example of an established standard for exchanging billing information on roaming subscribers between operators, is TAP. The Transferred Account Procedure is a standard from the GSM Association [2], [3]. TAP is used by GSM operators while Cellular Intercarrier Billing Exchange Roamer (CIBER) is the billing format of ANSI-41 users. ANSI-41 is an industry standard from the American National Standards Institute which is used mostly by North American networks.

TAP describes a file format for exchanging billing information. TAP files contain rated call information according to the operator's Inter Operator Tariff (IOT), plus any bilaterally agreed arrangements or discounting schemes. The transfer of TAP records between the visited and the home mobile networks may be performed directly, or more commonly, via a Clearinghouse. Invoicing between the operators then normally happens once per month [2].

TAP explained by an example In order to explain the mechanism of TAP, consider this example (following [2]): A customer of a Swiss GSM Public Mobile Network (PMN) calls a fixed phone in Canada. The Swiss PMN operator does not have to negotiate a price with the Canadian fixed network operator, but with a Swiss fixed network operator. International fixed operators normally have negotiated agreements among themselves. So the Swiss fixed network operator charges the Swiss PMN for the fixed network connection from Switzerland to Canada. Those charges will be recouped from the Swiss PMN to the mobile phone subscriber. Because the caller is a subscriber of the network where the call originate, no roaming is needed and no TAP is used.

Let's assume that the customer travels to Germany and connects with a German PMN, which has a roaming agreement with his provider at home. He calls again to a Canadian fixed phone. Now the situation is different. The German PMN now charges the Swiss PMN for the costs incurred by the Swiss subscriber. This form of inter-PMN accounting is where TAP comes into its own. Information about such roaming calls are collected as a TAP record at the German PMN operator in this example. Then TAP file is the transferred to the Swiss PMN, which uses this information to bill his customer. On the other hand the TAP file serves as a basis for the financial settlement between the two network operator. Depending on the form of agreement they have, financial clearing is done by a financial clearing house.

Versions of TAP Parallel to the developments of mobile communication technology, since the first specification of TAP in 1989 and the initial standard of 1993, there have been a number of progresses which lead to the following versions of the standard:

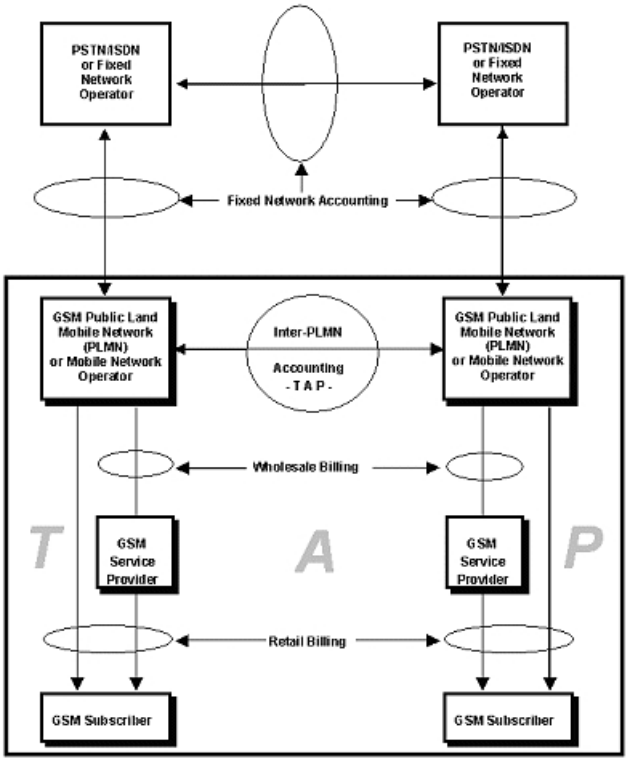


Figure 7.6: Transferred Account Procedure. Source [2]

- TAP1
- NA[1] TAP 2
- TAP2
- TAP2+
- TAP3

TAP3 is the current standard used by most GSM operators, although some earlier versions are still in use.

From the beginning, since TAP1, the three basic GSM service categories, Voice, Fax and Supplementary Services, have been supported. SMS roaming with third party Short Message Service Centres is supported since TAP2+. TAP3 differs from its predecessors on many ways: The file format became variable and contains more information than the predecessors, which were not expandable. TAP3 supports HSCSD and GPRS roaming. Value Added Services (VAS) are also supported since TAP3. Customised Application Mobile Enhanced Logic (CAMEL) phases 1 and 2 are supported in the current version of TAP3.

Full support of Inter Operator Tariff in TAP3 allows independent price determination. The visiting operator can give call level discounts and the home operator can then verify

[1]designed to meet the specific billing needs of North American GSM operators

that the discounts have been correctly applied. All GSM Phase 2+ services can be billed to roaming customers.

Rejects & Returns Process The Returned Accounts Procedure (RAP) was introduced in 2001 and is supported by TAP3. TAP files are validated at the home operators ARP. This standardized method for handling erroneous TAP files allows operators to reject call event details that do not conform to the TAP standard or to the terms of the roaming agreement.

RAP rejects not a whole TAP file but single calls. A TAP file with erroneous call data can still be processed which allows the operators to bill their subscribers in a timely manner.

Comparison TAP/CIBER In the following, some of the main differences between the billing record formats TAP and CIBER will be pointed out.

CIBER is used only by operators in the United States and Canada. The record format is maintained by CIBERNET Corporation. Since in the United States a lot of national roaming is needed, CIBER was developed for roaming among US carriers. TAP was originally developed for international roaming.

- Currency

  The transaction currency for CIBER is US dollar. TAP is used in (potentially) 220 countries. Hence currencies and exchange rates are an issue. Therefore the Special Drawing Right (SDR) is used in TAP files. SDR is a virtual currency set by the International Monetary Fund (IMF). SDR is valued based on the exchange rate of a basket of international currencies and therefore exchange rates are more steady. The SDR rates, which are used to convert the local currency to SDR, are issued every month [21].

- Subscriber identification

  In the TAP standard subscriber identification is done using the 15-digit International Mobile Subscriber Identity (IMSI). It has a 3-digit Mobile Country Code (MCC) that is assigned to a single country, and a 1-3 digit Mobile Network Code (MNC) that is unique to a carrier in that country.

  In a CIBER record, a subscriber is identified by the 10-digit Mobile Identification Number (MIN) which is not a worldwide unique number. For example, a MIN 2022441234, assigned to a subscriber in Washington, DC in the US, may look like a MIN assigned to a Brazilian subscriber [1].

  TAP3 has the capability to also carry the MIN, when available, to facilitate multi-standard roaming [3].

- Rejects and returns process

  Both TAP and CIBER have a rejects and returns process in place. For TAP it is standardized between clearinghouses. How a clearinghouse should handle rejects and returns with the operator over the private interface has not been defined.

- Cycle

  CIBER operators still send billing data by tape, either weekly or monthly. TAP operator transmit billing data electronically on at least a daily basis. A rapid data exchange is needed for calling plans that provide „buckets“ of minutes.

- SID/BID

  System Identifier (SID) and Billing Identifier (BID) are used by North American operators. This information allows to ascertain in which town a call was made. Without this information, only the country is known. TAP (since TAP3) and CIBER support SID/BID. For TAP it is not a mandatory field.

Data Clearing

The process of managing and exchanging roaming billing records with hundreds of partners, in different countries, in different timezones, with different currencies, with different settlement cycles and above all different data formats is very cumbersome.

That is why billing data exchange between operators is done by a data clearing house, more precisely by a worldwide network of data clearing houses. Such a clearing house serves as an operator's Authorized Receipt Point (ARP).

An operator sends its billing records to its ARP through the so called private interface. Such a clearing house processes billing data in many formats (see Figure 7.7) The data can be in any proprietary format, CIBER, or any version of TAP. The data can be on tapes or it can be electronically transferred. It is the clearing house's job to convert data to other formats and to validate it.

The clearing houses among each other normally exchange the latest version of TAP (TAP3) through the so called public interface. The latest version of TAP is always downward compatible with the older versions of TAP and also with CIBER.

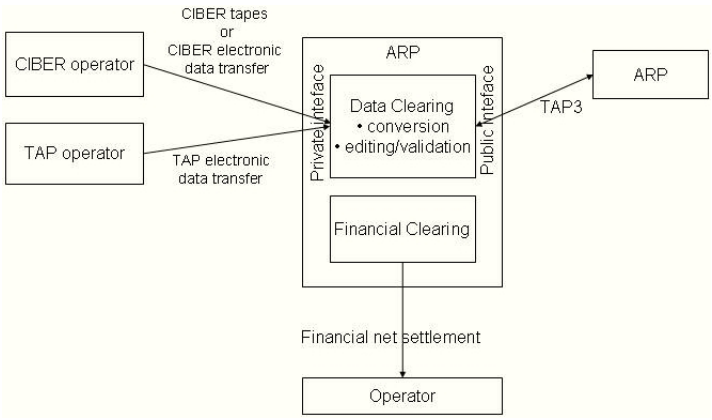


Figure 7.7: Illustration of some of the components of a global clearing system., according to [3]

Thus billing records are sent regularly from the foreign network operator to the home operator of a roaming subscriber. TAP files are even sent regularly when they are empty.

The billing records serve as a basis for the retail and wholesale billing. The home network provider uses the data for billing its subscriber.

Besides data clearing, financial net settlement is accomplished by the ARP. Depending on the roaming agreement, financial clearing is done either by billing all consumed communication services or by netting the balances on a regularly basis.

7.5 Losses resulting from fraud

Fraud is one of the expanding problems of international roaming. The worldwide losses resulting from fraud are estimated by CFCA (Communication Fraud Control Association) between $35 billion and $40 billion [19].

There are many different kinds of fraud, some on a technical level others base on social engineering and some are a mix of both. This section shall show you the different kinds of fraud exist and the developed mechanisms to combat fraud.

Different kinds of fraud

Cloning fraud: At the time of analog mobile communication this was one of the easiest possibilities to avoid paying for a service used. The aim of this attack is to steal the identity of a mobile phone (ESN/MIN) and program them into another phone. For stealing the identity of another phone the fraudster just have to scan the airwaves during the registration process between the cell phone and the base transceiver station. The cloned phone can be used now and all the charges which accrue will be added to the legitimate subscriber's bill. Consider this way of gathering the identity from the airwaves is only possible in analog technologies. Technologies like GSM encrypt the data channel of the air interface [20].

Subscription fraud: Subscription fraud bases on social engineering. The fraudster signs the wireless service contract with a wrong or stolen identity and will never pay for the ordered service. If the used identity exists the fraud is called true name subscription fraud [1].

Employee or reseller agent fraud: If the own employees are involved in selling secret subscriber identities to some suspects it mainly needs the most effort to ban this kind of fraud. The operators have to observe their employees and monitor all critical processes, to avoid this kind of fraud. A more social and trustful way to prevent employee fraud is to train and sensitise the staff about this matter. In this case the employees know how to react if one of their workmates commits fraud and they don't turn a blind eye on the problem [3].

Ways to avoid or detect fraud

As many different kinds of fraud exist as many methods to cope with them has been developed. The following types of fraud protection shall give you an impression about the possible techniques which exist.

The beating fraud techniques can be separated into fraud prevention and fraud detection. Prevention bases more or less on technical barriers to make fraud as difficult and expensive as possible and detection on analytical techniques for monitoring the traffic.

The following techniques are used to prevent fraud:

Authentication: To avoid illegal net access each subscriber gets authenticated during the login process with the Mobile Switching Center. For this the MSC generates a 128 Bit long random number, which will be sent to the mobile station. The cell phone uses the random number and the authentication key, which is stored on the SIM card, to generate the Signed Response (SRES). The mobile station sends the generated SRES to the MSC, where it will be compared with the number generated by the MSC. If the two numbers matches the subscriber will gain access to the network. After the authentication part the MSC generates a 114 bit key from the SRES and the user's authentication key to encrypt the following data stream.This technique is used within the GSM network.

RF fingerprinting: RF fingerprinting is used within ANSI networks. The technique uses each phones unique signal fingerprint together with the MIN/ESN. If the combination of these two properties is correct the subscriber is allowed to use the mobile network. This is an effective approach to avoid fraud. But due to some necessary hardware adjustments it is a cost intensive solution as well.

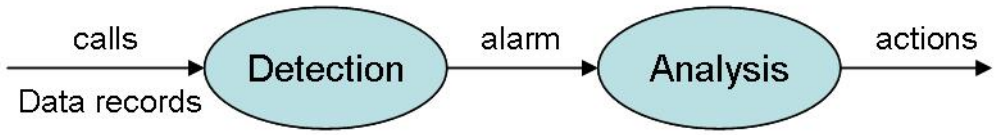


Figure 7.8: Detection technique

The detection technique can be separated in the parts detection and analysis:

Detection: Call data records are processed by looking for calling patterns or behaviour that is unusual for the subscriber. For example if a subscriber normally uses his mobile phone very seldom and from one day to the other the mobile phone gets used for many calls, the new behaviour contradicts with the stored pattern of this subscriber (smart thresholds). Other kinds of patterns are suspicious international country, black list checking or collision (call overlap). When the detection can determine an anomaly it generates an alarm.

Analysis: In this part each alarm from the detection component will be reviewed in detail. The system compares the alarm with other alarms from this subscriber and calculates the possibility of fraud. According to the system used for detecting and analysing fraud there are all different kinds of analysing features thinkable [18].

7.6 Summary and conclusions

Mobile communication is nowadays possible across borders of countries and across borders of different technologies. Offering international roaming is essential for mobile network operators to be competitive, but also pays off through additional revenues from the extra customers. Consumers are willing to pay high prices for roaming services and a price war for roaming charges does not look to break out. Primarily international travelling business people, being reachable using the same device, able to talk to customers, colleges and family in almost every corner of this planet is a priceless service.

Establishing a roaming agreement with another network operator is still a long winded process. The complexity can be reduced by not directly negotiating with every single roaming partner but rather use the possibilities of piggyback, consolidators, alliances (consortium) or go via a broker.

A roaming agreement mainly consists of the following issues:

- pricing
- technical infrastructure for the air communication
- signalling and identification of the device
- interconnection with different kinds of communication
- data clearing for the settlement.

Pricing The pricing of the international roaming is based on roaming agreements between the providers. There are different roaming agreements. Each time the provider wants to enlarge his roaming area he has to sign roaming agreements. This can last a long time depending on the roaming agreement. For costumers however the way of knowing the exact price they have to pay for a call making it or receiving it, is very difficult because of the lack of information either because the price is not published or the IOT has changed recently or because of the surcharge. Providers do have some possibilities to define the prices and costumers do have different kind of prices to pay. At the end the costumer pays a combination of the IOT and the given surcharge from the operator.

The IOT was first introduced to promote fairness. The bargaining of the price should have become fairer and easier with the introduction of the IOT. The IOT is a public price offer from which the operators can give rebates but also let costumers to pay too high prices. The IOT will not be sunk as there is no motivation in doing so.

Technical infrastructure If an operator is looking for a roaming partner in a foreign country the technology used plays an important part. To establish roaming services for his customers an operator has to look for a partner who is using the same or compatible technology. If no operator in the favoured roaming area is compatible, the operators have to build conversion points, where the different protocols will be translated into the format used by the partner. Normally this solution is too cost intensive, because today operators have to offer worldwide availability to their customer which results in many roaming partners. To avoid this immense hardware and operational costs the operators can outsource the roaming related business to companies where this is their core competence.

For GPRS roaming a network infrastructure is needed. Operators have to take part in a GPRS Roaming Exchange network (GRX). GRX networks are worldwide connected together and thus make GPRS roaming worldwide possible.

Settlement For the settlement of roaming charges between operators standard formats are used. The standard for GSM operators is called Transfered Account Procedure (TAP). North American operators use CIBER. In order to enable roaming across standards data clearing is needed.

Bibliography

[1] International Forum on ANSI-41 Standards Technology (IFAST): International Roaming Guide 1.4, (http://www.ifast.org/files/IRG-1V1_4.pdf), October 2003.

[2] GSM World: Tapping the Potential of Roaming, (http://www.gsmworld.com/using/billing/potential.shtml).

[3] Syniverse Technologies Inc.: International Roaming Guide 2.c, (http://www.syniverse.com/pdfs/IntlRoamGuide.pdf), January 2002.

[4] ETSI: Digital cellular telecommunications system (Phase 2+); Event and call data. GSM 12.05 version 7.0.1, Release 1998.

[5] GSM Association: Transferred Account Procedure Data Record Format, Specification Version Number 3, 26 May 2005.

[6] GSM Association: Press Release 'PacNet Operates Asia's First Neutral Peering Point for GPRS Roaming: Peering Singapore', http://www.gsmworld.com/news/press_2004/press04_17.shtml, February 2004.

[7] Telecommunications Policy Research Conference, 23-25 September, 2000, Alexandria, VA, USA.

[8] Intec Telecom Systems PLC: Interconnection - an introduction (http://www.intec-telecom-systems.com/its/siteservices/downloads/?archive=/its/pressroom/whitepapers/.

[9] Definition of Roaming, http://www.webopedia.com/TERM/r/roaming.html.

[10] Article Alley: Cordless Phone Systems, http://www.articlealley.com/article_8556_45.html.

[11] Africa and middle east Telecom-week: Mozambique, issue 11, 10 January 2002 http://www.telecom-daily.com/samplereport.pdf.

[12] Proximus: Tarifs internationaux, http://customer.proximus.be/download/Roa_Voice_FR.pdf.

[13] TIM Hellas Telecommunications S.A, http://www.tim.com.gr.

[14] Comfone: Key2roam, 24 August 2004, http://www.comfone.com/_main_pages/news/newsletter/04_aug_23/testbox.htm.

[15] Weissbuch Mobilkommunikation, 5. Technische Grundlagen der Mobilkommunikation: Streunen in fremden Netzen, Fast „grenzenlose“Kommunikation, 29.Juni 2001.

[16] Situation of the Swiss Telecommunication market in an international comparison, A study carried out on behalf of the Federal Office of Communication, 29 April 2002.

[17] Klaus, R.: Kommunikationstechnik Teil 1, Zürcher Hochschule Winterthur, 2002.

[18] Hewlett-Packard: HP Fraud Management System, Verion 8.1, 2003.

[19] Hewlett-Packard: HP Helps Mobile Operator Optimus Combat Wireless Fraud - and Minimize Revenue Losses, April 2005, http://www.hp.com/hpinfo/newsroom/press/2005/050426a.html.

[20] Chorleywoods Publications: GSM Cloning Fraud - Still a Wireless Threat, http://www.3g.co.uk/PR/August2002/3876.htm, August 2002.

[21] International Monetary Found: Factsheet; Special Drawing Rights (SDRs), http://www.imf.org/external/np/exr/facts/sdr.HTM, September 2005.

[22] Cibernet: Ciber Records, http://www.cibernet.com/pdfs/ciber.pdf.

[23] Comfone: Comfone GRX Service Overview, www.comfone.ch/_main_pages/news/_oldfiles_comfone/details/grx_so.pdf.

[24] Alon Barnea of Starhone: Roaming's SOR point, 8 December 2005 www.totaltele.com/View.aspx?ID=77587&t=4.

[25] Deutsche Telecom Completes Acquisitions of VoiceStream: Wireless and Powertel www.t-mobile.com/company/pressroom/pressrelease17.asp.

[26] European Commission: Working Document on the initial findings of the sector inquiry into mobile roaming charges, Brussels, 13 December 2000.

[27] Informa telecoms & media: Falling roaming prices benefit costumers and operators, says Thomas Wehrmeier, 9 August 2005.

[28] MEMO/05/44: Mobile telephones/international roaming-frequently asked questions: What is international roaming?, 10 February 2005.

[29] INTUG: Speech prepared for a joint meeting of the Committee on Industry, Research and Energy (ITRE) and the Committee on Internal Market and Consumer Protection (IMCO) of the European Parliament on March 2005.

[30] Wik consult: International Roaming: A way forward, 14-15 October 2004.

[31] Swisscom: Miscellaneous brochures der Swisscom, 2005.

[32] Ewan Sutherland, Executive Director, International Telecommunications Users Group: Speech prepared for a joint meeting of the Committee on Industry,Research and Energy (ITRE) and the Committee on Internal Market and Consumer Protection (IMCO) of the European Parliament on 16 March 2005.

Kapitel 8

Software Patents: Innovation Killer or Innovation Supporter

Domenic Benz, Sascha Nedkoff, Jonas Tappolet

Diese Arbeit soll einen Überblick über verschiedene Aspekte von Softwarepatenten geben. Dazu werden zuerst die Unterschiede zwischen Urheberrecht und Softwarepatenten betrachtet. Anschliessend wird die aktuelle Rechtslage in verschiedenen Ländern, darunter auch die Schweiz betrachtet. Im Weiteren wird auf die verschiedenen Interessensgruppen im Zusammenhang mit Softwarepatenten und ihre Argumente eingegangen sowie die möglichen Auswirkungen von Softwarepatenten auf die Wirtschaft, die Forschungstätigkeit, die Wissenschaft sowie die Open Source Bewegung diskutiert.

Inhaltsverzeichnis

8.1 Einleitung

Softwarepatente sind - nicht nur in der IT-Branche - ein ebenso aktuelles wie auch umstrittenes Thema. Nicht selten werden die bisweilen sehr hitzigen Diskussionen jedoch mit emotionalen Äusserungen über ideologische Grundsätze statt mit sachlichen Argumenten geführt. Dass dies nicht im Sinne der Sache ist und bestimmt nicht zu einer Einigung oder zumindest einer Annäherung der Positionen führt, steht ausser Frage. Dieses Problem ergibt sich auch daraus, respektive wird dadurch bestärkt, dass viele Betroffene zu wenige Kenntnisse des Patentwesens haben, um sich auf dieser Grundlage ein eigenes Urteil zu bilden. Aus diesem Grund sollen in dieser Arbeit zuerst einige Grundlagen des Patentwesens betrachtet werden, um danach aufgrund dieser Kenntnisse fundierte, analytische Aussagen machen zu können.

Befürworter von Softwarepatenten führen gerne den Erfolg des herkömmlichen Patent-Systems als Argument an. Sie sind überzeugt davon, dass Softwarepatente die Investitionen in Forschung und Entwicklung ankurbeln und so zu mehr Innovationen führen würden. Durch den Patentschutz sollen sich grössere Investitionen in die Forschung vor allem auch für kleinere Unternehmen lohnen, da sie dann ein Monopol auf Zeit auf ihre Idee hätten. Es wäre durch das Patent sichergestellt, dass niemand sonst ohne Einwilligung des Patentinhabers von dessen Forschung profitieren könnte. Kritiker jedoch sehen in den Softwarepatenten Wettbewerbsnachteile vor allem für kleinere Unternehmen und selbstständige Programmierer, da diese nicht die Mittel hätten, um die ihnen durch die Patente zustehenden Rechte auch durchzusetzen respektive sich gegen Patentrechtsklagen von grösseren Unternehmen zu verteidigen.

Ziel dieser Arbeit ist es daher auch, die aktuelle Situation zum Thema Softwarepatente zu beleuchten, sowie die Vor- und Nachteile von Softwarepatenten zu diskutieren. Es sollen im Rahmen dieser Arbeit auch die möglichen Auswirkungen von Softwarepatenten auf verschiedene Bereiche erläutert werden.

8.2 Patente vs. Urheberrechte

Wozu Softwarepatente? Reicht der Schutz des geistigen Eigentums im Rahmen des Urheberrechts nicht aus? Zur Beantwortung dieser Frage sollen zuerst einige Grundlagen des Patent-Systems erläutert werden sowie nachfolgend der Zweck und die Unterschiede des Urheberrechts und des Patentrechts betrachtet werden.

8.2.1 Grundlagen Patentrecht

Definition Patent: [5] „Das dem Erfinder oder seinem Rechtsnachfolger vom Staat für sein Gebiet erteilte, zeitlich begrenzte Monopol für die wirtschaftliche Nutzung einer Erfindung.“

Damit die Diskussion der Vor- und Nachteile von Softwarepatenten auf einer soliden, sachlichen Grundlage geführt werden kann, soll an dieser Stelle eine kurze Einführung in das Patentwesen gemacht werden. Dabei soll der Fokus auf dem Begriff der Erfindung als Gegenstand eines Patents sowie auf den Kriterien zur Patentierbarkeit einer solchen liegen. Allgemein gilt es noch zu sagen, dass es sich bei dem Patentrecht um ein gewerbliches Schutzrecht handelt, welches territorial begrenzt ist. Ein erteiltes Patent hat folglich nur in jenem Land oder Wirtschaftsraum Gültigkeit, für welchen das ausstellende Amt zuständig ist.

Geschützer Gegenstand eines Patents - die Erfindung

Der zentrale Punkt im Zusammenhang mit Patenten ist die Erfindung, welche durch das Patent geschützt werden soll. Es gilt folglich, den zentralen Begriff der Erfindung zu definieren und Kriterien zur Patentierbarkeit einer Erfindung zu finden. Da es nicht nur einem Laien schwer fällt, genau zu bestimmen, wann es sich bei einer Idee oder einem Produkt um eine Erfindung handelt, wird im Patentgesetz nicht genau beschrieben, was exakt eine Erfindung ausmacht, sondern es wird geregelt, was keine Erfindung im Sinne des Patentrechts darstellt. Es existieren jedoch einige Kriterien, welchen eine Erfindung genügen muss, damit sie patentiert werden kann. Zu bemerken ist an dieser Stelle noch, dass auch die nachfolgend beschriebenen Kriterien nicht eindeutig sind und selbst wiederum ziemlich grosse Spielräume bei der Anwendung offen lassen. Unter diesem Gesichtspunkt erstaunt es auch nicht weiter, dass in der Praxis teils nicht unerhebliche Diskrepanzen zwischen den verschiedenen Patentämtern - mitunter auch innerhalb des gleichen Patentamts - bei der Erteilung von Patenten existieren.

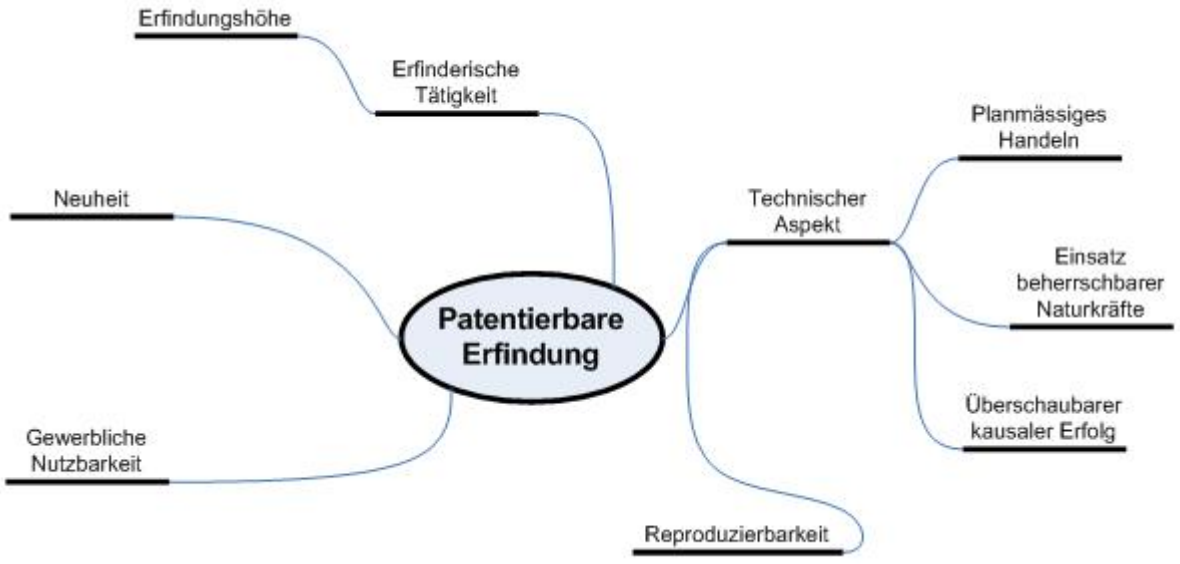


Abbildung 8.1: Aspekte einer patentierbaren Erfindung, Quelle: Eigene Darstellung

Charakteristika einer patentierbaren Erfindung

- **Neuheit**
  Damit eine Erfindung patentiert werden kann muss sie neu sein. Dies bedeutet, dass sie zum Zeitpunkt der Patentanmeldung nicht zum Stand der Technik gehören darf. Der Begriff „Stand der Technik“ bezeichnet dabei alles, was zum Zeitpunkt der Anmeldung in irgendeiner Form, schriftlich sowie mündlich, der Öffentlichkeit zugänglich war. Dazu gilt es zu sagen, dass dies nicht bedeutet, dass der Erfinder, welcher das Patent beantragt von einer allfällig bereits erfolgten Veröffentlichung gewusst haben muss [2].

- **Erfinderische Tätigkeit**
  Es muss im Weiteren eine erfinderische Tätigkeit erkennbar sein, damit ein Patent ausgestellt werden kann. Dies bedeutet, dass die Erfindung, welche patentiert werden soll, eine ausreichende Erfindungshöhe aufweisen muss. Dies ist im Allgemeinen der Fall, wenn ein (fiktiver) Fachmann mit durchschnittlichen Kenntnissen auf dem jeweiligen und einigen angrenzenden Gebieten die Erfindung nicht in nahe liegender Weise aus dem Stand der Technik herleiten kann.

- **Gewerbliche Nutzbarkeit**
  Da das Patentrecht ein gewerbliches Schutzrecht ist, ist es zwingend notwendig, dass eine patentierbare Erfindung auch gewerblich nutzbar ist. Um dieses Kriterium zu erfüllen, reicht es jedoch aus, wenn es denkbar wäre, die Erfindung in irgendeiner Weise wirtschaftlich zu nutzen. Ausgeschlossen werden durch dieses Kriterium meist jedoch Erfindungen, welche (noch) nicht umsetzbar sind oder schlicht nicht funktionieren.

- **Technischer Aspekt** Damit für eine Erfindung ein Patent ausgestellt werden kann, muss sie aus planmäßigem Handeln unter Einsatz von beherrschbaren Naturkräften bestehen und zur Erreichung eines kausal übersehbaren Erfolges dienen [3]. Zusätzlich muss sie reproduzierbar sein. Damit ist im Wesentlichen gemeint, dass eine Erfindung genau kontrolliert ablaufen muss und einen im Voraus beschriebenen Erfolg, welcher zwingend aus den vorherigen Handlungen entsteht, zur Folge haben muss.

Patentkategorien

Das Patentwesen kennt zwei Kategorien von Patenten, in welche ein neues Patent bei der Zuteilung eingeteilt werden kann. Es existieren einerseits Erzeugnispatente, welche die Idee respektive das Konzept hinter einem Produkt unter den Schutz des Patents stellen und andererseits Verfahrenspatente, welche ein Verfahren, welches zu einem definierten Erfolg führt, sowie alle Produkte und Zwischenprodukte, welche bei der Durchführung des Verfahrens entstehen schützen [3].

8.2.2 Urheberrecht

Da Software als literarisches Werk traditionell durch das Urheberrecht geschützt ist, soll an dieser Stelle das Urheberrecht kurz beschrieben werden. Die nachfolgende Betrachtung bezieht sich auf das Schweizerische Urheberrecht [4]

Definition Urheberrecht [5] „Das eigentumsähnliche Recht des Schöpfers eines Werks der Literatur, Musik, Kunst, Fotografie oder von Computerprogrammen an seinem Werk (geistiges Eigentum).“

Geschützter Gegenstand im Sinne des Urheberrechts ist immer das Ergebnis, welches aus schöpferischer Tätigkeit eines Menschen entsteht. Diese Schöpfung, im Kontext des Urheberrechts Werk genannt, muss dabei einen individuellen Charakter besitzen. Das Urheberrecht kommt mit dem „erstmaligen Festhalten auf einem Medium“ zustande und muss folglich nicht speziell beantragt werden. In der Praxis ist somit eine Software vom Zeitpunkt der Implementation an automatisch geschützt. Das Urheberrecht kennt zweierlei Ziele. Es dient einerseits dem Persönlichkeitsschutz des Urhebers, was bedeutet, dass niemand ausser dem Urheber die Erstellung des Werks und den damit eventuell verbundenen Ruhm für sich in Anspruch nehmen darf. Andererseits dient das Urheberrecht ähnlich dem Patentrecht dem wirtschaftlichen Schutz. Dem Urheber allein steht es zu, sein Werk wirtschaftlich auszunutzen und Benutzungsrechte an seinem Werk zu vergeben.

8.2.3 Vergleich

Nach dieser kurzen Einführung in das Patent- respektive Urheberrecht sollen nun nochmals die wichtigsten Unterschiede und Gemeinsamkeiten der beiden Systeme aufgeführt werden.

Die wichtigsten Rechte, welche der Urheber nach [4], respektive der Patentinhaber besitzen sind:

- Ausschliesslichkeitsrecht:
  Der Inhaber des Rechtes kann andere von der Verwendung seines Werks oder seiner Erfindung ausschliessen.

- Auskunftsrecht:
  Der Inhaber des Rechts hat im Falle der Verletzung seines Rechts das Recht zu erfahren, woher die verletzenden Materialien stammen.

- Vernichtungsanspruch:
  Der Inhaber des Rechts hat Anspruch auf die Vernichtung von Materialien, welche seine Rechte verletzen, sofern dies für den Verletzenden zumutbar ist.

Kriterium	Urheberrecht	Patent
Zweck & Ziele	- Persönlichkeitsschutz - Wirtschaftlicher Schutz	- Wirtschaftlicher Schutz - Offenlegung der Idee - Innovationsförderung durch Umgehungsinnovationen
Schutzgegenstand	- Konkrete Implementierung/ Produkt des Urhebers	- Erfindung (Idee/Konzept) des Patentinhabers
Voraussetzungen	- geistige Schöpfung - individueller Charakter	- neue Erfindung - Erfindungshöhe - Gewerbliche Nutzbarkeit - Technischer Aspekt
Zustandekommen	- Automatisch durch „erstmaliges Festhalten auf einem Medium“	- Auf Antrag bei zuständigem Patentamt
Kosten	- Kostenlos	- Kostenpflichtig - Jährliche Kosten
Schutzdauer	- Endet 50 Jahre nach Tod des Urhebers (für Software, übrige Werke: 70 Jahre)	- 20 Jahre

Abbildung 8.2: Vergleich Urheberrecht - Patentrecht

8.2.4 Konfliktpotential Urheberrecht - Softwarepatente

Das grösste Problem, welches durch die gleichzeitige Existenz von Urheberrecht und Softwarepatenten entsteht ist die fehlende rechtliche Sicherheit des Programmierers in Bezug auf die Rechte an einer von ihm eigenständig entwickelten Software. Während er sich in einer Situation ohne Softwarepatente sicher sein kann, dass alle Rechte an der von ihm entwickelten Software ihm zustehen, sofern er keinen fremden Code kopiert. Wenn nun jedoch zusätzlich Softwarepatente existieren, hat er zwar das Urheberrecht an der entwickelten Software, verletzt jedoch unter Umständen gleichzeitig und ohne zwingend darüber in Kenntnis zu sein fremde Patente. Die Konsequenz daraus wäre, dass er als Urheber nicht mehr frei über seine geistige Schöpfung verfügen kann.

8.3 Rechtliche Situation

Nachfolgend sollen einige rechtliche Aspekte betreffend Softwarepatente beleuchtet werden. Das Patentrecht ist ein Grundpfeiler einer funktionierenden Wirtschaft und findet daher schon seit dem 19. Jahrhundert Eingang in die Gesetzbücher vieler Staaten. Wird ein Patent angemeldet, so ist es das Ziel des Antragstellers dass seine Erfindung möglichst überall Geschützt ist. Dazu ist eine gewisse Abstimmung zwischen den verschiedenen Staaten von Nöten. Hierzu wurde schon 1883 die Pariser Verbandsübereinkunft (PVÜ, vgl. [12] abgeschlossen, welche das grundlegende Verfahren zur Patenterteilung international regelt. Ein weiteres internationales Abkommen ist TRIPS (Agreement on Trade-Related

Aspects of Intellectual Property Rights, vgl. [10] welches als Bestandteil des Welthandelsabkommens (WTO) für dessen Mitgliedsstaaten (darunter auch die Schweiz) gilt. Darin wird unter Anderem von den Mitgliedsländern gefordert, ihre Patentrechte untereinander zu harmonisieren.

Art. 27 TRIPS: Subject to the provisions of paragraphs 2 and 3, patents shall be available for any inventions, whether products or processes, in all fields of technology, provided that they are new, involve an inventive step and are capable of industrial application. 5 Subject to paragraph 4 of Article 65, paragraph 8 of Article 70 and paragraph 3 of this Article, patents shall be available and patent rights enjoyable without discrimination as to the place of invention, the field of technology and whether products are imported or locally produced.

Nach Art. 27 schliesst also TRIPS die Patentierbarkeit von Computerprogrammen nicht aus.

8.3.1 EU

Die Zusammenarbeit im Rahmen der EU weitete sich ab 1973 auch auf die Erteilung von Patenten aus. Damals einigten sich die EG-Staaten auf die Europäische Patentübereinkommen (EPÜ vgl. [9]. Dieses ist die rechtliche Grundlage für das Europäische Patentamt (EPA) welches dadurch Patente vergeben konnte, welche in den Mitgliedsstaaten dieselbe Gültigkeit haben, wie national vergebene Patente. Das EPÜ kennt keine Softwarepatente, schliesst diese sogar explizit aus:

Art. 52 EPÜ: Europäische Patente werden für Erfindungen erteilt, die neu sind, auf einer erfinderischen Tätigkeit beruhen und gewerblich anwendbar sind.

Als Erfindungen im Sinn des Absatzes 1 werden insbesondere nicht angesehen:

(...)

c) Pläne, Regeln und Verfahren für gedankliche Tätigkeiten, für Spiele oder für geschäftliche Tätigkeiten sowie **Programme für Datenverarbeitungsanlagen;**

Der Gesetzgeber hat also klar verankert, dass Patente auf Software (Programme für Datenverarbeitungsanlagen) nicht möglich sind.

Mit der wachsenden Bedeutung der Informationstechnologie während der 70er und 80er Jahre des 20. Jahrhunderts hat sich auch der Blickwinkel auf Softwarepatente verschoben. Die Erfindungen und neuen Entwicklungen die zu dieser Zeit gemacht wurden, hatten oft als integralen Bestandteil Computersysteme, und damit Software für deren Steuerung und Kontrolle. Immer mehr wurde klar, dass die wirkliche Neuerung der Erfindung die Software oder das Computersystem war. Am Beispiel von ABS (Antiblockier-System, Antilock Brake System) soll dies verdeutlicht werden:

Das Prinzip von ABS ist seit den 1930er Jahren bekannt. Es sollten, bei einer starken Bremsung eines Fahrzeugs, die Räder mittels Reduzierung der Bremskraft daran gehindert werden zu blockieren. Um dies zu realisieren wurde mittels aufwendiger Steuermechanik versucht das gewünschte Verhalten zu erzeugen. Später wurde die Steuerung mittels analoger Schaltungen optimiert. Wirklich funktionsfähig wurde ABS erst, als die Steuerung mittels Digitaltechnik implementiert wurde. Ein Mikroprozessor erhält Sensordaten ob und wie schnell die Räder drehen. Wenn während einer Bremsung die Räder blockieren, gibt die Software den Befehl an ein elektrisches Steuerventil zur Reduzierung des Bremsdruckes. Diese neue Möglichkeit durch Digitaltechnik machte das Produkt ABS erst Serienreif und verhalf dem System zum weltweiten Durchbruch. Die Innovation liegt hierbei nicht bei den Steuerventilen oder Sensoren. Es ist grösstenteils die Software der Steuerung die das gesamte Verhalten des Bremssystems ausmacht, welche aber nach Art. 52 EPÜ nicht patentierbar ist.

Das Europäische Patentamt musste daher seine Vergabepraxis für Patente überdenken. Es gab berechtigte Patentansprüche, die sich aber auf Software bezogen und daher nach Gesetz keinen Anspruch auf Patentschutz haben. Das Europäische Patentamt begann, den Art. 52 EPÜ neu zu interpretieren. Das EPA unterschied ab Mitte der 1980er Jahre zwischen „Computerprogrammen als solche“ und „Computerprogrammen als Teil einer technischen Erfindung“. Erstere konnten nach wie vor nicht Patentiert werden, bei Letzteren wurden neu Patente vergeben um auch Erfindungen wie ABS (s. O.) zu schützen. Grundsätzlich erteilte das Europäische Patentamt nun Patente für Computerprogramme, sofern diese an der Steuerung für ein technisches Verfahren beteiligt sind. Es wurde als das Patent als ganzes betrachtet, und sofern dieses ein technisches Verfahren war, konnte ein Patent dafür erteilt werden, egal ob letztendlich die einzige Neuerung in Form von Software bestand. Später ging das EPA dazu über, noch mehr Restriktionen im Bezug auf die Patentierbarkeit von Software aufzuheben. So wurde nun der Begriff des „technischen Beitrags“ eingeführt. Neu konnten Patente für Software angemeldet werden, sofern diese einen technischen Beitrag liefert. Dies entpuppte sich als sehr dehnbarer Begriff unter welchem praktisch jedes Softwarepatent angemeldet werden konnte.

Beispielsweise wurde ein Patent für die Reiter- oder Karteikartendarstellung bei Software vergeben. Dabei werden klassische Register in Ordnern als Metapher virtuell zur Bedienung von Software abgebildet. Der technische Beitrag unter dem das Patent auf dieses Konzept vergeben wurde, wurde beschrieben als: Die Reiterdarstellung dient der Einsparung der knappen (physischen) Ressource Bildpunkte die auf einem Computermonitor zur Verfügung stehen.

Da es die Kategorie des Software-Patents nicht gibt, kann die Anzahl der bisher vergebenen Patente auf Software nur geschätzt werden. Der FFII (Förderverein für eine Freie Informationelle Infrastruktur) hat einen Kriterienkatalog für Softwarepatente erarbeitet, und systematische die bisher vergebenen Patente des Europäischen Patentamts damit überprüft. Laut Zählung des FFII sind bin Anhin etwa 30.000 Softwarepatente in Europa vergeben worden. Je nach dem wie man den Kriterienkatalog festlegt, schwankt die Zahl der Patente, und die FFII als vehementer Gegner von Softwarepatenten hat vermutlich die Kriterien so festgelegt, dass eine möglichst grosse Zahl von Softwarepatenten resultiert. Gemäss FFII sind viele dieser Softwarepatente so genannte Trivialpatente. Also Patente mit einer geringen Erfindungshöhe die einer gerichtlichen Überprüfung kaum standhalten

würden. Der einzige Nutzen der ihr Inhaber hat ist, dass er sich Erfinder der patentierten Sache nennen darf. Ein direkter wirtschaftlichen Vorteil kann daraus nicht gezogen werden. Ein weiterer interessanter Punkt bei den bestehenden Softwarepatenten ist die Herkunft der Erfinder (Firmen). Dazu hat wiederum der ffii eine Statistik veröffentlicht:

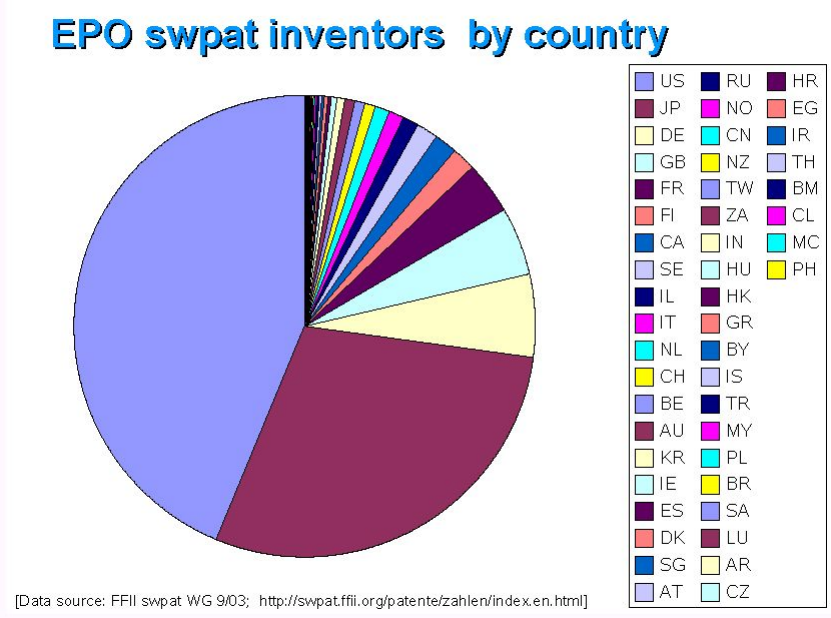


Abbildung 8.3: Softwarepatente nach Anmeldeländern

Daraus geht hervor, dass fast 75% der bestehenden Patente auf Software von Personen und Firmen aus den USA und Japan angemeldet wurden. Beide Staaten kennen Softwarepatente in ihrem Rechtssystem. Dieser grosse nichteuropäische Anteil hat auch in der Debatte in der EU das Argument der Softwarepatent-Gegner bestärkt, dass die wahren Profiteure einer Zulassung von Softwarepatenten nicht europäische, sondern amerikanische und japanische Firmen wären.

Die Debatte in der EU

Aus oben genannten Gründen, aber auch durch andere, neue Bereiche wie Gen- und Biotechnologie hat sich die europäische Patentorganisation 1998 dazu entschieden, eine umfassende Reform des Europäischen Patentwesens vorzunehmen.

Dazu wurde unter Anderen auch das Europäische Patentamt beauftragt, einen Richtlinienvorschlag für die „Patentierbarkeit computerimplementierter Erfindungen“ zu erarbeiten. Dieser sprach sich weitgehend für Softwarepatente aus und wurde im Jahr 2002 von der Europäischen Kommission der Öffentlichkeit vorgestellt [7], [6]. In diesem Vorschlag wurde der Art. 52 EPÜ dahingehende geändert, dass die Einschränkung von Abs 2: „nicht patentierbar sind ‚Programme für Datenverarbeitungsanlagen‘“ ganz gestrichen wurde. Der Art. 52 wurde noch in der Form erweitert, dass er nun Grundsätzlich Patente für alle Bereiche der Technik zuliess. Dies ist in etwa derselbe Wortlaut wie im TRIPS Art. 27. Die Idee war auch, TRIPS und EPÜ auf dieselbe Linie zu bringen. Die Frage der Softwarepatente

rückte nun in das öffentliche Interesse und es formierten sich Befürworter- sowie Gegnergruppen. Die Befürworter, meist Grosskonzerne, versuchten die Debatten nicht durch Öffentlichkeitsarbeit zu beeinflussen, sondern betrieben Lobbying zur Beeinflussung der Entscheidungsträger. Ein Indiz dafür, dass die Befürworter ihre Anliegen hinter verschlossenen Türen vorbrachten, ist die Internetseite Patents4Innovation.org welche genau einen Monat online war und danach wieder verschwand. Der Inhalt kann in einem Webarchiv betrachtet werden (z.B. www.archive.org/web).

Im Gegensatz dazu haben die Gegner versucht, die Debatte in die Öffentlichkeit zu tragen um sich Gehör zu verschaffen. Als vehementeste Gegnergruppierung ist hier sicherlich der „Förderverein für eine freie informationelle Infrastruktur“ (kurz: ffii, swpat.ffii.org) zu nennen. Als Vertreter von Privatpersonen und Kleinfirmen hat der Verein nicht die Finanziellen Mittel der Konsortien von Grossfirmen. Sie mussten sich daher auf die Information im Internet, Demonstrationen und verschiedene Aktionen die an Guerilla-Marketing erinnern, beschränken. Ein Beispiel dafür ist Präsenz mit Booten und Transparenten auf einem Fluss vor den Büros von EU-Abgeordneten.

Das Europäische Parlament befasste sich nun mit dem Richtlinienvorschlag und debattierte diesen im September 2004. Das Parlament legte seinen Standpunkt fest, dass die Technizität einer Erfindung gegeben sein muss und diese eine Wirkung auf die Naturkräfte haben muss. Zu einer ähnlichen Ansicht kam 1969 auch schon der deutsche Bundesgerichtshof im „Rote Taube Entscheid“ [1]).

Im Mai 2004 wurde ein gemeinsamer Standpunkt entschieden welcher Softwarepatente weitgehend ermöglichte, womit Gerichte auch bestehende Patente anerkennen müssten. Für die 2. Lesung wurden insgesamt 256 Änderungsanträge eingereicht. Sogar die Befürworter von Softwarepatenten sprachen sich nun gegen diesen Vorschlag aus, da der Richtlinienvorschlag durch die verschiedenen Debatten und Änderungsanträge so stark verstümmelt wurde, dass diese ihre Interessen darin nicht mehr vertreten sahen. Dieser stark veränderte und revidierte Vorschlag kam am 6. Juli 2005 vor das EU-Parlament zur Abstimmung: Mit 95% Mehrheit sprachen sich die Abgeordneten gegen den Vorschlag aus, und beendeten (vorläufig) die Debatte um Softwarepatente in der EU.

Aktuell gilt immer noch der Status Quo. Die rechtliche Situation hat sich nicht verändert, und das Europäische Patentamt vergibt weiter nach ihren unveränderten Richtlinien Patente auf Software. Diese haben aber durch das Scheitern des Richtlinienvorschlags vor dem Parlament einen schweren Stand vor Gericht, da die rechtliche Grundlage für ein Softwarepatent fehlt. Inhaber von Softwarepatenten werden also genau abwägen müssen, ob sie eine Patenverletzung vor Gericht einklagen wollen, da die Erfolgschancen aktuell noch eher gering sind.

