Cellular Assisted Heterogeneous Networking

Inauguraldissertation
der Philosophisch-naturwissenschaftlichen Fakultät
der Universität Bern

vorgelegt von
Marc Danzeisen
von Basel

Leiter der Arbeit:
Prof. Dr. T. Braun
Institut für Informatik und angewandte Mathematik

Cellular Assisted Heterogeneous Networking

Inauguraldissertation
der Philosophisch-naturwissenschaftlichen Fakultät
der Universität Bern

vorgelegt von
Marc Danzeisen
von Basel

Leiter der Arbeit:
Prof. Dr. T. Braun
Institut für Informatik und angewandte Mathematik

Von der Philosophisch-naturwissenschaftlichen Fakultät
angenommen.

Der Dekan:

Bern, den 02. Februar 2006 Prof. Dr. P. Messerli

Preface

The work presented in this thesis was performed during the years I spent as research and lecture assistant at the Institute for Computer Science and Applied Mathematics (IAM) of the University of Bern, Switzerland. The research work has been done in tight collaboration with Swisscom Innovations, who also funded most of the work that I describe in this dissertation.

I would like to thank all the people who have supported me throughout this work. Firstly, many thanks to Prof. Dr. Torsten Braun, head of the Computer Network and Distributed System group (RVS), for supervising this work and for his insightful advices.

I am also very grateful to Prof. Dr. Martina Zitterbart for having accepted to read and judge this work. I would also like to thank Prof. Dr. Hanspeter Bieri who was willing to be the co-examinator of this thesis.

I am particulary grateful to Beat Perny, Daniel Rodellar, Jan Linder, Walter Steinlin, and Urs Rötlisberger for the chance I got to join the Swisscom Innovations team. I would also like to thank Roger Lagadec and Stefan Mauron for all the fruitful discussions and projects in the domain of heterogeneous networking. Many thanks go to my colleagues of the RVS group for our various interesting discussions about all kinds of topics. Special thanks go to Florian Baumgartner, Attila Weyland, Ruy De Oliveira, Matthias Scheidegger, and Marc Steinemann. Special thanks also go to Ruth Bestgen for managing all administrative issues.

I especially would like to thank Simon Winiker, Ehsan Maghsoudi, Isabel Steiner, Felix Aeschlimann, Christian Cueni, and Amélie Saulnier who worked with me and helped a lot in developing and implementing.

Last but not least, I am deeply grateful to my family and friends for their support throughout my studies: Victor Danzeisen, Josian Danzeisen, Natalie Danzeisen, Ursula Grau-Danzeisen, Marc Heissenbüttel and Manuel Haag.

This thesis is dedicated to Tina Wasserfallen for all the support, courage and strength she gave me during the last few years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The nature of the Internet enabled a tremendous variety of data applications to be developed, especially in the last decade due to the exponential increase in Internet users. The requirements on the underlying network are often strongly depending on the type of application, terminal, and location. Some applications make sense only if high bandwidth is available whereas others may rather be dependent on the level of security or mobility offered by the network. The layered structure of the ISO/OSI communication stack has enabled high flexibility by decoupling the underlying network technologies from the applications to certain extend. However, most of the applications are not independent of the communication characteristics provided by the underlying network. Applications often require a certain level of bandwidth, delay or security to work properly. Analyzing further these applications reveals that their requirements also depend on the user's situation. The level of mobility has an impact on the way applications (or services) are used. The capabilities of the devices used in the different situations is very much influencing the requirements on the service and therefore on the underlying network. A single network can not cope anymore with the different and often changing requirements. The development and deployment of the 3G cellular network clearly showed the complexity to build up one network, which fits all the requirements of nowadays and future mobile applications. To optimally meet the applications need, different communication technologies have to collaborate.

This heterogeneity also raises new challenges on the mobile terminal side. First, the management of the various networking technologies and devices should be hidden from the end user. Secondly, having several alternatives available to connect nodes to the Internet or to other nodes requires an intelligent network selection. Today, it is often up to end user to select the network of choice and initiate the connection. There are situations, especially when thinking about mobile applications, where this manual network selection is far from being optimal. This can be due to lack of knowledge about the different available technologies or simply because the requirements of an application may change during an ongoing session. Additionally, the networking environment of the peer node has to be considered as well to assure the most appropriate network selection.

Dynamically assigning the most appropriate networking resources to each node, depending on its actual needs and capabilities, may increase the customer satisfaction and the overall performance of the heterogeneous network at the same time. Dynamic resource handling is further required to cope with mobility. Mobility poses an additional challenge by constantly changing the network environment, e.g., used networks become unavailable or new ones get detected. Communicating nodes can come close enough to establish direct links based on short range and infrastructure-less communication technologies. These links have the potential to deliver data rates which are orders of magnitude higher than infrastructure-based links will ever do. The integration of infrastructure-less connections between communicating nodes to enhance the performance is very challenging. The missing infrastructure to properly and automatically negotiate required parameters to establish a secured and optimized direct links is imposing some major hurdles that have to be taken before users can benefit from higher performance without compromising security.

Even though, heterogeneity will help to overcome many limitations imposed by homogeneous network, there are a lot of challenges to face to assure that both, the end user and the network operator can benefit from it.

1.2 Motivation

Despite big efforts done by the manufacturers to make the handling of the different technologies simpler and more user friendly, the increasing number of different communication technologies makes it nearly impossible for user to properly handle them. Furthermore, the different technologies are not designed for seamless interworking. Heterogeneity is therefore often perceived as a hurdle instead of an enabler for being always connected.

The dynamic selection and configuration of the most appropriate technology is by far too complex for the end user. Different communication technologies require often different settings and therefore a certain level of knowledge is needed to successfully connect them. This is especially true for direct node to node communication technologies, where no centralized system is available to manage the communication setup. Several parameters have to be set before communication can occur. When talking about secure communication the setup gets even more complicated due to the additional key negotiation and management. Because of this complicated connection establishment such direct node-to-node links are often not used at all, although they would provide data rates order of magnitude higher than infrastructure-based links.

Additionally, depending on the movement of the nodes and the applications in use, the initially chosen communication technology might become suboptimal or even inappropriate. This missing management of heterogeneous networking resources is not only preventing the users to profit from the different communication technologies, but also reducing the overall network performance. Suboptimal selection of networking technology not only degrades user experience but can also result in low network efficiency. Allocating expensive broadband network resources for applications requiring only small data rates for example, can considerably reduce the number of servable nodes.

Power management is a further problem in a heterogeneous network environment. Staying always connected to a wide range broadband data network is still not practicable due to power limitations of mobile devices. To safe energy, most data communication technology interfaces are only powered up if data has to be transferred. Consequently, nodes are not reachable if no data has to be transmitted. Communication technologies have first to be manually switched on before any end-to-end connection can be established.

In this thesis we address all the issues listed above to enable convenient and resource efficient heterogeneous data communication. The main resulting contributions are presented in the next section.

1.3 Contributions

The contributions of this thesis are highly related to the identified issues of handling heterogeneous data communication discussed in the prior section and can be summarized as follows.

To simplify the complexity of heterogeneous data communication, the proposed system is providing an abstracted view on the underlying communication technologies to the user. The abstraction is achieved by introducing logical sessions on top of the actual data sessions. Logical sessions are related to human-to-human interaction. Consequently, logical session endpoints are addressed with human readable identifiers, whereas the data sessions remain IP address based. The system architecture allows the utilization of mobile phone numbers to identify the logical session endpoints, making the establishment of heterogeneous data sessions as intuitive and simple as voice calls. The abstraction provided allows users to establish communication sessions to other users instead of addressing specific devices belonging to the user. The flexible handling of the relation between logical and actual data session endpoints enables users to dynamically choose the most appropriate device to be used for each communication session.

The system architecture proposed in this thesis allows the independent routing of signaling and data related information over different communication technologies. Unlike the inband signaling of IP, where the data channel has to be established before the first signaling messages can be exchanged and the ability to first exchange signaling information on a dedicated signaling channel is highly beneficial in heterogeneous environments. The separated treatment of signaling and data related information allows optimal resource management. During the data session establishment the nodes can learn about the networking capabilities and actual environment, such as the currently available networks at the peer's location.

Although our system architecture can be used to dynamically route the signaling messages over any type of communication technology, we argue that the use of the existing cellular mobile network as the primary signaling plane has several advantages.

First, the ability to reuse the human readable and widely accepted name space (i.e. the mobile phone number) as the primary identifier of communication end points, is very much facilitating the abstraction of heterogeneous data communication. Heterogeneous data sessions can be initiated by inviting peers based on their mobile phone number, for instance.

Second, the well established and power optimized location, paging, and mobility management services can implicitly be shared for other communication technologies and networks, which are lacking such functionality. Especially the low power characteristic of the cellular mobile network is beneficial when addressing the power management of heterogeneous communication. The ability to reach any node through the low power cellular network allows switching power demanding broadband data channels to sleep mode if no actual data session is going on. Reachability is no longer coupled with power demanding broadband IP connectivity. Waking up data channel interfaces only on-demand can considerably increase the power efficiency of the mobile devices.

Furthermore, the security relation between the cellular subscriber and the network operator together with the roaming relations between the operators, offer a secured communication channel which can be used to securely exchange our protocol messages. Especially if the establishment of infrastructure-less communication channels is considered, the secure exchange of configuration and security related parameters is absolutely mandatory to securely bootstrap the connections. Having established a secured initial communication between the communicating entities, all other parameters required to establish further communication channels can securely be negotiated. The concept of Cellular Assisted Heterogeneous Networking (CAHN) provides the missing part to securely extend the scope of heterogeneous networking. The ability of CAHN to securely bootstrap infrastructure-less communication between nodes enables the system to consider also direct node-to-node links when evaluating the most appropriate connection.

To evaluate the potential impact of the proposed CAHN concept on the performance of heterogeneous networking environments, simulations have been conducted. Our main interest was focused on the ability to wake up power demanding broadband interfaces on-demand and switch ongoing data sessions to direct node-to-node links, whenever the communicating nodes come close enough to each other. The simulation results showed that considerable improvements in terms of throughput, power consumption, network resource efficiency, and overall network capacity can be achieved with our presented CAHN architecture. The combination of both features, namely the possibility to switch unused IP interfaces to sleep mode and the ability to handover ongoing infrastructure-based data sessions to direct ad-hoc links, is reducing the energy consumption up to 80% and the network efficiency up to 40% in certain scenarios. In small areas like university or enterprise campus the average throughput can be increased by a factor of up to 4, if communicating nodes can automatically switch to direct node to node links.

The developed system and protocol allow the negotiation of any kind of configuration and security parameters and is therefore a promising framework to solve resource management problems on other layers as well. The framework will be used in further projects addressing the optimization of the dynamic radio resource allocation to increase the efficiency of radio spectrum. The concept of cellular and therefore operator assisted resource management raised high interest in the telecommunication community.

1.4 Thesis Outline

In Chapter 2, we analyze existing data communication technologies and protocols with respect to heterogeneous networking. The different technologies are quickly introduced focusing on the delivered performance, their availability, and how they are currently used by end users. This chapter aims at providing a better understanding of heterogeneous networking, and how each of the existing technologies can contribute to increase the value of heterogeneity. Chapter 3 provides a comprehensive overview of related work in the domain of seamless mobility, heterogeneous networks, and session management. The most popular published proposals are introduced in further detail and analyzed with regards to our envisioned heterogeneous networking architecture. In Chapter 4, we elaborate on the issues of managing heterogeneous IP sessions. The missing parts needed to optimize data communication using direct connections between mobile nodes are identified and we propose appropriate mechanisms to seamlessly integrate these high performance links into the vision of being always best connected. We conclude by introducing our new concept of logical and physical session management, which may help to solve most of the identified issues. Chapter 5 introduces our Smart Multi-Access Communications (SMACS) architecture and explains how it can deliver an abstracted view on underlying communication technologies. In this chapter we also introduce our simulator designed to quantify the improvement potential of our concept. In Chapter 6, the CAHN component, handling infrastructure-less connections, is presented. The CAHN protocol used to exchange configuration and security related connection parameters is presented together with the design, the implementation details, and the evaluation of the prototype. Chapter 7 summarizes the main findings and concludes the thesis and elaborates on potential improvements and gives an outlook on possible future work.

Chapter 2

Heterogeneous Networking

2.1 Introduction

The trend towards heterogeneous networking is mainly motivated by new emerging technologies delivering better performance for specific applications and the fact that no single technology meet all the different requirements. For the end users this heterogeneity could result in being always best connected. Depending on the actual requirements of the used application, the most appropriate technology could be automatically selected. Heterogeneity might also be a way for network operators to reduce capital expenses. Deploying different access technologies in specific locations can avoid expensive large scale deployment. Especially for emerging technologies with short life cycles, coming from the IT environment like WLAN, this dedicated deployment will have major impact on rollout costs and time to market. So finally, both the users and the network operators can benefit from heterogeneity. However, convenience will still be key, demanding for seamless integration of the different access technologies. Network operators will be interested in hiding the heterogeneity to avoid customers of getting tired of adopting new technologies. With the commercial deployment of wireless LAN, cellular network operators started to integrate this new access technology into their existing network. Emerging technologies like WLAN are disturbing the well planed evolution path of the legacy cellular networks and therefore forcing the operators to be innovative and flexible. Nevertheless, the integration of disruptive access technologies into the existing network is very challenging and happens stepwise. Early adopters start to use new technologies very early even if the level of integration is low. Those technically versed users compensate the missing convenience with technical know-how. With advanced level of integration, resulting in increased convenience, also less technically skilled users can benefit from the new communication technology, which further motivates network operators to push the integration process.

This chapter gives an introduction to existing communication technologies and protocols, which are related to heterogeneous data communication. The communication technologies and protocols are analyzed in terms of their potential contribution to enable seamless heterogeneous networking.

2.2 Overview of Communication Technologies

For the sake of better understanding a short overview of the mostly referenced data communication technologies and protocols is given in the following sections. The introductions are mainly focusing on characteristics which are relevant for the work done and presented in the scope of this thesis.

Throughout this document the different communication technologies are separately treated depending on their capability to establish connections without any infrastructure. Technologies, which can be used to directly interconnect nodes, are referred to as infrastructure-less, ad-hoc, or node-to-node. If any kind of infrastructure is required to establish a data communication the technology is called infrastructure-based. With regards to the vision of convenient and seamless heterogeneous networking presented in the prior chapter, it is interesting to analyze in further details the level of user interaction that is required to utilize the different communication technologies. When it comes down to the offering of secure session mobility on heterogeneous data networks, Mobile IP and IPsec are considered as the most promising candidates. Both protocols are introduced in the second half of this chapter.

2.2.1 Global System for Mobile Systems (GSM)

The GSM, as a representative of the second generation of mobile communication networks, was primarily designed to deliver voice services and not data services. Nevertheless, a dedicated mode for data transmission was defined as Circuit-Switched Data (CSD). The data rates offered by CSD were limited to $9.6\,kbit/s$. The rapidly increasing capabilities of the mobile phones stimulated data applications based on the Wireless Application Protocol (WAP) [143] and motivated the development of new technologies to increase the capabilities of the cellular system to deliver higher data rates.

High-Speed Circuit-Switched Data (HSCSD)

To enhance the transportation of data over the GSM air link, HSCSD was introduced. Like in the original Circuit-Switched Data (CSD), channel allocation is done based on circuit switched mechanisms. HSCSD enables the usage of different coding schemes which allow data rates up to $14.4\,kbit/s$ instead of the $9.6\,kbit/s$ delivered by the native CSD. Additionally to the better encoding schemes, HSCSD allows the aggregation of several time slots offering data rates up to $57.6\,kbit/s$. HSCSD requires the time slots being used to be fully reserved for a single user. It is possible that either at the beginning of the call, or at some point during a call, it will not be possible for the user's full request to be satisfied since the network is often configured so that normal voice calls take precedence over additional time slots for HSCSD users. Apart from the fact that the full allocated bandwidth of the connection is available for the HSCSD user, HSCSD also has an advantage in GSM systems in terms of low average transmission latency. However, for nowadays applications, demanding for high data rates, HSCSD is not appropriate anymore. Although its low latency and low resource utilization makes the HSCSD a valuable communication channel for specific low bandwidth applications. In chapter 4, 5, and 6 we will introduce a system where such low power and low bandwidth data channels can perfectly serve for the exchange of signaling messages.

Short Message Service (SMS)

SMS [3, 4] is defined by the 3GPP [2]. SMS defines types of messages, namely the Point-to-Point (SMS-PP) and the Cell Broadcast (SMS-CB). The former enables message delivery between mobile nodes, whereas the later addresses the broadcast of messages to all mobile users in a specified geographical area. SMS is based on the store-and-forward mechanism deployed on the Short Message Service Center (SMSC), which will attempt to send the message to the recipient and possibly retry if the user is temporarily not reachable. The SMS is best effort, meaning that the delivery of the message is not guaranteed. Message payload is limited to 140 *bytes*, either 160 7-bit characters, or 140 8-bit characters. Larger messages can be segmented over multiple messages. The standard permits up to 255 segments. SMS uses the signaling system of the GSM network and is therefore independent of ongoing voice or data sessions, which makes the SMS also interesting for the transmission of signaling messages.

Unstructured Supplementary Service Data (USSD)

Similar to SMS, the USSD [1] is based on the signaling system of the cellular network and supported by all GSM mobile devices. However, there is no store-and-forward mechanism available for USSD. Communication based on USSD is generally faster than SMS based message delivery. USSD is typically used as a trigger to invoke independent calling services such as interactive menuing services. USSD supports two types of sessions, terminal and network initiated. Due to its session based communication, USSD is much more reliable than SMS. The initiator of the session immediately gets notified whether the session is established or not. For exchange any information directly between mobile nodes some additional functionality needs to be added. Since USSD is not natively supporting node-to-node sessions, dedicated boxes have to intelligently combine terminal and network initiated sessions to form a sort of node-to-node session. However, USSD is a very promising technology to exchange signaling information between mobile nodes and network components. Like SMS the USSD support is mandatory for all mobile devices that access 2G and 3G networks and therefore ideal for the quick introduction of new services.

2.2.2 General Packet Radio Service (GPRS/EDGE)

To cope with the increasing demand for high data rates the ETSI standardized the GPRS. Later on the standardization efforts were handed over to 3GPP. GPRS is integrated into GSM standard releases starting with Release 97. In contrast to CSD, where a data connection establishes a circuit reserving the full bandwidth of that circuit during the whole session, GPRS is packet switched and therefore allows multiple users to share the same transmission channel. This results in dynamic assignment of the total available bandwidth to those nodes actually sending at any given moment. Especially for applications requiring intermittent data transfers benefit from sharing the available bandwidth. GPRS allocates unused time slots to provide data connections. Hence, the number of active voice connection in the cell, is determining the number of slots that can be assigned to GPRS sessions and therefore influences the achieved data rates. The theoretical limit for packet switched data is approximate 170 $kbit/s$, whereas a realistic bit rate is between 30 and 70 $kbit/s$. A change of the radio part of GPRS called Enhanced Data Rates for GSM Evolution (EDGE), allows the higher data rates of $20 - 200\, kbit/s$. The maximum data rate is extremely dependent on the coding scheme used and the number of assigned slots in the Time Division Multiple Access (TDMA) frame. Schemes with low error correction provide high throughput put require very good signal quality. Four encoding schemes (CS) are defined within the GPRS standard. The fastest (21.4 $kbit/s$) but least robust encoding scheme is CS-4 and only available near the Base Transceiver Station (BTS) while the most robust encoding scheme (CS-1) is offering the slowest data rates (9.05 $kbit/s$) and used for nodes that are far away from the BTS. Consequently, the connection speed drops with distance from the base station. This is not an issue in heavily populated areas with high cell density, but may become an issue in sparsely rural areas.

GPRS defines different classes of terminals. The classes are mainly determined by the number of down- and upload slots that can be used. GPRS class 8 provides 4 download and 1 upload slots (4+1). Class 10 is also known as 4+2, meaning 4 download and 2 upload slots. In the same way class 6 (3+2) and 4 (3+1) are defined. For industrial usage there is also the (4+4) class standardized. Although more than two upload slots are considered a health hazard for nearby user. Hence, GPRS is mainly used for asymmetric connections with higher data rates for download than for upload.

Since GPRS requires at least one time slot to provide IP connectivity, capacity might become an issue if all cellular subscribers want to benefit from staying connected and reachable for IP communications anytime and anywhere. Furthermore, the very limited data rates offered with GPRS compared to broadband technologies like WLAN prevents the adoption of GPRS for Internet applications. However, for dedicated applications requiring only low data rates like messaging or adapted e-mail notification systems used to provide user with partial information (e.g., without attachments), GPRS is still a reasonable way to go. Although, the capacity issue raising form the occupation of at least one time slot even if no traffic has to be sent or received might prevent its usage.

2.2.3 Universal Mobile Telecommunications System (UMTS)

Where GPRS and EDGE are often referred to as 2.5G and 2.75G, UMTS is clearly defined as 3G. UMTS is standardized within 3GPP and based on W-CDMA as the underlying radio standard. With the migration from TDMA to W-CDMA, data rates can be achieved of about 1920 $kbit/s$ in theory and 384 $kbit/s$ in practical experience. In most of the real networks the uplink is limited to 64 $kbit/s$, which makes UMTS preferably used for asymmetric application like Web browsing and content download. Similar to GPRS, nodes occupy resources if attached to the network, independent whether data is transmitted or not. In contrast to GPRS the resources are codes and not time slots. Due to the nature of W-CDMA transmitting nodes are increasing the noise level for other nodes resulting in a rapid degradation of available bandwidth in highly populated cells. W-CDMA defines different classes of codes, determining the achievable data rates. With the dynamic assignment of codes to the nodes, it is possible to equally share the available capacity among the nodes. The more nodes have to be served by a cell, the lower the class of codes assigned and the less capacity is available for each node. Consequently the usage of UMTS links to stay connected and hence reachable for IP sessions is not advisable due to waste of overall networking resources.

2.2.4 Wireless LAN (WLAN)

Wireless LAN was initially defined in the IEEE 802.11 [90] released in 1997 specifying two raw data rates 1 and 2 $Mbit/s$ to be transmitted via infrared (IR) signals or in the Industrial Scientific Medical (ISM) frequency band at 2.4 GHz. Although IR remains a part of the standard, it has no actual implementations.

In 1999 the 802.11b [92] was ratified providing a maximum raw data rate of 11 $Mbit/s$ using the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) media access method defined in the original standard. Due to the overhead of CSMA/CA the throughput that an application can achieve is between 5.5 and 7.5 $Mbit/s$, depending on the packet size and whether TCP or UDP is used. 802.11b can operate at 11 $Mbit/s$, but will scale back to 5.5 $Mbit/s$, 2 $Mbit/s$ and 1 $Mbit/s$ if the signal quality becomes bad. This mechanism is also known as Adaptive Rate Selection. 802.11g defines extensions (e.g., channel bonding and burst transmission techniques) in order to increase speed to 54 $Mbit/s$ staying backwards-compatible with 802.11b.

The 802.11a [91], standardized also in 1999 uses the same core protocol as the original 802.11. It operates in the 5 GHz band and uses a 52-subcarrier Orthogonal Frequency-Division Multiplexing (OFDM), delivering a maximum raw data rate of 54 $Mbit/s$. In practice this results in an achievable throughput of about 25 $Mbit/s$. Compared to the heavily used 2.4 GHz band the 802.11a can benefit from less interference in the 5 GHz band. However, the higher frequency carrier restricts the use of 802.11a to almost line of sight. Indoor penetration is rather poor compared to 802.11b, requiring more access points to achieve the same coverage, assuming the same power.

The WLAN standard proposes a RC4 based security framework to provide *Wired Equivalent Privacy* (WEP). WEP uses the stream cipher RC4 for confidentiality and the CRC-32 checksum for integrity protection. Two key sizes can be used: 40 bit and 104 bit. Both keys are used as shared secrets to derive the actual session key. After having revealed several security flaws in WEP [22] it was superseded by Wi-Fi Protected Access (WPA) in 2003 and then by the 802.11i standard in 2004. 802.11i, which is often referred to as WPA2, incorporates mainly a proper key management to enable per user and per session key and proposed the migration from the RC4 to AES encryption algorithm. The 802.1x in combination with the Extensible Authentication Protocol (EAP) offers a framework to perform user authentication and key management between the 802.11 client and access point.

All versions of the 802.11 WLAN offer two modes of operation. The *infrastructure* and the *ad-hoc mode*.

Infrastructure Mode

This mode of WLAN bridges a wireless network to a wired Ethernet network. It also supports central access points serving WLAN client nodes. To successfully join a WLAN the nodes have to associate with one of the serving access points. The clients have to know the Service Set Identifier (SSID), which can be considered as the name of the wireless network. In infrastructure mode the nodes can not directly communicate with each other. All the communication goes through the access point. This is a must especially with the 802.11e [93], which specifies the QoS functionalities. The central control of all communication is allowing dedicated resource allocation to the clients. To allow communication also without any access point the *ad-hoc mode* was defined.

Ad-hoc Mode

The ad-hoc mode allows direct communication between nodes that are within the radio range by forming an ad-hoc network. It is offering the same data rates as the infrastructure mode depending on the signal quality. In contrast to the infrastructure mode, where the access point interacts with a complete infrastructure offering user authentication, key management, address assignment and billing, the ad-hoc mode treats all interacting nodes equally. Therefore, these nodes have to agree on several settings before they can securely communicate with each other. Whenever two or more nodes want to interconnect using WLAN ad-hoc mode, at least one node has to choose the SSID of the ad-hoc network. This SSID is then broadcasted so that the other nodes can easily scan for that specific SSID and connect to that ad-hoc network. Nodes sharing that SSID can communicate with each other on the MAC layer, not yet starting TCP/IP sessions. Hence, the nodes have to agree on IP addresses. The whole setup procedure becomes further complicated, if the connection has to be secured.

WEP and its successor WPA2 offer enough protection to enable secure networking also in the ad-hoc mode. However, key management has to be handled properly. Especially in ad-hoc mode, where no infrastructure is available to enforce security mechanisms, the proper management of keying material is an issue. In ad-hoc mode, only shared secrets can be used for authentication of the peer and encryption of the transmitted data. If a malicious node within the radio range can learn this shared secret, it can easily intercept the packets and decrypt the content. WLAN is not offering the possibility to further authenticate participating nodes if they know the shared secret. Therefore, the proper exchange of this shared secret is crucial to guarantee secured communication.

If enhanced security mechanisms are required it has to be handled on the higher layers, for instance with IPsec 2.5.4. Nevertheless, the establishment of direct and secured links with WLAN ad-hoc mode is not simple and not practicable for most of the users.

Power Management

Due to the fact that 802.11 based systems are not relying on a slotted medium access mechanism makes the power management rather complicated. If stations are entering a power save mode, they risk missing packets from other stations.

Within the standard, the general idea is to synchronize the stations to wake up at the same time. At this time the sender announces buffered frames for the receiver. The receiver of such an announcement frame stays awake until the buffered framed were delivered. With help of the Network Allocation Vector (NAV) each station announces the time required to transmit the data frames. This allows the other stations to enter the power save mode until the data frames have been delivered to the destination. To do so, all stations have to be synchronized. In infrastructure networks, where there is a central access point, which is able to synchronize all stations. In ad-hoc networks this synchronization has to be done in a distributed manner. After each transmission, all nodes have to wake up to learn if the next frame is dedicated to them. Since the power save mechanism is based on the NAV announcement, it is mainly beneficial for heavily used networks. If no traffic is to be sent for a longer period, the stations have to regularly wake up to check for waiting data and avoid performance degradation. Most of the power save implementations allow different wake up intervals, resulting in different power save and performance levels.

2.2.5 Bluetooth

When using Bluetooth to interconnect mobile nodes, the connection setup process is somehow more user friendly. Bluetooth offers service detection functionality which reduces the user interaction to node scanning and key management. Whenever nodes want to securely connect using Bluetooth, a PIN has to be entered on all nodes. This key is then used for shared secret authentication and to derive a session key for traffic encryption. So Bluetooth basically delegates the key exchange problem to the user, which might severely weaken the security level. Most of the users even disable this security feature to simplify the usage of Bluetooth. However, the integration of the Bluetooth Service Discovery Protocol (SDP) makes the establishment much simpler than it is for WLAN in ad-hoc mode. Connections are defined as services (aka Bluetooth Profiles) and therefore handled by the SDP. The environment can easily be scanned for nodes and their provided services (profiles). Profiles are defined for any kind of communication that can occur between Bluetooth nodes, and new profiles become available to provide new service. To interconnect nodes with the Internet Protocol the Personal Area Network (PAN) Profile was defined, which is supported by almost any Bluetooth enabled device. Further information about Bluetooth and its profiles can be found in [16, 14].

2.2.6 Ultra Wide Band

Ultra Wide Band (UWB)[183] is a very low power communication technology delivering very high data rates for short transmission ranges. Due to its very large radio bandwidth of up to $7.5\,GHz$ (currently present in the 3 to $10\,GHz$ band), the IEEE 802.15.3a working group, concerned with standardizing a UWB physical layer, expects data rates greater than $100\,Mbit/s$ for a $10\,m$ communication range. Since UWB operates in overlapping frequency bands with other technologies like 802.11a WLANs, the basic idea is to limit the UWB transmission power to a level so low that it will not cause significant interference. Currently, there are two main technical approaches to realize UWB radio. The Direct Sequence UWB (DS-UWB) most closely resembles what is traditionally understood when talking about UWB radio (i.e. very short pulses with at least 500 MHz bandwidth). Several such pulses are sent for each bit to be transmitted. The second approach is technically very different and is based on Orthogonal Frequency Division Multiplexing (OFDM) and referred to as Multiband OFDM (MBOFDM). This approach is based on standard OFDM techniques and results in a 500 MHz bandwidth. While there are a few months of difference regarding the technical maturity of currently available DS-UWB and MBOFDM devices, the MBOFDM enjoys much larger industry support, giving it the edge in terms of raw market potential.

2.3 Availability and Usage

All the communication technologies briefly introduced in the prior sections are or become available in the near future for the mass market. Most of them are about to become standard equipment, already built-in at shipping of the devices (e.g., WLAN and Bluetooth in laptops, PDAs, and Smart Phones). On the other hand, operator started to provide access through different technologies some years ago with the introduction of WLAN. However, the adoption of WLAN for public use was not progressing as expected. Most of the WLAN users were to be found in the corporate environment. The deployment of WLAN extensions to the corporate network promised liberty to freely move between meeting rooms without loosing connection. Due to the limited data rates compared to the existing fixed LAN, quite a lot of users came back to the good old cable, whenever possible. Nevertheless, WLAN usage has enormously gained popularity at universities, where thousands of students have to connect to the network without having fixed workplaces. Consequently, students also motivated the deployment of WLAN in cafes located close to the university campus. Hoping that this trend would also pass over to the business environment, operators started to equip public places like train stations, hotels, restaurants and convention centers with WLAN access. The adoption of WLAN for business users happens at a lower pace than it did for students because of more strict security policies and longer life cycles for IT equipment (i.e. laptops). Lot of companies are concerned about data security and therefore impose extensive restrictions on the usage of public Internet access. The success of VPN solutions to provide secured access to corporate data is significantly releasing these constraints. The fact that communication technologies like WLAN become a standard feature of mobile devices guarantees that business users automatically get WLAN enabled when replacing their IT equipment.

Despite the fact that HSCSD and GPRS are available since quite a while, their usage was limited due to restricted data rates. This limited data rates clearly reduces the attractiveness of using bandwidth demanding applications like e-mail or even browsing on HSCSD and GPRS. The situation changed with the introduction of UMTS, offering easily up to eight times the bandwidth of GPRS. The trend towards flatrate makes it affordable also for non-business users.

Bluetooth, finally, succeeded to be ubiquitous. Laptops, PDAs, mobile phones, and even entertainment devices like mp3 players, photo cameras, and gaming consoles start to have Bluetooth integrated. Unfortunately, the provided data rates are quite limited. Bluetooth is not really appropriate to exchange large files. This might change with new versions of Bluetooth radio, offering higher data rates.

Unlike Bluetooth, the data rates delivered by WLAN in ad-hoc mode are reasonable to transfer large files between devices. However, due to its complicated handling, WLAN is rarely used in ad-hoc mode. In this thesis we will propose a framework to increase the usage of such technologies. The high data rates envisioned by the developers of UWB can further leverage such direct communication links.

2.4 Multi-Access

The different access technologies are not yet integrated to form an unified network. Users have to deal with each networking technology to benefit from its services. Hence, to optimally utilize the technologies that are available already today, users have to manually handle the different interfaces and processes required to get connected. Due to missing session mobility, users have to select the right network before starting the actual session. Depending on the complexity of the processes required to set up the connection (including powering up the correct device, authenticate, and potentially establish a secure VPN connection), this manual network selection and management can prevent users to use the service at all. Hence, the efforts to get connected to the different access technologies should be minimized. Furthermore, the end-user should notice as little as possible when changing the access network. Communication sessions (data, voice or video) should not get interrupted. This requirement is already fulfilled in today's cellular networks where an end-user making a voice call on his cellular handset, will not notice a network handover when he moves to another cell. This type of handovers is called horizontal because it basically happens from one access point (e.g., base station) to another, both of the same technology. As soon as an inter-technology handover occurs, it is referred to as vertical handover. Whereas horizontal handovers are often very much supported by the access technology itself, the challenge is to implement the seamless handovers across heterogeneous networks and services (i.e. vertical handovers).

To enable seamless data communication across different technologies, a form of session mobility is required. Without the ability to seamlessly switch between networks, the users are still aware of the heterogeneity of access technologies offered by the operators. This heterogeneity might be perceived bothersome rather than beneficial. Without session mobility the applications have to be restarted after changing communication technology, which might impose an unnatural user behavior. Instead of starting mobile data applications whenever needed, even if the available network is not the most appropriate one, users would wait until the latter becomes available, or even abandon the attempt. Without the ability to handover ongoing sessions from one access technology to another, users might hesitate to start any session if not sure that the session will be terminated before they have to leave the hotspot. Especially, when considering access to corporate networks often requiring time consuming authentication processes (i.e. VPN).

2.5 Mobility and Security

2.5.1 Mobility Models

Different mobility models can be found in the literature, whereas only a few are regularly used for simulations. A survey of the most common mobility models is given in [23]. The mostly referenced mobility model is probably the random waypoint mobility model, which is treating each individual node independently. To better reflect the fact that nodes are often moving in groups like it is the case in trains, buses or cars, we also used the reference point group mobility model, where nodes are traveling with a certain probability in groups rather than individually. Both mobility models are explained in further detail in the following.

Random Way Point

The random way point mobility model is very simple and defined through three parameters only. The interval between υ_{min} and υ_{max} defines the range velocity of the nodes. The nodes are initially placed randomly in the simulation area and select a velocity value within the given interval. With this speed they move towards a random selected destination point. Having arrived, the nodes pause for a randomly chosen *pause-time* between $0\,s$ and $pause_{max}$, which is the third parameter required to define the mobility model. There are some specific characteristics of the random way point mobility that have to be considered when using it for simulations. First, due to the limited simulation area, the random selection of the destination point is forcing the nodes to tentatively move rather close to the center than being uniformly distributed within the whole simulation area. Secondly, the distribution of the actually selected velocity of the nodes is only uniform at the beginning of the simulation. With time, the distribution of the velocity gets more and more inversely proportional to the speed, resulting in lowering the average speed on the nodes. As a consequence of these two characteristics there is warm-up phase required to get a stable mobility model. The authors of [203] show that the expected average speed of the nodes approaches 0, if $\upsilon_{min} = 0\,s$. It is therefore generally recommended to define $\upsilon_{min} > 1\,s$.

Reference Point Group Mobility

The reference point group mobility is keeping nodes moving together with a certain probability. This is achieved by defining reference points. The average size of the groups, its standard deviation, and the maximum distance to the group center can be explicitly specified. To allow the nodes to change from one group to another, a group change probability can be defined. Whenever groups come close enough to each other, nodes can decide to join the other group. Like the random waypoint mobility model, the reference point group mobility model offers the possibility to define υ_{min}, υ_{max} and a *pause-time*. The sum of the motion of the reference point $\overrightarrow{GM}$ and the relative random motion around the reference point $\overrightarrow{RM}$ (i.e. within the defined maximum distance d to the group center). The reference point group mobility model is often used for the simulation of two-level mobility, where the nodes are only moving within a limited space, which is also moving. Scenario examples are military maneuver or public transportation, i.e. trains and buses. Fig. 2.1 illustrates how the nodes calculate their individual movement.

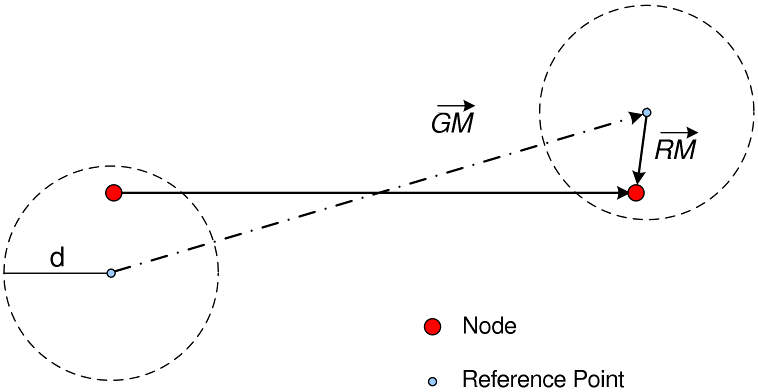


Figure 2.1: Reference Point Group Mobility

2.5.2 Mobile IP (v4/v6)

The problem of session mobility is based in the routing mechanisms that are used in the Internet. The current IP architecture has an implicit assumption that hosts in the network are stationary. But Internet hosts have become mobile with the advent of laptops and PDAs having a wireless Internet connection. The IP stack was not designed with host mobility in mind. Internet addresses are bound to the physical equipment making up the Internet and, thus, are bound to physical locations. When an Internet host (e.g., a laptop) moves to a new location, it has to acquire a new address. This does not have to be an issue since there are automated ways of configuring a new address (e.g., DHCP [50]). However, if the device moves between the networks during ongoing sessions, and the Internet address changes, all TCP and UDP sessions will break down. Mobile IP solves this in an elegant way by tunneling the topologically incorrect packets, allowing the mobile host to keep its address while visiting different network locations. Mobile IPv4 was designed much later than IPv4, which resulted in a sort of unnatural evolution (e.g., deployment of Foreign Agents). Mobile IPv4 was created to deal with the missing mobility support of the legacy version of the Internet Protocol. Differently, Mobile IPv6 represents a basic functionality in the new IPv6. IPv6 was from the beginning designed to cope with mobility, which is clearly reflected in the design of Mobile IPv6. With regards to the work done within this thesis, Mobile IP route optimization only will be discussed in further details. For a comprehensive introduction to Mobile IP, please refer to the literature [99, 150, 136, 102].

Mobility Management

Mobile IP offers seamless mobility for IP connections by offering a constant home IP address for layer four sessions. The mobile node requires connectivity to the home agent to perform the handover. Binding updates containing the new Care-of Address (CoA) and the home address have to be sent to the home agent whenever the CoA becomes topologically incorrect. The handover process is finished when the home agent sends back an acknowledgement. Therefore, the complete handover takes as long as the round trip time between the mobile node and the home agent. Depending on the location and the type of connection of the mobile node, this can be a rather long time making the handover slow. To enable nevertheless seamless handovers across heterogeneous networks, Mobile IP is mostly deployed in a *make before break* manner in combination with the *Simultaneous Bindings* option offered by the Mobile IP standard. By having the new and the old access technology enabled, a so-called soft handover can be performed, given a minimum overlap of coverage of both access technologies (see Fig. 2.2).

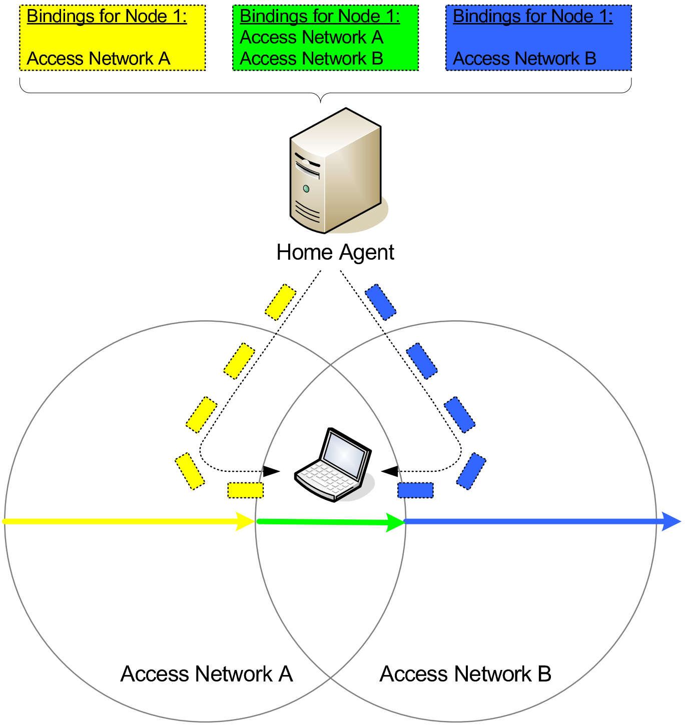


Figure 2.2: Mobile IP Soft-Handover with Simultaneous Bindings

Networking interfaces which are not active could theoretically be switched to a low power mode as long as they are able to detect the availability of any access network. Although the time required switching to full operational mode (including establishment of the IP context) is reducing the amount of time, which is available to perform the handover. Fig. 2.3 illustrates this circumstance in which t_1 marks the first possible detection of the availability of the access technology B and t_2 the latest moment to send the new binding update to the home agent. Hence, to be perceived as seamless, the handover has to be finished before t_2. Between t_2 and t_3 the interface for technology A could be switched back to the low power state. The whole handover procedure starts again when reaching t_3, switching from technology B to C.

Depending on the size of the overlapping section between technology A and

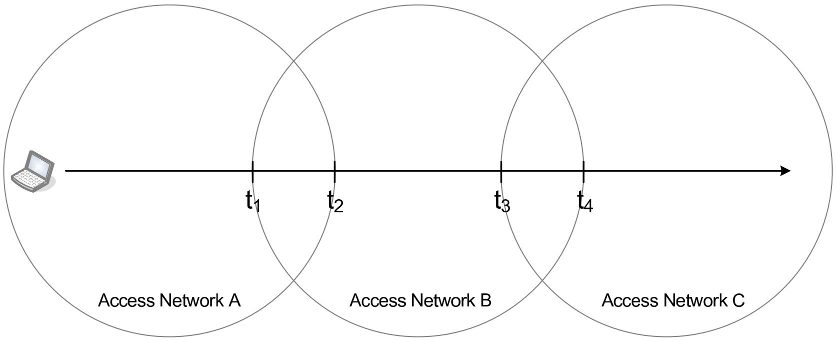


Figure 2.3: Low Power Handover

B, and the velocity of the mobile node, the time period between t_1 and t_2 might become very short. In cases where the round trip time between mobile node and the home agent is high and the process of preparing IP connectivity on the interface B is taking much time, the period between t_1 and t_2 might become too short to perform a seamless handover. There is a lot of research work going on to reduce the time required to perform the mobile IP handover[1]. The IEEE 802.21 [96] working group is preparing a standard to decrease the time required to prepare IP connectivity, unifying the *Service Access Points* (SAPs) and associated primitives to control the underlying communication technologies in a media independent way. However, these unified functions, especially the ones providing information about the availability of the access network, are still dependent on the capability of the underlying devices to permanently scan the environment. Vertical handovers will remain subject to *make before break* and therefore require means to learn quickly about the availability of communication networks.

The signaling of Mobile IP is mainly used to securely handle the binding update messages between the mobile node and the home agent. Signaling messages are transported within UDP packets addressing the port number 434. Mobile IPv4 defines that binding updates have to be sent using the new CoA, whereas Mobile IPv6 offers the possibility of registering alternate CoA. This is especially helpful for the *Route Optimization* feature (see next section). Using this alternate CoA registration it is possible to register a different CoA than is used for the actual signalization. Theoretically, this allows running an out-of-bound signaling for Mobile IP binding updates.

Route Optimization

In contrast to Mobile IPv4, forcing any communication between the correspondent and the mobile node to pass through the home agent, Mobile IPv6 offers a dedicated service to optimize the data flow. In IP version 4, the support of mobility is optional, which requires the mobile nodes to collaborate with the home agent. Correspondent nodes do not know anything about the changing location of the mobile node. All data packet intended for the mobile node are sent to its home address. The home agent has then to forward those packets

[1] A comprehensive comparison of the different proposals to improve the Mobile IP handover can be found in [83]

to the registered CoA. In theory, the packets sent from the mobile node to the correspondent nodes could be directly routed, resulting in a sort of triangle routing of the packets between the mobile node, the correspondent node, and the home agent. However, due to firewall restrictions (e.g., ingress filtering) these packets are normally dropped before reaching the corresponding node. This packet dropping occurs because of the topologically incorrect source address, namely the mobile node's home address. The only practicable solution to this problem is the *reverse tunneling*, wherein the mobile node tunnels the packets with the incorrect source address back to the home agent using the correct CoA. The home agent then decapsulates and forward the original packets to the correspondent node, having now the topologically correct home address. In the extreme case where the mobile and the correspondent node are close to each other but far away from the home agent, this reverse tunneling introduces a unnecessarily long routing path.

To optimize this path, a dedicated feature was included in Mobile IPv6, called *Route Optimization*[2]. In contrast to Mobile IPv4, the mobility support is mandatory for all nodes in version 6. This allows the introduction of special features supporting the route optimization on the correspondent node, to finally decrease the dependency on the home agent situated far away in the home network. With the separation of the communication address (i.e. CoA) and the logical address provided to the higher layers (i.e. home address), the major limitations of IPv4 were broken. The introduction of the *Home Address Destination Option* guaranteed the coexistence with routers that perform ingress filtering. Packet sent from the mobile to the correspondent node can now use the topologically correct CoA and include the home address into the home address option. The receiving node uses the home address when delivering the packet to the layer four. Consequently, the changing CoA is not having any impact on the ongoing layer four sessions. The use of the CoA as the source address, and hence as the actual communication address, also simplifies routing of multicast packets sent by a mobile node. With Mobile IPv4, the mobile node had to tunnel multicast packets to its home agent in order to transparently use its home address as the source of the multicast packets. With Mobile IPv6, the use of the home address option allows the home address to be used, but still be compliant with multicast routing, which is partially based on the packet's source address.

To secure the notification in case of changing CoA, the route optimization defines two way signaling between the mobile and the corresponding node. Correspondent nodes do not generally have a security association with the mobile node. Instead, a method called *Return Routability Procedure* is used to assure that the right mobile node is sending the message. The Return Routability Procedure offers the possibility to prove that the claimed CoA address belongs to the appropriate home address.

First, the mobile node sends the *Care-of Test Init* message directly addressed to the corresponding node. The *Home Test Init* message is routed through the home agent to the correspondent node. In most situations these messages are routed over different network segments. Both messages contain a *cookie*, which consists of a random value used to identify the messages.

[2] Route optimization was initially also designed for Mobile IPv4 (see [151]).

The correspondent node replies with the *Home Test* and the *Care-of Test* messages. Again, the first one is sent through the home agent and the second is addressed directly to the mobile node. The Home Test and the Care-of Test messages both contain keying material (i.e. home keygen token and care-of keygen token), with index values that the correspondent node will use when it receives the binding update from the mobile node.

The keying material from the Home Test and the Care-of Test messages are used by the mobile node to calculate a cryptographic key to secure the binding update message. This key is referred to as *Binding Management Key* or K_{bm}.

The binding update message consists of the home address, a sequence number, the home nonce index, the care-of nonce index and the message authentication code (MAC) to authenticate the prior values. For the structure of the K_{bm}, the two keygen tokens and the MAC is shown right after the Fig. 2.4.

The correspondent node uses the home nonce index and care-of nonce index values sent with the binding update to look-up the keying material it sent to the mobile node. The correspondent node uses the keying material to form a value for the binding management key (K_{bm}). The correspondent node uses the authentication value for the binding update to verify that the mobile node generated the same value for the binding management key (K_{bm}). The correspondent node optionally sends a binding acknowledgement message back to the mobile node.

The verification of the authentication value and binding management key (K_{bm}) proves that the mobile node received data-packets sent through its home agent and sent directly to its proposed care-of address (return routability).

Normally the CoA is used as the source address of the IPv6 header carrying the binding update. However, there are situation, where the CoA is not reachable by the home agent (i.e. due to link local addresses or firewalls). For those cases a different care-of address can be specified by including an alternate care-of address mobility option in the binding update. When such a message is sent to the correspondent node and the return routability procedure is used as the authorization method, the *Care-of Test Init* and *Care-of Test* messages must be performed for the address in the alternate care-of address option (not the source address). Consequently, the nonce indices and MAC value have to be based on information gained from these test messages. The complete Return Routability Procedure is illustrated in Fig. 2.4.

The different security related message fields are calculated as follows:

- **Node key (K_{cn}):** Secret key that every correspondent node has.

- **Nonce:** A secret random value used only once to compute the *keygen tokens*. Nonces are stored in an array and indexed with the *Nonce Index*.

- **Init cookie:** A nonce random value used to identify the Test message.

- **Home keygen token:**
  First(64, HMAC:SHA1(K_{cn},(home address|nonce |0)))

- **Care-of keygen token:**
  First(64, HMAC:SHA1(K_{cn},(care-of address|nonce|0)))

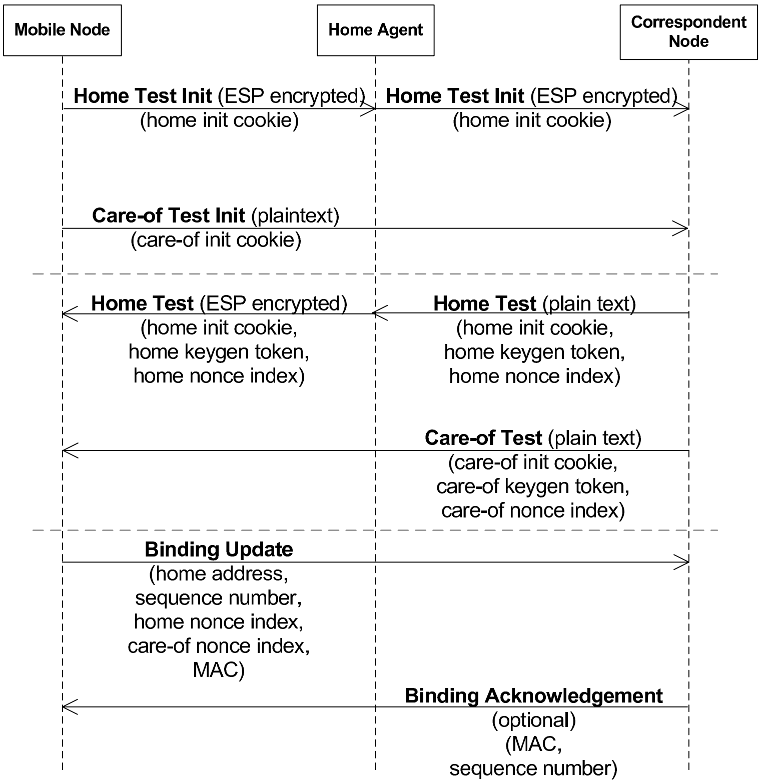

Mobile Node
Home Agent
Correspondent Node
Home Test Init (ESP encrypted) (home init cookie)
Home Test Init (ESP encrypted) (home init cookie)
Care-of Test Init (plaintext) (care-of init cookie)
Home Test (ESP encrypted) (home init cookie, home keygen token, home nonce index)
Home Test (plain text) (home init cookie, home keygen token, home nonce index)
Care-of Test (plain text) (care-of init cookie, care-of keygen token, care-of nonce index)
Binding Update (home address, sequence number, home nonce index, care-of nonce index, MAC)
Binding Acknowledgement (optional) (MAC, sequence number)

Figure 2.4: Return Routability Procedure

- **Binding management key (K_{bm}):**
  SHA1(home keygen token|care-of token)

2.5.3 Session Initialization Protocol

The *Session Initiation Protocol* (SIP) [164] was designed to handle setup, modification, and teardown of multimedia sessions. In combination with the *Session Description Protocol* (SDP) [69], SIP is used to describe the session characteristics to potential session participants. SIP provides basically the functionality for user location service, session establishment, and session participant management.

A major feature of SIP is the support of multi-device leveling and negotiation. If a node initiates video and voice and the invited node is not supporting video, the voice can still be transmitted. Since SIP itself is only specifying how a session should be managed and not anything about the type of session, it can be used for an enormous number of applications like gaming, music, and video on demand as well as voice, video conferencing and many more.

SIP defines four major components, namely the *User Agent* (UA), the *Registrar Server*, the *Proxy Server* and the *Redirect Server*:

- The UAs are located on the end-user devices. The User Agent Client initiates the messages and a User Agent Server responds them.

- The Registrar Server is mainly a database that contains the location of all UAs within a domain. In SIP messaging, these servers retrieve and send participants' IP addresses and other pertinent information to the SIP Proxy Server.

- Proxy Servers accept session requests made by a SIP UA and query the Registrar Server to obtain the recipient UA's addressing information. If the recipient is located within the same domain, the invitation is forwarded directly to it, or to a Proxy Server if the UA resides in another domain.

- The Redirect Servers allow SIP Proxy Servers to direct SIP session invitations to external domains.

To assure reachability, nodes have to register their IP address with their Registrar Servers. Whenever a invitation is received at the Proxy Server, the Registrar Server is asked about the current IP address of the invited node, and delivers the invitation to the mobile node's current location. In all possible scenarios supported by SIP the receiving nodes have to be registered with their Registrar Servers. SIP registrations, requests, and responses are generally sent using UDP, although TCP is also supported. Therefore, SIP requires permanent IP connectivity to allow users to be reachable for incoming requests. In order to support IP mobility, SIP has to offer the ability to change the location (IP address) during a traffic flow. If the mobile node moves during a session, it must send a new INVITE message to the correspondent node using the same call identifier as in the original call setup. To redirect the data traffic flow, it indicates the new address in the SDP field, where it specifies the transport address. This re-invitation has to be done for all active calls or SIP sessions.

SIP together with SDP, focusses very much on application layer relevant information to be negotiated. Being a pure application layer protocol itself, SIP is not designed to interact with the lower networking layers. Hence, SIP is not able to interact with the networking stack, which makes it not applicable for our purpose of making heterogeneous networking seamless. However, the way SIP is addressing users as signaling endpoints was very much motivating our proposed abstraction of heterogeneous networking sessions described in Chapter 4. Therefore, we will briefly explain in Section 6.2.1 how our system can interwork with the SIP, if a SIP infrastructure is available.

2.5.4 IP Security (IPsec)

The initial design of the Internet Protocol was not caring about security issues. It was not designed for public use and did therefore not yet consider how privacy and security could be guaranteed. But the application of IP also for commercial usage, especially for the interconnection of corporate networks, raised the demand for means to protect the transport of sensitive data. Therefore, so-called *Security Gateways* building IPsec connections between private corporate network have been deployed. The resulting interconnection of the private networks through the public Internet is called *Virtual Private Network* (VPN). Later on, the application of IPsec was extended to mobile nodes. IPsec software clients on the mobile nodes allow them to connect to the security gateways from abroad and hence get access to corporate data. This mobile extension of VPNs is also often referred to as *Mobile VPNs* (MVPN), even if rather nomadicity is provided than mobility. IPsec connections have to be re-established whenever the mobile node's IP address changes. IPsec provides a complete protocol suite to handle the different aspects of security. The protocol suite consists of three protocols, namely the *Internet Key Exchange* (IKE), the *Authentication Header* (AH) and the *Encapsulating Security Payload* (ESP). The three components are identical for both versions of IPsec, version 4 and 6. With regards to the focus of this thesis on IPv6 networks offering route optimization, only IPsec version 6 is further treated in this document. For further information about the IPsec version 4 refer to the literature [100].

IPsec offers the following functions:

- Data origin authentication (non-repudiation)
- Data integrity
- Data confidentiality

The IPsec framework was designed to be independent of the algorithms used for authentication and encryption, which simplifies the integration of new cryptographic algorithms.

Internet Key Exchange

To handle the keys used for the encryption and authentication of the sensitive data sent, IKE was designed. IKE provides the functionality to establish manually or automatically a *Security Association* (SA) between nodes. These SAs define the set of parameters required to successfully exchange secured data. Communicating nodes have to agree on keys and selection of algorithms used to encrypt and decrypt the sensitive data packets. When a SA is created manually, each communicating node has to be configured with the appropriate keying material. Manual configurations are applicable only for small and static VPNs, but not for MVPNs including a large number of mobile nodes. To some extend, the manual configuration process can be used to setup direct links between nodes, but requires deep knowledge on the specific parameters of the SA. Hence, for scalable and convenient setup of IPsec links, the automatic configuration was introduced. The automated key management protocol is defined in three IETF RFCs:

- RFC 2407: Internet IP Security *Domain of Interpretation* (DOI) for ISAKMP
- RFC 2408: Internet Security Association and Key Management Protocol (ISAKMP)
- RFC 2409: Internet Key Exchange (IKE)

Whereas the ISAKMP defines the common framework and packet formats, IKE the key exchange procedures, and DOI how IKE and ISAKMP are used to negotiate the SA for IPsec.

IKE operates in two phases. In the first phase, it performs mutual authentication and establishes an IKE security association that can then be used to efficiently establish the IPsec SAs in phase two. The IKE parties authenticate each other by using a shared secret, public keys or digital signatures. The two latter methods require trust on both parties' public keys, which requires regularly access to a Public Key Infrastructure (PKI) (see Section 4.4.2). During phase two, the negotiating parties agree on the encryption and the authentication algorithms, on the keying material, and on the lifetime of the new SA. IKE has a number of deficiencies from which the three major are the high number of round trips, its vulnerability to denial of service attacks, and the complexity of its specification. Especially this complexity has led to interoperability problems between different implementations. At the time of this writing a new version of IKE is about to be standardized (IKEv2) [107]. IKEv2 is based on IKE, but the protocol is lighter and simpler. The number of messages and available options is decreased to ease the implementation and the interoperability. In IKEv2, the attention is focused on the protocol's robustness and security against denial of service attacks.

The most evident advantage of IKE (v1 and v2) is that it automatically negotiates IPsec security associations (SAs) and enables IPsec secure communications without costly manual pre-configuration. The following list summarizes the main features provided by IKE:

- Eliminates the need to manually specify all the IPsec security parameters in the SA at both peers

- Allows to specify a lifetime for the IPsec security association

- Enables dynamically changing encryption keys during IPsec sessions

- Offers support for certificate based authentication (PKI)

- Allows dynamic authentication of peers

The framework provided by IKE can also be used to handle end-to-end security for heterogeneous data sessions. For further information about IKE, refer to [70].

Authentication Header

IP Authentication Header (AH) provides data origin authentication and data integrity for all end-to-end data transported in IP datagrams. The AH offers two modes of operation: The *Transport Mode* is used for end-to-end security. Only the communication endpoints can perform the authentication of the datagrams. The *Tunnel Mode*, on the other hand, is used to secure a part of the end-to-end path only. This mode is especially useful if not all communicating nodes support the required security features (i.e. a specific algorithm), but also when several IPsec sessions can be aggregated between intermediate nodes (e.g., Security Gateways). For detailed information about the tunnel mode of the AH refer to [112].

Encapsulating Security Protocol

The third protocol in the IPsec suite is the ESP. It offers the ability to guarantee privacy and data integrity for IP packets. Similar to the AH, the ESP defines two modes of operation, namely the transport and the tunnel mode. Again, tunnel mode is only used in combination with security gateways (i.e. access firewalls) and therefore not further considered for our heterogeneous session management. For further details on the tunnel mode refer to [113].

ESP in transport mode is typically used to protect a layer four segment such as a TCP or UDP segment, containing application level data. The ESP header is inserted immediately prior to the transport layer header (e.g., TCP, UDP, or ICMP). In the case of IPv6, if a destination option header is present, the ESP header is inserted immediately prior to that header. Transport-mode operation may be summarized as follows:

1. At the source, the data block consisting of a trailing portion of the ESP header plus the entire transport-layer segment is encrypted, and the plaintext of this block is replaced with its ciphertext to form the IP packet for transmission.

2. This packet is then routed to the destination. Each intermediate router needs to examine and process the IP header plus any plaintext IP extension headers, but does not need to examine the ciphertext.

3. The destination node examines and processes the IP header plus any plaintext IP extension headers. Then, on the basis of the SPI in the ESP header, the destination node decrypts the remainder of the packet to recover the plaintext transport-layer segment.

Transport-mode provides protection for any type of datagram independent of the application. Thus, it avoids the need to implement privacy in every individual application. The added overhead is minimal due to encryption of the payload only.

2.6 Conclusion

In this chapter we briefly introduced the most important components that form the heterogeneous networking environment that has to be faced by mobile data users. Therefore, we quickly discussed the different communication technologies in terms of availability and how they are used today in a multi-access fashion. The analysis of the Mobile IP, IPsec, and SIP revealed the strengths and weaknesses each of these protocols has with regards to the realization of seamless heterogeneous data communication. In the next chapter we will address the related work done in the research community to further approach the unification of the various communication technologies.

Chapter 3

Related Work on Heterogeneous Networking

3.1 Introduction

The design of our envisioned framework is motivated by several areas of research in mobile computing. When thinking of integration of heterogeneous networks, there are some key issues that have to be addressed. Session mobility is supposed to glue the different networking technologies together by enabling seamless session handovers from one network to the other. We will study some major proposals from the research community in further details, especially with regards to their capability to cope also with infrastructure-less communication links. Beside the ability to dynamically switch ongoing sessions between networks, further basic functionality is required, including resource management and location management. Although this thesis is not focusing on those aspects, it is important to understand how our developed system has to interact with the proposed solutions. Our framework should optimally fit into the big picture of future heterogeneous networking. In the second section, we also discuss some related work focusing on the evolution path of wireless data networks towards 4G and Ambient Networks. In the third section of this chapter we concentrate on the research efforts done to enable simple and flexible end-to-end communication in heterogeneous networks. The most popular concepts are presented and analyzed to extract further requirements for the design of our framework presented in the next chapters.

3.2 Mobility Management

3.2.1 Introduction

To enable host and hence session mobility in nowadays heterogeneous networking environments, the following four issues have to be considered carefully[1]:

- **Addressing:** Due to the hierarchical definition of the Internet Protocol routing and addressing, the validity of IP addresses are limited to certain

[1] This classification is highly motivated by the work of Henderson *et al.* [77]

domains. When nodes move from one network to another, their used addresses become topologically incorrect.

- **Location Management:** When nodes change their IP address, they become unreachable for other nodes until they communicate the new IP address.

- **Session Management:** Since transport protocol sessions are defined based on the IP address (and layer four port) they break if the IP address is changed. Furthermore, extended periods of disconnection can cause higher-layer applications to abort, even if the transport layer session is successfully handed over to the new IP address.

- **Security:** Security associations have to be established between communicating peers and maintained upon moving.

To overcome the limitation of topologically bound IP addresses, several proposals are available in the research community. The most straight forward approach is incorporated in the design of IPv6. The tremendous extension of the IP address space is theoretically allowing the allocation of fixed addresses to every mobile node. This basically obsoletes the conflict between host identifiers and communication addresses (see Section 3.4.6). Together with the mobility offered by Mobile IP, IPv6 addresses can be used as unique and static host identifiers even if the nodes are mobile. Several proposals are focusing on the acceleration of Mobile IP based handovers to eventually offer seamless mobility purely on the networking layer (see Section 3.2.3).

In the case of IPv4, having private addresses requiring Network Address Translation[2] (NAT) [176] widely deployed, the IP addresses can not be used as unique host identifiers. Alternatively, the second available name space, namely the Fully Qualified Domain Names (FQDN) [135, 155, 54], which, together with its address resolution mechanisms, DNS could be used to provide host mobility. However, the existing mechanisms to resolve addresses are not appropriate for dynamic use. They have been designed for rather static address assignment. A third possibility is to create a new name space and corresponding address resolution architecture.

Independent of the name space used to identify the hosts, the peering node has to be informed about the change of IP address. This can happen either directly, indirectly (depositing the new IP address into the network infrastructure so that it can be accessed as needed), or not at all. The third case requires an intermediate node, which is making the IP address change transparent (e.g., Mobile IP home agent).

The research work found in the literature and discussed in the following sections was analyzed having these four major issues in mind. We categorized the different solutions according to the ISO/OSI layer they are addressing. Proposals focusing on layer four are extending the TCP session management to handle changing underlying IP addresses. IP mobility concepts are allowing the provisioning of static IP addresses to the upper layers by encapsulating the topologically incorrect home IP address.

[2] Information about the advantages of NAT and the resulting dilemma between NAT and IPv6 is discussed in [84].

3.2.2 Transport Layer Mobility

Since the major hurdles of IP mobility are based on the fact, that layer four connections are uniquely identified by a 4-tuple (*source address, source port, destination address, destination port*), some researchers argue that mobility management has mainly to be handled on that layer. In contrast to IP, TCP is end-to-end session oriented. If mobility is handled on layer four, both end-points of the sessions have to collaborate to perform the handovers. This might also be an advantage, since session end-points know most about the ongoing session and can therefore act more intelligently, compared to end-to-end session agnostic layer three mobility solutions. The related work presented hereafter is focusing on that end-to-end aspect, which is highly relevant to the work presented in this thesis. Thereby, we discuss two proposals in detail. Both, the *TCP Migrate* and the work presented by *Seamless and proactive end-to-end mobility solution*, were selected as representatives for layer four session mobility solutions because they propose a straight forward system architecture. Other proposals focusing on layer four and end-to-end session mobility can be found in [61, 128, 154, 117, 114].

TCP Migrate

In [174, 175], the authors proposed extensions to TCP to enable the migration of sessions from one IP address to another. To locate mobile hosts as they change their network attachment point, they take advantage of the widely deployed Domain Name System [135] and its ability to support secure dynamic updates [53, 187]. They argue that this dynamic resolution of hostnames to IP address at the beginning of each connection is already happening in standard Internet applications and therefore no additional overhead is introduced. When nodes change their network attachment point (i.e. IP address), they send a secure DNS update to one of the name servers in their home domain updating their location. Two communicating peers must securely negotiate a change in the underlying network layer IP address and then seamlessly continue communication. Since network layer moves may be quite sudden and unpredictable, the nodes require learning the new IP address before a move occurs. The authors presented a new end-to-end TCP option to support the secure migration of established TCP connections across an IP address change. Using this option, a TCP peer can suspend an open connection and reactivate it from another IP address. In this protocol, security is achieved through the use of a secret key negotiated through an Elliptic Curve Diffie-Hellman (ECDH) [7] key exchange during initial connection establishment. It requires no third party to authenticate migration requests, thereby allowing the end-points to use whatever authentication mechanisms they choose to establish a trust relationship.

The authors rely on the arguments presented in [167], which observed that functionality is often best implemented in a higher layer (i.e. end-to-end) at an end system, where it can be done according to the application's specific requirements. The handling of the mobility on an end-to-end basis is enabling higher layers like TCP and HTTP to learn about mobility and adopt to it. As an example, the authors proposed to restart TCP transmission from slow start or a window-halving [103], or adapt the transmitted content to reflect new network conditions, after a network route change, since the bottleneck might have changed. The authors claim that these optimizations could be made naturally if mobility is handled end-to-end since no extra signaling is needed. Research in the domain of mobility-aware applications [105] should be able to benefit from their proposed architecture.

The TCP migration options are included in the SYN segments by identifying a SYN packet as a part of a previously established connection, rather than a request for a new session. This migrate option contains a token that identifies a previously established connection on the same destination (*address, port*) pair. The token is negotiated during the initial connection establishment.

To secure the TCP migration the authors proposed the use of IPsec or to encrypt the connection token with a secret connection key (e.g., with *crypto-based identifiers* [137]). The authors compare their solution with Mobile IP route optimization (see Section 2.5.2) in terms of security. They argue that in contrast to Mobile IP, their solution requires only a trust relationship between the end-points (i.e. the mobile nodes) and not any further trust relationship between to end-points and the home agent.

The proposed solution demonstrates very well the advantages of handling mobility issues on an end-to-end basis. Focussing on the layer four is very straight forward, since it is the first layer with regard to the ISO/OSI stack, which is providing end-to-end characteristics. Furthermore, it is still low enough to be generic and hence independent of the applications. The biggest limitation of this approach is that both peers can not move simultaneously. Because the proposed scheme does not have an anchor point like Mobile IP's home agent, any IP address change must be completed before the peer node can proceed with the next IP address change. In [194], the authors address this problem of simultaneous mobility of IP host in further details, and proposed stationary proxies that can be queried if both nodes loose each other due to simultaneous movement.

Seamless and proactive end-to-end mobility solution

In [65], the authors presented a very similar approach than proposed with TCP Migrate [174] focusing on the extension required to offer layer four mobility. Unlike the architecture proposed for TCP Migrate, the system presented in [65] implicitly addresses the problem of simultaneous mobility and NAT by deploying a dedicated subscription service. The proposed system integrates a Connection Manager (CM) that intelligently detects the condition of wireless networks and a Virtual Connectivity-based mobility management scheme that maintains connection's continuity. By using MAC-layer sensing in addition to physical-layer sensing, network parameters such as available bandwidth and access delay when roaming from wireless wide area networks (WWAN) to WLANs can be obtained to enable a pro-active reaction to roaming events. Virtual Connectivity (VC) consists of a Local Connection Translation (LCT) to make mobility transparent to upper layer applications and Subscription/Notification (S/N) service to successfully handle mobility under NAT and simultaneous movement. The system architecture is shown in Fig. 3.1 and illustrates very well the introduction of the intermediate connection translation module (LCT) between the IP and the TCP layer.

The VC is responsible for two major operations, namely the *peer negotiation* and *connection maintenance*, which is performed per connection. During

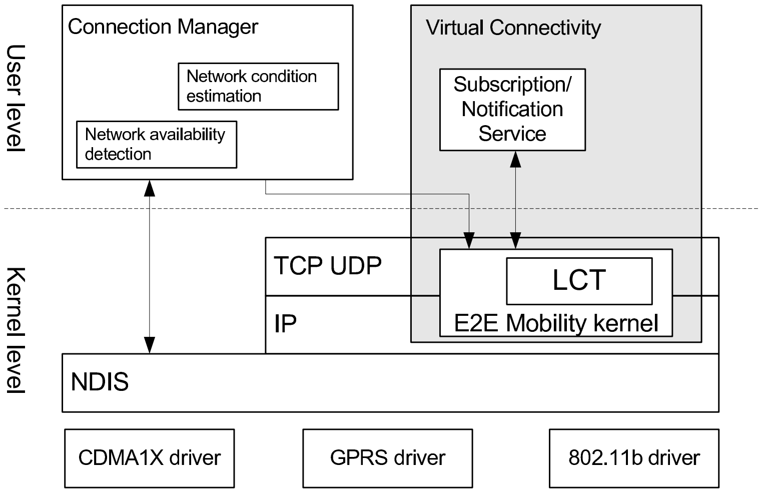


Figure 3.1: End-to-End Layer 4 Session Management Proposed by Guo *et al.*

the peer negotiation, communicating nodes have to agree on items that will be needed for secure and accurate mobility management before mobility events happen.

The following information has to be exchanged during the peer negotiation:

- Shared Secret and Connection Identification: The shared secret is used to protect the connection maintenance. The authors proposed Diffie-Hellman key agreement to derive the shared secret between nodes. To uniquely identify each connection a Connection Id (CID) is created.

- Original IP and Port Number: Knowing the original IP address and comparing it with the source IP address of the packet, the receiving node can check whether its peer is behind a NAT.

- Capability and Preferences: When both hosts have explicit information about their peers, they can make more efficient decisions. For example, if a host is publicly addressed and does not move, no S/N will be needed. The nodes can also exchange certificates to build up a trust relation for further session maintenance.

To maintain ongoing connections if the IP address changes, the VC provides following primitives:

- Connection Update (CU): Is used to update the peer if the IP address changed.

- Connection Update Acknowledge (CUA): Is an acknowledgement message for a received CU.

- Connection Update Challenge (CUC): Can be used to check whether the mobile node is reachable at its claimed current IP address.

- Connection Update Challenge Response (CCR): In combination with the CU and CUC, it can be used to perform a three-way handshake to check the return routability.

Within [65], these messages are supposed to be securely exchanged prior to the first movement. In contrast to Mobile IPv6 route optimization, where mobile nodes have a security relation to their home agents, which can then be used to build up a trust chain between the nodes, here the nodes can not rely on any pre defined security association. In [65] it is not further explained in detail how this required initial trust relationship can be established between the nodes. The authors are working towards the extension of the CM to include wireless personal area networks (WPAN) and other communication technologies like WCDMA and 802.11a/g. Furthermore, they are investigating how P2P networks could be used to decentralize their S/N service.

The basic concept of providing layer four end-to-end mobility minimizing the dependency on infrastructure to perform seamless handover is the same way than presented in [174]. Nevertheless the extensions proposed in this work to handle simultaneous movements and cope with NAT are reflecting the dependency on fixed infrastructure shown by [194]. Furthermore, the planned work to integrate also WPANs is interesting and further motivates the comparison with layer three mobility solutions addressing mobility management for mobile routers.

3.2.3 Network Layer Mobility

Most of the existing solutions and proposed architectures for heterogeneous all-IP networks [87, 210, 52] are based on layer three mobility (i.e. Mobile IP and IPsec). As a representative, we briefly present a promising combination of Mobile IPv4 and IPsec v4 called *Secured Mobile IP* (SecMIP). More details can be found in [38, 37] and [36, 45, 44, 43].

Secured Mobile IP (SecMIP)

Since Mobile IPv4 and IPsec v4 were designed rather independent of each other and both much later than the native IPv4, the integration of them is not as natural as it is in their version 6. In contrast to Mobile IPv6 the version 4 is not offering means to ensure strong security. Some weak authentication for the binding updates is provided to avoid replay attacks but encryption of the tunneled data is not specified. Due to the IP tunneling used to send the data packets from the home agent to the mobile node and vice versa, the usage of IPsec in transport mode to protect the sensitive data when traveling through the Internet seems to be the most evident method of combining Mobile IP and IPsec. However, there are reasons to protect the Mobile IP signaling (e.g., Binding Updates) as well. The binding update messages reveal both, the home address and the CoA of the mobile node, which allows tracing. Especially in public networks offering mobility, this is not acceptable due to privacy issues. Furthermore, Mobile IPv4 is vulnerable to denial of service attacks. Deploying robust IPsec gateways to authenticate mobile nodes before accepting any Mobile IP binding updates, can solve this problem. Further security issues of Mobile

IPv4 can be found in [38]. A simple and straight forward way to protect both the signaling and data transmitted by Mobile IPv4 is the complete encapsulation in IPsec tunnel mode. Fig. 3.2 is illustrating the packet structure when using IPsec tunnel mode to protect the Mobile IPv4 packets.

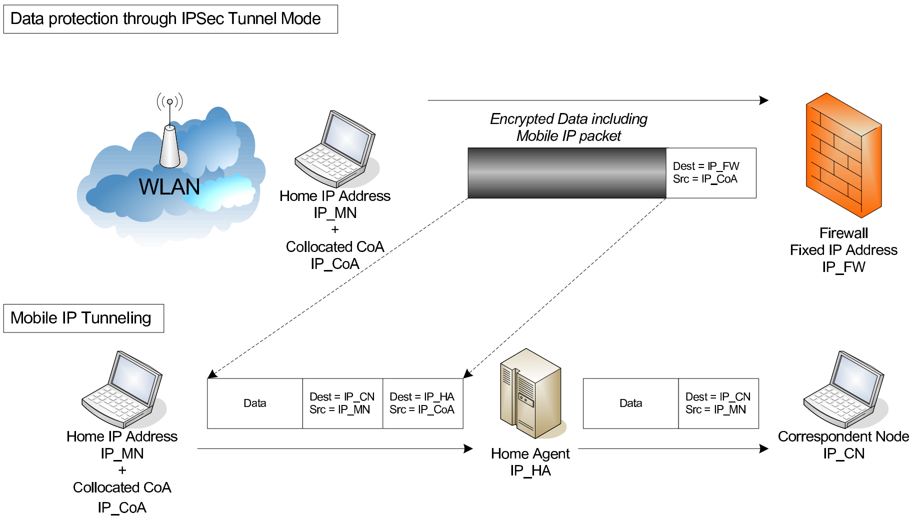


Figure 3.2: Packet Structure of Secured Mobile IP (SecMIP)

Whenever a mobile node detects a new access network, it initiates an IPsec tunnel to the VPN gateway protecting the home agent. Only after successful authentication against the VPN gateway binding updates can be sent to the home agent.

In IPv6, both Mobile IP and IPsec have been designed to interwork. Therefore, there is much less tunneling required to achieve secured mobility over heterogeneous networks. Mobile IPv6 uses reverse tunneling, encapsulating packets between the mobile node and the home agent and vice versa. Where the tunnels need to be protected, they are replaced by IPsec tunnels. The home agent then acts as a security gateway terminating the IPsec tunnel. The SA is set up between the mobile node's CoA and the home agent's IP. Like in IPv4 this leads to the re-establishment of the SA whenever the CoA changes.

Basically, the evolution towards IPv6 is quite straight forward. Due to the better integration of Mobile IP and IPsec, the deployment becomes easier and the route optimization offers means to overcome the drawbacks imposed by Mobile IPv4, namely the tunneling of all data back to the home agent, independent of the location of the mobile and the correspondent node.

Mobile IP defines, that whenever an handover occurs, the mobile node has to send a binding update to the home agent. If the mobile node is far away from the home agent, this can lead to significant delays. Especially, if a lot of handovers happen due to small ”cells” or high mobility, this can considerably reduce the performance of Mobile IP. This motivated researchers to declare Mobile IP as mobility protocol for *macro-mobility* and develop new solutions for *micro-mobility*. The introduction of mobility domains and localized mobility solutions offering mobility within these domains (micro-mobility), limited Mobile IP to handle Mobility between mobility domains (macro-mobility). The most known micro-mobility solutions are Telecommunication-Enhanced Mobile IP Architecture (TeleMIP) [47, 30], Cellular IP [8, 24, 130], HAWAII [161], and Edge Mobility Architecture (EMA) [144]. All proposed solutions adapt the access networks to support micro-mobility within a specific micro-domain without requiring the mobile node to send binding update to the home agent. In [163, 25], a overview of the different micro-mobility protocols and architectures can be found. A comprehensive comparison of the ongoing work in the domain of host mobility in IP networks can be found in [76, 166]. Other proposals to improve the handover performance of Mobile IP can be found in [204, 208, 172], where as the authors of [172] address also paging issues (see Section 3.3.4).

3.2.4 Conclusion

In this section, we analyzed existing proposals for secure session mobility. Layer four solutions can be deployed without changes to the applications. In contrast to layer three solutions, they can benefit from the end-to-end session management. However, both, network and transport layer based mobility solutions require means (i.e. infrastructure) to establish a security relation between the communicating peers to eventually secure the session migration. All analyzed solutions propose the utilization of IPsec to provide data security and protect the session management. Simultaneous movements of the communicating nodes are a generic issue, especially for end-to-end oriented mobility mechanisms. To locate lost nodes in case of simultaneous movements or enable mobility also with NAT, fixed infrastructure is required. To provide seamless heterogeneous networking both, layer four and three solutions could be used. However, the standardized and widely accepted Mobile IP (with its route optimization) is the most advanced mobility management solution for IP communications. The adoption of Mobile IP as the defacto standard mobility scheme for future all-IP telecommunication networks (3GPP [2]) is further pushing layer three solutions. The required infrastructure to enable strong security and simultaneous movements is further favoring Mobile IP with its home agent. Nevertheless, there is a major advantage of layer four based mobility when considering inter-device handovers. The ability to migrate ongoing sessions from one IP address to another might be used to transfer sessions from one node to another as well. An appropriate security framework would be required to securely move the session context information to the new node. To the best of our knowledge no work on that has been published so far.

3.3 Heterogeneous Networks

3.3.1 Introduction

In contrast to horizontal handovers occurring between base stations (or access point) of the same technology, vertical handovers cross network boundaries. With the help of session mobility solutions, these handovers can be performed seamless. However, the adoption of such session mobility architectures to unify heterogeneous networks has just started, which is also retarding the operational integration of the different access networks. Network operators treat each access network autonomously without taking any advantage of having an integrated network. The network selection is still done manually by the users based on their preferences. With the introduction of seamless mobility solutions the boundaries between the different access networks disappear, enabling the network operators to handle network resource management commonly for all their networks. Research work related to the aspects of handover decision taking and heterogeneous resource management is discussed later in this section.

A further issue of heterogeneous networking is location management and reachability. The ability to combine location and environmental information provided by the different communication technologies has the potential to further increase the value of heterogeneity.

Some perspicacious researchers started already to describe future networks beyond what is referred to as 4G. Terms like *Ambient* or *Ubiquitous Network* are reflecting the envisioned characteristics that such future networks should incorporate. As described in the related publications in Section 3.3.5, the integration of wireless personal networks including all kind of communicating devices that users will carry with them, will be a major component of such ambient network.

3.3.2 Connection Managers and Dashboards

Mobile IP and IPsec are very promising protocols to offer seamless and secured communication over heterogeneous IP networks. Together with the efforts going on to automate and therefore simplify the authentication required to get access to the different communication networks, there is a big trend towards user convenience. Despite the fact that technologies like WLAN already exist for almost a decade, the adoption is happening slower than many operators expected. The introduction of HSCSD and GPRS clearly showed that user convenience is essential for getting a data access network accepted and widely used. If the usage is too complicated, people tend to not use it, even if they see the potential of the service. Convenience is often determining whether a service becomes a success or a failure. This is especially true for mobile communication services, where the users have be able to easily access the service anytime and anywhere without struggling with configuration issues. The introduction of GPRS capable mobile phones and its usage as modems to connect laptops and also PDAs to the Internet raised the quest for easy connection management. The major hurdles preventing seamless access to a heterogeneous network environment are related to seamless authentication and interface configuration. The standardization bodies are investing a lot of efforts to simplify the connection setup process of the different networks. The migration from Web based [10] to integrated authentication like EAP-SIM [72, 73] is a big step in the right direction. With EAP-SIM the simplicity of SIM based authentication has been extended to Public WLAN hotspots making their use much more user friendly. Dedicated applications have been developed to handle the connection establishment to the different access networks. These so-called *Connection Managers* have been extended with further functionality to ease the access to operator services. SMS client, shortcuts to browser and VPN client are only examples of extensions that made connection managers evolve to *Dashboards*. These Dashboards offer a unified graphical user interface for data services. Solutions like the Swisscom Mobile's "Unlimited" [64] product are bundling various access technologies in a transparent way for the end user. The mobile device (laptop) gets connected to the best available network (in this case the choice is between GPRS, UMTS or WLAN) in terms of signal quality and maximum bandwidth. The use of Mobile IP allows a seamless handover between the different access technologies. Thanks to the Dashboard user interaction is minimized. In [66], Gustafsson *et al.* proposed the "always best connected (ABC)" concept that offers the most appropriate connectivity over multiple-access technologies depending on the actual user needs to enhance the networking performance. This work provided the framework of ABC; but mobility management was not discussed.

3.3.3 Handover Decision and Resource Management

With the adoption of WLAN as a supplementary access network for cellular network operator and the ability to offer seamless session mobility beyond the network borders, a further dimension for network resource management has been opened. Resource allocation schemes can now be extended to heterogeneous resource management taking all available networks into account. Dashboards combined with Mobile IP have the ability to logically merge different access networks so that they are perceived as one network. The introduction of EAP-SIM simplified the integration the different networks in terms of authentication and billing, but from a network resource perspective the networks are still treated separately. Due to the *make before break* characteristic of mobile IP based handovers between the access networks, the different networks have to overlap (see Section 2.5.2). Furthermore, data services like GPRS or EDGE built on top of the cellular voice access networks have often a very high degree of coverage compared to pure data networks like WLAN and WiMAX [95] Hotspots. Hence, these high coverage access networks are often used as backup connections even if other technologies are available. Nodes allocate resources in both networks, which results primarily in waste of resources, since the actual user data is only sent through one network keeping the other network idle but occupied. For scarce and precious networking resources like the UMTS, this greedy resource allocation is very unfavorable. In nowadays implementations the handover decision is taken on the mobile node (i.e. Dashboard software) or even manually by the user.

Metrics used for Network Selection

A variety of metrics have been employed in mobile data networks to decide on handovers. Primarily, the received signal strength measurements from the serving point of attachment and neighboring points of attachment are used in most of these networks. In [168], a roaming scheme that considered only the signal strength was proposed, resulting in sub-optimal handover decisions. More detailed performance description of [168] can be found in [57]. The authors of [202] consider the relative bandwidth of WLAN and GPRS to decide about handovers, but no technical details on how to obtain the actual bandwidth were provided.

In [21] a system is presented, which includes handover decision policy profiles to allow users to influence the handover decisions in an intuitive and simple way. They proposed an approach in taking vertical handover decisions, which are not anymore exclusively based on the knowledge of the available access networks' characteristics but also on higher level parameters which fall in the transport and application layers. The aim of the work presented is to balance from the end-user point of view, the overall *cost* of vertical handovers (e.g., delays, change of bandwidth and power consumption) with the actual benefits they bring to his actual networking needs. To this extent, in this paper a model has been realized and simulations have been run in order to evaluate the impact of the vertical handover and its frequency on a set of typical user's network applications/services. The results show that, dependent on the handover performance (i.e. dropped packets, handover delay, etc.), and the requirements of the application, the networking experience is better if less handovers are performed. Especially in the case, where applications needs are satisfied with the current access network, any handover is considered as bothersome, even though the new network would provide better QoS characteristics. Hence, handover decision might be very dependent on the actual user's networking needs.

Alternatively or in conjunction, the path loss, carrier-to-interference ratio (CIR), signal-to-interference ratio (SIR), bit error rate (BER), block error rate (BLER), symbol error rate (SER), power budgets, and cell ranking can be employed as metric in certain mobile voice and data networks. In order to avoid the ping-pong effect, additional parameters are employed by the algorithms such as hysteresis margin, dwell timers, and averaging windows. To handle all these different indicators to finally take an adequate handover decision, the authors of [147] introduced neural-network-based algorithms for handover decision in heterogeneous networks. The results are very promising when thinking of the increased complexity when extending the number of metrics to application or user specific requirements. The complexity of providing always best connected (ABC) capabilities was also analyzed in [63], where the authors show that the ABC problem belongs to the class of NP-hard combinatorial optimization problems.

The authors of [186] defined a metric to estimate the user satisfaction for different handover algorithms. The presented network model is limited to UMTS and WLAN, whereby the UMTS is considered to offer full coverage and the WLAN is covering hotspots. The authors assume that bandwidth is most important for the user, and that a handover from UMTS to WLAN is more satisfactory than being handed over from WLAN to UMTS. Furthermore, the satisfaction is decreased with increasing number of occurred handovers. The evaluation are done using two basic mobility models, one for corporate users, having little to no mobility and the other users, having full mobility. The idea of measuring handover algorithms in terms of user satisfaction is promising but the definition of realistic satisfaction function seems to be quite challenging. The authors stated that further parameters like application requirements, network operator interests, battery power level, user location and many more have to be considered to realistically reflect the user satisfaction.

Resource Negotiation

All kinds of solutions providing session mobility across different networks require the prior establishment of the new link to allow a seamless transition. From a network perspective, the different simultaneous connections are not perceived to belong to the same session, even if it would be possible to do the mapping based on the common authentication credentials (e.g., SIM). It is therefore not possible to consistently handle networking resources provided by all deployed networks. The operator has hence to over-provision his network to be able to assign resources from different networks to the same session. To optimize the resource allocation in heterogeneous networking environments, there is a missing link between the mobile nodes (i.e. the Dashboard handling all available connections and deciding about the potential handovers) and the network resource management system. The concept of MIRAI [196, 87, 89] addresses this problem by defining dynamically one channel to be used for signaling information and negotiation of handover decisions. The authors proposed an agent based platform that provides location-based information on available access networks through a so-called basic access signaling, which is assumed to have a larger coverage than all other access networks. This concept is very beneficial, especially if the basic access signaling channel is a low power channel[3]. In [55] it has been shown that such a signaling channel does not have to provide high data rates. Depending on the implementation, it might even be possible to store most of the required information on geographical availability of the different access networks locally on the mobile node. This would allow reducing the required data to be exchanged between the network and the mobile node. Obviously, the accuracy of that information on geographical availability of the different networks is reflected in the efficiency of such out-of-band signaling. If the used information is wrong, it will result in either an unsuccessful scan or a missed network. The proposed system consists mainly of a resource management component, which is aware of all available networking resources. With the help of a further component on the mobile node, the network resource management component can get all the needed information about the applications that are used and their requirements on the underlying network. Very similar, the middleware presented in [20] and [13] is able to gather context information to optimize the handover decision. Both approaches use software agents to process the collected context information and finally get the optimized resource allocation for each node. However, all these proposals concentrate on infrastructure-based networks only. When considering infrastructure-less communication technologies as well, these centralized management of handover decisions might become limited.

On one side, the optimal selection of the access network might depend on the requirements of the actual applications and the capabilities of the used device. But there is also the possibility to enhance the application performance by making it mobility aware. In [18] a complete architecture is presented to provide adaptive content mediation based on an agent collecting characteristics of the actual underlying network and deducing the optimal content presentation with the help a proxy (i.e. content mediation server). Similarly, an intelligent service mediation system is proposed in [148] to adapt content in a location and mobility aware manner.

[3] The concept proposed by MIRAI is further analyzed in 3.3.4.

More information about the generic approach to smartly integrate existing communication technologies to form what is called *Beyond 3G* (B3G) or lately also more and more 4G, can be found in [12, 11]. A tutorial on the design and performance issues for vertical handoff in an envisioned multi-network fourth-generation environment is presented in [132].

3.3.4 Location Management and Reachability

In mobile networks, nodes move from one base station (or access point) to another. In heterogeneous mobile networks, the nodes can even change the access technology when moving around. Existing wireless wide area networks (aka WWAN) like GSM or UMTS deploy sophisticated location management systems to enable fast localization of mobile nodes to assure the fast delivery of incoming communication sessions. There are two basic ways for location management solutions to handle this issue. First, the mobile node sends so-called location updates (LU) reporting its actual position (e.g., by associating to a specific base station or access point). Second, the network sends a solicitation to trigger the node to send a LU. This second process is referred to as paging. It was mainly introduced to save power and radio resources if no communication sessions are going on with that specific mobile node. Different approaches exist to reduce the number of LU and regions to send the paging requests. Especially, when mobile nodes are in idle or sleep mode, the sending of LU can significantly increase the power consumption, depending on the size of the paging area. In [195], an overview of the different strategies sending LU for homogeneous networks is presented.

IP Paging

When considering heterogeneous networks, location management becomes even more complex. Depending on the different paging capabilities of the underlying networks, there might be the need to provide paging mechanisms also on the IP layer. A basic problem statement of IP paging in homogeneous and heterogeneous networking environments is given in [109, 110]. There is a lot of work published on the location management in Mobile IP based networks, especially addressing the different aspects of IP paging. The authors of [32] analyze in further details the signaling costs of LU and paging in Mobile IP based networks. To minimize the signaling cost for location management, they proposed an architecture called 'Combinatorial Mobile IP' introducing hierarchical paging functionality, not only separating micro-mobility from macro-mobility, but also active mode from idle mode. A similar approach is presented in P-MIP [206, 205]. The authors proposed extensions to Mobile IP to indicate paging capabilities both, in the foreign agent advertisements and in the registration request message. Additionally, the agent advertisements contain a paging ID, which can be used by the node to determine if it changed paging area and hence has to update its location by sending a registration message. Paging is done by broadcasting a paging request containing the home address of the searched node. In order to simulate whether IP paging can increase the signaling performance compared to Mobile IP, the authors of [111] implemented a dedicated simulator. The presented results show that IP paging is only decreasing the signaling costs if the number of cells per subnetwork is small, limiting the geographical coverage of the subnetwork and hence increasing the number of Mobile IP binding updates.

Lot of research work tackles the question about the optimal size and shape of the paging area to minimize the signaling overhead. In [159, 158, 160], three different approaches to offer IP paging are analyzed in terms of delay and signaling cost. The differentiation is done depending on the location, where the paging service is deployed. Home agent paging is supposed to be initiated by the home agent, whereas foreign agent paging is managed by the foreign agent. In domain paging, paging state is distributed among the routers and base stations in a domain rather than at one fixed node such as the foreign or home agent. The authors conclude that the domain paging has to highest potential to minimize the signaling costs for paging. In [27, 26] the authors proposed the definition of adaptive per-host paging areas to further optimize the ratio between LU and paging messages.

In [200], it is proposed to combine session setup with paging. The author proposed to use RSVP Path messages instead of dedicated paging messages. Although this approach is beneficial for high RSVP session arrival rate, it is limited to communication sessions using RSVP. In [140], Bloom Filters [15] are used to encode multiple node ID into a single paging message. This approach is promising in IPv6 networks, where the home address is used to identify mobile nodes (each IP address requiring 128 *bits*).

Paging in Heterogeneous Networks

All discussed paging solutions are primarily focusing IP paging and do not address the potential benefit of having heterogeneous environments. To the best of our knowledge, there is only one approach taking advantage of the different existing network to enable efficient paging for IP mobility. The MIRAI project [87, 89] addresses the design and implementation of an architecture to efficiently use existing wireless networks according to the capabilities and requirements of the mobile nodes. One of the key concepts proposed by the authors is the definition of an out-of-band signaling channel. This concept would also allow using non-IP networks for the paging of the mobile nodes. In [85], the same authors presented a mobility management scheme to reduce power consumption in IP-based wireless networks. In addition to the separation of active and idle mode, introduced by [32], the scheme described in [85] is distinguishing between different states depending on the level of activity of the mobile node. The proposed solution manages communicating, attentive, idle, and detached states for efficient management of battery power, radio resources, and network load. Whenever a node terminates a communication session, it switches stepwise back to power save mode. Therefore, the node's status can first transit from communicating to attentive/idle, where the network still knows exactly to which access point the node is attached. After a certain timeout, the node switches to attentive/paging-area-connected state, reporting only the paging area to the paging agent. To further save energy, the node can detached all radio connections except a dedicated broadcast channel and a paging channel. Detached mode is finally entered when the node is switched off. It neither sends location registration messages nor responds to paging.

MIRAI

In [127, 126, 88], more information is provided about the Basic Access Channel (BAC), which is used in MIRAI to transmit signaling information (e.g., LU and paging). MIRAI uses a dedicated Basic Access Network (BAN) (providing the BAC) to perform all signaling functions required to handle heterogeneous access networks. Fig. 3.3 shows the general network architecture proposed by MIRAI.

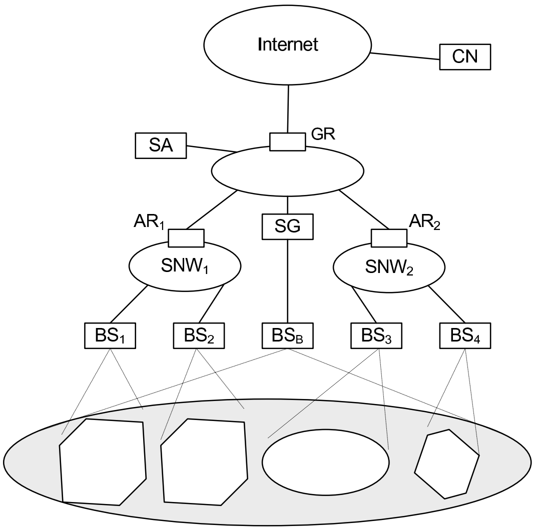


Figure 3.3: MIRAI Architecture: SG Signaling Gateway, SNW Sub Network, AR Access Router, GR Gateway Router, SA Signaling Agent

The BAN is assumed to be a high coverage and highly reliable network, since all signaling information is exchanged through this network. When a node comes into the area of the BAN, it can initiate the registration procedure. The node sends a LU including its position, capabilities, and preferences through the BAC to the Signaling Gateway (SG). Within the system discovery, the network informs the node about available RANs at its position[4]. After successful access to one of the available RAN, the Mobile IP binding update is performed in the home agent (GR of the home network). The handover from one RAN to another is supported by the BAN. Preparation of the new RAN resources and packet duplication over the BAN can considerably increase the handover performance of Mobile IP.

When a correspondent node (CN) from Internet sends a packet to the mobile node, it is intercepted by the home agent in the home GR. Depending if there exists a binding entry for that mobile node, the home agent can directly tunnel the packet to the actual CoA, or must first initiate paging through the BAN. The SG initiates the actual paging signal towards the BS (of the BAN) under its control. After receiving the paging signal, the terminal sends a paging response, which includes the LU. The authors proposed to send back the list of available RAN with the correspondent network prefixes to enable the mobile node to compose the CoA. After validation of the CoA by the correspondent RAN, the

[4] Depending on the resource management policy, a dedicated RAN can be proposed.

node registers that CoA with the home agent and receives the packet from the CN.

The authors initially proposed the deployment of the new dedicated signaling network (BAN), but then generalized their signaling to work on any IP access network. Since MIRAI focuses on IP access networks, any network that is selected as signaling channel can also be used for data transmission. This allows best support of heterogeneous nodes. Whatever communication interface requires the least power can be used as signaling channel and trigger further radio interfaces whenever more bandwidth is required. However, this signaling channel is basically used to avoid power consuming scanning for the various access networks that are available at a specific location. The nodes have only to connect to the network used as BAN and get the information about other available networks. Depending on the requirements of the applications, the capabilities of the node and the disposable resources in the access networks, further data channels can be negotiated.

To secure the signaling information transmitted through the BAN, the authors proposed the usage of a symmetric key between the mobile node and the network. This symmetric key is also used to derive further keying material for the actual data communication through the different RANs. Therefore, the authors refer to the Internet Key Exchange (IKE) protocol discussed in 2.5.4.

MIRAI is addressing infrastructure-based network technologies only. The concept of having a dedicated signaling network offering wide area coverage to negotiate the resource allocation also for the access networks used for the actual data transmission could also be extended to infrastructure-less communication technologies.

3.3.5 Towards Ambient Networks

The integration of different wireless IP networks into the core of existing mobile networks to make the Internet mobile determined the direction to go for future heterogeneous networking. The standardization work in development at 3GPP is only considering the integration of WLAN APs with the cellular network; they do not consider WLANs operating in ad-hoc mode as part of the architecture, and consequently do not address issues related to multi-hop routing. However, lot of researchers address the combination of ad-hoc links to increases the coverage and capacity of cellular networks. In [29] an overview of the issues in integrating cellular networks, WLANs, and Mobile Ad-hoc Networks (MANET) [101] is given. The authors proposed a stepwise evolution towards full heterogeneous networks by starting with the integration of WLANs (e.g., IEEE 802.11a/b/g), wireless WANs (e.g., GPRS, UMTS), WPANs (e.g., Bluetooth, IEEE 802.15.1/3/4), and wireless MANs (e.g., IEEE 802.16) by observing a common characteristic of one-hop (single-hop or infrastructure) operation mode, wherein users access the system through a fixed base station or AP connected to a wired infrastructure. The extension to multi-hop operation, interconnecting nodes without the use of base station or AP, providing alternative connections inside hotspot cells, is considered to come in a second step.

Single-hop extensions of the cellular network is already becoming commercially available by integrating WLAN APs and multi-mode terminals supporting

WLAN in infrastructure mode. There are several proposed architectures in the literature for the second step of the evolution, namely the integration of multi-hop connections to enhance and extend the coverage and capacity of cellular networks. Recently, many researchers addressed the capacity and connectivity of hybrid networks [34, 49, 62, 81, 80, 121]. All of them conclude that there is a benefit in integrating infrastructure-based networks with multi-hop functionality. However, they also state that the overall performance increase in terms of throughput and robustness strongly depends on the type of traffic that has to be routed through the network, the topology, and the density of nodes (mobile nodes and base stations/APs). In [80, 82], the authors analyzed the benefit of node-to-node (i.e. ad-hoc) connections compared to centralized cellular links, stating that the throughput and power consumption can considerably be enhanced, especially if source and destination nodes are within the same cell (which is the case for WPAN and communities, see Section 3.3.5). Most of the work concentrates on the network layer, since it integrates all technologies. We discuss the most relevant architectures and protocols integrating multi-hop MANETs with infrastructure networks.

The IST-Ambient Network project [86, 142] is addressing the strategic objective of mobile and wireless systems beyond 3G networks. The project defines Ambient Networks as a dynamic composition of networks, providing access to any network, including mobile personal networks, through instant establishment of inter-network agreements. The project is defined over three phases and is about to enter its phase two (2006-2007). The work done in the first phase mainly addressed the automatic management of Service Level Agreements (SLA), when dynamically interconnecting ambient networks. Concepts focusing on the connection management are expected for the second and third phase of the project.

Multi-hop Access to Base Stations/Access Points

The Unified Cellular and Ad-hoc Network architecture (UCAN) [124] basically aims at the usage of multi-hop routing to increase the throughput of downlink channels when the signal between the mobile node and the base station becomes weak. UCAN considers dual-mode terminals equipped with cellular and WLAN interface, which is used for ad-hoc mode operation only. UCAN does not consider the deployment of WLAN APs. Mobile nodes can interact as proxies and relays for nodes facing poor cellular coverage. The overall network capacity can be increased due to the fact that CDMA-based systems deliver more throughput if the distance between base station and mobile node is small. The use of WLAN ad-hoc to relay the downlink traffic from the proxy located close to the base station to the mobile node that is far away reduces the average distance to serve all nodes. The authors do not explain how the election of the proxies and relays happens and how it can be motivated.

The authors of Two-Hop-Relay architecture [191] proposed the use of multi-hop links to increase the throughput of the downlink, similarly to UCAN. To reduce the system complexity and avoid inefficient routing of ad-hoc networks, the hop limit is set to two. The presented architecture considers not only cellular base station, but also WLAN APs. The relay gateways (RGs) can be nodes placed by the operator or dual-mode terminals able to act as RGs. RGs broadcast their relay service periodically. Mobile nodes can then dynamically decide whether they want to communicate directly with the base station/AP or using the relay service. All authentication and accounting functionality is controlled by the cellular system, which might considerably increase the manageability of the proposed solution.

The Hybrid Wireless Network (HWN) Architecture presented in [79] defines two operation modes for each cell (base station). The single-hop mode is used for sparse topology and the MANET mode is only deployed if the node density is high enough. The mobile nodes periodically send location updates based on their GPS coordinates to the base station, which runs an algorithm to decide on the operation mode to maximize the throughput. A major drawback of the proposal is the centralized selection of the operation mode for all connections, which may not be optimal. A better option may be to choose the operation mode on a per connection basis.

iCAR [153, 197] addressed the efficient load balancing between neighboring cells. Therefore, the authors proposed Ad-hoc Relay Stations called ARS to be deployed by the network operator. The ARS are equipped with two interfaces, one to communicate with the cellular base station and another to communicate in MANET mode with other ARS. Hence, the MANET operation is only used to interconnect the ARS. The ARS are used to divert the traffic from an overloaded cell to a neighbor cell being less utilized. Similar to iCAR, the Mobile-Assisted Data Forwarding (MADF) [199, 56] aimed at load balancing enabled by multi-hop connections as well. The main difference is that MADF proposes to use the mobile nodes instead of dedicated stations (ARS). The same applies to ACENET (Ad-hoc Cellular Networks) presented in [201].

The ad-hoc Global System for Mobile communications (A-GSM) architecture [5] allows dual-mode mobile nodes to relay packets in MANET mode. The aim of the system is not only to increase the overall performance, but also provide connectivity in dead spot areas. The proposed MANET mode is a GSM interface, which has been adapted to handle beacons used to announce the relay service. The relaying itself is based on encapsulating the GSM layers between the mobile node and base station. When considering pure data traffic, this encapsulation might result additional overhead, especially when using all-IP wireless technologies like WLAN. A detailed comparison of the different solutions providing multi-hop access to base stations can be found here [29].

Beside the motivation to provide higher data rates and appropriate connections for different applications, the authors of [207] describe the benefits in terms of robustness, when deploying heterogeneous networks. The presented thoughts on the combination of different technologies is somehow different from the main research streams, addressing network convergence aiming at eliminating duplicated functionalities and components. The authors declare heterogeneous networking as a new survivability paradigm, driving the systematically increasing of heterogeneity without sacrificing its interoperability.

Ad-Hoc and direct Node-to-Node Connections

The evolution towards heterogeneous networking, enabling to be Always Best Connected, is building the basis for ubiquity. Multi-mode nodes equipped with intelligent software choosing transparently the best suited available network depending on the actual needs of the user (i.e. its current applications), come already quite close to what is also referred to as Ambient Networking. In [193], the authors describe the evolution towards an open all-IP network architecture including support for Personal Area Networks (PANs) and ad-hoc networks. The introduction of wireless PANs clearly brings in a new dimension of ad-hoc networking, interconnecting nodes within and between PANs. With the introduction of cheap, but powerful short range communication technologies like Bluetooth and WLAN, the interconnection of nodes belonging to the same logical entity enable the formation of moving networks. These ad-hoc, spontaneous, and often also moving networks, deployed in the unlicensed bands will form "PAN bubbles", as the authors call them. This bubble is expected to be highly dynamic depending on the actual needs of its owner. Virtual Private Networking technologies like IPsec (see Section 2.5.4) enable even the interconnection of distributed bubbles. Hence, users driving in their car can stay connected to their devices and services deployed in the home or office networks. The authors introduce the term Virtual Terminal Equipment (VTE) to better reflect the communication capability of that bubble. The traditional (ad-hoc) PAN network concept is extended from a composition of devices that form ad-hoc connections between themselves to a composition of devices that have cellular and wireless access capabilities. The authors proposed a middleware abstracting the capabilities of the VTE to enable applications and services to seamlessly interwork with the actually underlying PAN devices.

MObile grouPEd Devices (MOPED) [115] is very much aiming at the same extension of simple terminals to complete PANs offering the flexibility required to interact with the variety of applications and their different requirements in terms of capabilities. The authors address the challenges to handle ever increasing heterogeneity in available networking technologies and devices that users are carrying with them (e.g., mobile phone, PDAs). The proposed cooperation approach presented in [115] should bring the potential for increasing bandwidth and better connectivity to users moving through different environments by exposing to all devices the aggregation of services available to each individual device. The goal of the authors is to fill the gap from communication to cooperation among nodes belonging to the same PAN. To integrate the MOPEDs into the Internet, they proposed to create a representative presence on the Internet. All communication to a user is then directed through this presence. If the presence has a unique network name, a single IP address, the user is in essence built into the network infrastructure. All communication destined for that presence is addressed to a unique identifier.

The internal routing within the MOPED is relying on NAT mechanisms, hiding the internal structure of the MOPED while making all devices reachable through one single (public) IP address. The management of the internal MOPED network topology is done by the Multipath Layer, which is responsible to maintain a partial graph of the MOPED and use its tracking information along with the possible application input to choose external interfaces by which packets enter or leave the MOPED. The MOPED is considered as a single private routing domain. With the help of Mobile IP, a tunnel is maintained to the proxy, which is acting as a counterpart to the Multipath Layer. To circumvent the MOPED proxy, the functionality has to be implemented in the correspondent node (i.e. the destination MOPED). The authors also allow native IP communication from any component of the MOPED that has an external networking interface, bypassing the structure of the MOPED.

In Ambient Network [142], three major innovations extending all-IP towards ambient networks are identified, namely network composition (beyond simple internetworking), enhanced mobility, and effective support for heterogeneity in networks. The authors consider the dynamic, instant composition of networks as a basic mechanism for ambient networking. Similar to the proposed PAN support presented in [115], the spontaneous formation of (personal) networks is identified as crucial to face the challenges of future multi-mode and multi-node communication. In dynamically composed network architectures, mobility of integrated and localized communications (e.g., in PANs) has to be supported as well. A vehicular network of devices requires a new kind of mobility compared to mobility offered in nowadays networks. Especially, when considering device-device interactions across compositions. Ambient networks will be based on a federation of multiple networks of different operators and technologies. Ambient networks take a new approach to embrace heterogeneity visible on different networking levels. The authors further define three design principles for ambient networks. The first is to remove architectural restrictions on whom or what can connect to what. Services should be offered to end-networks rather than to end-nodes. This is motivated by the observation that current node-centric designs fail for many scenarios (e.g., PANs, moving networks, or sensor networks). The second design principle addresses the self-composition and self-management. Simple networking offering full flexibility in terms of device and service selection is considered as very important for future ambient networking. And finally, the ambient network functions have to be added to existing networks. The authors of [6] describe a control plane to provide end-to-end communication in heterogeneous internetworking environments. The introduced control plane focusses on the abstraction of the underlying networks and the dynamic mapping between flows and bearers. The abstraction enables the decoupling of networking layer and the application and service layer, which is considered as essentially to provide the characteristics of ambient networking.

Spontaneous Heterogeneous Networking

The envisioned flexibility to handle devices and network communication technologies in a heterogeneous environment also includes the ability to easily interconnect nodes with short range technologies to form what is called spontaneous networks.

In [120], the authors define a spontaneous network as a small-scale ad-hoc network intended to support a collaborative application. They explore some of the unique challenges that need to be faced in building such environments. The fact that hosts are not pre-configured raises several problems like host name and address management, handling available services and the nodes where they are hosted, and finally the provision of security related functions. Especially security mechanisms usually rely on availability of a trusted key management infrastructure (see Section 2.5.4). A further challenge is to cope with the actuality that users are not experts. Operations must be intuitive to non-technical users. Users are notoriously bad at configuration, especially in an environment that poses complex security issues.

Some researchers combine the efforts done to enable spontaneous networking with the vision of ambient networking including wireless PANs. In [118, 119], a user centric architecture for spontaneous networking is presented. Similar to ambient networks, the authors define communication spaces which consist of the services provided by the computing facilities for a specific user. These computing facilities are not limited to computers and mobile terminals, but include vehicles, buildings, consumer electronic devices, and sensors. The aim of the presented architecture is to offer a system, where the user can spontaneously interact with any such facilities, network, services, and content in the communication space. However, users should also be able to communicate with any other communication space, which might belong to another user, institution or a group of users. The authors do not limit the type of interconnection to direct node-to-node links, but envision also the integration of ad-hoc networks. The authors define a Service Gateway (SG) for each communication space, offering access to the internal services provided by the different facilities of the communication space. These SG also connect the communication space to other ad-hoc networks or fixed access infrastructures. The proposed SG is based on the Open Service Gateway Initiative (OSGi) [146]. If remote users want to connect to any service offered by a node within a communication space, it has to connect to the SG, which is handling all services registered by the different nodes. This initial communication between the remote node and the SG is based on SIP and requires therefore the SG to have IP connectivity to its SIP registrar server to update its registration, whenever it moves. The SG then relays the remote request to the specific node within the communication space. To enable mobility for ongoing communications between the remote node and the communication space, the system uses Mobile IP. Therefore, the SG registers its home IP address with the home agent.

The authors of [108] tackle the issue of spontaneous networks formed by the different devices of a user by introducing a communication gateway. In contrast to the work presented in [119], the communication gateways are mainly used to offer access to network services through different access networks. Hence, the nodes within the user environments are considered as pure service clients, not offering any service by themselves. Their solution is highly motivated by the automotive industry, where communication gateways built into cars can offer access to heterogeneous networks to the client devices being either installed fixed in the car or belonging to the person sitting in the car.

3.3.6 Conclusion

The analysis of ongoing research work in the domain of heterogeneous resource management and handover strategies showed that context information has to be taken into account to optimally assign network resources to mobile and multimode nodes. Existing proposals require the nodes to inform the network about available networks, ongoing application requirements, and user preferences to intelligently handle resource management. Any system that aims at providing an abstraction layer to heterogeneous and complex underlying networks has to offer means to deliver this required information to the resource management components.

We believe that the concept of separating signaling and data plane proposed by MIRAI is very promising, when handling heterogeneous networks including location management and paging. We even go one step further, taking advantages of having paging mechanisms already in place, when considering the cellular system to be part of the heterogeneous network. This offers two main advantages, namely the ability to power down energy demanding IP connectivity interfaces whenever there are no active IP sessions and reuse of existing highly sophisticated, globally deployed location management mechanisms. We further studied related work proposing the integration of personal area and mobile networks. This led us to spontaneous networking, addressing simple and convenient communication between mobile nodes and entities of the personal networking environments, using direct node-to-node and ad-hoc connections.

3.4 Mobility Management for Heterogeneous Data Sessions

3.4.1 Introduction

The evolution towards heterogeneous networks raised a lot of research interest in the domain of communication session management. In Section 3.2, we discussed related work addressing the problem of session mobility across different IP networks. Some of the proposed solutions, especially those acting on the transport layer, are also focusing on the end-to-end aspect. In contrast to layer three mobility solutions that are purely concentrating on the correct IP packet delivery to and from mobile nodes, the end-to-end awareness of higher layer solutions allow a more intuitive session management. IP packets belonging to the same TCP session are treated accordingly, enabling a more session- and hence also application- and eventually even person-oriented mobility management. However, the implementation of session mobility on the application layer implicates some considerable disadvantages. The required adaptation of each application to correctly handle IP address changes is probably the most impending issue. The following subsection is further treating the limitations of application layer session management (e.g., SIP) in heterogeneous networks. The rest of this section addresses some concepts proposing the deployment of mobility functionality somewhere between the transport and the application layer.

3.4.2 SIP for Heterogeneous Networking Sessions

When talking about session management, SIP is the most referenced protocol. Although there are several proposals how SIP could be used to offer to a certain extend mobility even in heterogeneous networks, they are all focusing on application mobility only[190, 171, 141]. This is mainly due to the fact that SIP is situated on the application layer and therefore lacks any capabilities to interact with networking layers. The authors of [51] analyzed in further detail the issues of SIP multimedia sessions in heterogeneous access environments. With regards to SIP signaling setup including several intermediate steps (Invite, Provisional responses, Ack, OK), it may happen that the mobile node changes its access network even before the session setup is complete. After a certain timeout, SIP

UA tries to re-transmit the Invite message. The invitation can only be correctly forwarded after the destination has successfully updated its registrar server. To do so, the SIP UA is dependent on the underlying layers to re-establish IP connectivity, since it can not handle this by its own. SIP UA are not foreseen to handle multiple IP interfaces. For fast moving nodes, changing IP addresses frequently and having temporarily multiple interfaces available, the registration procedure is critical. Registered IP addresses have to be reliably handled, especially the de-registering process for obsolete addresses has to be done in a consistent way.

Comparing these issues with the problems that Mobile IP has to face, makes it obvious that a combination of both protocols could deliver enhanced session management. So, one may argue that Mobile IP is inappropriate for real-time applications due to the rather high delays introduced by the binding updates that have to pass via the home agent, even if using route optimization. A further drawback of Mobile IP is the enormous overhead introduced by packet encapsulation used to redirect small IP packets to the actual CoA of the mobile node (e.g., VoIP)[5]. However, Mobile IP is providing transparent mobility which is needed to keep TCP connections alive as the user is moving. SIP, on the other side, is preferably used for real-time applications relying on small UDP packets. More details on the different proposed architectures combining Mobile IP with SIP to achieve best performance for both, real-time and long-lived TCP sessions can be found here [190, 189, 58, 152, 106, 116].

3.4.3 Personal Mobility

The authors of [165, 9, 129] tackle the problem of personal mobility by introducing a *person layer* on top of the application layer, emphasizing the idea that the person, rather than the device, is the communication endpoint. Each person is identified by a globally unique *personal online ID*. Each person has a *personal proxy* performing person-level routing, including location tracking, accepting communications on person's behalf, converting the communications into different application formats according to the preferences, and forwarding the resulting communication to the person. Although these proxies are independent of the telecommunications infrastructure, one drawback is that regardless of where the person is, all traffic is routed to the proxy first. This can yield to inefficient routes if the person is located far away from his personal proxy.

A similar approach is presented in [188]. The authors were also motivated by the idea of personal mobility, but focussed on the design of a centralized architecture to provide converged voice and data services.

The costs of this personal/session mobility scheme are modifications to applications (unlike Mobile IP) and an indirection infrastructure. Furthermore relying on centralized proxies is in contradiction to any node-to-node and route optimization efforts done to increase the efficiency of heterogeneous communications.

[5] Mobile IPv4, applying IPinIP encapsulation, is adding an IP header of 20 *bytes* to every data packet forwarded to the mobile node. For VoIP, using very small UDP packets, the overhead can make up to about 30%.

3.4.4 Host Identity Protocol (HIP)

During the last years, the flexibility required to provide session mobility is severely inhibited by the overloading of the IP address with additional semantics. IP addresses, initially designed to serve as communication addresses only, are often also used as host identifiers. A recent development proposes a clear separation of the host identification and the IP addresses by introducing an additional layer between the network layer and the transport layer. This Host Identity Protocol defines the Host Identity (HI) name space for creating cryptographic end-to-end connection identifiers to complement standard routing identifiers (i.e. IP addresses). This HI is a public key, making authentication of protocol transactions very easy and enabling the protocol robust against man-in-the-middle attacks. HIs can be stored in data bases like a PKI or used anonymously. If used anonymously, HIP can only protect ongoing sessions from hijacking.

Since the long public key is not practical in all actions, a 128 *bit* long Host Identity Tag that is generated from the HI is introduced[6]. A major strength of HIP is its seamless integration with IPsec transport mode. To decouple the IPsec SA from the IP address, the SA is bound to the HIT instead. Fig. 3.4 illustrates the extended protocol stack when using HIP.

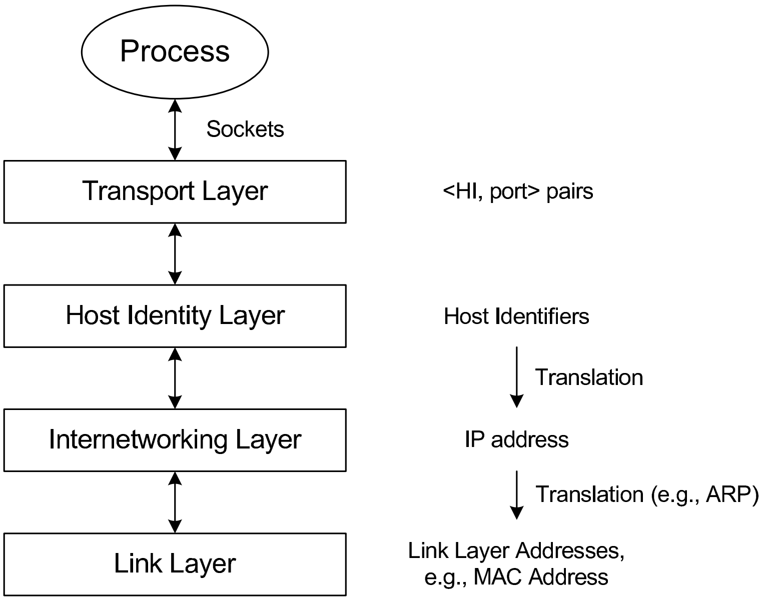


Figure 3.4: HIP Extended Protocol Stack

When packets are leaving the host, the correct route is chosen and the HI is replaced with the corresponding IP address of the networking interface that is used to actually send the IP packet. HIP does not specify how the path is chosen, which leaves full flexibility to routing in heterogeneous environments.