8.3.2 Schweiz

In der Schweiz ist Software per Gesetz durch das Urheberrecht geschützt. Die aktuelle Vergabepraxis des IGE (Eidgenössisches Institut für Geistiges Eigentum) lässt keine Softwarepatente zu. Durch die Mitgliedschaft der Schweiz in der Europäischen Patent

Gemeinschaft (EPG) gelten in der Schweiz Patente die nach dem Recht der Europäischen Patentübereinkunft (EPÜ) vergeben wurden wie nationale Patente. Die Schweiz hat ein Mitspracherecht in den EPG und kann aktiv an der Gestaltung eventueller Änderungen der rechtlichen Übereinkunft teilnehmen [8]. Sie ist aber an die Entscheidung der EU Staaten gebunden, da diese gemeinsame Entscheidungen in der EPG treffen und daher die Stimme der Schweiz kein Gewicht hat. Wenn die EPG sich für Softwarepatente aussprechen würde, wäre die einzige Möglichkeit der Schweiz, aus der Europäischen Patentorganisation auszutreten, mit der Konsequenz, dass Europäische Patente in der Schweiz keine Gültigkeit mehr hätten. Der Preis eines Alleingangs wäre demzufolge zu hoch und die Schweiz wird sich den Entscheidungen der EU-Staaten beugen müssen.

8.3.3 USA

Bereits im Jahr 1980 hat der Oberste Gerichtshof der USA die Patentierung von Software mit dem Entscheid „Diamond vs. Diehr“ ermöglicht. Ähnlich wie in der EU wurde die Notwendigkeit erkannt, dass gewisse Erfindungen die durch die neue Möglichkeit von Mikroprozessoren und Software erst möglich wurden, auch einen ausreichenden Patentschutz erfordern. Dazu machte der Gerichtshof die Einschränkung, dass Computerprogramme nur patentiert werden können, wenn diese einen engen Bezug zu industriellen Prozessen haben (Beispiel: Software die die Abläufe bei einer Verpackungsmaschine steuert). Später wurde diese Einschränkung nicht mehr wörtlich ausgelegt, und 1999 wurde durch das Bundesberufungsgericht im Entscheid „State Street Bank“ [11] die Patentierbarkeit von Software ganz ermöglicht. Neu konnten auch Geschäftsprozesse zum Patent angemeldet werden. In der folgenden Grafik wird der Prozentuale Anteil von Softwarepatenten an allen Patenten über die Jahre aufgezeigt.

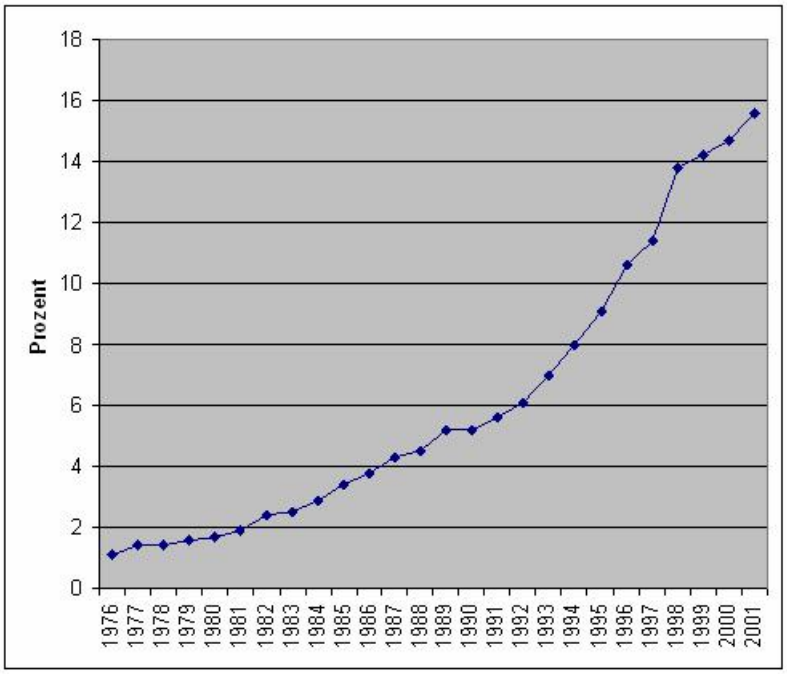


Abbildung 8.4: Entwicklung der Softwarepatente in den USA

Erfahrungen mit Softwarepatenten

Da in den USA der Gerichtsprozess ein Bestandteil des täglichen Geschäfts ist, zeigen sich nun auch vermehrt Patentstreitigkeiten (überwiegend zwischen Grossfirmen) in Millionenhöhe. Es haben sich auch einzelne Patentkanzleien auf Softwarepatente spezialisiert und kaufen diese vorwiegend aus Nachlass insolventer Softwarefirmen. Diese Kanzleien nutzen das Patent selber nicht, sie treten höchsten ein Nutzungsrecht unter Zahlung von Lizenzgebühren ab. Eine weitere Einnahmequelle ist die Geltendmachung von Ansprüchen vor Gericht, sofern jemand das Patent verletzt hat.

8.3.4 Indien

Der Art. 27 der TRIPS Verträge, wonach die Patente für alle Bereiche der Technik vergeben werden sollen, kann im Konflikt mit bestehenden Verboten zu Softwarepatenten stehen. Es ist Auslegungssache: Entweder ist man der Ansicht, dass Software kein Bereich der Technik ist, und daher auch nicht patentierbar sein muss. Ein weiterer Schluss den man daraus ziehen könnte ist, dass Software patentierbar sein muss, sofern sie für einen technischen Zweck (Technizität) eingesetzt wird. Wenn Software als Bereich der Technik angesehen wird, ist ein Verbot von Softwarepatenten eine Verletzung von Art. 27 der TRIPS Verträge. Zu ebendiesem Schluss kam Indien Ende 2004. Angeheizt durch die Debatte in der EU wurde durch eine Notverordnung der ordentliche Gesetzgebungsprozess und damit das Parlament umgangen und Softwarepatente eingeführt. Später wurde das ganze vom Parlament wieder Rückgängig gemacht.

8.4 Pro & Contra Softwarepatente

8.4.1 Vorteile und Chancen

Offenlegung von Erfindungen

Ohne Patente gibt es ein Spannungsverhältnis zwischen wirtschaftlicher Verwertung und wissenschaftlicher Veröffentlichung einer Erfindung. Denn in der Geheimhaltung der Funktionsweise einer Erfindung besteht die einzige Möglichkeit, seine Exklusivrechte zu schützen. Die Auflösung dieses Interessenskonfliktes soll im Interesse der Gesellschaft dadurch erfolgen, einem Erfinder im Gegenzug zur Offenlegung seiner Erfindung einen zeitlich befristeten Rechtsschutz zu gewähren. Patente sollen somit den Stand der Forschung und Technik dokumentieren und vorantreiben [14].

- **Sicherung der Erfindereinkünfte aus Gerechtigkeitserwägungen**
  Der wesentliche Grund, warum der Begriff des Patents in der breiten Bevölkerung positiv belegt ist, liegt im allgemeinen Konsens, dass Erfindern eine angemessene

Entlohnung für ihre geistige Leistung zusteht. So ermöglicht der Schutz eines Patentes die Kosten für die Innovationstätigkeit zu decken und zusätzlich Gewinne zu erwirtschaften [14].

- **Sicherung der Erfindereinküfnte als Investitionsanreiz**
  Zum moralischen Aspekt des Schutzes von erfinderischer Leistung tritt das wirtschaftspolitische Ziel, Investitionen in Innovationen zu fördern. Das Patentwesen erfüllt den Zweck der Innovationsförderung, wenn dank seiner Existenz eine höhere Investitionstätigkeit in Forschung und Entwicklung stattfindet. Eine Rechtfertigung kann unabhängig vom Gesamtvolumen schon darin bestehen, dass bestimmte einzelne Forschungs- und Entwicklungsvorhaben finanziert werden, die für die Gesellschaft oder Wirtschaft von hohem Interesse sind (z. B. Medikamente) [14].

8.4.2 Nachteile und Gefahren

- **Transaktions- und Opportunitätskosten der Patentanmeldung**
  Die patentamtlichen Gebühren für eine Patentanmeldung haben an den gesamten Kosten, die mit dem Erwerb eines Patentes verbunden sind, im Normalfall einen sehr kleinen Anteil. Wesentlich höher ist der Aufwand für die Formulierung der Anmeldung, wofür sowohl Arbeitszeit auf Erfinder- als auch auf Patentanwaltsseite benötigt wird. Dieselben Mittel könnten ansonsten beispielsweise in die Forschung und Entwicklung selbst fliessen. Am potenziell kostspieligsten sind Patentverletzungsklagen, bei denen man sich mit Ansprüchen anderer Parteien auseinander zusetzen hat [14].

- **Zeitliche Verzögerung und Effizienzverluste**
  Soweit zur Vermarktung einer Erfindung eine Patentgewährung abgewartet werden muss, entstehen durch oftmals mehrjährige Entscheidungszeiträume beträchtliche Verzögerungen. Zu Patentanmeldungen sind fast nur Unternehmen fähig, denn eigenständige Entwickler in Einzelunternehmen verfügen abgesehen von den Ressourcen und Wissen auch nicht über die entsprechenden Prozesse. Oftmals werden nicht einmal die Folgen etwaiger Veröffentlichungen auf den Patentanspruch bedacht. Denn durch eine frühzeitige Veröffentlichung der Erfindung kann der Patentanspruch nicht mehr geltend gemacht werden, da die Erfindung nicht mehr neu ist [14].

- **Willkürliche Einschränkung der Wettbewerbsdynamik während der Patentgültigkeitsdauer**
  Ein Patent stellt ein exklusives Recht dar, über das vergleichbar frei wie über anderes Eigentum verfügt werden kann. Dies beinhaltet ein hohes Mass an Willkür. Ein Erfinder kann für die Nutzung seines Patentes beliebige Gegenleistungen erwarten oder diese ansonsten untersagen. Es steht ihm frei, niemand anders eine Lizenz einzuräumen. So entsteht ein Potential für Missbrauch [14].

- **Errichtung von finanziellen Eintrittsbarrieren**
  Die Kosten eigener Patentanmeldungen sowie auch alle Kosten, die aus der Anmeldung von Ansprüchen durch andere Parteien resultieren können, bedeuten für die Entwickler von potenziell patentierbaren Technologien eine zusätzliche Anforderung

an ihre Kapitalausstattung. Die Zahl der Marktteilnehmer wird damit unabhängig vom Geltungsbereich des Exklusivrechtes einzelner Patente ganz allgemein begrenzt. Das Prozesskostenrisiko wirkt sich für einen Marktteilnehmer dann am stärksten aus, wenn aufgrund eines Kräfteungleichgewichtes ein grösseres Unternehmen eine unüberschaubar grosse Zahl von Patenten besitzt und mehrfache Ansprüche gleichzeitig geltend machen kann [14].

- **Be- oder Verhinderung von inkrementeller Innovation während der Patentgültigkeitsdauer**
  Wenn Innovation auf vorherige Leistungen ähnlich einem Baukastensystem aufbauen, dann schränkt die Existenz von Patenten auf die zu Grunde liegenden Bausteine jede inkrementelle Innovation wirtschaftlich ein. Scheitert der Versuch auch nur für eines der benötigten Patente das Nutzungsrecht zu erwerben, hat dies die Nichtvermarktbarkeit der inkrementellen Erfindung zur Folge. Dieses Risiko hält von Innovation ab. Das Schlagwort „Trivialpatent“ bezeichnet den Extremfall, in welchem ein Patent entgegen der üblichen gesetzlichen Formulierungen auf eine einfache, in zahlreichen Erfindungen genutzte Technik gewährt wird. Eine Be- oder Verhinderung inkrementeller Innovation setzt jedoch nicht Trivialpatente in Reinkultur voraus, sondern beginnt bereits, wenn eine gemeinhin benötigte Technik mittlerer Erfindungshöhe patentiert ist [14].

- **Umgekehrt proportionales Verhältnis von Erfindungshöhe und Wert eines Patentes**
  In der Theorie sollte ein Patent, hinter welchem eine besonders grosse erfinderische Leistung steht, besonders hohen kommerziellen Wert haben. In der Praxis verhält es sich jedoch leider genau umgekehrt. Am besten vermarkten oder als strategisches Blockadeinstrument einsetzen lässt sich ein Patent, wenn es gerade möglichst trivial ist. Denn dann benötigen es möglichst viele potenzielle Lizenznehmer, Käufer oder Konkurrenten. Inhaltliche Trivialität kann durch geschickt aufgeblähte Darstellung manchmal verschleiert werden [14].

- **Risiken aus nicht schuldhafter Handlung**
  Die Verletzung eines Patents kann verheerende wirtschaftliche Konsequenzen nach sich ziehen, ohne dass ein schuldhaftes Handeln oder Unterlassen vorliegt. Damit ein Patentsystem funktionieren kann, ist somit Voraussetzung, dass eventuelle unabsichtliche Verletzungen mit angemessenem Rechercheaufwand erkannt bzw. vermieden werden können. Unabhängig davon besteht stets das Risiko, gegen Patente zu verstossen, die bereits angemeldet, aber noch nicht veröffentlicht sind, diese sind unabhängig vom getriebenen Aufwand nicht recherchierbar [14].

8.4.3 Interessensgruppen

Aus der unterschiedlichen Wahrnehmung und Gewichtung der oben genannten Vor- und Nachteilen, sowie unterschiedlichen Interessen verschiedener Gruppen, ergeben sich sehr unterschiedlichen Positionen bezüglich Softwarepatente. So gehen die Meinungen bei der Diskussion um Softwarepatente sehr stark auseinander. Dabei werden zum Teil sachliche

Fakten durch emotionale Äusserungen ersetzt, deren argumentativer Inhalt weder der Sache dienlich ist, noch von einem Verständnis über das Patentwesen zeugt. Dabei ist es von entscheidender Bedeutung, die Interessenslagen der Kleinen und Grossen, die zum Teil auch durch lautstarke Gruppen vertreten werden, voneinander und von den Dingen abzugrenzen, die langfristig der Gesellschaft, der Wirtschaft und der Informatik als Wissenschaftsdisziplin von nutzen sind. Im letzten Positionspapier der Gesellschaft für Informatik [16] vom Juli 2005 wurden drei unterschiedliche gesellschaftliche Interessensgruppen angegeben, die an dieser Stelle in Bezug auf Softwarepatente dargestellt werden sollen. Dies betrifft die Gebiete der Wirtschaft, Wissenschaft und Open Source Community. Alle diese Gruppen sind dabei mehr oder weniger stark von Softwarepatenten betroffen, so dass sich hinsichtlich ihres Tätigkeitsfeldes unterschiedliche Interessen beobachten lassen [21].

Wirtschaft

Die Wirtschaft ist sicher der gesellschaftliche Bereich, der am stärksten von Softwarepatenten betroffen ist. Aber auch in diesem Bereich driften die Meinungen bei der Frage nach Softwarepatenten auseinander, weil verschiedene Interessenslagen und Grössenordnungen unterschiedliche Meinungen zum Problem von Softwarepatenten induzieren. So müssen diese verschiedene wirtschaftliche Grössenordnungen unterschieden werden: Grossunternehmen, KMUs und Freiberufler Es ist aber anzumerken, dass sich Interessenslagen nicht immer eindeutig diesen verschiedenen Gruppen zuordnen lassen. Es gibt Unternehmen jeder Grösse, die sich für Softwarepatente stark machen, ebenso wie solche, die sie ablehnen. Die Haltung zur Patentierungsfrage lässt sich also nicht allein an der Zahl der Mitarbeiter oder dem Umsatz eines Unternehmens festmachen. Gleiches gilt für Freiberufler, auch hier finden sich sowohl Gegner als auch Befürworter von Softwarepatenten [16], [21].

Grossunternehmen

Die Produkte der Fertigungsindustrie basieren in immer grösserem Ausmass auf Software. Dabei fallen nach Schätzungen zu Folge bis zu 60% des Aufwands für Forschung und Entwicklung von technischen Produkten auf Software. Aus Sicht der international agierenden Unternehmen ist der Patentschutz für Eigenentwicklungen unverzichtbar. Denn zum einen müssen die Innovationen gegenüber der Konkurrenz abgesichert werden, und zum anderen müssen auch weiterhin Anreize bestehen, in neue Entwicklungen und Technologien zu investieren [16]. Sollten in Europa IT-Produkte vom Patentschutz ausgeschlossen sein, so befürchten europäische Unternehmen mit internationalen Marktanteilen gravierende Wettbewerbsnachteile gegenüber Konkurrenten in den USA und Japan, da die dortigen Unternehmen auch weiterhin Patentschutz für ihre Innovationen erlangen können. Bei Patentverletzungsklagen stehen europäische Firmen somit zweifellos in der schlechteren Verhandlungsposition, da ihnen auf Grund des fehlenden Patentportfolios eine aussergerichtliche Einigung in den meisten Fällen verwehrt bleiben dürfte [21]. In Grossunternehmen dienen Patentportfolios in erster Linie der Absicherung gegenüber Konkurrenten gleicher Grössenordnung, um sich bei Verletzungsansprüchen oder Kooperationen in einer guten Verhandlungsposition zu befinden. Denn nicht selten werden unbeabsichtigte Patentverletzungen mit einer Gegenforderung begegnet [21]. Der Schutzanspruch von Softwarepatenten richtet sich dabei weniger gegen Kleinunternehmen oder Konsumen-

ten. Auseinandersetzungen wegen Patentrechten gab es in der Vergangenheit meistens nur mit direkten Wettbewerbern [16]. So ist es auch nicht verwunderlich, dass in den Reihen der Grossunternehmer die Mehrheit der Patentierungsbefürworter zu finden sind, die eine grundsätzliche Patentierbarkeit von Software fordern. Für Grossunternehmen ist ein gut funktionierendes Patentwesen, auch auf dem Gebiet von Softwareinnovationen, sehr wichtig. Die dabei entstehenden Mehrkosten für Patentrecherchen und Patentanmeldungen fallen zumindest aus Sicht der Grossunternehmen nicht so stark ins Gewicht, da auf Grund der Grösse der Projekte und dem technologischem Vorsprung gegenüber der Konkurrenz diese Mehrkosten am Markt schnell wieder eingespielt sein dürften. Fraglich bleibt, ob die Unternehmen die erhöhten Kosten auf den Endverbraucher umlegen [21].

KMUs
Die Meinungen zu Softwarepatenten sehen bei den kleinen und mittleren Unternehmen anders aus. Viele erwarten bei einer Erweiterung der Patentierungsmöglichkeiten von Software erhebliche wirtschaftliche Nachteile, oder halten sie sogar für existenzbedrohend. Selbst der momentane Status quo bei der Patenterteilungspraxis wird oft scharf kritisiert. KMUs befürchten, dass der Verlust von Arbeitsplätzen und ein drastischer Innovationsrückgang bis hin zum Innovationsstopp drohen [13]. So halten die meisten KMUs den Schutz von Softwareerfindungen durch das Urheberrecht als vollkommen ausreichend. Für sie bietet er gegenüber dem Patentrecht sogar wesentliche Vorteile. Erstens ist Software automatisch und gebührenfrei durch das Urheberrecht geschützt, zweitens kann man mit einer selbst entwickelten Software nicht das Urheberrecht einer anderen Partei versehentlich verletzen, so dass die Risiken und Folgen von Patentrechtsklagen entfallen. Laut Studien zählen für sie Patente sogar zu den am wenigsten effizienten Methoden des Investitionsschutzes [13], [16].

Das grösste Problem, welches KMUs in Bezug auf die Erweiterung der Patentierung von Software sehen, ist der Kostenfaktor. Der Schutz eigener Ideen durch Patente ist teuer, ebenso wie der Aufbau eines Patentportfolios, wie es zur Verteidigung gegen Patentansprüche von Mitbewerbern notwendig ist. Um versehentlichen Patentverletzungen aus dem Weg zu gehen, ist darüber hinaus ein nicht zu vernachlässigender Rechercheaufwand notwendig. Werden dabei relevante Patente entdeckt, kommen noch entsprechende Lizenzkosten hinzu. Diese lassen sich unter Umständen vermeiden, wenn mit zusätzlichem Aufwand und Kosten um existierende Patentansprüche herum entwickelt wird. Es verbleibt aber immer ein Restrisiko, Patente zu verletzen und deswegen später verklagt zu werden. Entweder werden entsprechende Patente bei der Recherche übersehen, oder sie waren zum Zeitpunkt der Recherche zwar angemeldet, aber noch nicht veröffentlicht. Was leicht passieren kann, da Anmeldeverfahren relativ lange dauern [16].

Die Kosten und Risiken fallen zwar auch bei grösseren Firmen an, sie fallen dort wegen grösserer Softwareprojekte und einem insgesamt deutlich höheren Budget relativ betrachtet weniger ins Gewicht. Urheberrechte und andere Wettbewerbsregelungen erlauben es bereits kleinen und mittelständischen Software-Firmen, trotz immenser Betriebsmittelnachteile gegenüber grossen Firmen zu bestehen. Aus der flexiblen und wachstumsfreudigen Softwarebranche könnte eine schwerfällige Industrie werden, weil ein Einstieg in diese Branche nur noch über weitläufige Absprachen und Vereinbarungen mit Grossunternehmen möglich sein wird, und weil etliche juristische Hürden genommen werden müssen. Auch wenn es einige kleine und mittelständische Unternehmen in einem derartigen Umfeld

schaffen könnten, werden es sehr viele wahrscheinlich nicht schaffen. So würden allgemein die Aussichten auf eine freie und leicht zugängliche Softwareindustrie innerhalb Europas verschlechtert und die Dominanz gegenwärtig herrschender Marktführer gestützt [13].

Diesen Bedenken könnte allerdings entgegengehalten werden, dass diese Kosten auch von anderen Unternehmen, die nicht in der Softwarebranche tätig sind, seit jeher getragen werden. Der Nachteil, der durch die erhöhten Kosten entsteht, könnte durch den Wettbewerbsvorteil eines eigenen Patents ausgeglichen werden. Denn gerade KMUs sind auf Investoren angewiesen, die Kapital zur Verfügung stellen. Diese Kapitalgeber wollen aber eine Sicherheit für ihre Investitionen, die häufig bei kleineren Unternehmen nicht gegeben ist. So ist es nicht unüblich, dass ein Patent bzw. Patentportfolio als Sicherheit für das investierte Kapital genutzt werden kann. Ein weiterer Vorteil wäre, dass KMUs ihren Patentanspruch gegenüber Grossunternehmen gerichtlich durchsetzen könnten. Dadurch wird es möglich, dass kleinere Unternehmen ihre Innovationen und Marktposition gegenüber der Konkurrenz absichern können [21].

Freiberufler
Die dritte und letzte Kategorie der hier untersuchten Wirtschaftsbereiche stellen die Freiberufler dar. Im Gegensatz zu den anderen hier vorgestellten Wirtschaftszweigen hätten Softwarepatente höchstwahrscheinlich weitreichende negative Folgen für diesen Berufsstand.

Ein wichtiges Merkmal von Freiberuflern und ihren Dienstleistungsverträgen ist, dass sie für von ihnen verursachte Schäden haften, und zwar mit ihrem Privatvermögen. Dies im Gegensatz zu anderen Gesellschaftsformen, wie der GmbH oder AG. Des Weiteren müssen ihre Arbeitsergebnisse frei von Ansprüchen bzw. Forderungen Dritter sein. Bezüglich des Urheberrechts ist die Gefahr der Forderungen Dritter unkritisch, denn solange nichts illegal kopiert wird, gibt es keine Ansprüche Dritter. Dies ist anders bei Patenten, um hier festzustellen, ob eigene Softwareideen nicht schon patentiert sind, müssten aufwendige Patentrecherchen durchgeführt werden. Ein Aufwand, der das wirtschaftliche Arbeiten eines Freiberuflers in Frage stellt, da die zeitliche und finanzielle Belastung wegen der hohen und unüberschaubaren Zahl von Patenten nicht zu bewältigen ist [16], [21].

Zusätzlich kommt erschwerend hinzu, dass Freiberufler nicht mit Hilfe von Cross-Licensing Abkommen auf die Patentbestände anderer zugreifen können, da sie selbst kaum über eigene Patente verfügen dürften. Die fehlenden eigenen Patente bringen sie ausserdem im Falle einer Verletzungsklage in eine schlechte Verhandlungsposition. Da sie als Freiberufler mit ihrem gesamten Vermögen haften und es wohl kaum eine Haftpflichtversicherung geben wird, die ein solch hohes Risiko tragen würde, wäre jede Verletzungsklage existenzbedrohend [21].

Weiter ist ein Patent ein kleines Monopol, also ein Arbeitsbereich, auf dem ohne kosten- und zeitaufwändige Lizenzvereinbarung nicht legal gearbeitet werden darf. So bedeuten Softwarepatente, dass Freiberufler in einem beachtlichen Bereich nicht mehr ohne weiteres arbeiten können und dürfen [16].

Wissenschaft

Wie bei den anderen Interessengruppen sind auch in der Wissenschaft unterschiedliche Motivationen erkennbar. Massgebend für die jeweilige Interessenlage ist die Ausrichtung der Forschung - grundlagen- oder anwendungsorientiert - und die damit verbundenen Finanzierungsmöglichkeiten. Ebenso spielt aber auch die Einstellung der Wissenschaftler gegenüber der eigenen Tätigkeit eine bedeutende Rolle. Auf der einen Seite stehen die Ideale der Offenheit und Freigebigkeit und die moralische Verpflichtung gegenüber der Öffentlichkeit, auf der anderen Seite Begriffe wie Wissen und Innovation als Wirtschaftsgut, Wirtschaftsstandort, Arbeitsplätze [16].

Technische Hochschulen und anwendungsorientierte Forschung

Anwendungsorientierten Forschungseinrichtungen in der Informatik, die zum Teil hohe Drittmittelaufkommen besitzen, erarbeiten normalerweise keine fertigen Softwareprodukte, sondern ingenieurmässige Lösungskonzepte, die in der Regel formal, empirisch oder durch Prototypenbau evaluiert werden. Solche Lösungskonzepte, wie in anderen Ingenieursdisziplinen auch, fallen dabei nicht unter die vom Urheberrecht geschützten Werke. Vielmehr können solche Arbeitsergebnisse nur über das Patentrecht vor der Benutzung Dritter richtig geschützt werden, oder über strikte Geheimhaltung, was jedoch die Innovationsdynamik und den technischen Fortschritt massiv behindern würde.

Da solche Forschungseinrichtungen im zunehmenden Masse von öffentlichen und privaten Auftraggebern und somit von finanziell starken Investoren abhängig sind, spielt der Schutz des geistigen Eigentums eine immer stärker werdende Rolle. Ein Patent bietet dabei die Möglichkeit, die investierten Forschungsgelder durch Lizenzeinnahmen zumindest Teilweise wieder zu refinanzieren. Würde die entwickelte Innovation dagegen ungeschützt bleiben, wäre ein Rückgang der Investitionsbereitschaft auf diesem Gebiet die Folge, was wiederum kontraproduktiv gegenüber dem technischen Fortschritt wäre. So ist es nicht verwunderlich, dass Auftraggeber aus der Wirtschaft nur noch dann Aufträge vergeben, wenn die Ergebnisse als geistiges Eigentum geschützt werden können [16], [21].

Wenn die Informatik gegenüber anderen Ingenieurwissenschaften wesentlich eingeschränkte Patentierungsmöglichkeiten erhält, so schwächt das deren Position im Wettbewerb der Fächer, aber auch die Stellung der europäischen Informatikforschung und den Innovationsstandort im weltweiten Wettbewerb. Dies insbesondere deshalb, weil Innovationen in der Informatik weitgehend globalisiert stattfinden und sich erfolgreiche Innovationen sowieso im weltweiten Wettbewerb und damit in unterschiedlichen Patentsystemen bewähren müssen. Eine Reihe europäischer Forschungsinstitute ist angesichts der Situation in Europa seit einiger Zeit schon dazu übergegangen, ihre Erfindungen grundsätzlich zuerst in den USA anzumelden, was sicher auch nicht im Interesse Europas liegen kann [16].

Universitäten und Grundlagenforschung

Anders sieht die Situation an Einrichtungen aus, die im wesentlichen Grundlagenforschung betreiben. Diese Einrichtungen sind meist ausnahmslos Grundfinanziert und die Ergebnisse ihrer Arbeit werden der Öffentlichkeit zugänglich gemacht. Das ursprüngliche Wesen von Wissenschaft und Forschung besteht darin, Ideen und Ergebnisse nicht zu schützen,

sondern im Gegenteil der Öffentlichkeit zugänglich zu machen und so auch eine Überprüfung durch andere Wissenschaftler zu ermöglichen. Wissenschaftlicher Fortschritt ist existenziell darauf angewiesen, dass Ergebnisse nutzbar gemacht und von der Gemeinschaft weiterentwickelt werden. Forschung lebt vom Austausch, nicht von der Abkapselung aus wirtschaftlichen Interessen [16].

Die Verpflichtungen der staatlich geförderten Einrichtungen zur Veröffentlichung ihrer Forschungsergebnisse, die mit Hilfe von Steuergeldern erarbeitet wurden, und zur Weitergabe des Wissens an die Allgemeinheit, sind mit Softwarepatenten schlecht vereinbar. So fürchten Forschenden in solchen Einrichtungen auch praktische Auswirkungen. Konkret erwarten Vertreter dieser Interessengruppe eine Verzögerung der wissenschaftlichen Entwicklung durch eine Erweiterung von Schutzrechten. Dass diese Befürchtung begründet ist, scheint durch einige Studien, die seit der Erweiterung des Patentschutzes auf Software in den USA eine Stagnation der Forschungsintensität zeigen, belegt zu sein. Der finanzielle Aufwand, der nötig wäre, um nach einer Ausweitung der Patentierungspraxis mit gleicher Effizienz weiter forschen und lehren zu können, kann von den meisten öffentlich finanzierten Einrichtungen nicht erbracht werden. Dies würde eine erhebliche Einschränkung der täglichen Arbeit bedeuten, in der der offene Austausch von Lösungskonzepten und Verfahren einen besonderen Stellenwert einnimmt [16].

Open Source Community

Die Open Source Community ist die dritte und letzte Interessensgruppe, die im Hinblick auf Softwarepatente untersucht werden soll. Sie zählt zu den energischsten Gegnern von Softwarepatenten, denn die Community sieht darin eine existentielle Bedrohung der gesamten Bewegung und der dahinterstehenden Idee. Das Ziel von Open Source Projekten, die Öffentlichkeit an der Entwicklung und Bearbeitung von Software teil haben zu lassen, steht dabei im totalen Gegensatz zu den Gedanken des Patentrechts, Dritte ohne Lizenzzahlungen von einer Erfindung auszuschliessen. So verträgt sich die Idee, die hinter der Open Source Bewegung steht, nicht mit den Grundsätzen des Patentwesens. Aus diesem Grund spricht sich die grosse Mehrzahl der Anhänger der Open Source Bewegung dafür aus, Software ausschliesslich über das Urheberrecht und nicht über das Patentrecht zu schützen. Die Gründe dafür sind offensichtlich, wenn man die Vorteile des Urheberrechts und die Nachteile von Patenten für Open Source näher betrachtet [16].

Die einfache Handhabung des Urheberrechts hat viel Vorteil für Open Source Projekte, denn so lange sich alle an die Gesetze halten, kann man sich auf die Entwicklungsarbeit konzentrieren. Erst bei Rechtsverletzungen ist die Unterstützung von Juristen notwendig, deren Kosten der Rechtsbrecher zu tragen hat. Patente dagegen erfordern Verwaltungsaufwand für Recherche und Anmeldung schon während der Entwicklung [16]. Das grosse Problem was dabei entsteht, sind die gigantischen Kosten, die im Verhältnis zu den Einnahmen eines Open Source Projektes stehen. Zum einen ist es für die Community beinahe unmöglich selber Patente anzumelden, denn die dabei entstehenden Gebühren für Anmeldung und Honorar eines Patentanwalts würden das Budget um ein Vielfaches überschreiten. Und zum anderen sind auch aufwendige Patentrecherchen und allfällige Lizenzzahlungen auf Grund der geringen Finanzmittel nicht möglich [21].

Zudem sind Patente immer dann kritisch, wenn sie die Nutzung von Datenformaten oder Protokollen betreffen. Es ist dann für Open Source Software nicht mehr möglich, interoperabel mit der zugehörigen kommerziellen Software zu sein [16]. Ein weiteres Problem bei Open Source Software ist die grosse Gefahr, bedingt durch die Offenheit des Quellcodes, dass Patentverletzungen sehr schnell erkannt und nachgewiesen werden können. Dies ist besonders problematisch, da Patentverletzungsprozesse für einen Privatmann schnell existenzbedrohend werden können. Je mehr Open Source Software Lösungen in grosse Unternehmen und Behörden eingesetzt werden und eine Alternative zu den Produkten grosser Anbieter darstellen, desto mehr werden diejenigen, die sich diesem Wettbewerbsdruck nicht mehr über Qualität und Preis stellen können bzw. wollen, in Versuchung geraten, ihre üblicherweise sehr umfangreichen Patentarsenale einzusetzen, um durch Patentverletzungsklagen die Open Source Software vom Markt zu verdrängen [14]. Ebenso ist die in Europa existierende Veröffentlichungssperre vor der Patentbeantragung eine weitere Hürde, die sich nicht mit Open Source verträgt, da dort die Entwicklung in der Öffenlichkeit stattfindet [16]. So ist auf Grund der verschiedenen Grundphilosophien zwischen Open Source Bewegung und Patentwesen eine Lösung der Problematik mit geltendem Patentrecht kaum zu erwarten. Es lassen sich höchstens Wege finden, um die negativen Auswirkungen zu begrenzen. Zum Beispiel bestünde eine Möglichkeit darin, so früh wie möglich die entwickelten Computerprogramme von Open Source Projekten zu veröffentlichen. Denn solche gerade im Internet veröffentlichten Programmlistings sind für jede Patentanmeldung neuheitsschädlich [21].

8.5 Auswirkungen von Softwarepatenten

Generell ist zu sagen, dass die Auswirkungen von Softwarepatenten schwierig zu bestimmen sind. Studien ist es bisher nicht wirklich gelungen, eine Notwendigkeit oder eine positive Wirkung von Softwarepatenten auf die Volkswirtschaft nachzuweisen, welche die Position der Befürworter untermauern könnte. Es ist allerdings auch weitgehend unbestritten, dass ausserhalb des Gebietes von Softwarepatenten Innovationen durch das Patentsystem gut geschützt und gefördert werden, so dass es für Kritiker schwierig ist, nachzuweisen, dass dies auf dem Gebiet der Softwarepatente nicht der Fall ist. Währendem die theoretische Literatur zu Patenten wenig geeignet ist, in einem bestimmten Sektor zwischen einem starken oder weniger starken Schutzrecht zu differenzieren, hinterlassen die Erkenntnisse der theoretischen Modelle Zweifel an der Effizienz starker Schutzrechte in der Softwarebranche. Insbesondere wegen der Häufigkeit von sequentiellen Innovationen, die durch Patente blockiert werden können, sowie höheren Marktzutrittsschranken durch solche blockierende Patente.

8.5.1 Wirtschaft und F&E

Die wirtschaftlichen Auswirkungen von Softwarepatente müssen vor dem Hintergrund der ökonomischen Ziele des Patentsystems gesehen werden. Grundsätzlich wäre es aus gesamtwirtschaftlicher Sicht wünschenswert, wenn existierende Erfindungen von jedermann genutzt werden könnten. Allerdings wäre es dann dem Erfinder zumeist nicht möglich, die

Kosten seiner Innovationstätigkeit zu decken und Gewinne zu erwirtschaften, was seine Innovationsanreize reduzieren würde [18]. So ist die übliche ökonomische Begründung für Patente, der Schutz potentieller Innovatoren vor Imitationen, damit ein Anreiz besteht, Innovationskosten auf sich zu nehmen. Mögliche Konkurrenten müssen also davon abgehalten werden, eine Erfindung zu imitieren, damit der Erfinder genügend Gewinne macht, um seine Innovationskosten abzudecken [17]. Aus diesem Grund wird dem Patentinhaber ein zeitlich begrenztes Schutzrecht gewährt. Dabei werden aber auch negative Effekte in Kauf genommen. Erstens wird die Erfindung weniger breit verwendet, zweitens werden darauf aufbauende weitere Innovationen behindert, und drittens führt dieses Schutzrecht zu einem temporären, potentiell ineffizientem, Monopol. Das ökonomische Ziel des Patentsystems muss es nun sein, die innovationsförderlichen und die innovationsschädlichen Effekte des Patentschutzes optimal gegeneinander abzuwägen [18].

Es spricht einiges dafür, dass die negativen Effekte bei Softwarepatenten besonders stark sind. Innovationen im Softwarebereich sind häufig sequentiell, das heisst sie bauen, mehr als in anderen Branchen, auf vorhergehenden Innovationen auf. Aus diesem Grunde wiegt der negative Aspekt von Patenten, weiterführende Entwicklungen zu erschweren oder sogar zu blockieren, bei Software besonders schwer. Lizenzierungen lösen dieses Problem aufgrund der mit ihnen verbundenen Transaktionskosten nur sehr bedingt. Da der Softwaremarkt ein Markt mit extrem geringen Zutrittskosten ist, können schon kleine Transaktionskosten zu grossen Veränderungen des Marktzutritts führen [16].

In der Studie vom MIT (Massachusetts Institute of Technology) von James Bessen und Eric Maskin [17] wird anhand eines einfachen Modells aufgezeigt, dass Patentschutz in dynamischen Industriezweigen, wie der Software-, Halbleiter- und Computerbranche, zu einem Rückgang der Gesamtheit der Innovationen und infolgedessen zu einer Verringerung des sozialen Nutzens führen kann. Es gibt gute Gründe anzunehmen, dass in diesen Industriezweigen, entgegen der oben genannten üblichen ökonomischen Sichtweise, Imitationen Innovationen fördern und starker Patentschutz, das heisst eine lange Schutzdauer und grosser Schutzbereich, Innovationen hemmt. Das liegt daran, dass in diesen Industriezweigen Innovationen häufig sowohl sequentiell als auch komplementär sind. Mit „sequentiell“ ist gemeint, dass jede Erfindung auf einer vorangegangenen aufbaut, wie zum Beispiel Windows auf DOS aufbaut. Mit „komplementär“ ist gemeint, dass jeder potentielle Innovator einen etwas anderen Forschungsansatz wählt und dadurch die Gesamtwahrscheinlichkeit erhöht wird, ein bestimmtes Ziel innerhalb einer vorgegebenen Zeit zu erreichen. So haben beispielsweise die vielen unterschiedlichen Ansätze bei der Entwicklung von Spracherkennungssoftware die Einführung von erschwinglichen Produkten auf dem Markt beschleunigt. Ein Unternehmen, das sein Produkt in einer Welt sequentieller und komplementärer Innovationen patentiert, kann verhindern, dass seine Konkurrenten dieses Produkt oder ähnliche Ideen zur Entwicklung weiterer Innovationen verwenden. Sollten diese Konkurrenten, nicht aber das ursprüngliche Unternehmen über wertvolle Ideen zur Weiterentwicklung dieses Produkts verfügen, könnten weitere Erfindungen durch das Patent gebremst oder verhindert werden. Ausserdem könnte es, unabhängig davon, ob Patentschutz besteht oder nicht, für ein Unternehmen von Vorteil sein, wenn andere Unternehmen seine Produkte imitieren und mit ihm in Wettbewerb treten. Obwohl Imitationen den gegenwärtigen Gewinn des Unternehmens verringern, erhöhen sie die Wahrscheinlichkeit weiterer Innovationen und verbessern somit die Aussicht des Unternehmens auf eine weitere gewinnbringende Entdeckung. So bedeutet dies, dass bei

sequentiellen und komplementären Innovationen Standardargumente über Patente und Imitation auf den Kopf gestellt werden können. So dass Imitationen einen Ansporn für Innovation bedeuten, wohingegen starke Patente ein Hindernis darstellen [17].

Die Veränderungen, die sich ergaben, als der Patentschutz in den 80er Jahren auf die Softwarebranche in den USA ausgeweitet wurde, werden zur Verifizierung dieses Modells herangezogen. Die übliche Argumentation läuft ja darauf hinaus, dass sich der Forschungs- und Entwicklungsaufwand und die Produktivität bei Unternehmen, die Patente anmelden, erhöhen sollten. Sie zeigen aber, dass die stärkeren Eigentumsrechte in den USA gerade nicht eine Welle von Innovationen hervorbrachten, sondern im Gegenteil dazu führten, dass die F&E-Aktivitäten gerade bei solchen Branchen und Unternehmen, die viele Patente anmeldeten, stagnierten, wenn nicht sogar zurückgingen. Darüber hinaus wird das Modell gestützt durch die Tatsachen, dass in diesen Industriezweigen Cross-Licensing auffällig häufig und die Anzahl der Markteintritte proportional zur Anzahl der Innovationen ist. In der Softwarebranche sind Cross-Licensing Abkommen von Patentportfolios unter direkten Wettbewerbern häufig zu beobachten. Insbesondere umfassen diese Lizenzabkommen oft auch zukünftige Patente. Ein solches Verhalten ist schwer im Rahmen eines klassischen Patentmodells zu erklären. Es passt eher zu einer dynamischen Branche, in der Unternehmen nicht versuchen, im Rahmen von Lizenzeinnahmen F&E-Ausgaben zu kompensieren, sondern um die Blockierwirkung der gegnerischen Patente besorgt sind. Ebenso deutet das Verhältnis von Innovationen und Markteintritten auf eine dynamische Branche hin, wo die Innovationsrate steigt, wenn sich viele neue Unternehmen am Markt beteiligen. Dies steht ebenfalls im Gegensatz zum statischen, klassischen Patentmodell. Während der Lebensdauer eines durch Innovation entstandenen Produktes kommt es zu einer Schwankung bei der Anzahl der Markteintritte, die mit dem Patentschutz zusammenhängen. Anfangs besitzt typischerweise ein einzelnes Unternehmen eine Monopolstellung, die durch Patente gestützt ist. Wenn die Patente ablaufen oder aus anderen Gründen Markteintritte möglich werden, können sich andere Unternehmen am Markt beteiligen. Im Laufe der Zeit werden dann zusätzliche Innovationen entwickelt, die das Produkt verbessern. Durch die Patentierung dieser Verbesserungen wird es allmählich wieder schwieriger für Unternehmen sich am Markt zubeteiligen. Dies führt schliesslich dazu, dass sich gegen Ende des Lebenszyklus keine neuen Unternehmen mehr am Markt beteiligen, weniger erfolgreiche Unternehmen aus dem Markt aussteigen und sich die Branche stabilisiert. Wie die Studie zeigt, ist die Innovationsrate weder während der Monopolstellung noch gegen Ende des Lebenszyklus besonders hoch. Tatsächlich treten die höchsten Innovationsraten in der Phase auf, in der sich neue Unternehmen am Markt beteiligen und die Zahl der Unternehmen am grössten ist. Dies bedeutet, dass die Anzahl der Innovationen proportional zur Anzahl der Markteintritte ist [17].

Problematisch an Softwarepatenten ist zudem die vielfach geringe Erfindungshöhe. Es gibt zahlreiche Beispiele dafür, dass Ideen, die für einen kundigen Programmierer offensichtlich sind, durch Patente geschützt wurden. Damit werden die ökonomischen Ziele des Patentsystems ad absurdum geführt. Denn es kann nicht darum gehen, Patente an den schnellsten Anmelder offensichtlicher Ideen zu vergeben. Nur wenn der durchschnittliche Aufwand für eine Erfindung in einem vernünftigen Verhältnis steht zu der erwarteten Rendite aus deren patentgeschützter Nutzung, wird die intendierte Anreizwirkung des Patentsystems erreicht. Andernfalls entsteht ein höchst ineffizientes Dickicht aus banalen Patenten [18]. Durch die Vielzahl zum Teil offensichtlicher Patente entsteht ausserdem eine

hohe Unsicherheit. Es wird schwieriger festzustellen, ob ein Computerprogramm irgendein bestehendes oder gar ein angemeldetes und noch nicht erteiltes Patent verletzt. Diese Unsicherheit stellt gerade für kleine Unternehmen ein Risiko dar. Eine kleinere Firma, die von einem grossen Unternehmen wegen vermeintlicher Patentverletzung verklagt wird, kann schnell in wirtschaftliche Schwierigkeiten geraten. Umgekehrt schützt ein eigenes Patent eine kleinere Firma nur begrenzt vor Verletzungen durch grössere Wettbewerber, die im Zweifel bessere Anwälte und stärkeren finanziellen Rückhalt haben. Nur wenn das Patent so eindeutig ist, dass die Firma Verletzungen leicht nachweisen und Prozesse sicher gewinnen kann, nützt der Patentschutz dem Unternehmen [18].

Ergebnisse einer empirischen Studie

Das Fraunhofer Institut für Systemtechnik und Innovationsforschung hat in Zusammenarbeit mit dem Max-Planck-Institut für ausländisches und internationales Patent-, Urheber- und Wettbewerbsrecht eine empirische Untersuchung durchgeführt, um die Mikro- und makroökonomischen Implikationen der Patentierbarkeit von Softwareinnovationen zu untersuchen [15]. Dazu wurden 263 deutsche Unternehmen einschliesslich der freien Softwareentwickler befragt. Die Studie wurde im Frühjahr 2001 auf einer Internet basierenden Befragung durchgeführt und hat Software erstellende Unternehmen unterschiedlicher Grösse und Einstellung zu Patenten aus der Primärbranche (reine Softwareanbieter und IT-Dienstleister) und der Sekundärbranche (verarbeitendes Gewerbe mit eigener Softwareentwicklung) zu den konkreten Auswirkungen von Softwarepatenten befragt. Unter anderem wurden die Unternehmen nach den Konsequenzen einer stärkeren, an der Praxis der USA orientierten Patentierung von Software für das eigene Unternehmen und für die jeweilige Branche als Ganzes befragt (siehe Abbildung 8.5). Diese Einschätzung durch die Unternehmen selber, kann hier als wichtiges Indiz für die möglichen Auswirkungen von Softwarepatenten dienen.

Während die Sekundärbranche Vor- und Nachteile stärkerer Patentierung wahrnimmt, sieht die Primärbranche überwiegend negative Konsequenzen. Offensichtlich wären die Erhöhung der Kosten die unmittelbarste und eindeutigste Konsequenz, und zwar für beide Branchen. Ökonomisch bedeutsamer dürften die Folgen für die Konkurrenzfähigkeit, die Produktqualität, die Entwicklung von Open Source und die Innovationsdynamik sein. Die Primärbranche sieht überdies im Durchschnitt negative Auswirkungen für die Rechtssicherheit, die Beschäftigung und die Konkurrenzfähigkeit im In- und Ausland. Die Antworten der Sekundärbranche sind ambivalenter. So nehmen die Unternehmen der Sekundärbranche die Möglichkeiten zur Patentierung auch als ein effektives Instrument zur Verbesserung der eigenen Konkurrenzfähigkeit wahr. Jedoch befürchten auch sie einen Rückgang der Innovationsdynamik, der sich aber wohl nicht negativ auf die eigene Konkurrenzfähigkeit auswirkt [19].

Zu den Auswirkungen von Patenten auf das Innovationsverhalten ergab die Studie weiter, dass die Primär- und die Sekundärbrache einer sehr hohen Dynamik auf der Angebots- und Nachfrageseite gegenüber stehen und die Entwicklungsdauer für neue Innovationen sehr knapp ist. In der Softwarebranche sind sequentielle Weiterentwicklungen sehr häufig und im Wettbewerb sind rasche Innovation und effektive Entwicklungsprozesse entscheidend. So haben hier Einschränkungen durch Patente intensive wirtschaftliche Auswirkungen auf

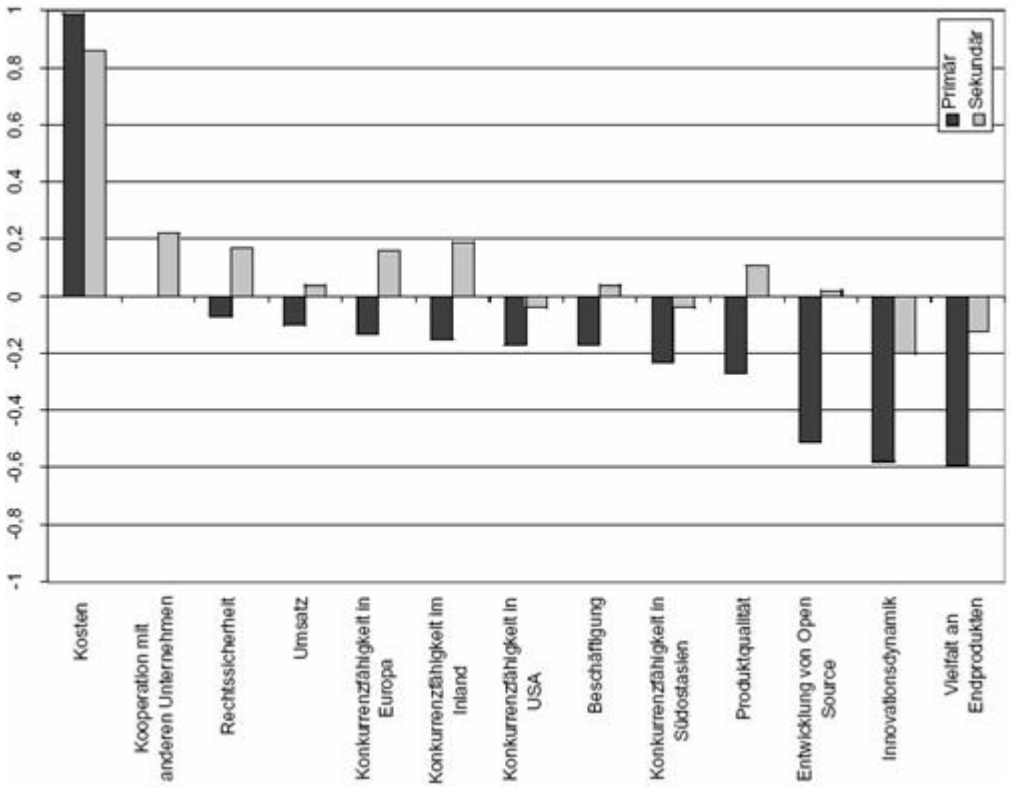


Abbildung 8.5: Erwartete Konsequenzen durch Softwarepatente für das eigene Unternehmen (+1 = steigt, -1 = sinkt), Quelle: [19]

das Innovationsverhalten. Deshalb hat die Primärbranche vor allem Bedenken, dass die Innovationsdynamik gehemmt wird und die Kosten und Unsicherheiten, die mit der Patentanmeldung verbunden sind, die innovationsfördernde Wirkung übersteigen. Die Sekundärbranche, die auf Erfahrung im Bereich der Patente aufbaut, befürchtet die mangelnde Nachweisbarkeit von Patentverletzungen, die Durchsetzbarkeit und damit die Schutzwirkung von Patenten im Softwarebereich. Diese negative Sichtweise begründet auch auf Erfahrungen mit Rechtstreitigkeiten, da fast 20 Prozent der Unternehmen der Primärbranche und 40 Prozent der Unternehmen der Sekundärbranche bereits in Rechtsstreitigkeiten verwickelt waren und die Durchführung eigener Entwicklungstätigkeiten aufgrund von Patenten eingeschränkt war [20]. Es wird auch deutlich, in welcher Weise Patente auf die Marktstruktur Einfluss nehmen. Die Unternehmen erwarten, dass die Wettbewerbsintensität abnimmt, eine Reihe von Unternehmen vom Markt verschwinden und sich Monopolpositionen verstärken. Das sind in der Einschätzung beider Branchen die negativsten Konsequenzen. Damit einhergehend würde in der Erwartung der Unternehmen der Primärbranche die Vielfalt an Endprodukten und an Komponenten sinken, die Interoperabilität vermindert sowie weniger in die Entwicklung von Open Source investiert. Vor dem Hintergrund der stark zunehmenden Bedeutung von Open Source Software liefe so eine stärkere Patentierung einem wichtigen Trend der Softwareentwicklung zuwider. Da die Nutzer von Open Source die Vorteile hauptsächlich in der Produktqualität sehen, sind die negativen Erwartungen für die Produktqualität in der Primärbranche nur folgerichtig [20]. Die Unternehmen der Sekundärbranche sehen die potenziellen Auswirkungen auf ihre eigene Branche etwas weniger kritisch. Bei zunehmender Konzentration erwarten

sie Umsatzwachstum, Beschäftigungswachstum und stärkere Rechtssicherheit und sehen keine Abschreckung zur Erstellung von Open Source Software. Nach Ansicht der Sekundärbranche wird auch die Interoperabilität durch Patentierung verbessert. Dies zeigt einen fundamentalen Mentalitätsunterschied der beiden Branchen. Einerseits die Befürchtung der Schliessung eines bisher offenen Systems und anderseits die Aussicht, über die Patentoffenlegung in der Branche leichter kompatible Produkte herstellen zu können. Aber vor allem innerhalb kleinerer Unternehmen wird diese Informationsfunktion von Patenten kaum wahrgenommen [19], [20].