If a HIP node is having multiple IP addresses configured on different interfaces connected to different networks, it can notify the peer node of some or all of these addresses. Therefore, the HIP mobility and multi-homing proto-

[6] Having the same length as an IPv6 address, it can directly be used for IPv6 applications.

col defines a readdress (REA) parameter that contains the current IP address. This allows a full flexibility of handling data paths between HIP enabled nodes similar to Mobile IPv6 nodes.

The problem of simultaneous movement is handled similar to [65] discussed in Section 3.2.2 by using a fixed infrastructure like the DNS with dedicated HIP records. However, there may be performance drawbacks of using DNS in this manner, particularly relating to the heavy reliance on caching for scaling. To solve this issue, the author of HIP proposed a special *rendezvous* server that has a non-changing DNS record but which is used to relay the first HIP packet to the current IP address; this concept resembles a Mobile IP home agent.

In [77], the authors compared HIP with Mobile IP route optimization and conclude that HIP provides potential performance, security, and addressing realm interworking benefits over Mobile IP in some scenarios, particularly in heterogeneous mobile networks with nodes having multiple simultaneously used interfaces. The strength of Mobile IP is clearly the ability to handle entire subnets (routers), where HIP is not deployed, or where infrastructure to support pervasive IPsec has not been deployed. Another interesting combination of both protocols is to use HIP to secure the Mobile IP binding update.

A possible combination of HIP and TCP Migrate [174] is presented in [76]. The combination is compared to Mobile IP route optimization concluding that the operational differences, on an end-to-end basis, do not appear to be significant. The authors proposed a complementary deployment of both approaches. HIP may, for example, offer a cleaner solution to binding update authentication than currently offered by return routability. However, the authors also state that more development of HIP is necessary to better assess its potential to complement Mobile IP in this way. For more information on HIP, the HIP architecture, and how HIP can offer mobility, consult [139, 138, 75].

An other proposal that introduces a similar idea than HIP is in [33]. The authors presented FARA which stands for *Forwarding directive, Association, and Rendezvous Architecture* and describes an abstract architectural model. The main objectives of FARA are to decouple the host identifier and location information without introducing a new global name space. The authors of FARA encompass an interesting part of the design space, while leaving many details unconstrained. M-FARA is finally defined to instantiate the FARA model, binding free parameters in FARA and dining mechanisms. FARA could make use of the HIP when the node identifications are verified. Consequently, HIP could be part of a particular FARA instantiation.

A separation of the identity and the routing information is also proposed by the Internet Indirection Infrastructure (I^3) [178]. The authors proposed to identify the data instead of the hosts using an identifier. Receivers register their IP address and the identifiers of particular data with the rendezvous server, which is then forwarding data identified with that identifiers to all subscribed receivers. The proposed solution concentrates on multicast environments and therefore not applicable for generic end-to-end mobility management. The authors of ROAM [209] reused this idea and defined extensions to allow mobile nodes to control the placement of indirection points (rendezvous servers) in the infrastructure. The introduction of so-called triggers that can be set on the indirection points to relay specific data to any kind of end point (i.e. users or hosts) is enabling the I^3 architecture to serve also for unicast traffic. Simulations done by the authors, show that their approach can reduce the overhead and transmission delay, if compared to Mobile IP.

3.4.5 Session Maintenance Protocol (SMP)

The problem of having names coupled with IP addresses is also addressed in [131] by introducing a *Session Maintenance Protocol* (SMP) with a new local *Session Identifier* (SID) supporting transparent changes to IP addresses. Furthermore, the authors proposed replacing DNS, which simply returns the IP address associated with a name, with a *Logic Name Server* that routes messages to the *Location Server* (LS) associated with a name. Similar to the proposals presented in Section 3.2.2 to locate mobile nodes with existing DNS, using uncacheable DNS entries, the delegation of the location management to the LS situated close to the mobile node allows higher degrees of mobility. Only the LS have to be informed if changes occur to IP addresses, whereas the LNS has only to be informed if the mobile node changes LS. Similar to HIP, the SMP approach introduces an additional layer to separate host names from communication addresses. This additional layer with its unique SID might prevent the adoption of the solution.

3.4.6 End-to-End Addressing

Although there is not much doubt that the utilization of such an intermediate layer is the way to go to clearly separate names from mobility management, it is questionable, whether the introduction of a new and additional infrastructure to resolve the addresses is feasible. Therefore, several research proposals have suggested the use of DNS to enable mobility and/or routing across multiple addressing realms. In [156] and [31] the authors proposed NAT-based architectures that support routing by name, and are somehow similar to HIP sine they introduce an additional protocol layer between the layer three and four as well. The architectural implications of using a domain name rather than a cryptographic host identity are discussed in detail in [156], as well as the possibility combination with HIP to provide stronger security for that approach.

In [182, 181] the authors proposed the concept of 4+4 addressing, which represents two IPv4 addresses that are concatenated. The public address is used for routing in the Internet, whereas the private is used within NATed domains. NATed nodes use the public address of the NAT server as their public address. Nodes having public addresses set their private address to 0.0.0.0. The concatenation of the two addresses is done using minimal encapsulated IP [149]. When a packet is crossing the boarder between the Internet and a private network, the NAT server sets the outer address accordingly. This allows standard routers to process the 4+4 packets without knowing about the inner address. This concept is allowing the unique end-to-end addressing of IPv4 nodes without requiring changes within the network (i.e. beside the NAT servers). The authors also provide a concept for the migration path from standard NAT to 4+4 addressing. A similar concept is proposed by [19], where the private address uses within the NATed domain is treated similarly to the Mobile IP home address and the public address similar to the CoA. Therefore, the NAT server is also performing home agent functionalities. The approach is offering mobile node reachability and limited mobility within the NATed domain. The authors proposed the combination with standard Mobile IP for macro-mobility.

Researchers have proposed novel internetworking architectures that have implications on how mobility management is performed. The authors of Nimrod [28] proposed a complete re-engineering of the Internet addressing and routing architecture, separating node addresses from the interface identifiers. Although Nimrod reached the state of a RFC [28] and even the alignment of Nimrod with the concepts of Mobile IP was documented in a RFC [157] as well, the novel architecture was not really adopted.

3.4.7 Conclusion

All discussed session management solutions propose to introduce a new name space (either by extending and combining existing, or creating completely new ones). Most of the solutions try to combine the benefits provided by NAT with the ability to promote end-to-end addressing to enhance and simplify the end-to-end session and mobility management. The success of such solutions is strongly dependent on the adoption of IPv6. If IPv6 is introduced, most of the problems the Internet has to face today can be solved with Mobile IPv6 and its route optimization. Although, the unfortunate utilization of the IPv6 addresses as host identifiers still remains. If, however, IPv4 continues to persist, or if the network address translation does not die out, alternatives that do not rely on use of public routable address as an end-point identifier may ultimately prevail.

The concepts of personal mobility discussed in Section 3.4.3 are very interesting with regards to our envisioned user centric approach to offer simple heterogeneous data connections. The creation of the additional personal layer is enabling users to be communication endpoints. In contrast to SIP, the concepts presented integrate user centric addressing with network layer functionality to control the data sessions. This cross-layer integration between the personal layer and the physical connection layers may be a basic enabler for seamless and intuitive networking.

3.5 Conclusion

In this chapter, we discussed the most relevant research work related to our envisioned framework. In the first section we analyzed the different published proposals to offer seamless session mobility across different communication technologies. The study of the different proposals the need for infrastructure to establish initial security relations between nodes and overcome the problem of simultaneous movements that pure end-to-end session management systems have to face. Network layer mobility provided by Mobile IP, especially with its version 6 including standardized route optimization features is definitively the most advanced mobility solution. Its support by the 3GPP consortium is further stimulating wide research efforts to improve handover performance. With regards to our envisioned scenario integrating infrastructure-less communication channels to increase the networking performance, Mobile IP combined with IPsec serves as an ideal session mobility management framework. The second section addressed the requirements of heterogeneous networking in terms of usability, handover decisions, resource and location management. The analysis of the evolution path from existing heterogeneous (but infrastructure-based) access networks to hybrid networks integrating ad-hoc and multi-hop links and finally also wireless personal area network enabling ambient networking helped to identify further missing parts in the big picture of future networks. In the third section, we elaborated on some related work introducing middle layers to better handle mobility for IP nodes. All presented solutions address the same problem, namely the misuse of the IP address as a host identifier, which is, especially with private IPv4 addresses and NAT, prohibiting a proper end-to-end session management.

Taking the issues discussed in these related work into account, we will analyze the integration of infrastructure-less connections into existing seamless mobility solutions in the next chapter. Based on these additional findings, we will present a new concept of user centric heterogeneous session management utilizing the existing mobile phone numbers in Chapter 4, 5 and 6. A key feature of our proposed architecture is the separation of the signaling and the data plane for heterogeneous IP data sessions, which solves the problem of simultaneous movements and enables the usage of existing low power signaling offering paging and authentication mechanisms to efficiently manage power demanding broadband data channels.

Chapter 4

Management of Heterogeneous IP Sessions

4.1 Introduction

In the previous chapters, a variety of communication technologies were analyzed in terms of their different characteristics and settings that have to be made to successfully use them. The seamless integration of infrastructure-based access technologies was discussed in Chapter 3. One of the major challenges raised by such a seamless integration is to do it in a way that the end user is not aware of the underlying heterogeneity and complexity. Heterogeneous IP communication has to become a commodity and therefore intuitive and convenient. This chapter studies the requirements on a signaling framework to manage heterogeneous networking.

4.2 Separation of Signaling and Data Plane

As discussed in the first chapter of this thesis, the trend in telecommunications nowadays is to offer communication systems where the users can be always reached, everywhere and anytime. The network should be able to learn about the user needs and provide always the most appropriate connection to satisfy them best. To do so, the network has to have means to learn quickly about these user needs, which can change rapidly. Environmental changes can highly influence the requirements on communication sessions. So it is not surprising that many people do prefer to use SMS or e-mail instead of making video calls when they are in public places where others can listen to their communication. But also the choice of the communication device has an impact on the selection of the communication technology. Broadband communication is still not very suitable for small devices due to limited screen size and power restrictions. Hence, having a possibility to learn more about the actual user needs and situations can ease the proper management of networking resources and therefore increase the overall network performance. To efficiently obtain these user needs prior to the actual session setup, separation of signaling and data plane is a powerful concept. Especially in heterogeneous environments, where the different communication technologies may have different characteristics in terms of availability, coverage, security, power consumption, capacity and price, the selection of a dedicated technology for the signaling independent of the technology used for the actual data can be highly beneficial. To allow fast communication session recovery in case of data network loss, it makes perfectly sense to keep the signaling on a robust and widely available technology and using the local area network for the actual data traffic only. The separation of the sensitive signaling channel from the data channels releases the security requirements for the data channels. If the signaling channel is offering a secured environment to exchange sensitive keying material to protect the transmitted data, the data channels themselves do not have to be highly secured, which makes the deployment much simpler.

While this separation of signaling and data plane is state of the art for the PSTN [98], it looks different for IP communications. IP is not distinguishing signaling information from data and therefore not offering dedicated channels/routes for signaling packets. Signaling for IP is realized with the Internet Control Message Protocol (ICMP), which is transported using standard IP packets. Being a packet switched network protocol, IP deduces the transport path purely from the destination address of the packets. It is not foreseen to select different routes to the same destination, depending on the type of information that has to be sent. Therefore, the complete signaling is done using the same communication channel that is used for the data transfer. This is referred to as inband signaling. This inband signaling was considered as a big advantage of IP and made the deployment of networks as well as applications much easier. But IP was designed without consideration of heterogeneous sessions. In heterogeneous environments the situation is different. Whenever data channels are precious and therefore subject to careful resource management, an autonomous narrow band signaling channel is desirable. Especially in mobile networks, inband signaling for IP sessions is not optimal. When looking at GPRS or UMTS, the nodes have to maintain a PDP context (see Chapter 2) to stay reachable for IP signaling, even if no actual data session is going on. Keeping the PDP context does not only require more radio resources, but also consumes more of the precious battery power. Introducing a low power out-of-band signaling concept for IP sessions on GPRS/UMTS can significantly reduce the power consumption by offering broadband connectivity on-demand. Consequently, the PDP context is only established when data sessions have to be started. In Chapter 5 this concept is presented and discussed in detail.

Besides reducing power consumption, out-of-band signaling can also help to handle scarce networking resources. As mentioned before, standard fixed telecommunication systems transport the signaling using narrow band signaling channels and the data via dedicated data channels, which support higher bandwidth. This separation of the signaling and data plane allows dynamic resource allocation. When setting up a session, the complete signaling message exchange is done using a narrow band channel, while the broader data channels can be allocated whenever required. The next section analyzes heterogeneous IP session for a better understanding of how out-of-band signaling can be applied.

4.2.1 The Need for a New Signaling for Heterogeneous Sessions

In the ISO/OSI networking model, the different layers are responsible to deliver services to the upper layer using the services provided by the lower layers. Whenever data has to be sent from one node to another, control is passed down the layers to set up the required connection. Therefore, each layer communicates with its peer at the destination node to properly deliver the services to its upper layer. This signaling is done among instances of the same layer. Hence, there are several signaling planes on top of each other. This signaling happens separately for each communication stack and independently from other communication technology. For homogeneous communication sessions this makes perfectly sense to keep the autonomy of the different networking interfaces. The usage of a WLAN interface, for instance, should not be dependent on the status of an UMTS interface. Although, heterogeneous sessions can profit from a tight collaboration between the different communication stacks. Having the possibility to handover active IP sessions from one communication interface to another, demands for a better synchronization of both interfaces. Lower layers have to be prepared correctly before the handover can occur. This preparation may include network detection and authentication. The higher layers have to be possibly adapted to cope with the new transport characteristics. For video applications, for example, the frame rate or resolution might be changed to suit the actual bandwidth.

A straight forward approach to manage this is to add a new signaling plane that covers all interfaces. The integration of infrastructure-less communication technologies has significant effects on the design of the signaling for heterogeneous sessions. Centralized radio resource allocation and authentication of the nodes, as well as address management and billing issues are solved and well deployed in infrastructure-based communication networks, while there is still a lot of research work ongoing to provide good solutions in the domain of pure ad-hoc networking. When considering heterogeneous networks offering both infrastructure-based services and ad-hoc flexibility, the signaling has to become heterogeneous as well. It has to be able to use existing functions for authentication, billing, and resource management and provide new functionality to securely cope with ad-hoc connections. To ensure a smooth integration with existing infrastructure, the signaling protocol for heterogeneous networks has to be designed as an extension of existing signaling systems.

Chapter 5 is addressing the impact of heterogeneous session management on session duration, bandwidth, energy, and network resource consumption. The simulations done focus on the impact of an integration of ad-hoc and direct node-to-node links with infrastructure-based access networks, especially in terms of networking performance.

4.2.2 Signaling Requirements for Heterogeneous Sessions

A common method to specify system requirements is taking use cases to deduce the system entities, components and processes that have to be supported. The use cases used to derive the system requirements are somehow extracted from the vision of seamless heterogeneous networking. The vision describes a system that cares about all available communication technologies on behalf of the user.

The system offers simple connectivity to any destination abstracting and hiding the heterogeneity of underlying communication technologies, devices, and networks. The envisioned system provides a simple tool to initiate heterogeneous sessions by inviting peers. These peers are ideally associated with human readable identifiers. Users usually want to communicate with other users and therefore do not care about FQDNs, IP, and MAC addresses identifying computers and terminals. When receiving an invitation for a heterogeneous session, the choice of device is often dependent on different conditions like type of application, actual location, and situation. Therefore, the system has to allow a flexible selection of the session endpoints during the session setup phase. Finally, the provided connection has to be secured and based on the most appropriate available communication technology. In other words, the connection has to seamlessly switch from one technology to another, whenever required. This might be the case, if the used network becomes unavailable or a new, more appropriate communication technology pops up. To map these key requirements to signaling specific features, the most important entities and use cases have to be analyzed in further detail.

Beside the requirements that can be deduced from the description of the user experience, the designed system is entitled to exploit the maximum benefit out of heterogeneous networking, in terms of network and energy resource management. Basically, the framework has to support signaling and data sessions over nearly any type of channels to assure full flexibility in terms of heterogeneity. To efficiently use heterogeneous network resources, the signaling bootstrapping has to be executable on any available communication technology.

4.3 Signaling Framework for Heterogeneous Data Sessions

To design a signaling framework offering the required features to enable seamless and convenient heterogeneous networking, as described in the beginning of this thesis, actors and use cases have to be extracted from the vision.

4.3.1 Actors and Signaling Entities

The vision portrayed in the introduction is very much focusing on the end user. Hence, it is obvious that actors are primarily persons that want to communicate with each other. These persons may have multiple devices forming a so-called personal area network (PAN) [94]. Fig. 4.1 shows such a PAN including a mobile phone, a PDA, and a laptop. All PAN devices are considered as personal devices belonging to the same person. Therefore, they are supposed to belong to the same administrative domain. In other words, the person owning the devices belonging to his PAN can easily configure them in terms of security and connectivity. So, secure communication among members of the same PAN is taken for granted. This assumption holds throughout the whole work done and presented in this thesis.

Even if the vision is user oriented, the designed framework should not be limited to communications between persons. The system should support person to machine and machine to machine communication as well. Even though the concept presented here is not restricted to person to person communication,

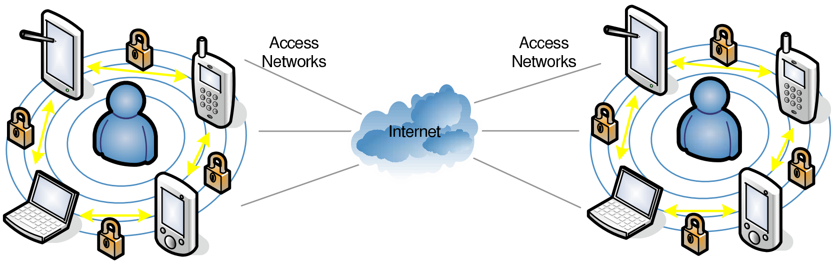


Figure 4.1: Personal Area Networks

the interfaces between the system and the actors have to be easily adaptable to allow machines to perform as actors. So to be precisely, there are two type of actors identified, persons and machines. For simplicity reasons the later will not be treated separately in the following considerations.

The vision gives evidence of the desired flexibility to choose for each communication session the preferred device to be used. Both, the initiator and the invitee can select a device out of their PAN to handle the communication session. Since the users do not care about the device selected by the communication peer, the communication can be abstracted in two sessions: One communication between the users, and one between the actual devices selected by the users. The session between the users is referred to as logical communication session, and the one between the devices as physical communication session. This abstraction reflects very much the main goal of the system, namely the simplification of heterogeneous networking. The variety of access interfaces of the devices is hidden from the users as much as possible. Users only have to select one device out of their PAN and do not care about the different devices of the peer's PAN. For the signaling system, this ends up in having two kinds of signaling entities, the users and the devices. The users serve as signaling entities for the logical and the devices for the physical communication session. The term physical session is only representing the fact that physical devices are involved and not limiting the session to direct physical connections.

4.3.2 Logical and Physical Session Signaling

The separation into a logical and a physical session has also to be reflected in the signaling system. Therefore, two signaling layers were created, the logical session signaling (LSS) for the human related and the physical session signaling (PSS) for the device related session setup. The two layers are shown in Fig. 4.2. Both layers are tightly correlated. The LSS layer is offering a initial signaling channel between the involved users (1), whereas the PSS layer is used afterwards to exchange parameters specific to the physical session management (3). When looking at the different use cases, it will be quite evident that the user is somehow acting as connecting link (2) between the LSS and the PSS. When receiving a session request (LSS), the user selects the most appropriate device and initiates thus the corresponding PSS to set up the final communication links between the devices.

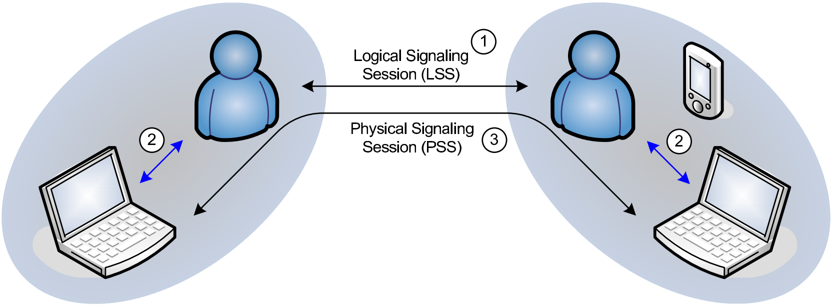

Logical Signaling Session (LSS)
1
Physical Signaling Session (PSS)
3
2
2

Figure 4.2: Logical and Physical Session Signaling: 1 LSS offers initial signaling between the involved users, 2 dynamic selection of the communication device, 3 PSS establishes the actual link

4.3.3 Identifiers and Communication Addresses

To successfully exchange information between signaling entities, an addressing scheme has to be available. With regard to the two signaling layers specified in the prior section, requirements to these addressing schemes are different for each layer. The separation of the signaling plane into two layers has been done to abstract the different physical connection by offering a sort of human to human communication session, which is then relayed to some dynamically chosen devices. This abstraction hides the complexity of the underlying communication technologies to the user. Therefore, the addressing scheme used for the LSS should be human readable and intuitive as well. The most popular addressing schemes designed to be human readable are the E.164 [97] defined by the International Telecommunication Union (ITU) and the Network Address Identifier (NAI) being standardized by the IETF, used in the telecommunication and the IT domain, respectively. The main difference between these two addressing schemes, with regards to heterogeneous networking, is the level of availability. Thanks to the wide spread cellular telephony networks subscribers are nearly always reachable through their E.164 address (aka MSISDN) and the according signaling system SS7 [98], whereas NAIs are often used for nomadic applications due to the required IP connectivity to be reachable for the signaling (e.g., SIP 2.5.3). The signaling should be performed in an inband manner, if and only if IP connectivity is available anyway. However, if IP is available the LSS can also use NAI instead of E.164 phone numbers. Any other low power and low bandwidth channel should be usable for the proposed heterogeneous session signaling otherwise. As long as IP connectivity is not permanently available for signaling of heterogeneous sessions, we propose to use the SS7 network for the LSS.

Using the E.164 addresses for the LSS is not only beneficial in terms of availability in combination with the cellular network, but helps also to enhance the user convenience. Re-using standard phone numbers to identify and address the peer of a communication, like it is done when performing a voice call, is essentially increasing the degree of the acceptance by the end users. Heterogeneous networking has to become as easy as making voice calls. People are used to identify peers based on their phone numbers. Having electronic address books on nearly any communication device is even simplifying the mapping of names to phone numbers. The MSISDN is gaining more and more momentum in daily life to identify and contact people, much more than fixed phone numbers. The mobile phone number gets even more important than any other means to contact a person. The fact that a cellular subscription is nearly almost treated as something personal and therefore non-transferable, makes the MSISDN an appropriate identifier also for heterogeneous session management (i.e. for the LSS). Furthermore, mobile subscribers are used to have their mobile phone always switched on and always with them, which guarantees a maximum of reachability. Throughout the rest of this document, the E.164 addresses used for the LSS entities are referred to as identifiers to clearly separate them from the addresses used for the PSS.

In contrast to the LSS, messages belonging to the PSS are of relevance for the devices only and not necessary visible for the users. Consequently, the PSS can easily use standard machine readable address schemes used in the IT environment like Bluetooth, MAC, and IP addresses to distinguish signaling entities. For the sake of clarity, the term *communication address* will be used as representative for any of those physical session related addresses, whenever possible.

4.3.4 Address Resolution

As mentioned in the Section 4.3.2 the dynamic selection of the device for each communication session is determining the relation between the static identifier (e.g., MSISDN) of the LSS and the currently valid communication address (e.g., IP) of the PSS on the fly. This process is known as *address resolution* and normally done in a centralized manner. DNS [135] is probably the mostly used address resolution system for IP networks. With the extensions proposed by ENUM [59] called NAPTR [133]) the DNS can also be used to resolve E.164 phone numbers (e.g. MSISDN) to URIs [35] and hence NAI or IP addresses. ENUM is bringing the world of telecom and the Internet closer to each other by enabling a simple mapping of phone numbers to URIs and therefore enabling E.164 addresses to be used as primary identifiers for IP applications. The most obvious application taking advantage of this mapping is Voice over IP (VoIP) where the usage of E.164 phone numbers to identify the peers is crucial for a co-existence with the public switched telephony system (PSTN). DNS is handling the records to resolve the IP address of a URI or E.164 phone number (with the help of ENUM) in a rather static manner. Although there are proposals to allow end nodes to dynamically change DNS records, the centralized architecture can not handle changes fast enough to serve realtime modifications. It is not foreseen to use DNS/ENUM to handle IP mobility. DNS/ENUM is mainly used to bridge phone numbers with NAI for example, which makes the use of MSISDN and NAI interchangeable for the LSS. Unlike DNS and ENUM, the address resolution for our LSS/PSS system requires much more flexibility and has therefore to be done decentralized to allow dynamic and fast resolution even during the session setup procedure. Furthermore, the process required here is somehow different from standard address resolution, since the initiator does not care about which device will be used for the communication session. The initiator only expects that the PSS request is forwarded to any device belonging to the invitee. Therefore, it is rather more similar to an any-cast routing in IPv6 than to standard DNS.

It is worth mentioning that existing protocols like SIP are not addressing the problem of such a highly dynamic identifier to address resolution, which has furthermore to be controllable by the destination node on a per session base. SIP assumes IP connectivity to keep the presence information on the registrar server up to date. SIP users regularly have to inform their centralized registrar server about their availability and reachability (i.e. IP addresses). Mobile IP has similar constraints, since the home agent has to be informed about the actual CoA of the mobile node. Hence, communication sessions can not be initialized if no IP connectivity is available. This limitation is very much related to the inband signaling of Mobile IP. Although SIP would theoretically allow the signaling to happen on a different IP address than the one used for the actual application session, the signaling itself requires permanent IP connectivity.

4.3.5 LSS and PSS Deployment

The vision reflects an easy and simple system to establish, maintain and tear down heterogeneous communication sessions between users having different devices. To a certain extend, the users should not care about the session maintenance, for example the handover from one technology to another. Whenever this handover is mandatory (i.e. network initiated) the users should not be bothered with it. Only if there are options for the users to choose, an interaction is desired, especially, if the choice has an impact on quality or cost of the communication. Users only have to interact with the LSS, whereas the PSS is handling the heterogeneous communication sessions in terms of connectivity and security. Hence, the use cases can be limited to the user interactions with the LSS. User interaction is further required when setting up and tearing down a communication session. To guarantee the highest level of reachability the entity handling the LSS has to be always operational and close to the user. Furthermore, the LSS has to be always reachable for incoming session requests. These requirements are very similar to the ones of a mobile phone, which is not surprising when considering the objective of making heterogeneous communication sessions as simple as voice calls. The node hosting the LSS, and being therefore connected to the network used as primary signaling plane, is also interacting with the user. This node is referred to as *supernode*. This term reflects best the privilege of that node to interact with the user on behalf of all other devices present in the PAN. In the case where the actual physical session is terminated on the supernodes as well, both, the LSS and the PSS are located on that device. In terms of system architecture, this is only affecting the interprocess communication between the LSS and the PSS. Either both processes are located in the same device or distributed in different devices belonging to the same PAN. Assuming regular connectivity among the devices, the inter-process communication can be done using standard IP sockets in both cases. The IP routing is then taking care of the proper information exchange between the LSS and the PSS. For the sake of completeness Fig. 4.3 and 4.4 show both, the setup in the case of distributed and collocated LSS and PSS. In the distributed case, the LSS is supposed to run on the mobile phone and the PSS on the other PAN devices (Laptop, PDA, etc.). In the collocated case, both signaling layers are on the laptop having access to the LSS plane. Since there is no big difference concerning the system architecture, both cases are used interchangeable. Depending on the aspect of interest one or the other visualization will be used in the subsequent sections. Whenever the PAN aspect is of importance the distributed scenario will be chosen. Otherwise the second scenario is used for sake of clearness.

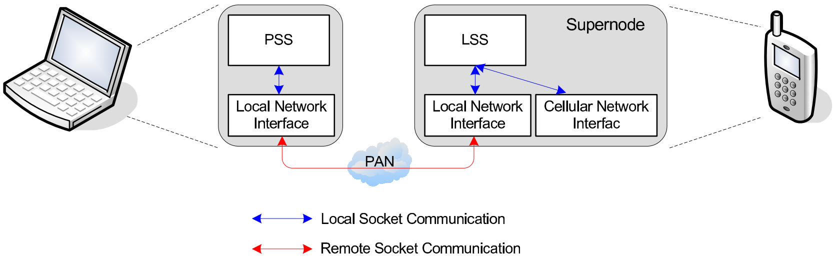

PSS
Local Network Interface
LSS
Supernode
Local Network Interface
Cellular Network Interfac
PAN
Local Socket Communication
Remote Socket Communication

Figure 4.3: Distributed LSS and PSS

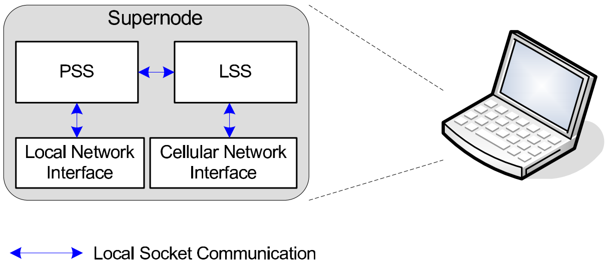


Figure 4.4: Collocated LSS and PSS

4.3.6 Session Maintenance

The session maintenance is probably the most demanding part of the system. When talking about session maintenance in the domain of heterogeneous networks, it mostly comes down to session handovers among different technologies. The technical hurdles that have to be taken to successfully manage handovers have been discussed in Chapter 3 and some of the economical aspects will be studied in Chapter 5. The system designed should mainly offer the required functionality to cope with the session mobility mechanisms identified in the previous chapter, namely Mobile IP and an interface to learn about the different available data channels. In other words, there is a cooperation between the designed system, Mobile IP and any future system to manage heterogeneous network resources. The user's preferences together with the preferences of the resource management system and the available networking technologies are influencing the handover decision. Handover decision is based on information provided by the PSS layer and therefore not dedicated to end users. This makes it difficult to involve the user when taking the handover decision. To keep the heterogeneous session management simple and convenient, the handover alternatives have to be presented in a way that the user can decide without having to understand the technical details. Like in session mobility solutions for infrastructure-based access networks, the alternatives have to be characterized in terms of connection quality, costs and maybe power efficiency. To do so, the physical characteristics of the different alternative communication technologies reported by the PSS have to be classified relative to the actually used technology. Whenever a new communication technology becomes available, this relative comparison has to be done and presented to the user if he can benefit. There are several approaches to automate this handover decision on behalf of the user but also on behalf of the network operator. Most of them rely on so-called profiles reflecting the users preferences to further reduce the user interaction (see Section 3.3.3). Independent of which component finally decides about any handover, the PSS has to provide all the required information. Fig. 4.5 is illustrating the intra-device handover.

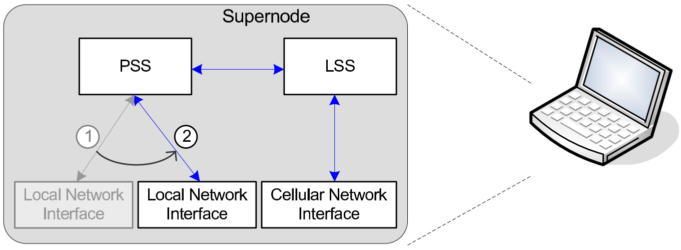


Figure 4.5: Intra-Device Session Handover

Another kind of handover can occur between devices instead of communication technologies. This might be desirable for specific applications like multi-media sessions or streaming. From a system perspective, such inter-device handover is nothing else than selecting another PSS device during an ongoing session. Compared to inter-technology handover addressed before, the reselection of a new PSS entity has to be done prior to the actual session handover. This reselection can be done independent of the ongoing session until the new device is ready to take over the session. Therefore, standard session setup procedures can be used between the LSS and the new PSS entity. Whenever the new device is prepared a standard handover message can be sent to the peer to fulfill the device handover. For layer four sessions this kind of inter-device handover requires a restart, since the Mobile IP home address can not be transferred from one node to the other without session interruption. In Chapter 3, we discussed some solutions allowing to migrate ongoing session from one IP address to another. To seamlessly perform inter-device handovers such solutions are required. However, the system presented in this thesis offers a framework for those session migration solutions to easily exchange signaling information between the involved nodes. Fig. 4.6 illustrates the inter-device handovers that can occur during an heterogeneous communication session.

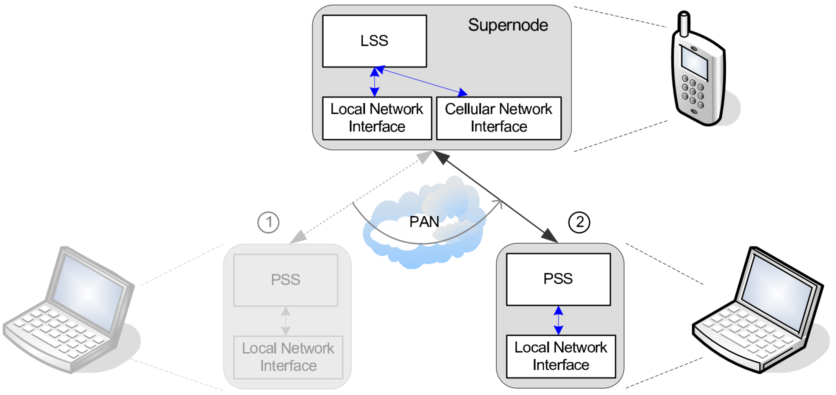


Figure 4.6: Inter-Device Session Handover

4.4 Optimizing the Data Path

Nowadays short range communication technologies like Bluetooth, WLAN, and in near future also UWB, offer very high data rates compared to wider range technologies like EDGE or UMTS. When referring to ad-hoc and direct node-to-node links in this thesis, exactly those short range technologies are of interest. Most of these communication technologies become more and more available in portable and mobile devices. However, the utilization of those short range communication technologies to exchange data between nodes is too complicated for most end users. A minimum of knowledge is required to successfully interconnect nodes. When requiring a secure connection, the setup becomes even more complicated. Due to the decentralized structure of these technologies the configuration needed to setup secured links comprise the negotiation of several parameters. The design of Bluetooth and other PAN capable communication technologies have been done aiming at providing simple and intuitive handling. Even if this was successfully done in terms of service detection (see Section 2.2.5), the establishment of secure sessions was not adopted by the users. The process of entering a PIN on all devices (aka pairing) to protect the link with minimal encryption, is not convenient and therefore not widely used. Consequently, the number of attacks in highly populated locations, like train stations and airports, increased rapidly. Attackers connected to mobile phones having Bluetooth security disabled, and accessed personal information like contacts and call lists. Connecting to a GPRS or UMTS enabled phone, attackers could even start broadband Internet connections on victim's account. In [104, 67, 68] the different types of attacks are discussed. The Bluetooth security vulnerabilities, coming mainly from improper handling of the security mechanisms, lead the users to simply disable it. WLAN has to face similar problems. The usage of WLAN in infrastructure mode to access private and public access networks is widely adopted, but nearly no one is using the ad-hoc mode to directly interconnect mobile devices. Unlike Bluetooth, WLAN ad-hoc mode requires even more configuration efforts to be done, before any data can be transmitted or received. WLAN is not offering any service discovery mechanisms to scan for available communication services. WLAN is only offering pure layer two connectivity. It is up to the end user to correctly set the communication stack on the different nodes to establish a connection. Hence, powerful short range communication technologies like Bluetooth and 802.11 (ad-hoc mode) are not yet appropriately considered when talking about heterogeneous data connections. Nevertheless, the potential benefit of integrating these ad-hoc links in terms of connection performance is remarkable (see Section 5.3 and [42]) and will be substantial when thinking about the enormous data rates offered by technologies like UWB. The ability to seamlessly switch ongoing data sessions to such an high bandwidth short range technology, offering easily up to order of magnitude higher data rates than infrastructure-based systems, would therefore reuse existing and unused resources to enhance the overall networking performance. To better understand the functionality required to extend the *always best connected* approach introduced in Section 3.3.2 to infrastructure-less links, some further aspects have to be considered. We first analyze what is missing to enable Mobile IP route optimization to perform session handovers also to direct node-to-node links. In the second part of this chapter, we describe how the reuse of an infrastructure-based network like the cellular network could solve the identified issues and eventually enable Mobile IP to successfully handover the sessions.

4.4.1 Infrastructure-less Connections

The route optimization feature of Mobile IPv6 combined with the ability to register any link local CoA using the alternate CoA mobility option enables the switching between infrastructure-based access networks like the cellular and ad-hoc based links like UWB. If two nodes being attached to the cellular network would come close enough to use direct communication, the route optimization could be used to reroute the data packet directly between the two nodes. This switching over from the infrastructure to the ad-hoc mode is illustrated in Fig. 4.7.

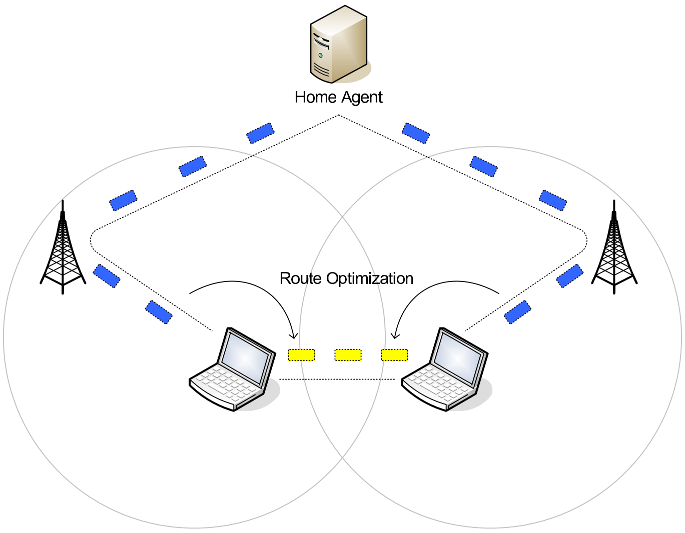


Figure 4.7: Mobile IP Route Optimization to Perform Handovers between Infrastructure-based and Ad-Hoc Links

Unfortunately, Mobile IP by itself is not able to prepare the direct link for route optimization. Mobile IP is a pure layer three protocol and hence does not have any facilities to handle lower layer functionality. In the special case of infrastructure-less communication, where no centralized instance is managing the connection setup, communicating nodes have to agree on configuration and security parameters to successfully establish the IP link required for the route optimization process. Before there is no IPv6 address assigned to the network interface used for the ad-hoc links, the Mobile IP Route Optimization is not able to communicate that link local CoA to the correspondent node. It requires therefore some external mechanism to detect the peering nodes and setup the communication stack of the ad-hoc interface up to layer three.

Fig. 4.8 illustrates the failure of the Route Optimization process to switch over to ad-hoc link even if nodes are within the vicinity.

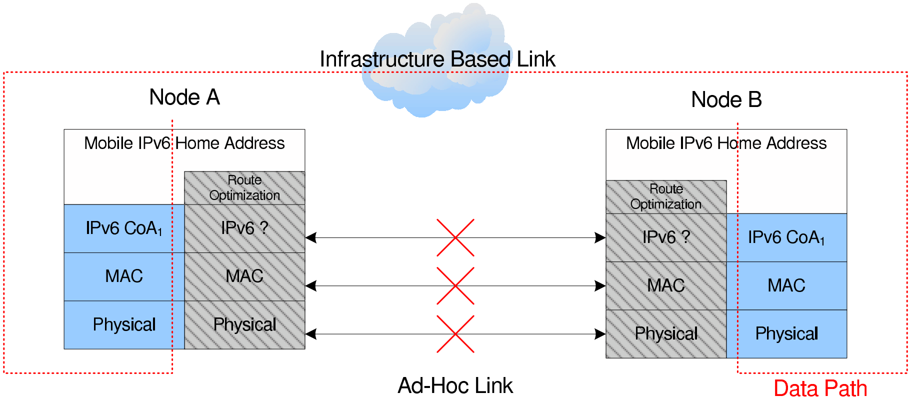


Figure 4.8: Failure of Route Optimization in the Case of Direct Links

With the help of an external mechanism taking care of the configuration of the lower layers (i.e. including layer 3, CoA_2), Mobile IP can be triggered to send a binding update including CoA_2 to the correspondent node. Afterwards, the Route Optimization can be used to redirect the traffic through the ad-hoc link using the Route Optimization process.

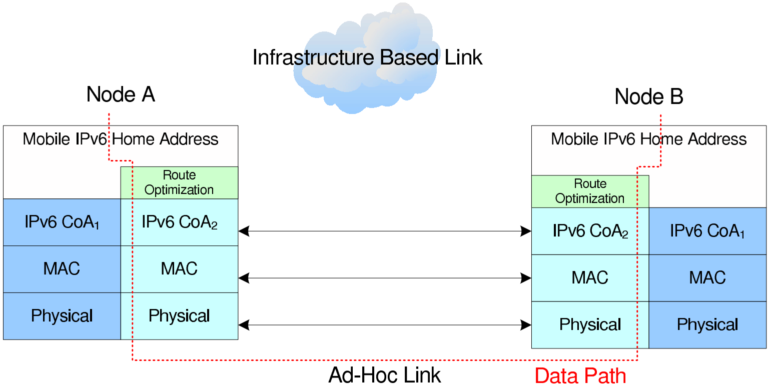
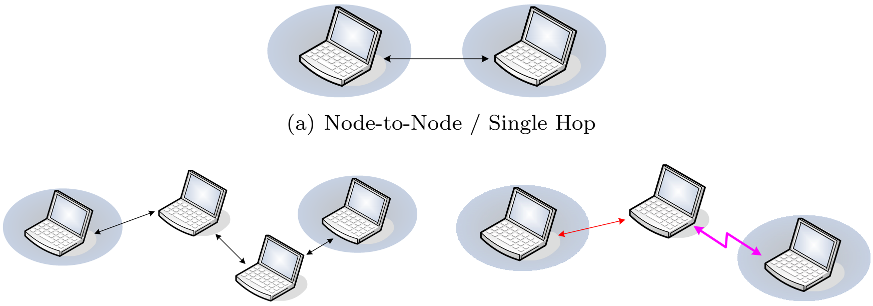


Figure 4.9: Route Optimization with External Bootstrapping Process

IPsec and Mobile IP offer the ability to handover sessions seamlessly and securely from one access technology to another. However, since both protocols are acting on the networking layer they can not prepare the lower layers to eventually perform the handover on layer three. Especially in the wireless environment this preparation of the lower layers requires several actions. Radio parameters like modulation and medium access schemes need to be correctly set to enable basic communication. Finally, properly set IP configuration is required to use IPsec and Mobile IP, which often requires authentication processes. This bootstrapping problem of Mobile IP and IPsec as pure layer three protocols becomes even more noticeable when considering infrastructure-less communications. The missing infrastructure is making the negotiation of the parameters required to successfully establish secured data channels very complicated.

The most straightforward way to solve this bootstrapping problem is to introduce a dedicated channel between the mobile nodes to securely negotiate the configuration and security parameters. In the context of heterogeneous networking including also infrastructure-based access networks this dedicated channel is not disruptive. The architecture proposed in this thesis offers such a dedicated signaling channel to securely establish direct links enabling Mobile IP route optimization to use optimized paths, namely the ad-hoc and direct connection between two nodes (see Chapter 4, 6). The rest of this chapter is hence further investigating on the required functionality that has to be provided by this external process to enable seamless switching to infrastructure-less connections.

For the Mobile IPv6 route optimization, the exact structure of the new path is not relevant. Only IP connectivity is required to successfully perform the route optimization. Therefore, any type of interconnection can be used to enhance the characteristics of the data exchange between nodes. Fig. 4.10 illustrates the three major types that have been considered as important with regards to seamless and convenient heterogeneous networking.



(a) Node-to-Node / Single Hop

(b) Multi-hop / Ad-Hoc

(c) Heterogenous Links

Figure 4.10: Types of Interconnections

To establish pure IP connectivity, the three types of connections require different configuration and security parameters to be negotiated.

Node-to-Node / Single-Hop

Using *Single-Hop* links is the most simple way to interconnect nodes. Only the session endpoints are required to establish the links. This makes the negotiation of configuration and security parameters very straight forward. Knowing any layer two identifier of the peering node allows the scanning process to detect whether the peer is within communication range. All IEEE 802 based technologies offer unique layer two addresses based on the 48 *bit* MAC address specification. Hence, the MAC address can be used to identify the detected nodes. This assumes that the MAC address of the communication peer is known prior to the connection establishment. In Chapter 4, a new session management is proposed to dynamically learn the MAC address of the interfaces of the peer. After having correctly detected the peer node, the connection can be established. Security related credentials have to be negotiated to secure the link prior to route sensitive data packets over it. Depending on the technology, different parameters have to be negotiated to successfully interconnect nodes (see Chapter 2). Finally, the IP address assignment needs to be coordinated between the nodes. In single hop connections, this address assignment can be done rather easily because no actual IP routing occurs. As soon as the direct link is established, the Mobile IP route optimization can handle the redirection of the data packets.

Multi-Hop

In *Multi-Hop* environments, the situation looks by far more complicated. Intermediate nodes are required to guarantee IP connectivity by forwarding IP packets. The management of multi-hop networks is out of scope of this thesis. A lot of research work is going on in the domain of multi-hop ad-hoc networks. Routing mechanisms proposed in the literature to assure the packet forwarding and efficient routing in such multi-hop ad-hoc networks are offering IP connectivity between participating nodes [101, 180, 198, 74]. With regards to route optimization in heterogeneous networking environments, the multi-hop ad-hoc networks interconnecting the end nodes are not explicitly perceived as such. The multi-hop link offered by the ad-hoc network is considered as an alternative link to interconnect the communicating nodes as illustrated in Fig. 4.11.

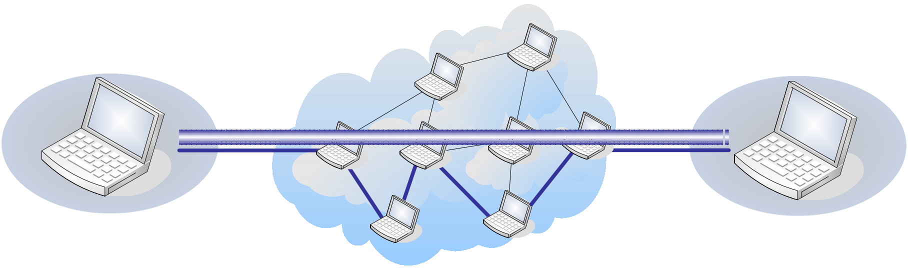


Figure 4.11: Multi-hop as an Alternative Heterogeneous Link

Ad-hoc routing protocols combined with IP address assignment mechanisms for ad-hoc networks provide IP connectivity if the network between the communication peers is not fragmented. The IP addresses used for the ad-hoc network are considered as CoA for the Mobile IP route optimization process. Unlike single-hop connections, where link layer security can be established between the communicating peers, end-to-end security is favorably implemented on layer three (e.g., IPsec). With respect to the envisioned framework for seamless and convenient heterogeneous networking, multi-hop links are considered similarly to infrastructure-based links. Compared to single-hop links, no layer two configuration or security parameters have to be negotiated and exchanged between the end-points for multi-hop ad-hoc connections. The ad-hoc routing process tries to get access through an ad-hoc network similarly to an infrastructure-based network.

Heterogeneous Links

Single-hop connections are implicitly built on homogeneous links. All communicating nodes require the same technology to establish direct links. Multi-hop ad-hoc links could theoretically rely on heterogeneous links, but this would drastically limit the flexibility of the network topology and some nodes would have to act as technology bridges to avoid the fragmentation of network into homogeneous parts. Such multi-mode nodes, being capable to communicate on several technologies and therefore taking over the role of technology gateways, may also enable communication between nodes lacking a common communication interface. Hence, such heterogeneous multi-hop ad-hoc networks can be considered very similar to homogeneous multi-hop ad-hoc networks, but having a smaller number of links if not all nodes are supporting all communication technologies. With respect to our envisioned integration of infrastructure-based and direct links, there are no further requirements compared to homogeneous links.

The three mentioned types of infrastructure-less connections are very similar from a data path optimization perspective. IP connectivity needs to be established before Mobile IP can interact. Pure IP connectivity provided by any other means is enabling Mobile IP route optimization. For multi-hop links we can rely on dedicated ad-hoc routing protocols and architectures, whereas there is no bootstrapping mechanism for single-hop connections.

4.4.2 Configuration and Security Issues

The structure of the message exchange used in the Mobile IPv6 Return Routability allows the approvement that the origin of the binding update is the mobile node. But it does not help to establish an end-to-end security relation between the mobile node and the correspondent node. The two-way sending of the *Home Test* and the *Care-of Test* messages through the home agent and directly to the correspondent node, respectively, reuses the security relation between the mobile node and the home agent and the topologically correct routing to proof the binding between the home address and the new care-of address. Mobile IP does not explicitly specify how the initial security relation has to be established. Mobile IP only guarantees that if a connection has been established to a node with a specific home address, only that node can send valid binding updates. The establishment of the end-to-end security relation between the mobile and correspondent node can be done with IPsec. The integration of Mobile IPv6 and IPsec is further pushing the use of IPsec for mobile nodes. Nevertheless, the problem of the initial security relation remains. IPsec requires a key management system, which takes care of that. IPsec can handle shared secrets and public/private keys to set up the initial SA, whereby shared secret key management is generally not considered as scalable. Authentication based on public/private key pairs and certificates [78] require a Public Key Infrastructure (PKI). To assure the authenticity of the peering node, the certificate of the corresponding public key has to be validated before starting any secured communication with the peer. The PKI provides certificate revocation lists (CRL) where invalid and bogus certificates are listed. Consequently, when using IPsec with public/private keys to set up the end-to-end SA, mobile nodes require access to the PKI prior to the connection establishment to the peering node. Unlike to the actual data exchange between the communicating nodes, this verification process requires only narrow band communication channels. The ability to access the PKI through the cellular network to establish IPsec protected connections between end-nodes highly motivates the marriage of infrastructure-based and ad-hoc networks.

Once a SA is available between the communicating nodes, Mobile IPv6 route optimization can securely perform the switching between infrastructure-based and direct communication. Through the established SA the nodes can negotiate the required parameters to successfully build up the network layer connectivity through the infrastructure-less interface. The infrastructure-based network (e.g., the cellular network) can be used to securely exchange the initial parameters like the layer two addresses required for the bootstrapping on the direct node-to-node connection. In combination with the latest proposals of IKEv2 [107], which allows to change IP addresses of ongoing IPsec sessions, the heterogeneous end-to-end session can be handed over from one communication technology to another. If needed, the SA can also be used to derive security parameters for the direct link in the case of single-hop connections. This might be of interest to reduce the computational efforts imposed by IPsec if the security mechanisms offered by the layer two are offering sufficient protection. Fig. 4.12 illustrates the relation of SA, PKI, direct link and end-to-end session security.

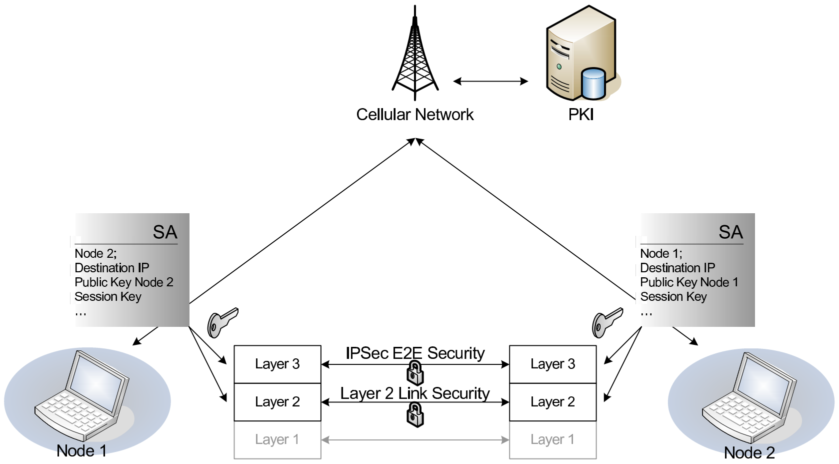


Figure 4.12: Security Framework to Protect Heterogeneous Sessions Based on Public/Private Keys

The cellular network guarantees permanent access to the PKI to verify certificates and establish SA whenever needed. These SA can then be used to protect any direct node-to-node connection between the communicating nodes. The framework developed in thesis mainly addresses the signaling required to establish and manage heterogeneous and secured end-to-end sessions. The security framework depicted in Fig. 4.12 is one possibility to bootstrap the SA required to protect the layer two and three of the communication stack. The establishment of the secured direct links between the communicating nodes, to seamlessly switch over the ongoing data session from infrastructure-based to infrastructure-less communication, increasing the networking performance, is novel. The utilization of the cellular system to initiate the SA between the nodes would also allow the use of shared secrets instead of a complete PKI. The fact that each cellular node has a security relation with the operator enables a trust chain between the nodes. Hence, if the nodes trust the cellular operator simple shared secrets can be exchanged to form the SA protecting the communication session. Similar to the SA built on public/private keys, the shared secret SA can serve as a basis to derive the actual link and session keys (see Fig. 4.13).

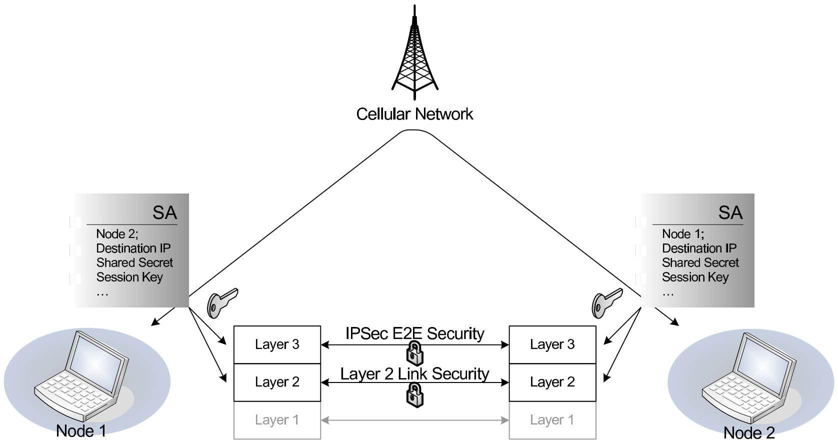


Figure 4.13: Overall Security Framework to Protect Heterogeneous Sessions Based on Shared Secrets

The system developed and presented in Chapter 4, 5, and 6 focuses on the definition of a signaling framework, and does not restrict the selection of the security mechanisms used to establish the SA between the nodes. However, the two presented mechanisms to adopt puplic/private and shared keys to derive link and session security parameters are fully supported by our signaling framework.

Additionally to the security parameters, some communication technologies require also configurations to bootstrap the connectivity. To interconnect nodes with WLAN, for instance, special parameters like channel number, SSID, and mode of operation have to be set correctly. Similar to the security related settings, these configuration parameters can be negotiated during the initial session setup process (see Chapter 6) using the cellular link and the signaling framework proposed in the next chapter.

4.5 Overall System Architecture

The analysis of heterogeneous communication management made in the previous sections motivated the separation of signaling and data channel management. The signaling can advantageously further be separated in LSS and PSS related functionality. Addressing logical session management towards the end users, the LSS is delegating the physical session management between the devices to the

PSS. This abstraction allows a clear separation between human and hardware related functions, which is also enabling a modular handling of the heterogeneous networking, if properly reflected by the system architecture. The hierarchical structure of our system architecture abstracts stepwise the various network technologies and interfaces by providing a generic session management, which is completely technology independent. Through this generic session management, any connection is represented similar to the user, independent whether it is infrastructure-based or infrastructure-less. This abstraction is achieved with the introduction of the *Smart Multi-Access Communication Service* (SMACS) layer, which is representing the logical session management.

SMACS

SMACS is using the functions offered by the underlying layers to provide simple handling of the logical sessions towards the user. Logical sessions can be initiated and terminated by the user. Incoming session requests can be rejected or silently dropped. The connection abstraction offered by the logical session management allows the user to initiate a session by selecting a destination based on its human readable identifier (e.g., E.164 or NAI). After the user has selected one of his devices to be used for the session, SMACS is using the functions of the lower layers to form the connection request, which is then sent to the session peer. If the peer accepts the request, the physical session management is taking care of the actual connection establishment between the two end devices. There are two modules providing the required functionality of the physical session management to the SMACS layer called CAHN and SecMIP.

CAHN

The *Cellular Assisted Heterogeneous Networking* (CAHN) is offering two major functions. First, it is taking care of all negotiation with the peer to successfully establish the physical session. It therefore relays the signaling messages required to exchange configuration and security parameters to the destination using the most appropriate signaling channel. A dedicated signaling protocol called *CAHN protocol* is introduced to properly handle the signaling information exchange between the nodes. Second, CAHN is handling the network interfaces that are able to establish ad-hoc and direct node-to-node connections. Both functions together enable CAHN to solve the bootstrapping problem discussed in Section 4.4. The configuration and security parameters required to securely establish the infrastructure-less connection between the nodes can be exchanged with the signaling mechanisms offered by CAHN.

SecMIP

The SecMIP module, introduced in 3.2.3 is taking care of the Mobile IP and IPsec functions. It hence allows SMACS to seamlessly and secure handover active sessions between networking technologies. Together with the bootstrapping of infrastructure-less connection by CAHN, Mobile IP Route Optimization can be used to seamlessly switch ongoing sessions between infrastructure-based and infrastructure-less technologies.

Fig. 4.14 shows the overall system architecture, including the SMACS, the CAHN, and the SecMIP modules. The separation between the logical session management between end-users and physical session management between the devices is reflected by the functionality provided by the three different modules.

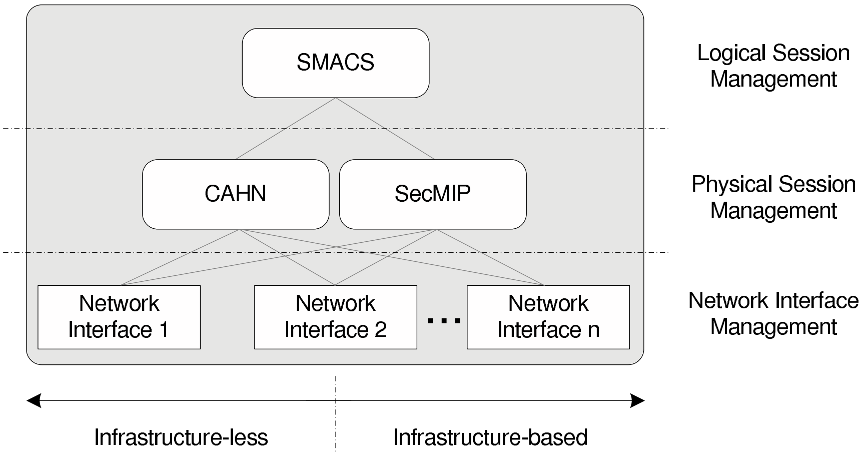


Figure 4.14: Interaction of SMACS, CAHN, and SecMIP

The SMACS and CAHN modules are explained in detail in the Chapters 5 and 6, respectively.

4.6 Conclusion

This chapter analyzed the requirements of a heterogeneous session management in terms of signaling and data channels. Actors and signaling entities where identified and the separation of the heterogeneous session into a logical and a physical session was proposed for a better abstraction of the heterogeneity of communication technologies.