Die Studie zeigt ebenfalls, dass der Umgang mit Schutzrechten in der Softwareindustrie noch gering institutionalisiert ist. Unternehmen, die einen solchen Bedarf sehen, decken ihn meist über externe Beratung ab. Auch das Wissen über Schutzrechte, insbesondere Patente, ist sowohl in der Primär- als auch in der Sekundärbranche noch schwach ausgebildet. Dadurch wird deutlich, dass dem Schutz durch Softwarepatente von allen Möglichkeiten zum Schutz eine relativ geringe Bedeutung beigemessen wird. In beiden Branchen werden Patente nur als Ergänzung der gesamten Palette an möglichen formellen und informellen Schutzmassnahmen gesehen. Patente sind bis heute die am wenigsten verbreitete Schutzmassnahme. Bei weitem am häufigsten werden verschiedenen Formen der internen Geheimhaltung genutzt, gefolgt von Strategien zur Erhaltung eines zeitlichen Marktvorsprungs und zur Kundenbindung. Die Primär- und die Sekundärbranche haben in der Tendenz also eine skeptische Haltung zu Softwarepatenten, wobei die Primärbranche durchwegs skeptischer ist. Die Primärbranche befürwortet überwiegend eine generelle Ausschliessung von Software aus der Patentierung, die Sekundärbranche ist in ihrer Meinung weniger eindeutig. Von einer breiteren Patentierung nach dem Vorbild der USA würde aber die Mehrheit fast ausschliesslich negative Konsequenzen für ihr eigenes Unternehmen wie auch für die gesamte nationale Branche erwarten, insbesondere für die Innovationsdynamik, die Produktqualität, die Vielfalt der Produkte. und die Interoperabilität [19], [15].

8.6 Zusammenfassung und Ausblick

Weder gibt es auf wirtschaftstheoretischer Ebene stichhaltige Argumente für die unbedachte Verstärkung des Rechtsschutzes von Software, noch gibt es auf empirischer Ebene Belege dafür, dass eine solche Verstärkung gesamtwirtschaftlich wünschenswerte Effekte hätte. Empirische Analysen deuten eher darauf hin, dass eine solche Verstärkung, wie in den USA, nicht mit der damit gewünschten, gesellschaftlich effizienten Verstärkung der Forschungsintensität einhergeht, sondern sogar im Gegenteil ein Abflachen oder Abnehmen der F&E-Intensität zur Folge haben kann. Da ein Softwarepatent nur einer von mehreren Schutzmechanismen ist, und oft nicht der effektivste, ist es durchaus nicht so, dass eine Innovation ohne Patentschutz automatisch in Allgemeinbesitz übergeht. Schutzmechanismen, wie Geheimhaltung, Schnelligkeit der Umsetzung sowie Vertriebswege oder Serviceangebote werden in vielen Branchen als bessere Schutzmassnahmen angesehen. Im Softwarebereich sind Urheberrecht sowie Geheimhaltung sehr wirksam, und werden vielfach als ausreichend angesehen. Zudem ist im internationalen Wettbewerb der strategische Nutzen von Softwarepatenten sehr gering und kommt nur für relativ wenige grosse

Unternehmen zum Tragen. Allgemein sind die langfristigen Kosten für eine breitere Patentierung höchstwahrscheinlich höher als die Beeinträchtigung der Innovationsdynamik. Es ist nicht einmal sicher, ob überhaupt eine innovationsfördernde Wirkung von Softwarepatenten ausgeht.

Wenn denn Softwarepatente vergeben werden, sollten zumindest die Prüfungsrichtlinien der Patentämter verbessert werden und nicht nur eine Überprüfung, ob die Software die Eigenschaft der Technizität erfüllt, überprüft werden. Ebenfalls dürfen Patente nur an wirklich neue und nicht-offensichtliche Erfindungen vergeben werden. Andernfalls werden Innovationen sicher nicht gefördert, sondern gebremst. Bei der Entscheidung über Softwarepatente und der Ausgestaltung des Patentsystems, muss vor allem immer die Innovationsdynamik, die Sequenzialität und die Interoperabilität in der Softwareentwicklung berücksichtigt werden.

Literaturverzeichnis

[1] Deutscher Bundesgerichtshof: Rote-Taube Entscheid, http://www.jura.uni-freiburg.de/institute/prurh/downloads/BGH\%20RoteTaube.pdf, aufgerufen am 13.11.2005.

[2] Ilzhöfer, V.: Patent-, Marken- und Urheberrecht, 4. Aufl., Verlag Franz Vahlen, München 2000.

[3] Eisenmann, H.: Grundriss Gewerblicher Rechtsschutz und Urheberrecht. C. F. Müller Verlag, Heidelberg 2001.

[4] Bundesgesetz über das Urheberrecht und verwandte Schutzrechte, http://www.admin.ch/ch/d/sr/231_1/index.html, aufgerufen am 25.11.2005.

[5] Der Brockhaus: in 15 Bänden. Permanent aktualisierte Online-Auflage. Leipzig, Mannheim: F.A. Brockhaus 2002, 2003, 2004.

[6] EU: Zusammenfassung - Patentierbarkeit computerimplementierter Erfindungen, http://europa.eu.int/scadplus/leg/de/lvb/l26090.htm, Juli 2004, Aufgerufen am 13.11.2005.

[7] Amtsblatt der Europäischen Gemeinschaften: Vorschlag für eine Richtlinie des Europäischen Parlaments und des Rates über die Patentierbarkeit computerimplementierter Erfindungen, http://europa.eu.int/eur-lex/pri/de/oj/dat/2002/ce151/ce15120020625de01290131.pdf, 20.2.2002, Aufgerufen am 13.11.2005.

[8] Justiz- und Polizeidepartement (EJPD): Antwort auf die Anfrage von Ruedi Noser zu „Übernahme der Software-Patentrichtlinien der EU durch die Schweiz“ , urlhttp://www.parlament.ch/afs/data/d/gesch/2004/d_gesch_20041094.htm, September 2004, Aufgerufen am 24.11.2005.

[9] Europäisches Patentamt: EUROPÄISCHES PATENTÜBEREINKOMMEN, 11. Auflage, http://db1.european-patent-office.org/www3/dwld/epc/epc_2002_v1.pdf, Juli 2002, Aufgerufen 25.11.2005.

[10] Agreement Establishing the World Trade Organization, Annex 1c: AGREEMENT ON TRADE-RELATED ASPECTS OF INTELLECTUAL PROPERTY RIGHTS, urlhttp://www.wto.org/english/docs_e/legal_e/27-trips.pdf, April 1994, Aufgerufen am 13.1.2006.

[11] U.S. Fed Circuit Court of Appeals: STATE STREET v SIGNATURE, http://laws.findlaw.com/fed/961327.html, Juli 1998, Aufgerufen am 25.11.2005.

[12] Rechtstext: Pariser Verbandsübereinkunft zum Schutz des gewerblichen Eigentums, Systematische Sammlung des Bundesrechts, http://www.admin.ch/ch/d/sr/0_232_04/, Juli 1967, Aufgerufen am 9.1.2006.

[13] Confédération Européenne des Associations de Petites et Moyennes Entreprises: Stellungnahme zum Richtlinienvorschlag der EU-Kommission zu Software-Patenten, http://swpat.ffii.org/papers/eubsa-swpat0202/ceapme0309/ceapme-pr0309.de.pdf, 15.09.2003, Aufgerufern am 16.11.2005.

[14] Florian Müller: Open Source und Softwarepatente: Exklusives Oder für Politik und Gesellschaft? http://ec01.et-inf.fho-emden.de/isos2/images/stories/isos2004/isos_2004_proceedings.pdf, September 2004, Aufgerufen am 20.11.2005.

[15] Frauenhofer Institut für Systemtechnik und Innovationsforschung und Max-Planck-Institut für ausländisches und internationales Patent-, Urheber- und Wettbewerbsrecht: Mikro- und makroökonomische Implikationen der Patentierbarkeit von Softwareinnovationen: Geistige Eigentumsrechte in der Informationstechnologie im Spannungsfeld von Wettbewerb und Innovation, http://www.juergen-ernst.de/download_swpat/studie_bmwi.pdf, September 2001, Aufgerufern am 20.11.2005.

[16] Gesellschaft für Informatik e.V. (GI): Positionspapier der Gesellschaft für Informatik e.V. (GI) zur Patentierbarkeit rechnergestützter Erfindungen, http://www.gi-ev.de/fileadmin/redaktion/Patente/patentierung2005.pdf, 4. Juli 2005, Aufgerufen am 16.11.2005.

[17] James Bessen, Eric Maskin: Sequentielle Innovation, Patente und Imitation, http://www.researchoninnovation.org/patentde.pdf, November 1999, Aufgerufen am 16.11.2005.

[18] Joachim Henkel: Zuviel Schutz schadet - warum Patente auf Software problematisch sind, http://www.inno-tec.bwl.uni-muenchen.de/forschung/henkel/SW-Patente_JHenkel_2002-11.pdf, November 2002, Aufgerufen am 16.11.2005.

[19] Knut Blind, Jakob Edler, Michael Friedewald: Wer braucht eigentlich Software-Patente? Ergebnisse einer empirischen Untersuchung, http://www.friedewald-family.de/Publikationen/GI2001.pdf, 2001, Aufgerufen am 20.11.2005.

[20] Michaela Glaser: Die Sicherung von Rechten des geistigen Eigentums: Konflikte und Interessen der Softwareindustrie um das TRIPS-Abkommen, http://www.wifo.ac.at/Stefan.Schleicher/down/da/DA_Glaser.pdf, Mai 2003, Aufgerufen am 16.11.2005.

[21] Sebastian Aisch: Analyse und Bewertung der Software-Patentsituation, http://ivs.cs.uni-magdeburg.de/sw-eng/agruppe/forschung/diplomarbeiten/aisch.pdf, 25. September 2005, Aufgerufen am 16.11.2005.

Kapitel 9

Die ökonomischen Einflüsse der geistigen Eigentumsrechte in der Informations- und Kommunikationstechnologiebranche

Claudia Bretscher, Ursula D'Onofrio, Lukas Eberli

Diese Arbeit untersucht die ökonomischen Einflüsse der geistigen Eigentumsrechte auf die Branche der Informations- und Kommunikationstechnologie (IKT). Für die IKT-Branche sind vor allem Patente, Marken, Industriedesigns und Urheberrechte relevant. Geistige Eigentumsrechte können einen Beitrag zur Wohlstandsbildung und zu wirtschaftlichem Wachstum leisten. Patente verstärken die Motivation für neue Erfindungen, Unternehmen können sich durch starke Marken von der Konkurrenz abheben und Urheberrechte schützen Werke wie z.B. Software vor unautorisierter Reproduktion. Neue Technologien veränderten in den letzten Jahrzehnten das Umfeld der IKT-Branche und führten zu neuen Problemen und Herausforderungen. Durch die Internationalisierung des Internets wächst das Risiko der Verletzung von nationalen Gesetzen und der Wunsch nach internationaler Zusammenarbeit und Angleichung der Rechtsvorschriften wird grösser [1]. Der illegale Tausch von urheberrechtlich geschützten Werken wurde durch das Internet vereinfacht. Könnte die Softwarepiraterie-Rate um nur wenige Prozentpunkte verringert werden, würde dies die Entwicklung der IKT-Branchen nachhaltig fördern und zu höherem Wirtschaftswachstum führen [2]. Softwarehersteller versuchen zusehends, ihre Produkte durch neue Technologien vor urheberrechtsverletzenden Aktionen zu schützen. Dadurch werden aber nicht nur Urheberrechte geschützt sondern auch die Freiheit der Benutzer eingeschränkt. Es existieren bereits heute Lösungsansätze für diese neuen Herausforderungen: Verbesserte, flexiblere Verschlüsselungstechnologien sollen helfen, die richtige Balance zwischen Schutz und Einschränkung zu finden. Neue internationale Gesetze und Vereinbarungen können dazu beitragen, die Folgen des Territorialprinzips zu verringern und Registrationen von geistigen Eigentumsrechten zu vereinfachen.

Inhaltsverzeichnis

9.1 Einleitung

Diese Arbeit führt kurz in das Thema der geistigen Eigentumsrechte ein. Da sich diese Seminararbeit hauptsächlich mit den ökonomischen Einflüssen dieser Rechte auf die Branche der Informations- und Kommunikationstechnologie (IKT) beschäftigt, werden diese Rechte hier nur kurz beschrieben.

Geistige Eigentumsrechte können, wie der Name erahnen lässt, in zwei Hauptkategorien aufgeteilt werden [3]: Gewerbliche Rechte zum Schutz des Eigentums und Urheberrechte. Die erste Kategorie besteht aus Patenten, Marken, Industriedesigns und geographischen Herkunftsbezeichnungen. Die zweite Kategorie beschreibt urheberrechtliche Bestimmungen und verwandte Schutzrechte.

`Patente` [4]: Patente sind von der zuständigen Behörde erteilte Schutztitel für Erfindungen. Eine Erfindung im rechtlichen Sinne löst ein technisches Problem mit den Mitteln der Technik und ist nur dann patentierbar wenn sie neu ist, nicht nahe liegt und gewerblich angewendet werden kann. Das Patent verschafft seinem Inhaber das ausschliessliche Recht die Erfindung gewerbsmässig (z.B. Herstellung, Verwendung, Verkauf oder Einfuhr) während einer beschränkter Dauer (in der Schweiz 20 Jahre) zu benützen. Es steht dem Inhaber offen, anderen dieses Recht zu übertragen, sei es durch Verkauf des Patentes oder durch Lizenzverträge. Als Gegenleistung für das erhaltene Schutzrecht muss die Erfindung offen gelegt werden. Das heisst sie muss so genau beschrieben werden, dass ein Fachmann in der Lage wäre, die Erfindung nachzuvollziehen.

`Marken` [3]: Eine Marke ist ein unverwechselbares Zeichen, welches bestimmte Produkte oder Dienste als solche identifiziert, die von einer bestimmten Person oder Unternehmung hergestellt wurden. Marken dienen dazu, Waren eines Unternehmens von jenen anderer Unternehmen zu unterscheiden. Die registrierte Marke gibt dem Benutzer das exklusive Recht diese zu benützen, Produkte oder Dienste damit zu kennzeichnen oder einem anderen dessen Benutzung gegen Bezahlung zu erlauben. Eine Eintragung ist in der Schweiz während zehn Jahren wirksam und kann beliebig oft jeweils für weitere zehn Jahre verlängert werden.

`Industriedesigns` [4]: Industriedesigns stellen die bestimmte äussere Formgebung von etwas Zweidimensionalem (Muster) oder von etwas Dreidimensionalem (Modell) dar. Die Formgebung eines Modells darf nicht allein von der Funktion bestimmt sein, für die der Gegenstand gedacht ist. Voraussetzungen für den gesetzlichen Schutz sind, dass die gewerblichen Muster und Modelle neu und originell sind. Eine Form kann in der Schweiz höchstens während 25 Jahren geschützt werden, aufgeteilt in fünf Schutzperioden zu fünf Jahren.

`Geographische Herkunftsbezeichnungen` [4]: Geografische Herkunftsbezeichnungen sind Orte, Gegenden oder Länder, die die territoriale Herkunft von Erzeugnissen identifizieren. Es können auch Regionen sein, denen die Qualität, der Ruf oder eine andere Eigenschaft des Erzeugnisses im wesentlichen zugeschrieben wird. Herkunftsbezeichnungen sind Hinweise auf die geografische und nicht die unternehmensmässige Abstammung eines Produkts. Sie weisen auf eine bestimmte Gegend und damit auf alle Unternehmen hin, die sich dort mit der Anfertigung dieser Waren oder Dienstleistungen befassen.

`Urheberrechte` [3]: Urheberrechte sind Rechte welche dem Ersteller (Inhaber des Urheberrechts) an seinen Werken der Literatur und Kunst zustehen. Darunter fallen literarische Werke (Gedichte, Romane, Computerprogramme, etc.), Filme/Musik, gestalterische Werke (Zeichnungen, Fotografien, Skulpturen, etc.), Architekturen, Karten oder technische Zeichnungen. Das Urheberrecht gibt dem Inhaber das Recht zur Reproduktion, für öffentliche Aufführungen oder Aufnahmen, beispielsweise auf CD oder DVD. Auch moralische Rechte werden eingeräumt wie z.B. das Recht, Änderungen am eigenen Werk zu verhindern die der Reputation des Urhebers schaden könnten. Urheberrechte müssen nicht registriert werden und bleiben in der Schweiz bis 70 Jahre (bei Computerprogrammen 50 Jahre) nach dem Tod des Erstellers erhalten.

`Verwandte Schutzrechte` [4]: Verwandte Schutzrechte sind Rechte die um die Urheberrechte herum entstanden sind. Damit werden ausübende Künstler, Hersteller von Ton- und Tonbildträgern sowie Sendeunternehmen (insbesondere TV und Radio) geschützt. Der Schutz erstreckt sich auf Darbietungen und Wiedergaben auf Ton- und Tonbildaufnahmen und auf Sendungen. Die verwandten Schutzrechte erlöschen in der Schweiz 50 Jahre nach der Darbietung des Werkes durch den ausübenden Künstler, der Herstellung der Ton- oder Tonbildträgern oder der Ausstrahlung der Sendung.

9.2 Ökonomische Einflüsse aus Sicht verschiedener Anpruchsgruppen der Informationstechnologiebranche

Dieses Kapitel erklärt die wirtschaftlichen Auswirkungen der geistigen Eigentumsrechte auf die IKT-Branche. Meinungen und Sichtweisen von verschiedenen Anspruchsgruppen werden dargestellt und erläutert. Behandelt werden Patente, Marken, Industriedesigns, Herkunftsbezeichnungen und Urheberrechte.

9.2.1 Die ökonomischen Einflüsse von Patenten

Wie in der Einleitung bereits erwähnt, sind Patente von der zuständigen Behörde erteilte Schutztitel für Erfindungen [4]. Erfindungen und Patente sind also unweigerlich miteinander verbunden. Patente schützen Erfindungen und ohne Patente gäbe es nur geringen Anreiz für neue Erfindungen.

Wird eine Erfindung patentiert, erhält der Erfinder - während einer beschränkten Zeitspanne - das exklusive Recht auf die Herstellung, Benutzung und den Verkauf von Produkten, welche auf seiner Erfindung basieren. Er erhält so einen Wettbewerbsvorteil und Schutz vor den Kräften des Marktes. Auch wenn dieser Schutz relativ klein ist, reicht das doch aus, um neue Produkte zu entwickeln und diese erfolgreich zu vermarkten [1]. Der Erfinder erhält so einen Vorsprung auf die Konkurrenz. Dies heisst jedoch nicht, dass der Wettbewerb im Markt reduziert wird. Konkurrenten dürfen zwar die Erfindung nicht

kopieren, haben aber die Möglichkeit andere, kompetitive Erfindungen und Produkte patentieren zu lassen und ebenfalls auf den Markt zu bringen [1].

Patentierte Erfindungen sind nicht nur von Nutzen für den Erfinder selbst, sonder auch für die Allgemeinheit. Patentierte Erfindungen müssen offen gelegt werden und sind nach Ablauf der Patentfrist für alle anderen frei zugänglich. Dadurch werden sie zum Grundstein neuer Erfindungen.

Das Patentsystem erlaub es dem Erfinder, Profite auf drei Ebenen [1] zu erzielen.

- Über Verkauf von Produkten kann dieser die Kosten decken, welche bei der Entwicklung seiner Erfindung entstanden sind.
- Durch die Patentierung seiner Erfindung erhöht sich die Wahrscheinlichkeit, dass der Erfinder nicht nur seine Kosten decken, sondern auch Gewinne generieren kann. Dies hängt davon ab, wie gross die Steigerung der Attraktivität des Produktes durch die Erfindung ist und wieviele Substitute und Alternativen zum Produkt und der Erfindung existieren.
- Der Erfinder kann seine Erfindung jemand anderem lizenzieren oder weiterverkaufen. Dies ist für den Erfinder vor allem dann attraktiv, wenn der Lizenznehmer Märkte abdeckt, die für den Erfinder unattraktiv sind oder zu denen er selbst keinen Zugang hat. Zusätzlich kann der Erfinder auch seine Erfindung mit neuen Erfindungen kombinieren oder neue Produkte auf den Markt bringen, die auf seiner Erfindung basieren.

Wenn der Erfinder finanziell von seiner Erfindung profitieren kann, ist er motiviert diesen Prozess zu wiederholen und Teile seiner Gewinne in die Forschung und Entwicklung neuer Erfindungen und Produkte zu investieren. Er kann Leute einstellen und ausbilden, Geschäfte mit Dritten abwickeln welche wiederum - motiviert durch den finanziellen Anreiz - neue Erfindungen und Produkte hervorbringen können [1].

Der Prozess der Erfindung und Patentierung wird zum Kreis und wiederholt sich ständig (siehe Abbildung 9.1, [1]).

In der IKT-Branche gibt es verschiedene Möglichkeiten, Erfindungen durch Patente zu schützen. Die folgenden Abschnitte befassen sich mit der Patentierung von Hardware, von Software und von elektronischen Geschäftsmethoden.

Patentierung von Hardware

Die Erfindung des Mikrochips von Jack S. Kilby 1959 hatte für die Entwicklung der IKT-Branche weitreichende Auswirkungen. Die technische Entwicklung mikroelektronischer Halbleitererzeugnissen schritt in horrendem Tempo vorwärts und ist auch heute noch nicht beendet. Damals, in der Mitte des 20. Jahrhunderts, war man der Ansicht, dass der Entwurf von Mikroprozessoren keine erfinderische Leistung darstelle und somit nicht den

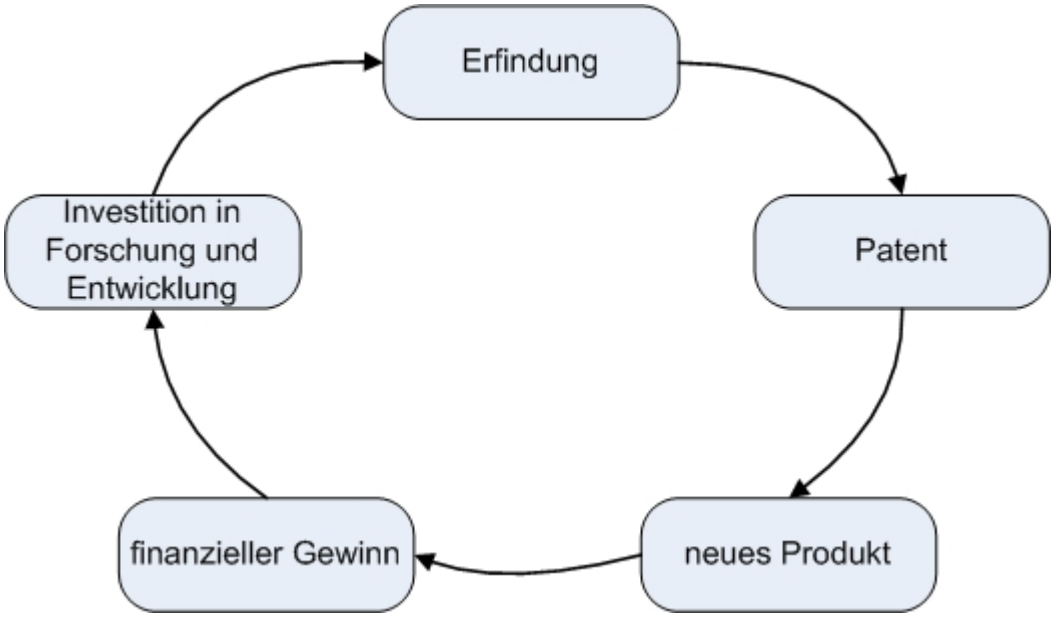


Abbildung 9.1: Kreislauf der Erfindung [1]

Schutz von Patenten verdiene. So wurde der Aufbau von Mikrochips und Mikroprozessoren von speziellen Gesetzen geschützt [1]. Auch heute wird diese Frage noch diskutiert.

In einigen Ländern gibt es solche spezielle Gesetze zum Schutz von Halbleitererzeugnissen, in anderen werden diese durch Patente geschützt. In der Schweiz z.B. sind Halbleitererzeugnisse nicht patentierbar, sondern werden vom Topographiengesetz geschützt. Topographien sind dreidimensionale Strukturen von Halbleiter-Erzeugnissen (integrierte Schaltungen bzw. Chips). Der Registereintrag einer Topografie begründet eine Schutzdauer von bis zu zehn Jahren [4].

Softwarepatente

Eine Erfindung muss im rechtlichen Sinne ein technisches Problem mit den Mitteln der Technik lösen [4]. Dabei ist das Kriterium der technischen Erfindung ausschlaggebend dafür, dass Software als solche nicht durch Patente, sondern durch Urheberrechte geschützt ist. Es existieren jedoch Formen von Software, die dieses Kriterium dennoch erfüllen und somit ebenfalls patentierbar sind. Bewirkt z.B. ein Programm in Kombination mit einer Datenverarbeitungsanlage, dass die Anlage aus technischer Sicht andersartig arbeitet, so kann die Kombination patentfähig sein, denn sie liefert einen technischen Beitrag zum Stand der Technik [5]. Zum Beispiel sind elektronische Steuerungen patentierbar [4]. Patentschutz für Software ist also nur sehr eingeschränkt erreichbar [6]. Diese Regelungen kommen in den meisten europäischen Ländern zur Anwendung.

In verschiedenen Ländern ist es jedoch möglich, Patente für „reine Software“ zu erteilen. Z.B. wurden in letzter Zeit durch das schnelle Wachstum des Internets und des elektronischen Handels vor allem in den USA sogenannte elektronische Geschäftsmethoden patentiert. Elektronische Geschäftsmethoden bestehen hauptsächlich aus softwarebasierten

Systemen und Methoden, welche die elektronischen Geschäfte im Internet vereinfachen. Die wachsende Anzahl und Komplexität von Geschäftsmodellen widerspiegelt die immer höher werdende Bedeutung von Geschäftstechnologien in der heutigen Wirtschaft [1].

Z.B. liess Amazon.com [7] ihre one-click Einkaufsmethoden patentieren. Das Patent beschreibt ein Onlinesystem, welches dem Benutzer erlaubt, die Kreditkartennummer und Adressinformation nur einmal eingeben zu müssen, damit es bei späteren Besuchen der Webseite nur einen einzigen Mausklick benötigt, um Produkte aus dem Onlineshop zu erwerben [8]. Neben Geschäftsmethoden für Online-Shopping wurden auch Methoden für Finanztransaktionssysteme oder Werbeverwaltungssysteme patentiert [1].

Ein weiteres Beispiel sind die vom Suchmaschinenbetreiber Google [9] patentierten Systeme und Methoden für die Hervorhebung von Suchresultaten (Systems and methods for highlighting search results). In der Beschreibung des Patents heisst es, dass das System für einen Suchbegriff eine Suchanfrage generiert, welche eine Liste von Verweisen zu Dokumenten im Internet returniert. Nach der Auswahl eines Verweises ruft das System ein Dokument ab, in dem der Suchbegriff hervorgehoben wird [10].

Im Sommer 2005 entschied die Europäische Union, Softwarepatente vorerst nicht zu legalisieren. Beim Thema, ob Patente auf Software erlaubt werden sollten, gehen die Meinungen weit auseinander. Im folgenden werden die wichtigsten Argumente der Befürworter und Gegner dargelegt. Die Befürworter von Softwarepatenten glauben, dass eine Harmonisierung des EU-Rechts erforderlich ist, weil verschiedene nationale Patentgesetzgebungen zu Verwirrung führen [11] und aufgrund der Internationalität des Internets das Risiko der Verletzung von nationalen Gesetzen besteht [1]. Zudem sei nach bisherigem Recht eine rein mechanische Lösung rechtlich bevorzugt gegenüber einer modernen elektronisch gesteuerten elektronisch-mechanischen Lösung, weil diese nicht patentiert werden können. Als weiterer Grund wird angefügt, dass Patente bisher auf dem Gebiet technischer Innovationen ihre Wirksamkeit bewiesen haben, die Innovation fördern und die Wirtschaft stärken [12]. Die Einführung von Softwarepatenten würde kleine und mittlere Unternehemen besser gegenüber grossen Konkurrenten abgrenzen und könnten so ihre Produkte besser vermarkten [11].

Anhänger von Softwarepatente sind der Meinung, dass in der heutigen sehr wissenslastigen Welt innovative Ideen oft die wertvollsten Ressourcen sind, um im Wettbewerb bestehen zu können. Dies gelte vor allem für Internet-Unternehmen, die kaum materielle Anlagen besitzen und deren Erfolg wesentlich von innovativen Ideen und anderen immateriellen Anlagen wie z.B. Geschäftsmodellen abhängt [1]. Dadurch, dass in einigen Ländern Geschäftsmodelle patentierbar sind und in anderen nicht, besteht aufgrund der Internationalität des Internets das Risiko der Verletzung von nationalen Gesetzen. Es ist also dringend nötig in dieser Frage eine internationale Kooperation anzustreben [1].

Die Kritiker von Softwarepatenten hingegen argumentieren, dass das Urheberrecht bisher ausreichte um Software zu schützen und die Softwarebranche sich damit sehr gut entwickelt hat. Zudem bewiesen die bisherigen Erfahrungen aus den USA, dass Softwarepatente die Wirtschaft schwächen, Innovation beeinträchtigen würden und nur zum Vorteil grosser Software- und Computergiganten seien. Es wird argumentiert, dass sich grosse Firmen zu Patent-Pools zusammentun, aus denen sich Grosskonzerne bedienen dürfen und mittelständische Firmen ausgeschlossen werden, sobald eine Konkurrenzsituation entsteht

[12]. Auch würden Patente nur angemeldet als Mittel, Wettbewerber zu bremsen. Grosse Software-Unternehmen könnten kleine Firmen, welche Patente besitzen, bis zum Bankrott bekämpfen oder diese aufkaufen. Patente für Schnittstellen und Kommunikationsprotokolle würden möglich und alle die diese verwendet wollten, müssten dem Patentinhaber Lizenzgebühren bezahlen. So wäre die Entwicklung freier Software praktisch unmöglich, weil alle „Grundlagen“ bereits patentiert worden wären [12].

9.2.2 Die ökonomischen Einflüsse der Marken

Im Folgenden wird die Bedeutung der geistigen Eigentumsrechte in Bezug auf die Marke näher betrachtet. Für die weltweit operierenden Konzerne ist die Marke ein wichtiger Aktivposten und muss dementsprechend gut geschützt werden.

Die Bedeutung der Marke als ökonomischer Aktivposten in der IKT-Branche

Marken dienen zur besseren Identifikation des Produktes unter Konkurrenzprodukten auf dem Markt. Sie erleichtern den Entscheidungsfindungsprozess der Konsumenten: die Opportunitätskosten der Konsumenten, die durch das Suchen, Finden, Vergleichen und Bewerten von Konkurrenzprodukten entstehen, werden durch Marken erheblich gesenkt. Der Konsument kauft mit der Marke eine vorfabrizierte Sicherheit, was besonders bei Erfahrungsgütern von grösster Wichtigkeit ist. Andererseits machen sie Investitionen in die Entwicklung qualitativ hochwertiger Güter für Eigner attraktiv [1]. Denn eine Marke mit qualitativ hochwertigen Produkten schafft Kundenloyalität, schützt vor der Konkurrenz und bringt den Eigner in eine attraktivere Marktposition mit höheren Gewinnen und besserer Wettbewerbsfähigkeit. Deshalb sind Marken ein entscheidender ökonomischer Aktivposten einer jeden Firma.

Unter den sechs wertvollsten globalen Marken des Jahres 2005 sind fünf aus der IKT-Branche: mit Microsoft auf Platz 2 der prominenteste Vertreter aus dieser Branche mit einem Markenwert von 59’941 Mio. Dollar für ihr Microsoft Logo, dicht gefolgt von IBM (Platz 3), GE (Platz 4), Intel (Platz 5) und Nokia (Platz 6), (Tabelle 9.2). Zu den fünf aufsteigendsten Marken des Jahres 2005 gehört mit eBay, Samsung und Apple ebenfalls die Mehrheit der Marken der Informations- und Kommunkationstechnologiebranche an [13]. Die Firma eBay allein steigerte im letzten Jahr den Marktwert ihrer Marke um 21 Prozent. (Tabelle 9.1)

Tabelle 9.1: Die sechs besten globalen Marken des Jahres 2005 [14].

	2005 Rang	Marke	2005 Markenwert (in Millionen US Dollars)
Microsoft	2	Microsoft	59941
IBM	3	IBM	53376
GE	4	GE	46996
Intel	5	Intel	46996
Nokia	6	Nokia	26452

Tabelle 9.2: Die fünf aufstrebenden globalen Marken des Jahres 2005 [14].

	Rang	Wachstum des Markenwerts im Vergleich zum Vorjahr (in Prozenten)
eBay	1	21
HSBC	2	20
Samsung	3	19
Apple	4	19
UBS	5	16

Damit wird klar, dass gerade für die IKT-Branche die Marke von steigender ökonomischer Bedeutung ist. Die Gründe dafür sind schnell ersichtlich: gerade Software und verwandte Produkte sind häufig immaterielle Güter und Dienstleistungen. Der Kunde kauft das Produkt also unter Unsicherheit und möchte natürlich sein Risiko beim Kauf minimieren. Er verlässt sich somit stärker auf die Marke als in anderen Branchen. Ausserdem sind Standards sehr wichtig: die Gewinne, die aus Netzwerkexternalitäten resultieren, sind in dieser Branche enorm, da der Zusatznutzen überproportional steigt, je mehr Leute das Produkt kaufen. Dieser Zusatznutzen resultiert zum Beispiel aus verbesserter Interaktionsmöglichkeit und/oder einem besseren Angebot an Komplementen. So entsteht eine installierte Basis, welcher Firmen, die neu in den Markt eintreten, nichts entgegenzusetzen wissen (Intel [15], Microsoft [16] und Sun Microsystems [17] verfolgen diese Strategie seit Jahren mit grossem Erfolg) [18]. Ein gutes Beispiel für die Registrierung und den daraus entstehenden Wert einer Marke ist sicher das Unternehmen Intel Inc.: Intel produzierte die sehr erfolgreiche Reihe von X86 Mikrochips, doch Intel liess diesen Namen nicht schützen und bald begannen die Firmen AMD [19], Chips and Technologies (diese Firma wurde später von Intel Inc. übernommen) und Cyrix (1998 von der Firma National Semiconductor übernommen [20]) ebenfalls den Namen X86 zu benutzen. 1991 realisierte Intel ihren Fehler und lancierte eine neue Generation von Computerchips mit dem geschützen Intel inside Logo. Um das Logo zu verbreiten, gewährte Intel den Absatzmittlern einen Rabatt von drei Prozent auf ihre Prozessoren, wenn diese als Gegenleistung das Intel inside Logo auf ihren Verpackungen gut sichtbar platzierten. In einer Periode von 18 Monaten erschien das Intel inside Logo in über 90000 Werbeseiten, die Erkennungsrate bei den Endnutzern stieg von 46 auf 80 Prozent. Ein Jahr nach der Lancierung der Intel inside Kampagne, waren die Verkäufe von Intel Inc. global um 62 Prozent gestiegen. Die prominente Vertretung des Logos hatte die Komsumenten beeinflusst, so dass sie dachten, der Intel Mikrochip sei besonders gut und erstrebenswert [21].

Die Veränderungen in der Handhabung des Markenmanagements

Durch das Internet und die steigende Komplexität der Produkte in der IKT-Branche entstehen neue Trends. Einerseits beeinflusst die Marke mit den neu geschaffenen Möglichkeiten die Martketingstategie der Firmen enorm, andererseits wird sie aber selbst durch die neuen Arten der Präsentation der Produkte über das Internet beeinflusst.

1. Die steigende Komplexität der Produkte durch die Globalisierung und deren Konsequenzen in der IKT-Branche

Hardware und Software werden immer komplexer und vielschichtiger. Die Marke gewinnt dadurch mehr und mehr an Bedeutung, da die Kunden unmöglich alle Teile und Lieferanten überblicken können, aus denen ein komplettes Gerät zusammengesetzt wird. Gerade die heutigen Computer sind eine Ansammlung verschiedenster Teile von unterschiedlichen Lieferanten, die global auf allen sechs Kontinenten, manchmal sogar mit anderen Namen und Vertriebswegen, operieren. Mit dem Zunehmen der Komplexität werden aber nicht nur Marken wichtiger, sondern auch Zertifizierungen und Gütesiegel, um dem Kunden zu signalisieren, dass das Produkt gewisse Standardtest erfolgreich bestanden hat. Das Einhalten technischer Standards wird häufig durch ein Logo auf der Packung signalisiert: ein Beispiel wäre der ENERGY STAR, der ein Gütesiegel für wenig Stromverbrauch darstellt (Bild: 9.2) [22].

Abbildung 9.2: Der ENERGY STAR als Gütesiegel

Apple Computer [23] registrierte die Marke ColorSync, die mit Software und Hardware vertrieben wird, um zu zeigen, dass die Produkte eine proprietäre Software von Apple benutzen, die die Farben korrekt aufeinander abstimmt [1]. DOLBY DIGITAL [24] ist ein weiteres Beispiel im Bereich der digitalen Musik, das beweist, wie etabliert diese neuen Gütesiegel in der IKT-Branche sind. Das komplizierte Konzept des Erfüllens einer technischen Spezifikation wird durch diese neuen Logos für den Konsumenten einfach verständlich, visuell schnell erfassbar und attraktiv dargestellt. Heutzutage kreiert man mit der Marke nicht nur ein aussergewöhnliches Produkt, sondern tendiert mehr und mehr dazu, Produktefamilien von technologisch verwandten Produkten zu schaffen, häufig sogar unter an sich unabhängigen und selbständigen Firmen. Eine besonders erfolgreiche Strategie verfolgt die Firma Microsoft derzeit mit ihrer Flagge als Logo. Das vielfarbige, eingängige Design ist auch auf Komplementen (zum Beispiel in der Spieleindustrie) gut ersichtlich und garantiert damit die reibungslose Konnektivität verwandter Produkte. Die Zertifizierungen und Gütesiegel dienen einerseits dazu, dem Konsumenten von der Industrie akzeptierte Produkte zu liefern und andererseits trotzdem zu vermitteln, das Produkt sei innovativ und etwas vollkommen Neues [1].

2. Das Aufkommen des Internets, Cyberbesetzung und schärfere Gesetze zum Schutz des geistigen Eigentums

   Wie bereits erwähnt, ist die Immaterialität von Produkten ein entscheidender Grund, weshalb Marken in der IKT-Branche eine wichtige Rolle spielen. Hält der Kunde beim Softwarekauf wenigstens noch die CD in der Hand, sind die neuen Formen der Dienstleistungen, die übers Internet angeboten werden, nun vollkommen virtuell. Deshalb werden Marken noch essentieller im Zusammenhang mit dem globalen e-commerce. Die globale Vermarktung eines Produktes bring viele Konsequenzen mit sich: Die Marke wird mehr und mehr zu einer kulturellen Ikone und vermittelt eher einen Lebensstil, als den effektiven Nutzen eines Produktes. Firmen beginnen mehr Geld mit ihren Marken, denn mit ihren Produkten zu generieren, da gerade in der betrachteten Branche die Qualitäten und Spezifikationen der Produkte schwer zu

vermitteln sind [1]. Auch aufgrund von technisch sehr ähnlichen oder sogar gleichen Leistungsmerkmalen setzen die Unternehmen mehr und mehr auf Emotionen, innere Bilder und Erwartungen der Konsumenten in der Gestaltung und in der Verbreitung ihrer Marken. Ein gutes Beispiel für diesen Trend ist die Firma Samsung: mit einer Erhöhung des Markenwerts um 19 Prozent allein im Geschäftsjahr 2005 und mit einer Steigerung von insgesamt 186 Prozent über die letzten fünf Jahre ist sie eine der erfolgreichsten aller Marken [13]. Gerade hat Samsung eine neue Markenkampagne realisiert: auf der Website der Firma wird der Kunde nun mit viel Farben, Licht und Musik, sowie kleinen Kurzfilmen mit der Marke vertraut gemacht. Mit ihrem Slogan „With Samsung, it's not so hard to imagine", dem digital freedom Logo mit dem Surfer oder der prominenten Vertretung des Markenlogos auf dem Dress der englischen Fussballspieler von Chelsea hat Samsung weltweit grossen Erfolg [25] [26]. Die Nachricht ist klar: das Produkt wird um den Kunden gestaltet, der Kunde steht im Mittelpunkt. Die Marke will jung, sportlich und attraktiv sein. Man will dem Kunden alle Freiheiten geben, so dass sich dieser frei entfalten kann. Das Leben, zum Beispiel mit Samsung-Produkten, wird einfacher, besser und interessanter als jemals zuvor. So wird der Besuch der Internetseite zu einem multisensualen Erlebnis, das ureigene Bedürfnisse der Kunden stimuliert, über die eigentlichen Produkte aber wenig aussagt. Damit nimmt die Bedeutung der Internetseiten enorm zu. Heutzutage ist eine global operierende Firma ohne Website unvorstellbar. Ihre Gestaltung ist ein wichtiger Teil der Marketingstrategie. Daraus folgt eine weitere interessante und vollkommen neue Entwicklung: die Beziehung zwischen Marke und Domainname wird enorm wichtig. Um die Marke im Internet eindeutig finden zu können, muss der Domainname ebenfalls eindeutig zugeordnet werden können. Das bringt verschiedene Probleme mit sich: findige Benutzer und Kleingeschäfte registrierten Domainnamen mit dem Markennamen grosser Firmen mit dem Zweck, sie nachher dem eigentlichen Unternehmen gewinnbringend verkaufen zu können [27]. Die sogenannte Cyberbesetzung [28] griff 1999 rasch um sich und ein neuer Schutz für Marken in Verbindung mit Domainnamen musste schnell etabliert werden. Ein prominentes Beispiel ist eine Privatperson, die sich am 23. September 1998 den Domainnamen www.pentium3.com sicherte, pornographische Fotos darauf veröffentlichte und verkündete, er verkaufe die Seite an den Höchstbietenden. Das höchste Angebot lag bei 9,350.12 USD, als die Intel Corporation schliesslich einschritt und Anklage erhob [29]. Die ICANN (Internet Corporation For Assigned Names and Numbers) erliess deshalb im Oktober 1999 die UDRP Resolution, zum besseren Schutz der Marken in Kombination mit den zugehörigen Domainnamen. Neu müssen nun Personen, die einen Domainnamen, der gleich oder mit hoher Ähnlichkeit zu einer geschützen Marke ist, nur mit der klaren Absicht des Weiterverkaufs oder in anderer kommerzieller Absicht erstehen, den Domainnamen an die berechtigte Firma ohne Entgeld abgeben. In besonders gravierenden Fällen wird auch eine Busse verlangt [27].

Die Revision des Markenschutzgesetzes in der Schweiz

Im Januar 1991 wurde das Markenschutzgesetz (MSchG) revidiert [30]: neu wurde es den Programmierern erlaubt, nicht nur Computerprogramme, sondern auch Informatikdienstleistungen (Schulung, Wartung und Erstellung) zu schützen. Firmen können das Logo so-

wohl auf der Packung einer Software anbringen, als auch im Programm selbst, so dass es bei dessen Benutzung erscheint. Damit lässt die Schweiz den Markenschutz im Zusammenhang mit Waren und Dienstleistungen zu [31]. Verschiedene Schutzmechanismen wurden weiterentwickelt und etabliert. Somit wurde auch in der Schweiz die Sonderstellung der Marke in Bezug auf die immateriellen Güter der IKT-Branche gesetzlich anerkannt.

9.2.3 Die ökonomischen Einflüsse des Industriedesigns

Das Industriedesign kann, wie in der Einleitung bereits erwähnt, ein Muster, eine Form oder eine Oberfläche eines Produktes sein. Das umfasst zum Beispiel sowohl die Verpackung einer Software, als auch die Gestaltung einer Spielkonsole oder eines Mobiltelefons. Es werden mit der Form und Oberfläche, sowie den Linien, Mustern und Farben drei- und zweidimensionale Aspekte der Produkte berücksichtigt. Unerlaubtes Kopieren oder Imitieren der so geschützten Produkte durch Dritte ist nicht gestattet. Die Bedingung, um eine Gestaltung eines Produktes schützen zu können, ist, dass diese neu und originell sein muss [32].

Da die Leistungsmerkmale der Produkte in der IKT-Branche immer ähnlicher werden, ist die Gestaltung, neben der Marke, ein Differenzierungsmerkmal für den Kunden auf dem Markt und damit ein entscheidendes Marketinginstrument, das in den Marketing Mix einer Unternehmung integriert werden muss. Die Produkte müssen sich auf den ersten Blick unterscheiden und dem Kunden einen attraktiven Trend signalisieren. Durch neuartige Materialien, CAD Programme (Computer Aided Design) und innovative Produktionstechniken sind heutzutage bei der tatsächlichen Produktion in der IKT weniger Schranken gesetzt als noch vor ein paar Jahren. Gerade bei Hardware wird ein starker Trend zur Miniaturisierung festgestellt, was den Gestaltern immer mehr Raum zum Experimentieren bietet. Bei Luxusprodukten, wie der moderne Laptop eines ist, wird neu auch das Gewicht des Endproduktes in die Gestaltung und Produktion einbezogen. Auch die Verpackung gewinnt an Bedeutung [33]. Ebenso spielt das Industriedesign bei der Einführung neu entwickelter Technologien und zur Durchsetzung neuer Standards eine grosse Rolle: die bekannten Abspielgeräte für digitale MP3-Musik haben sich auch deshalb so schnell durchgesetzt, weil sie durch ihre Grösse und ihr gut durchdachtes Design in jede Hosentasche passen, sowie schnell und intuitiv zu bedienen sind. Das sogenannte menschenfreundliche Design muss wesentliche Dinge sofort sichtbar machen [33]: Wo befindet sich zum Beispiel der Abspielknopf auf dem MP3 Gerät und wo ist der Batterieschacht zu finden? Auch müssen die richtigen Botschaften übermittelt werden, um somit die Benutzerfreundlichkeit eines Produktes deutlich zu steigern: signalisiert durch einen Pfeil muss beim Öffnen des Batterieschachtes des Gerätes zuerst der Deckel in eine bestimme Richtung verschoben werden, damit der Kunde die Batterie nachher richtig plazieren kann.

Ein gutes Beispiel, wie wichtig Industriedesign heutzutage geworden ist, ist die Firma Apple. Im Jahre 1990 konnten die Kunden einen Apple Computer immer weniger von einem Computer aus dem Hause IBM unterscheiden. Als die Firma Microsoft dann auch noch mit ihren Computern das Betriebssystem Windows 3.0 mit der Look and Feel-Benutzeroberfläche auslieferte, fürchtete Apple um ihre Existenz und die Unternehmung

stürzte in eine tiefe wirtschaftliche Krise, die beinahe im Bankrott geendet hätte [34]. 15 Jahre später gehört Apple unter anderem auch dank innovativem und attraktivem Industriedesign ihrer Produkte wieder zu den Top Unternehmen in der IKT-Branche. Mit ihrer ganz in weiss gehaltenen iBook- Linie oder dem edlen, in Silber gehaltenen Powerbook, gehören Apple Computer zu den Produkten, die Kunden am schnellsten als solche identifizieren können. Mit der Gestaltung neuer Produkte, wie zum Beispiel dem iPod, schafft es die Firma, ein Image aufzubauen, das weltweit einzigartig ist [35]. Ausdruck findet die Wichtigkeit dieser Sparte auch in den jährlich vergebenen Preisen für gelungenes Industriedesign: Im Jahre 2003 erhielt Apple von der IDEA (Industrial Design Excellence Awards) in der Sparte Computerprodukte die Goldmedaille verliehen, als Unternehmen mit dem besten Industriedesign. Ein Jahr zuvor holte sich Apple ebenfalls Gold, und das nicht nur für die Gestaltung ihrer Computerprodukte, sondern auch für das beste Team von Designern in einer Firma weltweit und gleich zwei ihrer Produkte, das iBook und der iPod, wurden zu den innovativsten Produkten des Jahres 2002 gekürt [36]. (Bild: 9.3)

Abbildung 9.3: Der Apple iPod: Ein mit dem Industriedesignpreis ausgezeichnetes Produkt [36].