We also analyzed the issues related to the optimization of the data path using Mobile IP and IPSec on infrastructure-less connections. The fact that both Mobile IP and IPsec are acting on layer three only makes them also dependent on an external mechanism that can prepare the lower layers to finally enable a seamless session handover. The integration of these ad-hoc and direct node-to-node communication technologies with infrastructure-based access networks has the potential to overcome these hurdles. Reusing the centralized services deployed to provide authentication and data protection for infrastructure-based access network also for the establishment of secured direct links can decrease the inhibition threshold that these ad-hoc and infrastructure-less technologies have to face nowadays.

Furthermore, the difference between identifiers and communication address was introduced to better understand the process of authentication when having multiple communication addresses. This lead to the conclusion that an initial security relation is required between nodes to bootstrap the establishment of a secure communication. This initial security relation can then be used to securely perform address resolution and exchange keys to set up secure communication using ad-hoc and node-to-node links.

Chapter 5

Smart Multi-Access Communications SMACS

5.1 Introduction

The introduced concept of separating the signaling and the data plane is allowing to combine the best out of two networking paradigms, namely infrastructure and ad-hoc networking. The vision of being always-on and always best connected introduces a major conflict between the interests of a network operator and the ones of the end users. The most obvious discrepancy is probably the motivation to use direct links between nodes. Offering better performance in terms of bandwidth and costs, it might become the first choice to exchange large amounts of data between nodes. However, simulations presented later in this chapter are showing that the benefit of using direct links is considerably increased if infrastructure-based networks can take over the session whenever the direct link is lost. Introducing an intelligent resource management for heterogeneous network resources could help to reduce the cost of a data session. But the usage of infrastructure-less links might also become economically interesting for the network operator. If the price of broadband connectivity is further falling, the operators will be forced to reduce the costs of their infrastructure. The possibility to offer connectivity through infrastructure-less technologies, which are already built in nowadays devices, may offer an interesting option. Due to the missing authentication infrastructure, the provisioning of security features in pure ad-hoc networks is still challenging. In hybrid networks combining infrastructure-based and ad-hoc elements some of the problems can be solved. Therefore, efficient and secured signaling over infrastructure-based networks will probably be a key value proposition, even if the actual data is transported using ad-hoc networks free of charge. With the help of the concept and architecture proposed in this thesis, such a general signaling service could be offered to handle any type of transport network[1]. To further increase the efficiency of the signaling of heterogeneous end-to-end sessions, we propose to do only the bootstrapping through the valuable cellular network, if no other secured IP connectivity is available. Furthermore, the signaling messages are sent inband

[1] The proposed system is not limited to mobile wireless broadband networks, even if the fixed and wired technologies are not explicitly treated in this document.

after successful data channel setup, whenever possible. The end-to-end session management, together with the separation of signaling and data related information is facilitating the concept of being reachable without wasting valuable networking resources. By using low power and low bandwidth signaling channels to notify a node about connection requests, scarce broadband resources can be saved without loosing the advantages of being always reachable. Section 5.2.3 is addressing the possibilities of such a *broadband on-demand* in further detail. Detailed evaluations on the improvement potential of an intelligent end-to-end session management for the end user experience and for the operator are presented in Section 5.3.

5.2 SMACS Framework

SMACS enables the smart usage of all available communication technologies. Using the two underlying modules CAHN and SecMIP it can immediately establish a secured ad-hoc and direct connection if the peer node is within vicinity and trigger Mobile IP to seamlessly handover the data traffic to that optimized path. This ability to handle heterogeneous infrastructure-less connections is further explained in next section. The signaling protocol offered by CAHN enables the exchange of configuration and security related parameters prior to the data channel selection, which might be used for heterogeneous network resource management. The ability to reach a node through a narrow band signaling channel without requiring the node to be connected permanently to a broadband communication network allows to use broadband channels only on-demand. These three major functions enables by SMACS, CAHN, and SecMIP are explained in further detail in the following sections.

5.2.1 Integration of Infrastructure-less Connections

In Chapter 3, solutions have been discussed to seamlessly manage infrastructure-based communication technologies. When considering the interconnection of any two nodes, the optimal data path might be very heterogeneous. Depending on the available networks, the optimal end-to-end data path between the nodes can consist of infrastructure-based and ad-hoc links (see Chapter 4.4). Fig. 5.1 illustrates a scenario where one node is attached to a fixed infrastructure (e.g., DSL) and the other is moving abroad connecting to UMTS and Wireless LAN (e.g., Public Hotspot). Initially, node 1 and node 2 are communicating using the cellular and the DSL link, respectively (step 1). Then the node 1 changes its point of attachment to WLAN and sends a binding update to the node 2 (step 2). If node 2 moves towards the same WLAN access point, the session path can be optimized (step 3). If both nodes come close enough to each other, the session is switched to direct node-to-node communication (step 4). During the whole session the SMACS layer is querying the underlying SecMIP and CAHN components about available infrastructure-based and infrastructure-less networks, respectively. If the peer is reachable through any infrastructure-less link, CAHN is establishing the link and notifying the SMACS layer about the resulting settings.

The SMACS layer can then decide, whether further end-to-end session protection has to be deployed. In the case where the nodes are not located within

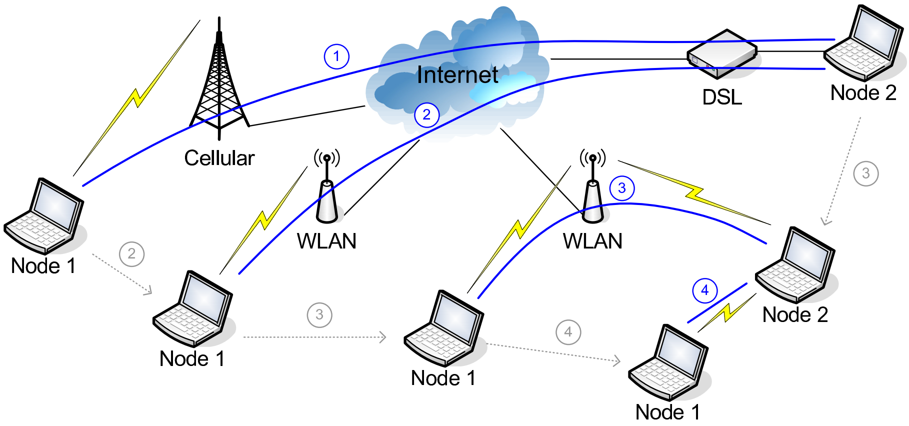


Figure 5.1: Handover between Infrastructure-based and Infrastructure-less Communication

the range of direct communication, the SecMIP module is connecting to the best available access network and reports the acquired (mobile) IP address to the SMACS layer, which is taking care of the secured end-to-end session establishment. Therefore, further parameters can be negotiated between the nodes. The SMACS signaling channel is offering means to securely exchange sensitive data (i.e. keying material) required to establish an end-to-end IPsec session.

The different steps involved in a handover process between infrastructure-based and infrastructure-less technologies is shown in the Fig. 5.2 and 5.3.

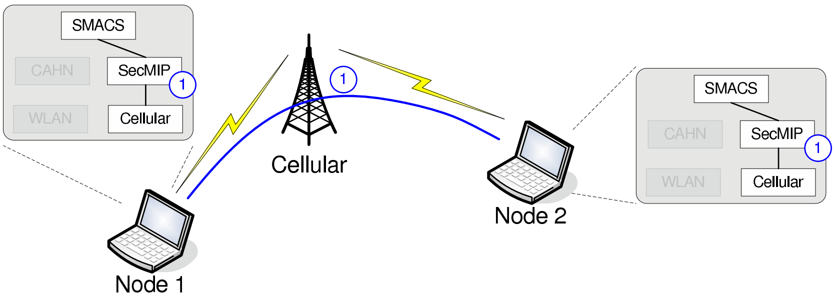


Figure 5.2: Infrastructure-based Session

First, the nodes are connected through an infrastructure-based communication technology (e.g. cellular). The connection is provided by the SecMIP module, which is controlling the cellular network interface. Node 1 and node 2 can exchange further information about available infrastructure-less capabilities using the signaling messages provided by the CAHN protocol (see 6.3). In this case, both nodes would learn about the WLAN interface of the peer. If now the node 1 is approaching node 2 (step 2), such that the CAHN module is detecting the WLAN interface of the peer, the negotiation required for the establishment of the direct link can be done through the cellular network (step 3). After having established the direct link between the two WLAN interfaces, SMACS can trigger SecMIP (i.e. Mobile IP Route Optimization) to switch the ongoing session from the cellular to the WLAN interface (step 4).

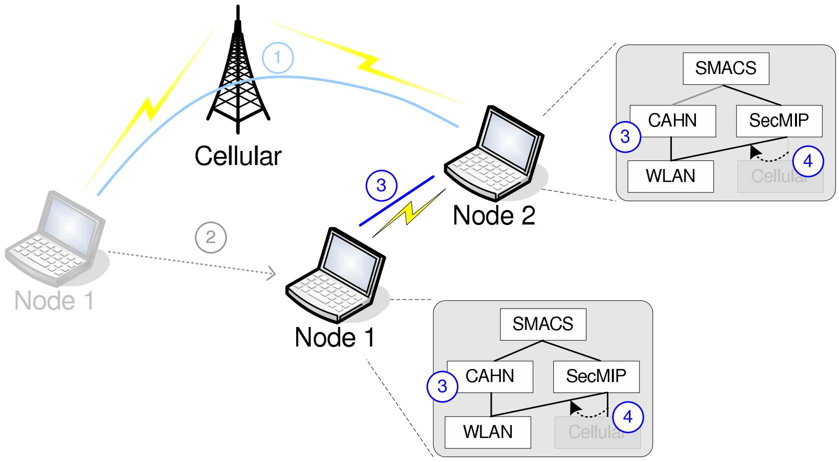
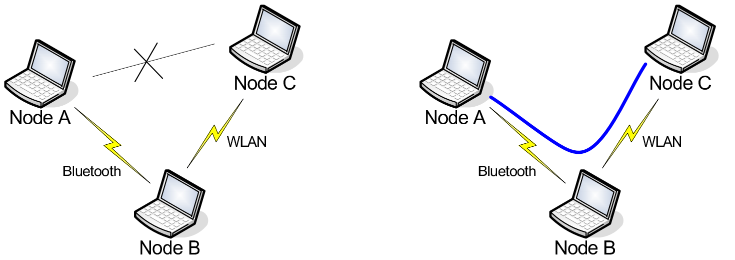


Figure 5.3: Infrastructure-less Based Session

Thinking about sessions among more than two nodes, the SMACS architecture can further support the connection establishment, especially if the interconnection of all nodes require several links to be set up. Fig. 5.4(a) illustrates three nodes with different communication technologies. Node A and B are Bluetooth enabled, and node B and C are WLAN capable. Since node A and C do not share a common communication technology, they can not directly communicate with each other. Hence, node B can be configured to act as a *technology bridge*, relaying the IP packets between node A and C, which is shown in Fig. 5.4(b).



(a) Node A and Node C can not communicate

(b) Node B as Technology Bridge

Figure 5.4: Heterogeneous Spontaneous Networking

The fact that all SMACS enabled nodes are reachable for signaling messages all the time allows to securely set up multi-hop sessions. Intermediate nodes can be used to gather information about neighboring nodes to learn about the network topology. The ability to rely on a signaling plane, which is always available and offering simple location management, address resolution, authentication, and billing mechanisms, might help to solve a lot of problems of end-to-end session management.

5.2.2 Enabling Heterogeneous Network Resource Management

When deploying IP mobility based on Mobile IP, the handover decisions are taken on the mobile node. To ensure that all available access networks are detected, the mobile node permanently scans on all communication devices. Whenever a new access network is detected the mobile node tries to get connected and, if successful, acquires a new care of address like explained in Section 2.5.2. Depending on the local priority list, the care of address of the chosen access network is registered with the home agent. In terms of power management this mode of operation is not optimal at all. To allow scanning for available access networks, all communication interfaces of the mobile node have to be powered up permanently. This might be very energy wasting, for example in the case of a WLAN interface, when thinking at rural environments, where nowadays only wide area networks like GPRS or UMTS are available, but no WLAN. To avoid such unthrifty scanning, several approaches have been proposed in the literature (see Section 3.3.3). To the best of our knowledge, all research effort is addressing resource management for infrastructure-based access networks only. The ability to seamlessly switch to high bandwidth direct communication considerably increases the average throughput and hence shortens the time required to transfer a certain volume of data. Enabling the automatic use of energy efficient short range communication channel increases the overall battery lifetime.

The separated treatment of signaling and data plane proposed in Chapter 4 is enabling another dimension of resource management. The ability to use low bandwidth and therefore low power signaling channels to set up power consuming high bandwidth data channels only if required is enabling the possibility to be always reachable without wasting valuable network resources. Section 5.2.3 is further addressing this concept offering *broadband on-demand* capability. The low bandwidth signaling channel can additionally be used to exchange context information to increase the accuracy of resource management. Depending on application requirements or user preferences, the network selection might be different. However, when analyzing scenarios where two or more mobile nodes are interconnected via different access technologies, the choice of the appropriate communication technologies becomes more difficult. For example, in the case where two communicating mobile nodes are connected through two different access technologies (like UMTS and GPRS), the one being connected to the lower capacity access network (i.e., GPRS) limits the maximum transfer speed. Hence, allocation of expensive resources for the peering node does not increase the connection performance, but may result in waste of network resources. If nodes would be informed about the limited connection capacity of the peering nodes, they could adapt their connection accordingly. It the case mentioned above, the node attached to UMTS could downgrade its connection to GPRS and thus safe valuable UMTS resources. Hence, when focusing on node-to-node communication in heterogeneous networks, the end-to-end session consideration becomes crucial to implement optimized resource management.

Finally, the handover decision has to consider best the interests of the operators and the users, which is not always easy, especially when taking into account that most direct links are free of charge. From a user perspective it might be preferable to use these direct links whenever possible if session duration, power consumption, and connection costs can be decreased. One of the motivations for the operators might be the ability to seamlessly take over the session whenever direct communication is not possible. Furthermore, if the signaling of the direct links is enabled by the network operator, resource management can be highly influenced. Knowing the allocation of channels in the unlicensed band like the 2.4 GHz used by WLAN at commercially operated hotspots, the direct links can be set up on adjacent channels to avoid interference. Both, the operator and the users utilizing the direct connection could profit from better performance.

In contrast to conventional homogeneous access networks, the resource management in heterogeneous access networks has to be much more comprehensive. Depending on the location of the users, the communication capabilities or even battery level of the actually used device, the most appropriate access technology may vary. The almost unlimited number of factors influencing the appropriate selection of networking resource make the domain of heterogeneous network resource management extremely complex and by far to big to be evaluated thoroughly within this thesis. The proposed system was designed to enable heterogeneous network resource management by offering an appropriate signaling framework. The actual resource negotiations and management processes are clearly out of scope. However, a simulator was built to identify and quantify the possible gain that both, the end user and the network operator could get, if the heterogeneity is intelligently and seamlessly managed with the help of the presented framework. Therefore, ideal and optimal resource management is assumed. More details about the simulator and the results obtained in terms of heterogeneous connectivity, battery, and network resource savings, are presented in Section 5.3.

5.2.3 Broadband on Demand

With existing mobile devices the battery lifetime is still very limited, especially when using energy demanding broadband communication technologies[2]. Therefore, it is still not practicable to stay connected all the time. Nevertheless, there is a big desire to be always reachable for any type of data. Introducing the separation of signaling and data plane for heterogeneous sessions with SMACS and CAHN, it is possible to offer the same reachability as available for mobile voice services. The ability to send a communication request to a mobile node without requiring energy demanding broadband connectivity (infrastructure and ad-hoc and direct node-to-node based) results in considerable energy savings. Depending on the number and duration of the sessions, using a narrow band and low power channel to handle connection requests can substantially increase the battery lifetime and save valuable network resources. Especially for GPRS or UMTS where a certain level of network resources are allocated for attached nodes even if they do not transmit or receive any data[3], the *broadband on-demand* approach can significantly improve the resource management. Therefore, also extreme scenarios have to be considered, where nodes exchanging only occasionally bursts of data and switch to sleep mode between the broadband sessions. This scenario might be representative for applications involving

[2] Tests [177] showed that broadband communication can take up to 28% of the total energy consumption of a standard laptop

[3] GPRS assigns at least one timeslot for each connected node. UMTS nodes reserve a CDMA code when attached.

machine-to-machine communication or sensor networks (i.e. gateways between sensor networks and backbones). Scenarios requiring bursty broadband communication might very much profit from the low power signaling of broadband sessions.

5.3 SMACS Simulation

To study the potential of the idea of *Smart Multi-Access Communications* in terms of throughput, power and resource management, simulations have been conducted. In all simulated scenarios the integration of ad-hoc links, whenever possible and the ability to switch broadband network interfaces on, only if a data session is active, were of special interest. For sake of clarity we will call the first feature *ad-hoc mode* and the second *on-demand mode*. The main motivation to perform simulations was to estimate the relative benefit in terms of throughput, power and resource management. To estimate the relative benefit of the ad-hoc and on-demand mode as such, without being limited to existing technologies, several scenarios were to be simulated without knowing anything about the lower layers. We were interested to see the impact of the ad-hoc and on-demand mode dependent on the relative bandwidth provided by infrastructure-based and ad-hoc links. Therefore, we defined three infrastructure-based links and one ad-hoc purely based on their bandwidth relative to the slowest link. A scenario referred to as 1:10:100:1000 considers the second infrastructure-based link to be ten times and the third hundred times faster than the first one. The forth number represents the infrastructure-less link, which in this particular scenario delivers thousand times higher bandwidth. This abstraction allowed us to observe different possible scenarios for the evolution of both networking paradigms and get a rough idea about the potential of our two modes, independent of actual technologies. The further estimate the potential of the ad-hoc and on-demand mode in terms of resource management, we limited the available capacity of the infrastructure-based links.

5.3.1 Requirements on the Simulation Tool

To simulate the envisioned scenarios the used simulation tool should support the definition of abstract communication technologies, purely based on the provided bandwidth. Communicating nodes should permanently consider all available links and always handover to the link providing the highest bandwidth. To estimate the potential of the ad-hoc mode the switching of ongoing sessions between infrastructure-based and ad-hoc links should also be supported. A thorough analysis of existing simulation tools [177] ended up with the conclusion that heterogeneous networks are not yet supported. Existing network simulators (ns2 [184], Qualnet [170], OpNet [145], etc.) do not provide appropriate support for the simulation of heterogeneous networks with dynamic vertical handovers during runtime, end-to-end communication between nodes using different wireless technologies simultaneously, and switching between infrastructure and ad-hoc mode of operation. Furthermore, these simulators either do not yet implement certain wireless technologies, e.g., GPRS in Qualnet, or implement different technologies for different incompatible versions, e.g., UMTS for ns-2.26 and GPRS for ns-2b7a, which is basically making the simulation of heterogeneous networks impossible. All mentioned simulation tools are packet based, which is not ideal for our purpose because this does not allow the definition of abstracted communication links. Flow based simulations are very much appropriate for technologies like GPRS and UMTS, where the maximum available bandwidth is assigned per node and time and not packet based. Since the simulations should help to estimate the overall potential of the ad-hoc and on-demand mode proposed in this thesis, the exact definition of the lower layers is not fundamental. Modeling the bandwidth assignment and the maximum capacity conservatively is soundly reasonable for first estimations.

5.3.2 The Heterogeneous Network Simulator HNS

To meet best our requirements and because of the severe limitations of existing simulation tools when simulating heterogeneous communication sessions, a new simulator, called Heterogeneous Network Simulator (HNS) has been developed. Heterogeneous sessions are modeled as a sequence of homogeneous sub-sessions. To decompose a heterogeneous session into its homogeneous sub-sessions, the simulation is performed in two steps. First, the complete simulation scenario is pre-processed to calculate the different events like mobility changes, handover, session start and termination. These resulting events are listed in a log file and serves to extract the sub-sessions, which are determined by the handover events occurring between the actual session start and termination. Second, each sub-sessions is emulated as a flow separately, considering the corresponding bandwidth assignment model. Finally, the heterogeneous sessions are evaluated by merging the modeled the corresponding sub-sessions. The bandwidth assignment for the different communication technologies can either be set fixed or modeled. Modeling the technologies has been soundly compared to simulating, in terms of implementation effort and impact on the overall simulation results. The fact that the main interest of the SMACS simulation is directed towards the relative improvement potential introduced by the developed concepts of heterogeneous networking, strongly favored the modeling of the data session. Applying the extensive simulation of the different technologies would undoubtedly result in more accurate values for the sub-sessions, but probably not change the relative improvement on the heterogeneous sessions effectuated by SMACS. However, the delegation of the simulation of the sub-sessions might be of interest for future evaluations of heterogeneous networking. Thus, the HNS was designed to easily interact with state of the art simulators addressing homogeneous networking.

Modeling of Communication Technologies

The modeling does not account for any physical propagation medium properties or MAC layer functionality and simulates sessions between peer mobile nodes at the application level, i.e., no packet transmissions are simulated. The models define the available bandwidth for each sub-session depending on the number nodes attached to the same access point or base station and its maximum capacity. The amount of data transmitted is derived from the time attached to a certain technology and its bandwidth.

UMTS To model the bandwidth assignment for a UMTS node, we considered the up- and downlink channel separately. The uplink bandwidth is statically set to $64\,kbit/s$ per node, independently of the assigned downlink bandwidth. The overall capacity m offered by the base station is assigned to the attached nodes n until there is no capacity left to assign.

$$\text{bandwidth}_{up} = \begin{cases} 64\,kbit/s, & \frac{m}{n} \geq 64 \\ 0\,kbit/s, & otherwise \end{cases}$$

The downlink bandwidth is equally distributed to the attached nodes. The UMTS provides different bandwidth rates, namely $384\,kbit/s$, $128\,kbit/s$, and $64\,kbit/s$. If the n is the number of nodes and m is the maximum capacity offered by the base station, the assigned bandwidth is modeled as follows:

$$\text{bandwidth}_{down} = \begin{cases} 384\,kbit/s, & \frac{m}{n} \geq 384 \\ 128\,kbit/s, & \frac{m}{n} \geq 128 \\ 64\,kbit/s, & \frac{m}{n} \geq 64 \\ 0, & otherwise \end{cases}$$

GPRS is modeled based on TDMA slots. The coding scheme (CS) is statically set to CS4, which is providing $21.4\,kbit/s$ per slot. In our model we assume that CS4 can be used for up- and downlink independent of the distance between the node and the base station. We further assume class 10 devices, allowing 4 downlink and 2 uplink slots maximum. The number of assigned slots is depending on the availability of slots. The network tries to assign the maximum number of slots supported but dynamically adapt the assignment to guarantee equal distribution of the available resources. For the uplink this results in following slot assignment model, where n is again the number of nodes and m the maximum capacity offered by the base station.

$$\text{bandwidth}_{up} = \begin{cases} 42.8\,kbit/s, & \frac{m}{n} \geq 42.8 \\ 21.4\,kbit/s, & \frac{m}{n} \geq 21.4 \\ 0\,kbit/s, & otherwise \end{cases}$$

The downlink slot assignment is modeled similarly, but allowing up to 4 slot per node.

$$\text{bandwidth}_{down} = \begin{cases} 85.6\,kbit/s, & \frac{m}{n} \geq 85.6 \\ 42.8\,kbit/s, & \frac{m}{n} \geq 42.8 \\ 21.4\,kbit/s, & \frac{m}{n} \geq 21.4 \\ 0, & otherwise \end{cases}$$

WLAN The medium access mechanism of WLAN is aiming at the provisioning of equal bandwidth for all attached nodes. Unlike in GPRS or UMTS, the up- and downlink are not treated separately. Sending and receiving nodes are competing for the same medium. To simplify the modeling of WLAN we assume that no collisions occur and the maximum available capacity can equally assigned to the nodes without loss because of collision recovery mechanisms (e.g. backoff). The model used for resource assignment for WLAN nodes is consequently as follows, where n is the number of nodes competing for the medium and m the overall capacity of the medium:

$$\text{bandwidth}_{up/down} = \frac{m}{n}, kbit/s$$

Coverage The transmission ranges for the different wireless technologies are modeled as circles with varying radiuses with respect to their characteristics, e.g., small radius for local area technologies such as WLAN and UWB and larger radius for wide area technologies such as GPRS and UMTS. The signal strength is modeled very simple based on the distance between the source and destination and only used for the handover decision.

Mobility Models, Network Topology, and Handover Decision

The simulator implements the standard random waypoint mobility model and also the reference point group mobility model as introduced in Section 2.5.1. In the former model, nodes move independently of each other such that the period when two nodes are within transmission range and can communicate directly in ad-hoc mode is unrealistically short. On the other hand, the latter model allows the simulations of scenarios in which nodes move as a group such as on a train or a unit in a battlefield where the ad-hoc mode of SMACS is most beneficial.

All simulation parameters are specified in a configuration file, which is read by the HNS at the beginning of the simulation process. Mobility behavior, network topology, and session patterns can be freely defined. For our simulations the mobility pattern implementations provided by the BonnMotion [48] have been used. The BonnMotion library generates NS-2 compliant mobility scripts, which are then imported by the HNS. To describe the network topology, the location of the base stations and access points has to be indicated. To get results on the potential benefit of using ad-hoc links and offering broadband on-demand capabilities, these features can be enabled and disabled for the simulation. Three different basic handover decision algorithms have been implemented within the simulator. The network selection can be done either according to the best signal, the lowest cost, or the best available bandwidth. Since the simulations done within the scope of this thesis are addressing the influence of ad-hoc links and energy aware low power signaling over the cellular network, the handover decision algorithms are not optimized for the overall network resource management with perspective to the operator. Nevertheless, the structure of the simulator does not limit future extension to support more sophisticated algorithms taking further context information like traffic pattern and application requirements into account.

Implementation of the HNS

The Heterogeneous Network Simulator is programmed in Java. It is event-driven and flow oriented. Unlike packet oriented simulators, the HNS processes whole data sessions (i.e. flows), which are split into sub-sessions, whenever the network environment changes (e.g., number of nodes attached to the same base station or access point, handovers occur). All simulation parameters can be defined trough a configuration file. The scenario can be also be defined using the graphical user interface (GUI). Upon simulation is started, an event scheduler processes the events in the queue and synchronizes with the GUI to display movements and sessions in real time. For the evaluation, log files are generated at runtime. Fig. 5.5 summarizes the main simulation and evaluation procedures.

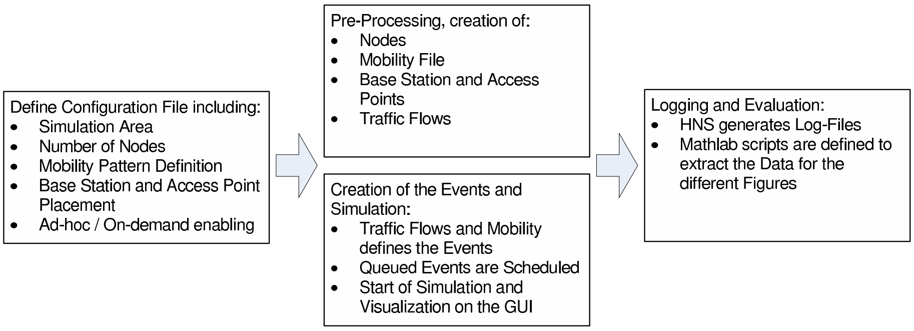

Define Configuration File including:
• Simulation Area
• Number of Nodes
• Mobility Pattern Definition
• Base Station and Access Point Placement
• Ad-hoc / On-demand enabling
Pre-Processing, creation of:
• Nodes
• Mobility File
• Base Station and Access Points
• Traffic Flows
Creation of the Events and Simulation:
• Traffic Flows and Mobility defines the Events
• Queued Events are Scheduled
• Start of Simulation and Visualization on the GUI
Logging and Evaluation:
• HNS generates Log-Files
• Mathlab scripts are defined to extract the Data for the different Figures

Figure 5.5: HNS procedures with pre- and post-processes

An overview of the package organization of the HNS is shown in Fig. 5.6. The main functionality of the different packages is explained in the figure.

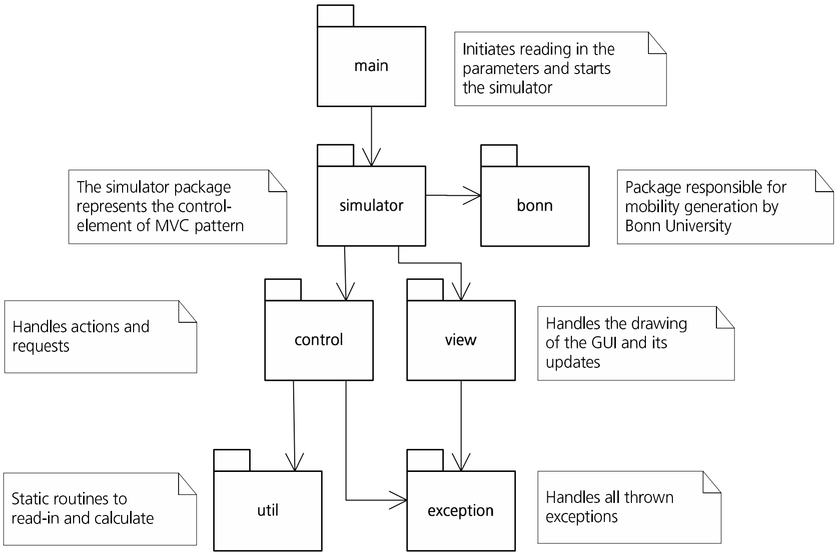

main
Initiates reading in the parameters and starts the simulator
The simulator package represents the control-element of MVC pattern
simulator
bonn
Package responsible for mobility generation by Bonn University
Handles actions and requests
control
view
Handles the drawing of the GUI and its updates
Static routines to read-in and calculate
util
exception
Handles all thrown exceptions

Figure 5.6: HNS package organization

Graphical User Interface

The GUI reflect the simulation in real time. It shows the base stations and access points, which are indicated by technology, location and range. A map can be loaded as background image to visualize the geographical topology. Ongoing communication sessions are indicated by straight dashed lines between the nodes and also sending and receiving activities are represented in the GUI. Handover points are shown as red dots. A screenshot of the GUI is shown in Fig. 5.7. For larger or script based simulations it is also the possibility to call a batch job on a folder structure which then processes every configuration file in the folder and storing all results without any graphical output.

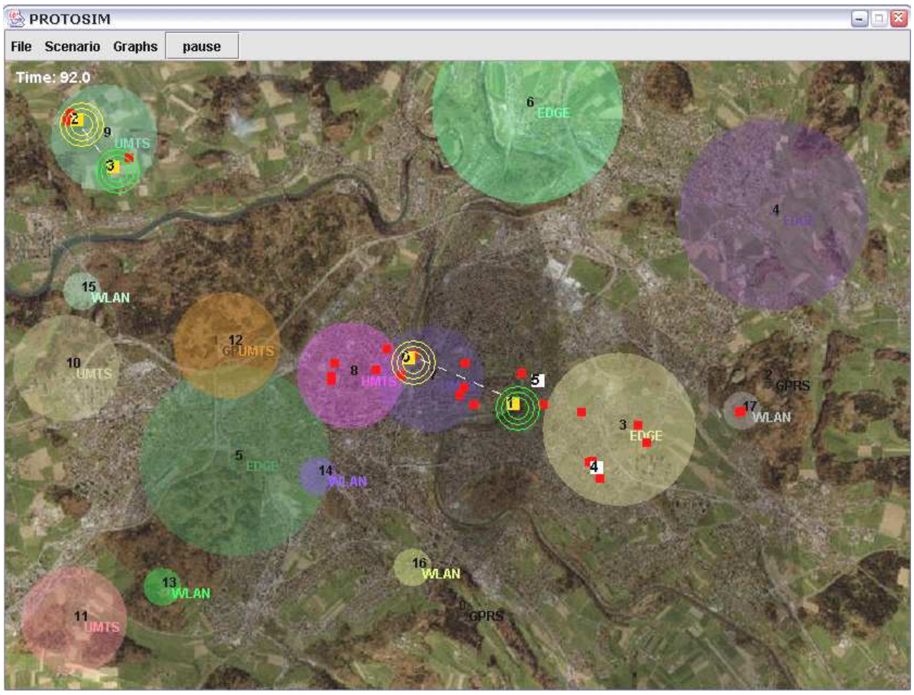


Figure 5.7: HNS Graphical User Interface

5.3.3 Simulation Scenarios

The conducted simulations were separated into two simulation sets. The first set focused on a rather generic view on heterogeneous networking, defining technology characteristics independent of existing technologies and deployments. The aim of these simulation series was to identify the potential of the ad-hoc and the on-demand feature in any kind of networking environment with different simulation areas, node and session densities, bandwidths, and coverage of infrastructure technologies. We mainly focused on the user relevant aspects like achieved throughput and energy consumption. In a second simulation set we considered actually available technologies and deployments. Infrastructure-based communication technologies like GPRS, UMTS, and WLAN have been defined with realistic bandwidth and capacity limitations. This allowed us to evaluate the impact of the ad-hoc and the on-demand feature on network operator relevant aspects as well. The simulations revealed how our concept is influencing the overall performance of the heterogeneous network.

Performance Evaluation in Generic Network Environments

In our first simulation set, 50 nodes move over a given simulation area according to either the random waypoint (RWP) or the reference group point (RPGM) mobility model. An overview of these models can be found in 2.5.1. Multiple random sessions are established between pairs of nodes where the session arrival rate is Poisson distributed and the amount of data to be transferred during a session is Pareto distributed. Each node always uses the available wireless technology with the highest bandwidth, i.e., a vertical handover occurs whenever a nodes moves into the range of a technology with a higher bandwidth. Consequently, the effective session transfer rate is the minimum bandwidth of the technologies, currently used by the two communicating nodes. Three different wireless infrastructure-based technologies are deployed over the simulation area that differ in their bandwidth, range, and coverage to model existing or possible future technologies such as GPRS, UMTS, WLAN. Furthermore, there is a infrastructure-less wireless technology that allows for direct node-to-node communication such as Bluetooth, WLAN or UWB.

We devised four simulation scenarios by varying the node density, the number of sessions, the ratio of the bandwidth among the available wireless technologies, and the ratio of their coverage. If not noted otherwise, the other simulation parameters are kept fixed and set to the values as follows. The simulations last for 4600 seconds and sessions between nodes are only established after an initial warm-up phase of the mobility model of 1000 seconds to reach a stable state, i.e., traffic is generated during exactly one hour of simulation time. The simulation area is $3000\,m$ x $3000\,m$. In the random waypoint mobility model and the reference point group mobility model, nodes move with a speed between 1 and $15\,m/s$ and have a pause time of $30\,s$. The average group size is set to 4 with a standard deviation of 3 and a maximal distance to the group center of $50\,m$ in the group mobility model. Furthermore, nodes have a group change probability of 0.3. The session arrival rate is Poisson distributed with 4 sessions per hour and source-destination pair, which yields 100 sessions for 50 nodes. The amount of data is Pareto distributed between 10 KB and 100 MB. Table 5.1 summarizes the simulation parameters used in the first simulation set.

The bandwidth ratio for the three infrastructure-based technologies are set to 1 : 10 : 100 where the coverage is 100%, 50%, and 5% of the total simulation area, respectively. Considering today's deployed technology such as GPRS, UMTS, and WLAN, we believe that these values provide a reasonable rough approximation. The base stations are deployed randomly all over the simulation area. The number of base stations and the transmission radii for the respective technologies are varied accordingly to obtain these coverage values. Considering currently available technologies for node-to-node communication such as 802.11g or UWB, we can reasonably assume that node-to-node communication is 10 times faster than the fastest available infrastructure-based wireless technology. The transmission range for the ad-hoc technology was set to $150\,m$. We simulated these four scenarios for the four cases when nodes have each of the

Simulation Time (s)	4600
Warm-up Phase (s)	1000
Simulation Area (m)	3000 x 3000
RWP/RPGM Speed (m/s)	1 - 15
RWP/RPGM Pause Time (s)	30
RPGM Group Size	4
RPGM Group Size Standard Deviation	3
RPGM Group Change Probability	0.3
RPGM Max Distance to Group Center (m)	50
Session Arrival Rate (sessions/hour)	4
Size of Data transferred per Session	10 KB - 100 MB

Table 5.1: Summary of the Simulation Parameters

two features of SMACS enabled/disabled, i.e., neither ad-hoc nor on-demand mode enabled, have either ad-hoc or on-demand enabled, and have both modes enabled. We measured the average session duration and estimated the energy consumption to quantify the possible benefits of the ad-hoc and on-demand mode, respectively. All simulation results are given with a 95% confidence interval.

To estimate the impact of the ad-hoc and on-demand feature on the battery life, we based the energy consumption for the three infrastructure-based and the ad-hoc links on average values found for nowadays technologies. As a reference we selected GPRS, UMTS, and WLAN. The values for WLAN have been also taken for the ad-hoc link. The values found in the literature for the energy consumption of the different devices are highly variable. The only consistent values were found for WLAN and thus, we tried best to estimate the average power consumption for the remaining types of devices, i.e. GPRS, UMTS [60, 122] relative to WLAN. Table 5.2 summarizes the factors taken for the energy consumption relative to WLAN in receiving mode.

	Receiving	Sending	Idle	Sleep
WLAN	1	2	1	0.05
GPRS	3	4	1	0.05
UMTS	2	3	1	0.05

Table 5.2: Used Relative Energy Consumption Values

The values found were not used for estimation of the absolute impact of our concept on the power consumption, but for the relative improvement potential. The values served as a basis for a rough idea about the relative power consumption evolution between different infrastructure-based and ad-hoc communication technologies.

Varying Node Density

In the first scenario, we evaluate the impact of the node density on the performance by varying the side length of the square of the simulation area from $1000\,m$ to $10000\,m$. For larger simulation areas, the probability that two peers can communicate directly in ad-hoc mode is smaller than when the nodes move in a smaller area and, thus, the benefit of the ad-hoc feature is reduced. This behavior is reflected in Fig. 5.8. Since the throughput is identical whether the on-demand feature is enabled or not, we did not considered the on-demand feature for the throughput evaluations.

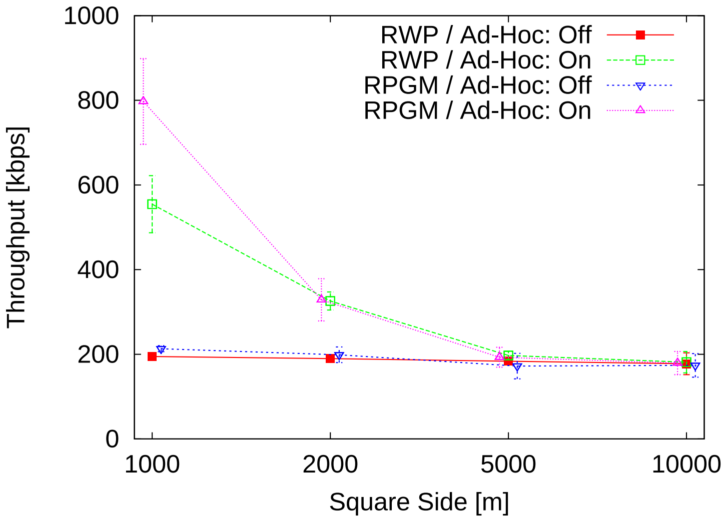


Figure 5.8: Throughput for Varying Node Density

For small areas the RPGM results in high probability that peering nodes come close enough to benefit from the high data rate ad-hoc link. For the smallest simulation area this results in average throughput increase from 200 *kbps* up to 800 *kbps*. If the ad-hoc feature is disabled, the throughput is quite constant for all simulation areas, which is expected since the relative coverage for the different technologies is constant independent of the size of the area. The average consumed energy per node for the different simulation areas is depicted in Fig. 5.9.

In terms of energy consumption approximately 20% can be saved if the on-demand feature is enabled. Another 20% can be saved if the system can switch to ad-hoc links. This is mainly due to the increased throughput, which in turn results in shorter session durations. (For the energy consumption evaluation, the RPGM and the RWP were not differing that much and therefore the results for the RPGM are not shown explicitly for sake of clarity.)

Varying Session Density

If the session density is very high the nodes are constantly transmitting and/or receiving data anyway such that the on-demand feature of SMACS is not really beneficial. However when nodes receive data only very infrequently, SMACS enables nodes to be in sleep mode and be nevertheless reachable for session invitations. Unlike for SMACS-enabled nodes, ”normal” nodes have to remain in idle mode to be reachable for incoming data all the time. In today’s devices however, the energy consumption in idle mode is significantly higher than in

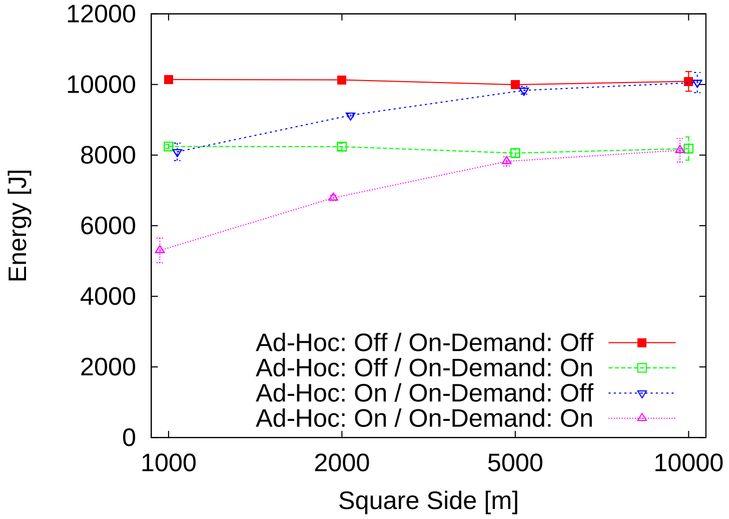


Figure 5.9: Energy Consumption for Varying Node Density

sleep mode. This allows SMACS-enabled nodes to reduce significantly the use of scarce battery power. In these simulations, we varied the amount of transmitted data by the session arrival rate which is Poisson distributed with 1 and 40 session per hour and source destination pair. In Fig. 5.10, the impact of the session density on the average throughput of the sessions is depicted.

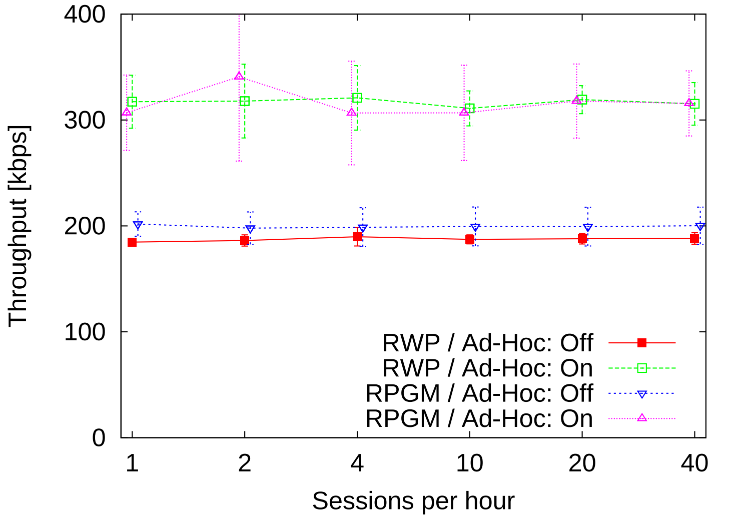


Figure 5.10: Throughput for Varying Session Density

Since here the simulation area was set to 3000 m x 3000 m, the improvement is consistent with the one depicted in Fig. 5.8. The difference between the RPGM and the RWP mobility model is not as big as expected, which is due to the small group size chosen for the RPGM. With bigger group sizes the probability that two communicating nodes are within the same group and hence able to use ad-hoc links is significantly higher. However, choosing the group size too big is not realistic neither.

Fig. 5.11 shows the energy consumption values for the different numbers of sessions. The potential energy savings strongly depend on the number of sessions, since the devices are only switched to sleep mode if no session are ongoing. Again, the increased average throughput with enabled ad-hoc mode is shortening the average session duration and thus the energy consumption.

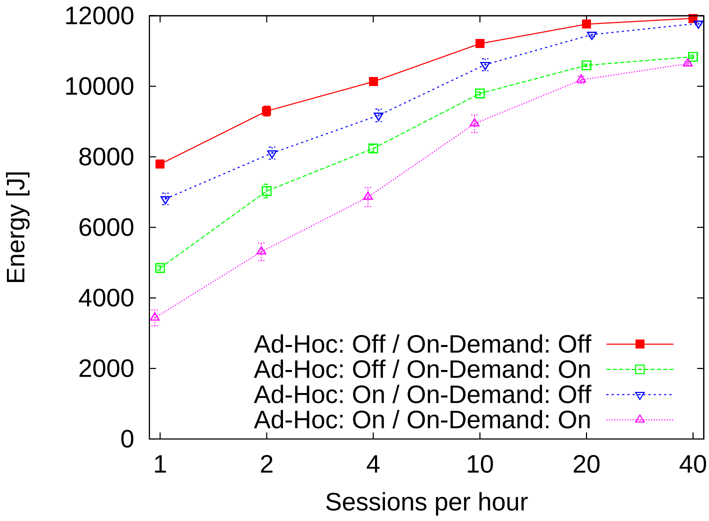


Figure 5.11: Energy Consumption for Varying Session Density

Varying Bandwidth Ratio

We basically distinguish between four kinds of wireless technologies in this chapter, three infrastructure-based and an infrastructure-less technology for node-to-node links. The first kind of technology provides almost full coverage but has only limited bandwidth such as GPRS, EDGE, or also satellite networks. The second kind of technology constitutes 3G wireless networks such as UMTS, HSDPA, which provide higher bandwidth, but are not yet as widely deployed as 2 and 2.5G networks, perhaps only within urban areas. Wireless broadband technologies are the third kind of infrastructure-based technology considered, which are commonly not area-wide deployed, but at specific locations only, such as 802.11b in so-called Hotspots. Fourth, nodes can communicate directly without any infrastructure in ad-hoc mode with certain technologies such as WLAN or UWB. Depending on the technologies in use, the current active users, the signal-to-noise ratio, and/or operator policies, etc., the ratio between these technologies may vary strongly. We evaluated two scenarios and set the bandwidth ratio of the technologies with respect to the first technology (e.g., GPRS) providing the highest coverage. In the first scenario the second technology provides twice the bandwidth of the first technology. The second (e.g., UMTS) 20 and the ad-hoc (e.g., UWB) 1000 times more than the first technology. The second scenario was simulated with a technology bandwidth ratio of 1 : 10 : 100 : 1000.

Fig. 5.12 shows the variation of the average throughput if the bandwidth ration of the different technologies is changed. The first scenario, having a very high discrepancy between the data rates offered by the infrastructure-based and the ad-hoc communication technology, is very much profiting from the ad-hoc feature. The throughput increase in the second scenario is only about 20% because of the similar bandwidth offered by the fastest infrastructure-based and the ad-hoc link.

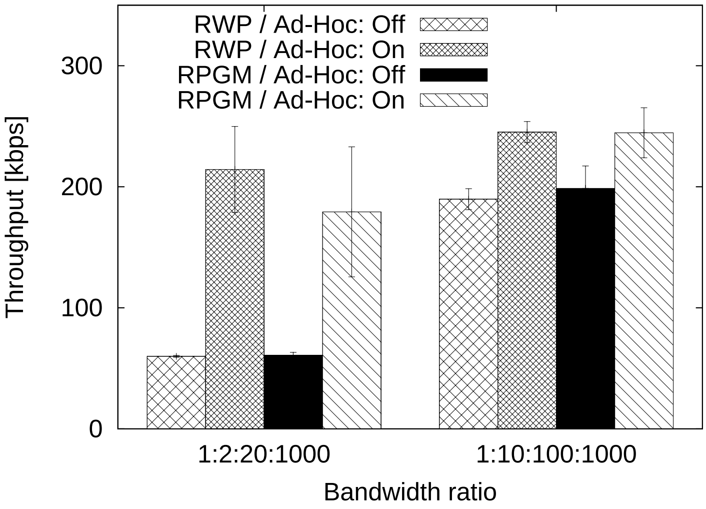


Figure 5.12: Throughput for Varying Bandwidth Ratio

Assuming further development of high bandwidth ad-hoc technologies like UWB offering data rates that are by orders of magnitude higher than the ones available at infrastructure-based networks, the capability of seamlessly switching to ad-hoc links becomes crucial to improve the average throughput. In our simulations the average throughput can be increased up to a factor of 4.

Varying Coverage of Infrastructure Technologies

In this last scenario, we analyze the impact of the coverage of the three different infrastructure-based technologies on the performance. We consider two specific cases where the coverage of each technology is very low and very high, respectively. In the first case, the coverage of the first, second, and third technology is $50\%, 25\%$, and 1% whereas in the second case the coverage was $100\%, 80\%$, and 10% of the whole simulation area, respectively. The impact of the variation of the relative coverage is depicted in Fig. 5.13.

For low coverage of infrastructure-based technologies the gain of the usage of the ad-hoc link is higher than for well covered areas. In the first scenario with low coverage the probability of having no or only very narrow band connection is rather high. Thus, even if the chance of having direct communication via the ad-hoc links is small as well, the impact on the session throughput is so much bigger in case of occurrence. When focusing on the energy saving potential the ad-hoc feature has nearly no effect compared to the on-demand capability for low coverage values as observed in Fig. 5.14.

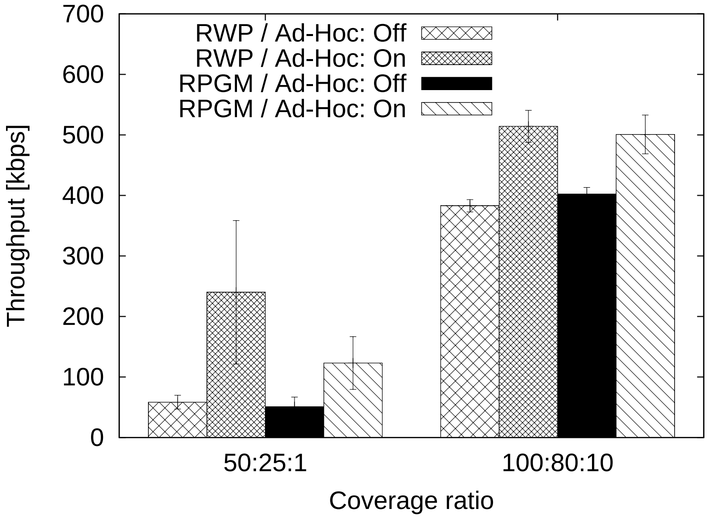


Figure 5.13: Throughput for Varying Coverage

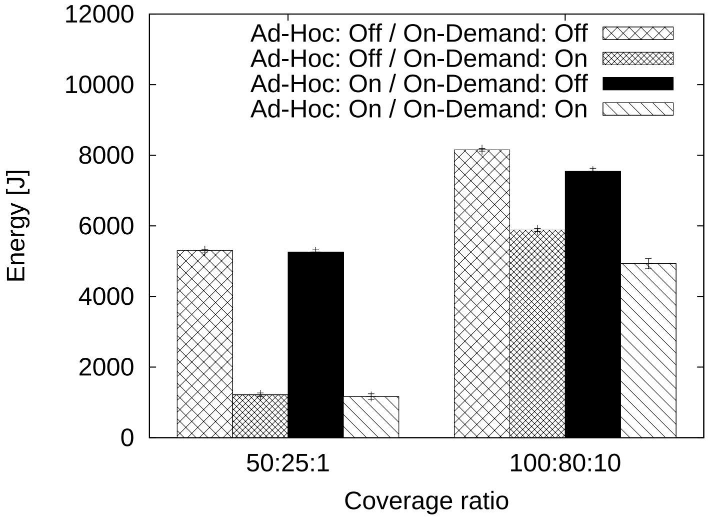


Figure 5.14: Energy Consumption for Varying Coverage

For high coverages values, the ad-hoc mode is more beneficial, but still not as advantageous as the on-demand feature.

Performance Evaluation based on Specific Network Models

The second simulation set was defined to further analyze the influence of SMACS on the network performance. The different access technologies were defined with limited capacity and adaptive data rates delivered to each node, depending on the actual load of the cell or access point. The data rates offered by WLAN were defined as explained in Section 5.3.2. The maximum capacity (m) per access point was set to $11\,Mbit/s$ as provided by the 802.11b standard. The overall capacity for GPRS nodes is adapted according to the number of nodes for each simulation. Assuming that network operators deploy enough bandwidth to serve all nodes with the minimal data rate of one TDMA slot for both, the up- and downlink, we define the overall capacity of GPRS as n up- and n downlink slots. These slots are equally distributed among the GPRS cells, resulting in blocked sessions if the nodes are not equally distributed among the cells. In all simulation scenarios GPRS is covering the whole simulation area. UMTS cells with coverage radius of $450\,m$ are supposed to offer a maximum capacity of $1024\,kbit/s$, which is equally distributed among the nodes according to the model defined in Section 5.3.2. The simulation area for the second simulation set was limited to $2000\,m$ x $2000\,m$. Two different coverages were defined for UMTS and WLAN, respectively. In the first simulation subset, we set the UMTS coverage to 80% and the WLAN coverage to 10%. In the second subset, UMTS was reduced to cover only 50% and WLAN 5% of the simulation area. Both scenarios were tested with different numbers of nodes, sessions and the two mobility models, RWP and RPGM. Similar to the first simulation set, the simulation time was set to 4600 seconds including a 1000 seconds warm-up phase for the mobility models. The parameters of the mobility models remain the same than shown in Table 5.1. The values that where adapted for the second simulation set are summarized in Table 5.3.

Simulation Area (m)	2000 x 2000
UMTS Coverage (%)	80/50
WLAN Coverage (%)	10/5

Table 5.3: Summary of the Simulation Parameters

The simulation results have been analyzed in terms of network load and efficiency, throughput, session block and drop rates, and transmission outage. The different terms are defined in the corresponding sections.

Network Load

The overall network load is calculated based on the load of each technology. Each load is weighted according to the coverage provided by that specific technology. This reflects the fact that the overall network load is mainly dependant on the load of technologies serving a large area. The network load is evaluated for different numbers of nodes and sessions. Both SMACS features were enabled or disabled to analyze the impact on the overall network load. Figure 5.15 shows the four resulting network loads in function of the number of nodes for 2 sessions per hour for the RWP mobility model. UMTS was set to 80% and WLAN to 10% coverage.

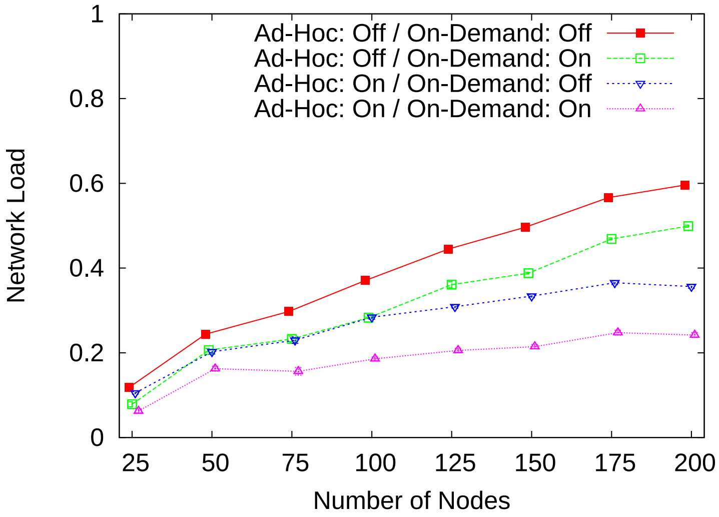


Figure 5.15: Network Load: RWP with 2 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

The ability to liberate network resources by switching ongoing sessions to infrastructure-less technologies whenever possible, reduces the network load by up to 24%. Thus, the network is able to serve more nodes with the same capacity if the ad-hoc feature is enabled. The on-demand feature is further increasing the number of nodes that can be served. With the on-demand feature, inactive nodes do not occupy network resources. The less sessions the nodes have, the higher the resource saving potential of this feature. With 2 sessions per hour about up to 12% of the network resources can be liberated due to the on-demand feature and eventually assigned to other nodes. With increasing number of sessions per hour, the benefit of the on-demand feature is considerably decreasing. The shorter the period between the data sessions, the less network resources can be liberated. Figure 5.16 shows that the benefit from the on-demand feature is nearly negligible if the amount of sessions is increased to 8. With 8 sessions per node the probability of resource shortages is very high. If some of these sessions are overlapping, they are competing for the same available resources. Waiting sessions are immediately occupying the liberated resources in a session can be offloaded to ad-hoc mode. Thus, the ad-hoc mode is not able anymore to discharge the network. The simulation results for the RPGM mobility model are shown in Figure 5.17 and 5.18 for both session rates 2 and 8.

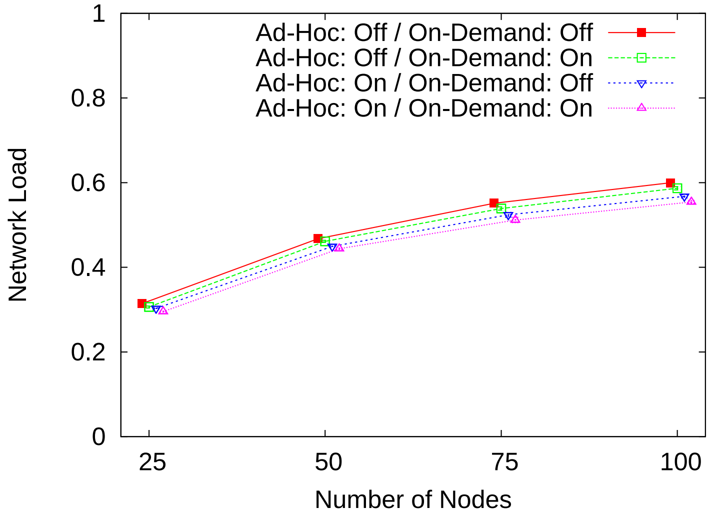


Figure 5.16: Network Load: RWP with 8 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

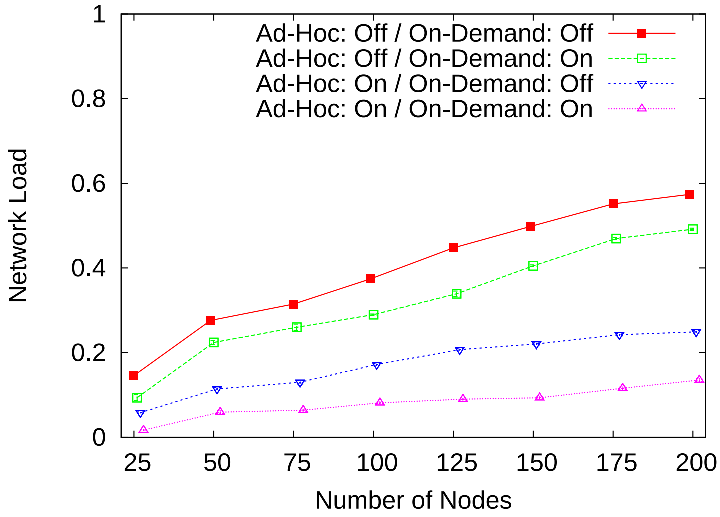


Figure 5.17: Network Load: RPGM with 2 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

The ad-hoc feature slightly benefits from the RPGM. If two communicating nodes are in the same group, the probability for them to stay close to each other is higher than with the RWP, although there is no relation between the selection of the group members and the selection of the session end-points. In the simulations there is hence not considered that group members are communicating among each other with a higher probability than they are with any other node without their group. Therefore, the ad-hoc feature would be even more beneficial if these community aspects would be considered when selecting session end-points. Even for 8 sessions per hour the beneficial impact of the RPGM is visible in Figure 5.18 The on-demand capability is almost independent on the mobility pattern, which is also reflected in the figure.