Die zunehmende Verbreitung des Internets und die fortschreitende Globalisierung stellen aber auch den Schutz des Industriedesign vor neue Herausforderungen. Wie alle Rechte des geistigen Eigentums wird das Industriedesign primär nur national geschützt. Um die einzigartige Gestaltung eines Produktes auch weltweit schützen zu können, wurde in der Haager Übereinkunft (seit 2004 in Kraft) vereinbart, dass man neu Industriedesign auch international schützen kann: Ein Produkt kann nun durch Abgabe eines Formulars bei der WIPO, mit minimalem administrativem Aufwand, in mehreren Staaten gleichzeitig geschützt werden [37].

9.2.4 Herkunftsbezeichnung und deren Rolle in der IKT-Branche

Die Herkunftsbezeichnung eines Produktes ist, wie auch eine Marke oder ein Design, ein Differenzierungsmerkmal für Kunden auf dem Markt. Die Identifikation der territorialen Herkunft kann unter Umständen als ein Zeichen für Qualität, gute Verarbeitung und technisches Know How stehen [4]. Herkunftsbezeichnungen sind Hinweise auf die geographische und nicht unternehmensmässige Abstammung eines Produktes. Sie spielen vor allem in der Herstellung und dem Vertieb von Lebensmitteln eine Rolle, in der der Ort der Produktion ein entscheidender Faktor für die Güte eines Produktes darstellt

(zum Beispiel die Qualität des Bodens oder der Milch, sowie geltende Produktions- und Umweltbestimmungen). In der IKT-Branche sind solche Faktoren unbedeutend, deshalb nehmen die Herkunftsbezeichnungen auch eine untergeordnete Rolle ein und werden hier nur der Vollständigkeit halber am Rande erwähnt. Obwohl das Silicon Valley, Berkeley oder Redmond wichtige Standorte darstellen, dienen sie nicht zur Bezeichnung der von dort stammenden Produkte. Da es also bis jetzt kein geographisches Gebiet gibt, das zur Herkunftsbezeichnung eines Computers, einer Software oder eines sonstigen Produktes in der IKT-Branche dient, ist ihre Bedeutung und ihre Auswirkungen im Zeitalter des Internets in der betrachteten Branche auch dementsprechend gering und wird hier nicht mehr weiter behandelt.

9.2.5 Die ökonomischen Einflüsse des Urheberrechts und der verwandten Schutzrechte

Das Urheberrecht betrifft nicht nur den Urheber, sondern auch die IKT-Branche und die Benutzer. Im Folgenden geht es um die Vor- und Nachteile des Urheberrechts und darum, wie das Urheberrecht die betroffenen Gruppen beeinflusst.

Die Einflüsse des Urheberrechts auf die IKT-Branche

Das Urheberrecht war ursprünglich zum Schutz von Kunst- und Literaturwerken gedacht. Seit dem Aufkommen von Computerprogrammen schützt das Urheberrecht auch diese. Es waren jedoch zusätzliche Normen zur Anpassung der bisherigen Regelungen nötig [38]. Zum Beispiel ist es grundsätzlich erlaubt, urheberrechtlich geschützte Werke zu vermieten, für Computerprogramme gilt diese Regelung aber nicht. (Art. 13 URG) Weitere Unterschiede sind in Tabelle 9.3 aufgelistet.

Zusätzlich schützt das Urheberrecht Datenbanken. Diese können jedoch nur als Sammelwerk geschützt werden, wenn ihre Auswahl oder Anordnung individuellen Charakter hat [38].

Der Inhaber des Urheberrechts hat das Recht auf Erstveröffentlichung. Bei Computerprogrammen gilt die Veröffentlichung des Sourcecodes als Erstveröffentlichung. Aus dem Urheberrecht leiten sich direkt die Urhebervermögensrechte ab, welche dem Inhaber erlauben, aus seinem Werk finanziellen Nutzen zu ziehen. Im Gegensatz zum Patent ist das Urheberrecht nicht übertragbar, es ist fest an eine Person gebunden [38].

Wie schon gesagt, schützt das Urheberrecht Kunstwerke, dazu gehören auch Bücher, Musikstücke und Filme. Die Verbreitung des Internets, welche die Möglichkeit brachte, uneingeschränkt Daten auszutauschen, vereinfachte den Austausch dieser erheblich. Dies wird dadurch ermöglicht, dass die Daten digital vorhanden sind, es verschiedene Möglichkeiten gibt, Daten zu kopieren (CDs rippen, DVDs kopieren), die Technologien zum Austausch (z.B. Peer-to-peer) laufend entwickelt werden und die Hardware dazu (Computer, CD/DVD-Brenner) mittlerweile leicht zugänglich ist. Das heisst, die Verstösse gegen das Urheberrecht haben mit dem Internet massiv zugenommen. 90% der Benutzer haben

Tabelle 9.3: Unterschiede des Schutzes von Computerprogrammen und sonstigen Werken [38]

	Computerprogramme	sonstige Werke
Vermietung durch Inhaber des Werkexemplars	nein (Art. 10 Abs. 3 URG)	ja (Art. 13 Abs. 1-3 URG)
Erschöpfung der Urheberrechte an rechtmässig in Verkehr gebrachten Werkexemplaren	Gebrauchs- und Veräusserungsrecht des Erwerbers (Art. 12 Abs. 2 URG/Art. 17 URV)	Erschöpfung der Rechte am Werkexemplar (Art. 12 Abs. 1 URG)
Kopieren zum Privatgebrauch	nein (Art. 19 Abs 4 URG)	ja (Art. 19 Abs. 1-3 URG)
Reverse Engineering	nur Schnittstellen für interoperable Programme (Art. 21 URG / Art 17 URV)	ja (falls überhaupt möglich)
Sicherungskopien	ja (Art. 24 Abs. 2 URG)	ja (Art. 24 Abs.1 URG)
Schaffung im Arbeitsverhältnis	Spezialregelung (Art. 17 URG)	Zweckübertragungstheorie (Art. 16 Abs. 2 URG)
Schutzdauer	50 Jahre über den Tod hinaus (Art. 29 Abs.2 lit. a URG)	70 Jahre über den Tod hinaus (Art. 29 Abs.2 lit.b URG)

mindestens ein Filesharing-Programm auf ihrem Computer, das sie zum Tauschen von Musik, Filmen, Spielen, Bücher oder Bilder verwenden. Laut Tagesanzeiger vom 9. November 2005 werden monatlich ungefähr 2 Milliarden Dateien ausgetauscht, von denen 90% gegen das Urheberrecht verstossen [39]. Somit sind intensivere Massnahmen zum Schutz dieser Rechte erforderlich geworden. Es sind also neue Aufgaben für die IKT-Branche entstanden. Aufgaben, welche durch die Fortschritte der Branche selbst nötig wurden [40]. Neue Technologien und Fortschritt bringen neue Möglichkeiten, aber auch potentiell neue Probleme. Welche Probleme es sind und wie man sie lösen will, wird im weiteren Verlauf dieser Arbeit erörtert.

Die ökonomischen Einflüsse des Urheberrechts auf die Distribution digital speicherbarer Medien auf dem Markt

Zu den meist verbreiteten digital speicherbaren Medien gehören Bücher, Musik und Filme. Bei den meisten Problemen mit dem Urheberrecht geht es heutzutage um Musik. Obwohl dies eindeutig gegen das Urheberrecht verstösst, ist der Austausch von Musik über das Internet weit verbreitet. Das bedeutet, dass vor allem die Musikindustrie daran interessiert ist, den Austausch von Musik einzuschränken. 1999 kam eines der ersten Peer-to-peer Netzwerke Napster auf, das schnell viele Benutzer anzog. Schon bald wurde Napster aber von der Musikindustrie wegen Verletzung des Urheberrechts verklagt und musste wenig später seinen Dienst einstellen [40]. Trotz dieser Niederlage wurden weitere Peer-to-peer Netzwerke aufgebaut, mit Technologien, die es schwieriger machen die Verantwortlichen zu identifizieren. Die Unterhaltungsindustrie versucht dennoch Wege zu

finden, die Verantwortlichen zum Zahlen zu zwingen. Der illegale Austausch von Medien über das Internet bringt die Unterhaltungsindustrie um einen grossen Teil ihres Gewinns. Das Urheberrecht verbietet zwar das Kopieren und Verbreiten von Medien ausserhalb des Freundeskreises (URG, Art. 19), es hindert jedoch nicht daran, es zu tun. Wie schon gesagt, ist es heutzutage sehr simpel z.B. eine CD zu kopieren und sie weiterzugeben oder sogar weiterzuverkaufen. Deshalb setzt sich die Industrie für die Entwicklung digitaler Schutzmechanismen für Medien ein. Bemühungen, einen solchen Schutz zu entwickeln, fallen unter den Begriff Digital Rights Management (DRM). Hier zeigt sich jedoch eine weitere Problematik. Nämlich die Frage, wie weit Schutz nur Schutz ist und wann die Grenze zur Einschränkung des Benutzers überschritten wird. Diese Frage wird im Teil 9.4.2 dieser Arbeit besprochen.

Bis jetzt wurde die digitale Distribution von Medien nur von der negativen Seite betrachtet. Die neuen Technologien bringen der Unterhaltungsindustrie zwar Probleme und Gewinneinbussen, sie bringen ihnen aber auch neue Vertriebs - und Verteilungswege. Einerseits können die altbekannten Bücher, CDs, DVDs usw. über das Internet in Onlineshops verkauft und direkt zum Kunden versandt werden, so wie Amazon das tut. Andereseits kann das illegale Herunterladen von digitalen Medien legal gemacht werden, indem man Gebühren verlangt [41]. Das wohl populärste Beispiel eines Musik E-Shops ist der iTunes Music Store. Über die Software iTunes von Apple kann man auf den Internetshop zugreifen und gegenwärtig gegen eine Gebühr von CHF 1.50 pro Musikstück oder CHF 15.00 pro CD, die gewünschten Dateien herunterladen. Dass dieses Angebot rege genutzt wird, zeigt die Tatsache, dass alleine im ersten Jahr 50 Millionen Songs über den iTunes Music Store verkauft wurden. Am 15. Juli 2005 war die Zahl der verkauften Musikstücke auf 500 Millionen angestiegen [42]. Seit kurzem gibt es im iTunes Music Store auch Filme zu kaufen. Innerhalb der ersten 20 Tage wurden eine Million Videos verkauft [43]. Digitalisierung von Medien bedeutet also nicht nur eine Zunahme der Piraterie, sondern auch eine grössere Zugänglichkeit der Werke zu verschiedenen Preisen. Ein Beispiel soll die Vielseitigkeit und Vorteile der Digitalisierung von Medien veranschaulichen. Das Buch „Death on the Nile“ von Agatha Christie ist in verschiedenen Formen erhältlich:

- Das gebundene Buch ist für CHF 51.65 in der Buchhandlung erhältlich.
- Das Taschenbuch ist ebenfalls in der Buchhandlung erhältlich und kostet CHF 26.35.
- Im Onlineshop wird das Taschenbuch für CHF 19.50 zuzüglich eines Portos von CHF 6 zum Kunden nach Hause geliefert.
- Das E-Book, also die digitale Version des Buches, kostet CHF 15.20.

Es werden also die Bedürfnisse mehrerer Kunden befriedigt. Der Inhalt bleibt derselbe, nur die Form ändert sich. Als Folge davon erreicht das Medium mehr Kunden und wird mehr verkauft. Das bedeutet mehr Einnahmen für die Unternehmen.

Das Urheberrecht aus der Sicht der Urheber

Im letzten Abschnitt wurde gezeigt, wie die modernen Informations- und Kommunikationstechnologien die Verbreitung von Werken fördern. Von Interesse ist diese Tatsache

nicht nur für die Industrie, sondern auch, wenn nicht vor allem, für die Künstler. Ihr Ziel ist es, so viele Leute wie möglich zu erreichen, möglichst bekannt zu werden. Dafür sind im heutigen Zeitalter alle technischen Hilfsmittel vorhanden. Es ergeben sich nicht nur Vorteile für Künstler, die schon eine Karriere gestartet haben, sondern besonders für jene, die noch unbekannt sind. Jeder kann den eigenen Film, das selbst komponierte Musikstück oder seinen Roman aufs Internet stellen. Auf diese Weise ergibt sich die Möglichkeit, Millionen von Menschen direkt zu erreichen [41]. Da der Vertrieb der Medien stark vereinfacht wurde, ist es zudem einfacher, die eigenen Werke direkt zu verkaufen, ohne Zwischenhändler. Ein Autor muss sich zum Beispiel nicht vertraglich an einen Verlag binden und sein Werk an die Wünsche des Verlags anpassen. Die künstlerische Freiheit ist somit gewährleistet. Dem Urheber bringt der direkte Vertrieb jedoch auch finanzielle Vorteile. Sein Ertrag wird weniger durch Abgaben an die Vertriebsfirma gemindert. Diese Sichtweise ist allerdings ziemlich eng, da sie die Produzenten als reinen Kostenfaktor des Vertriebs betrachtet. Sie haben jedoch auch andere Aufgaben, wie Marketing und Organisation, die nicht so einfach zu ersetzen sind und nicht immer vom Urheber des Werkes übernommen werden können.

Das Urheberrecht aus der Sicht der Benutzer

Sowohl die Probleme der Industrie mit dem Urheberrecht als auch die Vorteile für die Urheber sind nun betrachtet worden. Wie steht es um die Benutzer und um das Urheberrecht? Es sind die Benutzer, die meistens gegen das Urheberrecht verstossen. Benutzer, die ein Medium legal erworben haben, wollen damit tun, was sie wünschen, also auch Kopien machen. Für den privaten Gebrauch ist das erlaubt. Das Urheberrecht verbietet es jedoch, diese Kopien zu veröffentlichen. Wie schon erwähnt hindert das Urheberrecht die Benutzer nicht daran, digital speicherbare Medien übers Internet auszutauschen [40]. Es ist verständlich, dass der Kunde lieber ein Musikstück kostenlos herunterlädt, als dass er für 30 Franken die CD kauft. Der Preisunterschied zwischen herunterladen und kaufen ist zu gross, sogar wenn man die Internetkosten des Benutzers einbezieht. Ein weiterer Punkt, den man aus der Sicht des Benutzers betrachten muss ist das DRM. Zwar trägt dieses zum Schutz des Urheberrechts bei, der Käufer empfindet dies aber oft als Einschränkung. Beim Erwerb eines Mediums verlangt der Käufer die uneingeschränkte Verwendung. Diese ist beim Einsatz von Technologien des DRM, jedoch nicht immer gewährleistet [40]. Zum Beispiel ist es nicht immer möglich eine CD auf den Computer zu kopieren. Dies trägt zur Verhinderung von Piraterie bei. Was aber, wenn der Käufer die CD auf seinen MP3-Spieler laden will? Zuerst muss er die Musik ja auf seinen Computer speichern. In dem Fall fühlt sich der Benutzer eingeschränkt und das DRM wirkt kontraproduktiv. Diese Problematik wird im Abschnitt „Digital Rights Management: Schutz oder Einschränkung?“ wieder aufgegriffen.

9.2.6 Die Vor- und Nachteile der Rechte des geistigen Eigentums für unterschiedliche Anspruchsgruppen

Aus den vorhergehenden Ausführungen wird ersichtlich, dass Patente, Marken, Industriedesigns und Urheberrechte die IKT-Branche auf verschiedene Weisen beeinflussen. Im

folgenden Abschnitt werden nun die damit einhergehenden Auswirkungen für verschiedene Anspruchsgruppen näher betrachtet: Nicht nur für die Erfinder und Firmen, sondern auch für den Staat, die Gesellschaft und nicht zuletzt für die Kunden und Konsumenten bringen die Rechte des geistigen Eigentums Vor- aber auch Nachteile mit sich.

1. Die Firmen als privatwirtschaftliche Anspruchsgruppe

   Wie in den vorhergehenden Abschnitten bereits erwähnt wurde, dienen die Rechte des geistigen Eigentums den Firmen der IKT-Branche zum Schutz ihrer Erfindungen, Marken und Werke ihrer angestellten Künstler. Nicht nur wird mit der effizienten Vermarktung der Marken und des Industriedesigns erheblichen Mehrwert geschaffen, sondern es entstehen durch lukrative Lizenzverträge mit anderen Firmen auch neue Einnahmequellen. Durch die Möglichkeit, neu auch Geschäftsideen patentieren zu können, entstehen zudem Firmen, die sich in dem Bereich des e-commerce etablieren können und so die entstandene Marktnische für sich nutzen. Aus Sicht der Firmen ist also der Schutz der Rechte des geistigen Eigentums eine profitable Sache, aber auch die Nachteile dürfen nicht vernachlässigt werden: so ist heute noch, unter anderem wegen dem Territorialprinzip, die Patentierung neuer Erfindungen eine langwierige und rechtlich komplexe Angelegenheit. Der grössere administrative Aufwand, sowie die manchmal zusätzlich nötige rechtliche Abklärung der Verhältnisse schafft neue Kosten, für die die Firma aufkommen muss. Zudem werden durch die zunehmende Globalisierung die Produktlebenszyklen immer kürzer. Das heisst natürlich auch, dass eine Firma ihr Produkt möglichst schnell auf den Markt bringen will, um der Konkurrenz zuvorzukommen und somit wettbewerbsfähig zu bleiben. Dieses schnelle Vorgehen wird aber durch den Prozess der Patentierung verzögert und behindert.

2. Der Staat und die Gesellschaft als öffentliche Anspruchsgruppe

   Dadurch, dass für Erfindungen Patente beantragt werden können, erhält deren Erfinder einen temporären Schutz, um seine Erfindung zu vermarkten. In diesem Prozess tätigt der Erfinder neue Investitionen und schafft neue Arbeitsplätze. Der Wohlstand einer Gesellschaft steigt und der Staat kommt in den Genuss von höheren Steuerbeträgen. Indem das Wissen von Erfindungen nach der Schutzdauer jedem frei zur Verfügung steht, werden Erfindungen zur Grundlage für neue Erfindungen. Um diesen Fortschritt zu gewährleisten, führt der Staat ein öffentliches Patentverzeichnis. Fortschritt bedeutet auch, dass das Alltagsleben von Menschen einer Gesellschaft vereinfacht und deren Lebensniveau erhöht wird. Der Schutz von geistigen Eigentumsrechten kann auch negative Folgen haben. Eine Firma kann durch geistige Eigentumsrechte eine gewisse Marktmacht aufbauen und diese unrechtmässig nutzen, z.B. um ihre Monopolstellung im Markt gegenüber der Konkurrenz zu verteidigen. Monopole stellen meistens einen Wohlfahrtsverlust dar. Monopolhalter können durch Monopole zwar einen höheren Gewinn abschöpfen, für die Gesellschaft als Ganzes produzieren Monopole jedoch keinen Gewinn. Es entsteht ein Wohlfahrtsverlust.

3. Die Erfinder als individuelle und private Anspruchsgruppe

   Erfinder und Urheber erhalten durch die geistigen Eigentumsrechte die Möglichkeit, aus ihren Erfindungen und Werken wirtschaftlichen Profit zu schlagen. Eine weitere

Möglichkeit ist, ein Patent zu lizenzieren oder weiterzuverkauften. Für individuelle, private Halter von geistigen Eigentumsrechten kann der Aufwand für Administration und eventuelle rechtliche Abklärungen oder gar Gerichtsprozesse schnell ansteigen und untragbar werden. Würde Software in Zukunft patentierbar, würden für Programmierer unüberwindbare Hindernisse beim Entwicklen von Anwendungen entstehen. Patentgebühren müssten bezahlt werden, wenn patentierte Programmteile weiterverwendet würden.

4. Die Kunden und Konsumenten als individuelle und allgemeine Anspruchsgruppe

   Einige Rechte des geistigen Eigentums, wie zum Beispiel der Schutz der Marke, ist auch für die Konsumenten von grosser Bedeutung. Die Marke ist ein Merkmal für die Qualität und Herkunft der Produkte und ermöglicht dem Kunden eine eindeutige Identifizierung und Klassifizierung vozunehmen. Dieses Differenzierungsmerkmal erleichtert die Kaufentscheidung der Konsumenten und nimmt ihnen so das mühsame und aufwändige Vergleichen von Produkten mit ähnlichen Leistungsmerkmalen ab. Zusätzlich profitieren die Kunden von neuen Erfindungen, die wegen dem Schutz der Rechte des geistigen Eigentums, frei zugänglich werden. Durch kompetitive Firmen, die Konkurrenzprodukte auf den Markt bringen, entsteht zusätzlicher Wettbewerb, der die Ausgereiftheit des Produktes steigert und letztendlich dem Konsumenten nützt. Doch durch den von Firmen vorgenommenen Schutz und die Patentierung von zum Beispiel neuen proprietären Standards leidet die Konnektivität von Interaktionsprodukten enorm. Die Limitierungen durch Standards verunmöglichen es manchmal den Konsumenten, Einzelteile verschiedener Firmen zu kombinieren. Ein gutes Beispiel dafür sind die prorietären Standards von Sony [54]. Ein Konsument, der eine digitale Kamera von Sony erwirbt, muss die proprietären Sony Memorysticks als Speichermedium verwenden. Eine Alternative, wie zum Beispiel ein Memorystick der Konkurrenzfirma Canon, kann nicht verwendet werden, da durch den proprietären Standard die Konnektivität dieser Interaktionsprodukte verunmöglicht wird. Das stellt für den Kunden sicherlich einen erheblichen Nachteil dar.

9.3 Die neuen Problembereiche für die geistigen Eigentumsrechte mit dem Aufkommmen neuer Technologien

Dieses Kapitel zeigt exemplarisch zwei Probleme auf, welche mit dem Aufkommen von neuen Technologien im Zusammenhang mit den geistigen Eigentumsrechten entstanden sind. Zum einen das Territorialprinzip, welches mit der Entstehung und weltweiten Verbreitung des Internets an Bedeutung zugenommen hat. Zum anderen die Piraterie, welche durch die leichte Erhältlichkeit von Kopier-Hardware und ebenfalls durch das Internet eine neue Dimension erreicht hat.

9.3.1 Das Territorialprinzip im Internetzeitalter

Auch wenn geistige Eigentumsrechte ein wirksames Instrument zum Schutz von Erfindungen, Marken, Literatur- oder Kunstwerken sind, ist deren Wirksamkeit begrenzt. Nach dem sogenannten Territorialprinzip sind geistige Eigentumsrechte, welche angemeldet werden müssen, um sie in Anspruch zu nehmen (z.B. Patente, Marken), nur in diesen Ländern gültig, in denen sie bei der zuständigen Behörde registriert worden sind. Im Rest der Welt, d.h. in den anderen Ländern, in denen z.B. ein Patent für eine Erfindung nicht registriert worden ist, kann diese Erfindung uneingeschränkt benutzt werden [1].

Weiter sind die Bestimmungen für den Schutz der geistigen Eigentumsrechte nicht in jedem Land gleich. Jedes Land verfolgt seine eigene Politik und hat eigene Gesetze. Dies führt dazu, dass der selbe Tatbestand in verschiedenen Ländern unterschiedlich interpretiert werden kann. In der heutigen Zeit wächst der internationale Handel und Unternehmen, welche international tätig sind, möchten ihr geistiges Eigentum in der ganzen Welt auf die gleiche Art schützen. Auch aufgrund der Internationalität des Internets wächst das Risiko der Verletzung von nationalen Gesetzen und der Wunsch nach internationaler Zusammenarbeit und Angleichung der Rechtsvorschriften wird grösser [1].

Bereits 1985 versuchte die WIPO das Patentrecht zu harmonisieren. Das Resultat war 1991 der Entwurf des Patent Harmonization Treaty (PHT), dessen Bestätigung jedoch an den USA scheiterte, die nicht bereit waren, ihr nationales Recht entsprechend anzupassen. Deshalb wurden in der Folge die inhaltlichen Fragen ausgeklammert und die weiteren Verhandlungen beschränkten sich lediglich auf Verfahrensregelungen. Dieser Teil wurde im Juni 2000 mit der Annahme des Patent Law Treaty (PLT) erfolgreich abgeschlossen [44].

Ein weiterer Versuch zur internationalen Zusammenarbeit im Bereich der geistigen Eigentumsgesetze ist die Patent Cooperation Treaty (PCT) von 1970. Viele Länder, darunter auch die Schweiz und alle wichtigen Nationen, haben diese unterzeichnet. Sie eröffnet die Möglichkeit zu einer 'Weltanmeldung', indem das Schutzrecht durch eine einzige internationale Patentanmeldung mit Wirkung für alle PCT-Vertragsstaaten geltend gemacht wird. Spätestens nach 20 Monaten muss der Anmelder jedoch in jedem gewünschten Land einzeln ein nationales Erteilungsverfahren einleiten. Da dies mit erheblichen Kosten verbunden ist, werden sich sogar grosse Firmen diesen Schritt gut überlegen. Denn auch hier gilt das Territorialprinzip. Die Patentämter können keine Schutzrechte außerhalb ihres Hoheitsbereiches vergeben [44].

Eine Lösung dieser Probleme wäre die Einrichtung eines Weltpatentamts oder die gegenseitige Anerkennung der nationalen geistigen Eigentumsrechte über bilaterale Abkommen. Beides setzt einheitliche, von allen Ländern akzeptierte Standards der Schutzrechtserteilung voraus [44]. Die Schweiz z.B. hat eine ganze Reihe von Handels- und Wirtschaftszusammenarbeitsabkommen (HuwZ) und Freihandelsabkommen (FHA) mit Drittstaaten abgeschlossen, die unter anderem auch das geistige Eigentum abdecken [4]. Die Tabelle 9.4 [45] zeigt eine Liste dieser bilateralen Wirtschafts- und Kooperations-Abkommen der Schweiz mit Drittstaaten.

In der Europäischen Union gibt es für Patente auch die Möglichkeit, einen Antrag auf Erteilung eines europäischen Patents beim Europäischen Patentamt zu stellen. Hier kön-

Tabelle 9.4: Bilaterale Wirtschafts- und Kooperations-Abkommen der Schweiz mit Drittstaaten, Stand 2005 [45]

Staat	Aktueller Status des Abkommens	Typ
Albanien	In Kraft seit 1. August 1996	HuwZ
Armenien	In Kraft seit 1. Januar 2000	HuwZ
Aserbaidschan	in Kraft seit 1. August 2001	HuwZ
Belarus	In Kraft seit 1. August 1994	HuwZ
Bosnien und Herzegowina	In Kraft seit 1. Juni 2002	HuwZ
China	Unterzeichnet am 8. Juli 1992	Memo. of Und.
Estland	In Kraft seit 2. März 1994	FHA
Georgien	In Kraft seit 1. Januar 2001	HuwZ
Iran	Paraphierung am 10. Dezember 2003	HuwZ
Kasachstan	In Kraft seit 1. Juli 1997	HuwZ
Kirgisien	In Kraft seit 1. Mai 1998	HuwZ
Kroatien	In Kraft seit 1. Juni 2000	HuwZ
Lettland	In Kraft seit 1. März 1994	FHA
Litauen	In Kraft seit 1. März 1994	FHA
Mazedonien	In Kraft seit 1. September 1996	HuwZ
Moldawien	In Kraft seit 1. September 1996	HuwZ
Russland	In Kraft seit 1. Juli 1995	HuwZ
Serbien und Montenegro	In Kraft seit 1. Juni 2002	HuwZ
Turkmenistan	Paraphierung am 9. November 1998	HuwZ
Ukraine	In Kraft seit 1. Dezember 1996	HuwZ
Usbekistan	In Kraft seit 22. Juli 1994	HuwZ
Vietnam	In Kraft seit 19. Mai 2000	Schutz IP&Z

nen durch eine einzige Anmeldung mehrere Vertragsstaaten benannt werden, in denen das Patent seine Wirkung entfalten soll. Mit der Erteilung eines europäischen Patents hat dieses grundsätzlich in jedem Staat, für den es erteilt ist, dieselbe Wirkung wie ein in diesem Staat erteiltes nationales Patent. Daher zerfällt das europäische Patent in ein Bündel nationaler Patente [46].

9.3.2 Die Piraterie als ökonomischer Kostenfaktor

Durch die weltweite Verbreitung des Internets, von illegalen elektronischen Tauschbörsen und einfach erhältlicher Kopier-Hardware war das Erstellen und Beschaffen von illegalen Raubkopien noch nie einfacher als heute. In diesem Kapitel liegt der Schwerpunkt auf der Softwarepiraterie. Natürlich sind z.B. auch die Musik- und Film-Industrie betroffen.

Daten zur Softwarepiraterie im Jahr 2004

Eine von der Business Software Alliance (BSA) in Auftrag gegebene Studie über den Schaden von Raubkopien, durchgeführt vom Marktforschungsinstitut IDC, kommt zum

Ergebnis, dass im Jahre 2004 weltweit Umsatzausfälle im Umfang von 32,7 Mrd. US-Dollar (USD) entstanden sind. In den USA betragen sie 6,6 Mrd. USD, in der EU 12,1 Mrd. USD und in der Schweiz 308,8 Mio. USD, was ca. 386 Mio. CHF entspricht (Umrechnungsmittelkurs der Quartalsersten 2004: 1 USD = 1.25 CHF). Diese Pirateriestatistik vergleicht den Gesamtbedarf an Software mit den tatsächlich verkauften Lizenzen und errechnet daraus den Anteil von Raubkopien. 2004 betrug der weltweite Gesamtmarkt für Software 90 Milliarden USD. Die legalen Umsätze machten 59 Mrd. USD aus. In den nächsten fünf Jahren werden schätzungsweise 300 Mrd. USD an Umsätzen im Softwarebereich generiert werden, gleichzeitig aber Programme im Wert von 200 Mrd. USD raubkopiert werden [2]. Die Tabellen 9.5 und 9.6 zeigen die Top 10 der Softwarepiraterie und die Top 10 der Software-Legalität [2].

Tabelle 9.5: Die Top 10 der Softwarepiraterie 2004 [2]

Staat	2004
Vietnam	92%
Ukraine	91%
China	90%
Simbabwe	90%
Indonesien	87%
Russland	87%
Nigeria	84%
Tunesien	84%
Algerien	83%
Kenia	83%

Tabelle 9.6: Die Top 10 der Software-Legalität 2004 [2]

Staat	2004
USA	21%
Neuseeland	23%
Österreich	25%
Schweden	26%
UK	27%
Dänemark	27%
Schweiz	28%
Japan	28%
Finnland	29%
Deutschland	29%

Die wirtschaftlichen Auswirkungen von Softwarepiraterie

Die weltweit erste Studie über die gesamtwirtschaftlichen Folgen der Softwarepiraterie wurde vom Marktforschungsinstitut IDC im Jahre 2003 durchgeführt. Sie kommt zum Schluss, dass die IKT, angetrieben durch die Softwareindustrie, ein wichtiger Einflussfaktor für Wirtschaftswachstum und Wohlstand ist. Eine Reduktion der Softwarepiraterie-

Rate könnte der weltweit stagnierenden Wirtschaft neue Impulse geben. Neue Arbeitsplätze und Geschäftchancen würden geschaffen, welche die Ausgaben und Steuereinkünfte ankurbeln würden [47]. Als wichtigste wirtschaftlichen Effekte des IKT-Sektors werden folgende Punkte [47] genannt:

- `IKT-Wachstum bringt entscheidende wirtschaftliche Vorteile`: Weltweit arbeiten bereits heute Millionen von Menschen in der IKT-Branche und es werden immer mehr. Dies führt zu Wohlstand und grossen Steuereinnahmen. Zwischen 1996 und 2002 wuchs der Informatik-Sektor um 26% und generierte über 2.6 Mio. neue Arbeitsplätze weltweit.

- `Software trägt erheblich zum Wachstum der IKT-Branche bei`: 2003 wies der Anteil der Software-Industrie an der gesamten IKT-Branche 60% auf.

- `Je geringer die Piraterie-Rate, desto mehr wächst die IKT-Branche`: Länder mit tiefer Softwarepiraterie-Rate profitieren von grösseren ökonomischen Vorteilen der IKT-Branche als Länder mit hoher Rate. Je kleiner die Softwarepiraterie-Rate, desto mehr wächst die IKT-Branche und desto grösser sind die Vorteile für die Wirtschaft.

Die Studie untersuchte, wie sich eine weltweite Reduktion der Softwarepiraterie-Rate um 10% im Jahre 2003 in den darauf folgenden vier Jahren auf die Wirtschaft ausgewirkt hätte. Die wichtigsten Auswirkungen [47] wären folgende:

- `Die IKT-Branche wächst durch die Reduktion der Softwarepiraterie-Rate`: In fast zwei Drittel der untersuchten Länder würde sich die IKT-Branche während vier Jahren mehr als verdoppeln. Chinas IKT-Sektor würde sich verfünffachen.

- `Schnelleres Wachstum der IKT bringt neue Arbeitsplätze, mehr Steuereinkünfte und ein höheres Wirtschaftswachstum`.

- `Länder mit hoher Softwarepiraterie-Rate würden den grössten Nutzen aus einer Piraterie-Reduktion ziehen`: Acht der Länder die 2003 zu den Top 10 Piraterie-Nationen gehörten, würden vier Jahre später in den Top 10 der Software-Legalität auftauchen.

- `Die Wirksamkeit einer Reduktion der Softwarepiraterie-Rate wurde bereits bewiesen`: Nationen die in den letzten Jahren die Piraterie-Rate erfolgreich reduzierten, konnten bereits von bedeutenden ökonomischen Vorteilen profitieren.

- `Jede Region der Welt würde profitieren`: Vor allem Asien, Ost- und Westeuropa würden von einer solchen Reduktion profitieren.

- `Eine 10%-Reduktion der Softwarepiraterie-Rate ist erreichbar`: Fast zweidrittel der untersuchten Ländern konnten seit 1996 ihre Softwarepiraterie-Rate um 10% reduzieren.

Von einer Reduktion der Softwarepiraterie-Rate könnten viele profitieren - Konsumenten, Unternehmer, Arbeiter, Staaten und Wirtschaften. Z.B. trug die IKT-Branche 2002 mit mehr als 700 Mrd. USD Steuergeldern weltweit zur Finanzierung öffentlicher Dienste bei [47]. Jede 1-Prozentpunkt-Reduktion würde die Steuergelder zusätzlich um 6 Mrd. USD erhöhen. Eine Reduktion von 10% würde 64 Mrd. USD Staatseinkommen generieren, was reichen würde für ...

- mehr als 30 Mio. Computer für Schulen
- medizinische Betreuung für mehr als 32 Mio. Menschen
- einen Hochschulabschluss für 6.9 Mio. Menschen
- Internetanschlüsse für mehr als 20 Mio. Menschen über vier Jahre hinweg (inklusive Telefon- und Dienstanbieter-Kosten)
- die Grundausbildung von ungefähr 4 Mio. Kinder

Staaten könnten gezielte Massnahmen ergreifen um grössere IKT-getriebene Wirtschaftsvorteile freizusetzen. Sie könnten in der Politik und durch eine strengere Gesetzgebung die Software-Piraterie bekämpfen, Konsumenten informieren und als Resultat von wirtschaftlichen Vorteilen profitieren [47].

Piraterie-Profiteure

Nicht alle Personen sind mit den aktuellen Gesetzen der geistigen Eigentumsrechten zufrieden. Z.B. wies der Autor und Dokumentarfilm-Regisseur Michael Moore 2004 ausdrücklick darauf hin, dass es ihm egal sei, wenn eine Kopie seines Films „Fahrenheit 9/11“ auf dem Internet in elektronischen Tauschbörsen getauscht werde. Er schreibe seine Bücher und drehe seine Filme nicht, um reich zu werden, sondern um Missstände aufzuzeigen, wozu er auch das mit den Büchern und Filmen verdiente Geld größtenteils verwenden wolle [48]. Michael Moore: „Solange sie dafür kein Geld verlangen, also sich nicht mit meiner Arbeit bereichern, sondern nur meine Gedanken weiter verbreiten, wozu ich die Filme und Bücher ja gemacht habe, ist das für mich völlig okay. Ich bin mit den jetzigen Copyright-Gesetzen nicht einverstanden!“ [48]. Diese Aussagen verhalfen Michael Moore und seinem Film zu enormem Medienecho und nicht zuletzt deshalb spielte sein Film Rekordsummen ein.

9.4 Lösungsansätze für die festgestellten Herausforderungen und Probleme der geistigen Eigentumsrechte

Der Schutz des Urheberrechts wird immer schwieriger. Raubkopien werden vermehrt und schneller hergestellt und verbreitet. Dies hat zur Folge, dass neue Massnahmen zum Schutz des Urheberrechts benötigt werden. Im folgenden Kapitel werden zwei rechtliche und eine technische Massnahme vorgestellt.

9.4.1 Neue und stärkere Verträge zur Sicherung der geistigen Eigentumsrechte im Internet: WCT und WPPT

Wie schon erwähnt, gab es das Urheberrecht schon vor der Entstehung von Computerprogrammen und die bisherigen Gesetze mussten angepasst und erweitert werden. Die Unterschiede im Gesetz zwischen dem Schutz von Software und dem Schutz anderer Formen geistigen Eigentums wurden schon erläutert (siehe Tabelle 9.3). Diese Gesetze sind jedoch nicht ausreichend, da sie, zum Beispiel, nur in der Schweiz anwendbar sind und die Gesetze vor der Verbreitung des Internets eingeführt wurden. Deshalb hat die WIPO zwei internationale Verträge zur Sicherung der Urheberrechte ausgehandelt: Den WIPO Copyright Treaty (WCT) und den WIPO Performances and Phonograms Treaty (WPPT) von 1996 [1]. Den WCT haben 56 Staaten ratifiziert, den WPPT 55, darunter die Schweiz, die USA und die EU [49]. Mit Hilfe des WPPT sollen die verwandten Schutzrechte zeitgemäss gesichert werden [50]. Ziel des WCT ist es, die Urheberrechte weltweit zu sichern unter Berücksichtigung des öffentlichen Interesses. In der Präambel des WCT wird dies wie folgt ausgedrückt:

PRÄAMBEL

DIE VERTRAGSPARTEIEN

- *IN DEM WUNSCH, den Schutz der Rechte der Urheber an ihren Werken der Literatur und Kunst in möglichst wirksamer und gleichmässiger Weise fortzuentwickeln und aufrechtzuerhalten,*
- *IN ERKENNTNIS der Notwendigkeit, neue internationale Vorschriften einzuführen und die Auslegung bestehender Vorschriften zu präzisieren, damit für die durch wirtschaftliche, soziale, kulturelle und technische Entwicklungen entstehenden Fragen angemessene Lösungen gefunden werden können,*
- *IM HINBLICK AUF die tiefgreifenden Auswirkungen der Entwicklung und Annäherung der Informations- und Kommunikationstechnologien auf die Erschaffung und Nutzung von Werken der Literatur und Kunst,*
- *UNTER BETONUNG der herausragenden Bedeutung des Urheberrechtsschutzes als Anreiz für das literarische und künstlerische Schaffen,*
- *IN ERKENNTNIS der Notwendigkeit, ein Gleichgewicht zwischen den Rechten der Urheber und dem umfassenderen öffentlichen Interesse, insbesondere Bildung, Forschung und Zugang zu Informationen, zu wahren, wie dies in der Berner Übereinkunft zum Ausdruck kommt*

SIND WIE FOLGT ÜBEREINGEKOMMEN [...] [51]

Neu ist im WCT die Sicherung der Urheberrechte in Bezug auf das Internet. Eine wichtige Regelung ist in Artikel 8 verankert [1].

Artikel 8

Recht der öffentlichen Wiedergabe

Unbeschadet der Bestimmungen von Artikel 11 Absatz 1 Ziffer 2, Artikel 11 bis Absatz 1 Ziffern 1 und 2, Artikel 11 ter Absatz 1 Ziffer 2, Artikel 14 Absatz 1 Ziffer 2 und Artikel 14 bis Absatz 1 der Berner Übereinkunft haben die Urheber von Werken der Literatur und Kunst das ausschließliche Recht, die öffentliche drahtlose oder drahtgebundene Wiedergabe ihrer Werke zu erlauben, einschließlich der Zugänglichmachung ihrer Werke in der Weise, daß sie Mitgliedern der Öffentlichkeit an Orten und zu Zeiten ihrer Wahl zugänglich sind [51].

Dieser Artikel wurde eingeführt, um die Veröffentlichung von Medien über das Internet durch Nicht-Urheber zu verhindern.

Die zwei Verträge sind in der heutigen Zeit sehr wichtig. Da die Medien viel schneller und weiter verbreitet werden, ist es nötig internationale Abkommen durchzusetzen, was mit dem WCT und dem WPPT gelungen zu sein scheint.

9.4.2 Das Aufkommen neuer, verbesserter Verschlüsselungstechniken zum stärkeren Schutz der geistigen Eigentumsrechte

In diesem Abschnitt geht es darum, wie die Unterhaltungsindustrie versucht, Verstösse gegen das Urheberrecht zu verhindern und was die Konsequenzen davon sind. Es gibt im Wesentlichen zwei Arten, wie sich die Industrie wehrt: Einerseits versucht sie es mit rechtlichen Mitteln, anderseits setzt sie DRM-Systeme ein. Zuerst zu den rechtlichen Mitteln.

Über den letzten Gerichtserfolg der Musik- und Filmindustrie konnte man kürzlich in der Zeitung lesen: Grokster, eine der grössten Internet-Tauschbörsen, ist in den USA verurteilt worden. Bisher hatten die Tauschbörsen vor Gericht damit argumentiert, dass sie selbst nicht am Tausch von Medien beteiligt seien, sondern lediglich Peer-to-peer Software herstellten. Diesmal entschied das Gericht aber, dass dem nicht so ist und hat den Softwareherstellern Unrecht gegeben. Nach monatelangen Verhandlungen, versucht es Grokster nun wie Napster auf legalem Weg weiterzuexistieren [39]. Es scheint also, dass sich die Bemühungen der Unterhaltungsindustrie in diesem Bereich gelohnt haben. Wieso aber dieser Richtungswechsel nach vielen Misserfolgen vor Gericht? Laut Justin Hughes von der Cardozo-Rechtsfakultät der Universität New York, haben die Peer-to-peer Unternehmen kein brauchbares Geschäftsmodell entwickelt. Die Investoren seien nicht bereit, ihr Kapital in illegale Projekte zu investieren [39].

Trotz dieses einen Erfolgs ist es um die Internet-Tauschbörsen nicht geschehen. Es gibt immer noch viele und diese werden rege benutzt. Daraus lässt sich schliessen, dass die vorhandenen rechtlichen Mittel nicht ausreichen, um die verlorenen Gewinne wieder einzuholen. Die Firmen wollen das Urheberrecht aktiv schützen und setzen deshalb DRM-Systeme ein.

Digital Rights Management: Schutz oder Einschränkung?

Als Digital Rights Management Systems werden Technologien bezeichnet, die das Kopieren von digital speicherbaren Medien einschränken oder auch vollständig verhindern oder Zur Veranschaulichung der Anwendung eines solchen Systems nochmals das Beispiel iTunes aus 9.2.5. Die Firma Apple schützt die Musikstücke, die über den iTunes Music Store erworben werden, mittels einer Software mit dem Namen FairPlay [52]. Diese verhindert, dass die Medien auf mehr als fünf Computern gleichzeitig abgespielt werden können. Bevor der Benutzer ein über iTunes erworbenes Stück auf seinem Computer abspielen kann, muss er diesen dafür autorisieren, wobei maximal fünf Computer auf einmal für ein Stück autorisiert sein können. Will er das Musikstück auf einem sechsten PC abspielen, muss er zuerst eine der fünf ersten Maschinen „deautorisieren", um dem gewünschten PC zu erlauben, das Medium zu spielen. Trotz „Deautorisation" wird das Stück nicht von der Maschine gelöscht, es kann lediglich nicht auf dieser angehört werden, solange sie nicht dafür autorisiert ist. Für den iPod gibt es keine Beschränkungen. Die Medien können auf beliebig viele iPods gespeichert und angehört werden [53].

Das DRM von Apple wird insofern kritisiert, dass FairPlay nur Medien schützt, die über den iTunes Music Store gekauft wurden. Andere Stücke können mit iTunes kopiert werden und auf beliebig vielen PCs abgespielt werden. Natürlich haben Hacker es schon geschafft, den Schutzmechanismus von Apple zu umgehen, so dass die Titel ohne DRM-Schutz heruntergeladen werden können [43]. Was zeigt, dass DRM nicht bei allen Anklang findet. Das kommt vor allem daher, dass DRM Systeme nicht nur als Schutz, sondern sehr oft als Einschränkung empfunden werden. Dieser Aspekt soll anhand eines aktuellen Beispiels erläutert werden.

Wer in letzter Zeit die Nachrichten in der IKT-Branche verfolgt hat, hat höchst wahrscheinlich von den Problemen mit Sonys DRM-System gehört [54]. Das Problem ist das folgende: In den USA hat Sony CDs mit DRM-Software verkauft. Will der Käufer einer solchen CD, diese auf seinem Computer abspielen, kann er das nur mit dem auf der CD mitgelieferten Player tun. Er kann die Musik weder mit einem anderen Player abspielen, noch kann er die CD kopieren. Bisher ist nichts speziell an der ganzen Sache, es scheint wie ein übliches DRM-System zu funktionieren. Der Haken an der ganzen Sache ist, dass Sonys DRM-Software nicht nur die CD schützt und den Player installiert, sondern im Hintergrund, ohne Wissen des Benutzers ein sogenanntes „rootkit" installiert. Rootkits werden üblicherweise von Hackern verwendet, um den Zugang zu einem fremden Computer zu verstecken. Registry Einträge sowie Verbindungen und Dateien können mit Hilfe von rootkits vor dem Benutzer verborgen werden. Es ist zudem nicht einfach, ein rootkit zu deinstallieren [55]. Die technischen Details sind hier nicht weiter von Bedeutung. Wichtig ist, dass rootkits kein geeignetes Mittel sind, um ein DRM umzusetzen. Weder Sony noch eine andere Firma ist dazu befugt, unerlaubt Software auf einen Computer zu installieren, vor allem nicht solche, die einem Computer schaden kann. Sony hat zwar einen 'uninstaller' zur Verfügung gestellt, doch das Herunterladen ist ziemlich umständlich. Die Software soll zusätzlich nach der Installation über das Internet mit Sony Kontakt aufnehmen [56]. Es stellt sich die Frage, ob das noch DRM ist oder schon mit Spyware vergleichbar ist. Ist dies DRM, so sind die Nachteile klar. Der Käufer kann nicht frei über das Produkt verfügen, wird zudem von der Musikindustrie überwacht und verliert die

Kontrolle über seinen Computer. Eine solche Art von DRM ist nicht förderlich und wird sich langfristig nicht durchsetzen können. Wie man es schon im Fall von Sony sehen kann, schadet ein solcher Vorfall dem Image der Musikfirmen erheblich und fördert letztendlich die Piraterie.

9.5 Zusammenfassung und Schlussfolgerung

Erfindungen müssen von Patenten geschützt werden, um einen Beitrag zum wirtschaftlichen Wachstum leisten zu können. Der Erfinder erhält einen temporären Wettbewerbsvorteil und Schutz vor den Kräften des Marktes. Durch diesen Schutz kann dieser neue Produkte im Zusammenhang mit seinem Patent entwickeln und erfolgreich vermarkten. Patente bringen nicht nur dem Erfinder Vorteile, sondern auch der Allgemeinheit. Patentierte Erfindungen müssen offengelegt werden und stehen nach dem Ablauf der Patentfrist allen frei zur Verfügung. So werden Patente zum Grundstein für neue Erfindungen. Wichtig für die IKT-Branche sind vor allem Hardware- und Softwareschutz durch Patente. In einigen Ländern ist Hardware durch Patente geschützt, in anderen durch spezielle Gesetze. In den USA ist bereits heute die Patentierung von reiner Software möglich, z.B. der Schutz von sogenannten Geschäftsmethoden.

Die Marken, sowie das Industriedesign spielen als ökonomischer Aktivposten einer Unternehmung in der IKT-Branche eine grosse Rolle. Häufig ist der Erfolg oder Misserfolg einer Unternehmung abhängig von ihren Marken und der Gestaltung ihrer Produkte. Das trifft gerade in der IKT-Branche zu, da durch zunehmend gleiche Leistungsmerkmale der Produkte und durch die steigende Komplexität der Geräte, die Bewertung und Differenzierung für den Kunden immer schwieriger wird. Deshalb werden auch Gütesiegel wichtiger, um dem Kunden einfach und deutlich die Leistungsmerkmale der Produkte signalisieren zu können. Die zunehmende Globalisierung, sowie die Distribution und der Verkauf über das Internet haben an dieser Tatsache nichts geändert. Allerdings änderte sich die Art und Weise der Werbung: neu wird jetzt mittels passender Domainnamen und genau abgestimmtem Internetauftritt auf die entsprechenden Produkte aufmerksam gemacht. Der Internetauftritt wird für die Unternehmen wichtiger als die lokale Werbung in Zeitungen oder im Geschäft selbst. Durch diese neue Form der Werbung ist aber auch schnell eine neue Art des Missbrauchs entstanden: Privatpersonen erstehen sich bekannte Domainnamen, um sie anschliessend den Unternehmen gewinnbringend verkaufen zu können. Um die Rechte des geistigen Eigentums in diesem neuen Umfeld besser schützen zu können, erliess die ICANN die UDRP Resolution, die zum besseren Schutz der Marken in Kombination mit den zugehörigen Domainnamen dient. Zusätzlich ist die Haager Übereinkunft in Kraft getreten: sie ermöglicht es, ein Produkt nun durch Abgabe eines Formulars in mehreren Staaten gleichzeitig zu schützen.

Das Urheberrecht wurde lange bevor Computerprogramme an Bedeutung gewannen eingeführt. Mit der Verbreitung von Computern und des Internets entstanden jedoch neue Herausforderungen für das Urheberrecht. Die Gesetze mussten angepasst und ergänzt werden. Es stellte sich jedoch bald heraus, dass die Gesetze für digitale Medien keinen hinreichenden Schutz des Urheberrechtes bieten. Dabei geht es nicht nur um Software sondern allgemein um Werke, welche digital speicherbar sind. Nicht nur die Schöp-

fer der Werke, sondern vor allem die Industrie bemüht sich darum, die Urheberrechte zu sichern. Die Massnahmen sind einerseits legale Mittel, wie zum Beispiel Klagen gegen Internet-Tauschbörsen, andererseits DRM-Systeme zum direkten Schutz der Medien. DRM-Systeme sind umstritten, da sie von den Benutzern als Einschränkung der Nutzung der gekauften Medien angesehen werden. Zur weltweiten Sicherung des Urheberrechts wurden zudem von der WIPO die zwei internationalen Verträge WCT und WPPT eingeführt. Diese berücksichtigen und regeln unter anderem den Schutz des Urheberrechts unter der Verwendung des Internets.

Neue Technologien veränderten in den letzten Jahrzehnten das Umfeld der IKT-Branche und führten zu neuen Problemen und Herausforderungen. Durch die Internationalisierung des Internets wächst das Risiko der Verletzung von nationalen Gesetzen und der Wunsch nach internationaler Zusammenarbeit und Angleichung der Rechtsvorschriften wird grösser. Der illegale Tausch von urheberrechtlich geschützten Werken wurde durch das Internet vereinfacht. Könnte die Softwarepiraterie-Rate um nur wenige Prozentpunkte verringert werden, würde dies die Entwicklung der IKT-Branchen nachhaltig fördern und zu höherem Wirtschaftswachstum führen. Softwarehersteller versuchen zusehends, ihre Produkte durch neue Technologien vor urheberrechtsverletzenden Aktionen zu schützen. Dadurch werden aber nicht nur Urheberrechte geschützt sondern auch die Freiheit der Benutzer eingeschränkt. Es existieren bereits heute Lösungsansätze für diese neuen Herausforderungen: Verbesserte, flexiblere Verschlüsselungstechnologien sollen helfen, die richtige Balance zwischen Schutz und Einschränkung zu finden. Neue internationale Gesetze und Vereinbarungen können dazu beitragen, die Folgen des Territorialprinzips zu verringern und Registrationen von geistigen Eigentumsrechten zu vereinfachen.

Literaturverzeichnis

[1] WIPO: „Intellectual Property - A Power Tool for Economic Growth“, http://www.wipo.int/about-wipo/en/dgo/wipo_pub_888/index_wipo_pub_888.html, 7.11.2005.

[2] Business Software Alliance (BSA): „Schaden durch Raubkopien sinkt in der Schweiz auf 386 Mio. Franken“, http://www.bsa.org/switzerland/presse/newsreleases/BS055-08C.cfm, 18.5.2005.

[3] WIPO: „About Intellectual Property“, http://www.wipo.int/about-ip/en/, 20.10.2005.

[4] Eidgenössisches Institut für Geistiges Eigentum: „Eidgenössisches Institut für Geistiges Eigentum“, http://www.ige.ch/, 15.10.2005.

[5] Eidgenössisches Institut für Geistiges Eigentum: „Richtlinien für die Sachprüfung der Patentgesuche“, http://www.ige.ch/D/jurinfo/documents/richtpat.pdf, 7.11.2005.

[6] Koch F. A., Schnupp P.: „Software-Recht, Band 1“, Springer-Verlag, 2001.

[7] Amazon.com: http://www.amazon.com/,2005.

[8] Carr D. et al.: „Software Patents and their Implications - Amazon One-Click Shopping“, Stanford University, http://cse.stanford.edu/classes/cs201/projects-99-00/software-patents/, 5.6.2000.

[9] Google.com: http://www.google.com/, 2005.

[10] US Patent and Trademark Office (USPTO), Patent Full Text and Image Database: „Systems and methods for highlighting search results“, http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=/netahtml/srchnum.htm&r=1&f=G&l=50&s1=6,839,702.WKU.&OS=PN/6,839,702&RS=PN/6,839,702, 4.1.2005.

[11] BITKOM: „Presseinformation - Chance verpasst für eine einheitliche Patent-Praxis“, http://www.bitkom.org/files/documents/Presseinfo_BITKOM_Entscheidung_Softwarepatente_06.07.2005.pdf, 11.11.2005.

[12] Sommer U.: „Einführung zum Thema Softwarepatente“, http://www.ulisommer.de/computer/swp-begriffe.htm, 7.11.2005.

[13] Zintzmeyer und Lux: „The Best Global Brands 2005“, http://www.interbrand.ch/d/presse/, 1.8.2005.

[14] Robert Berner and David Kiley: „Global Brands“, BusinessWeek Magazine, 1.8.2005.

[15] Intel.com: http://www.intel.com/, 2005.

[16] Microsoft.com: http://www.microsoft.com/, 2005.

[17] Sun.com: http://www.sun.com/, 2005.

[18] Prof. Dr. Egon Franck: „BWL 3: Strategische Unternehmensführung“, Universität Zürich, SS 05.

[19] AMD.com: http://www.amd.com/, 2005.

[20] National.com: http://www.national.com/, 2005.

[21] Chiranjeev Kohli und Mrugank Thakor: „Branding Consumer Goods: Insights from Theory and Practice“,Journal of Consumer Marketing 14, No 3 (Frühling 1997).

[22] Energy Star Company: „ENERGY STAR“, http://www.energystar.gov/,2005

[23] Apple.com: www.apple.com, 2005.

[24] Dolby: „Dolby Digital“, http://www.dolby.com/, 2005.

[25] Allen: „Chelsea Football Club Announces Samsung As Official Sponsor“, http://www.mobiledia.com/forum/topic30099.html, 25.4.2005.

[26] Samsung: „The New Brand Campaign“, http://www.samsung.com/, 2005.

[27] ICANN: „Uniform Domain Name Dispute Resolution Policy“, http://www.icann.org/udrp/udrp-policy-24oct99.htm, 24.10.1999.

[28] Jack McCarthy: „Senate committee targets 'cybersquatters' by approving new bill“, http://www.cnn.com/TECH/computing/9908/02/cybersquat.idg/, 2.8. 1999.

[29] Anne Chasser: „Cybersquatting and Consumer Protection: Ensuring Domain Name Integrity“, http://judiciary.senate.gov/oldsite/72299ac.htm, 22.7. 1999.

[30] Bundesgesetz über den Schutz von Marken und Herkunftsangaben: „Markenschutzgesetz, MSchG“, 28.8.1992.

[31] Christoph Spahr: „Wirtschaftsinformatik: Internet und Recht“,vdf, 28.8.1992.

[32] WIPO: „What is an industrial design?“, http://www.wipo.int/about-ip/en/industrial_designs.html, 2005.

[33] Hans Peter Wehrli: „Marketing: Einführung“, Bütler und Partner AG, 5.Auflage 2003.

[34] Ed Tracy: „History of computer design: Apple IIc“, http://www.landsnail.com/apple/local/design/apple2c.html, Frühling 1998.

[35] Apple: „Apple and the Global Environment“, http://www.landsnail.com/apple/local/design/apple2c.html, 2005.

[36] Apple: „Product Design Awards“, http://www.apple.com/environment/design/awards.html, 2005.

[37] WIPO: „Hague Agreement Concerning the International Deposit of Industrial Designs“, http://www.wipo.int/treaties/en/registration/hague/, 2005.

[38] Wolfgang Straub: „Informatikrecht: Einführung in Softwareschutz, Projektverträge und Haftung“, vdf, 2003.

[39] Walter Niederberger: „Grokster: Funkstille für illegale Musikbörse“, Tagesanzeiger, 9.11.2005.

[40] WIPO: „Standing Committee on Copyright and Related Rights, Tenth Session, Geneva, November 3 to 5, 2003“, http://www.wipo.int/meetings/en/html.jsp?url=http://www.wipo.int/documents/en/meetings/2003/sccr/doc/sccr_10_2_rev.doc, 4.05.2004.

[41] Hal R. Varian, Joseph Farrell und Carl Shapiro: „The Economics of Information Technology“, Cambridge University Press, 2004.

[42] Wikipedia: „Die freie Enzykopädie“, http://de.wikipedia.org/wiki/Apple_iTunes, 1.11.2005.

[43] golem.de: „IT-News für Profis“, http://www.golem.de, 1.11.2005.

[44] Heise Online: „Der mühsame Weg zum Weltpatent“, http://www.heise.de/newsticker/meldung/17820, 17.5.2001.

[45] Eidgenössisches Institut für Geistiges Eigentum: „Bilaterale Wirtschafts- und Kooperations Abkommen der Schweiz mit Drittstaaten“, http://www.ige.ch/D/jurinfo/j13001.shtm, 11.11.2005.

[46] 123recht.net: „ Schutzwirkung und Geltungsbereich eines Patents“, http://www.123recht.net/article.asp?a=285&f=ratgeber_patentrecht_patentwas&p=3, 11.11.2005.

[47] BSA, IDC: „Expanding Global Economies - The Benefits of Reducing Software Piracy“, http://www.bsa.org/austria/piraterie/upload/IDC-2003.pdf 2.4.2003.

[48] Telepolis: „Michael Moore: Raubkopieren ist erlaubt, solange niemand daran verdient“, http://www.heise.de/tp/r4/artikel/17/17808/1.html, 5.7.2004.

[49] WIPO: „Treaties and Contracting Parties“, http://www.wipo.int/treaties/en/, 15.11.2005.

[50] WIPO: „WIPO Performances and Phonograms Treaty“, http://www.wipo.int/treaties/en/ip/wppt/trtdocs_wo034.html, 20.12.1996.

[51] WIPO: „WIPO Copyright Treaty“, http://www.wipo.int/treaties/en/ip/wct/trtdocs_wo033.html, 20.12.1996.

[52] iTunes: „iTunes Customer Service“, `http://www.apple.com/de/support/itunes/authorization.html`, 6.12.2005.

[53] iTunes: „iPod and iTunes Support“, `http://www.apple.com/support/itunes/musicstore/authorization/`, 13.11.2005.

[54] Sony.com: `http://www.sony.com/`, 2005.

[55] Sysinternals: „Mark's Sysinternals Blog“, `http://www.sysinternals.com/blog/2005/10/sony-rootkits-and-digital-rights.html`, 31.10.2005.

[56] Sysinternals: „Mark's Sysinternals Blog“, `http://www.sysinternals.com/blog/2005/11/more-on-sony-dangerous-decloaking.html`, 4.11.2005.

Kapitel 10

Charging Models in DiffServ Networks

Marc Eichenberger, Tariq Abdul, Visay Saycocie

Die schnell wachsende Entwicklung des elektronischen Marktes in den letzten paar Jahren führte zu einer rasanten Entwicklung und Verbreitung von multimedialen Internetapplikationen. Dies wiederum führte zu Überlastungen und Staus im Datenverkehr des Internets. Die Bandbreite wird aufgrund dieser Entwicklungen zu einer immer knapperen Ressource. Deshalb werden Stimmen laut, die nach einer Möglichkeit zum Verrechnen der genutzten Bandbreite schreien.
Unser Paper führt in einem ersten Teil in Funktionsweise und Komponenten von DiffServ-Netzwerken ein. Es wird auf die verschiedenen Möglichkeiten zur Implementierung von Serviceklassen in einem DiffServ-Netzwerk eingegangen. Zudem werden die Begriffe Quality of Service (QOS) und Traffic Conditioning Agreement (TCA) erläutert. In einem zweiten Teil gehen wir auf das Business Model, deren Phasen und die involvierten Parteien sowie deren Funktionen im Business Model ein. Zudem versuchen wir die Abläufe, die bei der Verrechnung von Dienstleistungen in DiffServ-Netzwerken anfallen, zu illustrieren.

Inhaltsverzeichnis

10.1 Einleitung

10.1.1 Technologische Umsetzung - Überblick

Wie bereits erwähnt ist DiffServ ein Quality of Service Mechanismus zur Behandlung der Verkehrsflüsse in einem oder mehreren Netzwerken. Um diese Anforderungen umzusetzen, ist es nötig, verschiedene Serviceklassen mit entsprechend verschiedenen Prioritäten zu definieren. Ein grosser Vorteil von DiffServ besteht darin, dass die Einteilung in die verschiedenen Klassen ohne eine aufwendige Signalisierung an jedem Router möglich ist. Die Einteilung geschieht bei den Boundary Routers (siehe Abbildung 10.1) zwischen den einzelnen DiffServ Domains und daher kann sich der Kern der Netzwerke vollständig den Routingaufgaben zuwenden. Jedes Paket kann entsprechend der gewünschten Dienstgüte markiert werden und wird fortan nur noch entsprechend seiner Markierung behandelt. Alle Pakete mit gleicher Markierung haben die gleiche Priorität und werden daher gleich behandelt. Ein DiffServ besteht nun hauptsächlich aus drei Elementen. Das erste Element sind die Per-Hop-Behaviors (PHBs), welche beschreiben, wie ein Paket von einem Router weitergeleitet werden soll. Ein weiteres Element sind die Traffic-Conditioner, welche den Verkehr messen und Schutzmassnahmen ergreifen (drop, Service Degradation), falls die Netzbelastung zu gross wird. Das dritte Element ist der Bandwidth Broker, welcher für das Aushandeln der Reservierungen zuständig ist.

Die DiffServ-Basisarchitektur sieht somit folgendermassen aus:

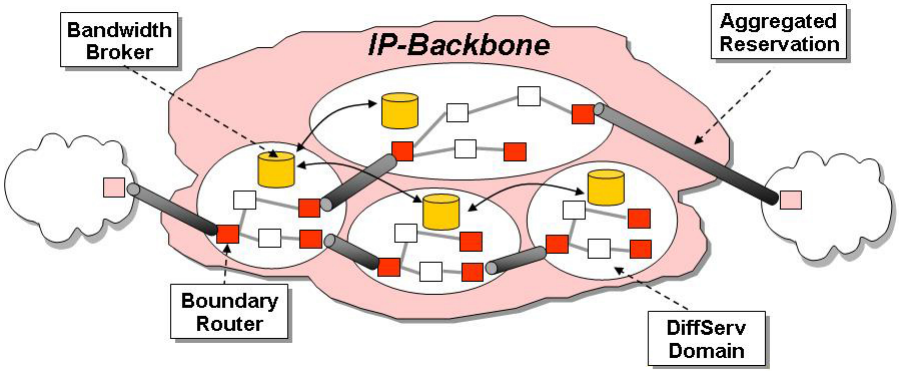


Abbildung 10.1: Basisarchitektur [1]

Die DiffServ Domain ist eine Menge von DS-Knoten mit einer gewissen Anzahl von PHB-Gruppen. Innerhalb der Domain werden die Pakete auf Grund der Per-Hop-Behaviors weitergeleitet. Mehrere DS-Domains zusammen bilden eine DS-Region.

Wie werden also nun die einzelnen IP-Pakete markiert? DiffServ verwendet eine neue Definition des IPv4 Type-of-Service Header-Feldes und IPv6 Traffic class Header-Feldes.

10.1.2 Per-Hop-Behaviors

Jeder PHB-Fluss ist durch einen Different Services Code Point (DSCP) bestimmt (siehe Abbildung 10.2). Vorgeschlagene PHBs sind:

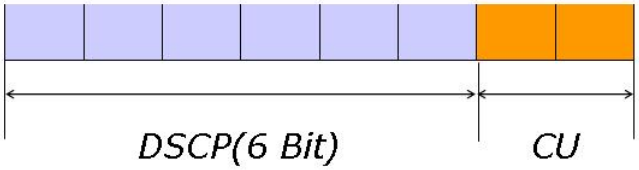


Abbildung 10.2: DiffServ Header [2]

- Expedited Forwarding PHB
- Assured Forwarding PHB
- Default PHB (Best Effort)

Ist das PHB zu einem gegebenen DSCP unbekannt, so wird das betroffene Paket dem Default-Fluss zugewiesen und wird somit nach der Best Effort Methode übermittelt.

Expedited Forwarding PHB

Das Konzept von Expedited Forwarding ist ziemlich einfach. Es sind zwei Dienstklassen verfügbar: Normal (regular) und Express (expedited). Der Grossteil des Datenverkehrs wird normal, ein kleiner Teil aber wird beschleunigt übertragen. Die Expresspakete sollten das Netz so durchqueren können, wie wenn keine anderen Pakete vorhanden wären. Die Idee dahinter ist, dass der Verkehr an allen Knoten so konfiguriert wird, dass Anzahl der ankommenden Pakete kleiner ist als jene der abgehenden. Diesen Grundsatz kann man zum Beispiel durch die Implementierung einer Warteschlange mit höchster Priorität mit einem Token-Bucket am Eingang erreichen. Dabei funktioniert der Token-Bucket-Algorithmus folgendermassen. Ein konstanter Strom von Tokens fällt in einen Behälter (Tocken-Bucket) und jeder Token repräsentiert ein Träger für eine vorgegebene Anzahl von Bytes. Werden nun zu einem Zeitpunkt nicht die volle Kapazität des Tokenstroms genutzt, so sammeln sich die Token im Tokenbucket, welcher aber höchstens b Token speichern kann. Diese nicht genutzte Kapazität ermöglicht es nun spontane Schwankungen, welche über der gegebenen Übertragungsrate liegen, auszugleichen, sofern sich genügend Token im Token-Bucket befinden.[3]

Eine weitere Möglichkeit, diese Strategie zu implementieren, ist, dass die Router zwei Warteschlangen für jede Ausgangsleitung haben, eine für Expresspakete und eine für normale Pakete. Kommt ein Paket an, wird es in die passende Warteschlange gestellt. Für die zeitliche Planung eignet sich ein Algorithmus wie Weighted Fair Queueing (WFG). Werden nun zum Beispiel 20% des Datenverkehrs Express übertragen und 80% normal, könnte man 40% der Bandbreite für den Expressdienst reservieren und den Rest für die normale Übertragung. Auf diese Weise bekommt der Expressverkehr doppelt so viel Bandbreite wie er eigentlich benötigt. Dadurch wird eine geringe Übertragungsverzögerung sichergestellt.

Expedited Forwarding PHB ist für Dienste geeignet, welche niedrige Paketverlustraten (loss), niedrige Varianz der Verzögerung (Jitter) und niedrige Verzögerung (latency) voraussetzen.

Assured Forwarding PHB

Die Idee hinter diesem Ansatz besteht darin, dass man vier verschiedene Güteklassen definiert. Jede Klasse reserviert dabei in jedem Knoten ein Teil von Ressourcen. Die Klassen enthalten ihrerseits drei drop-precedence Stufen, dies sind low, medium und high. Sie werden unter anderem am DS Domain-Ausgangsknoten vergeben (traffic conditioning). Die Prioritätsstufen kommen im Falle eines Staus zu Einsatz, wobei dort jene Pakete mit der Stufe low als erste verworfen werden.

Eine Implementierungsmöglichkeit wäre hier zum Beispiel eine RED-Queue (siehe Abbildung 10.3). RED steht für Random Early Drop und ist ein spezieller Mechanismus der unter anderem zur Pufferverwaltung eingesetzt wird. Dieser Mechanismus bewirkt, dass gezielt einzelne Pakete verworfen werden, noch bevor der Puffer zu überlaufen beginnt. Dadurch wird erreicht, dass der TCP-Regelmechanismus früher greift, die Senderaten der Quellen frühzeitig an die vorhandenen Ressourcen angepasst werden, somit grössere Verluste vermieden werden können und der mittlere Durchsatz durch das Netz erhöht wird. Bei der N-RED-Queue besitzen die n-verschidenen Klassen unterschiedliche grössen der Warteschlangen (siehe Abbildung 10.4).[4]

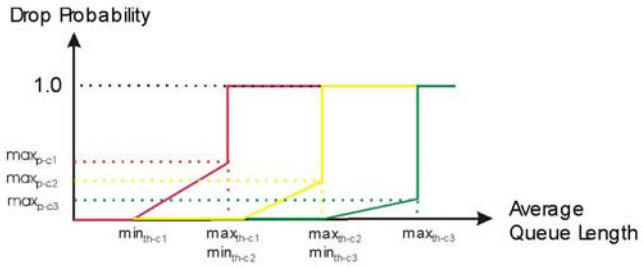


Abbildung 10.3: RED-Queue [2]

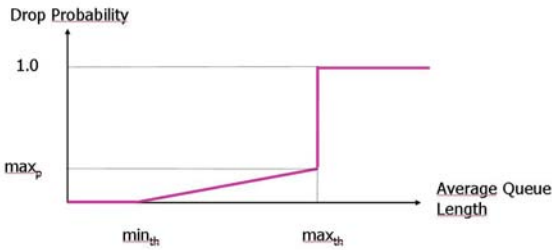


Abbildung 10.4: N-RED-Queue [2]

Die Verarbeitung der Pakete funktioniert nun folgendermassen.

In einem ersten Schritt werden Pakete einer der vier Prioritätsklassen zugeordnet (Classifier). Dies geschieht entweder auf dem Sendenden Host oder auf dem ersten Router. Geschieht die Klassifizierung auf dem sendenden Host, so hat dies den Vorteil, dass mehr Informationen verfügbar sind, welche Pakete zu welchem Datenfluss gehören.

In einem zweiten Schritt werden die Pakete durch den Marker entsprechend der Klasse markiert, in die sie eingeteilt wurden. Die Markierung erfolgt mittels des Diensttypfeldes (8Bit) im IP-Header.

In einem dritten Schritt werden die vier Datenströme mittels eines Shaper/Dropper-Filters, welcher einige Pakete verzögern oder verwerfen kann, in akzeptable Formen gebracht.

Der Meter misst die Geschwindigkeit des Verkehrsflusses (siehe Abbildung 10.5).

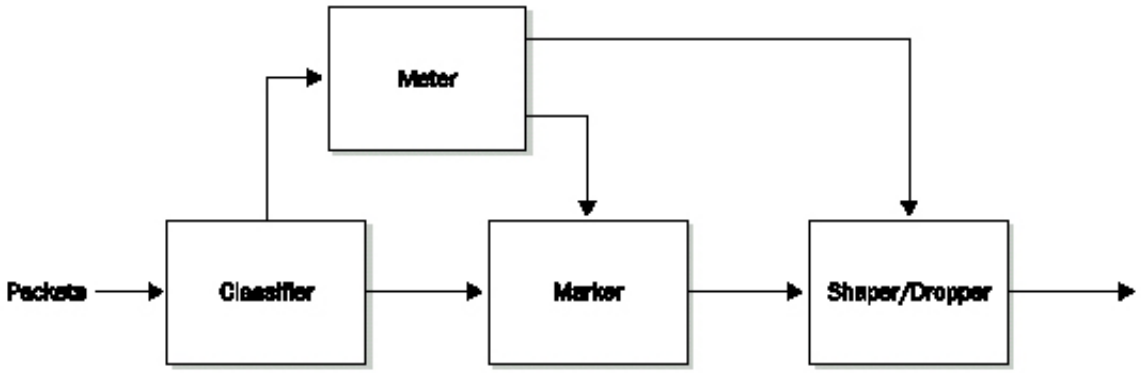


Abbildung 10.5: Assured Forwarding [5]

10.1.3 Bandwidth Broker

Die Funktion des Bandwidth Brokers in der DS Domain ist die des Ressourcenmanagers. Der Bandwidth Broker kann als Host, Router oder Software-Prozess am Edge-Router implementiert werden. Er besitzt eine Datenbank mit Informationen betreffend Service-Level-Agreements (SLA), Prioritäten und Restriktionen für einzelne Benutzer sowie die Daten für die Authentifikation. Er bearbeitet die Serviceanforderungen von Hosts und konfiguriert, sofern die entsprechenden Ressourcen vorhanden sind, den Firstrouter für Shaping, Policing und Marking.

Der Vorteil des Bandwidth Brokers ist dabei, dass die anderen Knoten in der DS Domain entlastet werden, da sie die Pakete nur noch entsprechend dem DS-Feld weiterleiten müssen.

10.1.4 SLA und TCA

SLA steht für Service Level Agreement und bezieht sich auf die individuellen Vereinbarungen zwischen dem Internet-Service-Provider (ISP) und dem Kunden. Sie spezifizieren die Service-Klassen und den entsprechenden erlaubten Grad des Verkehrs in jeder einzelnen Klasse. Je höher die SLA‘s umso höher ist die dem Kunden zugesicherte Dienstgüte. Dabei unterscheidet man zwischen den statischen und den dynamischen SLA‘s. Bei statischen SLA‘s muss noch zusätzlich das Updateintervall festgelegt werden. Die dynamischen SLA‘s verwenden Signalisierungsprotokolle. Hier kommt oftmals das Ressource Reservation Protokoll (RSVP) zu Einsatz.

TCA steht für Traffic Conditioning Agreement und enthält die Regeln für das Marking, Metering, Policing und Shaping. Diese Regeln werden vom Classifier auf die entsprechenden Paketströme angewendet (siehe Abbildung 10.6).

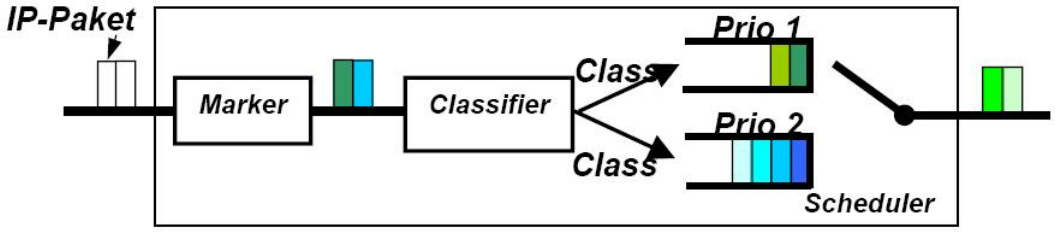


Abbildung 10.6: Schematischer Aufbau von einem Ausgang eines Netzknotens [6]

10.1.5 Assured Service

Ist ein bevorzugter Service gegenüber dem Kunden, welcher auch während einem Stau im Netzwerk gilt. Dieser Dienst ist aber nicht als hundertprozentige Garantie zu verstehen.

Die Service Level Agreements (SLA) spezifizieren, wie viel Bandbreite dem Kunden für diesen Service zur Verfügung gestellt wird. Der Kunde teilt nun die ihm zugewiesene Bandbreite selbst unter den verschiedenen Anwendungen auf. Somit markiert der Sender seine Klasse selbst.

Klassifikation und Policing wird an dem ISP Eingangsknoten durchgeführt. Wird die maximale Bitrate gemäss den SLAs nicht überschritten, gelten die Pakete als „in profile“ andernfalls werden sie als „out of profile“ klassifiziert. Die entsprechenden Pakete kommen nun in eine „in-profile“ und eine „out-of-profile“ Queue, welche die richtige Ankunftsreihenfolge beim Empfänger garantiert. Hier wird der RED oder RIO Warteschlangenalgorithmus verwendet.

10.1.6 Premium Service

Beim Premium Service wird dem Kunden eine Verbindung mit kleiner Verzögerung (low delay), niedrige Varianz der Verzögerung (Jitter) und fester maximaler Bitrate zugesichert. Wird diese in den SLAs vereinbarte maximale Bitrate überschritten, werden die Pakete vernichtet. Für das Traffic Conditioning wird das P-Bit gesetzt und das Shaping wird an den Grenzknoten durchgeführt durch zum Beispiel einen Token-Bucket-Algorithmus. Der Premium Service hat die höchste Priorität, das Problem ist aber trotzdem die Koexistenz des Assured Service. Daher ist es sehr wichtig, dass der Premium Service richtig konfiguriert wird. Die nicht benötigte Bandbreite kann von Assured Services und Best Effort genutzt werden. Der Premium Service findet z.B. bei der Internet Telephonie oder dem Video Conferencing Verwendung.

10.1.7 User Share Differentiation Service

Beim User Share Differentiation Service wird die verfügbare Bandbreite des ISP unter allen Benutzern aufgeteilt. Jeder Benutzer bekommt hierbei einen minimale Bandbreiteanteil. Der Rest der verfügbaren Kapazität wird nun den Teilnehmern anhand der von ihnen reservierten Bandbreite zugeteilt.

10.2 Geschäftsmodelle

Es gibt drei grundlegende Stufen von Abrechnungen und Buchhaltungen im Internetdienstleistungsbereich. Erstens die Businessstufe (Business Level), zweitens die Vertragsstufe (Contract Level) und drittens die Netzwerkstufe (Network Level). Das Business Model beschreibt auf der Businessstufe, welche Parteien beteiligt sind, welche Rolle und welche Geschäftsbeziehungen sie zueinander haben um ein elektronisches Geschäft durchzuführen. Ausserdem sind Verträge und Einverständnisse sehr wichtig um die Geschäftskonditionen und die Serviceleistung der Parteien zu bestimmen. Auf der Vertragsebene sind die Parteien verpflichtet unter den vordefinierten Konditionen die Serviceleistungen zu liefern und zu erbringen. Als letztes bilden die Netzwerkkomponenten und Architekturen die Basis der Netzwerkstruktur, die während dem Businessprozess aus der vordefinierten Vertragsphase im Business Model entstanden sind.

In einer Geschäftsumgebung herrscht ewig ein starker Konkurrenzkampf. Dadurch kommt es schnell einmal zu vertrauenswürdigen und nicht vertrauenswürdigen Geschäftsbeziehungen und auch zwischen Parteien, die einander nicht einmal kennen. Zum Beispiel kennt der Endbenutzer die als Zwischenknoten dienenden ISPs, die für die entstandenen Verbindungen zum ESP tangiert werden, gar nicht.

Hinzu kommen noch andere nicht berücksichtigte Einflussfaktoren darunter Risiken wie unrechtmässige Täuschung, nicht einhalten von Verträgen, usw. All dies und noch weitere können den involvierten Parteien gravierende Schäden zufügen. Auch Sicherheitsaspekte müssen berücksichtigt werden, damit sicheren Transaktionen gewährleistet werden können. Preise müssen richtig abgerechnet und Zahlungen korrekt durchgeführt werden. Alle diese Punkte müssen eingehalten werden, um ein funktionierendes Geschäft zu erhalten.

10.3 Business Model

Das Business Model beschreibt ein einfaches Ereignis innerhalb eines Wirtschaftssystems. Die einfachste Beziehung besteht zwischen dem Kunden und dem Verkäufer, die ein Geschäft über das Internet tätigen. In einem elektronischen Geschäft kann der Verkäufer als ESP (Electronic Commerce Service Provider) und der Kunde als Endkunde bezeichnet werden. Eine einfache End-zu-End Kommunikationsverbindung wie in Abbildung 10.7.

Der ESP offeriert Produkte, Inhalte und Dienstleistungen online übers World Wide Web (WWW) und repräsentiert den Verkäufer eines Gutes. Diese Produkte, welche vom ESP

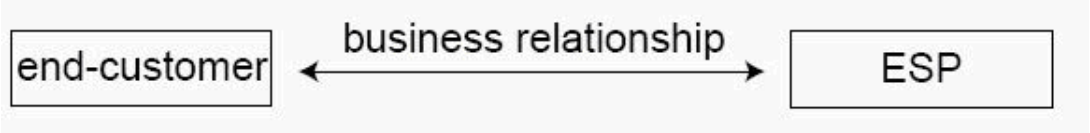


Abbildung 10.7: E-Commerce Business Beziehung [7]

offeriert werden, sind auch in normalen Geschäften erhältlich wie z. Bsp. Bücher, CDs, Autos usw. Auch nicht physikalische Produkte können vom ESP angeboten werden, wie elektronische Bücher und Hefte, die auf den heimischen Computer herunter geladen werden können. Ausserdem gibt es weitere Möglichkeiten wie Multimediainhalte z. Bsp. Video-On-Demand, Audio-On-Demand usw. Letztendlich kann der ESP seinem Kunden auch Backup-Speicherplatz zur Verfügung stellen, um so seine Daten von einem beliebigen Ort aus zu sichern. Diese Angebote des ESP können jederzeit und überall in der Welt über den Webbrowser bezogen werden.

Die Basis eines ökonomischen Systems, nämlich die Beziehung zwischen ESP und End-Kunde ist leicht und plausibel zu erklären. Schwierigkeiten können jedoch schon bei der Lieferung von Dienstleistungen eintreten. Performance- und QoS-Problemen sind Ursachen Nummer eins und können die Leistung verringern sowie Datenstaus verursachen. Die ISPs (Internet Service Providers) hingegen bieten physikalischen Infrastrukturen an. Auch technische Geräte wie Routers, Switches und Software Management Tools für die elektronischen Aktivitäten stellen sie bereit. Da die ISPs für den Datentransport zuständig sind, entstehen schnell einmal Engpässe bei starker Ausnutzung der Breitbandleitung. Deshalb müssen Transfergebühren erhoben werden um die Leistungsqualität bzw. den Leistungsdurchsatz zu sichern. Diese Gebühren müssen deshalb dem End-Kunden abhängig von der benutzten Leistung berechnet werden.

Diese Beziehung zwischen ESP, ISP und End-Kunden können durch ein E-Commerce Szenario dargestellt werden. In Abbildung 10.8 sieht man das Zusammenspiel zwischen ESP und End-Kunde. Dazwischen befinden sich die eingesetzten ISPs, die durchlaufen werden. Ausserdem kann eine eine dritte Partei hinzugezogen werden, die für die sichere Zahlung zuständig ist. Also ein Trusted Third Party (TTP) wie zum Beispiel eine Bank oder ein Kreditinstitut.

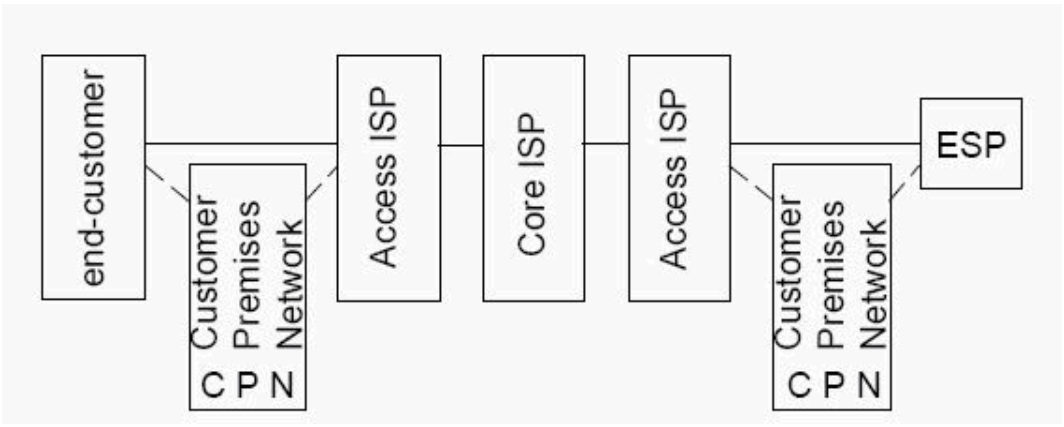


Abbildung 10.8: Beziehung zwischen den involvierten Parteien [7]

Unter den ISPs werden noch Unterscheidungen durchgeführt. Die einen sind die Access ISPs und die anderen die Core ISPs. Die Access ISPs unterstützen LANs (Local Access Networks) und die Internetverbindung zum Endkunden. Zwischendrin können direkt oder indirekt Transfers über das Customer Premises Network (CPN) gehen. Die CPN verbirgt den Endkunden im Internet. Somit bleibt der Endkunde gegen aussen anonym. Eingesetzt werden sie vor allem in Unternehmen und Universitäten. Ausserdem können diese CPN dem Endkunden zusätzliche Applikationen und Dienste zur Verfügung stellen. Die Core ISPs erhöhen die Erreichbarkeit der Access ISP und bilden das Rückgrat bzw. die Schnittstelle des Internet. Abhängig von der Reichweite können auch mehr als nur ein ISP bei der Übermittlung beteiligt sein.

10.4 Die Stufen eines Business Model

Um ein E-Commerce zu verwirklichen braucht es ein gutes Zusammenspiel sowie Kooperationen zwischen den beteiligten Parteien. Regeln und Geschäftskonditionen müssen dabei strikt eingehalten werden.

Dieser Business Prozess kann in vier Abschnitte aufgeteilt werden: Contracting Phase, Reservation Phase, Service Phase und Clearing Phase. In der folgenden Abbildung 10.9 sieht man die einzelnen Stufen.

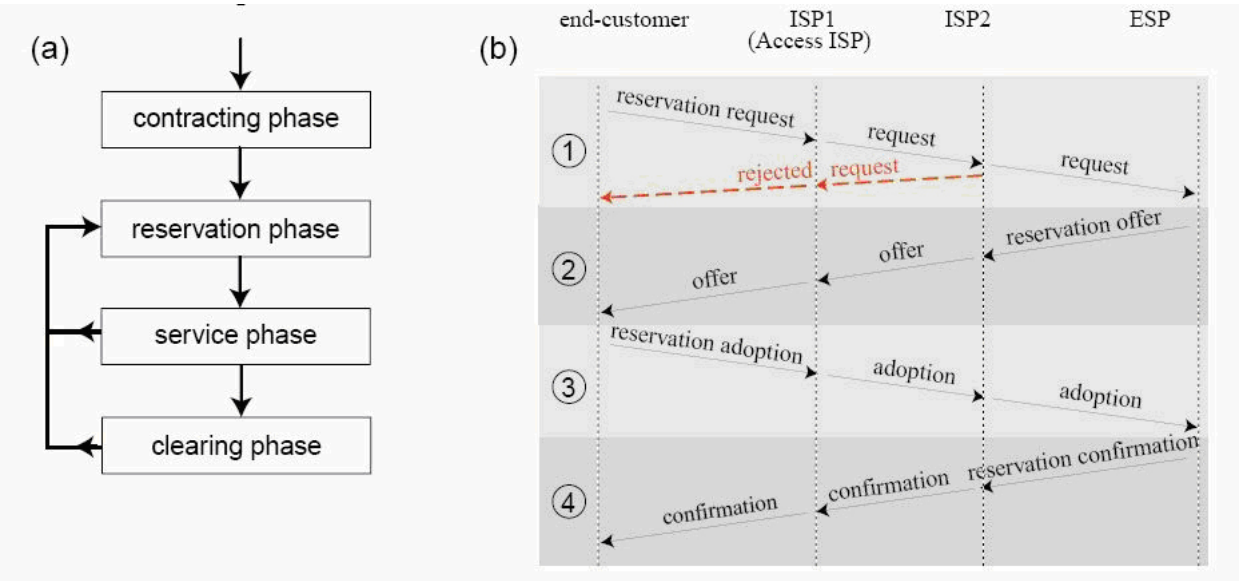


Abbildung 10.9: (a) Business Model in zeitlicher Anordnung (b) Reservationsmitteilungen zwischen den einzelnen Parteien [7]

10.5 Contracting Phase (Vertragsphase)

Die Vertragphase ist die Ausgangsphase für die Parteien, bevor ein Produkt oder Inhalt gekauft oder eine Dienstleistung durchgeführt wird. Zuerst müssen die kooperieren-

den Parteien sich besser kennen lernen. Demnach müssen zuerst Geschäftsverbindungen zwischen den Partnern hergestellt werden, um über eine allfällige Beibehaltung oder Erneuerung der Bedingungen zu entscheiden und Vertragsvereinbarungen für Geschäfte und Dienstleistungen auszuhandeln bzw. abzuschliessen.

Es gibt vier Geschäftsfälle die man unterscheiden kann, wobei die Verbindung über die CPN nicht berücksichtigt wird, also der Endkunde direkt in seiner Wohnung auf das Internet zugreift. Es existieren Beziehungen zwischen (1) Endkunden und ESP, (2) Endkunden und Accesss-ISP, (3) Access ISP und Core ISP und (4) Access ISP und ESP. Das Verhältnis zwischen dem Endkunden und ESP erfordert keine Verträge oder Vereinbarungen. Der Endkunden geht einfach auf die Homepage des ESP und bestellt sich das Produkt oder den Inhalt einer Dienstleistung. Der ganze Prozess könnte anonym durchgeführt werden, wobei der Endbenutzer keine persönlichen Informationen über seine Identität hinterlässt. Der ESP hat auch die Möglichkeit Kunden, die immer wieder bei ihm Produkte erwerben, registrieren zu lassen, damit sie beim nächsten Mal nicht alle persönlichen Daten erneut eingeben müssen. Ausserdem können die ESPs durch das Benutzerprofil der Endkunden, Aufzeichnungen über sein Kaufverhalten herstellen und ihm so ein für ihn zugeschnittenes Angebot an Produkten und Dienstleistungen offerieren.

Das Geschäftsverhältnis zwischen ESP und dem Access ISP ist sehr ähnlich wie das zwischen dem Endkunden und dem Access ISP. Der Hauptunterschied ist, dass der ESP dem Access ISP immer bekannt ist, weil beide fixe Orte haben. Wäre dies nicht der Fall, so würde die Geschäftsperformance nicht optimal sein oder sogar hinderlich wirken. Das Verhältnis zwischen dem Endkunden und dem Access ISP wird aufgebaut, sobald Daten zwischen und den Endkunden und dem ESP übertragen werden müssen. Dieses Verhältnis ist ganz speziell, da der Access ISP als Vermittler arbeitet, oder anders ausgedrückt eine dritte Partei darstellt, die dem Endkunden den Zugang zum Internet anbietet, ausserdem ist er selbst ein Kontaktpunkt über den jegliche Art von Datentransportdienste von und zum ESP stattfindet.

In einem Vertrag werden die Vereinbarungen zwischen dem Endkunden und dem Access ISP festgelegt und beinhalten alle nötigen und zukünftigen Geschäftskonditionen. Diese beinhalten ausserdem das Benutzerkonto des Endkunden für die Abrechnung und die Rechnungsstellung mit den dazugehörigen Informationen über die benutzten Dienste, wie monatlich verbrauchte Limiten oder spezielle Reservationen von Daten. Es besteht auch die Möglichkeit, dass der Endkunde aus einem bestimmten Grund unbekannt bleiben möchte, hierzu wird die Verhandlungsstufe übersprungen. Das Verhältnis zwischen Access ISP und Core ISP oder bzw. zwei generellen ISPs werden durch Vermittlung von Verkehrsverträgen und -SLAs gebildet, die auf dem Vertragsniveau die Menge des ankommenden und abgehenden Verkehrs durch das Netz eines ISPs regulieren.

Die ISPs bemühen sich um das aufrecht Erhalten einer dichten Netzinfrastruktur für ihre Kunden, um einen Datentransferdienst mit garantierten QoS zu sichern. Ebenfalls sollen statische SLAs zwischen den ISPs gebildet werden, bevor ein Datentransfer durchgeführt wird und gleichzeitig aber während des Betriebes dynamisch geändert und justiert werden kann, wenn sich gegebenenfalls die Netzsituation während der Service-Auslieferung ändert, oder der Endkunde diese Service-Parameter ändern möchte. Vertieftere Angaben zu SLAs können dem Kapitel „SLA und TLC“ entnommen werden.

10.6 Reservation Phase (Reservationsphase)

Hat der Endkunde entschieden, einen Inhalt oder eine Dienstleistung beim entsprechenden ESP zu erwerben, braucht er folglich einen Internetzugang um die Daten mit dem ESP auszutauschen. Innerhalb dieser Reservationsphase reserviert sich der Endkunde Netzwerkressourcen bei seinem Access ISP. In einer ersten Anmeldestufe beantragt der Endkunde bei seinem ISP die Übertragung der Daten. In der zweiten Stufe meldet wiederum sein ISP die Situation über die kurzzeitig verfügbaren Netzwerkverkehr, sowie Preisvergleiche anderer ISPs. Je nachdem, ob der Endkunde registriert oder anonym ist, kann diese Information zu seinem Vor - oder Nachteil sein.

In Märkten, wo Breitband- und Netzwerkressourcen angeboten und verkauft werden, können ebenfalls Auktionen eingesetzt werden, um die beste/billigste/kürzeste Verbindung zwischen dem Endkunden und dem ESP zu finden. Der Access ISP als Vermittler oder Vertragspartner für Endkunden ist für die Durchführung und Wartung der gelieferten Dienstleistungen verantwortlich. Verkehrsverträge und -SLAs wurden in die Vertragsstufe bereits aufgestellt, um den Datenverkehr durch die Netze der ISP zu organisieren. Die tatsächliche Reservierung eines Datentransportdienstes oder einer einfachen Preisanfrage findet innerhalb der Vermittlungsstufe, wie eine dritte Vorstufe der Reservationsstufe statt. Die Reservierung der Netzwerkressourcen für einzelne Datenflüsse auf dem Internet wird in vier Schritten entsprechend dem RSVP-Protokoll durchgeführt.

Im ersten Schritt sendet der Endkunde ein Signal zu seinem Access ISP aus um einen Reservationsantrag anzuzeigen. Dieser Antrag enthält Informationen über die Art der Reservation (zum Bsp. Anwendung, Zeitpunkt und Dauer), ein Schema über die Aufteilung der Kosten (Endkunde zahlt nur, nur ESP zahlt, beide beteiligen sich an den Kosten) und weitere Parameter, die das QoS für die Anwendung definieren. Diese Parameter können unter anderem die Bandbreite umfassen, welche die Leistung (bit/s) anzeigt, die Übergangsqualität (max. Störungs-Wahrscheinlichkeit, Verschlüsselung) und Parameter verlegen, oder die maximale verzögerte Zeit bis ein Datenpaket eintrifft. Der Antrag wird ausserdem über andere Core/Access ISPs (falls es anderes ISPs gibt), zum ESP geschickt, die mit einem Reservationsangebot antworten, wenn es den Antrag des Endkunden annimmt. Wenn ein ISP oder ESP den Antrag verweigert, kann dieser den Antrag zurückweisen und die Ablehnung zurück zum Endkunden senden. In diesem Fall beginnt die Vermittlung von neuem. In einem zweiten Schritt des Vermittlungsprozesses, antwortet der ESP auf den Reservationsantrag mit einem Reservationsangebot, das den Endkunden über die Annahme oder Ablehnung des Reservierungsantrags informiert. Das Reservierungsangebot wird auf die gleiche Weise zurück zum Endkunden geschickt, wo andere ISPs die erbetenen Ressourcen für diesen bestimmten Fluss reservieren, oder die Preisinformationen addieren, falls es nur ein Preisantrag war. Wenn ein Beteiligter das Reservierungsangebot nicht annimmt, muss der Endkunde den Reservierungsantrag mit den unterschiedlichen Parametern zurücksenden.

Der dritte und vierte Schritt der Vermittlungsphase (Reservierungsannahme und Bestätigung) sind wahlweise freigestellt und können verwendet werden um Absender oder Empfänger zu addieren, welche die elektronische Zahlungen zur Verfügung stellen, wenn zum Bsp. Reservierungsgebühren angefordert werden, oder Zahlungen im voraus geleistet wer-

den müssen. Diese vierstufige Vermittlungsphase muss für jeden erbetenen Dienst durchlaufen werden oder wenn der Endkunde die zur Zeit gebrauchte Dienstleistung verlängern möchte.

10.7 Service Phase (Dienstleistungsphase)

Nachdem der Datentransportdienst innerhalb der vorhergehenden Reservierungsphase aufgehoben worden ist, kann der tatsächliche Service innerhalb der Service-Phase durchgeführt werden. Der Datentransportdienst kann nun durch den Access ISPs und wenn nötig durch die einzelnen benötigten Core ISPs erfolgen.

Ein anderer wichtiger Punkt für die ISPs, ist die Zahlung der Dienste und ob der Endkunde schon bekannt ist und ihm folglich vertraut werden kann oder, ob er in einer früheren Phase entschieden hat, anonym zu bleiben. Falls der Endkunde mit seinem Access ISP im Vertrag registriert ist, erfolgt die Abrechnung und schließlich die Zahlung für den Dienst in der folgenden Clearingphase. Wenn der Endkunde ein anonymer Teilhaber ist, muss er vor dem Dienst, wegen fehlendem Vertrauen und Sicherheit, zuerst zahlen. Diese kann in Form von Zahlungen wie prepaid Konten, mit Schuldposten, Gutschrift oder irgendeiner prepaid Geldkarte oder durch irgendeine Art des elektronischen Geldes online geleistet werden. Abhängig von der Abrechnung müssen Endkunden und ESP ihre Anteile an den ISPs, bevor die zugeteilten Ressourcen für den reservierten Dienst freigegeben werden kann, zahlen. Nach der Anlieferung des Dienstes, kann der Endkunde zurück zu der Reservierungsphase gehen und um einen neuen Dienst bitten oder versuchen, den zurzeit durchgeführten Service mit ähnlichen Parametern QoS zu verlängern.

10.8 Clearing Phase (Rechnungsstellungsphase)

Die Clearing Phase folgt direkt auf die Service Phase in der Gesamtbetrachtung des Business Models. In dieser spezifischen Phase werden dem Kunden die von ihm in Anspruch genommenen Dienstleistungen verrechnet und durch diesen bezahlt. Während der Clearing Phase werden sämtliche vom Kunden beanspruchten Leistungen durch die Leistungserbringer beim Kunden in Rechnung gestellt. Dabei gibt es verschiedene Möglichkeiten, den Betrag für die erbrachten Leistungen beim Kunden einzufordern. Diese Möglichkeiten unterscheiden sich im Zeitpunkt der Verrechnung, als auch in den verschiedenen darin involvierten Personen und Dienstleistern. Auf die verschiedenen Verrechnungsmodelle und Zahlungssysteme möchten wir gerne in den folgenden Kapiteln genauer eingehen.