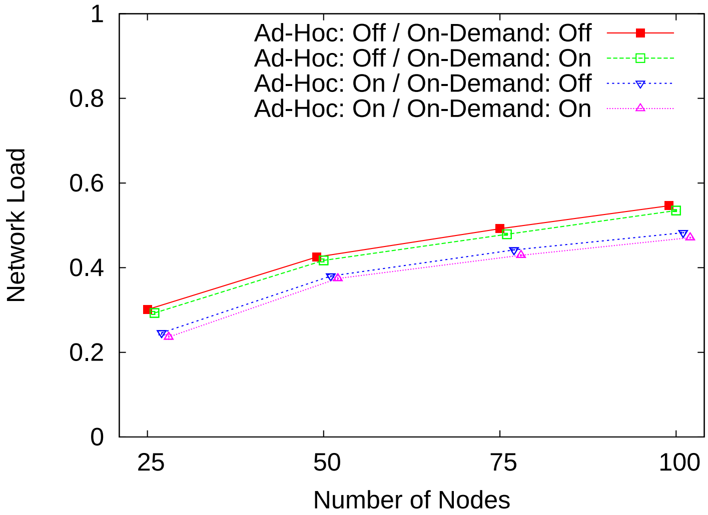


Figure 5.18: Network Load: RPGM with 8 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

To analyze the impact of the level of coverage of the different technologies on the network load, we repeated the four scenarios reducing the UMTS and WLAN coverage to 50% and 5%, respectively. Figure 5.19 and 5.20 are depicting the results with 2 sessions per hour for both mobility models. The reduction of the UMTS and WLAN coverages increases the network load in both cases, which is not surprising since we do not decrease the number of nodes or sessions. The impact of the on-demand feature on the overall network load is bigger compared to the high coverage scenario. The lower coverages increase the benefit of releasing unused resources. The figure for the network load is almost unchanged for RPGM compared to RWP if the nodes are sending 8 sessions per hour and therefore not explicitly shown. Waiting sessions are preventing the ad-hoc and on-demand mode to reduce the network load.

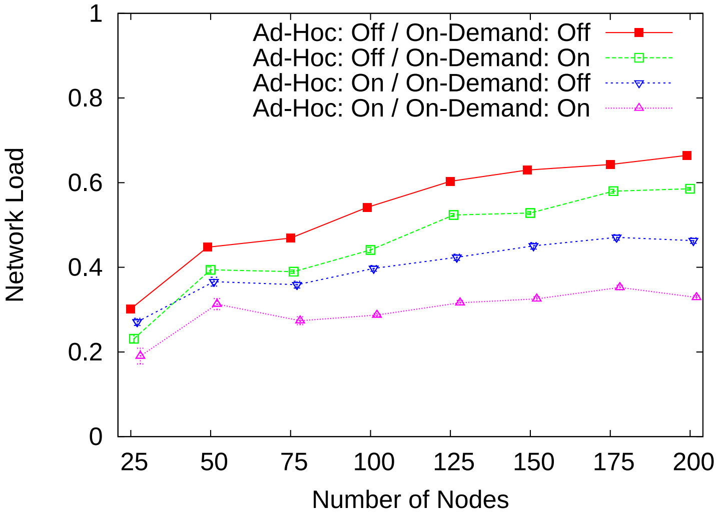


Figure 5.19: Network Load: RWP with 2 Sessions/h and 100% GPRS, 50% UMTS and 5% WLAN Coverage

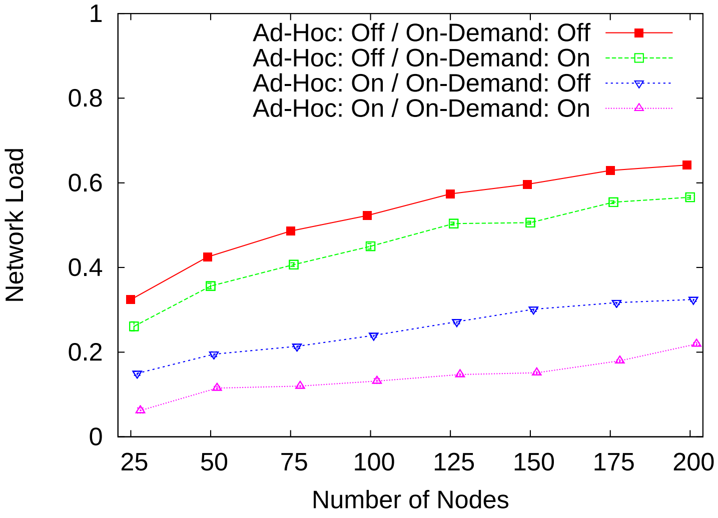


Figure 5.20: Network Load: RPGM with 2 Sessions/h and 100% GPRS, 50% UMTS and 5% WLAN Coverage

Network Efficiency

To further measure the influence of the ad-hoc mode we introduce a new metric called network efficiency. With regards to the ad-hoc feature, we define the network efficiency as the ratio between traffic sent using ad-hoc links and the overall traffic sent by the nodes. This indicates how much traffic the network could offload to the direct ad-hoc links. The bigger this ratio, the less operator resources are used to transfer the session data. This ratio was measured for both mobility models, 2 and 8 sessions per hour, and the two UMTS and WLAN coverages, respectively.

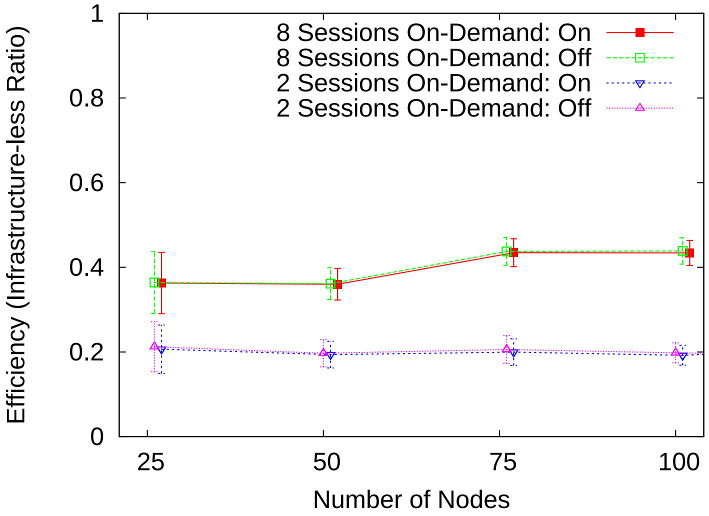


Figure 5.21: Network Efficiency: RWP with 2 and 8 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

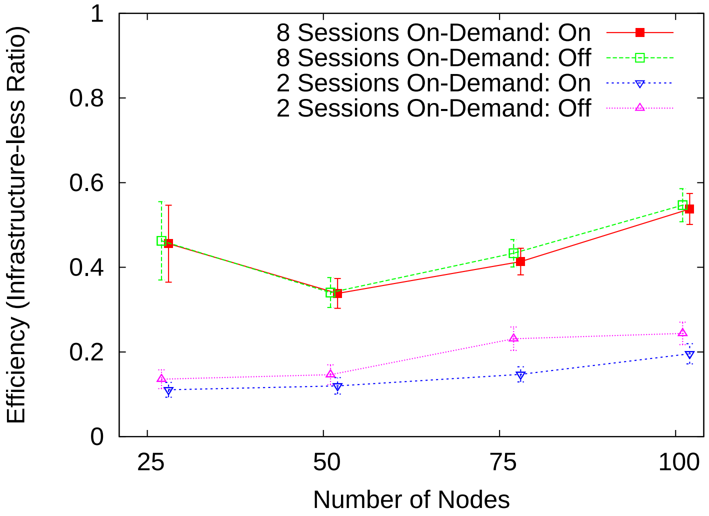


Figure 5.22: Network Efficiency: RPGM with 2 and 8 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

The figures reflect the fact, that the ratio of infrastructure-less connections is not dependant on the number of nodes. On the average, the selection of the mobility model is only slightly influencing the ratio. Similarly, the influence of the on-demand feature is almost negligible. For lower coverages and high number of sessions the efficiency is increasing up to 70%. This is mainly due to the high

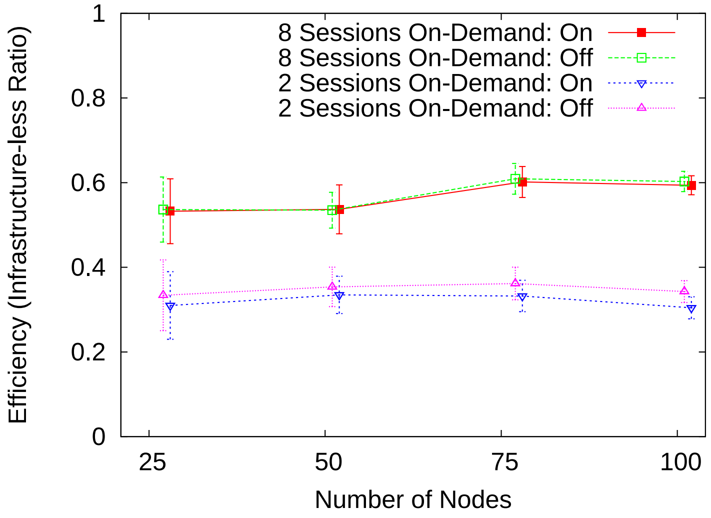


Figure 5.23: Network Efficiency: RWP with 2 and 8 Sessions/h and 100% GPRS, 50% UMTS and 5% WLAN Coverage

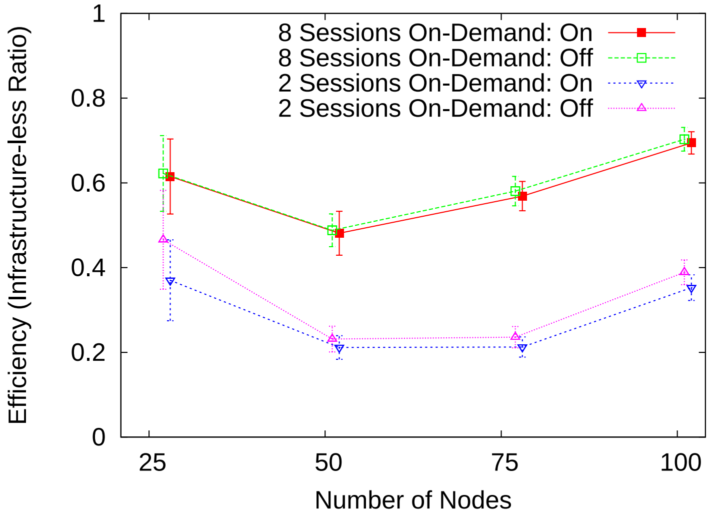


Figure 5.24: Network Efficiency: RPGM with 2 and 8 Sessions/h and 100% GPRS, 50% UMTS and 5% WLAN Coverage

network load, which decreases the average throughput of the infrastructure-based links and consequently increases the impact of the data transferred using the high data rates of the infrastructure-less connections.

Session Block Probability

Whenever a node is not able to start the session because of missing networking resources this is considered as a session block. The node is continuously trying to start the data transmission until either another node is releasing resources in the congested cell or the node is moving to another cell with available capacity. If the preferred technology is not available, the node is taking any other available networking technology, independent whether it is the most appropriate for that actual session. Thus, the network is overriding the node's preference if the overall network performance can be increased. Figure 5.25, 5.26, 5.27 and 5.28 depict the session block probability for the RWP and the RPGM mobility model for both, 2 and 8 sessions per hour.

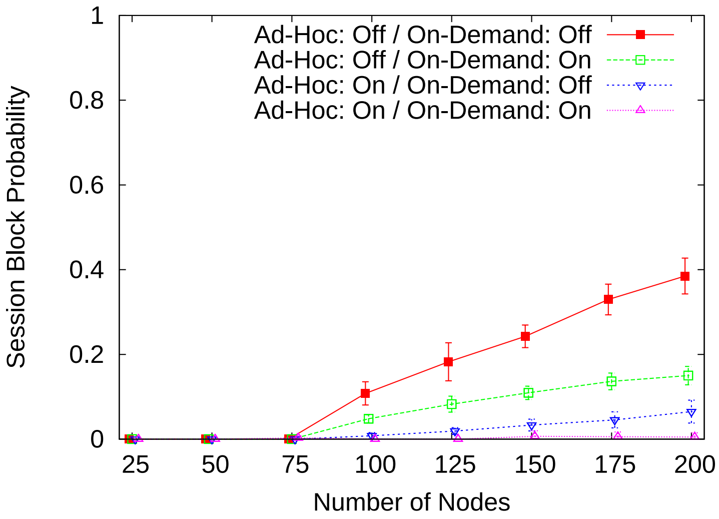


Figure 5.25: Session Block Rate: RWP with 2 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

For 2 sessions per hour, the network is able to serve approximately 75 nodes without blocking any resource request. The nodes are therefore distributed to all available networks. The network forces the nodes to switch to another technology if this increases the overall number of servable nodes. The ad-hoc mode has the biggest impact on the session block probability. Together with the on-demand feature, the blocking probability can be kept close to 0 for up to 150 nodes. With increasing session density, the influence of the on-demand mode on the session blocking probability is also considerably reduced, which can be seen in Figure 5.26.

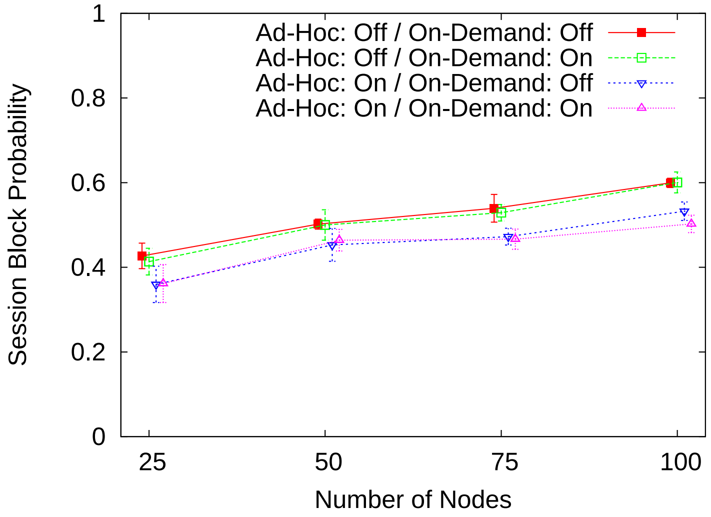


Figure 5.26: Session Block Rate: RWP with 8 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

Figure 5.27 shows that the same scenario results in an increase of the session block probability if the RPGM mobility model is used instead of the RWP. This is mainly due to the fact that the nodes are moving in groups which increases the risk of resource shortage if all group members are starting sessions at the same time. This group mobility also decreases the effect of the on-demand mode, but increases the benefit of the ad-hoc mode.

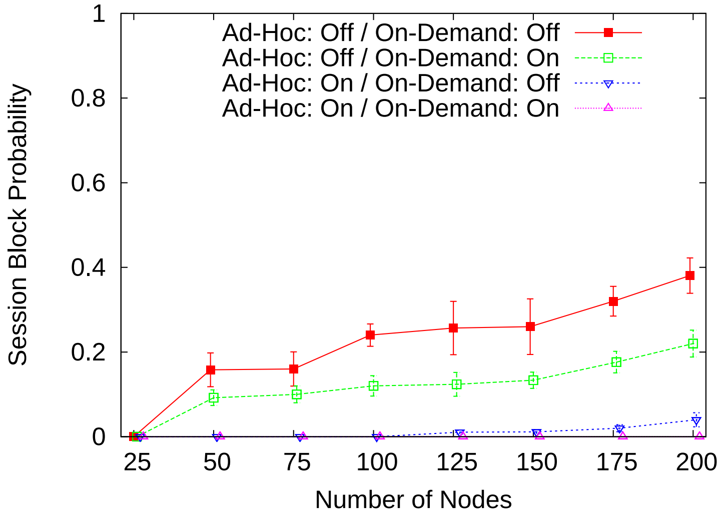


Figure 5.27: Session Block Rate: RPGM with 2 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

If Figure 5.28 and 5.26 are compared, the positive influence of the RPGM on the ad-hoc feature is clearly visible. The session block probability can be reduced by approximately 18% if the ad-hoc mode is enabled.

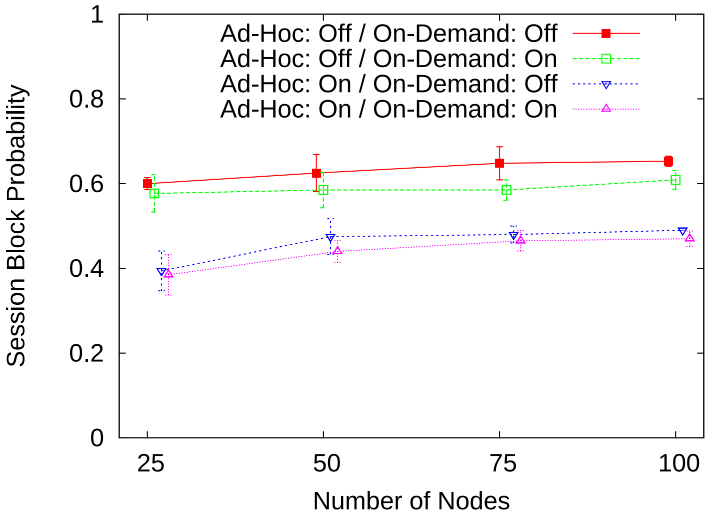


Figure 5.28: Session Block Rate: RPGM with 8 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

The session block probability is considerably increased if the UMTS and WLAN coverages are reduced to 50% and 5%, respectively. Figure 5.29 shows that the ad-hoc and on-demand features can together keep the session block probability low, but not assure that no session is blocked.

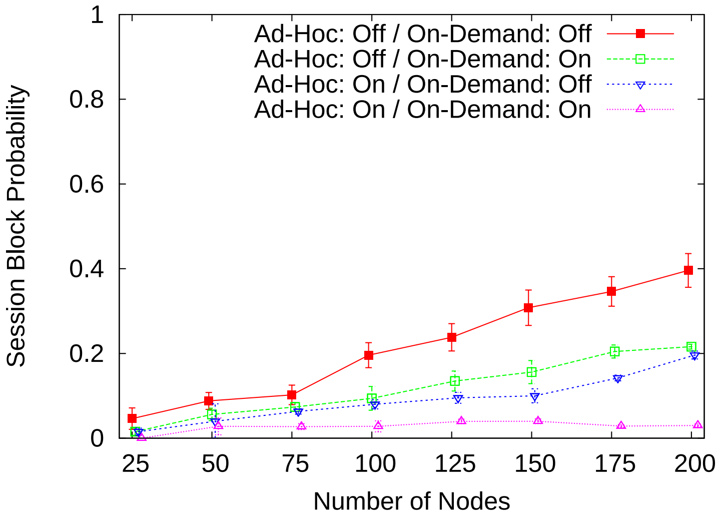


Figure 5.29: Session Block Rate: RWP with 2 Sessions/h and 100% GPRS, 50% UMTS and 5% WLAN Coverage

The RPGM is further increasing the session block probability because of the increased local network load introduced by the group mobility. Exactly this group characteristic is increasing also the probability of having infrastructureless connectivity and therefore enables the ad-hoc mode to sufficiently offload sessions and, thus, discharge the network load. Hence, if the ad-hoc mode is activated, the session block probability can be kept very low.

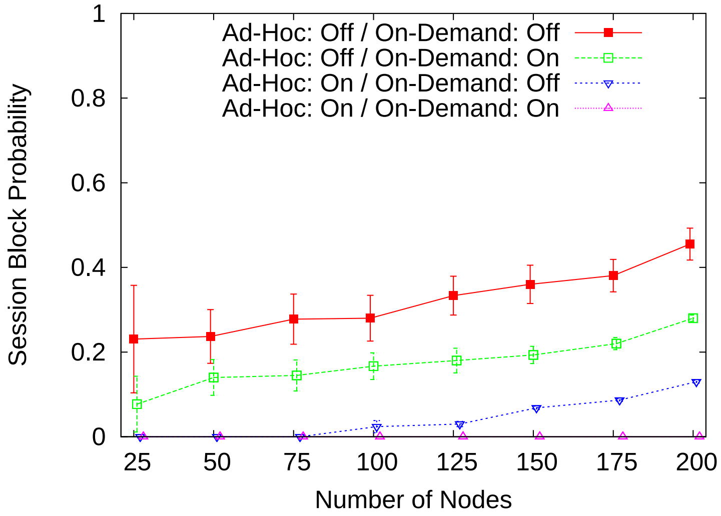


Figure 5.30: Session Block Rate: RPGM with 2 Sessions/h and 100% GPRS, 50% UMTS and 5% WLAN Coverage

Session Drop Rate and Outage

Whenever a communicating node comes into a congested cell the ongoing session is dropped. The cell is congested if according to the resource assignment model defined in Section 5.3.2 the remaining capacity of the base station is smaller than the minimum assignable bandwidth. The network tries to assign at least GPRS to a requesting node. Hence, the nodes can be forced to switch to another technology, if resources of that other technology are available. If even GPRS can not be assigned due to missing capacity, the node has to wait for available network resources to continue the session. During this waiting the ongoing session is hold, i.e. no further data can be transmitted. The number of drops per session has been analyzed for the different scenarios. Depending on the duration a node has to wait until it gets a network resource assigned after the session is dropped, the overall session duration is increased. This duration where the node is not able to transmit any data is called outage. To measure the outage relative to the overall session duration we define the outage ratio. In other words, the outage ratio represents the time a node spends waiting for network resources after the session gets dropped. If the network load is increasing, the drop rate is increasing together with the session block probability. As a further result of the overloaded network the outage ratio also increases. The nodes have to wait longer before they receive network resources again. During this outage, no drops can occur, which results in a decreasing number of drops per session again because the simulation period is limited. If the network load is high enough, the sessions do not terminate during the simulation period and tend thus to have less drops. In this stage the network is not able to serve additional sessions, which further decreases the average number of drops per session. This behavior is also represented by the following figures.

Figure 5.31 shows the number of drops that occur during a session if the nodes are moving based on the RWP mobility model, each of them starting 2 sessions per hour and having an UMTS and WLAN coverage of 80% and 10%, respectively. The GPRS coverage remains 100% like for all other simulations. If more than 75 nodes are using the network, the sessions get continuously dropped. If more than 125 nodes are present, the average number of drops per sessions reaches its maximum.

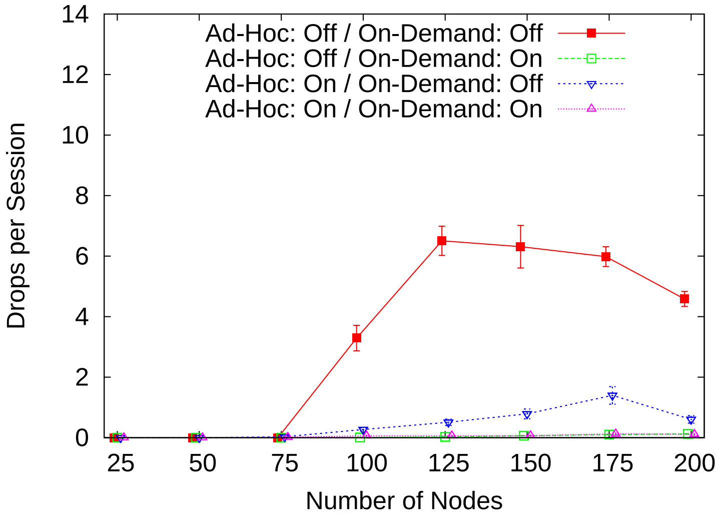


Figure 5.31: Session Drop Rate: RWP with 2 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

With the on-demand mode the drops can be avoided completely due to the released network resources between the sessions. The ad-hoc feature is considerably reducing the drop rate but is not able to avoid all session drops. Figure 5.32 shows the outage ratio for the same scenario. Without the help of the ad-hoc and on-demand mode the outage ratio is increasing with the number of nodes. The outage starts to increase simultaneously with the session block probability discussed before.

If the number of sessions is increased to 8 sessions per hour, the drop rate is also increased. With 25 nodes the network is already heavily loaded, which results in drop rates of about 100 drops per session. With 50 nodes, the drop rate reaches its maximum and decreases gently. For such a heavy loaded network the on-demand mode is not suited anymore. Only the ad-hoc mode can decrease the amount of drops. Figure 5.33 shows that the drop rates with 8 sessions per hour. The outage ratio is depicted in Figure 5.34.

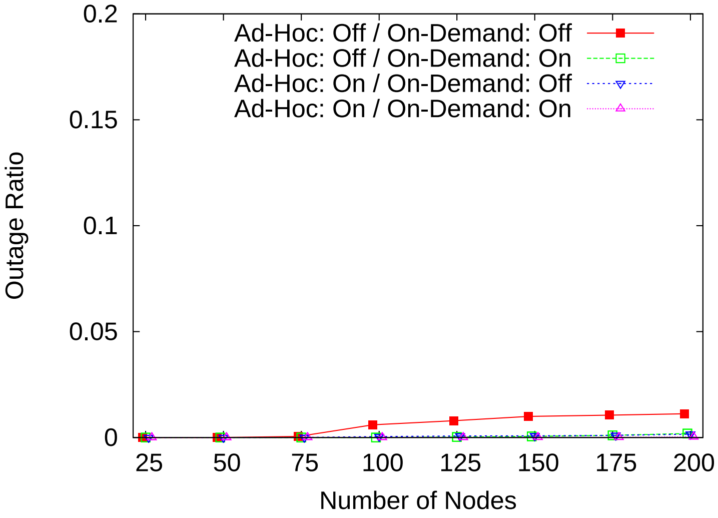


Figure 5.32: Session Outage: RWP with 2 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

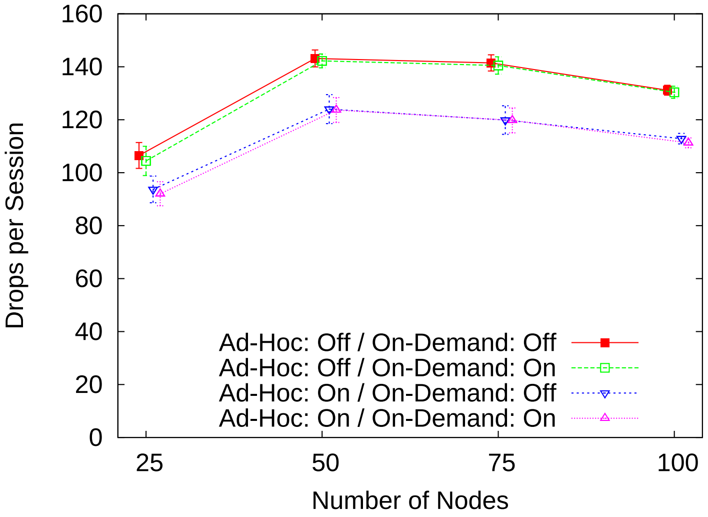


Figure 5.33: Session Drop Rate: RWP with 8 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

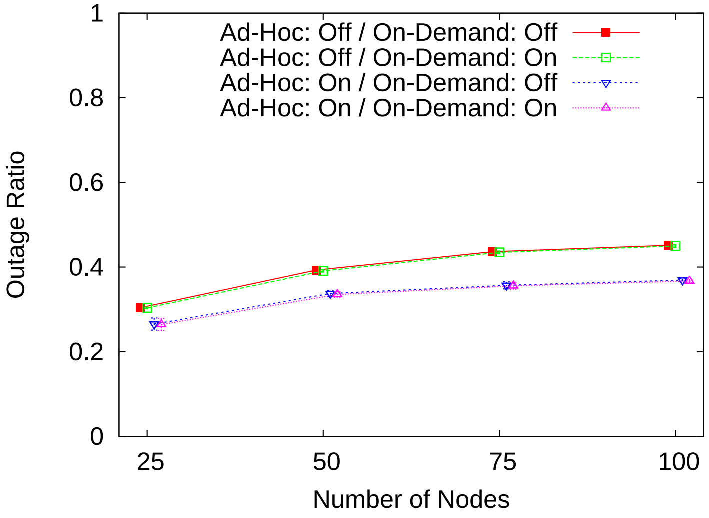


Figure 5.34: Session Outage: RWP with 8 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

The RPGM mobility model increases the number of drops per session like it did for the session block probability, because the nodes moving in groups competing for the same network resources. In Figure 5.35 and 5.36 the drops rate and outage ratio are shown, respectively.

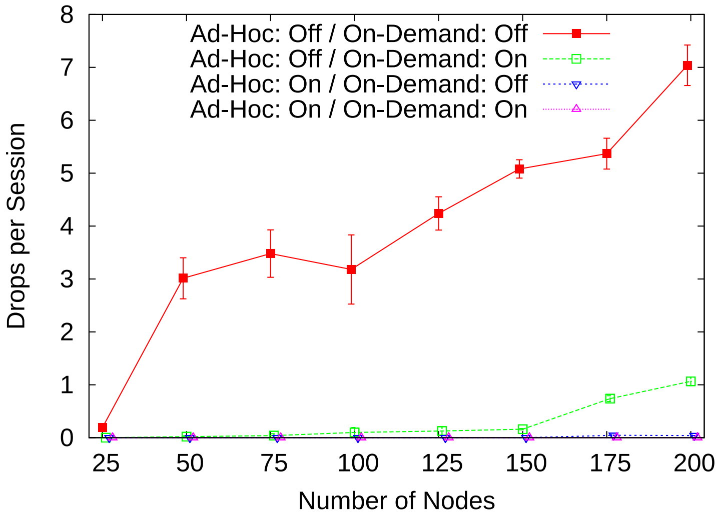


Figure 5.35: Session Drop Rate: RPGM with 2 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

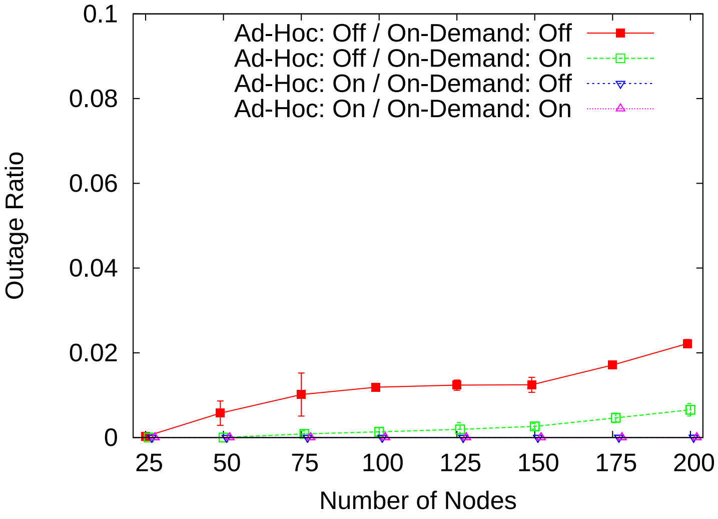


Figure 5.36: Session Outage: RPGM with 2 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

Figure 5.37 clearly shows that the on-demand feature is not influencing the drop rate for heavy loaded networks. Only the ad-hoc feature profits from the RPGM mobility model. With the ad-hoc feature the drop rate can be reduced up to 50%.

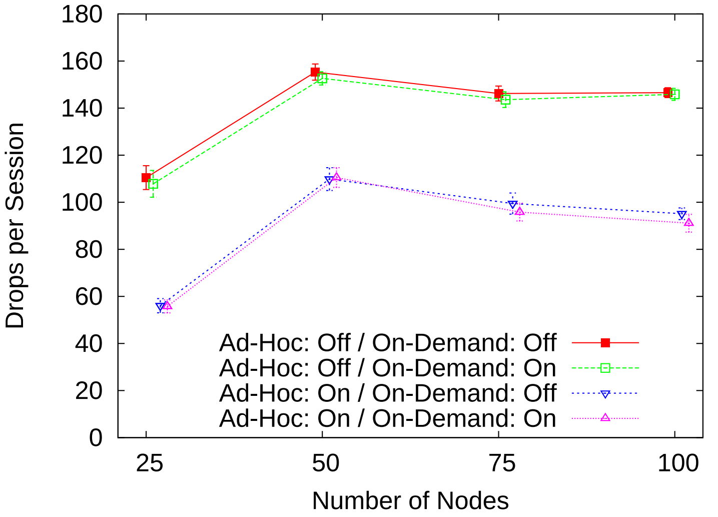


Figure 5.37: Session Drop Rate: RPGM with 8 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

The corresponding outage is depicted in Figure 5.38.

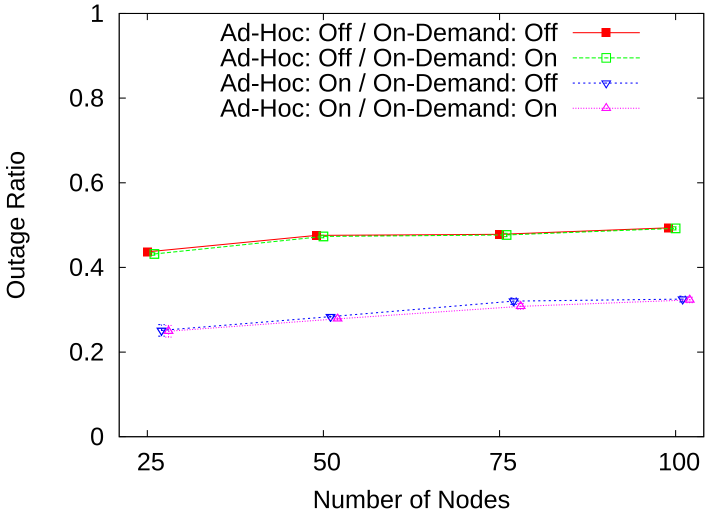


Figure 5.38: Session Outage: RPGM with 8 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

Figure 5.39 shows that the on-demand is strongly beneficial if the session density and coverages are low. Compared to the ad-hoc feature, which is only reducing the drop rate about a few drops per session, the on-demand is cutting the number of drops down by a factor up to 8. Figure 5.40 depicts the corresponding outage ratio.

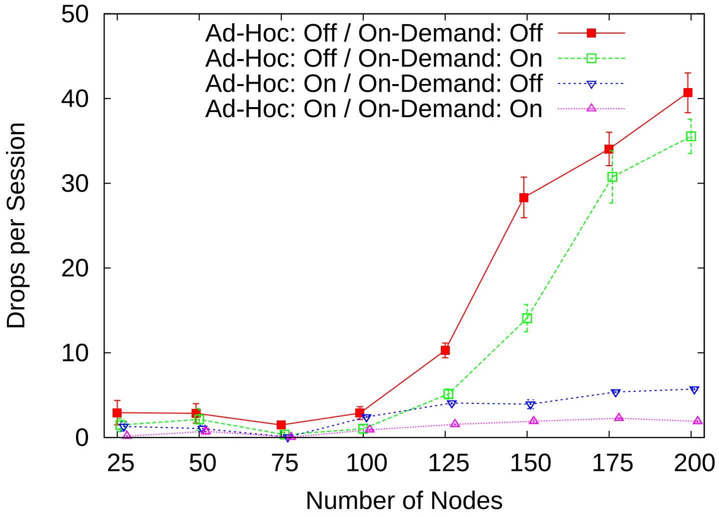


Figure 5.39: Session Drop Rate: RWP with 2 Sessions/h and 100% GPRS, 50% UMTS and 5% WLAN Coverage

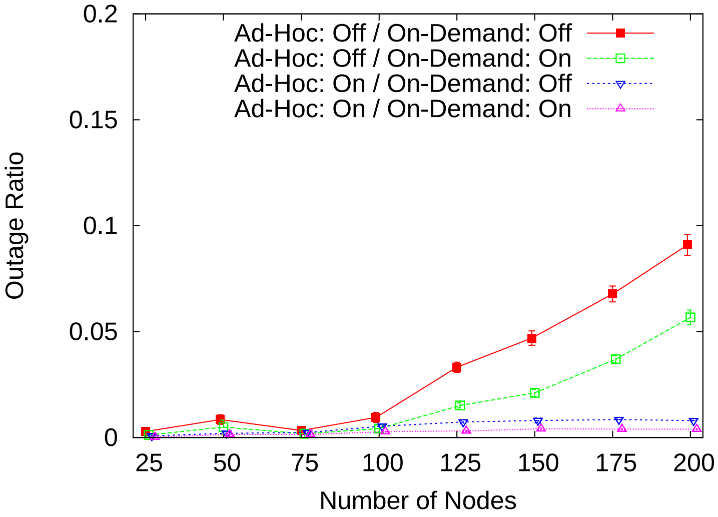


Figure 5.40: Session Outage: RWP with 2 Sessions/h and 100% GPRS, 50% UMTS and 5% WLAN Coverage

The same scenario but with RPGM is changing the picture. The ad-hoc is heavily decreasing the drop rate, whereas the on-demand is only slightly influencing the number of drops per session.

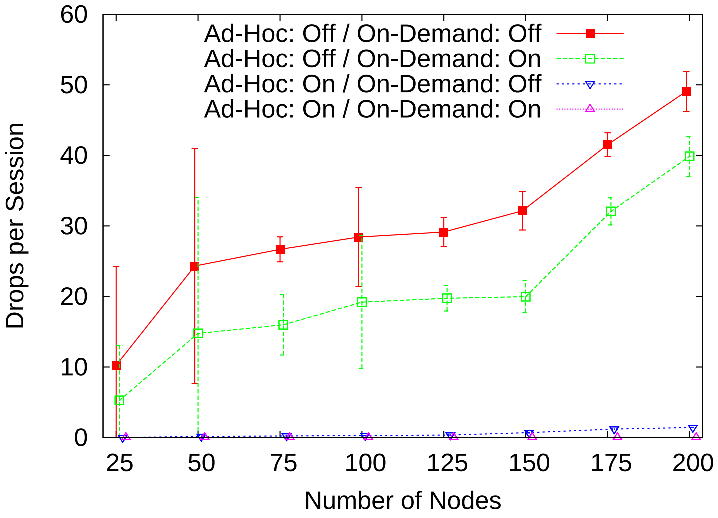


Figure 5.41: Session Drop Rate: RPGM with 2 Sessions/h and 100% GPRS, 50% UMTS and 5% WLAN Coverage

Session Throughput

Unlike in the first simulation set, where we evaluated the session throughput assuming unlimited network resources, the values obtained from the second simulation set are including the delays and outages imposed by the session blocking and dropping, which may occur if nodes have to share the network capacity. These delays and outages increase the duration of a session and, thus, decrease the average throughput of the session. Furthermore, the actual throughput is decreased whenever the network has to perform load balancing. If a node comes into a congested UMTS cell, the network assigns GPRS instead. This degradation did not happen in the first simulation set, where the UMTS resources were assumed to be unlimited. Figure 5.42 and 5.43 show the simulation results for the two session densities 2 and 8, respectively, assuming the high coverage scenario and RWP mobility. The ad-hoc mode is much more beneficial for the the throughput than the on-demand mode is. This is mainly due to the high data rates offered by the infrastructure-less links. Nevertheless, the negative impact of the number of sessions per hour on the on-demand mode is clearly visible when comparing the Figures 5.42 and 5.43.

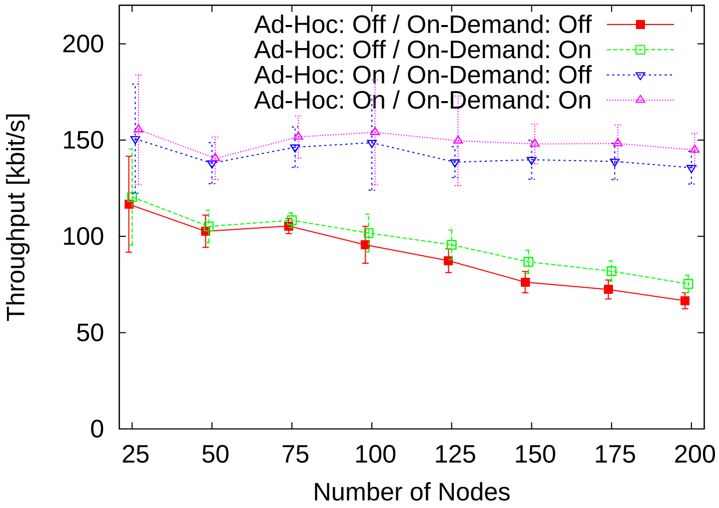


Figure 5.42: Session Throughput: RWP with 2 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

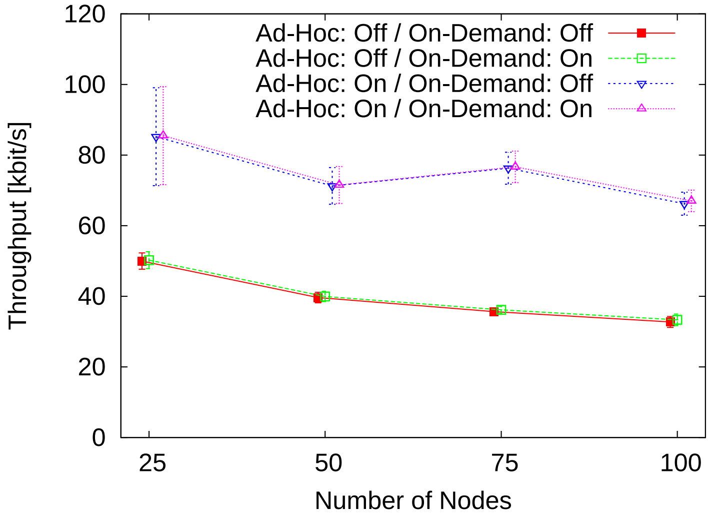


Figure 5.43: Session Throughput: RWP with 8 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

When choosing the RPGM mobility model, the average throughput is even further increased due to the ad-hoc mode. Considering the curve for the joint on-demand and ad-hoc feature in Figure 5.44, there is a positive influence of the ad-hoc mode to the benefit introduced by the on-demand mode. Whenever communicating nodes are moving within the same group, the network is further discharged, which increases the throughput of the infrastructure-based sessions.

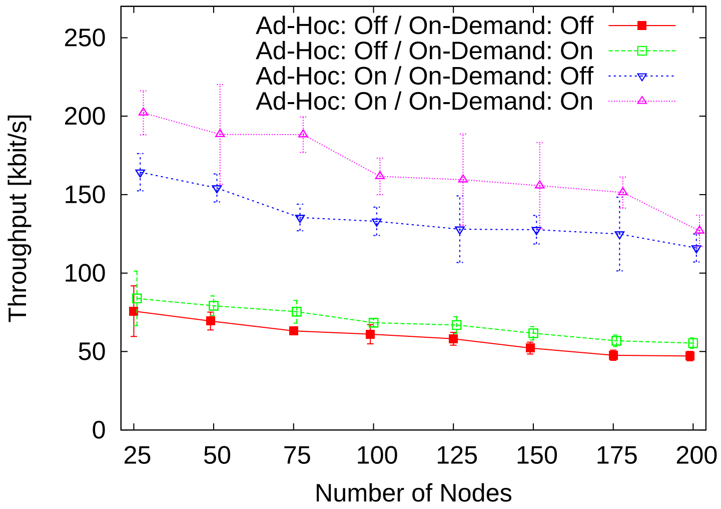


Figure 5.44: Session Throughput: RPGM with 2 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

Figure 5.45 depicts the throughput for 8 sessions per hour. Both, the ad-hoc and the on-demand feature are highly beneficial for low node densities.

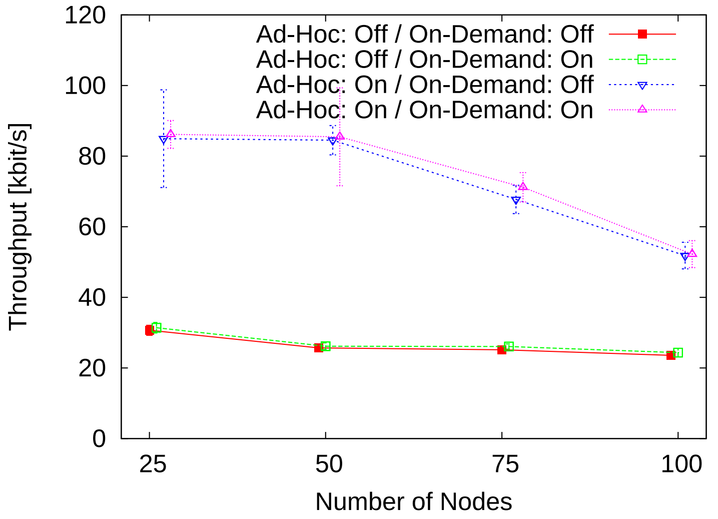


Figure 5.45: Session Throughput: RPGM with 8 Sessions/h and 100% GPRS, 80% UMTS and 10% WLAN Coverage

Figures 5.46 and 5.47 show the simulation results, when reducing the coverage for UMTS and WLAN for the RWP mobility model. For the lower session density, the on-demand mode can help to increase the throughput, but only for smaller number of nodes. Compared to the values obtained with the higher coverage scenario, the impact of the on-demand mode is negligible.

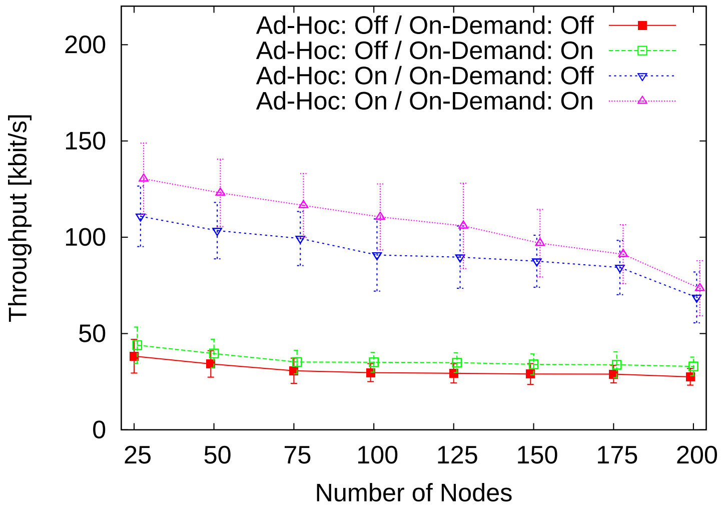


Figure 5.46: Session Throughput: RPGM with 2 Sessions/h and 100% GPRS, 50% UMTS and 5% WLAN Coverage

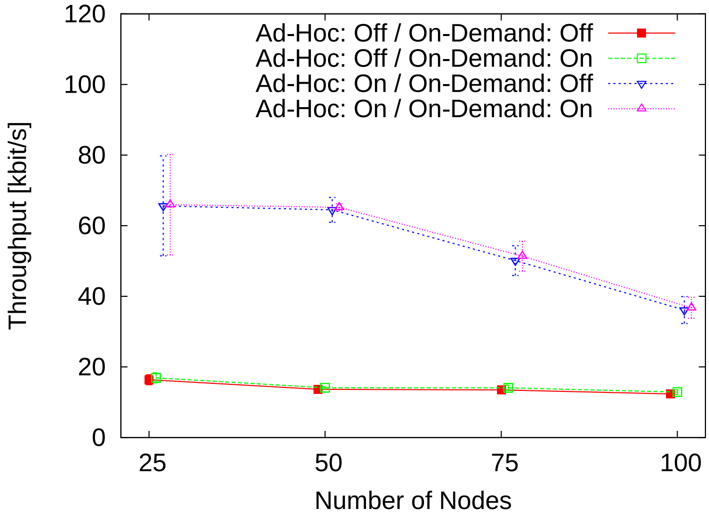


Figure 5.47: Session Throughput: RPGM with 8 Sessions/h and 100% GPRS, 50% UMTS and 5% WLAN Coverage

5.3.4 Summary of the Simulation Results

The simulations show that the ad-hoc mode is clearly profiting from the RPGM mobility model, where nodes are likely to move within groups sharing the same path. The characteristics of the RPGM were set on purpose to guarantee that the group members stay within the range of the infrastructure-less communication interface. Thus, whenever sender and receiver of a session happen to join the same group, a large part of the data could be sent using the infrastructure-less technology providing high data rates. The nodes changed the group with a pre-defined probability to assure a certain level of mobility of the individual node. Furthermore, the selection of the session end-points was done independent of the group membership. In reality there is probably a higher probability that nodes moving in the same group also start session among them. This behavior would for sure be beneficial for the ad-hoc mode. On the other hand the on-demand feature was generally suffering from the group mobility offered by the RPGM. The member of the group traveling the same path had to compete for all available resources. The amount of sessions per hour also dramatically decreased the benefit of the on-demand mode. The on-demand mode can only release network resources between the session when no data has to be sent or received. For higher session densities, where the period between the sessions is very short, there is not much time that the on-demand mode can release the assigned network resources. The simulations results showed that both, the ad-hoc and on-demand mode never decrease the network performance. Hence, there is no risk to enable both modes permanently. The simulations showed that in certain scenarios our concept of SMACS can help to increase the throughput by up to a factor of 4, the network efficiency up to 70%, and decrease the energy consumption close to 80%. In several cased the session block and drop probability can be reduced to zero due to efficient resource releasing, resulting in a reduction of the session outage ratio of data sessions in heterogeneous networking environments of up to almost 20%.

The modeling of the infrastructure-based technologies as defined in Section 5.3.2 is very conservative. In a real GPRS network, the achievable bandwidth values are considerably lower, because the CS-4 can only be used for nodes being very close to the base station and hence rarely used. More realistically, nodes use the CS-3, delivering actual speeds about 14.4 $kbit/s$ per slot, which would result in about 32% lower assigned bandwidth. Similarly, the UMTS model used for the simulations, is rather delivering more bandwidth than the actual UMTS network. Since CDMA is reacting very sensitive on high activities on the same carrier, the achievable data rates are normally lower than assumed for our simulations. Empirically, data rates of 384 $kbit/s$ are only achieved for one single node in a cell. Also the resource assignment model for WLAN we modeled is rather optimistic. Collisions and interferences can dramatically reduce the achieved data rates, for both, the infrastructure and the ad-hoc mode of WLAN. We are aware of the fact that modeling the technologies is not resulting in most accurate values, but modeling the infrastructure-based technologies very optimistic allows to estimate the potential of our proposed ad-hoc and on-demand mode for the worst case scenario.

5.4 Conclusion

In this chapter we addressed the aspects of Smart Multi-Access Communications by introducing our architecture that allows to integrate existing solutions proving seamless access to infrastructure-based access networks with infrastructure-less communication technologies. The estimate the potential of our architecture we implemented our own network simulator and conducted several simulations. The evaluation of the simulation results collected revealed the strengths and weaknesses of the two main features provided by our architecture. The possibility to switch ongoing sessions to infrastructure-less communication technologies whenever possible and the ability to power up resource demanding broadband communication interfaces only if required, thus, on-demand, has the potential to considerably increase the network performance of existing heterogeneous networks.

Chapter 6

Cellular Assisted Heterogeneous Networking CAHN

6.1 Introduction

The previous chapters were studying the hurdles that users have to take to securely get connected via the different communication technologies. For sake of clarity, this was done separately for infrastructure-based access networks in Chapter 2, and for ad-hoc and direct node-to-node links in Chapter 4.4. Due to rather high economical attention, the academic and industrial research and development efforts to simplify the usage of infrastructure-based access networks, was much bigger than for infrastructure-less networks. However, to deliver a true always best connected feature to the end users, these infrastructure-less communication technologies have to be considered as well. Many research work is going on in the domain of pure ad-hoc networking and even more issues are raised due to the missing central infrastructure to properly manage resources, guarantee fairness, and provide security features (see Chapter 4.4). On the other hand lot of research effort is spent to increase the performance of infrastructure-based access networks to cope with the steadily increasing demand for broadband data. When making a step back, the most promising evolution of heterogeneous networking is the integration of both paradigms. Taking advantage of the well controlled cellular environment and the high capacity of ad-hoc and direct node-to-node communication. The resulting hybrid networks are incorporating the best of both worlds. As mentioned before, the adoption of simple and convenient ad-hoc and direct node-to-node communication is not happening at the same pace as it is for infrastructure-based. Existing technologies like Bluetooth or WLAN provide quite powerful means to interconnect efficiently neighboring nodes. But the handling is by far too complicated for end users not caring about technical details.

As discussed in Chapter 2 and 3 seamless authentication and session mobility on heterogeneous infrastructure-based access networks is enabling seamless access. EAP-SIM allows the usage of SIM credentials also for WLAN and Mobile IP offers seamless handover between IP enabled technologies. But there is no similar trend going on for ad-hoc and direct node-to-node links. Mobile IP version 6 includes route optimization (see Section 2.5.2) to avoid inefficient packet routing through the home agent, if the mobile node and the correspondent node are close to each other. Whenever nodes have established an ad-hoc or direct node-to-node link, route optimization is enabling also the IP session to be routed directly between the nodes. There is no way to further optimize the routing. However, Mobile IP route optimization by itself would never establish the direct link between the nodes, because it is limited to layer three. Some other mechanism has to bootstrap the link by scanning the neighborhood for the peering node and build up the link. Then the Mobile IP route optimization can be triggered to perform route optimization to directly send the packets using the new link. And exactly this missing process doing the bootstrapping for Mobile IP route optimization is the subject of this chapter.

6.2 Cellular Network as Signaling Plane

The abstraction of LSS and PSS was basically done having two main drivers in mind. First, users do not care about the devices of their peer. Similar to voice calls, the inviter does not care about which phone the invitee is using. Phones are only tools to achieve the actual goal, the conversation between the users. The same applies for data communication. E-mails, for example, are sent to persons not to computers or mobile devices. So, one could state that the endpoints of the transaction are persons not devices. Secondly, flexibility is key when taking the variety of end devices that are available today into account. The trend to all-IP applications and IP capable devices with powerful operating systems is pushing the limits of dedicated devices. Smartphones and PDAs become multi-purpose devices and laptops get more and more mobile. With VoIP applications like Skype [123], laptops become already today the device of choice when making voice calls abroad. Generally spoken, IP applications should be dynamically terminated on a specific device depending on the user's preference. Therefore, the system should allow end users to decide on a per session basis which device they want to use. To do so, the system has to be able to prompt the user whenever a new session is started. This prompting has to happen on a device which is always on and always close to the user.

Mobile Phone as Enabler Heterogeneous Data Sessions

The most obvious device fulfilling these requirements is probably the mobile phone. There are further arguments for the selection of the mobile phone as the primary device for heterogeneous session management. First of all, the addressing scheme used for mobile phones is very popular to identify a person. Most of the people indicate their mobile phone number, when they have to be reachable. To achieve similar level of convenience for heterogeneous data session management than users have nowadays for voice calls, similar procedures have to be introduced. Addressing a peer to start a heterogeneous data session should be as easy as choosing a contact from the address book. And receiving a session request should be as simple as receiving a voice call. Designing the management of heterogeneous data sessions similar to voice call handling could considerably increase the convenience and therefore the usage of data communication networks.

Reusing the Security Mechanisms of the Cellular Network

The usage of the cellular network as primary signaling plane for heterogeneous sessions has further advantages in terms of security. Every cellular subscriber has a security relation with the network operator established during the subscription. With the help of this security relation represented in the SIM, the communication between the mobile handset and the cellular network can be protected. Reusing this secure channel to exchange sensitive information between nodes is simple but powerful [41, 39, 46, 40]. Mobile phones are always attached to the cellular network and can therefore easily be reached at anytime and anywhere. For the heterogeneous session management the reuse of the cellular network solves a lot of problems. The security relation between the nodes and the operator can be used to build up a trust chain between nodes to secure the exchange of configuration and security parameters mention in Section 4.4.2.

CAHN Signaling over the Cellular Network

When using the cellular network as signaling plane to initiate and maintain heterogeneous broadband sessions, there are several ways to do so. GSM and UMTS offer basically two different possibilities to transfer data (see Section 2.2.1). The most obvious is using the standard data channels like CSD, HSCSD, GPRS, EDGE or UMTS. Therefore, an IP data channel has to be established either doing a dial-up and running a PPP [173] session on top of the circuit switched based channels CSD and HSCSD or establishing a packet switched channel on GPRS, EDGE or UMTS. The resulting IP address could then be used for the LSS addressing. The major drawback of this approach is probably the permanent allocated radio resource for the data channel, which is only occasionally used to signal setup and maintenance of a heterogeneous data session. Using these IP data channels for the LSS is also not optimal because these IP addresses are normally assigned dynamically per session. Hence, if the data channel gets disconnected, it gets a different IP address assigned after reconnection, which forces the node to promote that new IP address. The second possibility to transfer data between nodes using the cellular network is based on the services offered by the signaling system (SS7). SMS and USSD are both offering simple and secured transmission of information, they are message and session based, respectively. SMS is a store and forward service and therefore not suited for realtime applications. However, the delivery delay of a SMS strongly depends on the dimensioning of the SMS-Center, which is the main serving entity for SMS. The USSD is session oriented and hence more appropriate to handle signaling for longer lasting data sessions (see Section 6.6.1).

Using signaling based on channels like SMS or USSD to bootstrap heterogeneous communication sessions has big potential to save energy. The complete signaling system of the cellular network was designed to be very power efficient, offering sophisticated energy saving algorithms, paging, and adaptive transmission power control. Using the cellular signaling system to enable on-demand setup of power consuming broadband communication technologies like GPRS, UMTS or WLAN can considerably increase battery lifetime depending on the number and length of data sessions (see Section 5.3).

6.2.1 Interworking with SIP

SMACS uses the CAHN protocol to exchange configuration and security parameters required to establish secured communication. The CAHN protocol messages can be transmitted through any type of network. Whenever nodes are attached to an IP network and successfully registered with their SIP registrar servers, the CAHN protocol messages can exchanged through SIP extensions. SIP is then used as transport medium to deliver the CAHN protocol messages to the right destination. To secure the CAHN protocol messages any additional security has to be used together with SIP like SSL or IPsec. For future cellular network releases which include SIP, the use of SIP extensions to transport CAHN protocol messages is straight forward. The same is true for wireless network supporting SIP signaling. The only requirement for the secure application of our SMACS/CAHN framework is the secure link between the nodes and the SIP server to assure the protection of the sensitive information exchanged by the CAHN protocol.

If nodes do not have IP connectivity (e.g., due to resource saving), the CAHN protocol messages can use any other non-IP signaling channel (i.e. cellular) to trigger the peer to get connected to an IP network. As soon as the nodes do have IP connectivity SIP can be used to exchange configuration and security parameter required to securely establish direct links if the communicating nodes come close enough to each other. In[169], we analyzed in further detail the utilization of SIP extensions to exchange CAHN messages to establish an IPsec link between nodes. To prove our concept, we implemented a simple testbed.

6.2.2 Using Multiple Identifiers

The advantages of using the cellular signaling system to handle the bootstrapping of heterogeneous end-to-end IP sessions exposed in the previous sections highly influenced the design of the system. The LSS addressing is based on the mobile phone numbers (MSISDN) and identifies the end users. To allow a maximum of flexibility, the LSS supports IP addresses and NAIs as well. Especially, when changing from out-of-band to inband signaling, the use of IP addresses simplifies the exchange of CAHN protocol messages. The nodes participating in a session keep a table of valid identifiers including the MSISDN for all other nodes. The MSISDN serves as primary identifiers, whereas the other identifier (e.g., IP addresses) are only temporally available and therefore rather considerable as secondary identifiers. There are mainly two possibilities to handle the identifier/address resolution based on these tables. One is decentralized and hence based on the information provided by the corresponding peer. Whenever a heterogeneous IP session is established, the peer informs the initiator about all the available communication addresses. The initiator updates its copy of the identifier table accordingly and starts the heterogeneous session setup using the most appropriate signaling channel available. Fig. 6.1 illustrates the decentralized identifier/address resolution. Since the IP addresses are only of

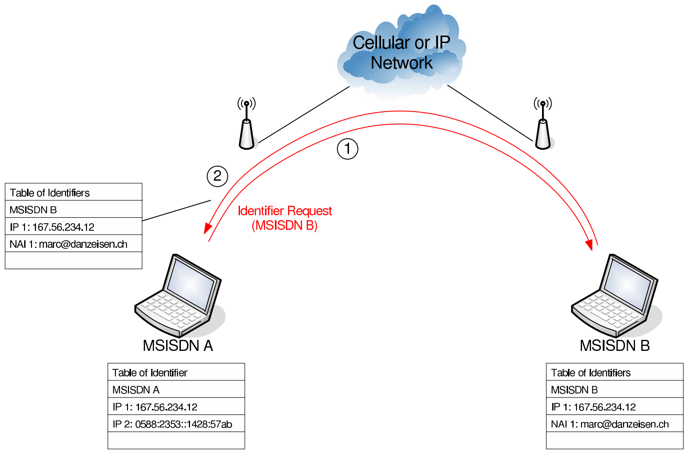


Figure 6.1: Decentralized Identifier/Address Resolution

temporal nature, the MSISDN[1] is the only identifier which is known prior to the connection establishment. Hence, the query is sent to the destination using the cellular network (step 1), where it is then processed. The destination then answers the request by sending a list of the currently valid identifiers that can be used to exchange CAHN protocol messages (step 2). The decentralized approach allows nodes to individually decide which identifiers should be revealed to the querying node, if at all. Especially to keep privacy and inhibit tracing this might be beneficial.

The second possibility is based on a centralized service where all available identifiers are stored. Before establishing the session, nodes can query that centralized service about alternative identifiers (i.e. temporarily acquired IP addresses) of the peer. In contrast to the reactive decentralized approach, nodes have to pro-actively update the central server, if any temporal identifiers get invalid or new ones become available (step 1). This might result is faster responses to queries but requires regularly signaling to keep the identifier tables consistent. The request is then sent directly to the central server (step 2), where a copy of the actual list of identifiers is sent back to the requesting node (step 3). Fig. 6.2 shows the centralized approach described above.

If there are other communication channels available than the cellular signaling system, which is the case whenever an heterogeneous IP session is already established, the session signaling can be done inband, and not using valuable cellular resources. However, the cellular signaling system is kept as backup signaling channel for the heterogeneous IP session in the case of loss of the actual data channel. Hence, the LSS is only relying on the MSISDN if no IP address is associated to the peer. This flexibility limits the use of the cellular network to bootstrap the session management. As soon as a secure IP session is established, the signaling can be done inband using this secured session. In case of unreliable connections, the signaling can be continued in an out-of-band manner using the cellular network. The next section is introducing a system to separately treat

[1] In the case of permanent IP connectivity SIP identifiers could be used as primary identifiers as well. For sake of clarity the figures are illustrating the use of the MSISDN as primary identifier.

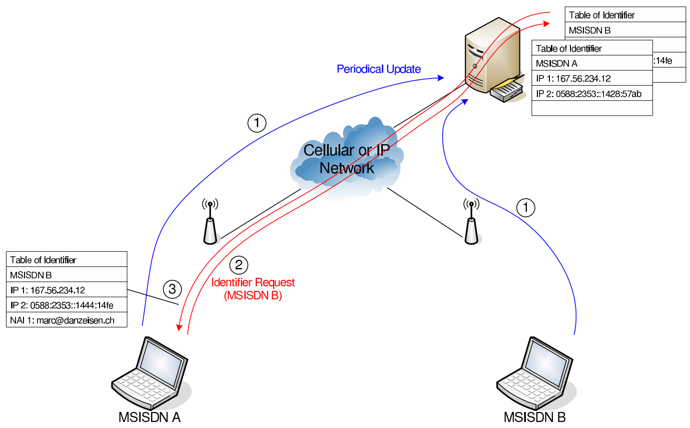


Figure 6.2: Centralized Identifier/Address Resolution

the routing of session signaling and data taking into account the variety of channels available in a heterogeneous environment. The separation of the routing decision based on the content of the packets instead of the destination, like it is done for standard IP communications today, is offering a higher degree of flexibility in heterogeneous IP session management. Packets delivering sensitive or time critical signaling information can be routed on appropriate communication channels like USSD or direct links between nodes.

6.3 CAHN Architecture

As mentioned in Section 4.5 the proposed architecture consists of two new logical layers. The CAHN component introduced in this chapter is mainly addressing the issues related to ad-hoc and direct node-to-node links and the provisioning of the CAHN protocol to exchange configuration and security related parameter required to bootstrap communication sessions. The CAHN functionality is required to securely setup ad-hoc and direct node-to-node links between nodes, whereas SecMIP is providing seamless and secured access to infrastructure-based networks. SMACS finally, is selecting the most appropriate signaling and data channels out of the ones offered by CAHN and SecMIP. Since the CAHN and the SMACS layer are both using the cellular network to bootstrap the connection, respectively the session, they are sharing some functionality. From an architectural point of view, SMACS is using the services provided by CAHN and SecMIP. In the special case, where the two nodes communicate over a single-hop link, SMACS is mainly relying on the functions of CAHN. Hence, SMACS is only required if infrastructure-based or multi-hop ad-hoc networks are involved in the physical communication path. For that reason the introduction of CAHN can be done without SMACS, if only direct node-to-node links are considered. Chapter 5 is then releasing this constraint and introducing SMACS by analyzing in detail which additional functionality is required to integrate infrastructure-based and multi-hop links as well.

Chapter 2, 3 and 4.4 showed that the heterogeneity embarrass simple usage of the various communication technologies. Despite having a theoretical increase of the overall performance of wireless networks, users can not really profit due to lack of knowledge required to choose and setup the most appropriate technology at the right time. Users would have to know about the characteristics of the running applications, the available networks, as well as how to manage the different communication interfaces. The system proposed in this thesis represents a framework to enable automatic and user friendly management of heterogeneous devices and networking technologies. As mentioned at the beginning of this document, the envisioned system should abstract the complexity of address resolution, authentication, key management, and session handover to make heterogeneous data sessions appear as simple as mobile voice.

Layers of CAHN

The CAHN component consists of three major layers (or modules) called *CAHN Communication Management* (CCM), *Connectors* and *Adapters*. The CCM is handling the LSS related functions, whereas the Connector is responsible for the appropriate configuration of the different networking interfaces. Because of the variety of network interfaces, there is one dedicated Connector for each interface type. An Adapter is adapting the messages exchanged between the CCMs of the peering nodes, according to the characteristics of the selected signaling channel. In the case of SMS, for example, the SMS Adapter is splitting the CCM messages to fit the length of a SMS. Incoming SMS have to be correctly combined to form the original CCM message. For sake of modularity, there is also one dedicated Adapter for each signaling channel. The presented architecture for CAHN is enabling the secure establishment of ad-hoc and direct node-to-node links between nodes, with the help of the cellular network. Fig. 6.3 depicts the general architecture with the three basic layers.

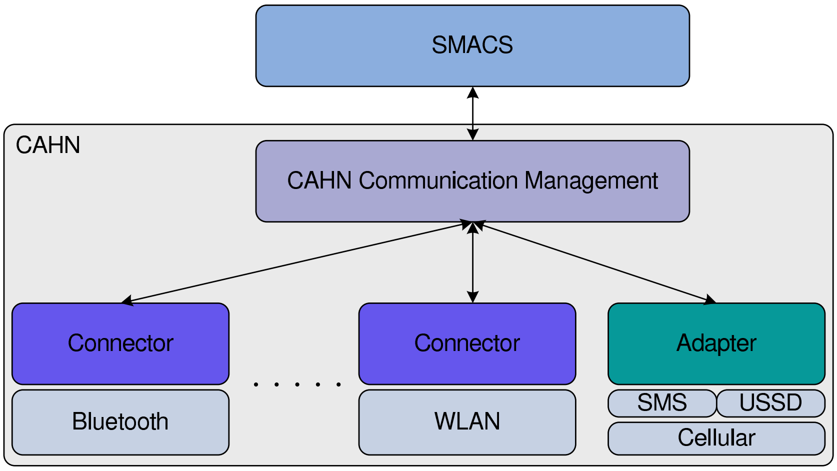


Figure 6.3: CAHN Architecture

Note, that these three layers can be located separately on the different nodes of a PAN depending on their role. The CCM is the main logic placed on the supernode of the PAN. The supernode is responsible for the user interaction and the handling of the management of the logical sessions. It requires therefore connectivity to the cellular network. Consequently, the supernode is also

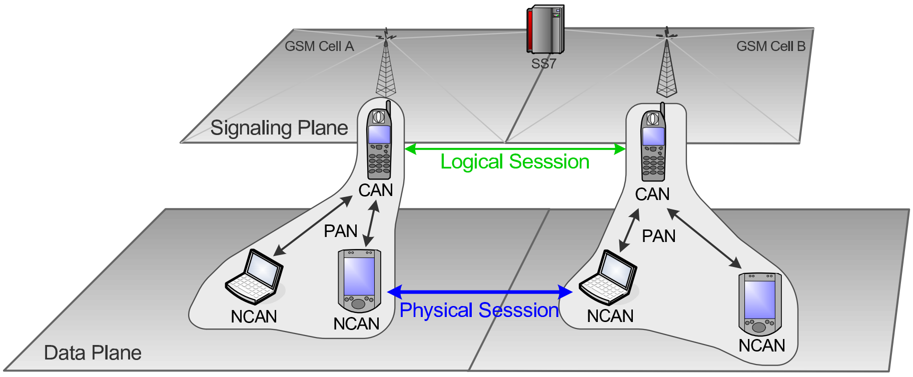


Figure 6.4: CAN and NCAN Nodes

referred to as *Cellular Aware Node* (CAN) in the context of CAHN. To extend the scope of CAHN, the CAN can act as a gateway for *Non Cellular-Aware Nodes* (NCANs) of the same PAN. Hence, the CAN can handle the logical session management and delegate the physical sessions to the NCANs. The layers required for the physical session handling, namely the Connectors, have to be present on the NCAN as well. Fig. 6.4 illustrates the difference between the CAN and the NCANs. The communication between the CCM and the underlying Connectors is based on standard communication sockets to guarantee full flexibility. Hence, for the CCM it is not of importance if the controlled Connector is on the CAN itself or on one of the NCANs of the PAN, as long as there is IP connectivity provided. This IP connectivity is assumed to be provided by the PAN and not further analyzed in the scope of this thesis. Consequently, the separation between CAN and NCANs is not relevant for the introduction of the CAHN architecture and for sake of simplicity not explicitly mentioned anymore.[2] Throughout the rest of this document, the three layers of CAHN are considered to be on the same device (i.e. a laptop having multiple communication interfaces available, including the connection to the cellular network).

6.3.1 CAHN Communication Management

The CAHN communication management is offering functions related to ad-hoc and direct node-to-node connection management towards the SMACS layer, based on the services provided by the Connectors and Adapters. All functions are called using messages, which enables an additional degree of flexibility. Hence, the SMACS layer can send a message either to the local or to the remote CCM to call a function. Realizing the communication between the two layers on standard sockets allows a simple and powerful remote function call. Of course the messages passed to a remote node have to be analyzed by the remote CAHN layer primarily to avoid any security risks (i.e. unauthorized manipulation of the remote Connectors). Fig. 6.5 visualizes the differentiation between local and remote message delivery.

[2] However, the separation between CANs (e.g., mobile phones) handling the user interaction and the logical sessions and the NCANs (e.g., PDAs, laptops, etc) providing the actual physical connection might be an essential enabler for commercial use cases.

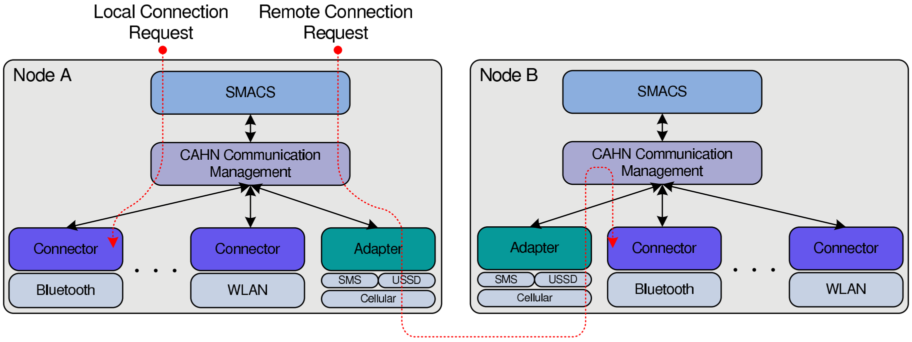


Figure 6.5: Local vs. Remote Message Delivery

Local requests are directly delivered to the local connectors. Remote connection requests are delivered to the remote node using the appropriate Adapter. A further advantage of the socket based architecture to handle function calls among the layers is the possibility to design the system architecture completely symmetric on all nodes. A connection request coming from the local SMACS layer (i.e. triggered by the local user, inviting another node), looks very much the same for the local CAHN layer, than a connection request received from the remote node, inviting for a session. Nevertheless, the CAHN layer can easily distinguish local and remote requests based on the different source addresses of the message. Beside the most obvious functions/messages like *Connection Request* and the *Connection Accept*, there are further messages defined to enable more sophisticated handling of ad-hoc and direct node-to-node links. They have been mainly introduced to provide more detailed information about the link states to enable the SMACS layer to react more intelligently when managing end-to-end heterogeneous sessions.

Fig. 6.6 is providing an overview of all functions required for the CCM to interact with the SMACS layer, the Connectors, the Adapters, and the IP stack of the operating system.

CCM-SMACS Interaction

The **Connection Request** is initializing the connection setup procedure and proposing a set of configuration and security related parameters to the invitee. The inviting node includes as much information as possible, to maximize the chance to meet the requirements of the invited node. This information includes concrete propositions of configuration for all available communication technologies. Specific settings like frequency, network ID, IP addresses, network prefixes, and network masks, but also security related parameters like encryption mechanisms and keys are inclosed in the connection request.

If the user accepts the invitation, the SMACS layer triggers the CCM to accept the connection request with the **Connection Accept** function. The SMACS layer can either accept the proposed connection parameters included in the connection request, or propose a new set of parameters in the case of collision with other communication interfaces (i.e. with other infrastructure-based connections).

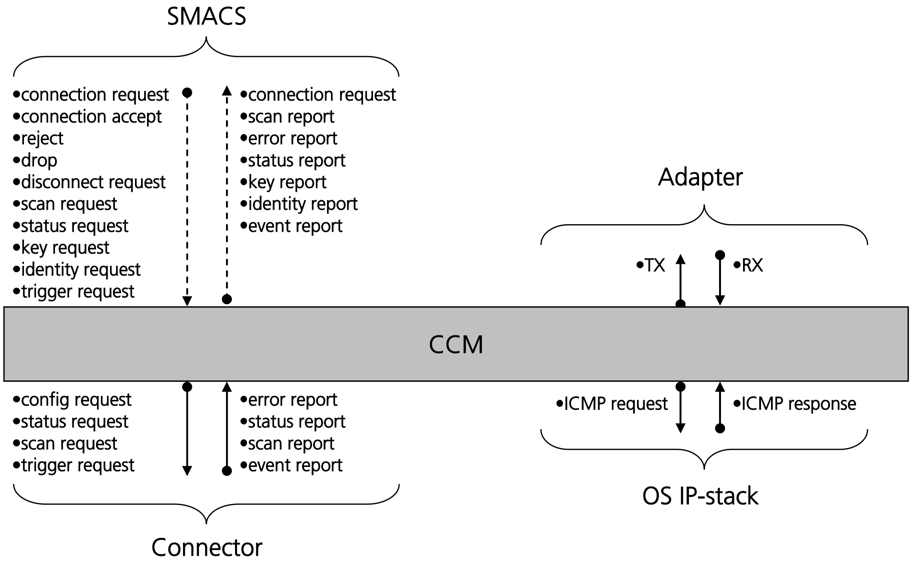


Figure 6.6: CAHN: Services and SAPs

If the connection request is not granted, the SMACS layer can reply with an **Error Report**.

To disconnect an ad-hoc or direct node-to-node link the **Disconnect Request** function was defined.

The introduction of the **Scan Request** function is allowing the SMACS layer to trigger a scan of its neighborhood. Due to the symmetric architecture of CAHN, this trigger can be sent to the peer nodes as well. This can be valuable to decide whether a direct connection is possible or not. It might also be used to get better location information about other members of the session or joining nodes. The requesting SMACS layer can ask for a specific network or node to be scanned for, or ask for a general scan. The scan result is then provided within the **Scan Report** function. If errors occur, they can be indicated with the help of the **Error Report** function.

Status Report is used to query information about a single interface, a specific technology, ad-hoc or direct node-to-node link or all available interfaces. The **Status Report** is revealing information about the type of the interfaces, their operation mode, IP addresses and netmasks used, signal strength or even about the reliability of the link.

To enhance the routing decision for signaling and data, made by the SMACS layer, traps can be set to get reports if certain threshold are reached or events are occurring. The **Trap Request** functionality is used to set traps. Again, due to the generic architecture of the system, these trap requests can be sent to the local but also remote CCM. Whenever a trap is released, the trap owner is notified by a **Trap Report**.

Especially for conference and meeting applications, the system offers the possibility to get identity information from the peering node. Therefore, an **Identity Request** can be sent. **Identity Reports** allow nodes to reply by providing identity information like username, affiliation, a.o.[3]

CCM-OS Interaction

Additionally, the CCM requires access to the IP stack of the operating system to check the established connections and routing. The simplest way to do so is by using the ICMP echo request and reply functions.

6.3.2 Connectors

The Connectors are providing a generalized interface to the CCM to control the different communication interfaces. Therefore, the technology specific functions have to be encapsulated in rather generic functions, which can then be used by the CCM. Whenever two nodes want to interconnect using a specific communication technology, the Connectors, being in charge of the corresponding interface on each node, are negotiating the required settings to successfully setup the link. The CCMs are providing the required secured signaling channel between the Connectors. When receiving a connection request for a specific communication interface, the CCM checks for appropriate Connectors (either local or within his PAN). If there are several potential Connectors available, a user interaction might be required to select the preferred interface or PAN device. The Connector offers functions towards the CCM to correctly handle physical session related settings. Fig. 6.7 depicts all Connector related functions towards the CCM, the Adapter and the network interface, it is in charge of.

The different functions are discussed in detail in the following section. Further information about how to use these functions is given in Section 6.4.1.

Connector-CCM Interaction

The most obvious function provided by the Connector is the **Config Request**, which is providing means to set the configuration parameters of an interface. Within one configuration request there are layer one to three related parameters that have to be passed to the Connector. Config requests are acknowledged by **Error Reports**, whereas an error code 0 stands for successful configuration.

The **Status Request**, **Scan Request**, and **Trap Request** features are identical to the ones offered by the CCM towards the SMACS layer. The CCM is dispatching the requests based on the destination address either to the local or the remote Connector. The same applies for the corresponding reports.

Each Connector provides these functions to the CCM in the same manner, to realize a common interface between the CCM and the different communication interfaces. Connectors are abstracting the variety of methods required to handle the different communication interfaces and allow therefore the flexible

[3] Note, that this identity information is not used for authentication reasons, but only for user convenience.

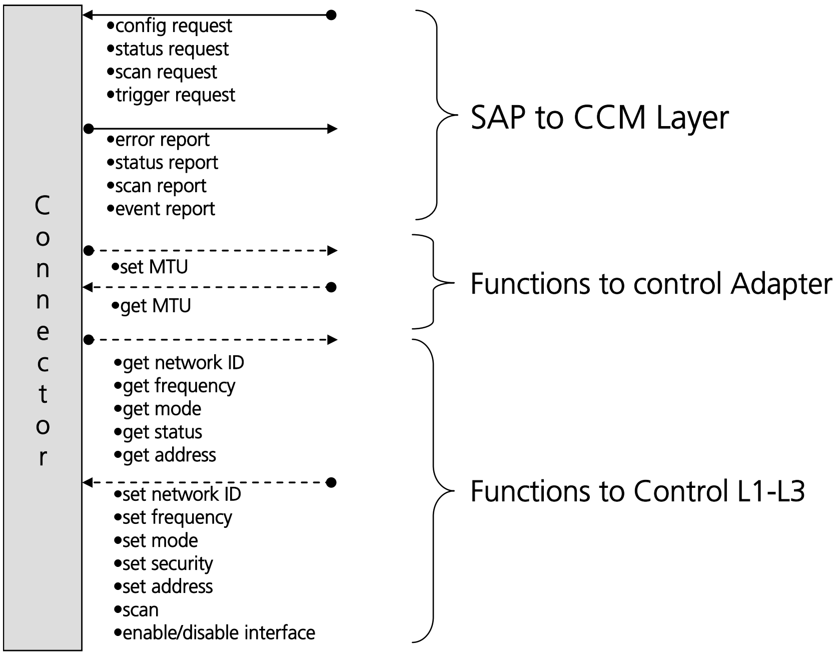


Figure 6.7: CCM Related Functions

introduction of new communication technologies, without requiring changes in the CCM.

Connector-Network Interface Interaction

The Connector is also interacting with the layer one, two, and three of the underlying network interface. Depending on the network interface functions, this might include functions to set and get parameters like *Network ID*, *Frequency*, *Mode*, *Status*, *Address*, but also to *enable* and *disable* the device.

6.3.3 Adapters

To enable support of different addressing schemes for the LSS, the system has to be modular. Therefore, *Adapters* were introduced to adapt the messages sent between the nodes. Dependent on the underlying channel that is used to transmit the signaling messages, a dedicated adapter is taking care of the fragmentation and flow control. This delegation of channel dependant manipulations of the messages to Adapters is enabling a flexible extension of the system, to cope with future signaling channels. Fig. 6.8 shows the Adapter related functions.

CCM-Adapter Interaction

The CCM interaction with the Adapters is very simple. There are only two functions required, one for sending and one for receiving signaling messages. The SMACS layer is making the routing decision for the transmission of signaling

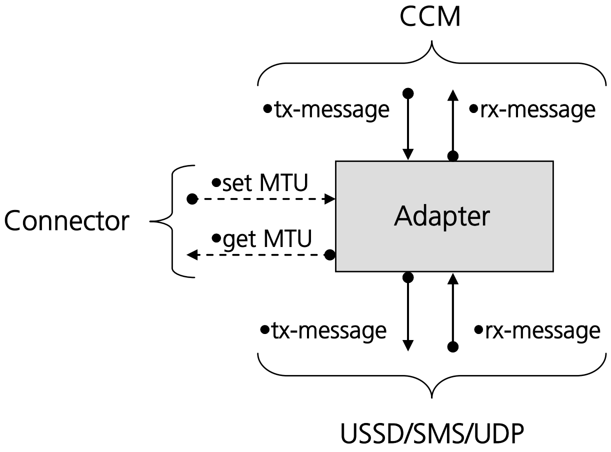


Figure 6.8: Adapter Related Functions

information and delegating the delivery of the messages to the CCM, which is in turn forwarding the messages to the appropriate Adapter.

6.3.4 Separated Signaling and Data Routing

The basic version of the CAHN concept relies on the cellular network, not only for the bootstrapping of the heterogeneous session but also for the monitoring of ongoing sessions. The reuse of the cellular signaling system is the simplest way to meet all requirements of a signaling channel. Basic functionality like authentication, paging, power management, and billing is already available and well established. However, to further improve the flexibility of the CAHN idea the signaling has not to be limited to the cellular network. To enhance the flexibility and further reduce the dependency on the cellular network, the signaling messages have to be routable over different channels. The architecture and protocol structure presented in this thesis is offering the required modularity to extend the system to cope with any type of signaling channel. The Adapter is the only component that has to be familiar with the actual channel used to transfer the signaling messages. Therefore, the mapping of the logical addresses to the physical addresses and the potentially required fragmentation is handled by the Adapter. Consequently, it is rather simple to map the CAHN messages to any available IP connection if only the IP address of the peer node is known. This enables inband signaling for further signaling as soon as an IP session has been established. To assure robustness of the signaling plane, both the cellular and the IP signaling channels can be used simultaneously. The different available signaling channels can be derived from the table of identifiers (6.2.2). The introduction of dedicated Adapters for each signaling channels allows a separate routing of the signaling messages. In contrast to standard routing mechanisms where the path is selected based on the destination address, the routing decision for the signaling messages happens before the actual IP routing. Depending on the state of the table of identifiers (i.e. the availability of signaling channels), the CCM selects the most appropriate Adapter. To allow the CCM to switch signaling channel for ongoing signaling sessions, there no session awareness at

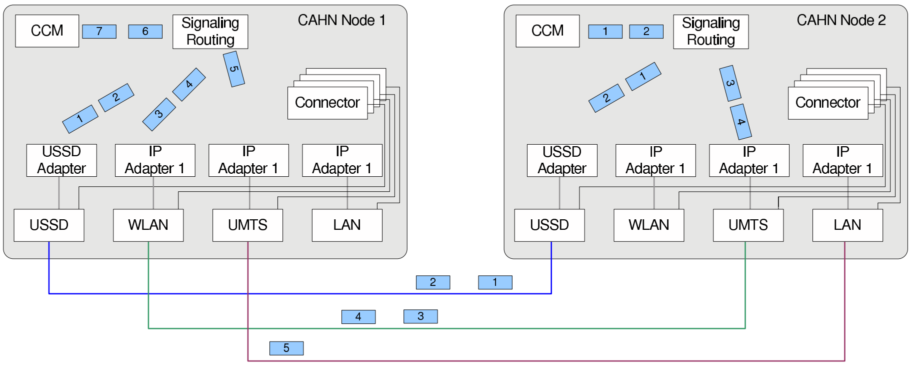


Figure 6.9: Routing of Signaling Messages

the different Adapters. Fig. 6.9 is illustrating the routing of signaling messages. The CCM of the node 1 forwards the signaling messages to the signaling routing which is dispatching the messages to the different Adapters according to the actual signaling channel selection. On the receiving node 2 the incoming signaling messages delivered by the Adapter are forwarded in the right order to the CCM.

Due to the CCM and the Adapters, the routing of the CAHN signaling messages can be handled without requiring any adaptation to the IP routing mechanisms. The IP routing takes care of the appropriate delivery of the UDP packets created by the Adapter in case of IP capable channels, and is not aware of any CAHN signaling session.

The flexible routing of the data packets is a little bit simpler because of the existing mechanisms provided by the Mobile IPv6 Route Optimization (see Section 2.5.2). The CCM can easily trigger the MIP to handover to a specific data channel. Hence, whenever an ad-hoc or direct link has been established by the CCM, a MIP route optimization is triggered. In the case that link-local addresses are used for the direct link, the *alternative CoA* descriptor has to be used for the binding update prior to the route optimization.

Thus, the selection of the communication channel to be used for the signaling can be done independently of the data routing. This allows the introduction of the presented separated signaling and data planes for heterogeneous sessions, without requiring changes in the existing communication stack. Considering the fact that SMACS/CAHN is handling all security related information prior to the actual optimization of the data path, it would even be possible to realize a simple tunneling between both communicating nodes also for IPv4. The only required information that has to be exchanged between the peers is the new communication address (e.g., collocated CoA of MIPv4). Due to the fact that this information is exchanged through the established secure link of SMACS/CAHN, no further verifications like performed with the HoTi/CoTi messages are required. However, the ability to rely on MIPv6 and its well defined route optimization feature eases very much the adoption of SMACS/CAHN.

6.4 CAHN Protocol

As mentioned already before, all communications between the layers are based on standard sockets. Therefore, the usage of messages to pass information between the layers is very convenient. Having the internal communication message based, makes the implementation of the system very modular. CAHN messages can be structured exactly the same way independent if the recipient of the message is local or remote. This symmetric structure is also very well reflected in the state machines, required to implement the defined protocol. In Section 6.4.8 the different state machines are explained in detail.

6.4.1 Message Structure

All messages are structured the same way. There is a mandatory part, which is mainly used to guarantee correct delivery and processing of the message. The optional part strongly depends on the type of message and the context, in which the message is sent. The following fields are mandatory for all types of messages:

- Source Address
- Destination Address
- Address Type
- Message ID (Signaling Session ID)
- Message Length
- Message Type
- Message Payload
- CRC

The source and destination addresses can be either IPv4, IPv6 or E.164 based, which has to be specified in the Address Type field. The future extension of the system to support further addressing schemes requires no changes to the message format. The message ID is used to assign the message to the correct signaling session. Thus, there can be several ongoing signaling sessions at the same time between two nodes. Sequence numbers are not needed since the different states of the signaling system and the message type are uniquely identifying the state transition (see Section 6.4.8). The message length is used for correct memory allocation. The type of the message is indicated in the message type field. And finally, the actual information of the message is transported in to payload field.

If the message has to be fragmented, the Adapter is adding a signal channel dependent header to the original message. This header is created according to the specific requirements of the selected signaling channel (i.e. USSD, SMS, UDP, etc.). To enable fragmentation (see Section 6.3.3) fragmentation related fields are included in the header. The *MSG-length* field is providing the overall message length and the *MSG-offset* indicates the offset of the current message fragment in bytes. Fig. 6.10 shows the message structure used to enable fragmentation.

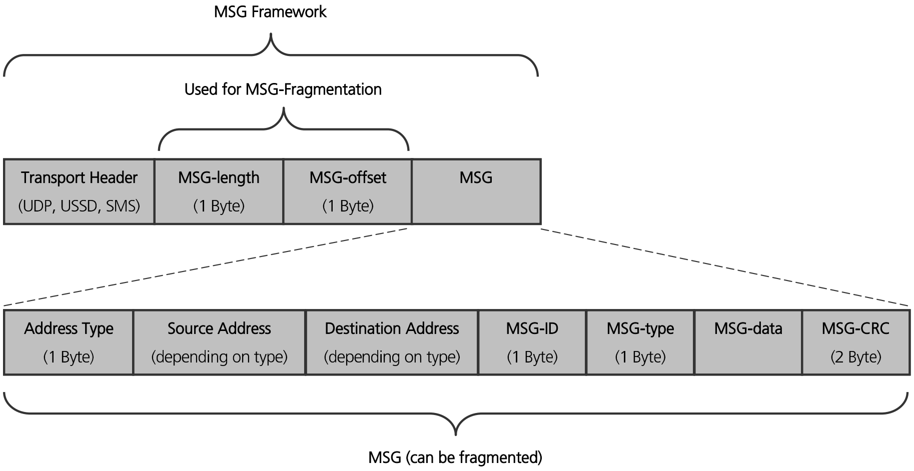


Figure 6.10: Message Structure

For sake of clarity, the specific messages are shown in the context in which they are used. Therefore, the messages are grouped in *session setup* and *teardown*, *network scanning*, *status reporting*, *trap setting* and *notification* and *identity exchange*. The messages required for the API are introduced in the Section 6.4.7.

6.4.2 Session Setup and Teardown

As briefly indicated previously, the session setup begins with the invitation of a peer. The SMACS layer triggers the CCM to form a complete *Connection Request*. The CCM together with the Connectors construct the *Connection Request* message and sends it using one of the Adapters, according to the routing for signaling defined by the SMACS layer (see Section 6.3.4). The request is created by the LSS (CCM) and the included connection and link related security parameters are provided by the PSS (Connector). Session related security parameters required for the end-to-end protection are handled by the SMACS layer (see Section 5.2.1). The *connection request* message is structured as followed:

Connection Request

The mandatory fields (address type, source address, destination address, MSG-ID, MSG-Type and MSG-CRC) are wrapping the actual message data (MSG-Data). Using Next Parameter fields indicating the type of parameter enclosed in the sequential field, all parameters are chained up in the message data part (see Fig. 6.11). Parameter Length is indicating the length of the parameter to clearly separate the different parameter sections.

The Next Parameter field consists of one byte defining the type of parameter. The following types are defined for the current system:

- **00: Network ID:** The name of the link (e.g., the SSID, in case of WLAN).
- **01: Network Dimension:** Number of host present on the link at the time the connection request is sent.

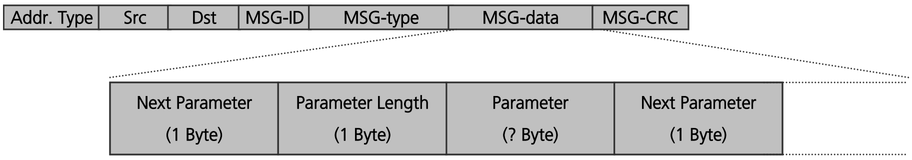


Figure 6.11: Connection Request

- **10: IPv4 Address:** Used to propose an IPv4 address (host address + netmask) the receiver should use for the interface configuration.
- **19: IPv4 Test Address:** Indicates the IPv4 address that has to be used to test the link after establishment. In most cases this might be the address of the inviting node.
- **20: IPv6 Network Prefix:** To form the IPv6 address if auto-address-configuration is used.
- **21: IPv6 Address:** Used to propose an IPv6 address (host address + network prefix) the receiver should use for the interface configuration.
- **29: IPv6 Test Address:** Same as the IPv4 test address, but for IPv6 connections.
- **30: Technology Vector:** The first few bits of the the vector are indicating, if a specific technology is available. Therefore, each bit is assigned to a specific technology, where a set bit represents the availability. The rest of the bits are binary coding the technology, whose configuration is proposed by the actual connection request (with respect to the order in the first part of the vector). Example: With a *Technology Vector* of one byte length the availability of 5 different technologies can be represented (first 5 bits, either set or reset) and with the remaining 3 bits the index of the actually proposed technology can be binary encoded.
- **31: Radio Frequency:** Frequency or Channel to be used for the corresponding technology.
- **32: Profile:** Parameter set in the case of pre-defined profiles. This might be of interest in the case of repeating link establishment, to avoid extensive parameter exchange.
- **33: Service:** Service to use. Some technologies might use services to interconnect nodes like Bluetooth (see Section 2.2.5).
- **40: Encryption:** Type of encryption and the cipher key.

Some of the parameter types are implicitly defining the length of the parameter field such as the **IPv4 Address** or the **Technology Vector**, and some might be defined as fixed length values like the **Network ID**, for example. Fig. 6.12 illustrates an exemplary *Connection Request.*

The CCM on the node of the invitee analyzes the received request, extracts the *Technology Vector* and compares it with the local vector, to verify whether there is at least one common technology available. If this is not the case, the

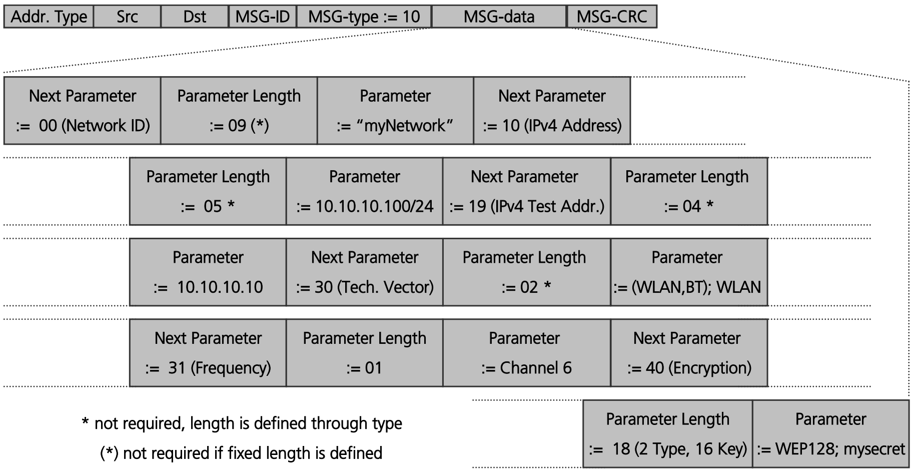

Addr. Type
Src
Dst
MSG-ID
MSG-type := 10
MSG-data
MSG-CRC
Next Parameter := 00 (Network ID)
Parameter Length := 09 (*)
Parameter := "myNetwork"
Next Parameter := 10 (IPv4 Address)
Parameter Length := 05 *
Parameter := 10.10.10.100/24
Next Parameter := 19 (IPv4 Test Addr.)
Parameter Length := 04 *
Parameter := 10.10.10.10
Next Parameter := 30 (Tech. Vector)
Parameter Length := 02 *
Parameter := (WLAN,BT); WLAN
Next Parameter := 31 (Frequency)
Parameter Length := 01
Parameter := Channel 6
Next Parameter := 40 (Encryption)
* not required, length is defined through type
(*) not required if fixed length is defined
Parameter Length := 18 (2 Type, 16 Key)
Parameter := WEP128; mysecret

Figure 6.12: Exemplary Connection Request

CCM is indicating this with an *Error Report* sent to the SMACS layer, which is then trying to setup a session based on infrastructure-based communication technologies. If the remote and the local *Technology Vector* match in at least on bit, there is a change to successfully establish an ad-hoc or direct node-to-node link between the nodes. Hence, the SMACS layer is prompting the user if the connection request should be accepted. Therefore, the LSS relevant information, like the identity (e.g., MSISDN) of the session initiator, is presented to the user. If the user accepts, the CCM is triggered to check further the proposed connection parameters. If the proposed parameters are acceptable, the local network interface is configured accordingly and a *Connection Accept* message is generated.

Connection Accept

The structure of that message is depending on the reliability of the signaling channel. In the case of a reliable channel there is no need to explicitly acknowledge all agreed parameters. If the channel is not considered as reliable the *Connection Accept* message includes a copy of the parameters proposed in the *Connection Request*. Fig. 6.13 illustrates the two principles, wherein the red part is optionally used in case of unreliable links.

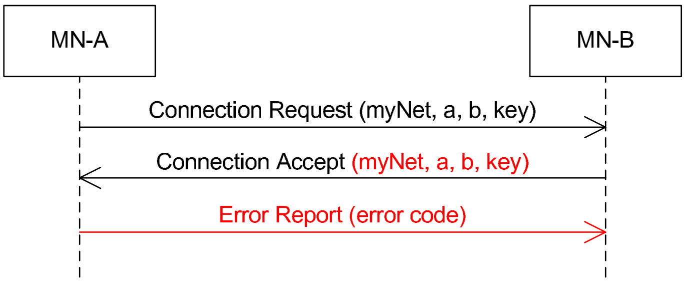
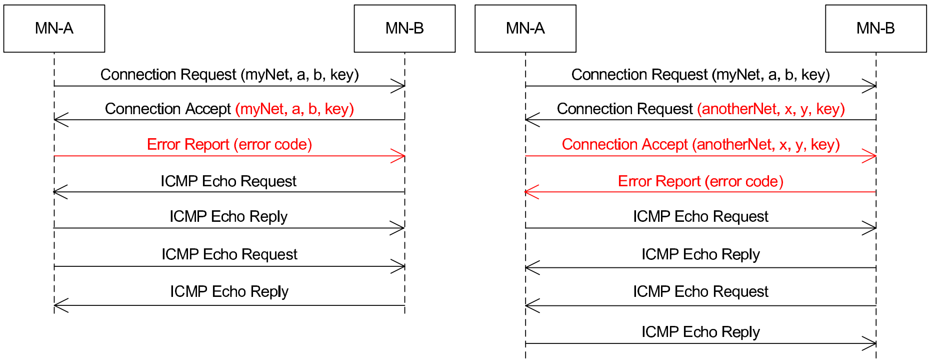

MN-A
MN-B
Connection Request (myNet, a, b, key)
Connection Accept (myNet, a, b, key)
Error Report (error code)

Figure 6.13: Connection Acceptance

If the proposed set of configuration parameters is accepted by the invitee, the connection setup is called simple. If not, the invitee can propose a new set of parameters by sending a *Connection Request* back to the initiator. Thanks to the extension based message structure, only the modified parameters have to be included into the new request. The initiator can then decide if he wants to accept the new parameters. The setup is aborted in the case of rejection and a new *Connection Request* has to be sent to restart the negotiations. In both cases, the simple and the advanced setup, the established link is tested with the help of ICMP echo requests and replies. The simple setup is depicted in Fig. 6.14(a) and the advanced setup, proposing a new set of configuration parameters, is shown in Fig. 6.14(b).


MN-A
MN-B
Connection Request (myNet, a, b, key)
Connection Accept (myNet, a, b, key)
Error Report (error code)
ICMP Echo Request
ICMP Echo Reply
ICMP Echo Request
ICMP Echo Reply
MN-A
MN-B
Connection Request (myNet, a, b, key)
Connection Request (anotherNet, x, y, key)
Connection Accept (anotherNet, x, y, key)
Error Report (error code)
ICMP Echo Request
ICMP Echo Reply
ICMP Echo Request
ICMP Echo Reply

(a) Simple Setup

(b) Advanced Setup

Figure 6.14: Connection Setup

Error Report

Whenever a node can not successfully configure its device according to the negotiated parameters it issues an *Error Report* with the corresponding error code. Errors can occur during the interface configuration, the CRC-check or if the target interface is already occupied by another session. The *Error Report* is identified by the MSG-Type 00. Depending on the error it might make sense to provide further information about the context in which the error occurred. Therefore, the *Error Report* message offers an optional description field. The structure of the message can be seen in Fig. 6.15.

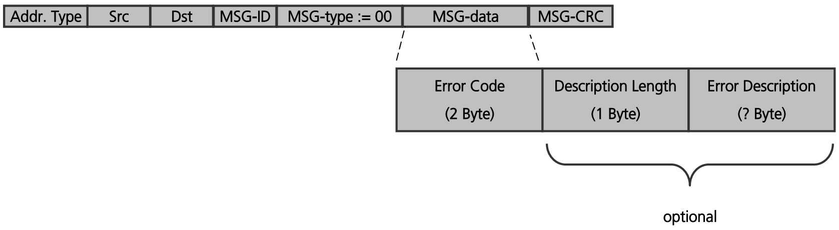


Figure 6.15: Error Report

The following error codes were defined to improve the recovery process, whereas the error code 0000 is indicating that there is no error and hence used for acknowledgements:

- **0000: No Error**
- **0001: Timeout**
- **0002: CRC**
- **0003: Invalid MSG-Type**
- **0004: Invalid MSG-ID**
- **0005: Interface Locked**
- **0006: Configuration Error**
- **0007: Connection Lost**
- **0008: Scan Error**
- **0009: Status Error**
- **0010: Request Rejected**

The error code 0010 is somehow special and introduced to allow the rejection of requests, without giving further reasons. Fig. 6.16 shows both, the simple and the advanced connection setup in the case of error occurrence.

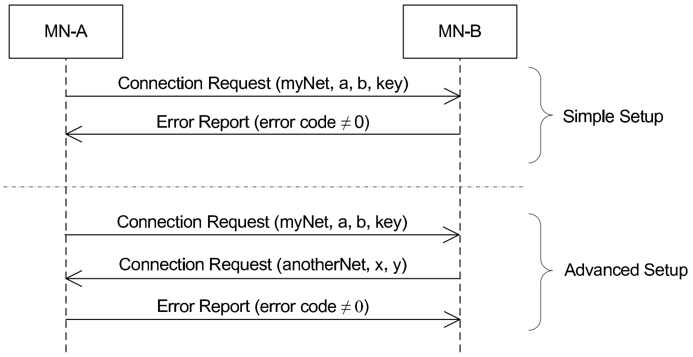

MN-A
MN-B
Connection Request (myNet, a, b, key)
Error Report (error code ≠ 0)
Simple Setup
Connection Request (myNet, a, b, key)
Connection Request (anotherNet, x, y)
Error Report (error code ≠ 0)
Advanced Setup

Figure 6.16: Faulty Setup

Disconnect

If a node wants to disconnect a link, it sends a *Disconnect* message to inform the peer node. Disconnect messages are sent over the interface that will be disconnected. In the case where multiple nodes are interconnected on the same link, the network ID and the identifier of the disconnecting node can be added to the message. This additional information can be included in the MSG-Data field, which is based on the Next Parameter structure, similar to the *Connection Request.* The complete *Disconnect* message is shown below. The MSG-Type is set to 21, only if the MSG-Data field is present. Otherwise, the MSG-Type has to be set to 20.

6.4.3 Network Scanning

Network scanning can be used to check whether a node is within the vicinity of the technology that is to be used for the connection. The scanning information might also be used to get environmental information of nodes not being within the vicinity.

Scan Request

The requesting node can ask for a specific network or node to be scanned for, or ask for a general scan. The scan result is then provided within the *Scan Report*. Both, the request and the report are structured similar to the connection request using Next Parameter fields. Fig. 6.17 illustrates the format of *Scan Request* message.

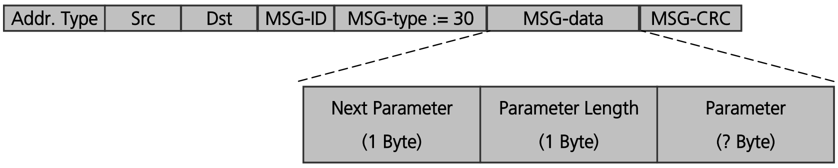


Figure 6.17: Scan Request

The MSG-Type for *Scan Request* is 30. The following list introduces the defined types of parameter:

- **00: Network ID:** The name of the ad-hoc network, e.g., the SSID in the case of WLAN.
- **01: Network IPv4:** The network address of an IPv4 network, e.g., 10.10.10.0/24.
- **02: Network IPv6:** The network address of an IPv6 network, e.g., FE80:0000:0000:0000:
- **20: Technology:** A technology vector indicating which technologies have to be scanned.
- **40: Host MAC:** MAC address of the host that has to be scanned for.
- **41: Host IPv4:** IPv4 address that has to be scanned for.
- **42: Host IPv6:** IPv6 address that has to be scanned for.
- **43: Hostname:** Hostname that has to be scanned for.
- **44: Host MSISDN:** If the node, requested to perform the scan, has already successfully resolved a communication address (e.g., MAC, IP) from the MSISDN of the searched node, the scan has to be done for those communication addresses.

Three examples of *Scan Request* messages can be found in Fig. 6.18.

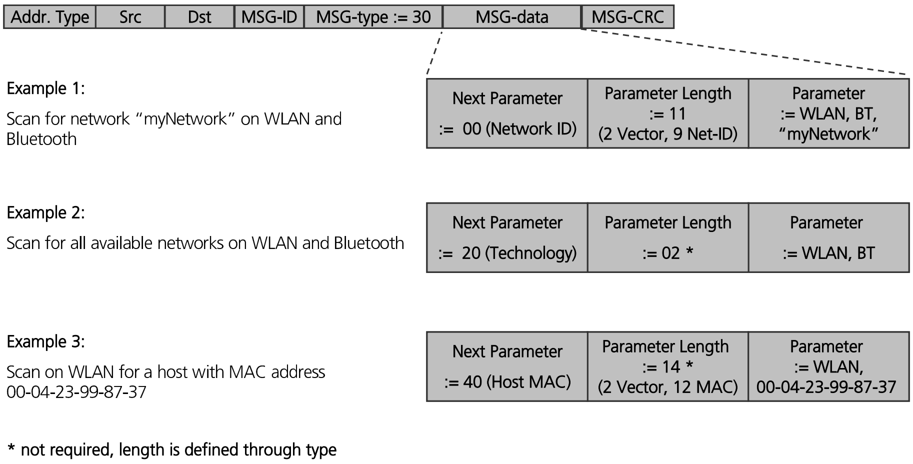

Addr. Type
Src
Dst
MSG-ID
MSG-type := 30
MSG-data
MSG-CRC
Example 1:
Scan for network "myNetwork" on WLAN and Bluetooth
Next Parameter := 00 (Network ID)
Parameter Length := 11 (2 Vector, 9 Net-ID)
Parameter := WLAN, BT, "myNetwork"
Example 2:
Scan for all available networks on WLAN and Bluetooth
Next Parameter := 20 (Technology)
Parameter Length := 02 *
Parameter := WLAN, BT
Example 3:
Scan on WLAN for a host with MAC address 00-04-23-99-87-37
Next Parameter := 40 (Host MAC)
Parameter Length := 14 * (2 Vector, 12 MAC)
Parameter := WLAN, 00-04-23-99-87-37
* not required, length is defined through type

Figure 6.18: Exemplary Scan Requests

Scan Report

Message type 31 defines the *Scan Report* message, whose MSG-Data section is structured very similar to the *Next Parameter* extensions used in the *Connection Request* and also in the *Scan Request*. The Parameter Length field is extended to two bytes to enable detailed reporting of the scan results. This might be important if lot of nodes are located in the vicinity of the radio interface performing the scan. Some examples of how MSG-Data fields could look like is shown in Fig. 6.19.

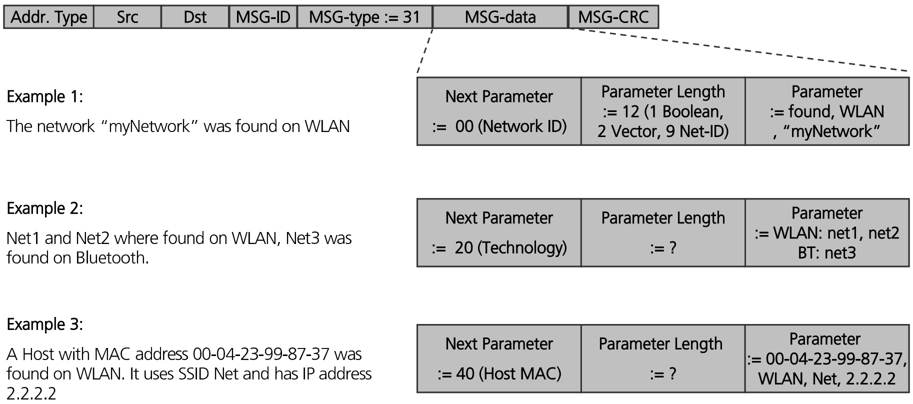

Addr. Type
Src
Dst
MSG-ID
MSG-type := 31
MSG-data
MSG-CRC
Example 1:
The network "myNetwork" was found on WLAN
Next Parameter := 00 (Network ID)
Parameter Length := 12 (1 Boolean, 2 Vector, 9 Net-ID)
Parameter := found, WLAN , "myNetwork"
Example 2:
Net1 and Net2 where found on WLAN, Net3 was found on Bluetooth.
Next Parameter := 20 (Technology)
Parameter Length := ?
Parameter := WLAN: net1, net2 BT: net3
Example 3:
A Host with MAC address 00-04-23-99-87-37 was found on WLAN. It uses SSID Net and has IP address 2.2.2.2
Next Parameter := 40 (Host MAC)
Parameter Length := ?
Parameter := 00-04-23-99-87-37, WLAN, Net, 2.2.2.2

Figure 6.19: Exemplary Scan Reports

Errors can be indicated with the help of an *Error Report* using the error code 0008 and the description field. If the requested scan is not performed due to privacy issues or because of any other secret reasons, the description field can be dropped.

6.4.4 Status Reporting

To enhance the reliability of the link management, collection of information about the environment, but also about the peering node itself, might be crucial. Radio signal levels are often asymmetric and therefore it is highly relevant to know the received signal quality at the peer. Only if such status information is regularly updated, an accurate management of heterogeneous sessions can be successful. Knowing exactly about the conditions at all communicating nodes increases the chance to execute a potential session handover to another technology at the right time.

Status Request

The *Status Request* is used to query information about a single interface, a specific technology, ad-hoc or direct node-to-node link or for all available interfaces. The message looks very much like a *Scan Request* but there are other parameters defined related to the status of the node and its communication technologies. As visible on the Fig. 6.20 the Parameter Length field takes also two bytes to allow rather detailed status reporting.

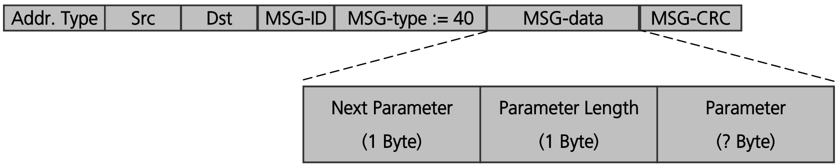


Figure 6.20: Status Request

Some of the defined parameter are rather of generic nature like the parameter with the number 40, which allows querying for specific capabilities of the peer. Similar is true for the parameter called *Services*. The idea behind those broadly applicable parameter declarations is to keep a maximum degree of liberty to use the status reporting, to get any kind of information that could be of interest and improve the system behavior. The possibility of querying the remaining power level of the peering node might also enhance the performance of the system. Routing decision for data and signaling data might be taken smarter if power constraints are taken into account as well. (ref power based routing) Here is the complete list of the defined parameters (for a detailed description of the parameters 00, 01, 02 and 20 refer to the Section 6.4.4):

- **00: Network ID**
- **01: Network IPv4**
- **02: Network IPv6**
- **20: Technology:** Indicates which technology should be reported.
- **40: Capabilities:** Indicates which capabilities should be reported.
- **41: Battery:** Queries the remaining battery energy level.
- **42: Services:** Indicates which services should be reported.

Some exemplary *Status Request* messages can be seen in Fig. 6.21.

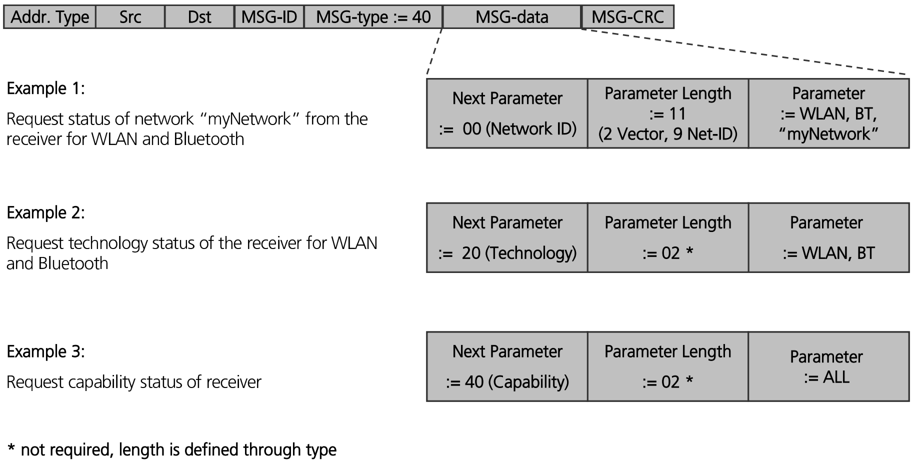

Addr. Type
Src
Dst
MSG-ID
MSG-type := 40
MSG-data
MSG-CRC
Example 1:
Request status of network "myNetwork" from the receiver for WLAN and Bluetooth
Next Parameter := 00 (Network ID)
Parameter Length := 11 (2 Vector, 9 Net-ID)
Parameter := WLAN, BT, "myNetwork"
Example 2:
Request technology status of the receiver for WLAN and Bluetooth
Next Parameter := 20 (Technology)
Parameter Length := 02 *
Parameter := WLAN, BT
Example 3:
Request capability status of receiver
Next Parameter := 40 (Capability)
Parameter Length := 02 *
Parameter := ALL
* not required, length is defined through type

Figure 6.21: Exemplary Status Requests

Status Report

If the queried node is willing to deliver the requested status information, it can reply with the *Status Report* message. The defined parameter types defined for the *Status Report* match the ones listed for the *Status Request*. Some examples how a report message could look like are presented in Fig. 6.22.

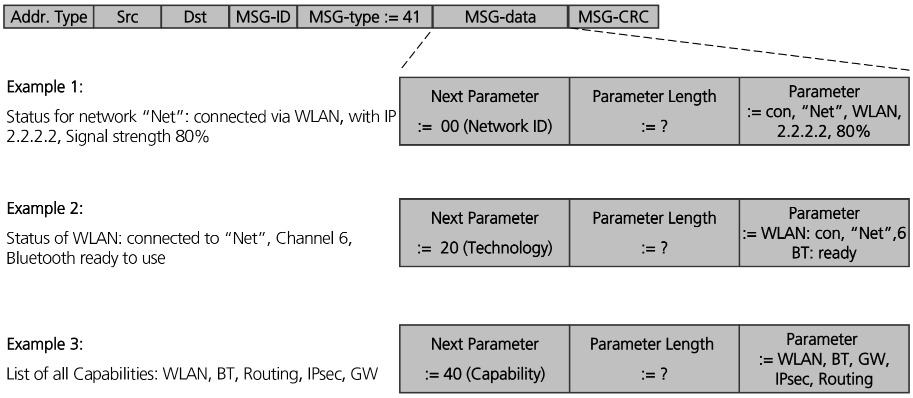

Addr. Type
Src
Dst
MSG-ID
MSG-type := 41
MSG-data
MSG-CRC
Example 1:
Status for network "Net": connected via WLAN, with IP 2.2.2.2, Signal strength 80%
Next Parameter := 00 (Network ID)
Parameter Length := ?
Parameter := con, "Net", WLAN, 2.2.2.2, 80%
Example 2:
Status of WLAN: connected to "Net", Channel 6, Bluetooth ready to use
Next Parameter := 20 (Technology)
Parameter Length := ?
Parameter := WLAN: con, "Net",6 BT: ready
Example 3:
List of all Capabilities: WLAN, BT, Routing, IPsec, GW
Next Parameter := 40 (Capability)
Parameter Length := ?
Parameter := WLAN, BT, GW, IPsec, Routing

Figure 6.22: Exemplary Status Reports

6.4.5 Trap Setting and Notification

Additionally to the status reporting functions, there is the possibility to set traps. Similar to the mechanisms in SNMP [71], the traps can be set to react on specific events or if certain thresholds are reached. The proposed support for traps and notifications is done on a modular and flexible basis using so-called *Trap Plug-ins*. These plug-ins can be defined and introduced without requiring any change to the CAHN protocol itself. The *Trap Request* and *Trap Notification* are offering a framework to initiate traps and exchange notifications between nodes. The actual handling of the traps is done by the plug-ins and therefore, to a certain extend, independent of the SMACS/CAHN system. It is also imaginable that plug-ins can be downloaded on-demand, if it is requested by a peer. However, to enable a differentiation between trap types already on the level of the CAHN messages, three classes were introduced: Network, technology, and host related traps.

Trap Request

To set a trap, a node can send a *Trap Request* to another node. A *Trap Request* is identified by the MSG-Type 60 and the data field is structured with the help of three different parameter blocks:

- **00: Network:** Defines network related traps.
- **20: Technology:** Used for technology related traps.
- **40: Host:** Indicates host related traps.

A trap can also be set locally. This allows the SMACS layer to get notified if a certain condition is fulfilled. The future introduction of trap plug-ins to monitor different conditions on the nodes, could considerably increase the efficiency of the signaling decisions taken by the SMACS layer (see Chapter 5).

The general structure of a trap request message is shown in Fig. 6.23.

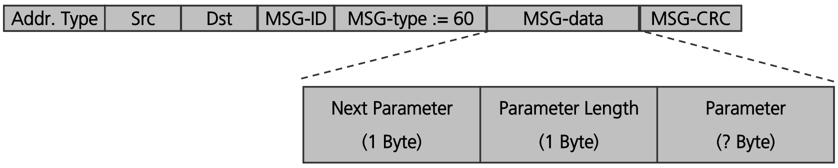


Figure 6.23: Trap Request

If the requested node accepts the *Trap Request*, it sends back an *Error Report* with error code 0000. From then on, the trap is set and continuously checking the corresponding condition.

Trap Report

Whenever a trap condition is fulfilled, a notification is sent to the trap owner. The notification is based on the *Trap Report* message and sent without any acknowledge. In contrast to *Status Reports*, the trap notifications are sent repeatedly until the owner disables the trap by sending the original *Trap Request* again. The report messages are structured as followed:

6.4.6 Identity Exchange

To enhance the user convenience of the proposed system, there is the possibility to request identity related information from the peer. Depending on the format of the identity provided, it can also be used to secure the session that has to be established. This is the case if certificates are exchanged. Other information types like the vCard are rather used for non-critical identity management (e.g., for address books).

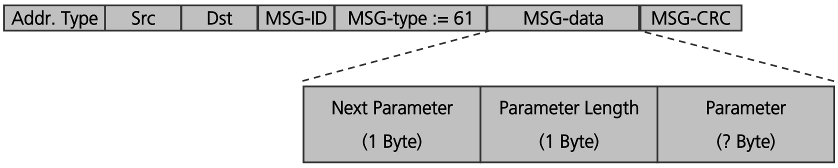


Figure 6.24: Trap Report

Identity Request

Hence, the *Identity Request* allows the exchange of further information about the user behind the peering node. Therefore, three different parameter types are defined:

- **00: XML:** If XML based identity information is requested.
- **10: vCard:** For vCard conform identities.
- **20: Certificate:** To get certificates from the peer.