Die Clearing Phase gilt als beendet, wenn jede involvierte Partei für die entgegengebrachten Leistungen bezahlt wurde. Der Kunde kann wieder zurück zur ersten Phase des Business Models, der Reservationsphase um erneut einen Service zu reservieren.

10.8.1 Verrechnungsmodelle

Verrechnung über Access ISP

Bei der Verrechnung über den Access ISP ist der Access ISP des Kunden für die Fakturierung der verschiedenen Leistungen zuständig. Er nimmt die Rechnungen sämtlicher vom Kunden beanspruchter ISP's entgegen, summiert die verschiedenen Beträge auf, stellt sie in einer Gesamtrechnung zusammen und übergibt diese dem Kunden. Dieser muss den Betrag nun beim Access ISP begleichen. Der Access ISP behält den von ihm fakturierten Betrag, und leitet den Rest an die verschiedenen ISP's weiter. Diese Weiterleitung der Gelder kann auf zwei verschiedene Arten erfolgen. Entweder der Access ISP nimmt den Betrag des Kunden entgegen, behält den ihm geschuldeten Betrag und leitet den gesamten Restbetrag an den nächsten ISP weiter. Dieser behält den ihm zustehenden Betrag und leitet den Rest wiederum weiter. Ein solches Zahlungsmodell ist in Abbildung 10.10 illustriert. Eine andere Möglichkeit wäre, dass der Access ISP den Gesamtbetrag vom Kunden entgegennimmt und diese einzeln auf die beanspruchten ISPs verteilt. Abbildung 10.11 zeigt ein solches Zahlungsmodell. Schlussendlich erhält jeder in Anspruch genommene ISP, den vom Kunden geschuldete Betrag.

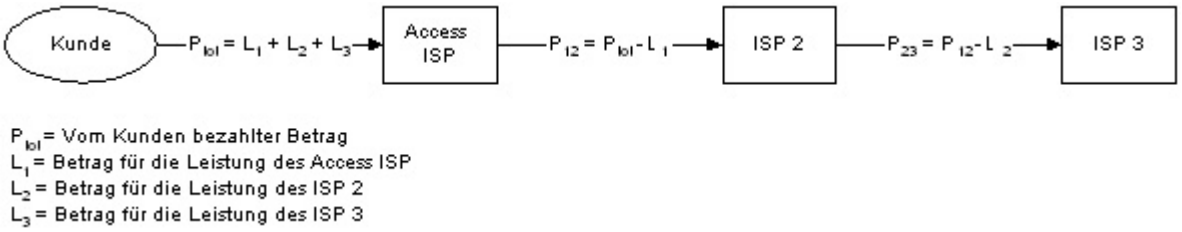


Abbildung 10.10: (a) Verrechnung über Access ISP [8]

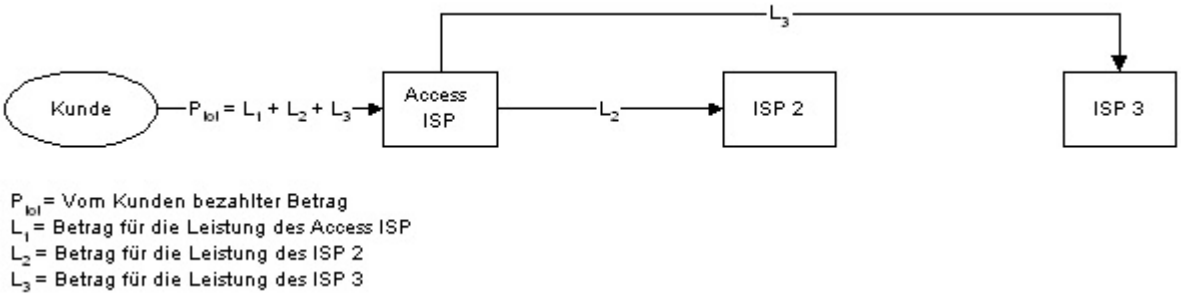


Abbildung 10.11: (b) Verrechnung über Access ISP [8]

Vorteile

Ein Vorteil ist ganz klar, dass der Kunde nur eine Rechnung zu begleichen hat, und somit der ganze administrative Aufwand bei einer Partei liegt. Der Access ISP ist für die korrekte Verrechnung der Leistungen verantwortlich. Zudem sind so sämtliche Daten, über die beanspruchten Leistungen bei einer Partei abgelegt, und dadurch im nachhinein einfach einsehbar.

Nachteile

Ein zentraler Nachteil, ist der massive Mehraufwand, der für den Access ISP entsteht. Er ist nicht nur für die korrekte Zusammenstellung der Kundenrechnung verantwortlich, sondern auch für die korrekte Weiterleitung des vom Kunden entgegengenommenen Geldbetrags. Der administrative Aufwand der dadurch entsteht ist nicht zu unterschätzen. Eine Möglichkeit für den Access ISP wäre es, diesen Mehraufwand dem Kunden weiterzuverrechnen. Was in der heutigen Zeit eigentlich auch gang und gäbe ist.

Verrechnung direkt bei den verschiedenen ISP

Bei der Verrechnung direkt durch die verschiedenen ISP's werden die vom Kunden beanspruchten Leistungen direkt durch die ISP's beim Kunden in Rechnung gestellt. Das heisst, dass der Kunde von jedem ISP deren Leistungen er in Anspruch genommen hat, eine Rechnung für diese kriegt, und diese direkt beim entsprechenden ISP zu begleichen hat. (Siehe Abbildung 10.12)

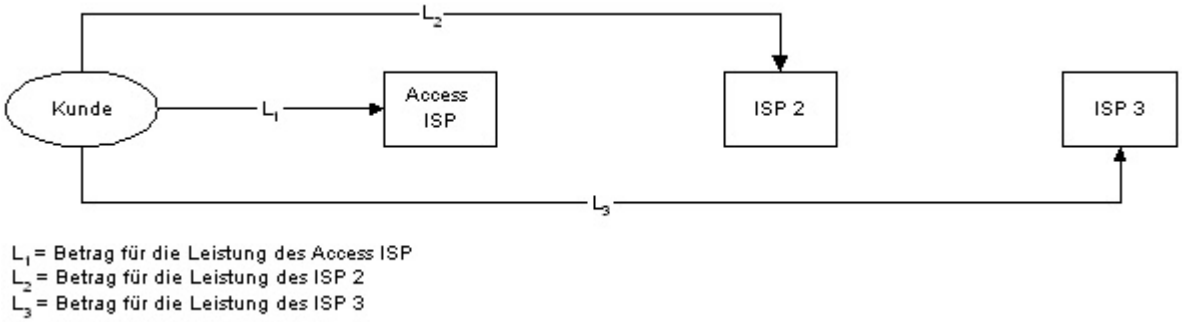


Abbildung 10.12: Verrechnung direkt bei den verschiedenen ISP [8]

Vorteile

Vorteile sind, dass der Kunde keinen Mittelmann hat und die Rechnung für die konsumierten Leistungen direkt beim tangierten ISP begleicht. So hat er Übersicht darüber welche ISP's für welche Leistungen wie viel Geld verlangen. Bei der Verrechnung über den Access ISP ist dem Kunden eine so detaillierte Einsicht in die Rechnungslegung unter Umständen untersagt. Weiterer Vorteil ist das Fehlen eines Mittelmannes, der zusätzliche Bearbeitungsgebühren verrechnen könnte. Zudem ist es immer ein Vorteil, wenn eine Person weniger existiert, der blind vertraut werden muss.

Nachteile

Der grosse Aufwand liegt jetzt nicht mehr beim Access ISP, sondern beim Kunden. Die übrigen ISP's müssen ihre Rechnung immer noch versenden. Ob sie diese dem Access

ISP oder direkt dem Kunden übergeben spielt keine Rolle. Das einzige was sich ändert ist die Verlagerung des Mehraufwands vom Access ISP zum Kunden. Da der Kunde nun sämtliche Rechnungen aller tangierten ISP's erhält, muss er diese nun alle einzeln bei den verschiedenen ISP begleichen. Ein weiterer Nachteil ist, dass falls er im Nachhinein gerne wissen würde, welche Leistungen er in einem bestimmten Zeitabschnitt in Anspruch genommen hat, die Daten nicht zentral bei einem Anbieter abgelegt sind, sondern er hat die angefallenen Rechnungen selbst archiviert oder muss bei jedem ISP einzeln nachfragen welche Leistungen er zu dem bestimmten Zeitpunkt beansprucht hat.

Verrechnung mittels Trusted Third Party (TTP)

Die Verrechnung mittels Trusted Third Party bezieht eine zusätzliche Geschäftspartei mit ein. Die TTP ist normalerweise eine unabhängige Institution deren tägliches Geschäft der Umgang mit Geld ist. Dieser TTP vertrauen alle in die Geschäftsaktivitäten involvierten Personen und über diese dritte Geschäftspartei wird die Fakturierung ausgeführt. Die TTP nimmt sämtliche monetären Ansprüche der ISP's entgegen, summiert diese auf und sendet sie in einer Rechnung an den Kunden. Dieser begleicht seine Schuld beim TTP und der TTP wiederum leitet die entgegengenommene Summe, entsprechend der entgegengebrachten Leistungen der ISP's, an diese weiter. (Siehe Abbildung 10.13)

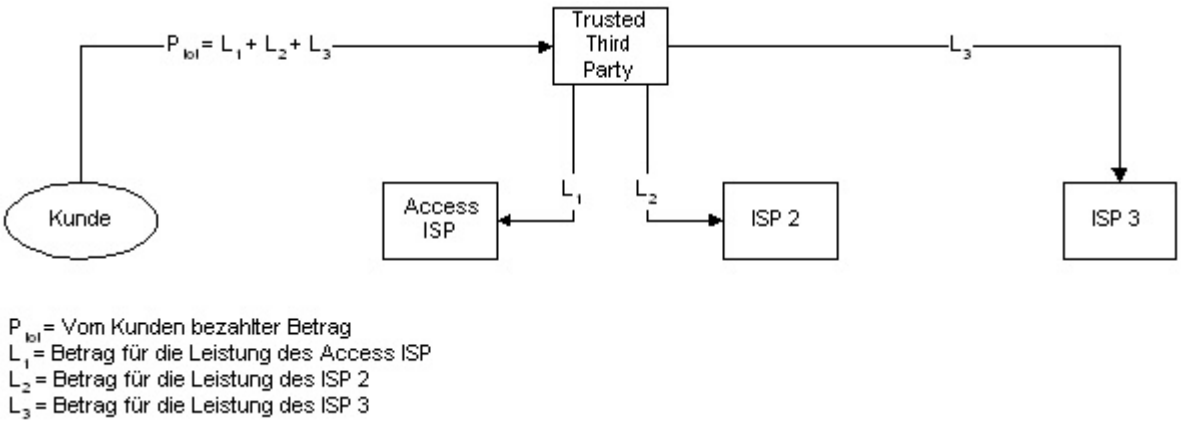


Abbildung 10.13: Verrechnung mittels Trusted Third Party [8]

Vorteile

Vorteile sind, dass sowohl ISP's als auch der Endkunde vom zusätzlichen Aufwand der Zusammenführung der einzelnen Rechnungen befreit sind. Zudem müssen weder ISP's noch Endkunde dafür sorgen, wer, wieviel von welchem Geld erhält. Sämtliche administrativen Aufgaben der Rechnungslegung werden vom TTP übernommen. Ein weiterer Vorteil ist, dass sich sämtliche involvierten Geschäftsparteien auf ihr Kerngeschäft konzentrieren können. Die ISP's können sich auf die Bereitstellung der vom Kunden benötigten Services kümmern, während sich die TTP auf ihre Kernkompetenzen, nämlich die Buchhaltung konzentrieren kann. Ein weiterer Vorteil ist die zentrale Ablage der Rechnungsdaten bei

der zuständigen Geschäftspartei, was einige Vorteile bei der späteren Einsicht archivierter Daten mit sich bringt.

Nachteile

Da die TTP keine Non-Profit-Organization ist, muss diese natürlich für ihren Arbeitsaufwand entschädigt werden. Dadurch entstehen Kosten, die entweder einzig von den ISP's oder dem Kunden getragen werden, oder die Kosten werden unter ihnen aufgeteilt. Ein weiterer Nachteil ist, dass eine weitere Partei in den Rechnungsstellungsablauf involviert wird. Dieser Partei muss natürlich vollumfänglich vertraut werden, was aber der TTP Möglichkeiten zur Hinterziehung von Geldern eröffnet.

10.8.2 Preismodelle

Es gibt verschiedene Ansätze bzw. Methoden von Verrechnungen und Preisen. Man unterscheidet generell zwischen fixen und variablen Preisen. Für den Kunden ist es extrem wichtig zu wissen, welche der beiden Arten ihm am besten bekommt.

Flat-rate Verrechnung ist sehr einfach und leicht zu implementieren. Dabei wird der Datenverkehr unabhängig vom bezogenen Volumen während einer Sitzung mit einem fixen Preis verrechnet. Solche Verrechnungsmethoden sind heute sehr verbreitet und Vorreiter des boomenden Internets. Jedoch hat dieses Model den Nachteil, dass er diejenigen Kunden bestraft, die darüber wenig Dienstleistung beziehen. Zum Beispiel geben heute viele Provider ihren Kunden einen Flat-rate ADSL Zugang zu einem fixen Monatsbetrag zur Verfügung. Wie auch immer ist die Effizienz dieses Ansatzes bei steigenden QoS-Nachfragen sehr schlecht. Deshalb ist diese Lösung ein schlechtes Schema für ein multi-Dienstleistungsnetzwerk, da sie das Netz zu stark belasten und eine schlechte Verzögerungszeit verursachen.

Andere Ansätze wie das **Volume-based** Preisschema, verrechnen dem Kunden nur den benötigten Datenverkehr pro Paket und/oder pro Reservationsgebühren. Je nach Provider können noch die Dauer einer Sitzung mit verrechnet werden. Der Vorteil dieses Schema ist, das der Kunde nur für das zahlt, was er effektiv auch bekommt. Ausserdem ist es eine gute Massnahme gegenüber Kunden, die wenige Dienstleistungen beziehen. Sie müssten ansonsten mehr zahlen. Studien zeigen auch, dass gewisse Kunden bereit wären zusätzliche Gebühren zu zahlen, wenn sie im Gegenzug bei einer QoS Nachfrage bevorzugt werden. Jedenfalls ist dieses Schema effizienter als die Flat-rate Methode. Ein Beispiel dazu ist der GPRS Dienst. Es wird hier nur für den Datentransport Gebühren erhoben.

Traffic-based Preismodel beruhen auf der Tatsache, dass die Bandbreite auf Verlangen für eine gewisse Zeit erhöht sowie eine höhere Priorisierung zugesichert werden kann und nach Gebrauch wieder herabgesetzt wird. Somit ist dieser Ansatz relativ gut geeignet für Video-On-Demand und ähnliche Dienste, die für eine gewisse Zeit höhere Datenverkehrsströme generieren. Der Nachteil liegt in der erhöhten Nachfrage, da dieser im vornherein

reserviert werden muss und so bei einer höheren Nutzerzahl, welche gleichzeitig die eine Brandbreiteerhöhung und bessere Prioriserung verlangen, zu einem Datenstau führen könnte.

Bei **Edge-Pricing** wird der Datenverkehr and den Rand der Domäne verschoben. Der Preis ist abhängig von den erwarteten Kapazitäten. Zuerst müssen zwei Vorbedingungen erfüllt werden bevor diese erwarteten Kapazitäten ermittelt werden können. Erstens wird ein Näherungswert pro Verbindung eines Tages und die QoS-Nachfrage berechnet. Zweitens muss die Kosten des aktuellen Weges durch einen erwarteten Weg ersetzt werden, wo die Gebühren nur noch von der Quelle(n) des Ziels des Datenflusses abhängig sind und nicht davon wo der Datenfluss durchgeflossen ist. Auf der Perspektive des Kunden ist es eine Anfrage von einem Punkt zu einem anderen. Der Weg der genommen wird, wird vom Netzwerk Router Algorithmus bestimmt. Wenn wir beide Näherungswerte kombinieren, ist der Preis durch die erwarten Ansammlung der Wege entlang der erwarteten Wege für das Quellpaket und das Ziel zu bestimmen. Deshalb kann der entstandene Preis bestimmt und die Gebühren lokal am Zugriffspunkt (zum Bsp. beim Provider, wo das Paket des Benutzers eintritt) erhoben werden. Dieses Schema wird auch lokales Edge-Pricing Schema genannt. [9]

10.8.3 Zahlungsmodelle

Bezahlen im Voraus

Oft wird der vom ISP angebotene Dienst vor der Inanspruchnahme bezahlt, wenn der Kunde anonym bleiben will. Das heisst er bezahlt im Voraus den Service, den er später benutzen wird. Eine solche Bezahlung im Voraus verlangt vom Kunden, die Kenntnis darüber wie oft, bzw. wie lange er den Service in Anspruch nehmen wird.

Basierend auf Anzahl beanspruchter Leistungen

Ein Tarifmodell bei dem der Kunde im Voraus bezahlt und dies anhand der Anzahl, wie oft er den Service in Anspruch nimmt, kann man mit einem Abonnement gleichsetzen. Der Kunde bezahlt dem ISP einen bestimmten Betrag, der wiederum dem Kunden eine bestimme Anzahl Zugriffe auf den von ihm bereitgestellten Service, freigibt. Ein Beispiel aus dem WWW ist die Seite www.livinghandy.de. Diese bietet die Möglichkeit einen Account einzurichten und sich so genannte „Prämientaler“ dazu zu verdienen, bzw. entgeltlich zu erwerben. Mit dem vorhandenen Betrag können dann die verschiedenen Dienste des Anbieters in Anspruch genommen werden, wie z. B. Klingeltöne, Games, Hintergrundbilder, etc. aufs Handy zu laden.

Basierend auf periodischen Zahlungen

Es besteht für den Kunden auch die Möglichkeit, den Service den er über eine bestimmte Zeitspanne in Anspruch nehmen wird, im Voraus zu bezahlen. Das heisst er bezahlt

Anfang Jahr, bzw. Monat dem ISP einen bestimmten Betrag, dadurch ist er berechtigt den Dienst während der vereinbarten Zeit zu nutzen. Werden die Zahlungen durch den Kunden eingestellt, wird ihm auch das Zugriffrecht auf den Service wieder entzogen. Ein Beispiel aus dem WWW ist die Seite www.best-price.ch. Auf dieser kann man für SFr. 5.00 einen Tagespass, für SFr. 10.00 einen Wochenpass und für Sfr. 30.00 einen Monatspass lösen, der es dem Kunden erlaubt sämtliche auf der Page eingetragenen Inserate, inklusive der dazugehörigen Kontaktdaten für die während der entsprechenden Zeit einzusehen.

Bezahlen im Nachhinein

Es kann auch vorkommen, dass der Kunde den Dienst in Anspruch nimmt und erst im Nachhinein bezahlt. Ist dies der Fall müssen dem Serviceanbieter, sichere Daten über den Kunden vorliegen. Denn wer bietet schon gerne einen Service an, der rege benutzt wird, aber dann niemand mehr da ist um ihn zu bezahlen. Deshalb ist es bei dieser Variante von Nöten, dass der Anbieter über die Identität des Kunden im Klaren ist, oder zumindest weiss wen er für die Benutzung des Dienstes belangen kann.

Basierend auf Anzahl in Anspruch genommener Leistungen

Bei dieser Methode wird nach einer bestimmten Zeitperiode, die Anzahl wie oft der Dienst in Anspruch genommen wurde, aufsummiert und dem Kunden dann die entsprechende Rechnung zugesandt. Das heisst der Kunde bezahlt zwar periodisch aber nicht für die gesamte Nutzungszeit sondern nur für die Anzahl wie oft er den Dienst effektiv genutzt hat.

Basierend auf Nutzungsperiode

Bei diesem Tarifmodell wird der Kunde im Nachhinein für die Nutzung eines Dienstes über eine bestimmte Zeitspanne berappt. Dies ist zum Beispiel der Fall bei einem Access Provider wie zum Beispiel der Bluewin oder der Cablecom. Man meldet sich für den von ihnen angebotenen Dienst an und bezahlt dann zum Beispiel jährlich die vom Anbieter für den entsprechenden Dienst vorgesehenen Kosten.

10.8.4 Zahlungssysteme

Für die verschieden auf dem Internet angebotenen Dienstleistungen stehen auch verschiedene Zahlungssysteme zur Verfügung. Bei der Auswahl des korrekten Zahlungssystems für die angebotene Dienstleistung sollte vor allem auf die anfallenden Kosten des gewählten Zahlungssystems geachtet werden. Das Zahlungssystem sollte wenn möglich keine allzu hohen Fixkosten in der Anschaffung aufwerfen, und auch keine allzu hohen Transaktionskosten generieren. Je geringer der zu verrechnende Betrag, desto geringer sollten auch die durch das Zahlungssystem anfallenden Kosten sein. Aufgrund der Beträge der Zahlungstransaktionen haben sich drei verschiedene Klassen von „Payments“ herauskristallisiert:

- **Micropayments:** 0.1¢ bis 0.1$
- **Small Payments:** über 0.1$
- **Macropayments:** über 10$

Die massgebenden Faktoren für die Entstehung der Transaktionskosten sind:

- Kommunikationsaufwand
- Rechenaufwand
- Speicherbedarf

Für den Kommunikationsaufwand verantwortlich sind die Initialisierung des Zahlungsprotokolls, die Transaktion der Zahlungsdaten, Verbindungen zu Dritten (Trusted Third Parties) und das Clearing mit der Bank. Der Rechenaufwand bemisst sich vor allem nach dem zeitlichen Aufwand für das Berechnen der kryptographischen Algorithmen (Signaturen, Chiffrieren, Zertifikate), aber auch für das Nachführen und das Clearing der Kontodaten. Der Speicherbedarf bezieht sich auf die Ablegung der Billing- und Accounting Daten und den digitalen Unterschriften in den Routern und Clearing-Servern.

Die Transaktionskosten können variiert werden, indem die aufgelisteten Kostentreiber angepasst werden. Je simpler zum Beispiel der kryptographische Algorithmus gehalten wird, desto kleiner wird der Rechenaufwand, und dies wiederum reduziert die Transaktionskosten. So können verschiedene Zahlungssysteme für die unterschiedlichen Klassen von Payments eingesetzt werden. [8]

Dies sollte aber nicht das einzige Kriterium zur Auswahl eines Zahlungssystems sein. Vielmehr sollten auch die Anforderungen des Kunden bei der Auswahl des Zahlungssystems berücksichtigt werden. Oft resultieren bei der Berücksichtigung aller Anforderungen sehr komplexe Systeme, welche sich nur für grössere Zahlungsbeträge eignen. Die meistgenannten Punkte sind:

- **Benutzerfreundlichkeit:** Die Bedienung des Zahlungssystems sollte für alle Teilnehmer möglichst einfach, komfortabel und transparent sein.
- **Anonymität:** Wahrung der Privatsphäre, Unverfolgbarkeit der Zahlungen, Einhaltung von Datenschutzgesetzen.
- **Geringes finanzielles Risiko:** Die Haftung bei Verlust des digitalen Geldes (wegen Ausfall der technischen Infrastruktur, Diebstahl oder Betrug) soll nicht (ausschliesslich) beim Kunden liegen.
- **Quittungen:** Transparenz der Transaktionen, Belegbarkeit im Falle von Streitigkeiten.
- **geringe Kosten:** kleine Fiskosten, tiefe Transaktionsgebühren.

- **Teilbarkeit:** Die Geldbeträge (electronic cash) können in kleinere Einheiten aufgeteilt werden.
- **Übertragbarkeit:** Die Geldbeträge können weiterverwendet werden und sind nicht an die ursprünglichen Besitzer gebunden.
- **Rückerstattung:** Nicht gebrauchtes elektronisches Geld kann zurückgegeben werden.
- **Verlusttoleranz:** Elektronisches Geld kann nach Verlust (oder Nichtgebrauch) zurückerstattet werden.
- **Stornierung:** Vollzogene Zahlungen können storniert werden.[8]

Im Folgenden werden wir ein bisschen genauer auf die meisteingesetzten Zahlungssysteme der heutigen Zeit eingehen.

Kreditkarte

Die Bezahlung von Internetdienstleistungen mittels Kreditkarte ist wohl die bewährteste und verbreitetste Möglichkeit. Bei der Bezahlung eines Dienstes mit der Kreditkarte, wird dem der den Service gewährleistet, die Kartennummer und die persönlichen Daten mitgeteilt. Der Betrag wird bei der Bank abgebucht, und dem Kunden am Ende des Monats in Rechnung gestellt.

Banküberweisung

Bei der Banküberweisung schickt der Serviceleistende ISP dem Kunden einen Einzahlungsschein, mit dem dieser dann über Online-Banking oder den Postschalter den geschuldeten Betrag begleichen kann.

Lastschriftverfahren

Im Gegensatz zur Überweisung wird der Zahlungsvorgang bei der Lastschrift nicht vom Kunden sondern vom ISP ausgelöst. Neben dem Kunden und dem ISP sind die Bank des Kunden und die Bank des ISP beteiligt. Der ISP erteilt der Bank des Kunden den Auftrag zum Einzug der Lastschriften. Dies wird auch als Lastschrifteinreichung, der ISP dementsprechend als Lastschrifteinreicher bezeichnet.[10]

PayPal

PayPal ist ein von der Firma eBay betriebenes Micopayment-System. Er stellt weltweit den grössten Online-Zahlungsdienstleiser dar. Ziel von PayPal ist, Überweisungen so einfach zu gestalten wie das versenden einer E-Mail. PayPal-Mitglieder können so Geld an jede beliebige Person in den unterstützten Ländern senden, die über eine E-Mail-Adresse verfügt. Um bei PayPal Geld an einen anderen Teilnehmer zu senden, gibt es drei Möglichkeiten:

- Man kann Geld direkt von dem PayPal-Kontoguthaben ausgehend versenden. Das Konto kann über Kreditkarte oder via Überweisung aufgeladen werden.
- Man kann, um eine Zahlung zu tätigen, (direkt) seine Kreditkarte verwenden. Hierzu muss das Geld nicht erst auf das Konto eingezahlt werden, sondern wird sofort dem Empfänger gutgeschrieben. Diese Möglichkeit besteht allerdings nur für Premium- und Businesskontoinhaber.
- Weiterhin ist es möglich, Zahlungen über das Lastschriftverfahren direkt vom eigenen Konto aus zu tätigen.[11]

Easyp@y

Um ein Produkt oder eine Dienstleistung im Internet mit Easypay zu bezahlen, wählt der Kunde in den beteiligten Internet-Shops die Zahlungsvariante Easypay und gibt im vorgesehenen Fenster der Website des Anbieters seine Easypay-Nummer ein, bestätigt mit „ok“. Der Kaufpreis wird vom aktuellen Kartensaldo abgezogen. Bei dieser einfachen Zahlungsabwicklung ist eine hohe Sicherheit gewährleistet: Es werden keine persönliche Daten wie Name, Kreditkarten- oder Kontonummern hinterlassen. Ebenso einfach und sicher ist die Bezahlung mit dem Handy. Eine einmalige Registrierung genügt. Angeboten wird Easypay in Werten zu CHF 25, CHF 50 und CHF 75. [12]

Paybox

Eine weitere Möglichkeit ist die Verrechnung eines in Anspruch genommenen Services über die Handyrechnung des Kunden. Diese Möglichkeit wird von Paybox gewährleistet. Als Payboxer gibt man einfach seine Handynummer oder Wunschnummer an. Paybox ruft die angegebene Nummer an, nennt Betrag und Empfänger und bittet um Bestätigung der Zahlung. Durch Eingabe der vierstelligen Paybox PIN auf Ihrem Handy bestätigen Sie die Zahlung, und der Betrag wird bequem von Ihrem Bankkonto - und nicht von Ihrer Handyrechnung - abgebucht. Dieser Dienst ist vor allem in Oesterreich verbreitet, stösst aber langsam nach Deutschland vor und wird die Schweiz in nächster Zeit höchstwahrscheinlich auch einnehmen. [13]

10.9 Cumulus Pricing Scheme (CPS)

Das Cumulus Pricing Scheme (CPS) basiert auf einem Flat-Rate-Schema, das heisst, die vom Kunden beanspruchten Leistungen werden ihm monatlich in Rechnung gestellt. Das Schema bietet einen Feedback-Mechanismus, der es erlaubt, die Marktkräfte zu berücksichtigen, sprich man kann das Verhalten der Kunden und deren Bedürfnisse analysieren. Zudem bietet das Schema eine immense Menge an Grundvoraussetzungen, für das Einbinden allfälliger Mess- und Verrechnungsmechanismen.

Anfänglich wird ein Vertrag zwischen dem Kunden und dem Anbieter der Leistung ausgehandelt, basierend auf einer Schätzung des Kunden, über die benötigten Leistungen. Dazu kommt ein neuartiger Feedback-Mechanismus, der es dem Anbieter der Leistung erlaubt über verschiedene Zeitspannen hinweg das Verhalten des Kunden zu protokollieren und allenfalls Voraussagen zu tätigen. In der kurzfristigen Sicht werden die vom Kunden beanspruchten Leistungen protokolliert, um dem Anbieter die Möglichkeit zu geben eine mittelfristige Prognose über die zukünftig vom Kunden in Anspruch zu nehmenden Leistungen zu tätigen. Diese kurzfristige Protokollierung von Leistungen wird mittels so genannter Cumulus Punkte vorgenommen. Die heutigen Modelle der dynamischen Preisbildung funktionieren so, dass bei einer Überbelastung des zugrunde liegenden Netzwerkes, die Preise für den angebotenen Service unmittelbar steigen. Dies führt dazu, dass die Kunden ihren Verkehr anpassen und der Anbieter den Andrang bewerkstelligen kann. Dieser Ansatz benötigt aber eine immense Flexibilität, sowohl auf Seiten des Kunden als auch auf Seiten des Anbieters.

Die Cumulus Punkte bieten hingegen einen Feedback-Mechanismus, der nicht direkt auf die Veränderungen reagiert. Als erstes wird ein Vertrag zwischen Kunde und Anbieter ausgearbeitet, der den Kundenanforderungen gerecht wird, und diese in Form von maximaler Bandbreite, Verzögerung etc. definiert. Zudem wird der monatlich zu leistende Preis für die Leistung definiert. Gemäss des Vertrages werden nun, die angebotene Leistung des Anbieters vom Kunden in Anspruch genommen. Es kann aber sein, dass die Voraussagen des Kunden über die Höhe der beanspruchten Leistungen nicht ganz korrekt waren. Sobald diese Diskrepanzen zwischen vorausgesagtem und tatsächlichem Verkehr einen gewissen Schwellwert übersteigen, erhält der Kunde ein Feedback betreffend der akkumulierten Cumulus Punkte. Diese existieren als grüne oder rote Flags: ein roter Cumulus Punkt indiziert, dass der Kunde seine Kapazitäten überstrapaziert hat, während ein grünes Flag darauf hinweist, dass der Kunde seine Ressourcen nicht ganz ausgenutzt hat. Je grösser die Diskrepanz zwischen der vertraglichen Vereinbarung und der der Realität, desto mehr Cumulus Punkte werden generiert. Die Cumulus Punkte behalten ihre Gültigkeit über eine definierte Anzahl von aufeinander folgenden Rechnungsperioden. Falls die Summe der gesammelten Cumulus Punkte einen gewissen Schwellwert erreicht, werden vom ISP gewisse Anpassungen des Vertrages vorgenommen, um diesen den veränderten Umständen anzupassen.

Abbildung 10.14 zeigt ein typisches Beispiel wie Cumulus Punkte zum Einsatz kommen. Der Kunde hat seinen erwarteten Bandbreitengebrauch mit x MB/s angegeben, doch die angegebene Bandbreite wird vom Kunden überzogen, dies sowohl im Januar, als auch massiv im Februar. Das bringt dem Kunden einen roten Cumulus Punkt im Januar und

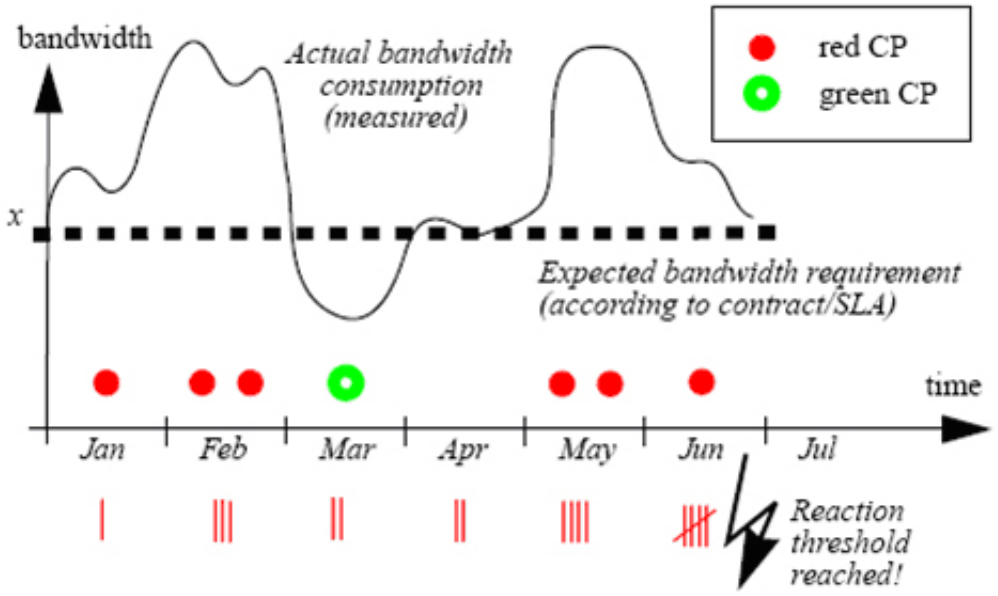


Abbildung 10.14: Rote und Grüne Cumulus Punkte und deren Kummulation über die Zeit hinweg

zwei rote im Februar ein. Danach fällt die benutzte Bandbreite unter den vorausgesagten Wert, wodurch der Kunde einen grünen Cumulus Punkt im März verdient. Im April wird genau die vereinbarte Bandbreite beansprucht. Im Mai und Juni wird die vertraglich vereinbarte Bandbreite erneut massiv überschritten, was dem Kunden erneut 3 rote Cumulus Punkte einbringt. Dies führt dazu, dass Ende Juni der Schwellwert von fünf roten Cumulus Punkten erreicht ist, was zu einer erneuten Verhandlung der vertraglich vereinbarten Bandbreitennutzung zwischen Kunden und Anbieter führt.

Eigenschaften des Cumulus Pricing Scheme:

- **Diskret:** Anstatt jede noch so kleine Fluktuation im Kundenverhalten zu protokollieren, erlaubt das CPS dem Kunden ein fluktuieren zwischen gewissen Grenzen. Das Feedback wird in einer quantifizierten Form gegeben.
- **Kumulativ:** Nicht einzelne Ausbrüche aus den prognostizierten Werten, veranlassen eine Reaktion, sondern kontinuierliche Überschreitungen.
- **Frühwarnung:** Falls einige Kunden ihre Verhaltensmuster ändern, wird dies keine unmittelbaren Konsequenzen haben, doch diese erfahren die Veränderung in einer sehr frühen Phase.
- **Vorhersehbar und transparent:** Die Kosten bleiben stabil über den langfristigen Zeithorizont hinweg, und notwendige Änderungen sind transparent für den Kunden, aufgrund des frühzeitigen Warnsystems des Feedbackmechanismus.
- **Flexibilität der Parameter:** Die Parameter über das Warnsystem lassen sich beliebig anpassen, und bieten somit dem Anbieter uneingeschränkte Flexibilität in der Festlegung der zugrunde liegenden Parameter.

- **Marktgleichgewicht:** Durch die vertraglichen Anpassungen, die zwischen dem Anbieter und Kunden vorgenommen werden, kann dem Kunden den für ihn optimalen Service zum optimalen Preis angeboten werden.[14]

10.10 Die Verrechnung von Dienstgüte in Transportnetzwerken

Diffserv basierte Dienste haben kein „flat rate“ Servicemodel und basieren auch nicht auf verbindungsorientierten Diensten. Aus diesem Grund sind weder fixe Zugangs- noch Verbindungsgebühren angebracht. Die Autoren schlagen ein Verrechnungsmodel vor, welches auf Intervallen beruht. Der Kunde wählt beim Vertragabschluss mit seinem Provider ein „traffic profile“, welches seiner spezifischen Serviceklasse entspricht. Die Auswahl dieses „traffic profiles“ erfolgt anhand der verschiedenen Preise, und wird für jedes Intervall erneut ausgewählt. Aufgrund des ausgewählten „traffic profiles“ erhält der Kunde bereits einen Anhaltspunkt der Kosten, welche er später zu bezahlen hat. Innerhalb des Intervalls wird ein nutzenbasierter Ansatz verwendet, der sich aber lediglich auf die Abweichungen vom gewählten „traffic profile“ beschränkt. Dieser Ansatz hat für den Provider den Vorteil, dass weniger Speicherplatz für das überwachen (monitoring) der Daten benötigt wird, da nicht der gesamte Datenstrom überwacht werden muss, sondern nur derjenige Teil, der oberhalb oder unterhalb des gewählten „traffic profiles“ liegt.

Für den Kunden liegt der Vorteil in der Verrechnung der effektiv genutzten Dienstgüte. Am Ende jedes Intervalls werden die Kosten für das gewählten „traffic profiles“ um die Kosten für die genutzten Kapazitäten oberhalb des Profile erhöht und für jene unterhalb des Profile verringert.

Durch dieses Vorgehen werden Abweichungen des Benutzerverhaltens mit berücksichtigt, ohne für den Kunden finanzielle Nachteile zu bringen. Die Lücke zwischen dem was bezahlt wird und dem was effektive genutzt wird, kann auf diese Art verkleinert werden. [15]

10.11 Schlusswort

In der heutigen Zeit ist das Netzwerk einem starken Zuwachs durch neue User ausgeliefert. Immer mehr Nutzer müssen sich die vorhandenen Bandbreite teilen. Die Folgen sind schlechte Latenz, Datenstaus, Paketverluste usw. Die Dienstgüte kann nicht mehr gewährleistet werden. Darunter leidet auch die Infrastruktur, welche langfristig dem ständigen Zuwachs angepasst werden muss, was nicht immer möglich ist. Die Kosten eines Ausbaus können sehr hoch ausfallen und die Mittel stehen nicht immer zur Verfügung. Somit bleibt oftmals nichts anderes übrig, als nach anderen Methoden einer Regulierung zu suchen.

DiffServ bietet in dieser Hinsicht mögliche Lösungen. Provider können zukünftig eine gewisse Dienstgüte anbieten und es sind neue Verrechnungsmodelle entstanden, welche diese Dienstgüte in die Abrechnung mit einbeziehen.

Immer mehr Benutzer sind bereit für eine bessere Dienstgüte mehr zu bezahlen. Die entstehenden Kosten können also von ISP in einem gewissen Grad dem Kunden belastet werden. Dieser erhält im Gegenzug eine zugesicherte Anzahl von Ressourcen und einen höheren Grad an Dienstgüte.

Der Vorteil von DiffServ liegt sicherlich in seiner Skalierbarkeit. Eine Implementierung ist erheblich einfacher als mit IntServ, weil nicht jeder Datenfluss verwaltet werden muss und eine aufwendige Signalisierung in jedem Router innerhalb der Domain entfällt. Man darf allerdings nicht vergessen, dass die Dienstgüte bei DiffServ nur innerhalb der entsprechenden Domain sichergestellt werden kann.

Obwohl DiffServ momentan noch in wenigen Anwendungen zu finden ist, hat dieser Ansatz für die Zukunft, speziell bei den VoIP Anwendungen ein nicht zu unterschätzendes Potenzial.

Literaturverzeichnis

[1] U.Thürmann - IPTEL2000 Workshop, 2000.

[2] Differented Services im Internet. http://www3.informatik.uni-wuerzburg.de, Recherchiert im Dezember 2005.

[3] Token Bucket. http://en.wikipedia.org/wiki/Token_bucket, Einsicht Dezember 2005.

[4] Random early detection. http://en.wikipedia.org/wiki/Random_early_detection, Recherchiert im Dezember 2005.

[5] Diff-Serv Network Elements. http://www.linktionary.com/d/diffserv.html, Recherchiert im Dezember 2005.

[6] Technische Grundlagen, Quality of Service in IP-Netzen. http://www.hfs.e-technik.tu-muenchen.de/ext/d12/qos_in_ip.pdf, Recherchiert im Dezember 2005.

[7] H. Kneer, U. Zurfluh, G. Dermler, B. Stiller: A Business Model for Charging and Accounting of Internet Services. Lecture Notes in Computer Science, Publisher: Springer-Verlag GmbH, 2000.

[8] Urs Kaiser: Sicherheits- und Kostenaspekte in elektronischen Zahlungssystemen für RSVP. Semesterarbeit April bis August 1998.

[9] Liang Ji, Theodoros N. Arvantis, Sandra I. Woolley: A New Charging Scheme for Multi-Domain DiffServ Networks. Global Telecommunications Conference, 2003. GLOBECOM'03.

[10] Lastschrift: http://de.wikipedia.org/wiki/Lastschriftverfahren, Einsicht Dezember 2005.

[11] PayPal: http://de.wikipedia.org/wiki/Paypal, Recherchiert im Dezember 2005.

[12] Einfach zahlen mit Easyp@y, Erstellt am 2. Juni 2004. http://www.swisscom.com/GHQ/content/Media/Medienmitteilungen/2004/20040602-01-EasyPay.htm.

[13] Paybox - ab sofort bezahlt ihr Handy. http://www.paybox.at/, Recherchiert im Dezember 2005.

[14] P. Reichl, P. Flury, J. Gerke, B. Stiller: How to Overcome the Feasibility Problem for Tariffing Internet Services: The Cumulus Pricing Scheme. Proceedings of IEEE ICC, 2001.

[15] C. Bouras, A. Sevasti. Pricing QoS over transport networks. Internet Research, Volume 14, Nr.2 - 2004.

Chapter 11

Grid Services and their Market Potentials

Sibylle Grimm

This paper discusses the basics of grid computing and the relating ideas and technologies as well as the economical aspects of grid computing. Grid computing evolved from meta-computing and distributed computing. It enables worldwide sharing of different resources like disk space, CPU, databases, software applications e.g. The Global Grid Forum suggests standards that should be used when generating a grid so that all grids use the same standards and can therefore deal with heterogeneous systems. This paper discusses the ideas, the technical background and the economical aspects of a grid system. From the economical point of view, grids have diverse potentials. Less hardware is needed as in case of any bottleneck additional ones can be rented. A smaller set of hardware causes less maintenance and consequently less costs. The hardest problem grids face is the allocation of resources. Currently, this problem is solved via different pricing strategies like the commodity market model, several auction systems, bid-based sharing models etc.

Contents

11.1 Introduction

In industries as well as in everyday life, users running an application on the computer often encounter limits of capacity or other resources. Reaching this point, they wish to expand the computer's capacity to run the application properly. But as these resources are very expensive and not needed everyday, it is not worth upgrading the computer. Instead, it is more appreciated to have the possibility to "'rent"' additional resources for the time they are needed. On this way a lot of money can be saved. That is what a grid is going to offer: the possibility to get access to additional resources whenever needed and paying it only as long as the resource is used.

The introduction part concentrates on the grid itself. The first subsection presents a general and abstract definition of a grid. The second part of the introduction explains the grid a bit more detailed by listing the capabilities of a grid and the third part of this chapter refers shortly to the history and development of grids. Section 11.2 explains the underlying technologies and Section 11.3 concentrates on the economical aspects. The last chapter then focuses on the potentials of a grid.

11.1.1 General Definition of a Grid

What is a grid? There exist many different definitions. They all have some main points in common. According to these definitions, a grid is a form of distributed computing and provides a service for sharing resources over the internet. These resources are computer power, storage capacity but also databases or software applications. [20] defined a grid in their paper as follows (p. 1): "'a Grid is a collection of distributed computing resources available over a local or wide area network that appear to an end user or application as one large virtual computing system. The vision is to create virtual dynamic organizations through secure, coordinated resource sharing among individuals, institutions, and resources. Grid computing is an approach to distributed computing that spans not only locations but also organizations, machine architectures and software boundaries to provide unlimited power, collaboration and information access to everyone connected to a Grid."'

11.1.2 Benefits of the Grid

A grid has a lot of capabilities whereof the most important ones are mentioned in this section. After this section it should be clear what a grid is and what it can be used for.

Exploiting underutilized resources The simplest thing a grid can do is running an application on a different machine. This might be necessary if the machine usually used for this application is busy with a more important task or other reasons. Then the mentioned application can be run on an idle machine somewhere in the grid. Two conditions need to be fulfilled for doing so. First of all, the application in question must be executable

remotely and second, the idle machine has to have the relating hardware, software or resources needed to be able to run the application.

To find these idle machines, grid computing offers a framework for exploiting the unused resources. So this framework helps to increase the efficiency of resource usage.

Parallel CPU capacity According to [13], the potential for high parallel CPU capacity is one of the most practical tools of the grid. Applications use algorithms that can be split in independently running parts. So a CPU intensive application can then be partitioned in smaller so called "'subjobs"' that are executed on different machines in the grid. Using this technology, a lot of time can be saved. If an application is perfectly scalable, it will for example be 5 times faster if 5 processors are used for executing the application. There exist some barriers to this perfect scalability. If an algorithm that is partitioning the application can only be split in a limited number of independently running parts, than this is a barrier. Sometimes the parts are not completely independent and can therefore not be run on different machines.

Virtual resources and virtual organizations for collaboration Grid provides an environment for collaboration for a much wider audience than for example distributed computing. As grid computing offers important standards, it is possible to connect heterogeneous systems and on this way forming a large virtual system. A user can be a member of several real and virtual organizations.

Access to additional resources In a grid, not only data can be shared (the so called "'data grid"') but also equipment, software, services, licenses etc. Therefore, a grid can be used to get access to special devices that are too expensive or too hardly needed to be worth buying them.

Resource balancing The grid merges the resources from different machines to one extensive virtual resource. The grid then enables the resource balancing by sending jobs to underutilized resources as shown in Figure 11.1.

This equalisation of resource using can be done in two ways ([2] p. 5):

1. An unexpected peak can be routed to relatively idle machines in the grid.

2. If the grid is already fully utilized, the lowest priority work being performed on the grid can be temporarily suspended or even cancelled and performed again later to make room for the higher priority work.

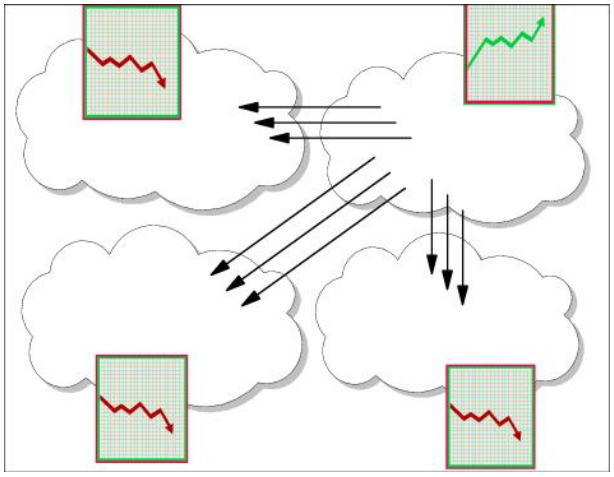


Figure 11.1: Resource balancing [13] p. 13

Reliability Today, computing systems are often equipped with more hardware than actually needed to increase reliability. They have for example double processors so that if one fails it can be replaced without turning the other off. Also the power supplies and the cooling systems usually are doubled. Even though this solution makes the system more reliable it is nevertheless very expensive. Therefore, new solutions are needed. One alternate way tries to use software to improve the system's reliability. So if there is any failure, the grid software automatically resubmits the job to another machine. If an application is critical (for example a real-time application), several copies of this job are run on different machines in the grid.

Management Grids ease the priority management among different jobs. Thanks to the larger view a grid offers, it is easier to control and manage several projects. If one project needs more resources and an other has idle ones, the grid runs jobs from the first project on the resources of the second. The administrators of a grid have a good overview over all the running applications and therefore can in any case of a bottleneck adjust the policies to achieve a better resource allocation.

11.1.3 History and Development

Grid has its name from the electrical power grid that is persuasive, easy to use and reliable. The grid's direct ancestor is metacomputing that describes projects aiming to interconnect supercomputers to combine the processing power of multiple supercomputers. The grid itself was born at a workshop in 1997 at Argonne National Laboratory. The motivation for this grid were resource intensive applications in the science that required more resources than one single computer could offer. One year later the first book "'The Grid: Blueprint for a New Computing Infrastructure"' has been published by Ian Foster and Carl Kesselmann [5], the two pioneers in grid computing. Since then, researchers are improving and expanding the grid technology almost every day.

11.2 The Grid Technology

This section explains the key technologies needed for a grid.

11.2.1 The Five Big Ideas

The grid technology mainly consists of five central ideas, that are discussed in this section.

Sharing of Resources The main goal of a grid is to share resources among different users. The idea is, that if one person has idle resources and an other user temporarly needs more resources than he owns, he can use the spare ones on another computer. This resource sharing is not a sipmle file exchange. Moreover it offers direct access to remote software, computers, data and storage space. It can even allow access and control of remote sensors, telescopes and other devices.

The fact that an application can - trough the grid technology - use resources from different users implies the problem, that - based on their different hardware, operating systems and software - the users have unequal security and access control policies. For this reason the users need to trust each other and if one of them is causing a lot of trouble, the others can manage to get him out of the grid community.

Security As a consequence of the different systems (see above), secure access is needed. This includes mainly three things: Access policy, Authentication and Authorization.

- Access policy: resource providers and users must define what is shared, who is allowed to share, and the conditions under which sharing occurs
- Authentication: a mechanism for verifying the identity of a user or resource is needed
- Authorization: a mechanism for determining whether an operation is consistent with the defined sharing relationships or not is essential

Resource Use The efficient use of resources is the third of the five big ideas behind the grid. There are always users offering resources and other ones consuming these idle resources. Using a mechanism that automatically allocates work to the offered resources, queues can be reduced and the resources are used efficiently.