For sake of fairness, the requesting node has to provide his own identity information in the *Identity Request*. However, the requested node can deny the request by sending an *Error Report* with error code 0010 (see Section 6.4.2). Fig. 6.25 illustrates how the *Identity Request* looks like.

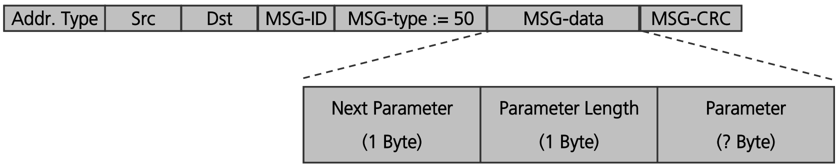


Figure 6.25: Identity Request

Identity Report

The *Identity Report* looks very similar to the request and is used to provide the requested identity information. The report can contain several types of parameters. It is up to the sender to decide which type of identity is provided. The type declared in the request is only indicating the preferred identity type of the requesting node. Due to the similarity of the request and report message the structure of the latter is not explicitly shown.

6.4.7 API Messages

The protocol messages and states defined in this chapter are mainly addressing the communication between CAHN enabled nodes. The interaction with the application layer (i.e. the user interface) is very much operating system dependent. Furthermore, the different user interfaces (e.g., terminal, GUI frameworks) are offering different possibilities to interact with. Hence, a comprehensive definition of the exact communication processes between SMACS/CAHN and the user is beyond the scope of this work. The SMACS layer is providing all required information about the ongoing session management so that it can interact with any type of APIs to fulfill a proper user interaction. However, for the basic concept of seamless and convenient heterogeneous networking the selection of the API framework is not relevant and therefore not further treated.

6.4.8 State Machine and State Transitions

After having introduced the different messages used to exchange the required information to successfully setup and maintain a link between nodes, the state machine of the protocol is presented in this section.

The complete CAHN protocol requires in total 22 states (client and server related). Fig. 6.26 shows the complete state machine of the CAHN protocol. The states on the left side are representing client related and the ones on the right side server related actions. Each CAHN node has to implement all states, since the system is designed to be fully symmetric. Every node can act as a client and server at the same time. This allows the simultaneous link establishment with multiple peers and technologies. The states marked with an asterisk are requiring user interaction. The arrows shown in the figure correspond to state transitions. Each transition is denoted with a number and a character: The number is the indicating the old state and the character is used for unique identification.

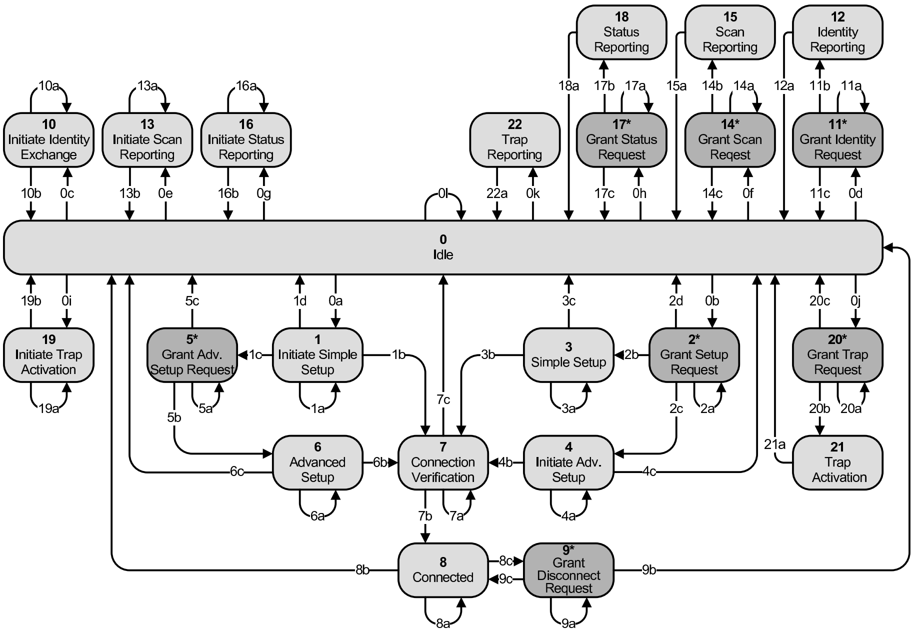


Figure 6.26: State Machine of the CAHN Protocol

State 0: Idle

For each state of the system, there is complete flowchart defined. Fig. 6.27 illustrates the flowchart for the idle state (state 0).

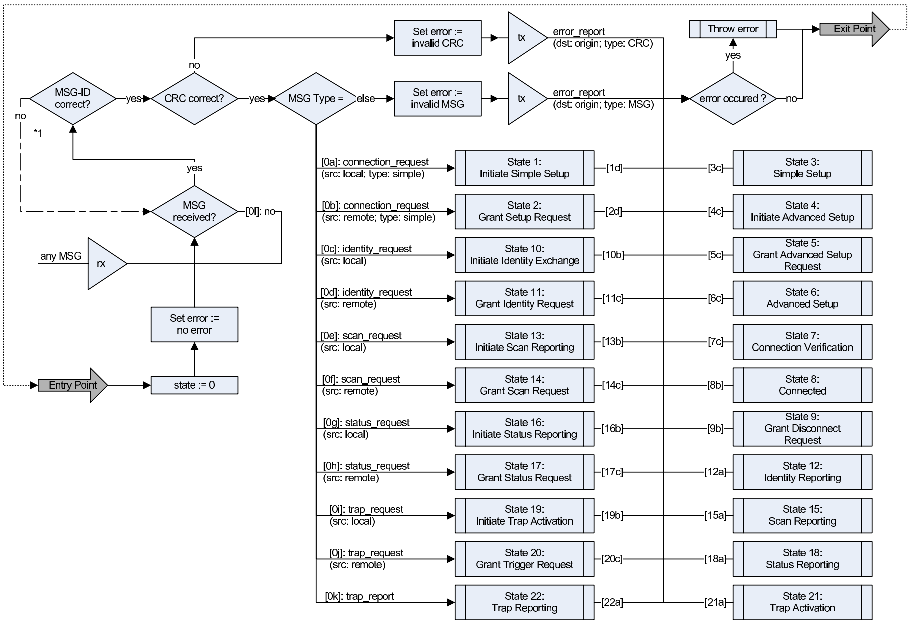


Figure 6.27: State 0: Idle

This idle state is used to analyze incoming messages, check for errors, and trigger the corresponding state transition. As long as there is no incoming message the system stays in the idle state, which is represented by the transition *0l*. If a message is received (indicated with *any MSG* at the rx symbol) the MSG-ID is checked. As mentioned earlier in this chapter the MSG-ID is used to distinguish different ongoing signaling sessions. In other words, there is a dedicated process running for every signaling session. Each of these processes can reach the idle state and is then waiting for incoming messages dedicated to it. As soon as a message is received, each signaling process is checking the MSG-ID, whether it is matching its own ID. The corresponding signaling process is taking care of the message and removing it from the rx queue. If there is no CRC error, the subsequent state is identified depending on the MSG-Type and the source of the message. In the case a received message fails the CRC test, the error type is set to *0002: CRC* and an *Error Report* is sent to the originator of the faulty message. Similarly, an appropriate error (*0003: Invalid MSG*) is reported if the MSG-Type is not matching one of the valid types for the actual state. For the state 0, for example, this is the case with message type 12 (*Advanced Setup Request*). To simplify and better coordinate the error handling, only state 0 is throwing exceptions. However, if an error occurs within another state, an error flag can be set indicating the type of error. Therefore, whenever a transition occurs from any other state to state 0, this flag has to be checked and if set, an appropriate error handling initiated.

State 1: Initiate Simple Setup

Whenever a local *Connection Request* is received, the system changes to *state 1: Initiate Simple Setup*. The structure of the actions belonging to state 1 are shown in Fig. 6.28. First, the state is set to 1 and the event message *Connecting* is thrown to indicate the connection establishment. Then, the initial *Connection Request* is forwarded to the remote destination and a timer is set. If the timer expires, the system error is set to *0001: Timeout* and an appropriate *Error Report* is sent to the remote node. The system then passes over to the idle state. If a message is received before the timeout, it's MSG-ID is compared with the signaling ID of the actual process. Messages with matching ID are further processed by analyzing the MSG-Type. The timer is stopped. Four message types are admissible at this stage; the *Drop* message coming from the local SMACS layer (potentially triggered by the user), the *Connection Request* (with *MSG-Type:12; Advanced*), the *Connection Accept* (with *MSG-Type:11; Simple*), and finally the *Error Report*. If the message received belongs to one of these types, the CRC is checked. The *Drop* and the *Error Report* messages force the system to transit to the idle state. The latter sets the system error according to the error type reported by the peer. The *Connection Request (Type 12)* leads to *state 5: Grant Advanced Setup*. When the peer accepted the connection request by sending a *Connection Accept* message, the system starts the configuration of the according communication interface. To avoid simultaneous access to the same interface by different signaling processes, each interface can be locked for other processes. Hence, an interface can only be configured if no other data session has already locked it. In the case, where the communication interface is not locked, the configuration process is trying to configure it, what is indicated by throwing an appropriate event (*Configuring Interface*). To assure that the link is really operational, the *state 7: Connection Verification* is triggered after successful configuration of the communication interface. In the case of unsuccessful configuration the system error is set accordingly. A timer is set before the configuration process (Connector), to assure liveness. After a certain timeout the interface is reset, the lock released, and the connection setup aborted.

State 2: Grant Setup Request

Fig. 6.28 (State 1) described the actions that have to be done if a connection request for a simple setup is sent to an invited node and accepted. State 2 defines what happens at the invited node receiving the connection request. The diagram on Fig. 6.29 shows the executed operations within *state 2*. The only way to enter the state 2 is by receiving a connection request (*MSG-Type:11 Simple*) at the idle state 0 (see state machine, transition 0b). The CCM forwards this connection request message to the SMACS layer of the local system, which is in turn using an API messages described in 6.4.7 to interact with the GUI and hence with the user. The decision taken by the user (or any other decision point attached to the SMACS layer in case of automation), is communicated to the SMACS layer. If no reply is received within a certain timeout, the setup is aborted and an *Error Report* (code 0001) is sent to the inviting node. The user can accept, reject or drop the proposed configuration parameter, or trigger the *advanced connection setup* by proposing a new set of configuration parameters. In most cases, the management of the advanced setup is done by the SMACS layer and not by the user itself. The advanced setup procedure is mainly introduced to enable a flexible connection management to avoid configuration conflicts (especially IP related), if there are several simultaneous sessions going on. The dropping is

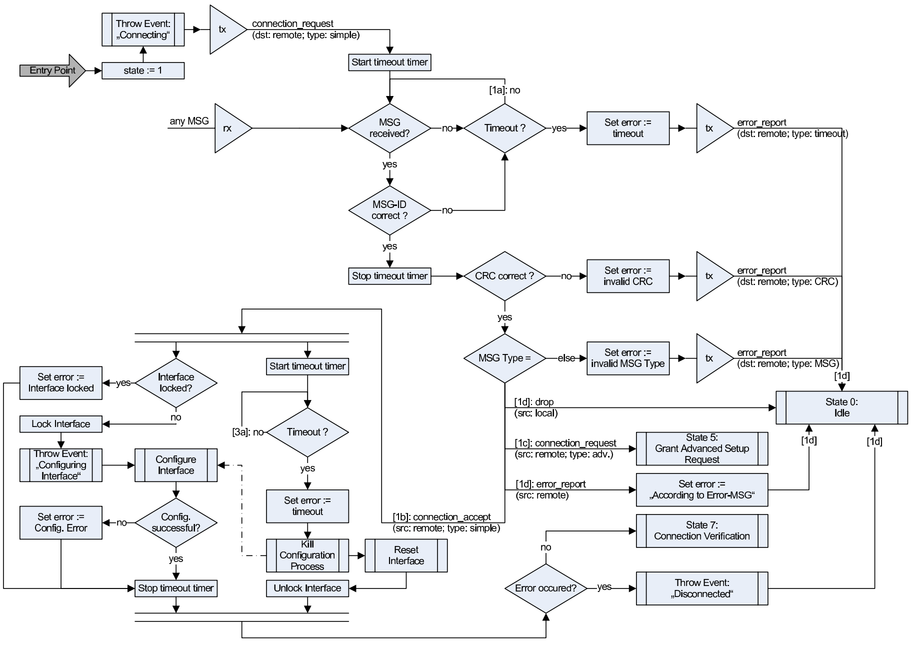


Figure 6.28: State 1: Initiate Simple Setup

done silently without notification of the peer by changing the system state to idle (0). The rejection decision is communicated by sending an *Error Report* (code 0010). Depending on the acceptance of the proposed parameter set, *state 3: Simple Setup* or *state 4: Initiate Advanced Setup* is initiated. Similar to the error handling described for state 0 and *1*, the CRC and type of the incoming messages are verified. Fig. 6.29 summarizes the actions defined within state 2.

All states requiring a user interaction are forwarding the requests to the SMACS layer, which is then communicating with the GUI. Depending on the decision taken by the user, different state transitions are initiated. However, the structure of these states is very much the same. They all have the part interacting with the SMACS/GUI to get the decision, on which the appropriate state transition can be initiated, and the standard message error handling. Therefore, the states 5, 9, 11, 14, 17 and 20 are not explicitly presented.

State 3: Simple Setup

State 1 handled the initiation of a simple setup, whereas state 2 brought about a decision whether the request should be accepted or not. In the case of acceptance of the request as is (with the originally proposed set of configuration parameters), the invitee has to fulfill the configuration of the interface accordingly. This is handled within the *state 3: Simple Setup* and the different actions involved are depicted in Fig. 6.30. Before configuring the interface, the *Connection Accept* message (*MSG-Type: 11 Simple*) is sent to the inviting node. This message is triggering the configuration of the interface of the inviting node as described in state 1.

The processes involved to configure the interface are equal to the ones used

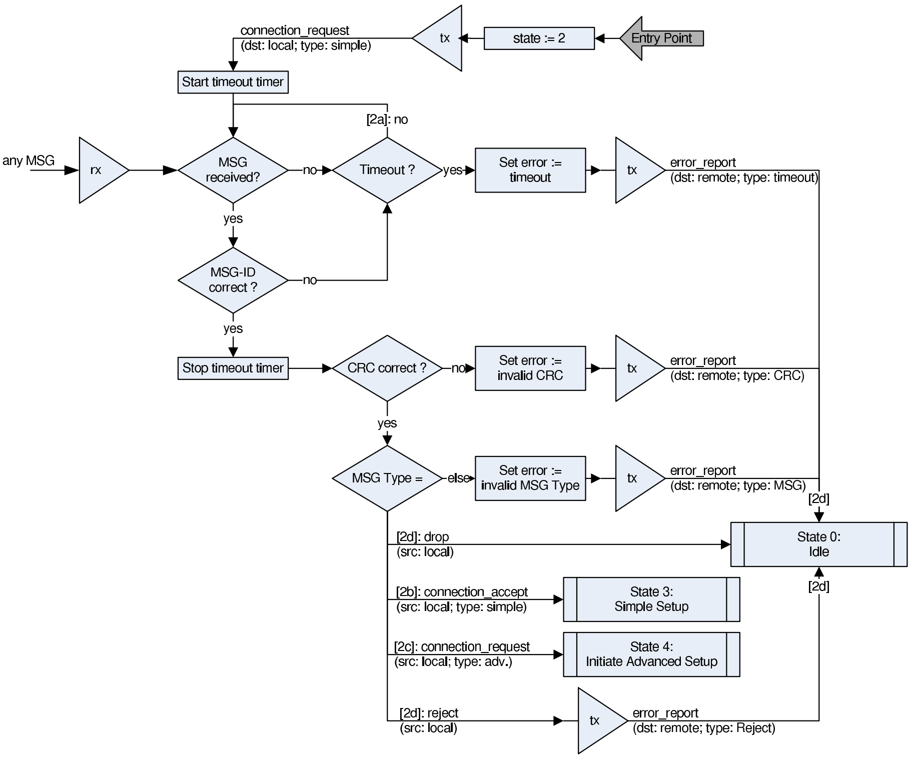


Figure 6.29: State 2: Grant Setup Request

within the *state 1: Initiate Simple Setup*, which is not really surprising. If the configuration is accomplished without any error or any timeout, *state 7: Connection Verification* is initiated. The same happens at the inviting node. With this verification of the actual connectivity, the connection establishment process is concluded and both nodes change into *state 8: Connected* until the connection is lost or disconnected on purpose.

State 4: Initiate Advanced Setup

The invited node can accept the invitation proposing a new set of configuration parameters by sending a *Connection Request* message with *MSG-Type: 12 Advanced*. After having sent the new proposal to the connection initiator, the invited node has to wait for a reply. Possible answers are *Connection Accept; Type Advanced* or *Error Report* in case of rejection of the new proposal or any other error indicated by the error code. Similar to the state 3, the corresponding interface is configured and the established link tested in the case of acceptance. If an *Error Report* is received, the system error is set according to the error code and the transition to state 0 (idle) is performed.

The drop event can terminate the connection setup process at any time and forces the system to cancel all actions and switch to the idle state. In contrast to the reject event, no notification is sent to the peer in case of dropping a connection setup. Fig. 6.31 illustrates the defined operations for state 4.

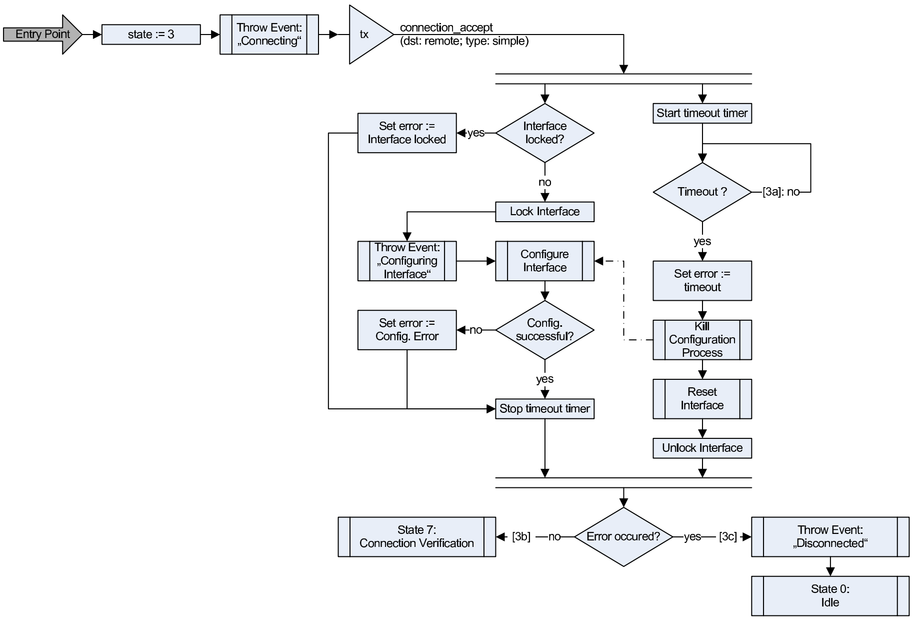


Figure 6.30: State 3: Simple Setup

State 6: Advanced Setup

If the request for an advanced setup, where the invitee is proposing a new set of configuration parameters, is accepted by the user (within *state 5: Grant Advanced Setup Request*), the inviting node changes to state 6. The advanced setup state mainly consists of the configuration of the interface according to the new set of parameters proposed by the invitee. Therefore, the state description looks very similar to the one of state 3, which is handling the simple setup if the initially proposed configuration is accepted. Nonetheless, Fig. 6.32 shows the details of state 6.

State 7: Connection Verification

Once the interfaces are properly configured on both devices the connection has to be verified. This is done using the ICMP echo request (aka ping). The connection verification is triggered after a simple or advanced setup, and leads to state 8 (connected), if the link is successfully tested and to state 0 (idle), in case of failure. A certain timeout is set when sending the ICMP echo requests to assure liveness. If no reply is received within this timeout, the interface is released and the *disconnect* event is thrown, followed by the transition to state 0. This behavior is represented by the flowchart on Fig. 6.33.

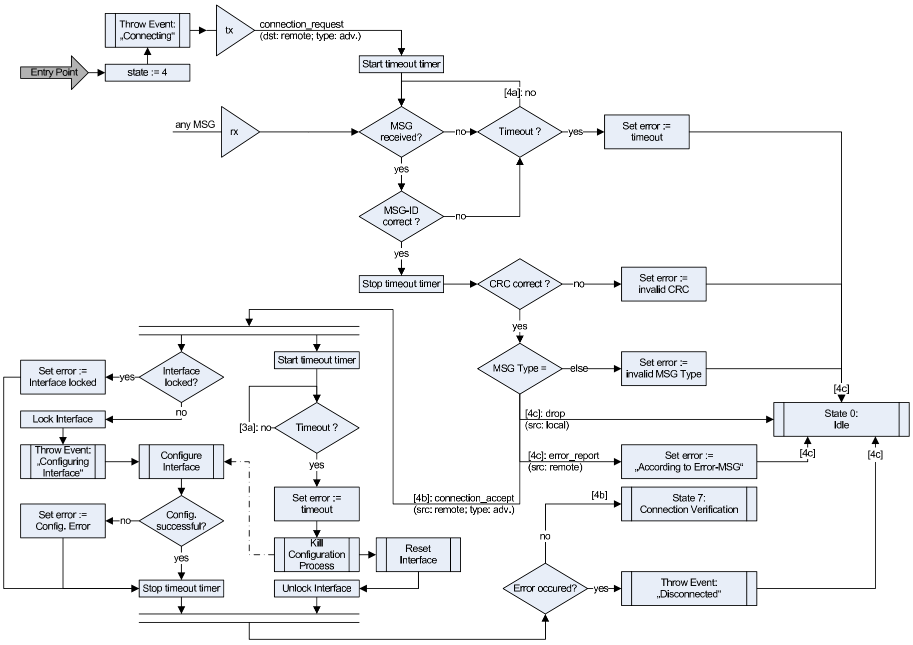


Figure 6.31: State 4: Initiate Advanced Setup

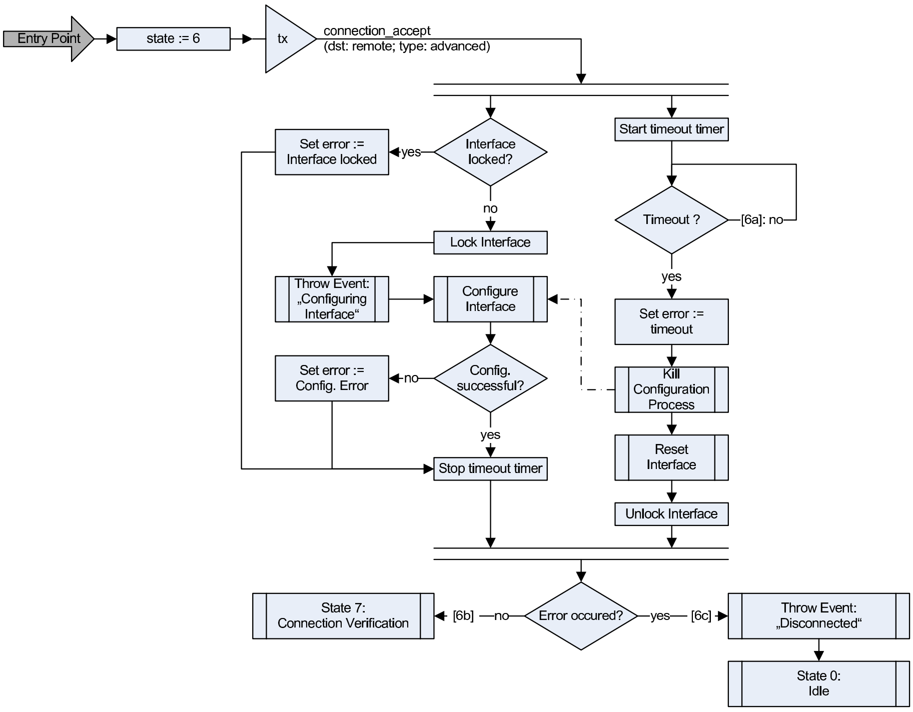


Figure 6.32: State 6: Advanced Setup

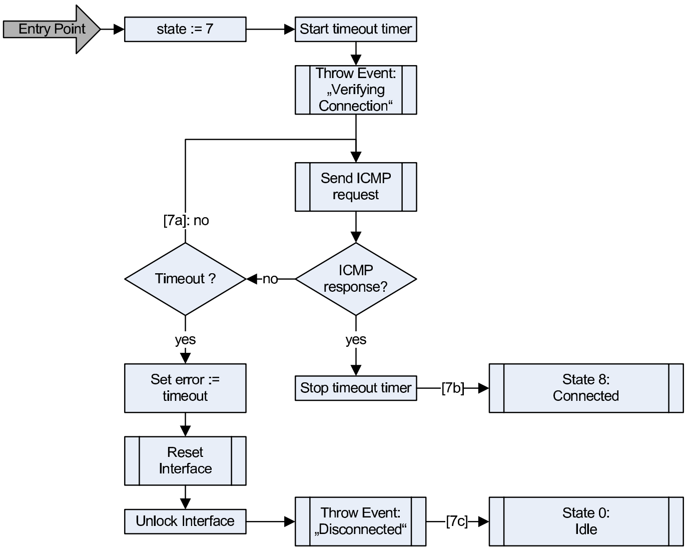


Figure 6.33: State 7: Connection Verification

State 8: Connected

Successful connection verification leads to the *state 8: Connected*. This state is responsible to monitor the established connection. After setting the corresponding system state, the status of the used interface is monitored until a disconnect message is received. A timer is started whenever the interface is down. If the interface is still down after the timeout, the connection is supposed to be lost and the system error is set to *Connection Lost*, the interface reset, and a *Disconnect* event is thrown followed by a transition to state 0. If the interface is up again before the timeout occurs, the timer is stopped and the monitoring of the interface continued. The received messages with the correct MSG-ID are verified (in terms of CRC) and depending on the MSG-Type triggering different actions. Drop messages coming from the local SMACS layer are forcing the system to immediately drop the connection by resetting and unlocking the used interface. After throwing a *Disconnected* event, the system switches over to state 0. Remote disconnect messages yield to the same behavior as local drop messages. If the source of the disconnect message is local (i.e. SMACS) the disconnect request is forwarded to the peer before the connection is dropped. Hence, the only difference between a drop and a disconnect is the notification of the peer in the latter case. Fig. 6.34 is summarizing the state description.

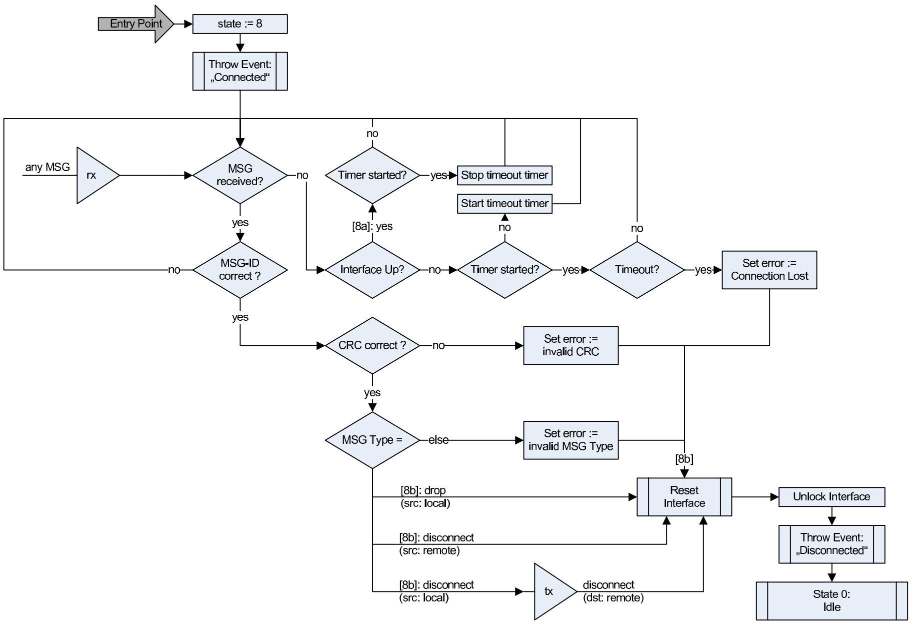


Figure 6.34: State 8: Connected

State 10: Initiate Identity Exchange

The identity exchange functionality is providing a framework to learn more about the peer. Therefore, an *Identity Request* can be sent to the peer. If no answer is received within a certain timeout, the system error is set accordingly and a transition to state 0 is performed. If the received message is an *Error Report* with error code 0010 (reject), a dedicated event is thrown to notify the initiating user. The other error codes are treated regularly. The peer can accept the request by directly sending back the requested identity information within an *Identity Report* message. The complete state actions are shown in Fig. 6.35.

State 12: Identity Reporting

The state used to sent an identity report is kept very simple. After receiving an identity request the *state 11: Grant Identity Request* is handling all required user interactions. Depending on the implementation, there is the possibility to select the file containing the identity information for each request or automatically sent a standard signature file. For the actual identity reporting state, it does not matter where the information to be transmitted comes from. As mentioned earlier, this identity information is not used by the system to authenticate the peer, it is only to enhance the user convenience. For sake of completeness, the state is depicted in Fig. 6.36.

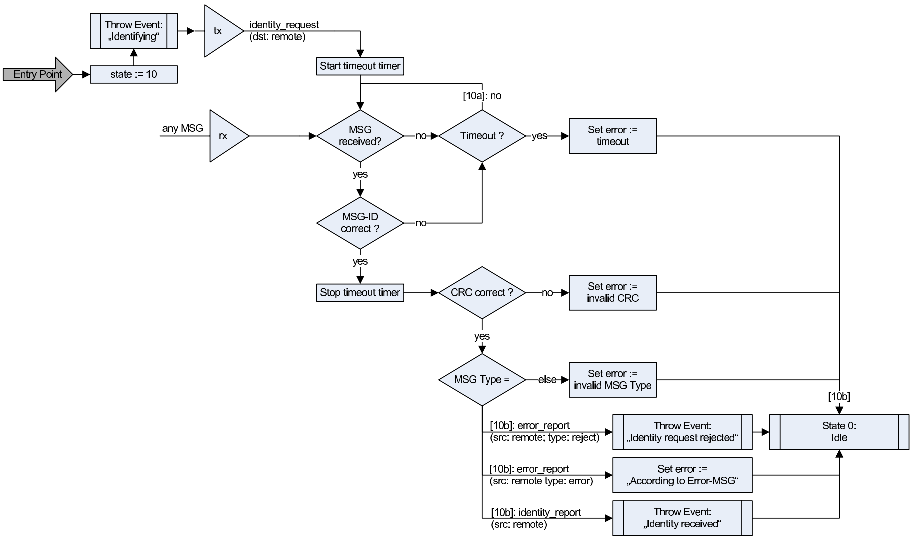

Throw Event: „Identifying"
tx
identity_request (dst: remote)
Entry Point
state := 10
Start timeout timer
[10a]: no
any MSG
rx
MSG received?
no
Timeout ?
yes
Set error := timeout
yes
MSG-ID correct ?
no
yes
Stop timeout timer
CRC correct ?
no
Set error := invalid CRC
yes
MSG Type =
else
Set error := invalid MSG Type
[10b]
[10b]: error_report (src: remote; type: reject)
Throw Event: „Identity request rejected"
State 0: Idle
[10b]: error_report (src: remote type: error)
Set error := „According to Error-MSG"
[10b]: identity_report (src: remote)
Throw Event: „Identity received"

Figure 6.35: State 10: Initiate Identity Exchange

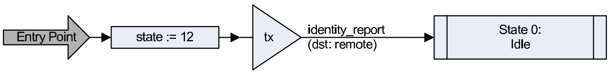


Figure 6.36: State 12: Identity Reporting

State 13: Initiate Scan Reporting

As mentioned earlier it might be helpful to know which nodes are within the range of the peering node. Especially, when thinking of route optimization using ad-hoc links whenever the peer is within the vicinity of any communication technology (see Chapter 5). To detect the peering node, *Scan Request* messages can be used. These messages are used to initiate both, local and remote scan reports. Hence, the CCM can receive a *Scan Request* from the local SMACS layer or from the remote CCM. However, if the request is coming from the local SMACS layer, it is granted anyway. But if the request is received from the remote CCM, the *state 14: Grant Scan Request* is getting the authorization from the user. Within the state 13 the *Scan Request* is forwarded to the destination being either a local *Connector* or a remote CCM. The processing of incoming messages is identical for both cases. After successful validation of the MSG-ID and CRC the MSG-Type is evaluated. An *Error Report* message with the error code 0010 is releasing a *Scan request rejected* event, whereas other error codes are setting the system error accordingly. If a *Scan Report* is received, the requester (i.e. the SMACS layer) is informed as well. The complete state activities are shown in Fig. 6.37.

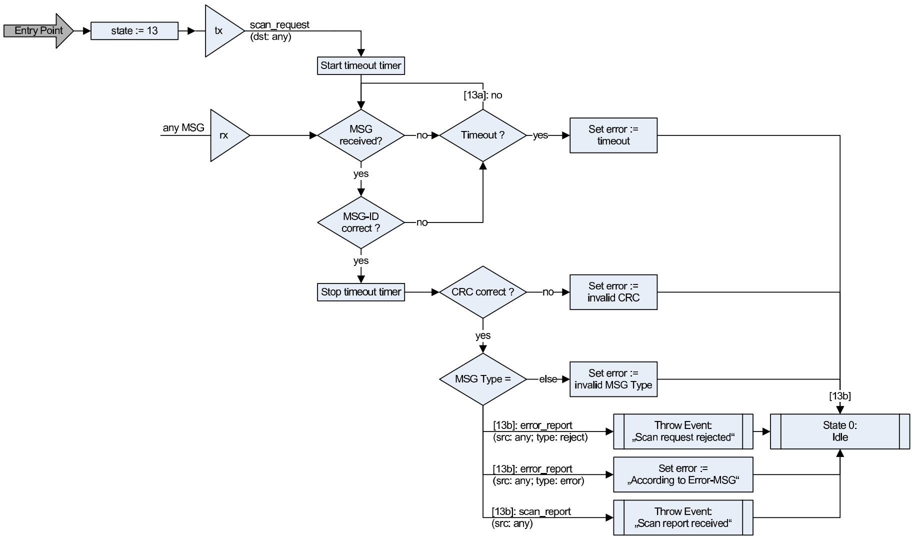


Figure 6.37: State 13: Initiate Scan Reporting

State 15: Scan Reporting

If a node accepted the *Scan Request* from either a local or a remote requester, it changes to *state 15: Scan Reporting*. Whenever a scan is started, a timeout is set in parallel to the scan process to assure liveness (see Fig. 6.38). If there is no scan result available after the timeout, the scan process is killed and the scan error is set accordingly. Both the scan results and the error reports are sent unacknowledged to the requestor.

State 16: Initiate Status Reporting

This state is defined like the *Initiate Scan Report* state and therefore not further elaborated. Since the CCM is only offering means to request and report status or scan information, it has no impact on the basic structure of that functionality. However the different states where introduced for sake of clarity. If further types of information is required to be exchanged within the CAHN protocol in future implementations, the introduction of a rather general reporting framework might be advisable. The differentiation of the type of reported information would then be indicated as a parameter within the general request and report message, instead of introducing separate messages types.

State 18: Status Reporting

The status reporting itself happens the same way as the scan reporting mentioned in 6.4.8. Both states call external functions to acquire the requested information, and deliver it to the initiator of the request.

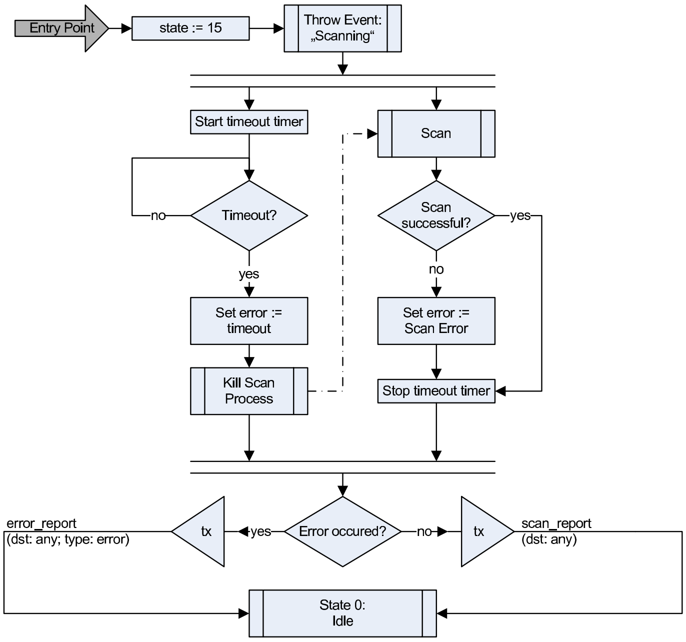


Figure 6.38: State 15: Scan Reporting

State 19:Initiate Trap Activation

The initiation of a trap is very similar to the initiation of a scan or status reporting. The requester of a trap activation is sending a *Trap Request* and waits for a reply within a certain timeout. In contrast to *state 13* and *16* a simple acknowledgement is expected indicating the acceptance or rejection of the request. The actual reporting is only done if the corresponding trap is released. Traps are active as long there is a connection maintained to the trap requester (aka trap owner) or until they are disabled. Hence, the receiver of the request replies with an *Error Report* either with code 0000 to indicate the acceptance or with code 0010 if the trap activation is not granted.

State 21: Trap Activation

Since the traps are based on plug-ins, the actual activation is handled by the external process responsible for the management of the traps. Therefore, from a CAHN perspective, the activation state looks similar to the *Scan-* and *Status Reporting*. The CCM informs the trap owner whether the plug-in is successfully installed and activated by this external process and switches to state 0. The actual trap notifications are triggered by the plug-ins, if the trap release conditions are fulfilled, and handled by the *state 22: Trap Reporting* introduced hereafter. Fig. 6.39 illustrates the procedure of the trap activation.

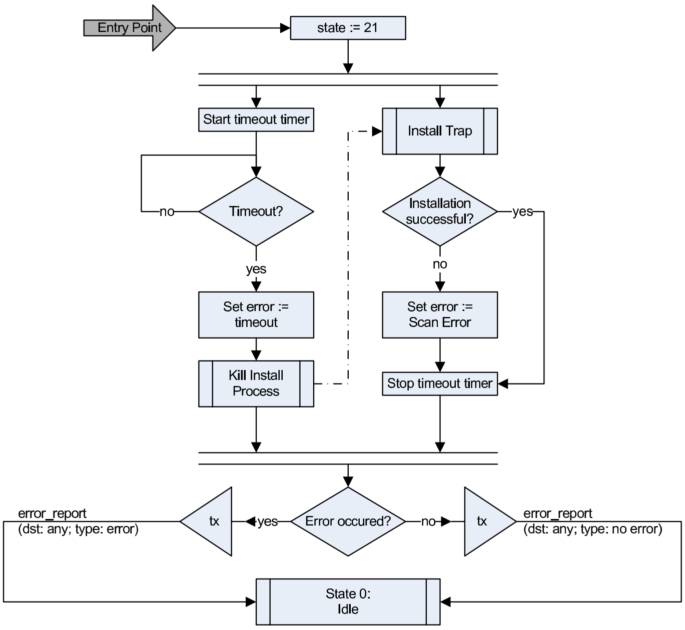


Figure 6.39: State 21: Trap Activation

State 22: Trap Reporting

If an active trap is released, the corresponding plug-in is triggering the CCM to switch to *state 22: Trap Reporting*, which is reacting different depending on whether the trap owner is local or remote. Local trap owners can be easily informed with the help of events, whereas remote owners have to notified sending the appropriate *Trap Report* message. The reporting is done without any acknowledgement and hence the CCM switches subsequently back to state 0, which also represented in Fig. 6.40.

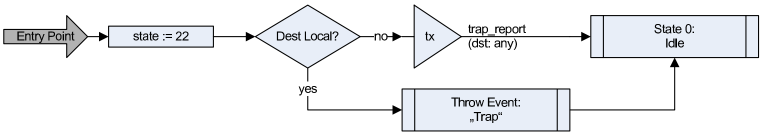


Figure 6.40: State 22: Trap Reporting

The states and transitions defined for the required communication between CAHN enabled nodes, allow the implementation of a flexible framework to securely establish ad-hoc and direct node-to-node links in a seamless and convenient way. Together with infrastructure-based seamless connectivity, it is possible to offer the real *Always Best Connected* experience using high performance and low cost direct links, whenever possible.

6.5 Connection Establishment with CAHN

This section addresses the deployment of the CAHN concept to securely establish infrastructure-less connections between nodes. We therefore focussed on the most known and widely distributed communication technologies, namely Bluetooth and WLAN.

6.5.1 CAHN for Bluetooth Networking

To integrate the provisioning of the MSISDN of the involved peers into the SDP of Bluetooth, a dedicated CAHN service (aka profile) has been designed and implemented. Bluetooth is offering a framework to scan for specific services within its radio vicinity. The human readable name of the service was set to *CAHN enabled device* to easily promote the ability to exchange sensitive data using the CAHN framework. Beside the service name, some additional service descriptors can be provided. To indicate the MSISDN as well, a service attribute of the type *String* was defined containing the MSISDN of the node providing the CAHN service. The scenario implemented and shown in Fig. 6.41 was called *Access Point Scenario*. It basically consists of a Bluetooth enabled node and a Bluetooth access point offering access to the Internet through the PAN service [94]. Both devices are CAHN enabled, which is illustrated with the mobile phone connected to the cellular network.

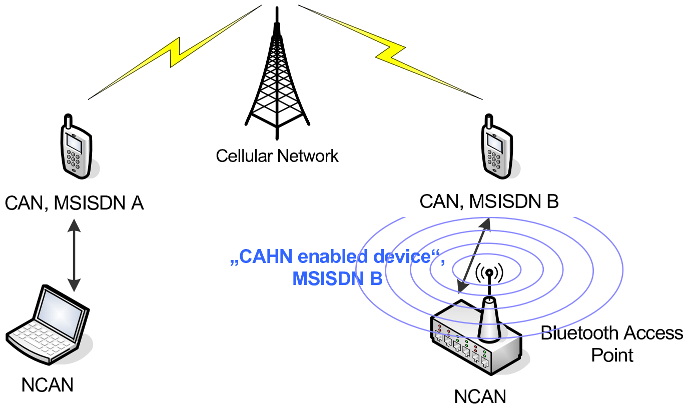


Figure 6.41: Access Point Scenario

The node is looking for Internet access and scans therefore its neighborhood for the Bluetooth PAN service. The access point is replying with the available services, including the CAHN service providing the MSISDN. From now on the service negotiation can be done using the secure cellular network. A CAHN *Connection Request* is sent to the MSISDN of the access point, proposing the use of Bluetooth to establish a connection. The access point enters the advanced setup sending another *Connection Request* including the proposed PIN to use. Using that PIN, the PAN service can be securely established. Beside the automation of the pairing the integration of CAHN authenticates both nodes against each other. The usage of the cellular link to deliver the PIN assures that both nodes provided the correct MSISDN, on which also the billing of the service might be done. It basically simplifies the pairing process of Bluetooth and may thus increase the level of security by assuring that the PIN is chosen randomly and never seen nor communicated by any one.

6.5.2 CAHN for WLAN Networking

When applying the CAHN based connection setup to WLAN links, the situation looks a little bit different because of the missing SDP. In contrast to Bluetooth, WLAN is not offering a built-in SDP implementation. Hence, there is no similar way to provide the MSISDN to the peer. Nevertheless, further investigations have been done and presented in [125] to integrate CAHN also with WLAN connections. The concept is basically the same as used for the integration with Bluetooth, but other means than SDP had to be found to promote the MSISDN to the peer.

Using the SSID to promote the MSISDN

The *Independent Basic Service Set* (IBSS) or *Ad-hoc Mode* of WLAN was designed for spontaneous and ad-hoc connections between nodes, and is therefore ideal for the the integration with CAHN. Wireless LAN uses the *Service Set Identifier* (SSID) to identify the different available networks. When starting an IBSS, the initiating node generates a random *BSS Identifier* (BSSID) and transmits it to other stations in the vicinity sharing the same SSID. Any node can listen for other SSIDs being broadcasted by other nodes. The SSID is a sequence of alphanumeric characters and has a maximum length of 32 bytes. Hence, the SSIDs can be used to promote the MSISDN. Therefore, every CAHN station announces its identity and CAHN ability by using the SSID. This works fine until two or more nodes decide to connect to each other. To communicate with each other, nodes have to join the same (I)BSS by associating with the (I)BSSID promoted by the initiator of the ad-hoc network. Unfortunately, the SSID of the network initiator has to be adopted as well, which forces the joining nodes to immediately stop promoting their MSISDN. Consequently, these nodes are not discoverable any more for other potential communication peers. In other words, the detection of the MSISDN, which is crucial for the efficiency improvement of CAHN, can only be done as long the node is not involved in any other (I)BSS. For small number of nodes joining the same (I)BSS, the SSID could be formed to include the concatenation of the MSISDNs of all involved nodes. However, due to the limited size of the SSID this approach is limited to about two to three nodes (depending on the length of the MSISDN). For the mentioned *access point scenario*, which is representing a typical client-server application, where the access point is providing connectivity to the Internet, it is not necessary to promote the client's MSISDN. If the access point is including its MSISDN into the SSID of the BSS, the clients can initiate the CAHN connection setup. Therefore, the limited number of MSISDNs that can be included within one SSID is not necessarily limiting the application of CAHN.

Coding Multiple MSISDNs with Bloom Filter

However, when addressing the direct node-to-node scenario, where all nodes can equally offer services and therefore have to announce their MSISDN, the situation looks different. Adopting the SSID of the node providing a requested service results in interruption of the provisioning of the own services. Thus, nodes can either act as servers or clients, but not both at the same time. Especially, when considering the route optimization offered by SMACS/CAHN, establishing direct links whenever possible, the announcement of the MSISDN becomes crucial as well. CAHN is initializing the direct link setup only if the destination node (i.e. its MSISDN) is detected within its vicinity. Therefore, nodes should be always detectable even if they are already joining another BSS. Whereas discovering a node's MSISDN is requiring the full MSISDN to be announced, the detection of a known MSISDN can be done with a lower level of scannable information. The scanning of fragments (i.e. some digits) of the MSISDN can already indicate the presence of that MSISDN with a certain probability. Consequently, if a non-zero probability of error is acceptable when scanning the SSID for a specific MSISDN, it is not necessary to promote the complete MSISDN. A *Bloom Filter* [15] is a space efficient, probabilistic algorithm that is used to quickly test the membership of an entity in a large set of data. The space efficiency (required due to the limited size of the SSID) is achieved at the cost of a non-zero probability of error. In case of error, the Bloom filter suggests that the tested element is member of the data set even if it is not. This error is also referred to as *false positive.* It has considerable impact on the signaling overhead of CAHN and is therefore further analyzed latter on.

Bloom Filter

Bloom filter use hash functions to reduce the volume of given information to a certain array of bits. Due to the characteristics of hash functions, the resulting message digest are not revealing anything about the initial data. When coding MSISDN with the help of Bloom filters, based on hash functions, the privacy can be guaranteed. The MSISDN of the searched node has to be known to successfully test if the node is member of the ad-hoc network (i.e. SSID). Some mathematical preliminaries might help to better understand the customization efforts required to optimally apply Bloom filters to CAHN. A Bloom filter is a method for representing a set $A = \{a_1, a_2, a_3, \dots, a_n\}$ of n elements (also called keys) to support membership queries in a m bits array. Therefore, a vector v of m bits is allocated, initially all set to 0. Furthermore, k independent hash functions $h_1, h_2, h_3, \dots, h_k$, each of them hashing uniformly to the range of $\{1, \dots, m\}$. It is important to use high quality hash functions to assure the hash values to be equally distributed over all possible values.

$$h_j : a_i \mapsto \{1, \dots, m\} \mid 1 \leq j \leq k, 1 \leq i \leq n$$

With regards to the length of the SSID, the vector v is defined to carry $256\,bits$ ($32\,bytes$). Hence, the selected hashes have to provide values between 1 and 256. For each element (e.g., key or MSISDN) $a \in A$, the bits at positions $h_1(a), h_2(a), h_3(a), \dots, h_k(a)$ in v are set to 1. A particular bit might be set to 1 multiple times, but only the fist change has an effect on the Bloom filter (OR):

$$\begin{aligned}\{(v_1, v_2, \dots, v_m)\,,\, h_j(a_i) = s \rightarrow v_s = 1 \mid \forall a_i \in A, \forall h_j, \quad & 1 \leq i \leq n, \\ & 1 \leq j \leq k,\end{aligned}$$

$1 \leq s \leq m\}$

To test if a MSISDN x is in A, the bits at positions $h_1(x), h_2(x), \ldots, h_k(x)$ have to be checked whether $h_j(x)$ are set to 1 for $1 \leq j \leq k$. If any of them is 0, then certainly x is not an element of the set A. Otherwise, if all bits $h_j(x)$ are set to 1, it can be assumed that x is in the set A. Depending on the parameters k and m, there is a certain probability for a *false positive.* A false positive happens if a particular bit is not set in the vector v:

$$\left(1 - \frac{1}{m}\right)^{kn} \approx e^{-\frac{kn}{m}}$$

And thus, the probability of a false positive with respect to the number of hash functions, the number of elements and the size of the array, is:

$$\left(1 - \left(1 - \frac{1}{m}\right)^{kn}\right)^{k} \approx \left(1 - e^{-\frac{kn}{m}}\right)^{k}$$

Consequently, the following three parameters are affecting the performance of a Bloom filter:

- Number of hash functions k
- Size of vector m
- Possibility of error $\left(1 - e^{-\frac{kn}{m}}\right)^{k}$

When applying Bloom filter to CAHN, the size of the vector m is given (SSID). Hence, for each number of nodes n, there is a number k, which minimizes the false positive rate $\left(1 - e^{-\frac{kn}{m}}\right)^{k}$. A high number of hashes increases the chance to find a 0-bit for a key (MSISDN), which is not member of the set A. On the other side, applying too many hash functions will increase the density of the Bloom filter, resulting in high probability of collisions (multiple hashes set the same bit of v). The optimal number of hash functions that minimizes the false positive rate (α) as a function of k, can be calculated as following:

$$\alpha = \left(1 - e^{-\frac{kn}{m}}\right)^{k} = e^{k \ln\left(1 - e^{-\frac{kn}{m}}\right)}$$

Hence, minimizing the probability of false positive $\left(1 - e^{-\frac{kn}{m}}\right)^{k}$ is equivalent to minimizing $k \ln\left(1 - e^{-\frac{kn}{m}}\right)$ with respect to k:

$$\frac{\partial}{\partial k} k \ln\left(1 - e^{-\frac{kn}{m}}\right) = \ln\left(1 - e^{-\frac{kn}{m}}\right) + \frac{kn}{m} \frac{e^{-\frac{kn}{m}}}{1 - e^{-\frac{kn}{m}}}$$

The function is minimized if the value of the derivation is 0, which is the case for $k = \frac{m}{n} \ln 2$. Since k is representing the number of hash functions used within the Bloom filter, it has to be an integer value. Furthermore, the number of hash functions has to be as small as possible to reduce the computational efforts required to apply the Bloom filter. So practically, $k = \lfloor \frac{m}{n} \ln 2 \rfloor$.

Finally, the selection of the optimal value for k results in a value of α which is:

$$\left(1 - e^{-\frac{kn}{m}}\right)^k = \left(1 - e^{-\ln 2}\right)^{\frac{m}{n}\ln 2} = \left(\frac{1}{2}\right)^{\frac{m}{n}\ln 2} = (0.6185)^{\frac{m}{n}}$$

If the size of the vector v corresponds to the size of the SSID the value m is fixed to 256, which results in defining k depending on the number of nodes n:

$$k = \frac{256}{n}\ln 2$$

performing with a false positive error rate α:

$$\ln \alpha = \frac{256}{n}\ln 0.6185 \Rightarrow \alpha \approx e^{-\frac{123}{n}}$$

With regards to CAHN, where the number of participating nodes of the ad-hoc WLAN network is known, the number of hash functions k can be quickly calculated by each node. However, to maintain a certain false positive, the nodes would have to calculate the corresponding number of hash functions whenever a node is joining or leaving the ad-hoc network. Fig. 6.42 is visualizing the effect of changing n on k:

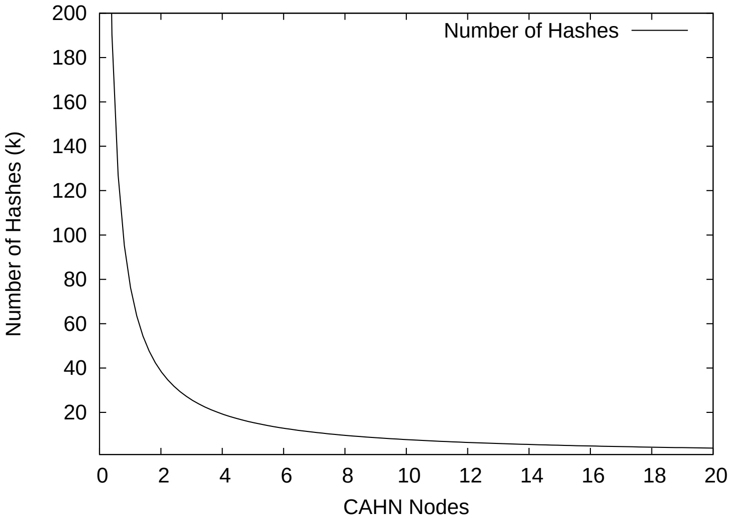


Figure 6.42: Optimal Number of Hashes

Recalculation of the SSID whenever a node is joining or leaving might be very inefficient, especially in highly dynamic networks. But choosing a fixed value for k is not appropriate neither, since the false positive error rate is very much depending on the number of nodes. Fig. 6.43 shows the false positive rate α under variation of the number of nodes n for $k = 18$:

Considering the possible consequences of a false positive error in the context of CAHN, a tradeoff between calculation effort and error rate is acceptable. Within the CAHN application Bloom filters are used to detect if a certain node is within the vicinity and thus if there is a chance to successfully establish a

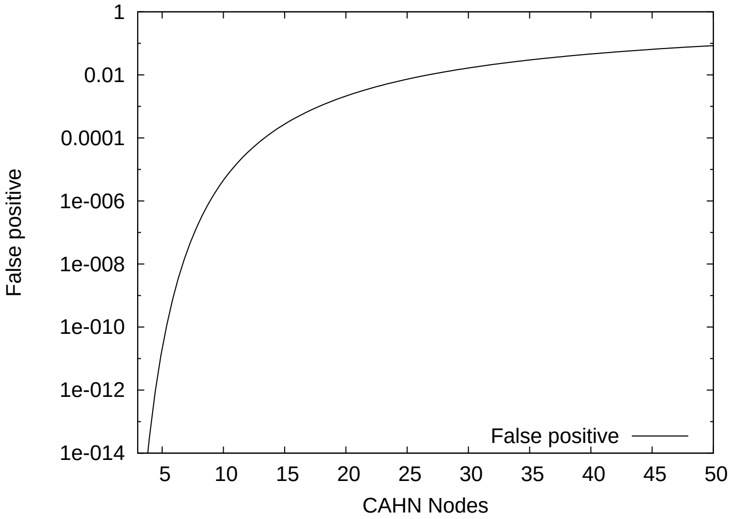


Figure 6.43: False Positive Rate as a Function of the Number of Nodes

direct link to optimize the data channel. If the testing of the scanned SSIDs results in a false positive, the CAHN signaling is started in vain. Introducing thresholds for the number of nodes within the same BSS considerably reduces the computation overhead by keeping an acceptable error rate. Table 6.1 shows the preferable values for k depending on the threshold for the number of nodes and the resulting false positive error rate.

Number of nodes n (Threshold)	Number of hashes k	False positive α
10	18	0.00001
15	12	0.0001
20	9	0.001

Table 6.1: Preferable Values for k Depending on n Thresholds

Using Bloom Filters for CAHN

The previous section showed that members of an ad-hoc network with up to ten nodes can use $k = 18$, resulting in a false positive error rate of about 10^{-5}, which is largely acceptable for CAHN. Hence, $k = 18$ is used a default value for each CAHN node. Initially, each node calculates its SSID using the same set of hash functions. Applying the hashes to MSISDN indicates which bits of the SSID have to be set to 1. Node B, for example, sets the bits number 6, 7, and 9 of his initial SSID accordingly (see Fig. 6.44).

The Bloom filter values for the peer node (here *Node B*) is calculated as well. To detect if *Node B* is within the vicinity to set up a direct link, the found SSIDs are analyzed. Therefore, the bits set by the Bloom filter value of *Node B* are tested in the scanned SSIDs. If the test is successful, the CAHN connection setup is initiated. Furthermore, if only the bits corresponding to the Bloom filter value of *Node B* are set, the node is not involved in any other

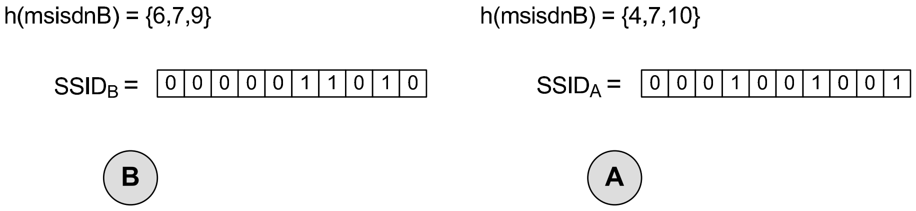


Figure 6.44: Initial SSID Calculation

ad-hoc network yet, since other MSISDNs result in setting other bits of the SSID. After having negotiated all necessary parameters, the SSID of the new ad-hoc network is set accordingly to represent both Bloom filter values. This is done by combining both values with the *OR* operation. Fig. 6.45 illustrates how $SSID_{AB}$ is formed based on $SSID_A$ and $SSID_B$.

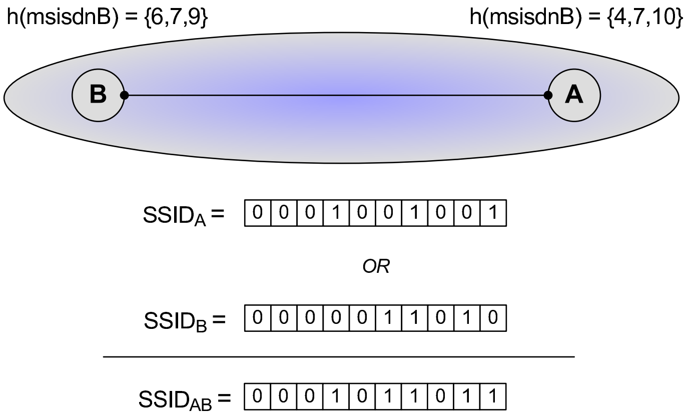

h(msisdnB) = {6,7,9}
h(msisdnB) = {4,7,10}
B
A
SSID_A = 0 0 0 1 0 0 1 0 0 1
OR
SSID_B = 0 0 0 0 0 1 1 0 1 0
SSID_AB = 0 0 0 1 0 1 1 0 1 1

Figure 6.45: Calculation of the Combined SSID

If the detected *Node B* is already member of another ad-hoc network and *Node A* should become member of that network too, all nodes have to update the SSID to include the Bloom filter value of the joining *Node A*. The SSID has also to be adapted if a node is leaving the ad-hoc network. Otherwise, the next test will turn into a false positive error because of the legacy bits set in the SSID.