The Death of Distance The next idea is that, using a grid, distance does not matter anymore. Due to today's high speed connections, even worldwide grids become possible. Some years ago it would have taken too much time to send data around the globe to get it processed somewhere else. But with today's technologies, using a grid saves a lot of time an nobody needs to care about the distance anymore.

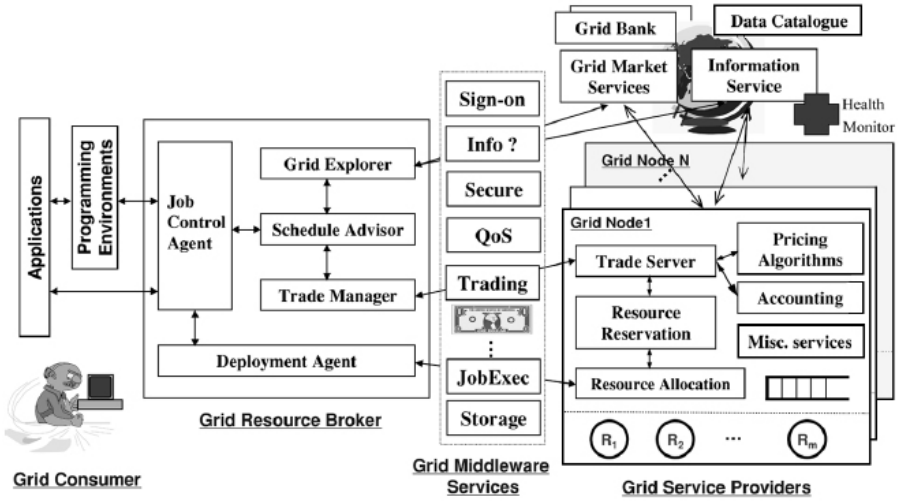


Figure 11.2: A Generic Grid architecture for computational economy (GRACE) [4] p. 702

Open Standards Open standards is the fifth key point in a grid. These allow an application of the grid to be run on another grid. If every company and grid has its own standard, the worldwide networking cannot be done anymore. As usual in discussions about standards, also in the grid technology the question about which standard to use arises. The Global Grid Forum developed grid-specific standards that should be used for every existing grid. At the moment it seems that the OGSA (Open-Grid Services Architecture) is going to be the key standard for grid computing. More about OGSA and Standards see Section 11.2.4.

11.2.2 Architecture

Figure 11.2 shows a distributed grid architecture for computational economy. It has several key components:

- Grid User with Applications
- Programming Environments
- User-Level Middleware and Tools (such as Grid Resource Brokers (GRBs))
- Core Grid Middleware
- Grid Service Providers (GSPs)

The Generic grid architecture for computational economy (GRACE) has services that support both, the resource users and the resource owners. The resource providers can use the mechanisms of GRACE to offer their resources in the grid and charging the resource consumers. The resource consumers interact with GRACE by defining the resources needed via the resource broker which tries to find a suitable resource to the cheapest price possible.

The core parts of this architecture are the Grid Resource Brokers, the Grid Middleware and the Grid Service Providers that are therefore discussed more detailed:

Grid Resource Brokers

The Grid Resource Broker (GRB) is a mediator between the user (consumer) and the resources offered. Its tasks are ([4] p. 702): "'It is responsible for resource discovery, resource selection, binding of software, data, and hardware resources, initiation computations, adapting to the changes in Grid resources and presenting the Grid to the user as a single, unified resource.'" As shown in Figure 11.2, the GRB consists of the following components:

- Job Control Agent: This is a control engine that leads a job through the system. It coordinates the schedule generation (with the schedule advisor), handles the creation of jobs, maintains the status of the jobs, interacts with the clients, the schedule advisor, and the dispatcher.
- Schedule Advisor: Has to discover resources offered, to select resources and to make sure that the selected resources meet user requirements.
- Grid Explorer: It discovers new resources, identifies the list of authorized machines and knows the resource's actual status information.
- Trade Manager: Identifies resource access costs and trades with Grid Service Providers (GSPs).
- Deployment Agent: Activates task execution and updates the status of the task exection regularly to send a report to the Job Control Agent.

More about the GRB follows in Section 11.2.3.

Grid Middleware

The grid middleware provides the core technologies to ensure secure and uniform access to the resources. These technologies are: security, single sign-on, remote process management, storage access, data management and information services. Usually these services are standardized. This allows the use of service-oriented architectures.

Grid Service Providers

The Grid Service Provider (GSP) is mainly in interaction with the following components:

- Grid Market Directory: It publishes the resources of the providers

- Grid Trade Server: This is the agent working for the resource provider and acts as a mediator between the resource owners and users. Its goal is to maximize the provider's profit and it's task is to direct the accounting system to charge the user the correct price. Therefore trade server needs to inform the accounting system about the agreed pricing strategy.

- Pricing Policies: There exist different ways to charge the user. Depending on different variables the pricing strategy of a resource provider can change depending on the market supply and demand or other influences. For more information see Section 11.3.2.

- Resource Accounting and Charging: As mentioned above, it is responsible for recording resource usage and billing the user according to the resources needed and the pricing strategy agreed.

The exact role of the GSP is described in Section 11.2.3.

11.2.3 Workflow

To use a grid for the first time, several steps need to be done. First of all, the new user has to enroll himself as a grid user. To do so, the grid may require an authentication because of security reasons to proof if a user is really the person it claims to be. Usually this is not done via the internet because the user would have to many possibilities to betray. A certificate authority is responsible for the verification of the user's authority. After that the user is provided a grid software, that he needs to install on his computer. He also receives a login that he needs to get access to the grid.

After the successful installation of the software, the user can log on to the system with his login and start submitting jobs to the grid. In fact, the user sends the job to be done by the grid to a resource broker. This broker represents the client in the grid and tries to find the most suitable resource for the job to the cheapest price possible. As he "'works"' for several users, he always has the best overview over the actual resource offers. The broker then queries the information service about available hard- and software. As soon as the broker finds suitable and free resources, the job will be done by them and the result sent back to the user via the broker. For the user self, this whole process looks like a single huge and powerful computer.

11.2.4 Key Technologies applied in Grid Services

As the main goal of grids is the sharing of resources, standards are needed to provide all users with different operating systems, hardware, and software access to the grid. The Global Grid Forum aims to define these standards that should be integrated in every grid.

OGSA

The Open Grid Services Architecture has been published by the Global Grid Form. The goal is to fulfill the requirements of a grid with open source standard software so that interoperability on heterogeneous system can be guaranteed and different types of systems can communicate and share information. So barriers because of heterogeneity of the system can be avoided.

OGSA offers a framework that supports integration, virtualization, and management, It's interface is described by WSDL (more to WSDL see 11.2.4), that defines how to use a service. Figure 11.3 shows the OGSA architecture. On the picture can be seen, that different operating systems can communicate together as they all use the same standard interface. The interface includes multiple bindings and implementations, like the programming languages Java and C#. OGSA also consists of a grid security mechanism that are responsible for the secure communication between services which are built on the Globus Toolkit.

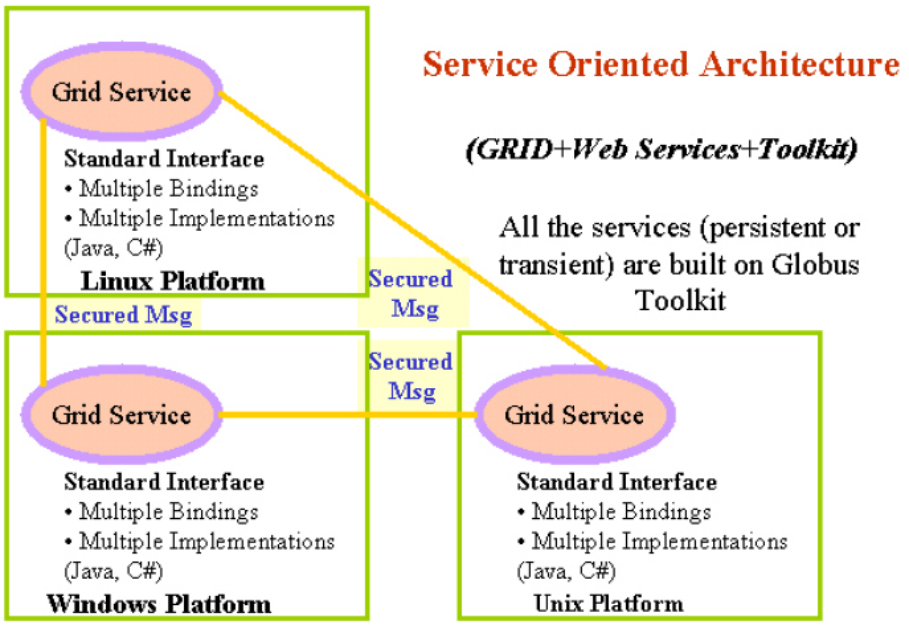


Figure 11.3: OGSA Architecture [20] p. 3

Globus Toolkit

The Globus Toolkit is an open source software used for building grids and applications. The goal is that all grid use the Globus Toolkit so that all grids use the same standards and hence all grids can be accessed independent of hardware, operating system and software. Figure 11.4 shows an overview over the Globus Toolkit and its components.

As can be seen in Figure 11.4, the toolkit includes software for: security, information infrastructure, resource management, data management, communication, fault detection, and portability.

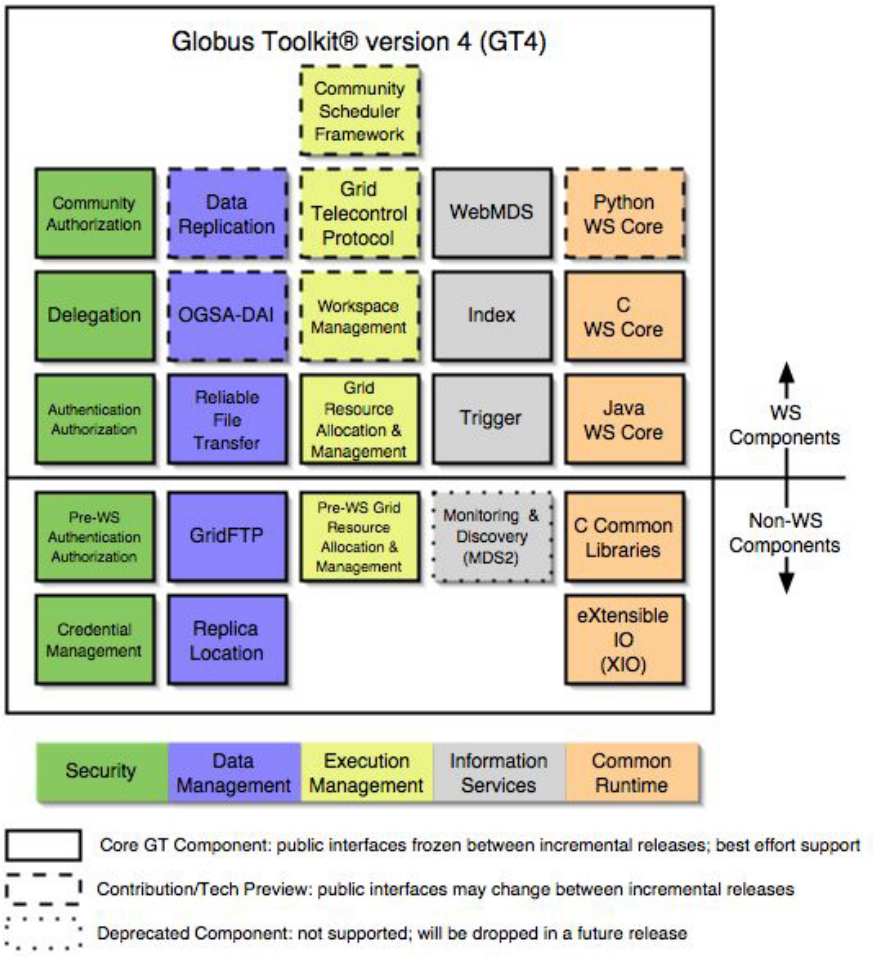


Figure 11.4: The Globus Toolkit [10]

Web Services

Web Services stands for an important distributed computing paradigm that focuses on Internet-based standards like SOAP, WSDL, and XML. These standards are needed to realize heterogeneous distributed computing. According to [6] p. 8 the Web Services "'define a technique for describing software components to be accessed, methods for accessing these components, and discovery methods that enable the identification of relevant service providers"'. The most important Web Services for a grid are described in this section.

SOAP The abbreviation SOAP stands for "'Simple Object Access Protocol"'. It is a standard based on several technologies and supports the messaging between a service provider and a service requestor. It provides:

- A message format for one-way communication with XML
- Conventions to implement the RPC Interaction Model
- Rules about what XML Elements an application needs to understand and what they need to do if they do not understand them

- A description how SOAP Messages have to be carried over http or smtp

As SOAP is independent of the used transport protocol, messages can be carried on HTTP, FTP, Java Messaging Services, e.g.

WSDL The Web Services Descripton Language (WSDL) is an XML format for describing the interface of a grid and defining how to use it. It contains functional information to the interface, the access protocol, and the deployment.

The services are defined through the following six XML-Elements:

- data types: definition of the data types that are used for the exchange of messages
- messages: abstract definition of the data to be transmitted
- port types: contains different types of messaging (one-way, request-response, solicit-response, and notification).
- binding: determines the concrete protocol and data format for the messages
- port: specifies the address of a binding
- service: merges a pool of related ports

XML The Extensible Markup Language, XML, is a W3 recommendation and was designed to describe data and uses therefore a Document Type Definition (DTD) or an XML Schema. Today, it is also used to exchange data. In XML the tags are not predefined, so own tags need to be invented.

XML has a lot of advantages:

- XML is free and extensible
- Data can be exchanged between incompatible systems
- With XML, plain text files can be used to share data
- With XML, plain text files can be used to store data

In grid computing XML is used to describe service interfaces and for the encoding of messages. This eases the integration of heterogeneous system into the grid.

11.3 Economical Aspect

11.3.1 Challenges

According to [4], Grid computing faces two main challenges, resource management and scheduling. The problem is, that the resource owners, called producers, and the resource users (consumers) have different goals and strategies. The fact, that the producers and consumers are distributed all around the world and therefore live in different time zones makes the problem of resource management and scheduling even harder. Different strategies have been developed to solve these challenges. Two approaches are so far used the most: a system-centric and a user-centric policy.

The system-centric policy tries to: "optimize system-wide measure of performance" ([4] p. 698) and is therefore mostly used in so called single administrative domains. A scheduling tool is responsible for the decision, which queries are to be done and which need to be refused. The decision is based on cost functions and the intention to improve the system throughput and utilization as well as the completion of the accepted jobs as soon as possible.

The goal of the user-centric approach, however, is to provide the highest utility possible to the users of the system. It's also called the economic-based approach and the scheduling is decided dynamically, based on the end-users requirements. There exist different competitive economic models for the allocation of the resources to the consumers. Some of them are presented in the next section.

11.3.2 Models for Resource Sharing and Allocation

Often a pricing system is used to solve the problem of resource sharing and allocation. The ones needing the resource the most are usually those willing to pay the most. So the problem is solved through the financial aspect. If a user has an unimportant job to be done via the grid, he is not willing to pay as much as a user under time pressure or having a very important application to be run on the grid. So the latter user usually can use the grid before the one with the unimportant task. There exist lots of different pricing models and the most important ones are discussed in this section.

Commodity Market

In the commodity market model, the Grid Service Provider offers a price, which is set depending on several factors. There exist four different pricing systems: Flat fee, Usage Duration (Time), Subscription and Demand and Supply-based. In the **flat fee concept**, the GSP sets a price that is after wards fixed for some time and does not change anymore, irrespective of quality, supply and demand. Applying the **usage duration model**, the consumers are charged a price according to the time they needed the resource. In the

demand and supply-based concept, the prices change very often according to the market demand and supply, trying to reach market equilibrium.

How does the model work? There exist different service providers which offer different prices and conditions. The GRB identifies these GSPs through the Grid Market Directory (GMD), that is a directory where the GSPs can publish their prices and the GRB tries to find a GSP providing the resources needed to the lowest price offered. As soon as the GRB found an appropriate GSP, the job is executed.

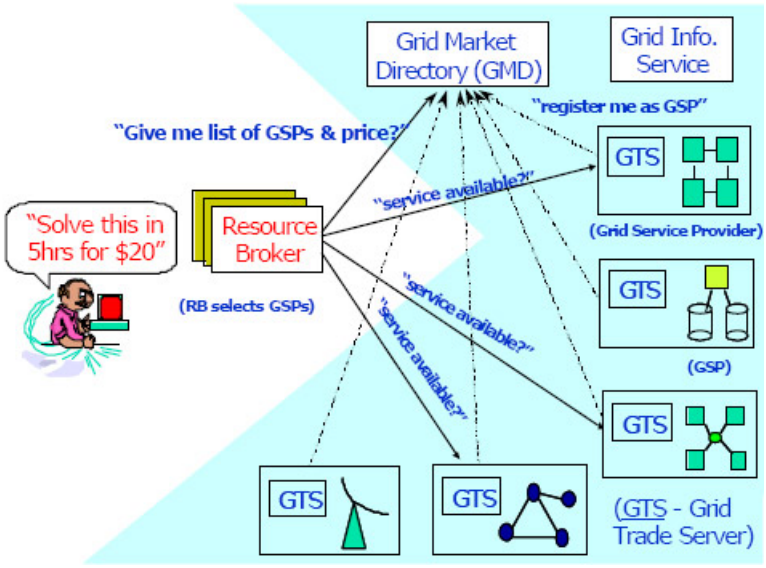


Figure 11.5: Commodity Market Model from [3], p. 6

Posted Price Model

The posted price model is very similar to the commodity market model. The only difference is, that in this model the GSPs make special offers. The aim of providing these special prices is to gain new consumers or to make existing consumers change to cheaper (time-) slots. Usually the special offers are coupled with some usage conditions, but if they meet a consumers need then they are very attractive to them.

How does the posted price model work? The functionality is again similar to the one of the commodity market model. Additionally, the GSP announce their special offers with the corresponding conditions in the GMD and the GRB tries to find a special offer that suits its requirements and then checks if the special service is available. The rest works similar to the commodity market model.

Bargaining Model

Systems using this type of resource management and scheduling do not have fixed prices - neither for the access nor for the usage duration. As the name bargaining says, the GRB and the GSP bargain with each other for low prices and long usage duration. The

bargaining lasts as long as needed to find a price that is agreeable for both sides or until one party is not willing to negotiate anymore. So they might end the bargaining without finding a common price. If the GRB takes a lot of risk, he has the chance to get the needed resources to a very low price. This can happen, when all GRB do not agree with high prices and therefore the GSP have a lot of unused resources. Then it is more economical for them to sell their resources to a low price instead of wasting them and not gaining anything at all.

Tender / Contract-Net Model

In [3] it is stated, that this model is "one of the most widely used models for service negotiation in a distributed problem-solving environment" (p. 8). In this model, the GRB is also called a manager and the resource owner a contractor. The GRB declares what he needs (which kind of resource and how long he needs the resource) and asks the GSP to make offers (bids). The GSPs receive this demand through the GMD and if a GSP has the adequate resources it can answer with a bid. The GRB then checks the different offers and decides which is the best for its needs. After that, the GRB and the GSP communicate directly, the GSP delivers its service and the GRB has to pay for it.

Auction Model

There exist many different auction models. The most important ones are ([3] p. 9):

- English Auction (first-price open cry)
- First price sealed-bid auction
- Vickrey (Second price sealed-bid) auction
- Dutch Auction
- Double Auction

Additionally to the seller (GSP) and the consumers (GRBs), they also need an auctioneer (instead of the GMD) which is also called a mediator (GMA - Grid Market Auctioneer). The auctioneer's responsibility is to manage the whole auction. A good and well-known example of an auctioneer is eBay.com. The different variations of auctions all have in common that they use a so called one-to-many negotiation. There is only one service provider and many different potential consumers, which bid a price for the resources offered by the GSP.

English Auction The English auction is handled as an open auction. This means that all bidders see the highest price offered at the moment for the resource. The GRB then have to decide if they want pay more than the actual price or stop bidding. The advantage of the open auction is that they do not have to start offering a very high price - they can increase their offer step by step. If none of them is willing to bid higher anymore, the

auction will be stopped and the GRB offering the latest and therefore highest price gets the access to the resource. Figure 11.6 shows the interactions between the auctioneer and the Clients.

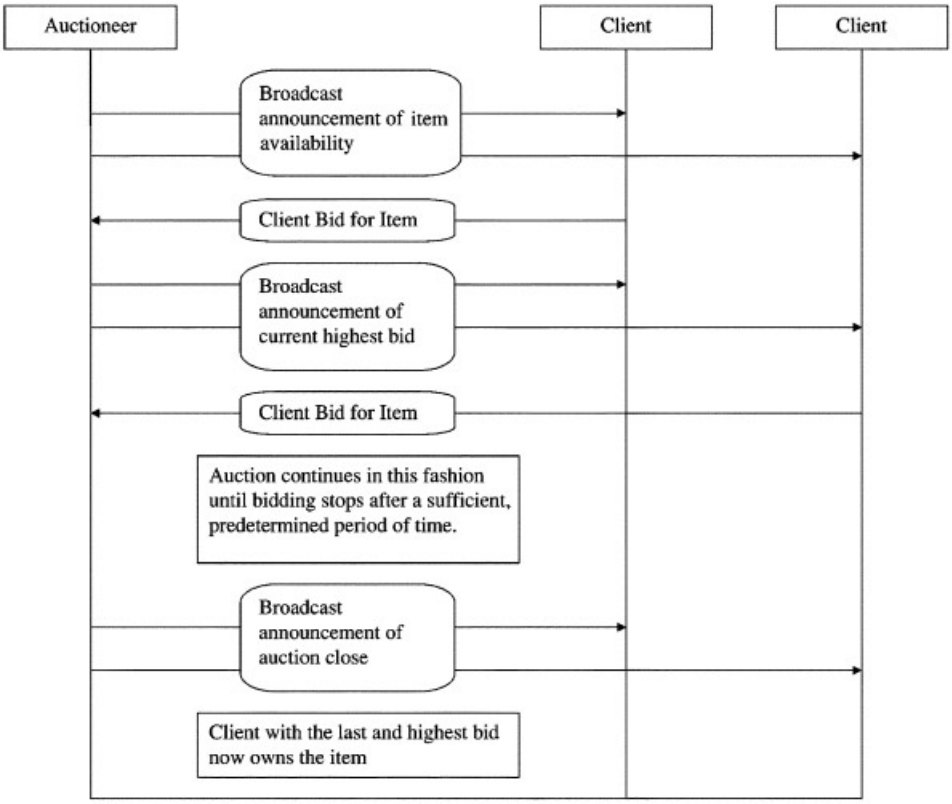


Figure 11.6: Model of the English Auction Interaction [4], p. 710

First Price Sealed-bid Auction Contrary to the English auction, this model is a closed auction, so the bidders do not know how much the other bidders offer. Hence, each bidder can only make one bid and the one willing to pay the most gets the resource at the price the bidder offered.

Vickrey (Second Price Sealed-bid) Auction The vickrey auction works the same as the first price sealed-bid auction. The only difference is that the highest bidder does not have to pay its offered price but the price offered by the second highest bidder.

Dutch Auction The Dutch auction is almost the opposite of the English auction. The auctioneer starts the auction by announcing a (too) high price for the service offered. Then the auctioneer lowers the price step by step until one bidder agrees to pay the actually offered price. This kind of auction is a very meaningful reflection of the market's demand and supply. If the demand is higher than the supply, the bidders are willing to pay a much higher price than during a period of high supply and low demand. Then the price accepted by the bidders is very low.

Double Auction The concept of the double auction is based the idea, that both - sellers and buyers - can offer a price for a service. As this model is a closed auction, only the GMA sees the prices for the offers and the bid prices. If one buyer bids a price that fits

the price of an offer, the sale of the service takes place. For this model, the auctioneer must be very trustful and guarantee that it does not match a high bid with a low-priced offer and keep the money surplus.

Bid-based Proportional Resource Sharing Model

Using this system, the users (GRBs) own tokens or credit points which they need for getting access to resources or services. If a user needs a resource, he can decide how many token he is willing to spend for this service. If other users need the same resource, they also need to communicate how many token they are offering for this resource. All GRBs willing to spend some token get a percentage of the resource, depending on the amount of token they offer. For example, if there are two users and one of them offers 4 tokens and the other 8, the first user gets 1/3 of the resource and the other 2/3. So the price is not fixed at all and just depends on the other users and consequently on the actual demand. Figure 11.7 shows the interactions between the user (respectively the GRB), the GMD and the GSP (which is represented by the GTS).

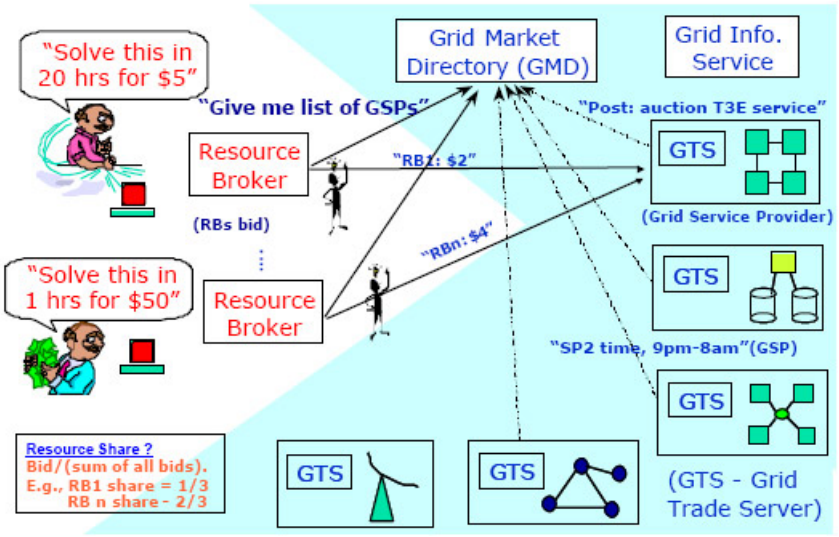


Figure 11.7: Model of the Bid-based Proportional Resource Sharing Interaction [3], p. 11

Community / Coalition / Bartering Model

Some people affiliate and share their resources with each other. So they build a community where all of them provide resources for the others but also can profit of this relationship and use the other's resources. If anyone wants access to some resources he has to contribute and also provide his own ones. There exist different possibilities to manage the system. One way is the token system, described in Section 11.3.2 or building accounts with credits etc.

Monopoly and Oligopoly

If only one GSP provides a certain service, this provider can profit from its position as a monopolist. In this case the user are addicted to the GSP and have no other possibility than accepting the GSP's price and conditions.

Usually the market situation is what is called an oligopoly: There are more than only one GSP, but still a small amount of them and much more users. So the GSPs still have the power to set the prices and conditions for their resources.

11.3.3 Benefit of Grid Economies

The Grid economy provides many different benefits to the consumers as well as to the providers. From the consumer's point of view, his main benefit of the Grid itself is that he does not need to buy the hardware. Instead of that he can rent the amount of the resources needed whenever he needs them. If he hardly needs more resources than he owns, its more economical for him to use the Grid. He neither has costs for the whole maintenance nor does he need a storage or server room.

The main benefit for the resource provider is, that he does not waste hardware or resources. Instead of leaving them unused, he can gain some additional money by participating in the Grid. This does not disturb his own work, because he only offers resources if he has some unneeded time slots. A list of benefits provided by the Grid economy is presented in [4] p. 699:

- It helps building a large-scaled Grid as it offers incentive for resource owners to contribute their resources for others to use and profit from it.
- It helps in regulating the supply and demand for resources.
- It offers an economic incentive for users to reduce their priority in favor of incurring a lesser expense and, thus, encourages the solution of time critical problems first.
- It provides a common basis for comparing conflicting needs by allowing users to express their requirements and objectives in currency terms.
- It offers uniform treatment of all resources. That is, it allows trading of everything including computational power, memory, storage, network bandwidth/latency, data, and devices or instruments.
- It helps in building a highly scalable system as the decision-making process is distributed across all users and resource owners.
- It supports a simple and effective basis for offering differentiated services for different applications at different times.

Conclusion Grid computing has a high potential for the future. It helps the industry to save time and money by acting as a consumer. So the enterprises do not have to wait until their own resources have executed a specific task neither do they need to buy additional resources. It is much more profitable for them to rent these resources and to pay no attention to the rest. From the view of the resource owner his profit of the Grid economy is to gain some (unscheduled) money without any additional effort. If they use a good system, the identification of unused resources as well as the whole allocation of these spare capabilities is done automatically. So the resource provider does not need to care for anything at all.

11.4 Usage of Grids Today

Today, grids are mostly used for testbeds like the Astrogrid, Birn, Datagrid, Eurogrid, Teragrid, Open Science Grid Consortium, etc[1].

The **Datagrid** is developed by the CERN, the European Organization for Nuclear Research, and different countries in Europe are involved in this project. Its goal is to enable next generation scientific exploration. This requires intensive computation as well as the analysis of shared large-scale databases.

The **Open Science Grid Consortium** is a project in the US. In this project they build a sustainable open national grid infrastructure for science in the U.S. that will make resources available at many labs and universities in the U.S. The accessibility is offered through a common grid infrastructure and provides access to shared resources for the benefit of scientific applications. The Open Science Grid consortium is cooperating with other national and international grid infrastructures to achieve global interoperability.

Grids are also often used for scientific projects like the Anthrax Research Project. It's goal was to find as soon as possible a suitable treatment for advanced-stage anthrax. Using a grid, for the screening of the 3.57 billion molecules only 24 h were needed. Without profiting from the advantages of a grid, this screening would have lasted at least four weeks.

11.5 Conclusion

Grid computing is a relatively new technology that is evolving very fast and has a broad acceptance in science. The main problem grids face is the allocation of resources and the correct charging for the consumed resources. Therefore, several market models have been developed that should help ease the allocation problem. Mostly the ones willing to pay the most get access to the resources first. If a resource is needed very urgently, usually the consumers pay a higher price and on this way get access to the needed hard- or software. Still there is the problem of the correct charging of the consumers. Actually

[1] more about these testbeds see http://gridcafe.web.cern.ch/gridcafe/gridprojects/testbed.html

the prices are due to the time the access is needed or the amount of disk space used etc. But as sometimes an application is split on different resources it is still not easy to share the price between the different resource providers as well as the consumer does not want to get several bills for one job he wishes to be done by the grid. New solutions for this problem need to be found in future.

Bibliography

[1] Alonso, G.; Casati, F.; Kuno, H.; Mchiarju, V.: Web Services: Concepts, Architectures and Applications, Springer Verlag, Berlin / Heidelberg, 2004.

[2] Berstis, Viktor: Fundamentals of Grid Computing, Redbooks Paper, IBM, 2002.

[3] Buyya, R.; Abramson, D.; Giddy, J.; Stockinger, H.: Economic Models for Resource Management and Scheduling in Grid Computing, Online Document: `http://www.csse.monash.edu.au/~davida/papers/jcpe.pdf`

[4] Buyya, R.; Abramsaon, D.; Venugopal, S.: The Grid Economy, Proceedings of the IEEE, Vol. 93, No. 3, March 2005.

[5] Foster, I. & Kesselman, C: The Grid: Blueprint for a New Computing Infrastructure, San Francisco: Morgan Kauffman, 1999.

[6] Foster, I.; Kesselman, C.; Nick, M. J.; Tuecke, S.: The Physiology of the Grid, An Open Grid Services Architecture for Distributed Systems Integration, Online Document: `http://www.globus.org/alliance/publications/papers/ogsa.pdf`, 2002.

[7] Foster, I.; Kesselman, C.; Tuecke, S.: The Anatomy of the Grid, Enabling Scalable Virtual Organizations, Online Document: `http://www.globus.org/alliance/publications/papers/anatomy.pdf`, 2001.

[8] Foster, I.; Tuecke, S.: Different Faces of IT as Service, ACM Queue, `http://acmqueue.com/special/ftpfiles/articles/foster.pdf`, July/August 2005.

[9] Global Grid Forum, `http://www.gridforum.org`.

[10] Globus Toolkit, `http://www.globus.org/toolkit/about.html`.

[11] GridCafé, Website about Grid computing: `http://gridcafe.web.cern.ch/gridcafe/index.html`.

[12] Grid Computing Info Centre (GRID Infoware), Website: `http://www.gridcomputing.com`.

[13] Jacob, B.; Brown, M.; Fukui, K.; Trivedi, N.: Introduction to Grid Computing, IMB Redbooks, 2004, Online Document: `http://www.redbooks.ibm.com/abstracts/sg246778.html?Open`

[14] Keller, S.: Service Oriented Architectures in Financial Services Companies, Semester Work, University of Zurich, Mai 2005.

[15] Open Grid Services Architecture, Website: `http://www.globus.org/ogsa`.

[16] Stiller, B.; Gerke, J.; Flury, P.; Reichl, P.; Hasan: Charging Distributed Services of a Computational Grid Architecture, IEEE International Symposium on Cluster Computing and the Grid, 2001.

[17] Strong, P.: Enterprise Grid Computing, From Enterprise Distributed Computing, Vol.3, No. 6, July/August 2005. Online Document: `http://www.acmqueue.com/modules.php?name=Content&pa=showpage&pid=324`.

[18] What is the Grid?, Webiste: `http://www.access.ncsa.uiuc.edu/witg/`.

[19] Wolski, R.; Plank, J. S.; Bryan, T.; Brevik, J.: G-commerce: Market Formulations Controlling Resource Allocation on the Computational Grid, International Parallel and Distributed Processing Symposium (IPDPS), San Francisco, CA, April, 2001.

[20] Zhang, L-J.; Chung, J-Y.; Zhou, Q.: Developing Grid computing applications, Part 1: Introduction of a Grid architecture and toolkit for building Grid solutions, 2002, Online Document: `http://www.106.ibm.com/developerworks/library/ws-grid1/`

Kapitel 12

New Business Models based on P2P

Jonas Alleman, Michel Hagnauer, Fabio Pérez Cina

Trotz dem schlechten Ruf der peer-to-peer (P2P) Technologie auf Grund von legalen Problemen, gewinnt dieser Ansatz immer mehr an Boden. Diese Arbeit führt verschiedene wesentliche P2P Technologien ein und erläutert die Hauptunterschiede zwischen strukturierten und unstrukturierten Netzwerken. Es wird ein Vergleich zwischen dieser Technologie und dem klassischen Client/Server Ansatz gezogen. Als Kern der Arbeit werden neue Geschäftsmodelle gesucht, die auf dieser Technologie basierend entstehen können und Geschäftsmodelle analysiert, die bereits existieren.

Inhaltsverzeichnis

12.1 Einleitung

Peer-to-peer (P2P) beschreibt grundsätzlich Netzwerke, deren Teilnehmer als Gleichberechtigt gelten. Am Anfang der Internet-Ära, war der Verbindungsaufbau in der Form eines P2P Netzwerkes konstruiert worden. Die Idee vom ARPA-Net war, dass viele verschiedene Computer sich direkt miteinander verbinden konnten, wobei jeder Knoten eine gleichwertige Rolle spielen würde. Zwei prominente Beispiele, die am Anfang der Internet-Ära entstanden sind und heute noch gebraucht werden, sind Usenet und DNS. Diese verfolgen weiterhin die Philosophie von einem P2P-Netzwerk. Durch das Aufkommen der Client/Server Architektur, wurde die Philosophie des Internets jedoch geändert. Es entstand ein Netzwerk, bei dem Millionen von verschiedenen Clients aus einer relativ beschränkten Anzahl von Anbietern, generell Daten runterladen. Es gab somit relativ wenig grosse Knoten, welche eine hohe Upload Bandbreite brauchten. Durch die fortgeschrittene Computer-Technologie, sowohl der Arbeitsleistung eines Computers, als auch der schnellen Verbindungen in den Netzwerken, werden heute manche Dienste sinnvoller über ein P2P-Netzwerk getätigt. Durch die peer-to-peer Verknüpfung können die grossen Server ihrer Last entnommen werden und jeder Computer der im Netzwerk angeschlossen ist, übernimmt die Rolle eines kleinen Servers. Durch dieses Konzept können die Verbindungen, als auch die Arbeitslast besser verteilt werden. Da in den letzten Jahren die Philosophie der Client/Server-Architektur sich überzeugend durchgesetzt hat und heutzutage die P2P-Technologie noch oft als eine illegale Form des Datenaustausches angesehen wird, ist es schwer ein standhaftes Business Modell im Rahmen der P2P-Technologie aufzubauen und dies in einer von Client/Server-Komponenten dominierten Umwelt umzusetzen.

In diesem Dokument werden einige Beispiele angeschaut, bei denen solche Business Modelle aufgebaut werden können. Es werden Vorschläge zu sehen sein, wie und wo das Ganze umgesetzt werden kann. Und am Schluss werden noch die Gefahren, Hindernisse und vor allem die Chancen aufgezeigt, was P2P heute und morgen für die IT-Welt bringen wird.

12.2 P2P Technologien

Wie bereits in der Einleitung erwähnt wurde, handelt es sich bei P2P Netzwerke hauptsächlich um Kommunikation unter Gleichen, was schon durch das Wort Peer, das aus dem Englischen übersetzt soviel wie Ebenbürtiger oder Gleichberechtigter bedeutet, deutlich wird. Im Gegensatz dazu steht die weit verbreitete Client Server Architektur, bei der das Netzwerk auf Bandbreite und Rechenkapazität von relativ wenigen Servern zu relativ mehr Clients aufbaut. Auf die Unterschiede zwischen diesen zwei Netzwerkarten wird später genauer eingegangen.

Im idealen P2P Netzwerk gelten alle Computer als gleichberechtigt und agieren sowohl als Client wie auch als Server. Das bedeutet, dass sie sowohl Dienste in Anspruch nehmen wie auch Dienste anbieten können.

12.2.1 P2P Klassen

Es wurde bereits angedeutet, dass P2P Netzwerke nicht zwangsläufig reine P2P Netzwerke sein müssen, in denen alle Teilnehmer in jeder Hinsicht gleichberechtigt sind. Deshalb unterscheidet man allgemein drei Klassen:

1. Zentrale P2P Netzwerke
    - zentraler Server hält Informationen über Peers und verarbeitet Anfragen zu diesen Informationen
    - die Daten, welche im Netzwerk vorhanden sind und eventuell zum Tausch angeboten werden, werden von den Teilnehmern gehalten, nicht vom Server
    - Teilnehmer teilen dem zentralen Server mit, welche Dateien sie zur Verfügung stellen
    - Teilnehmer sind verantwortlich für das Herunterladen von Ressourcen an andere Teilnehmer, die von diesen verlangt werden
2. Dezentrale P2P Netzwerke
    - Teilnehmer agiert als Client und Server
    - kein zentraler Server
    - Informationen über Teilnehmer werden nicht zentral gehalten
3. Hybride P2P Netzwerke
    - Besitzen sowohl zentrale wie dezentrale Eigenschaften, auch Super P2P Netzwerk genannt

Technisch gesehen sind reine P2P Netzwerke selten. Solch ein Netzwerk muss ausschliesslich Protokolle einsetzen, die das Konzept des Client und des Servers nicht kennen. Viele Netzwerke, die als P2P beschrieben werden, implementieren tatsächlich Elemente die nicht als reine P2P gelten. So sind zum Beispiel einige Teilnehmer in gewissen Netzwerken stärker (super-peer, super-node) und agieren als Server, um die sich die Client peers Sternförmig anbinden.

In hybriden Architekturen existiert mindestens ein Server. Alle Clients sind mit einem oder mehreren Servern verbunden. Die Daten, die vom zentralen Server gehalten werden, reichen im Fall von zum Beispiel file-sharing Netzwerken von der IP Adresse, Bandbreite und Portnummer des Clients bis hin zu der Liste der Dateien, die auf den jeweiligen Clients vorhanden sind. Alle Suchanfragen werden demzufolge an einen zentralen Server geleitet, der als Antwort eine Liste der Clients schickt, bei denen die geforderte Datei vorhanden ist. Wenn es darum geht, die Datei tatsächlich herunter zu laden, bauen die Teilnehmer eine direkte Verbindung auf. Vorteilhaft ist bei dieser Architektur die schnelle Suche dank den zentral gehaltenen Informationen über Dateien und Teilnehmer. Anderseits sind hybride dezentralisierte Architekturen anfälliger auf Angriffe, weil weniger Redundanz die zentralen Stellen schützt. Ausserdem sind diese Systeme an die limitierten Kapazitäten

bezüglich CPU Leistung, Bandbreite und Speicherkapazität des zentralen Servers gebunden.

Zur dritten Kategorie gehören zum Beispiel partiell zentralisierte Netzwerke. Im Gegensatz zu vollkommen dezentralisierten Netzwerken nehmen bei dieser Architektur gewisse Knoten eine wichtigere Rolle ein. Diese so genannten Super Knoten agieren als zentrale Indexe für eine Teilgruppe von normalen Knoten. Welche Knoten zu Super Knoten werden ist von System zu System verschieden. Da die Super Knoten jedoch meistens dynamisch zugewiesen und ausgewechselt werden, gelten sie nicht als single point of failure, das heisst sie verursachen nicht ein Systemkollaps, wenn sie ausfallen.

12.2.2 Vor- und Nachteile

P2P Netzwerke sind Systeme, die eine flexible Skalierbarkeit besitzen, da die Anzahl Teilnehmer wachsen und schrumpfen kann, ohne die Auslastung markant zu beeinflussen. Ausserdem brauchen reine P2P Netzwerke dank ihrer Dezentralisierung keine zentrale Koordinationsstelle. Im Vergleich zu einem Client/Server Ansatz, wie er zuvor eingeführt wurde, können für P2P Netzwerke folgende Vor- und Nachteile genannt werden.

Vorteile

- Geringe Kosten: Teilnehmer liefern Ressourcen, Bandbreite, Speicherplatz und Rechenleistung -> Erhöht sich Teilnehmerzahl, erhöht sich Leistung.
- Robustheit: durch replizieren der Daten über das Netzwerk. Im Fall von reinen P2P Netzwerke gibt es keinen Single Point of Failure, kein zentraler Server
- Fehlertoleranz: Beim Ausfall von Teilnehmer ist der Datentransfer meistens durch Zahlreiche benachbarte Knoten gewährleistet

Nachteile

- Langsamere Arbeitsplatzrechner, Z.B. im kleinen Heim- oder Firmennetz wegen der Beanspruchung von Ressourcen der zusätzlichen Arbeit im Netz
- Geringere Geschwindigkeit bei Suchanfragen, weil der Zugriff auf Daten nicht zentral geregelt ist
- Schwieriges Backup: Daten über mehrere Rechner verteilt
- Technisch anspruchsvolle Routing- und Suchalgorithmen

12.2.3 Sicherheit und Angriffe auf P2P Netzwerke

Zum einen betreffen Sicherheitsaspekte die Benutzer von P2P Netzwerken, wenn zum Beispiel beim Anwenden von Tauschbörsenapplikationen das Netzwerk als Instrument zum Angreifen benutzt wird. Anderseits können P2P Netzwerke selbst Opfer von Angriffen sein. Die zwei Szenarien werden in Folge vorgestellt.

Sicherheitsprobleme für Benutzer:

- Poisoning attackts: Ressourcen anbieten, dessen Inhalt von der Beschreibung abweicht (zum Beispiel in Tauschbörsen)
- Defection attack: Das Netzwerk benutzen, ohne eigene Ressourcen beizusteuern
- Viren (und andere schädliche Software) zum download anbieten (betrifft Tauschbörsen)
- Spyware (sammelt persönliche Daten über Benutzer) in der P2P Software einschleusen, ohne dessen Einverständnis (betrifft Tauschbörsen)
- Filtrieren von Seiten der Netzwerkoperatoren

P2P Netzwerke sind offen und arbeiten im Internet in einer feindlichen Umgebung. Angriffe auf P2P Netzwerke geschehen ständig und aus verschiedenen Gründen. In Folge werden einige Beispiele erklärt.

Ein üblicher Angriff ist der denial of service attack (DoS), insbesondere der distributed DoS (DDoS) bei P2P Netzwerken. Es handelt sich dabei um einen koordinierten Angriff von einer Vielzahl von Systemen auf ein Netzwerk oder Server, um dessen Dienste zum Beispiel durch Überlastung arbeitsunfähig zu machen. Dieser Angriff wird auch eingesetzt, um andere Angriffe zu tarnen oder maskieren.

Ein weiterer spezifischer Angriff auf die Vermittlungsschicht eines Netzwerkes sind Selective-Forwarding-Attacks. Dabei positioniert sich ein bösartiger Knoten zwischen zwei Knoten, dessen Kommunikation er stören möchte. Von dieser Stelle aus verschickt er Datenpakete zufällig oder nach einem bestimmten Schema. Um sich in solch einem Pfad zu integrieren, gibt es für den Angreifer zwei Möglichkeiten. Sinkhole-Attacks sind Angriffe, bei denen der bösartige Knoten versucht, einen Grossteil des Datenverkehrs um ihn herum an sich selbst umzuleiten. Das erreicht er dadurch, dass er angibt, optimale Routen zu verschiedenen Ziele zu besitzen. Eine weitere Möglichkeit Sinkholes zu erzeugen, ist durch die so genannten Wormhole-Attacks. In diesem Fall bilden zwei weit entfernte bösartige Knoten ein Tunnel, wodurch grosse Mengen an Nachrichten diesen Tunnel benutzen werden, wo sie dann zum Beispiel verworfen werden könnten.

Ein prominentes Beispiel einer Angriffsmöglichkeit auf P2P Netzwerke wird als Sybil-Attack bezeichnet. Die Bezeichnung wurde aus dem gleichnamigen Buch von Flora Rheta Schreiber entnommen, das von einem Mädchen mit 16 verschiedenen Persönlichkeiten, also multipler Schizophrenie, erzählt. Mit dieser absichtlichen Persönlichkeitsstörung im Fall eines Netzwerkteilnehmers können überproportionale Anteile eines Netzes kontrolliert

werden. Ein Knoten nimmt dabei mehrere Identitäten an und ist in der Lage Schutzmechanismen wie das Multipath-Routing, also das Schicken einer Nachricht über mehrere Pfade um Redundanz zu gewinnen, zu unterbrechen.

Die meisten der hier vorgestellten Angriffe können jedoch gegen Outsidern, also Angreifer ohne besondere Kenntnisse bezüglich Schlüssel, durch Verschlüsselung und Authentisierung vereitelt werden. Eine andere Abwehrmöglichkeit ist eine hohe Redundanz der Pfade beim vermitteln von Paketen einzubauen, um die Robustheit des Netzwerkes zu steigern.

12.2.4 P2P und die Illegalität bei Tauschbörsen

File Sharing ist wie bereits erwähnt die meist verbreitete Anwendung von P2P Netzwerken, bewegt sich jedoch auf Grunde der beliebten Musik- und Videotauschbörsen oft auch am Rande der Legalität. Die rechtlichen Aspekte von P2P Netzwerken entpuppen sich als einer der grössten Herausforderungen für das Überleben solcher Tauschbörsen. Die "Betamax desicion" [33], in den USA lange vor der Verbreitung des Internets gefallen, besagt dass Technologien nicht inhärent illegal sind, wenn damit wesentlicher legaler Gebrauch gemacht werden kann. Diese Eigenschaft trifft auch für P2P Tauschbörsen zu, da man sie für den Austausch von korrekt lizenzierten Dateien gebrauchen kann. Aber in der Praxis sind die meisten ausgetauschten Dateien urheberrechtlich geschützte Musikdateien. Als Konsequenz wurden die benutzten Applikationen als illegal und als eine Bedrohung für die etablierte Unterhaltungsindustrie betrachtet. Die Entwicklung der Software ging danach in die Richtung, dass sie sowohl technisch wie auch juristisch schwieriger angreifbar wurden. Da es Tauschbörsen gibt, bei der die Software an sich als legal betrachtet werden kann, machen sich die Benutzer, die Urheberrechtbestimmungen brechen, strafbar.

In den letzten Jahren sind Vertreter der Musikindustrie tatkräftig geworden und drohen mit Geld- und sogar Gefängnisstrafen. Das IFPI, der Verband der Schweizer Musikindustrie (Pendant zur Amerikanischen RIAA), hat bestätigt, dass bereits gegen 1200 professionelle und semiprofessionelle Raubkopierer Verfahren abgewickelt wurden. Gemäss den Aussagen von Beat Högger von IFPI wurde bisher nicht gegen Privatanwender vorgegangen [21]. Doch in Zukunft werde man vor allem gegen Benutzer vorgehen, die ein vergleichsweise grosses Volumen an geschützten Werken hoch- und runterladen. Die unbekannten Benutzer ausfindig zu machen ist wiederum eine schwere Aufgabe, da nur Ermittlungsbehörden die ISP (Internet Service Providers) auffordern können, den Datensatz von ihren Benutzer freizugeben. Zu dem können aus technischen Gründen viele ISP die dynamisch vergebenen IP Adressen der Benutzer nicht immer zuordnen. In der Schweiz werden P2P Tauschbörsen vor allem nach illegalem Inhalt wie harte Pornografie abgesucht [23]. Das kopieren von geschützten CDs und das Herunterladen von Tauschbörsen ist im Schweizer Gesetz noch nicht eindeutig verboten. Das Anbieten von geschützten Musikdateien im öffentlichen Internet ist jedoch strafbar, weil diese Tätigkeit weit über die Vergabe von Musik an den Bekanntenkreis hinausgeht [22].

12.3 Client/Server- und P2P-Architektur

12.3.1 Client / Server Architektur

Eigenschaften

Der Term Client/Server [20] wurde zum ersten Mal 1980, in Referenz zu Personal Computer (PC) im Netzwerk, gebraucht. Das heutige Client/Server Modell gewann Ende der 80er Jahren an Akzeptanz. Die Client/Server Softwarearchitektur ist "versatil" d.h. anpassungsfähig, Nachrichten basierend und hat eine modulare Infrastruktur. Somit versucht die Software, die Benutzbarkeit, Flexibilität, Interoperativität und Skalierbarkeit zu steigern. Ein Client ist als Serviceabfrager (requester of services) und ein Server als Provider eines Services definiert.

Beispiele

Als Beispiel kann man zwei der bekanntesten Architekturen von Client/Server, die häufig eingesetzt werden, erwähnen.

Two Tier Architekturen sind von Seitens der Software als zweischichtiges System aufgebaut. Die Rechenkapazität wird dabei weitgehend auf die Client-Rechner ausgelagert, um den Server zu entlasten. Auf dem Server läuft eine Datenbankanwendung. Die Clients übernehmen die Logik und die Darstellung der Benutzerschnittstelle.

Three Tier Architekturen Im Gegensatz zur zweischichtigen- oder Two Tier Architektur, bei der die Rechenkapazität weitgehend auf die Client-Rechner ausgelagert wird, um den Server zu entlasten, existiert bei der dreischichtigen Architektur noch eine zusätzliche Schicht, die Logikschicht, die die Datenverarbeitung vornimmt. Es gibt verschiedene Varianten um diese Middle Tier Architektur zu implementieren, zum Beispiel durch transaction processing monitors, message servers oder application servers. Zum Beispiel kann der Anwender eine Anfrage an die Middle Tier schicken und "entkoppeln", diese wird sich mit der Datenbank in Verbindung setzen und eine Antwort an den Clients schicken.

12.3.2 P2P - Unterschiede zwischen zentralisiert und dezentralisiert

In diesem Abschnitt werden die verschiedenen P2P Technologien, die in unserer heutigen Zeit benutzt werden, vorgestellt. Die Architektur und Protokolle werden analysiert, um Vor- und Nachteile sowie die Schwierigkeiten und Problematiken, die in Zukunft anfallen können, aufzuzeigen.

Einleitung

P2P ist weit davon entfernt, eine neue Technologie genannt zu werden. Die Server in vielen alten Technologien arbeiten in einer P2P-Weise zusammen, um benötigte Informationen auszutauschen. Nachrichten, E-Mail und IRC (Internet-Relay-Chat) fallen unter diese Kategorie.

In der P2P-Welt gibt es vollkommen unterschiedliche Kategorisierungen dieser Technologie, die von zentralisiert zu vollkommen dezentralisiert schwanken. Auf der einen Seite, Napster und Seti@Home, die zentralisiert sind, und auf der anderen Freenet und GNUtella, die dezentralisiert sind. Napster integriert einen zentralisierten Indizierungsserver, der auflistet, wo die Dateien sind und welche Arbeitsstationen Suchvorgänge starten (zeitlich geordnet). Freenet und GNUtella verteilen andererseits die Datenbank von gemeinschaftlichen Ressourcen über den Arbeitsstationen auf dem Netz, die einen Bedarf am zentralen Server beseitigt.

Zentralisiertes P2P

In dieser P2P Kategorie fällt die Napster und Seti@Home's Topologie. Es gibt weitere Applikationen, die eine ähnliche Topologie aufweisen, aber diese zwei gelten als die bekanntesten.

Napster [15] Das Programm "Napster" entstand im Januar 1999, als Shawn Fanning, ein Student im ersten Semester an der Northeastern University, Boston, Massachusetts, USA, eine Anwendung schrieb, um den Musikaustausch zwischen den Studenten im Wohnheim zu erlauben. Napster-Inc. wurde im Mai gegründet und gewann rasant an Anwendern (bis zu 21 Millionen). Im Dezember 1999 verklagte die Datenschutzindustrie von Amerika Napster für Copyrightmissbrauch. Obwohl Napster noch in Betrieb ist, existiert nur ein kleiner Bruchteil von der ehemaligen Benutzergruppe.

Napster basiert auf einer Client/Server Architektur Abbildung 12.1. Die Rolle des Servers ist, ein "suchbares" Register zu halten, das Eingaben von mp3's von allen gegenwärtigen verbundenen Clients (Arbeitsstationen) enthält. Der Server ist eigentlich vielfach sehr ausgelastet durch die Anfragen von Clients. Dieser Ansturm ans Service führt zur Anregung, ob man Maschinen in den Serververbund hinzufügen und sichern soll, falls Server scheitern, und sie somit ohne bedeutende Unterbrechung ersetzt werden können ohne den Dienst zu beeinflussen. Ebenso wird die Kommunikation zwischen Clients und Server implementiert. So werden die Server mit unterschiedlichen ISPs verbunden.

Die Arbeitsstationen können Metadaten mit den gemeinschaftlichen mp3s auf ihren eigenen Rechnern indizieren und assoziieren. Diese Information wird dann anschliessend den Napster-Servern geschickt, wenn man sich verbindet. An diesem Punkt kann der Client alle Clients durchsuchen, die auf Napster verbunden sind, indem man dem Napster-Server Search "queries" (Anfragen) schickt. Der Server wird sein internes Register von gegenwärtig gemeinschaftlichen (verfügbaren) Dateien durchsuchen und eine Rückmeldung generieren. Die Ergebnisse enthalten Metadaten über die Datei, die Stelle von der Datei und

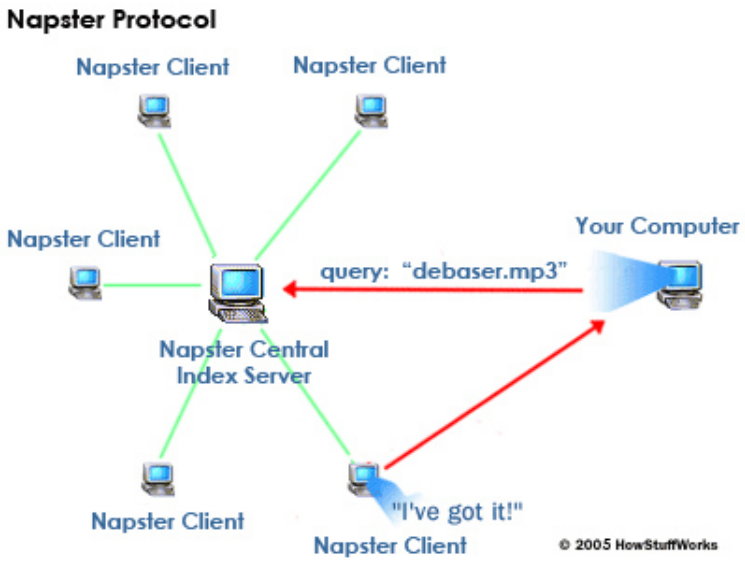


Abbildung 12.1: Napster Modell [35]

die Geschwindigkeit der Clients die Dateien zur Verfügung stellen. Wenn der Client eine der in den Suchergebnissen darin enthaltenen Dateien herunterladen will, schließt er eine direkte Verbindung mit den Clients und fängt den Download an. Die Datei selbst geht niemals durch oder wird auf dem Napster-Server gespeichert. Dies ist der P2P-Aspekt vom Protokoll.

Seti@Home [16] Seti@Home ist Teil des Suchvorgangs nach Außerirdischer Intelligenz, die von über zwei Millionen knirschenden Computern kommt. Die gesammelten Daten werden vom Arecibo-Radioteleskop in Puerto Rico heruntergeladen. Die Einrichtung dieses Projektes kostete Berkeley $500,000, es produziert über 15 teraflops [11]. Als Vergleich produziert, ASCI white, der mächtigste Supercomputer der Welt, 12 teraflops mit einem Kostenpunkt von $110 [12] Millionen. Das SETI@Home-Projekt wird als der schnellste Computer der Welt betrachtet. In der Tat hat das Projekt schon die größte sammelnde Berechnung bis zurzeit ausgeführt [7].

Vom Architekturstandpunkt gesehen, basiert Seti@Home auf den Client-Server Ansatz. Die zentralisierten Server halten enorme Datenmengen, die vom Arecibo-Radioteleskop gesammelt werden. Jene Daten müssen auf charakteristische oder ungewöhnliche Radiowellen genau untersucht werden, die auf außerirdische Nachrichttechniken andeuten könnten. Der Server vereint die Daten zu kleinen Paketen, die von den Clients heruntergeladen und genau untersucht werden. Anschliessend werden die Ergebnisse hochgeladen und überprüft. Die Server verfügen über eine hohe Bandbreite und genügende Verarbeitungsleistung, um die Ergebnisse der Clients nach "Fakes" überprüfen zu können.

Die Clients benötigen geringe Funktionalität. Sie haben die Fähigkeit, Kalkulationen auf den bereitgestellten Daten auszuführen, und können mit den zentralen Servern kommunizieren. Die Clients führen die Kalkulationen kontinuierlich mit einer geringen Priorität aus und kommunizieren nur mit dem Server, um Ergebnisse zu schicken und neue Daten anzufragen.

Diese Verhaltensweise scheint nicht P2P zu sein. Es sieht viel mehr aus, wie eine alte asymmetrische Client/Server Architektur, in der Clients Daten herunterladen, bearbeiten und dann zurück auf den zentralen Server hochladen. Es existieren keine Client/Client oder überhaupt P2P-Kommunikationen. Wenn man den Ablauf genauer analysiert, wird allerdings klar, dass die Clients nicht einfach stumme Browser sind. Sie nehmen eine aktive Rolle im Funktionieren des Netzes wahr. In der Tat würde das Netzwerk ohne sie nicht ganz funktionieren.

Dezentralisiertes P2P

Freenet [17] Im Juni 1999 beendete der Student an der Edinburg Universität (Schottland), namens Ian Clarke sein letztes Jahres-Projekt, Freenet, und stellte es der Welt übers Netz zur Verfügung. Das Projekt beschrieb ein verteiltes dezentralisiertes P2P resource-sharing Netzwerk, dass sich auf Anonymität konzentrierte, ohne jegliche Art von Steuerelementen. Zurzeit ist die Anwenderquote von Freenet relativ gering, da es schwierig ist, Ressoucen zu finden. Forschungen von Orwant [19] haben ergeben, dass nur 50% des Angebotes von Freenet legal ist, und ein grosser Teil der Anwendungen gegen die Copyrightrechte verstossen.

Freenets Architektur ist vollkommen dezentralisiert und verteilt, d.h., dass es keine zentralen Server gibt und dass alle Berechnungen und Dialoge zwischen Clients geschehen. Auf Freenet sind alle Kommunikationen zum Netzwerk gleich. Clients, die sich mit Freenet verbinden, werden nach dem Zufallsprinzip an den verfügbaren Client angeschlossen, dadurch entsteht eine unorganisierte zerstreute Topologie.

Die Kommunikation in Freenet Abbildung 12.2 entsteht beim Schicken einer Anfrage an die Arbeitsstation, mit der man verbunden ist. Es wird der Reihe nach an die dazu verbundenen Arbeitsstationen geschickt. Wenn ein Client ein Paket von einem anderen User aufnimmt, weiss man nicht, von wem und von wo das Paket kommt. Freenet erlaubt die Funktionalität “Ressourcen ins Netz einzufügen” und sie zu suchen und zu finden.

Um Ressourcen ins Netz einzufügen, werden diese gekennzeichnet, mit der Benutzung des Secure Hash Algorithm, SHA-1. Das SHA-1 generiert einmalige Schlüssel, die die Ressource identifiziert. Die Ressourcen werden dann mit jenem Schlüssel in Verbindung gebracht und örtlich gespeichert. Zum Beispiel könnte eine Textdatei mit politischen Anmerkungen nach den Suchkriterien Politik/Schweiz/aktuell, den Schlüssel MKL haben.

Bei Freenet muss jeder Benutzer auf seinen Rechner einen Speicheranteil zur Verfügung stellen. Dateien können somit in das Netzwerk eingefügt werden. Jedoch kann der Speicherort der Datei durch die Nachfrage ändern. Die Anonymität ist oberstes Gebot: Niemand weiss welche Teile von Dateien in seinem Rechner temporär gespeichert sind.

Anfragen und Uploads werden mit einer HTL (engl. Hops to live in Anlehnung an Time To Life, dass verhindert, dass unzustellbare Pakete unendlich lange weitergeroutet werden, das heißt: Wie oft darf noch weitergeleitet werden?) ausgestattet, die nach jedem Weiterleiten um 1 verringert wird. Es gibt eine obere Grenze für den Startwert, damit das Netzwerk nicht durch Aktionen mit unsinnig hohen Werten belastet wird. Der derzeitige

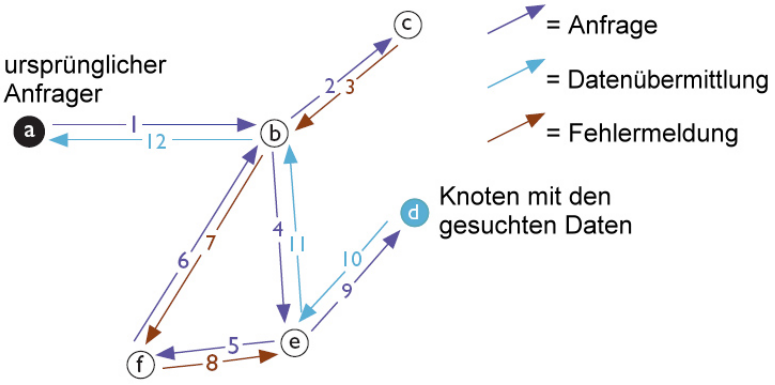


Abbildung 12.2: Freenet Modell [36]

Wert liegt bei 20: Wenn eine Anfrage nach so vielen Hops kein Ergebnis liefert, ist der Inhalt wahrscheinlich nicht vorhanden - oder das Routing funktioniert nicht, dagegen helfen höhere HTL aber auch nicht. Der Suchvorgang endet auch wenn ein Knoten während des Suchvorgangs zweimal gesehen wird, oder wenn die benötigten Ressourcen gefunden werden. Ähnliches gilt für das Hochladen: Nach 20 Hops sollte eine Information ausreichend verbreitet sein.

Ein kurzes Beispiel: Wir suchen den Schlüssel HGS. Wir sind mit anderen Freenet-Knoten verbunden, die die folgenden Spezialisierungen haben: ANF, DYL, HFP, HZZ, LMO. Wir wählen HFP als Adressaten unserer Anfrage, da dessen Spezialisierung dem gesuchten Schlüssel am nächsten kommt. Wenn der Adressat den Schlüssel nicht in seinem Speicher hat, wiederholt er die Prozedur, als ob er selbst den Schlüssel haben wollte: Er schickt die Anfrage weiter an den Knoten, der seiner Meinung nach am besten darauf spezialisiert ist und so weiter. Bei einem Treffer startet der Client mit den Ressourcen die Kommunikation. Alle Clients entlang des Weges werden die übermittelten Daten in den Speichercache aufnehmen, was zur Replikation von populären Ressourcen hilft. D.h., dass häufig verlangte Daten gecached sind, sich mehrmals im Netzwerk befinden und sich somit die Zugriffszeiten sich reduzieren.

Die Anonymität ist gewährleistet, da kein Client weiss, wenn die Ressourcen, die er herunterlädt, vom ursprünglichen Anbieter kommen und nach dem ursprünglichen Anfrager gehen, oder ob diese von einem anderen Verknüpfungsmitglied aus kommen oder gehen.

Eine wichtige Tatsache des Freenets ist, dass das File Sharing wie bei Napster nur eine der Nutzungsmöglichkeiten ist. Ian Clarke stellte sich ursprünglich vor, Freenet als Hilfsmittel zur Veröffentlichung und Abfrage von politischen Diskussionen und sensitive Angelegenheiten zu nutzen, wo Anonymität eine wichtige Rolle spielte.

GNUtella [18] Am 14. März 2000 veröffentlichte Nullsoft, eine Tochtergesellschaft von America Online, ein File Sharing- Anwendung, namens GNUtella, die Dateiauslagerung ohne Notwendigkeit eines zentralen Indizierungsserver, erlaubte. Frankels Arbeitgeber AOL zwang ihn jedoch am 10. April das Projekt aufzugeben und das Programm nicht weiter zu veröffentlichen. Es wurde jedoch von Tausenden von Anwendern vom Netz heruntergeladen, so dass es zu unzähligen Versionen kam.

GNUtella's Architektur ist ähnlich wie die von Freenet. Beide sind vollkommen dezentralisiert und verteilt, es gibt keine zentralen Server und alle Berechnungen und Dialoge passieren zwischen Arbeitsstationen.

Die Anonymität wird jedoch hier etwas benachteiligt, denn meistens wenn ein Packet geschickt wird, fängt es mit einer Time To Live (TTL) von 7 an. Aus diesem Grund erkennt man, dass wenn Client ein Packet mit TTL = 7 aufnimmt, dieses von seinem Upstreamnachbar kommt.

Beim Suchen in GNUtella schickt der Client seine Anfrage an alle seine Nachbarn. Diejenigen schicken die Anfrage weiter, wieder an alle Nachbarn. Dieser Vorgang geht weiter bis die TTL des Packets erreicht worden ist.

Auf eine Anfrage erstellen die gefragten Arbeitsstationen ein Paket, mit den notwendigen Informationen zur Lokalisierung der URL. Die Antworten laufen den Suchpfad zurück. Schließlich werden alle Ergebnisse zurück an den Client geschickt, der sie ursprünglich hinausschickte. Der Client entscheidet, ob er eine der erhaltenen Dateien herunterladen möchte.

Um eine Datei herunterzuladen, richtet der Client eine Direktverbindung zu einem anderen Client, der die Datei hat, und schickt ein HTTP-Paket, um diese Datei zu verlangen. Die Arbeitsstation mit der gesuchten Datei interpretiert dies und schickt eine standardmäßige HTTP-Reaktion. Allerdings beseitigt dies die Anonymität im System.

Diese direkte HTTP-Verbindung ist einer der Unterschiede zum Freenet. Ein weiterer ist, dass die Anfrage nicht an alle Nachbarn geschickt wird, sondern an den, der am ehesten in Frage kommt.

Fasttrack - KaZaA FastTrack ist eine der neueren Versionen des P2P. Es handelt sich hier um eine Architektur, die skalierbar ist und ein dezentralisiertes Design besitzt. Das FastTrack-Protokoll wird gegenwärtig von zwei Files Sharing - Applikationen KaZaA [31] und Morpheus [32] benutzt.

Die FastTrack-Architektur Abbildung 12.3 folgt einem 2-Stufen System, in dem die erste Stufe aus schnellen Kommunikationen zum Netz, und die zweite Stufe aus langsameren Kommunikationen zum Netz besteht. Clients auf der ersten Stufe werden als Superknoten und die auf der zweiten Stufe als Knoten erkannt. Diese Architektur lässt sich als Zweistufentopologie definieren.

Das Routen wird auf FastTrack durch das Broadcasting zwischen den Superknoten erfüllt. Wenn ein Knoten eine Suchabfrage an den Superknoten ausgibt wird es zum Beispiel an der Suchabfrage des Superknoten angeschlossen und dann an alle Superknoten, die, der Reihe nach, momentan verbunden sind, weitergeleitet. Der Suchvorgang wird weitergeführt bis der TTL null wird. Jeder Superknoten, der erreicht wird, durchsucht ein Register, das alle Dateien seiner verbundenen Knoten enthält. Wird das gewünschte File gefunden, wird es direkt von dieser Stelle (Client mit der Ressource) an den Anfrager via HTTP geschickt, es läuft nicht über den Superknoten zurücklaufen.

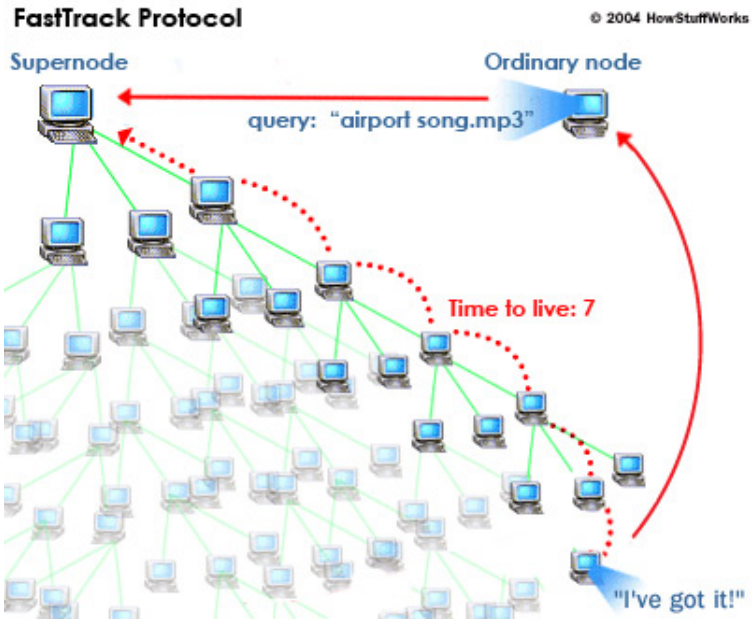


Abbildung 12.3: Kazaa Modell [35]

Durch das Broadcasting werden enorme Datenmengen erzeugt, die von Superknoten zu Superknoten weitergeleitet werden müssen. Da bei Superknoten Schnelligkeit gewährleist ist, entstehen keine großen Probleme wie bei GNUtella.

12.4 Peer-2-Peer Business Model

In diesem Abschnitt werden zuerst Anwendungsbereiche von P2P-Business Modellen vorgestellt. Danach werden einige Ideen angeschaut wie solche Modelle aussehen können und wie sie spezifischer schon eingesetzt werden. Am Schluss werden die Hindernissen von Business Modellen anhand der P2P-Technologie analysiert und Gründe besprochen wieso es heutzutage nur wenige etablierte P2P Business Modelle gibt.

12.4.1 Anwendungsbereiche

Die P2P-Technologie kann in sehr unterschiedlichen Bereichen zum tragen kommen. Anbei werden die bekanntesten dieser Anwendungsbereiche kurz vorgestellt.

P2P im kleinen Heim- oder Firmennetz

Einfache P2P Netzwerke können in Heimnetzen und Firmennetzen eingesetzt werden. Dabei handelt es sich meistens um eine kleine Anzahl von Teilnehmern. Diese Alternative ist kostengünstig, weil die Aufgaben des Netzes an die angeschlossenen Teilnehmer verteilt werden und keine Kosten für zusätzliche Server anfallen, und weil sie relativ einfach zu

verwalten sind. Dagegen wird der Arbeitsplatzrechner langsamer und der Zugriff auf Daten und Dienste wird schwieriger, da dieser nicht zentral verwaltet werden kann.

Filesharing in P2P-Netzwerken

Unter Filesharing verstehen wir das Austauschen von Dateien über ein Netzwerk. Dieses Einsatzgebiet ist heutzutage wohl das prominenteste Beispiel wie P2P-Netzwerke genutzt werden. Die Einfachheit in der Benutzung hat diesen Anwendungsbereich im privaten Umfeld sehr beliebt gemacht.

Im Jahre 1999 gewann zum ersten Mal eine Applikation zum Tauschen von Dateien namens Napster an verbreiteter Bekanntheit. Da vor allem Musik und Video Dateien über diese Wege getauscht wurden, rief es schnell die Unterhaltungsindustrie auf den Plan, welche versuchte Napster zu verbieten. Dass dies gelingen konnte, kann vor allem darauf zurückgeführt werden, dass Napsters Infrastruktur um einen zentral verwalteten Index Server gebaut war.

Da in den meisten Anwendungen jedoch mangelnde Zuverlässigkeit bezüglich Qualität und Sicherheit herrscht, ist der praktische Einsatz im Unternehmen in der bekannten Form weniger geeignet. Ausgereifte Business-Modelle tun sich deswegen auch schwer in diesem P2P-Anwendungsbereich Fuss zu fassen.

Instant Messaging und Internet-Telephonie

Unter Instant Messaging (IM) verstehen wir den Austausch von Informationen zwischen zwei oder mehreren Benutzern, wobei dies in Echtzeit geschieht. Die bekanntesten Beispiele von IM sind der MS Messenger und ICQ [24]. Dabei ist die Architektur im Sinne eines hybriden P2P-Netzwerkes aufgebaut. Es gibt einen zentralen Server über den sich die Benutzer anmelden und auf dem die Kontaktdaten der Mitglieder gespeichert sind. Der Server erstellt dann auch eine Verbindung mit dem kommunizierenden Peer her.

Das Routing der Nachrichten findet jedoch, in den meisten Systemen dieser Art, über den Server statt. Somit kann nur beschränkt von der Rede einer P2P-Architektur sein.

Während sich das IM noch hauptsächlich auf den Austausch von Nachrichten beschränkte, ist die Internet-Telephonie eine Weiterentwicklung, die auch nur durch die allgemein höhere Bandbreite im Internet entstehen konnte.

Ein Beispiel dafür ist Skype [25], welches durch die Kazaa-Gründer entwickelt wurde. Es ermöglicht ein Anruf von Rechner zu Rechner. Die Sprachübertragung mit Voice over Internet Protocoll (VoIP) als Datenpakete ist wesentlich günstiger als ein normales Gespräch via Festnetz oder Funk. Bei einer Paketvermittlung muss nicht ein ganzer Übertragungskanal für das Gesprächs besetzt werden. Durch die Einbindung von Gateways wird es auch möglich, einen Rechner mit dem klassischen Telefonnetz oder mit dem mobilen Funknetz zu verbinden.

P2P-Groupware und Collaboration Tools

Als Groupware oder Collaboration Tool bezeichnet man Software, zur Unterstützung der Zusammenarbeit zwischen verschiedenen Benutzer und Gruppen, die räumlich und zeitlich getrennt arbeiten. Das prominenteste Beispiel der Applikationsgruppe Groupware ist Lotus Notes [26], welches von IBM aufgekauft wurde.

Traditionell ist mit Groupware Videokonferenz und das zeitgleiche Bearbeiten von Dokumenten gemeint. Doch heute umfasst der Begriff weit mehr als das. Eine umfassende Groupware-Software enthält z.B. Email Client mit Kontaktverwaltungsfunktionalitäten, Instant Messaging, Terminplanungsfunktionalitäten, usw. Ein bekanntes Beispiel, das auf P2P Technologien aufbaut, ist Groove von Groove Networks [27], eine Firma die 2005 für 120 Millionen Dollar von Microsoft übernommen wurde.

Durch die Nutzung der P2P-Technologie in diesem Bereich, können sich Benutzer spontan zu Gruppen verknüpfen und diese ohne höhere kontrollierende Instanz administrieren. Dies kann über Unternehmensnetzwerke hinweg geschehen, da es sich nicht der Struktur des sich im Moment befindenden Netzwerks unterstellt. So entsteht ein dezentrales Gruppenmanagement.

Grid Computing

Nicht nur Daten können via P2P Netzwerke geteilt werden, sondern auch CPU- Rechenleistung, Bandbreite oder Speicherplatz. Unter Grid Computing verstehen wir das Verknüpfen von dezentral stationierten, geographisch verteilten Rechnern in einem Zusammenschluss um ein geteiltes Rechnen zu erlauben. Bei diesem Zusammenschluss können Rechenleistungen und Speicherplatz zu einem grossen virtuellen Rechner zusammengeführt werden. Es wird dadurch möglich, grosse Rechenaufgaben auf verschiedene Computer zu verteilen und parallel auszuführen. Ein Beispiel solcher Anwendungen sind SETI@Home [28], welches sich mit der Suche nach Signalen ausserirdischer Intelligenz beschäftigt. Ein weiteres Beispiel wäre im Bereich der wissenschaftlichen Berechnungen Folding@Home [29], welches zur Berechnung von Proteinen eingesetzt wird.

Web Services

Ein Web Service ist eine Software-Anwendung, die eindeutig über eine Schnittstelle ihre Funktionalität zur Verfügung stellt und mit Standard-Internettechnologien aufgebaut ist. Standard-Internetprotokolle wie zum Beispiel HTTP und SMTP können für den Transport verwendet werden, wobei als Basissprache des Protokolls XML verwendet wird.

Die Web Service-Architektur baut auf dem UDDI-Framework auf. UDDI (Universal Description, Discovery and Integration) stellt einen globalen Verzeichnisdienst dar, in dem die Web Services definiert werden. Web Service Architekturen, die mit UDDI arbeiten sind in Form einer hybriden P2P-Architektur aufgebaut.

12.4.2 Beispiele einiger P2P Business Modelle

Es wurden die Einsatzgebiete und -Möglichkeiten von Anwendungen einer P2P-Architektur besprochen. Im folgenden Abschnitt werden einige ausgewählte Beispiele von P2P Business Modellen vorgestellt und besprochen.

Dabei wird vermehrt auf die Idee der Modelle eingegangen, als dass der monetäre Nutzen analysiert wird. Das Zweitere würde den Rahmen dieser Arbeit sprengen. Wenn die unterschiedlichen Business Modelle betrachtet werden, wird festgestellt, dass die meisten der Modelle auf einer hybriden P2P-Architektur aufbauen und es kaum reine P2P-Modelle gibt. Dies liegt einerseits daran, dass in einem Business Modell eine gewisse Kontrolle und Administration vorhanden sein sollte und dies am Besten durch eine zentrale Instanz gelöst wird. Andererseits muss die Vernetzung zuerst aufgebaut und gepflegt werden. Dies kann auch am Besten durch einen Index-Service geschehen, der zentralisiert ist.

Kommerzielles Filesharing

Den Ursprung zum Filesharing in einer P2P-Umgebung brachte Napster. Als diese Anwendung jedoch in dem alten Stil verboten wurde, kamen andere Online-Anbieter von kommerziellen Musikdateien auf. Das heute bekannteste Online-Musikgeschäft ist iTunes von Apple.

Bei den heutigen Modellen, die Musikdateien kommerziell verteilen, handelt es sich jedoch weitgehend um Client-Server-Anwendungen. Weedshare [8] ist ein Beispiel für ein funktionierendes Geschäftsmodell im Bereich der Online-Musik, welches auf der P2P-Technologie basiert. In dem System existieren so genannte Weedfiles. Diese Musikdateien können gratis runtergeladen werden und bis zu dreimal abgespielt werden. Bei dem gratis abspielen muss jedoch eine Verbindung mit dem Internet bestehen, damit die Entschlüsselung des Songs geschieht. Wenn einem der Song dann gefällt, kann er über die Weedsoftware gekauft werden. Dieser kann dann auf bis zu drei PCs abgespielt werden. Es muss keine Verbindung mit dem Internet mehr existieren. Die Weedfiles können dann über P2P-Systeme geteilt werden. Wenn ein Nutzer diese Files zur Verfügung stellt, profitiert er davon, falls ein anderer Nutzer diesen Song von ihm runterlädt und dann kauft. Die Kosten werden folgendermassen aufgeteilt:

- 50% der Kosten stehen dem Urrechtsinhaber der Datei fest
- 15% wird vom Weed-System für dem Service verlangt
- 20% gehen an den Benutzer, von dem der Song runtergeladen wurde
- 10% gehen an den Peer, von dem dieser Benutzer den Song hat
- 5% gehen an den Quellpeer, von dem der zweite Peer den Song hat

Mobile P2P Content Distribution

Durch die bessere Auslastung in P2P-Systemen lassen sich in ansonsten sehr ausgelasteten oder an Bandbreite beschränkten Netzwerken, Konzepte erstellen die trotz diesen Einschränkungen umsetzbar sind. So kann zum Beispiel Content Distribution auf einzelne Geräte verlegt werden und muss nicht von einem zentralen Rechner aus verteilt werden. Dadurch können Geräte, denen es nicht möglich ist, eine direkte Verbindung mit dem ursprünglichen Verteiler herzustellen, trotzdem mit dem gewünschten Informationen versorgt werden, indem direkt mit anderen sich in der Nähe befindenden Geräten kommuniziert wird.

Die Nutzer, welche Content zur Verfügung stellen, können dann durch den Provider einen verbilligten Service angeboten bekommen, als Entschädigung ihres Einsatzes. Es besteht jedoch die Gefahr, dass ein Overhead durch die Analyse des Verkehrs und der Abrechnung auf einzelne Nutzer entsteht. Dies würde den Service wiederum verteuern. Der Service würde auch nur in beschränkten Szenarien zum Einsatz kommen, nämlich dann wenn keine direkte Verbindung zu einem Content Server hergestellt werden kann. Dafür muss jedoch ein genug grosser Benutzerkreis gefunden werden, der auch bereit sein wird, höhere Kosten in Kauf zu nehmen.

Als erläuterndes Beispiel kann hier der Aufbau eines Wireless Netzwerkes auf der Autobahn Abbildung 12.4 genommen werden. Dabei wird jedes Auto als ein Peer gesehen, dass über eine Wireless-Antenne verfügt mit einem Range von 200m. In gewissen Abständen befindet sich eine sogenannte Charging Zone. In dieser Zone kann eine Verbindung mit dem Internet hergestellt werden. Solange sich ein Auto nicht in der Charging Zone befindet, wird es versuchen, die Informationen von Autos zu bekommen, die sich in der Reichweite der eigenen Antenne befinden. So wird ein P2P-Netzwerk aufgebaut. Jeder Peer kann eine Informationsanfrage auch weiterleiten bis eine Verbindung in der Charging Zone mit dem Internet hergestellt werden kann.

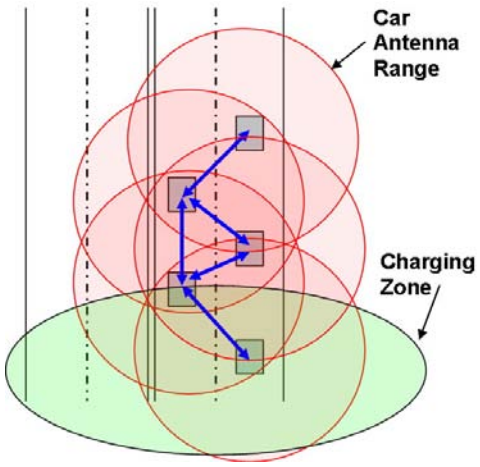


Abbildung 12.4: P2P auf der Autobahn

Bei diesem Modell müssen jedoch noch einige Aspekte berücksichtigt werden. Da das Szenario sich auf einer Autobahn abspielt, bewegen sich die Autos mit einer hohen Geschwindigkeit und das Netz wird sich sehr schnell wieder verändern. Es kann also keine

sehr grosse Datenmenge ausgetauscht werden. Ausserdem muss die Charging Zone genug lang sein, damit in der Zeit, in der ein Auto diese durchquert, genügend Informationen für sich selber und evtl. auch für Anfragen weiterer Teilnehmer zur ausgetauscht werden kann.

Das Szenario kann aber sehr interessant werden, wenn dieses Modell vor allem in Regionen aufgebaut wird, in denen viele Autos relativ nah zueinander stehen und sich nicht zu schnell bewegen.

Die Frage stellt sich nun, was für Informationen denn ausgetauscht werden könnten. Da lässt sich zum Beispiel vorstellen, dass aktuelle Strassenkarten für das GPS-System ausgetauscht werden können. Dies kann vor allem für reisende sehr interessant sein, da zusätzliche Informationen wie Daten von Hotels und Restaurants miteinbezogen werden können. Aktuelle Events und Veranstaltungen die in der Nähe stattfinden.

Es lassen sich weiter Informationen darstellen, wie Wetterlage, Stau- und Verkehrsinformationen, usw.

Business-to-Peer B2P

Als eine Erweiterung des Geschäftsmodells B2B kann die Idee des B2P (Business-to-Peer) oder Peer Portals gesehen werden.

Dabei treffen sich Unternehmen im virtuellen Raum um miteinander direkt Informationen und Services in standardisierter Form auszutauschen. Es muss nicht zwingend auf eine zentralisierte Serverstruktur zugegriffen werden. Die Teilnehmer können sich direkt miteinander verknüpfen und als Peers agieren. Wie schon erwähnt ist es sehr schwierig ein reines P2P Business Modell aufzubauen. Deswegen werden im Bereich B2P meist hybride P2P-Architekturen vorhanden sein, bei denen es zumindest einen Index-Server gibt, bei dem die Business Teilnehmer sich ausschreiben können.

12.4.3 Gefahren und Hindernisse bei P2P-Business Modellen

Momentan gibt es nicht allzu viele P2P Business Modelle, die sich in der Wirtschaft wirklich etablieren konnten. Im folgenden Abschnitt werden einige Aspekte besprochen, die die Ursachen begründen können.

Sicherheitsaspekte

Der Einsatz von P2P Technologien birgt viele Sicherheitsrisiken in sich. So werden bei den meisten IM-Tools keine Verschlüsselung der übertragenden Daten implementiert. Durch ein aufzeichnen der Kommunikation können Informationen direkt aus der Übertragung herausgefiltert werden.

Da in P2P Netzwerken keine zentrale Instanz agiert, ist generell keine einheitliche Security Policy vorhanden. Dadurch wird die Gefahr durch Malware infiltriert zu werden sehr gross, denn hat sich eine Malware einmal im Netzwerk installiert und verbreitet, ist es sehr schwer diese wieder zu entfernen. Jeder Peer muss sich um seine eigene Sicherheitsinstallation kümmern.

Der Einsatz von P2P-Systemen bringt indirekte Sicherheitsprobleme, wenn in Unternehmen die Mitarbeiter unsichere P2P-Systeme brauchen. So werden zum einen unnötig Bandbreite verschwendet, die keinen Business Value aufweisen. Zum anderen können durch diese Systeme die teuer aufgebauten Firewalls Abbildung 12.5 ausgeschaltet werden, indem durch Ports kommuniziert wird, die eigentlich für speziell vertrauenswürdige Applikationen reserviert sind und die nicht durch die Firewall analysiert werden.

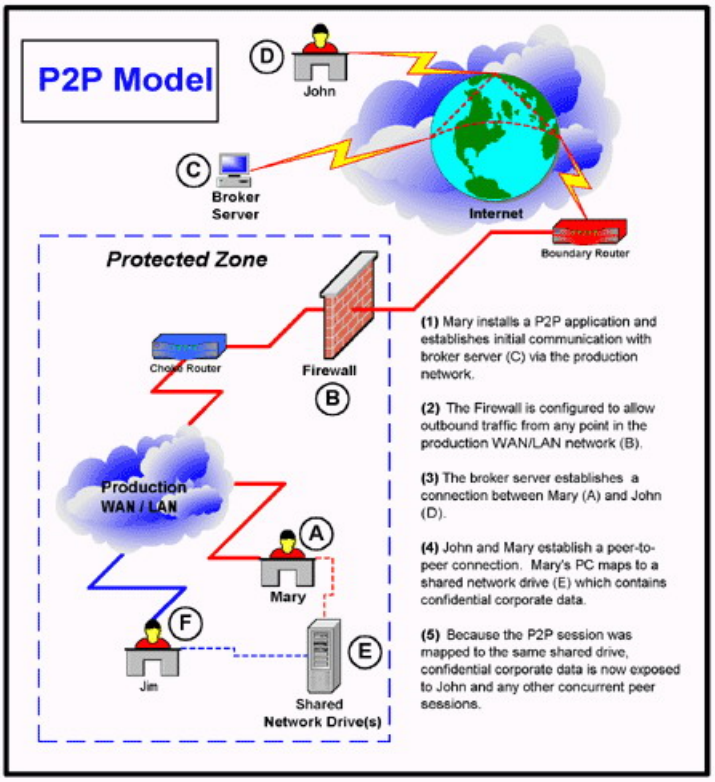


Abbildung 12.5: P2P Modell [34]

Die Kombination von P2P-Systemen und VPN in einem Unternehmen können schwere folgen haben (VPN steht für Virtual Private Network). Dabei wird von einem Rechner in einem ungeschützen Netz (Internet), eine gesicherte Verbindung mittels einem Tunnel, in ein geschütztes Netz gemacht. Durch die VPN-Verbindung befindet sich der Rechner auf logischer Ebene innerhalb des Netzwerkes und es wird keine Überwachung der Kommunikation zwischen dem externen Rechner und den anderen Instanzen des Netzwerkes betrieben. Dadurch wird die Wirkung der Firewall hintergangen. Wenn jetzt auf einem Rechner, welcher durch ein VPN mit dem Unternehmen verbunden ist, ein Verzeichnis oder gar das ganze virtuelle Laufwerk des Unternehmens durch ein P2P-System freigegeben wird, kann dieses durch die ganze P2P-Gemeinschaft zugänglich gemacht werden.

Einer Studie von AssetMetrix Research Lab [9], in der über 175000 Unternehmen untersucht wurden, kamen folgende Ergebnisse zustande

- P2P Applikationen wurden in etwa 77% der Firmen gefunden
- Einige Unternehmen hatten bis zu 58% ihrer Computer mit P2P-Applikationen
- In keinem der Unternehmen die mehr als 500 PCs hatten wurde keine P2P-Applikation gefunden

Deswegen wird sehr stark empfohlen, dass das Thema P2P in der Sicherheitspolitik eines Unternehmens sehr stark gewichtet wird. Es sollten aktive Massnahmen gegen den unerlaubten Einsatz von P2P-Systemen im Unternehmen getroffen werden.

Kontrolle und Transparenz

Eine grosse Herausforderung der P2P-Architektur ist die Kontrolle und Transparenz in der Nutzung innerhalb eines Netzwerkes. Daran scheitern momentan auch die meisten Business Modelle. Dadurch, dass versucht wird, eine zentrale, kontrollierende Instanz zu vermeiden wird die Analyse des Netzwerkes erschwert. Somit kann nicht genau definiert werden, wie die Last des Gebrauches auf die einzelnen Peers abgewälzt werden kann, um diese dann mit einem fairen System zu verrechnen.

Technische Probleme

Bei einem ersten Einloggen in ein P2P-Netzwerk muss zumindest ein Peer bekannt sein. Dies ist vor allem bei reinen P2P-Netzwerken problematisch. Deswegen werden bei den meisten Modellen mit Verzeichnissen auf zentralen Servern gearbeitet. Ein weiteres Problem kann entstehen, wenn sich Benutzer oft an- und wieder aus dem Netzwerk abmelden. Dadurch kann kein konstantes Netz geschaffen werden und die Ressourcen müssen immer wieder von neuem gesucht, und die Verbindung zu diesen aufgebaut werden.

Organisatorische Probleme

Durch das grosse Angebot an unterschiedlichen Ressourcen, entsteht das Problem des Auffindens der richtigen Information. Da die P2P-Netzwerke mehrere Millionen Nutzer haben können, erhöht sich damit auch der Datenverkehr bei einer Suche nach den richtigen Daten. Eine Verminderung dieses Problems, wird mit einer feiner Metadatenstruktur angesteuert. Dadurch verringert sich der Datenverkehr und die Antwortmenge für eine bestimmte Abfrage wird genauer.

Soziale Effekte

Die P2P-Systeme können eigentlich nur in einer gleichgesinnten Gemeinschaft überleben. In dieser Gemeinschaft agiert jeder als ein Peer (Gleichgestellter, Ebenbürtiger). Wie in vielen Gemeinschaften, entsteht jedoch das Problem der Trittbrettfahrer. Nutzer können auf die Kosten der Gemeinschaft kommen, ohne dabei selber eine Leistung zu erbringen. Einer Studie [30] zufolge, gibt es in den gängigen P2P-Filesharing Applikationen bis zu 70% von Benutzer, die keine Dateien teilen.

Der Nutzen eines P2P-Netzwerkes erhöht sich jedoch mit der Anzahl vollständig beteiligten Peers. Dadurch werden mehr Rechner verknüpft und können Beispielsweise mehr Dateien ausgetauscht werden.

12.5 Fazit

P2P Systeme stehen immer noch grossen Herausforderungen gegenüber. Wie dies in der Arbeit aufgezeigt wurde, stehen ausgereiften P2P Business Modellen Probleme bei der Umsetzung von Sicherheitslösungen im Wege. Wir haben auch weitere Aspekte gesehen, die in der technischen Umsetzung noch fehlen, wie eine verbesserte Auffindung der Ressourcen durch nützliche Metadatenstrukturen. Bei vielen Modellen ist es sinnvoll, das Netzwerk für eine beschränkte und übersichtliche Reichweite der angedockten Peers aufzubauen. Dies erlaubt einerseits eine gewisse Kontrolle über das P2P-Netzwerk zu haben und andererseits kann dadurch der Verbindungsaufbau mit den gewünschten Peers effizienter gestaltet werden.

Gartner hat 2001 eine Einschätzung der P2P Technologie für das Jahr 2010 gemacht. Darin besagte diese, dass P2P-Applikationen weitgehend eine Nischenrollle einnehmen werden und die CS-Architektur nicht ablösen würden.

Literaturverzeichnis

[1] *Definition P2P* http://searchnetworking.techtarget.com/sDefinition/0,,sid7_gci212769,00.html Definition P2P, 01.02.2006.

[2] *OpenP2P Homepage* http://www.openp2p.com/ Definition P2P, zuletzt besucht am 01.02.2006.

[3] *OpenP2P Homepage* http://www.openp2p.com/pub/a/p2p/2003/10/17/filesharing.html Filesharing P2P von Preston Gralla, 17.10.2003.

[4] *OpenP2P Homepage* http://www.openp2p.com/pub/a/p2p/2004/03/05/file_share.html Next-Generation File Sharing with Social Networks, von Robert Kaye 05.03.2005.

[5] *Answers.com*, http://www.answers.com/topic/peer-to-peer, peer-to-peer, zuletzt besucht am 20.01.2006.

[6] *InfoVis Cyberinfrastructure* http://iv.slis.indiana.edu/lm/lm-p2p-search.html Search Performance of P2P Networks, Version vom 08.08.2004.

[7] *NETWORKWORLD* http://www.networkworld.com/newsletters/techexec/2000/0828techexec1.html A peer-to-peer revolution, von Mark Eggleston 29.08.2000.

[8] *WeedShare.com* http://www.weedshare.com zuletzt besucht am 01.02.2006.

[9] *AssetMetrix* http://www.assetmetrix.com/news/pressReleases/press2003-07-16.asp AssetMetrix Research Labs reveals widespread P2P usage within corporations, Artikel vom 16.07.2003.

[10] *Peer-to-Peer Netzwerke und Geschäftsmodelle* http://www.fhso.ch/pdf/publikationen/dp04-04.pdf von Rolf Dornbergerge und Daniel Fuchs, Dezember 2004.

[11] *Seti@Home* http://setiathome.ssl.berkeley.edu/faq.html Frequently Asked Questions, zuletzt besucht am 28.11.2006.

[12] *Accelerated Strategic Computing Initiative (ASCI)* http://www.llnl.gov/asci/news/white_news.html ASCI-Weiß-Nachrichten, Artikel vom 29.06.2000.

[13] *PDOS Publications* http://pdos.csail.mit.edu/papers/chord:sigcomm01/ Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications, von Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan August 2001.

[14] *CacheLogic* http://www.cachelogic.com/p2p/p2punderstanding.php Understanding peer-to-peer, zuletzt besucht am 05.01.2006.

[15] *Napster.com* http://www.naspter.com/ Napster's page, zuletzt besucht am 01.02.2006.

[16] *SETI@Home* http://www.seti.org/ SETI Institute, zuletzt besucht am 01.02.2006.

[17] *The Freenet Project* http://freenetproject.org/ zuletzt besucht am 01.02.2006.

[18] *GNUtella's community* http://www.gnutella.com/community/ zuletzt besucht am 01.02.2006.

[19] *OpenP2P Homepage* http://www.openp2p.com/pub/a/P2P/2000/11/21/freenetcontent.html Was auf Freenet ist, von Jon Orwant 21.11.2000.

[20] *Communication Systems Group Homepage* http://csg.ifi.unizh.ch/teaching/ Prof. Dr. Burkhard Stiller: KV Verteilte Systeme und Kommunikation /Winter Semester 2005/ Modul 9.

[21] *IFPI Zürich, 15. November 2005 - Musikwirtschaft: "Game Over" auch für private Raubkopierer* http://www.ifpi.ch/docs/news.html\#20051115a

[22] *Lüthi, Nick: Der Urheberrechtler; Interview mit Bernhard Wittweiler; Espace Mittelland; 28.02.2004*

[23] *Kampf gegen Internet-Kriminalität; Bundesamt für Polizei* http://www.bap.admin.ch/d/archiv/medien/2003/08151.htm Artikel vom 15.08.2003.

[24] *ICQ* http://www.icq.com ICQ Homepage, zuletzt besucht am 01.02.2006.

[25] *Skype* http://www.skype.com Sype Homepage, zuletzt besucht am 01.02.2006.

[26] *Lotus Notes* http://www-306.ibm.com/software/lotus Lotus Notes Homepage, zuletzt besucht am 01.02.2006.

[27] *Groove Virtual Office* http://www.groove.net/home/index.cfm Groove Virtual Office, zuletzt besucht am 05.01.2006.

[28] *SETI@Home* http://setiathome.ssl.berkeley.edu/ SETI@Home Homepage, zuletzt besucht am 01.02.2006.

[29] *Folding@home* http://folding.stanford.edu/ Folding@home Homepage, zuletzt besucht am 01.02.2006.

[30] *Freeriding on Gnutella* http://www.firstmonday.dk/issues/issue5_10/adar/index.html von Eytan Adar and Bernardo A. Huberman am 27.09.2000.

[31] *Kazaa Homepage* `http://www.kazaa.com/us/index.htm` , zuletzt besucht am 01.02.2006.

[32] *Morpheus Homepage, Morpheus Server und Webseiten sind seit dem 27.01.2006 nicht mehr erreichbar* `http://www.morpheus.com/` zuletzt besucht am 26.01.2006.

[33] *Electronic Frontier Foundation, The Betamax Case: Sony Corp (USA) vs. Universal City Studios, 07.02.2006* `http://www.eff.org/legal/cases/betamax/` The Betamax Case, Zuletzt besucht: 09.02.2006.

[34] *SC Online Magazine* `http://athens.bitwise.net/isn/scmagazine/scmagazine/sc-online/2001/article/039/article.html` Artikel von Carlos Valiente Jr., September 2001.

[35] *howstuffworks Homepage, zuletzt besucht am 08.02.2006* `http://computer.howstuffworks.com/kazaa3.htm`

[36] *Wikipedia Homepage, zuletzt besucht am 08.02.2006* `http://de.wikipedia.org/wiki/Bild:Freenet_-_Ablauf_einer_Anfrage.png`