SSID Management

The proper management of the joining and leaving process for ad-hoc networks involving more than two nodes is very important to guarantee a low probability of node detection error. After each join, the new SSID has to the distributed to all participating nodes. When a new mobile node (slave node) enters the CAHN ad hoc network, the corresponding Bloom Filter needs to be updated by setting the matching hashed bits to one (see Fig. 6.45). Therefore, the peer node (referred to as master node) calculates the the new Bloom Filter and sends it to the other participants of the network using an update message via WLAN. The same has to happen if a node is leaving the network. Unfortunately, the corresponding bits in the Bloom filter can not be just reset to 0 because other

MSISDNs may be hashed to some of those bits, which would result in resetting bits of still existing participants. To avoid this problem, a counting Bloom filter has to be used. Each node in a CAHN ad hoc network has a list of stations. This list contains information about the number of currently participating CAHN nodes and the MSISDNs of those stations which are directly in its vicinity. Whenever a new node joins the network, its peer updates the list by adding the MSISDN of the new node and incrementing the number of nodes. It is important for all participant of the ad hoc network to know how many nodes belong to the same ad-hoc network because the choice of the number of hash functions depends on the number of nodes. To minimize the false positive error rate, the SSID has to be recalculated with a different number of hashes whenever a new station joins or leaves the network. To avoid too much calculation overhead due to continuous recalculation of the hash, a threshold value was defined. Fig. 6.46 illustrates the SSID distribution in a CAHN ad-hoc network. Node A wants to join the network and its peer node B sends the update message to its neighbors, which further distribute the new SSID. When changing the SSID the connections between the stations are broken until all nodes have successfully changed their SSID. To keep that outage as small as possible, the nodes send also an update acknowledge message back signalizing that they are ready to change their SSID. If node B receives the acknowledgement, it sends the CAHN response to the node A including the new SSID to use.

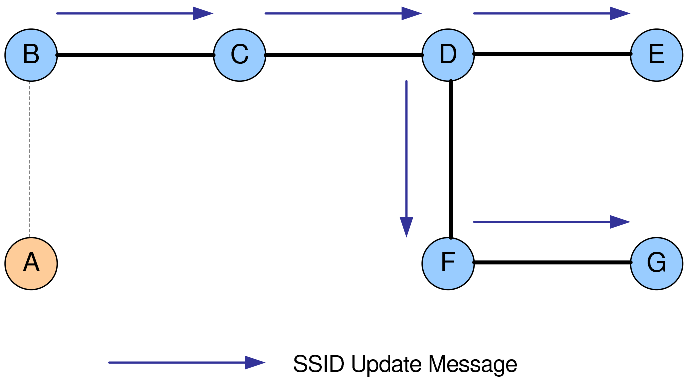


Figure 6.46: SSID Distribution in a CAHN Ad-Hoc Network

More details about the management of the SSID can be found in [125].

6.5.3 CAHN for Spontaneous Networking

When thinking of spontaneous networking applications like file exchange, synchronization, and ad-hoc collaboration tools, CAHN can also help to considerably increase the security and convenience level. The variety of communication technologies available in nowadays devices enabling spontaneous networking motivates the deployment of concepts simplifying the handling for end users. Therefore, this section concentrates on those kinds of applications relying on spontaneous networks, and the benefits of introducing CAHN to offer a simple and secure connection management.

Service Discovery

When addressing spontaneous networking applications, service discovery is crucial to promote, find, and share services. *Service Discovery Protocols* (SDP) aims at providing means to scan the neighborhood for services offered by other nodes. Originally addressing rather fixed networking environments, most of the efforts in the domain of service discovery are assuming friendly or secured environments. This makes perfectly sense when addressing common use cases, where the nodes are supposed to be authenticated before accessing the network. Especially corporate networks are supposed to be protected and consequently the service discovery process can focus on the management of services, instead of caring about security issues. But when considering rather open environment like spontaneous and ad-hoc networks, the security assumption becomes weak. In [192] we analyzed the different SDPs in terms of security and efficiency when applied on top spontaneous and ad-hoc networks. The work explored furthermore the integration of CAHN and SDP. CAHN is basically offering the secured signaling plane required to securely exchange SDP messages. The session is established between specific nodes (or users). The session setup starts with the selection of the peer and after successful connection establishment the available services can be discovered. Hereby, the efficiency can be increased by limiting the service discovery effort to one or a few dedicated nodes. Fig. 6.47 visualizes this destination centric service discovery.

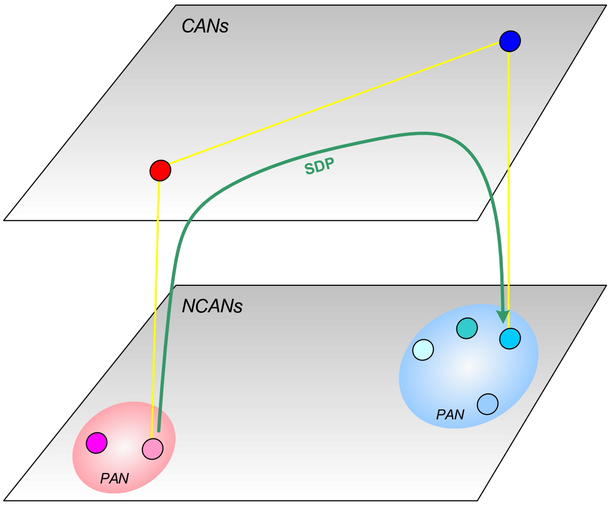


Figure 6.47: Destination Centric Service Discovery

This optimization of the service discovery process is automatically available when using SMACS/CAHN to setup IP sessions between nodes. The SMACS/CAHN system only sets up protected links between the invited nodes. These links form a kind of virtual and private overlay network among the involved nodes. When starting the SDP on top of that overlay network, only the involved nodes are visible for the service discovery. However, the typical use cases look a little bit different when focusing on the service discovery aspect. In the standard use case for SDP, the nodes search for a specific service and resolve the peer that is able to provide that service. Even if it is not crucial for the service discovery process itself to know who is finally offering the requested service, it is important to authenticate each other, especially when it comes down to billing. The combination of CAHN and SDP proposed in [192] is very straightforward, by extending the SDP to provide the MSISDN of the negotiating peers during the service discovery process. Knowing the MSISDN of the node providing the requested service, the sensitive data is then exchanged using the secured signaling channel offered by CAHN. In the case of Bluetooth the security mechanisms are based on a so-called PIN, which has to be provided during the *pairing* process. To setup a secured link between nodes, the same PIN has to be provided to all participating nodes and is then used to derive the shared secret used for symmetric encryption. Fig. 6.48 illustrates the described automated PIN exchange process.

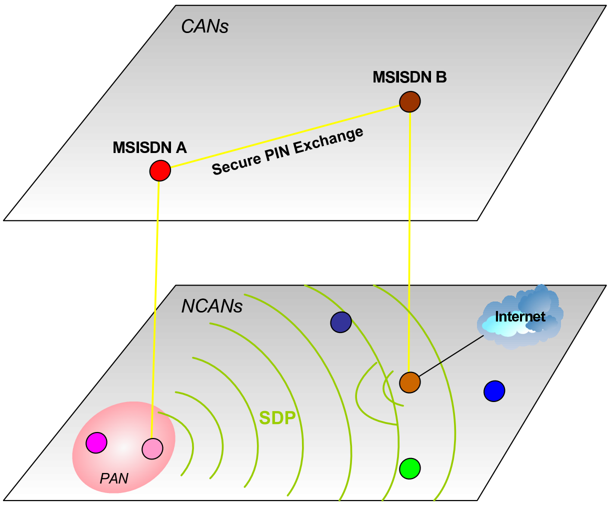


Figure 6.48: Automated PIN Exchange with CAHN

Defining the CAHN ability as a service that can be searched for increases the efficiency of the system. If the neighborhood can be scanned for a specific identity (i.e. MSISDN) of the peer before initiating the CAHN connection setup, networking resources can be saved. With the help of a local *Scan-* or *Trap Request* the SMACS layer can assure that it gets notified whenever a certain peer is detected within its range for direct communication. Hence, the CAHN layers are only trying to establish a direct link if the peer is within range. Thus, service discovery used for identity discovery may help to reduce the signaling overhead of CAHN. How this identity discovery can be realized with Bluetooth and WLAN is discussed in the following sections.

6.6 CAHN Implementation

The system and protocol presented in the previous sections to handle the establishment of direct links has been partially implemented to get some proof of concept. The implementation focused on the basic layers and protocol messages required to evaluate the feasibility and benefit of separated signaling and data channel management for heterogeneous communications.

6.6.1 Architecture on Linux

The prototype was implemented on GNU/Linux Fedora [162] Core 3 with Kernel 2.6. Opposite to the design presented in the previous section, the GUI is directly interacting with the CCM. Therefore, only the parts of SMACS required for the user interaction have been implemented in the GUI module. All three layers are written in C code. The CCM is implemented as a standalone application, which switches into background and waits for connection requests and responses. For incoming connection requests the CCM offers a Unix Stream Socket, which is enabling bidirectional communication between the CCM and the module sending the request. This allows the delivery of the response without requiring reestablishing a new socket. Due to the symmetry of our architecture, connection requests can originate either from the local GUI or from the remote CCM. For each received connection request, the CCM creates a signaling instance responsible for that connection request and a corresponding listening server socket for related incoming responses. Considering the fact that our signaling protocol is message oriented, these listening sockets were implemented with Unix Datagram Sockets. Local connection requests are treated by the CCM and forwarded to the Adapter which is taking care of the actually chosen signaling channel (see Section 6.3.4). Figure 6.49 is illustrating the communication between the GUI, the CCM and the Adapter.

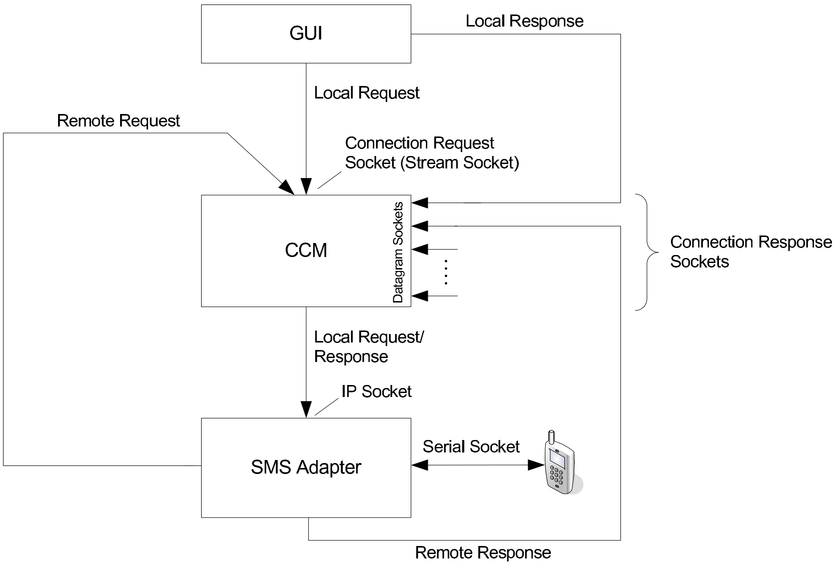


Figure 6.49: Inter-Process Communication Between the Graphical User Interface, the CCM, and the SMS-Adapter

The communication between the CCM and the Adapters is based on IP Sockets. The Adapter is transmitting the handled local request towards the remote CCM using the dedicated functions according to the underlying signaling channel. The main structure of all Adapters is identical. They only differ in how they send and receive signaling messages to and from the actual signaling channel they are in charge of. As a representative for all Adapters we describe the GSM SMS Adapter in further detail. The SMS Adapter is using AT Hayes commands over a Serial Socket to communicate with the GSM device offering the possibility to send and receive SMS. For IP based signaling networks the Adapter is simply relaying the CCM messages to the remote CCM. Incoming remote requests are handled and forwarded to the Unix Stream Socket of the CCM, where they are treated by the CCM. In the case of required user interaction, the CCM establishes the communication with the GUI. If the connection request is granted, the CCM creates a signaling process instance and a Unix Datagram Socket, similar than for local connection requests received from the GUI. If the Adapter receives a response, it forwards the message to the according listening server socket of the CCM, depending on the value of the MSG-ID field of the response (see Section 6.4.1), which is identifying the signaling instance it is belonging to. Local responses, coming from the CCM or the GUI are forwarded to the remote CCM, similar to local requests.

The utilization of a dedicated Unix Datagram Socket for each signaling instance allows full flexibility when routing signaling messages through different signaling channels. The Adapters are completely state-less. The routing of the outgoing signaling messages happens within the CCM. According to the actually selected signaling channel the CCM forwards the signaling messages to the corresponding Adapter. Incoming messages are forwarded by the Adapters to the Unix Datagram Socket of the CCM belonging to the right signaling instance. Hence, the remote CCM can dynamically decide on which signaling channel the message is to be sent, without requiring to inform its communication peer. Figure 6.50 depicts the initial connection setup using the SMS Adapter.

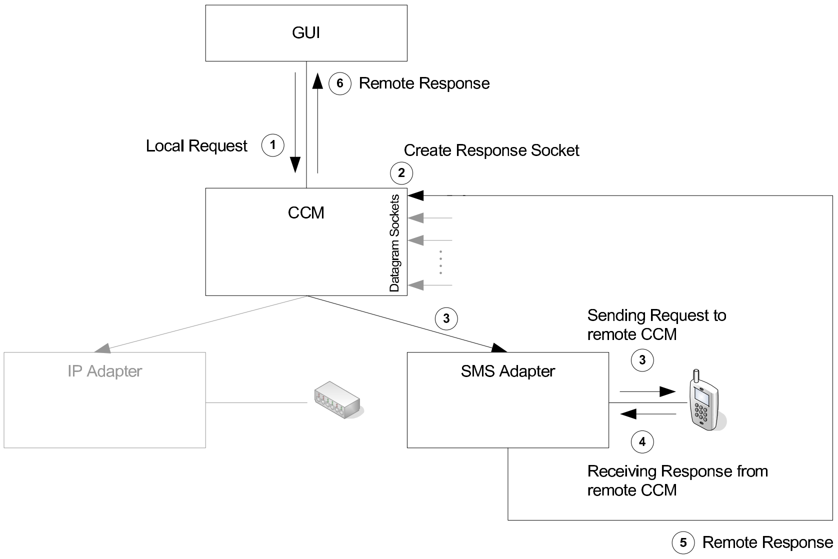


Figure 6.50: Initial Connection Setup Through SMS Adapter

First, the connection request is created based on the user's selection of the peering node. The CCM established the signaling context and initializes the response socket (Step 2). Based on the communication addresses available in the table of identifiers (see Section 6.2.2) for the peering node, the CCM selects the most appropriate signaling channel (here cellular MSISDN, because there is no other known identifier available at initial connection setup). The Adapter then sends the request to the peering node, which replies with a connection response (Step 4). Based on the MSG-ID of the response, it is relayed to the according response socket of the CCM (Step 5). The CCM finally informs the user about the successful connection establishment (Step 6).

After having successfully established the initial connection, the subsequent signaling messages can be routed through that secured channel. This inband signaling can be done without the explicit notification of the peer node. The local CCM can switch ongoing signaling sessions from one technology to another by forwarding the signaling messages to the new Adapter. This can happen without any user interaction and implicitly with the normal signaling process. The dynamic selection of the signaling channel does not require any additional message exchange between the communicating nodes. For every single signaling message that has to be sent to the peer node, the CCM can freely select the Adapter and hence the signaling channel to be used. Figure 6.51 illustrates the implicit change of signaling channel from SMS to an IP-based signaling channel. This IP channel may be the the actual data channel in the case of inband signaling or any other IP-based communication channel. Steps 1 to 3 depict a standard data connection update procedure. Because of the datagram based response sockets, the CCM is not even realizing that the response is coming from a different Adapter. However, the peering node is implicitly aware of the changed signaling channel and can react accordingly by updating its table of identifiers accordingly[4].

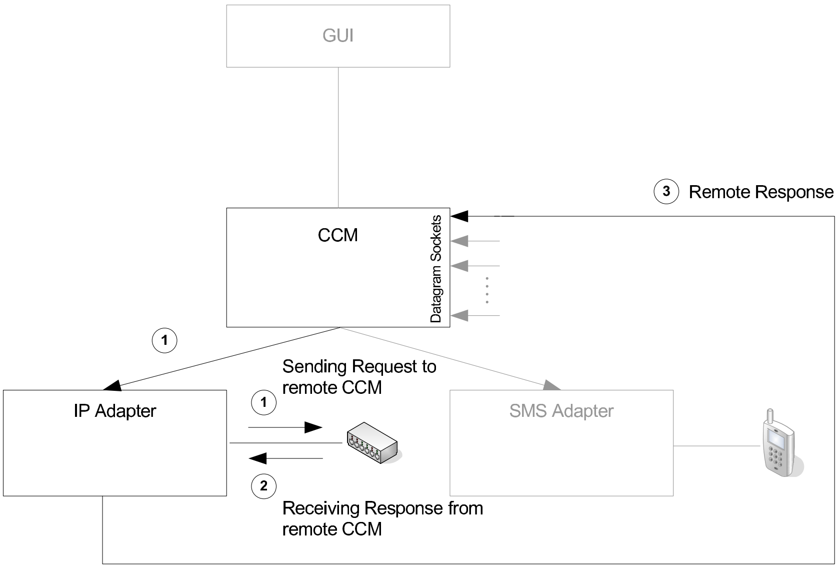

GUI
3 Remote Response
CCM
Datagram Sockets
1
Sending Request to remote CCM
1
IP Adapter
SMS Adapter
2
Receiving Response from remote CCM

Figure 6.51: Connection Update Through IP Adapter

The Connector is notified by the CCM after the reception of a remote request or response. In the case of a local request, the CCM is waiting until it receives the remote response before triggering the Connector to configure the communication interface accordingly. Although the connection setup process

[4] Note that the connection update message is only contain information relevant to the data channel and does not concern the signaling plane at all.

could be accelerated, if the local Connector would start to configure the interface right after the composition of the local connection request, this might result in overhead if the peer node proposes a new set of connection parameters by entering the advanced setup mode (see Section 6.4.8). This is especially true for complex and time consuming configuration processes that would have to be aborted immediately after receiving a new set of parameters. The communication between the CCM and the Connectors was implemented using linked functions. Therefore, the Connectors offer simple functions allowing the CCM to communicate the parameter set received from the SMACS layer or the remote CCM. Within the prototype, the CCM is creating the parameter set, since only one connection has to be handled. These functions defined in the header file of the Connectors are included in the C code of the CCM module. For future implementations, where several Connectors could be deployed in parallel, the introduction of socket based communication between the CCM and the Connectors would allow asynchronous processing of the configuration of the interfaces and further signaling messages. For the sake of completeness, Figure 6.52 illustrates both the Connector for WLAN and Bluetooth devices. The WLAN Connector is relying on the iwconfig [179] for the WLAN specific configurations like SSID, WEP keys, and mode of operation, and ifconfig [185] for IP related settings. For the configuration of Bluetooth specific parameters, the Bluetooth Connector uses the Logical Link Control and Adaptation Protocol (L2CAP) and the Host Controller Interface (HCI) defined by Bluetooth. Our Bluetooth Connector implementation relies on the Bluez [17] Bluetooth stack.

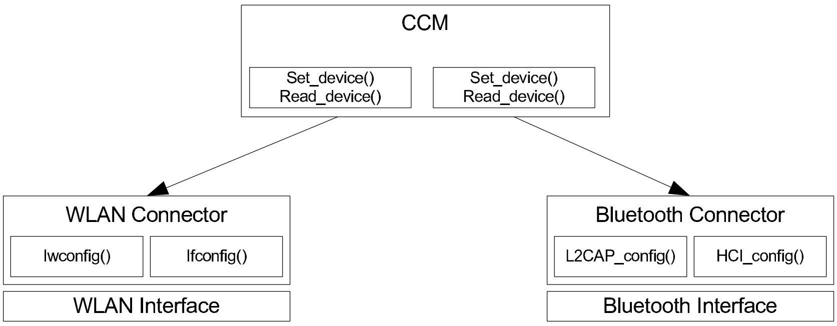


Figure 6.52: Bluetooth and WLAN Connectors

Further details on the implementation of the CAHN service for Bluetooth can be found in [192]. For sake of simplicity the proposed application of Bloom filter in Section 6.5.2 was not implemented in the prototype. However, the introduction of the Bloom filter to code several MSISDNs into one single SSID is not crucial to proof the concept of CAHN. Each node sets first his SSID according to its MSISDN and the concatenation of both MSISDNs is used to form the SSID of the joint ad-hoc network.

The initial version of our prototype used SMS to transfer the signaling information from one node to the other, which resulted in large delays due to the store and forward nature of the SMS service. However, this first version of the prototype successfully approved the concept of CAHN. The second version was extended to use USSD as the primary signaling channel to exchange the CAHN protocol messages.

SMS Adapter

The simplest way to interact with the cellular network signaling is using the SMS. Nearly all cellular devices offer an AT command interface through a serial link, like provided by Bluetooth, Infrared or cable. With the help of the AT commands, the messages to be sent can be spooled and incoming messages can be read out of the device. Both, the sender and the receiver are addressed based on the MSISDN. The SMS is transported with the help of the *SM-TP* (Short Message Transport Protocol) using a dedicated *SDCCH* (Stand-Alone Dedicated Control Channel) or using the *SACCH* (Slow Associated Control Channel) of an active call. Therefore, no data channel is used to send and receive a SMS, which perfectly reflects the concept of using low power signaling channels for CAHN. The SMS service is based on the store and forward principals, which may result in having delays when sending a SMS from one node to another. In [192], tests have been done to estimate the average transfer delays of a SMS sent from one node to another. The delay depends on the length of the message. For an empty SMS the average delay is about 7 seconds and for a full length SMS (160 character, 7 bit encoded) about 11 seconds. These values have been collected only to give a rough estimation on the relative delay distribution for CAHN connection establishment and are hence not considered as precise. In the case of roaming, the delivery can take even longer. Depending on the cellular devices, the spooling process used to send SMS, is taking up to 5 seconds before the SMS is actually sent. On the other hand, when receiving a SMS there is no mechanism to get notified through the serial link. Thus, the CAHN node has to regularly poll the cellular device for new messages. For the implementation a polling interval of 5 seconds has been chosen to reduce the processing efforts when waiting for connection requests.

USSD Adapter

The migration to USSD increased the performance considerably in terms of connection setup time and reliability. In contrast to SMS, the USSD is session oriented and transported over the *FACCH* (Fast Associated Control Channel), which is about five times faster than the SACCH. However, it is still based on the signaling of the cellular network and hence considered as a low power (and low bandwidth) communication. Opposite to SMS, the USSD was designed to transfer information between the *UE* (User Equipment) and the network and vice versa, and not between UEs. All USSD sessions are directly routed to the home operator's network (HPLMN), which enables CAHN functionality also for roaming nodes without requiring any changes in the visited network. However, there is no possibility to establish direct sessions between two nodes. To enable node-to-node communication, further functionality has to be added within the HPLMN. USSD is defining two basic modes of operation, being the *Network Initiated Dialogue* and the *Mobile Initiated Dialogue*. Combining those two modes allows the transmission of information between nodes. Therefore, a special node was designed to route the messages coming from the *Mobile Initiated Dialogue* through the appropriate *Network Initiated Dialogue*. The *USSD Gateway* is a standard component located in the cellular network and offering facilities to encapsulate the information transported through the USSD session into TCP/IP packets. Fig. 6.53 illustrates the overall architecture and information flow when using USSD for CAHN: A CAHN node sends a CAHN USSD Request (1) to the Home Location Register, which relays the request to the USSD Gateway via the SS7 link (2). The USSD Gateway routes the CAHN message to the CAHN USSD Router (3), which analyzes and relays it to the corresponding CAHN node (4, 5 and 6).

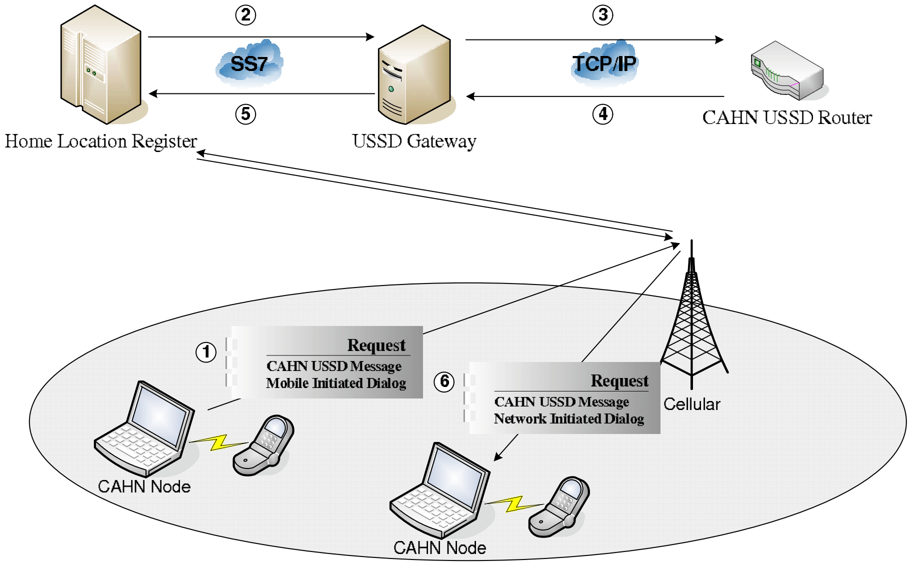


Figure 6.53: CAHN USSD Architecture

The *USSD Gateway* is set to forward all CAHN service relevant information to the CAHN USSD Router. In the first stage, this router is responsible to deliver the incoming (mobile initiated) information to the destination node (network initiated). Therefore, the router analyzes the received information and extracts the destination MSISDN (see Fig. 6.54).

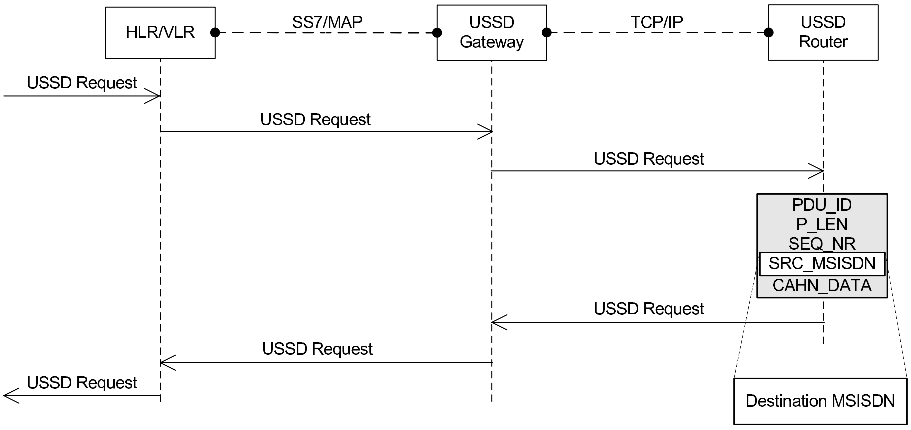


Figure 6.54: CAHN USSD Processing

Additionally to the message fragmentation handled by the USSD Adapter there are two further elements required to enable CAHN to cope with USSD. The component responsible for the actual USSD dialogues (Mobile- and Network initiated) is called *USSD Connector*, since it is handling USSD specific functions (see Section 6.3.2). A buffer (called USSD Pool) is introduced to enable the communication between the USSD Adapter and the USSD Connector. Whenever the USSD Connector finds a CAHN message in this pool it starts a *Mobile Initiated Dialogue* and forwards the message to the CAHN USSD Router. Network initiated USSD dialogues are handled by the USSD Connector as well. The CAHN relevant information is extracted from the USSD message and stored in the USSD pool, where it is further processed by the USSD Adapter, and finally delivered to the CCM.

The overall architecture that has been implemented is shown in Fig. 6.55.

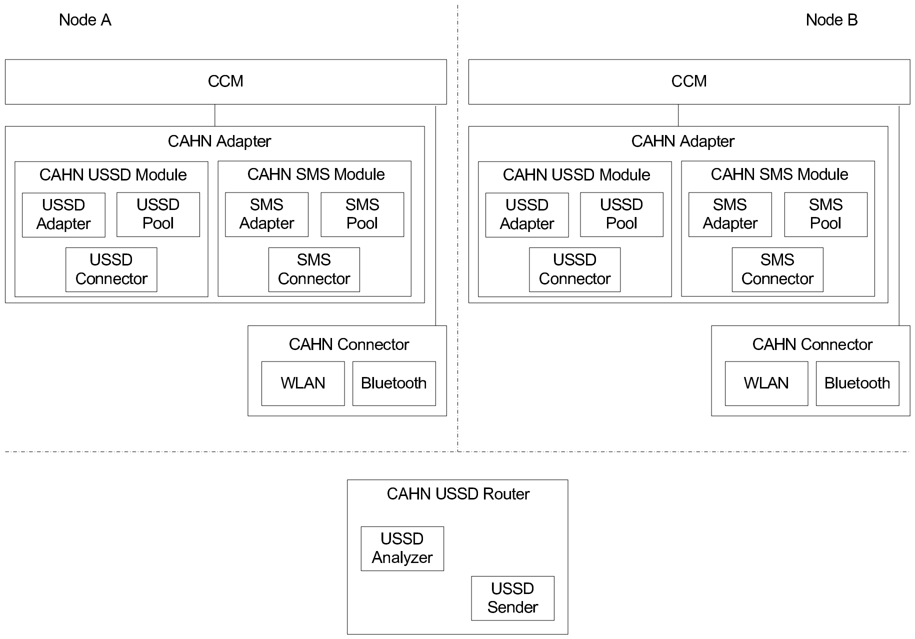


Figure 6.55: Implemented Prototype

The implementation of the USSD router required to properly handle terminal and network initiated USSD requests was done on a Linux server located within our testbed. The USSD gateway relays the SS7/MAP messages from the cellular network to TCP/IP based transport networks. The simplest way to exchange messages between the USSD Gateway and the USSD router is using HTTP. Whenever a CAHN node wants to transmit a message to another CAHN node, it sends a terminal initiated USSD message to our specific Service Code (SC). The USSD gateway is configured to relay incoming request from the cellular networks towards our USSD router using an HTTP POST message containing the original payload sent by the terminal in XML format. The reception of the HTTP POST message is done by the Apache server and a Python script located on our USSD router. The USSD router treats the HTTP POST message and creates the CAHN request that is sent to the destination node using a network initiated USSD request. Therefore, the USSD router forms an HTTP GET message including the actual CAHN connection request message as request parameters. Figure 6.56 illustrates the interaction between the CAHN nodes, the USSD Gateway and the USSD router with the according messages and protocols.

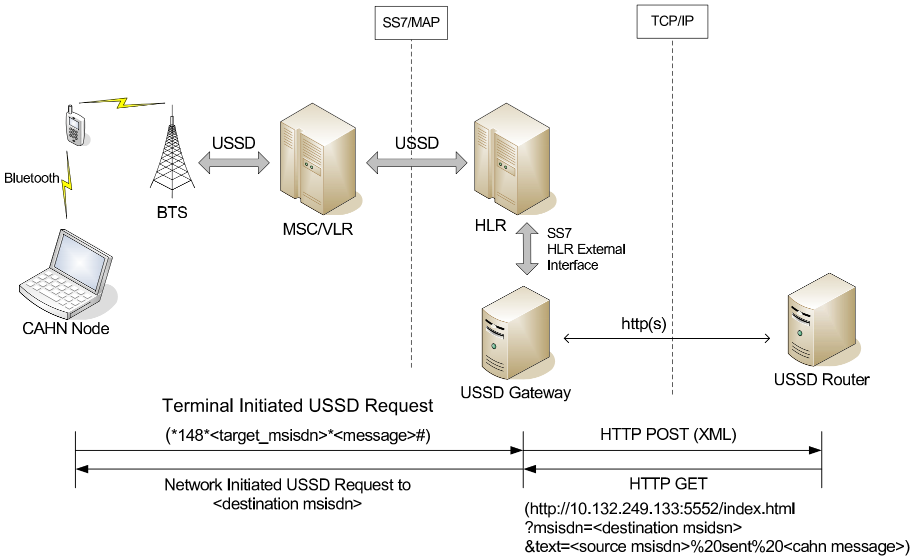

SS7/MAP
TCP/IP
Bluetooth
BTS
USSD
MSC/VLR
USSD
HLR
SS7
HLR External
Interface
CAHN Node
http(s)
USSD Gateway
USSD Router
Terminal Initiated USSD Request
(*148*<target_msisdn>*<message>#)
HTTP POST (XML)
Network Initiated USSD Request to
<destination msisdn>
HTTP GET
(http://10.132.249.133:5552/index.html
?msisdn=<destination msidsn>
&text=<source msisdn>%20sent%20<cahn message>)

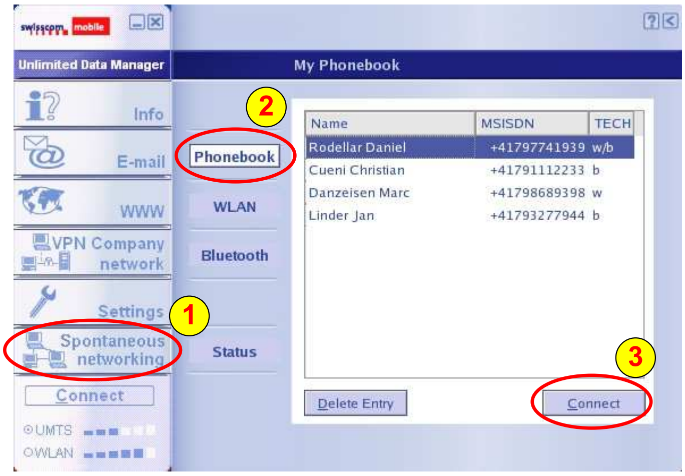
Figure 6.56: Interaction with the Cellular USSD Service Platform

The service number *148* was assigned for testing reasons by a network operator to the CAHN service and therefore relayed to our USSD router. The traffic between the USSD gateway and the USSD router can optionally be protected with HTTPS. The XML message carrying the CAHN connection request and response is structured as follows:

```
<?xml version="1.0"?>
<!DOCTYPE service SYSTEM "service.dtd">
<phonenumber>+41795972833</phonenumber>
<parameters>
<parameter id="1">
<name>message</name>
<value>CAHN message</value>
</parameter>
<parameter id="2">
<name>targetmsisdn</name>
<value>41795934446</value>
</parameter>
</parameters>
```

The connection request is split into two parameter triplets, one for the actual CAHN message and one for the destination MSISDN. Each triplet consists of parameter id, the parameter name, and its value. The source MSISDN is declared in the beginning of the XML file with the <phonenumber> tag. The ability to exchange XML based messages allows future extensions to provide more information to the USSD router to monitor CAHN connections. When thinking about providing CAHN like services in a rather centralized manner, it might be valuable to inform the USSD router about available networking technologies on the CAHN nodes. But also in a multi-party connection establishment process, the provisioning of complete lists of invited nodes might enable the USSD router to fork CAHN connection requests to reduce the number of messages that have to be sent by the inviting node.

The GUI of the prototype was designed to enable selection of the peer based on an address book. Fig. 6.57 shows a snapshot of the GUI, which is a graphical replication of the *Unlimited Data Manager* [64] provided by Swisscom. The GUI is underlining the envisioned abstraction of the underlying heterogeneous communication technologies. After having selected the spontaneous networking application (Step 1), the connection establishment is initiated by selecting the peer from the address book (Step 2 and 3). Optionally, the user can see what kinds of technologies are supported by the peering node (W: WLAN, B: Bluetooth).



Name	MSISDN	TECH
Rodellar Daniel	+41797741939	w/b
Cueni Christian	+41791112233	b
Danzeisen Marc	+41798689398	w
Linder Jan	+41793277944	b

Figure 6.57: CAHN Prototype Graphical User Interface

6.6.2 Implementation Challenges

In this section, we briefly address the main challenges faced when implementing the prototype.

Symmetry of the Architecture

The requirement to build a decentralized system, which is able to interact as client and server at the same time was imposing most of the challenges. Designing the CCM to offer a generic Connection Request Socket handling both, the local and remote requests allowed us to implement one only module for client and server activities. To facilitate the differentiation of the incoming requests, our protocol defines source and destination addresses for each single signaling message. Based on the source address, the CCM can distinguish local from remote requests. Local requests are forwarded to the remote destination address, whereas remote requests have to be treated explicitly (for example forwarded to the GUI, if the user has to decide whether the request should be accepted). Responses are analyzed in terms of their destination address. Remote responses are further treated by the local CCM. Local responses, on the other hand, are directly forwarded to the remote CCM. The definition of identical protocol messages for local and remote request and responses is considerably simplifying the implementation of the CCM. The symmetry of the architecture is not relevant for the Adapters and Connectors, since they offer simple client functionality to the CCM. The Adapters just relays messages between the remote and the local CCM by applying the required modifications to the messages according to the restrictions of the actual signaling channel. The Connector receives the configuration requests from the local CCM only and is therefore not impacted by the symmetrical design.

Routing of Signaling Messages

To enable flexible handling of the signaling channels, the whole protocol was designed to be message based. Since the signaling sessions are by definition session oriented, a dedicated mapping has to be done between the signaling sessions and the messages. This mapping function has to be implemented in the CCM to keep the selection of the actual signaling channel independent of the Adapters. The Adapters have to be able to handle the individual messages without keeping any states about ongoing signaling sessions. Therefore, the functions of the Adapters are purely defined based on the information found in the message header. Whereby outgoing messages are sent to the destination address found in the address field and the incoming messages relayed to the corresponding server socket of the CCM according to the signaling session ID (MSG-ID) field. The signaling ID corresponds to the port number of the listening server socket of the CCM. In order to respect the well defined server ports, a fixed number (greater than 1024) is added to the signaling ID shifting the port numbers.

SMS and USSD Interaction

The most straight forward method to interact with the cellular network is using PCMCIA cards. These cards offer serial ports that can be used to establish modem connections based on AT commands. To send and receive SMS, dedicated AT commands can be used. We used a serial socket to handle the interaction between the SMS Adapter and the PCMCIA card. Unfortunately, we did not find a PCMCIA card, which supports network initiated USSD sessions. So we had to rely on mobile phones to perform the exchange of USSD messages. On the other hand, we did not find a mobile phone supporting the direct delivery of incoming SMS to the serial connection. Hence we had to use a separate setup for SMS and for USSD. The communication between the laptop and the mobile phone was based on AT commands as well using a serial Bluetooth connection.

6.6.3 System Evaluation

The implemented prototype proved our general concept of cellular assisted heterogeneous networking. Both, the use of SMS and USSD worked fine to exchange configuration and security related information required to establish a Bluetooth and WLAN link between two nodes. Several tests have been done to get further information about the limitations of our proposed system.

Three major processes have been identified influencing the required time to successfully establish a connection between two CAHN enabled nodes. Namely the creation and processing of the signaling messages, the message exchange of the CAHN signaling messages (via SMS, USSD or any other signaling channel), and the configuration of the communication interface according to the negotiated parameter set. To assure that the peer is within the vicinity our system first scans the environment. Depending on the short range communication technology used, there is an additional delay imposed by the scanning process. The establishment of a Bluetooth link is identically with the establishment of a WLAN ad-hoc link, apart from neighborhood scan process, which takes much more time with Bluetooth than with WLAN. The duration of the Bluetooth scanning process can be set explicitly. Although setting the duration to low might lead to superficial scan results potentially missing neighbors. A good value for the Bluetooth scanning duration is $10\,s$. The delay of the scanning process of WLAN depends on the broadcast interval of the SSID, which can be manually set on most systems. For our tests we set the broadcast interval to $10\,ms$. This assures a quick discovery of neighboring nodes and does not unnecessarily slow down the connection establishment process. Considering that the connection establishment process is the same and the configuration is almost identical, independent whether a WLAN or a Bluetooth link is finally established, we focused our tests on WLAN only.

We first evaluated the overall performance of our prototype using SMS, USSD as signaling channels to establish a WLAN ad-hoc link between two nodes. To evaluate the processing overhead introduced by our framework, we built up a dedicated scenario. Then we made several tests with USSD to estimate the transfer delay introduced by our USSD Router allowing us finally to further decompose the overall connection establishment time.

Overall Performance

To evaluate the required overall connection establishment time of our prototype, we defined two different test scenarios. Figure 6.59 illustrates the SMS and Figure 6.58 the USSD related performance testbed setup. The SMS testbed was based on PCMCIA cards offering access to the cellular network. This allowed direct and fast access to the incoming SMS compared to using mobile handsets.

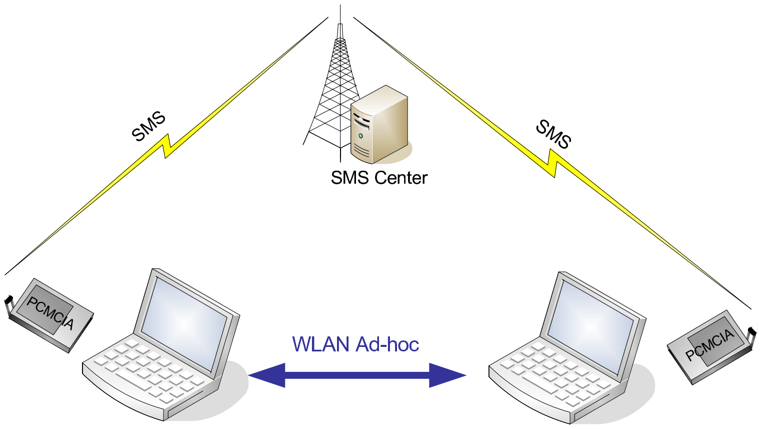


Figure 6.58: Testbed Setup for SMS

Network initiated USSD connection initiation was not supported by the PCMCIA cards used for the SMS testbed. Therefore, mobile handsets were connected through Bluetooth to send and receive USSD messages. Unlike SMS, USSD is not based on store and forward mechanisms, which eliminated the risk of having additionally delays between the handset and the computers.

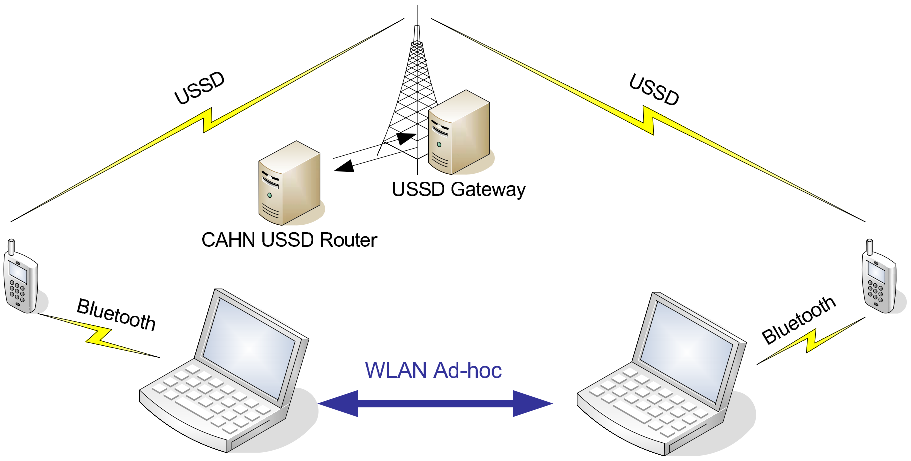


Figure 6.59: Testbed Setup for USSD

Using the two testbeds, we made 15 consecutive measurements with SMS and USSD, respectively. The measurements included the complete connection establishment, starting with the creation of the connection request message and ending with the successful configuration of the WLAN interface. Figure 6.60 shows the overall connection establishment time for the 15 test runs. The average values and confidence intervals are given in Table 6.2.

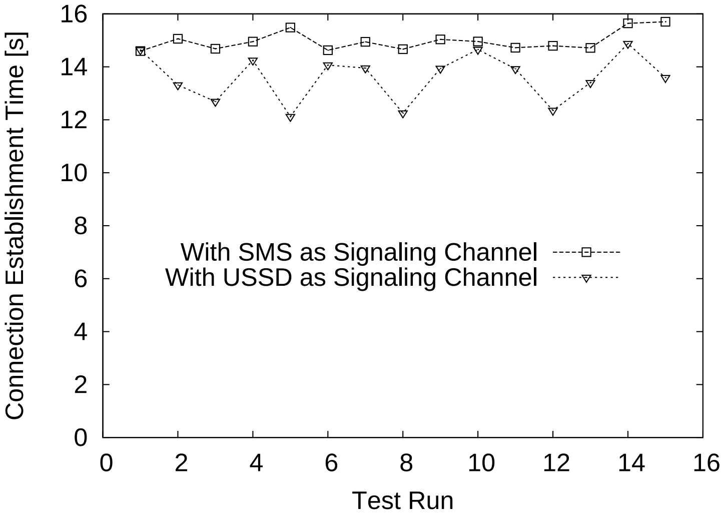


Figure 6.60: Overall Connection Establishment Time

	Signaling over SMS	Signaling over USSD
Average Time [s]	14.972	13.6024
Confidence Interval	0.184529	0.454669

Table 6.2: Overall Connection Establishment Measurement Results

To analyze the connection establishment time in further detail, we adapted the implementation code to log certain events in the system log. To assure the synchronization of the system clocks of both nodes, we used the Network Time Protocol (NTP) [134] via an Ethernet connection. The different events triggering a log entry are illustrated in Figure 6.61.

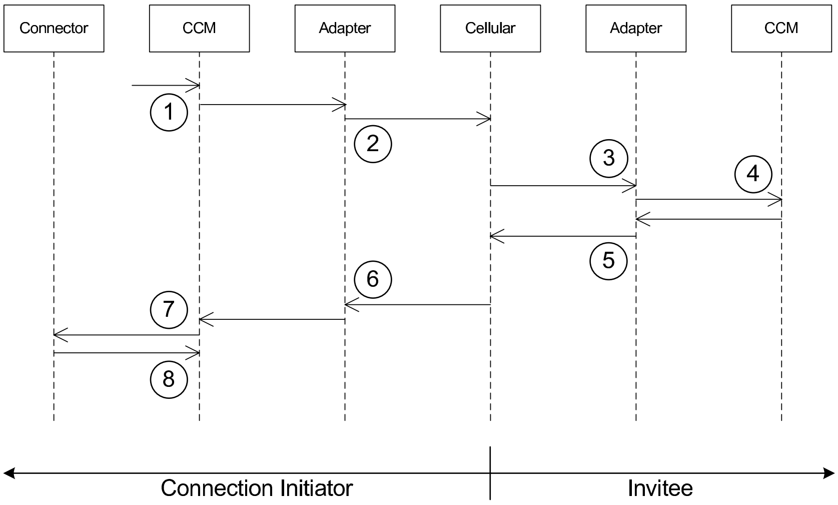


Figure 6.61: Logged Events

When the CCM is triggered to initiate the connection establishment, the first time stamp is logged (Step 1). The CCM forwards the connection request message to the Adapter, which is sending the SMS or USSD message to the cellular network (Step 2). Right after having received the acknowledgement of the cellular terminal, the Adapter creates a log entry. The Adapter of the peer node logs the receiving of the message (Step 3) and relays the connection request to the CCM, which is also logging the reception (Step 4). After processing the request and delivering the connection response message to the Adapter. The Adapter sends the response message through the cellular network and creates a further log entry (Step 5). The Adapter of the initiator is logging the arrival time of the response (Step 6) and relaying the message to the CCM. The CCM relays the connection relevant information to the Connector and immediately logs the start of the configuration procedure (Step 7). After having successfully configured the communication interface according to the negotiated parameter set, the last log entry is created (Step 8).

The eight logged time stamps on the two nodes determine the seven processes (P1-P7) required to perform a simple connection setup:

1. Creation of the connection request message by the CCM and send it through the Adapter: P1

2. Delivery of the request message through the cellular network to the peer node: P2

3. Processing of the request message by the Adapter and relaying it to the CCM: P3

4. Processing of the request, creation of the response message, and sending the response back to the initiator: P4

5. Delivery of the response message through the cellular network to the initiator of the connection: P5

6. Reception and processing the response by the Adapter and the CCM: P6

7. Configuration of the communication Interface according to the parameter set provided in the connection response: P7

Figure 6.62 shows the values measured for the seven Processes involved in the connection establishment. We made 15 test runs for SMS and USSD. The average values and the confidence intervals of the measured time for the different processes are given in Table 6.3.

Figure 6.62 clearly shows that the processing time required by the implemented prototype is nearly negligible compared to the time required to transfer the signaling messages from one node to the other using the cellular network. It takes only about 0.6% of the overall connection establishment time including the configuration of the WLAN interface. P1 and P4 include the delay introduced by the cellular device to sent the message. The Adapter relays the message to the cellular device using AT commands and waits for an acknowledge, which is only given after successful transmission of the message on the cellular interface. Therefore, the actual processing time required for the treatment of the CAHN

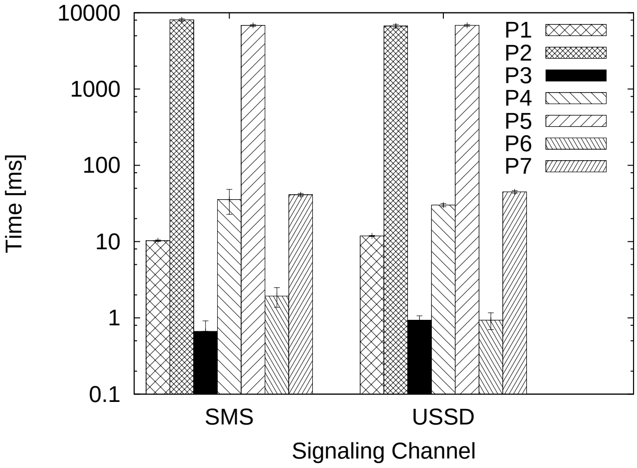


Figure 6.62: Distribution of the Overall Connection Establishment Time to the Seven Major Processes

	USSD		SMS	
	Average [ms]	Conf. Inter.	Average [ms]	Conf. Inter.
P1	11.86666667	0.178065379	10.26666667	0.231642985
P2	6679.133333	350.967522	8067.533333	289.5036413
P3	0.933333333	0.130664266	0.666666667	0.246932252
P4	30.13333333	1.403826681	35.6	12.81214358
P5	6834.466667	178.8992584	6814.933333	202.9267969
P6	0.933333333	0.231642985	1.933333333	0.55655703
P7	44.93333333	1.849191203	41.06666667	1.694322364

Table 6.3: Measured Delay for each Process P1-P7

protocol messages would be even smaller than measured with P1 and P4. To estimate the limitations of the prototype, we prepared a specific test setup to measure the processing overhead introduced by the CAHN components.

Processing Delay

The former measurements showed that about 99.4% of the connection establishment delay is due to the transmission of the signaling messages. To get a clearer picture of the performance of our prototype, we extended our implementation with a native Ethernet Adapter. We interconnected the two nodes directly with a LAN cat 5, crossed cable to assure that our tests are not influenced by other traffic load on the LAN. Since we were mainly interested in the delay introduced by our system, we implemented a dedicated dummy client to send consecutive connection requests to the peer node. The delays between the outgoing requests and the reception of the corresponding responses was logged. The peer node has been adapted to send the connection response without triggering the Connector to configure the WLAN interface. This was necessary to assure that the CCM is immediately available for the next connection request. A simple UDP server was implemented and measured for comparison. Figure 6.63 shows the measurement results for the hundred consecutively request/response pairs. In Table 6.4 the average response delay and its confidence interval is given.

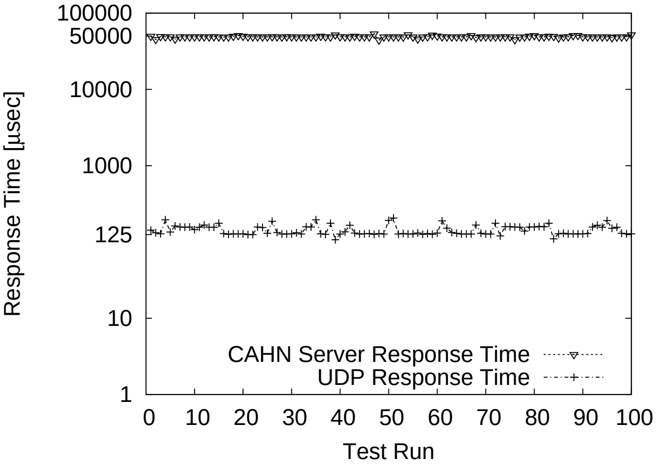


Figure 6.63: Comparison of the CAHN Server and Native UDP Response Time

	UDP Response Time	CAHN Server Response Time
Average Time [μs]	143.15	49040.33
Confidence Interval	4.14044515	246.196068

Table 6.4: Response Delay Measurement Results

The implemented UDP server waits for incoming packet and replies with a pre-configured UDP packet. No additional processing is performed. With this simple UDP server we analyzed the time required by the UDP stack to transmit a UDP packet over the Ethernet link. The CAHN server, on the other hand, processes each incoming message and creates a corresponding response. The comparison of the response times shows that 99.71% of the delay is imposed by the CAHN server. The average response time of about 49 ms would allow up to 20 connection request per second. Regarding the fact, that CAHN connection requests are mainly used to either initiate a spontaneous network between nodes or prepare the direct link to switch over ongoing data sessions from infrastructure-based connections, this response delay is almost negligible. In the first case, where user interaction is required to decide about the acceptance of the connection request, a delay of 49 ms is much smaller than the overall connection establishment time and therefore not crucial at all. If CAHN is used to prepare the direct link between SMACS nodes to optimize the data path of ongoing sessions, the handover is delayed upon the message exchange is fulfilled. This delay is not critical, if the infrastructure-based access network remains available until the CAHN connection establishment is terminated. Depending on the signaling channel used (e.g, SMS or USSD) the delay introduced by the message transfer between the nodes is considerably higher than the CAHN server respond time.

6.6.4 Improvement Potential

The different performance evaluations presented in the previous sections revealed several improvements that should be considered in a next version of implementation. Probably the most obvious limitation of the implemented system is the dependency on the transmission delay of the signaling channel used. The measurements done with SMS and USSD showed that about 99.4% of the overall connection establishment time is spent to transfer the connection request and response over the SMS and the USSD platform, respectively. Unfortunately, both, the SMS and the USSD message delivery had to be considered as a black box for the work done within this thesis. The SMS delivery is based on store and forward, which introduces additional unpredictable delays, depending very much on the actual load of the SMSC. USSD theoretically offers much faster data delivery but our tests evidence only about 9.2% shorter delays for USSD message transfer compared to SMS. Measurements of the round trip time between the USSD router and the USSD gateway showed that only a fraction of the USSD delivery time is used outside the cellular network. The rest of the time is spent by the USSD terminal and network initiated message transfer and can not further be decomposed without making measurements on the different nodes of the cellular network (e.g., MSC, VLR, HLR, and USSD gateway). Since the USSD channel should deliver near to real time data transmission, we have to assume that there is a lot of improvement potential to optimize the treatment of our USSD requests on the USSD gateway.

The implemented prototype supports any IP connection as well to exchange the CAHN protocol messages. Therefore, even if the transmission delay of the USSD channel could not be considerably decreased, the system could be extended to use USSD only if no other IP connection is established. It is also imaginable to make the USSD router reachable for native IP traffic, allowing the CAHN nodes to send IP CAHN protocol messages. In combination with a registration service similar to the centralized identifier tables presented in Section 6.2.2, the USSD router could act as an intelligent CAHN signaling gateway, relaying the messages on the most appropriate signaling channel. If the destination of an incoming connection request has not registered any communication address belonging to a faster signaling channel than USSD, the gateway is relaying the request using this low power channel. Preliminary tests have shown that the connection establishment time is reduced to less than 2 seconds if GPRS is used on both nodes. Although this considerably decreases the overall connection establishment time, the utilization of GPRS as a permanent signaling channel is unfavorable in terms of network and battery resource savings as shown in Section 6.6. Concluding from the issues just mentioned, a future version of the prototype should incorporate the capability of adaptively select the most favorable signaling channel and make use of a centralized signaling gateway.

6.7 Conclusion

Based on the findings of the prior Chapters, a new concept was developed, called *Cellular Assisted Heterogeneous Networking* or CAHN. This CAHN concept, reusing the existing cellular network to securely establish and maintain heterogeneous end-to-end IP sessions, seems to be a promising way to overcome various hurdles. The cellular network acts as signaling plane providing paging and authentication mechanisms and secure transfer of CAHN signaling messages. The three major layers of the CAHN component, being the CCM, the Connector and the Adapter have been introduced. Furthermore, a new signaling protocol has been developed offering the facilities to exchange required information to set up a secured IP session between nodes using heterogeneous networking technologies. The role of service discovery mechanisms in the domain of spontaneous networking has been explored, which lead to definition of a new Bluetooth Profile to facilitate the discovery of the peer and hence the connection setup process. Bloom filters were introduced to enable efficient CAHN node detection for WLAN based ad-hoc networks. Based on the designed concepts and system architecture, a prototype was implemented to prove the feasibility of the design work done. Various performance evaluations showed that the connection establishment time is acceptable for a first implementation and that the main limitations are imposed by the signaling channel used to exchange the CAHN protocol messages. The thorough analysis of the performance tests identified further improvement potential, if the system would be extended with a centralized signaling gateway, enabling the different CAHN nodes to use the most appropriate signaling channel depending on their actual status. Nodes being already involved in IP data sessions could reuse these data channels to perform inband signaling, whereas idle nodes could be reached through low power out-of-band channels like SMS or USSD.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this thesis we addressed the problem of heterogeneous networking that end-users have to face every day. Therefore, we studied the different communication technologies and protocols that are available today, and some that will be disposable soon to enable what is referred to as being always best connected. Interconnecting nodes with the most suited technology available is important to enhance the performance of existing networks and simplify the integration of future communication technologies. Standardization bodies, network operators, and researchers are working towards a seamless integration of access networks. Thinking further about being always best connected, there is a major part missing in the evolution plans of the network operators. High performance short range wireless communication is delivering data rates order of magnitudes higher than what is achievable with mobile wide area networks. Infrastructure-less links that can be established directly between nodes, without requiring expensive infrastructure, can hence considerably improve the networking performance. In this thesis, we proposed a system architecture to achieve a seamless integration infrastructure-less communication technologies in the vision of being always best connected.

In Chapter 4, we elaborated on the shortcomings of inband signaling used for the Internet Protocol, when considering heterogeneous communication networks. We further analyzed in detail the problem of bootstrapping direct node-to-node connections between peer nodes and how Mobile IP route optimization combined with IPsec could offer the seamless switching of ongoing communication sessions. The major issues preventing such infrastructure-less connections to be seamlessly established are related to configuration and security settings. Address management, key distribution, selection of encryption algorithm, and finally routing related adaptations are only a few example of what has to be done prior to successful communication. To hide this complexity from the user, we proposed a signaling framework to exchange configuration and security related parameters using a dedicated signaling channel like the cellular network. This separate signaling layer enables a simple management of the heterogeneous underlying networking technologies. By further separating the signaling plane in a logical- and a physical-session layer, the human-to-human communication session can be decoupled from the actual inter-device communication management. Users can invite users and do not have to care about the device used by the peer. This abstraction is highly motivated by the mobile telephony network, where phone numbers identify people and the mobile phones are interchangeable. Consequently, we propose to use the mobile phone numbers to identify the logical communication end-points for heterogeneous data sessions as well.

In Chapter 5 we presented the SMACS layer, which is based on the signaling framework for heterogeneous communication session proposed in Chapter 4. SMACS enables the integration of infrastructure-based and infrastructure-less connections and the ability to allocate broadband networking resources on-demand. Simulations done showed that especially in small areas like university or enterprise campus the average throughput can be increased by a factor of up to 4, if communicating nodes can automatically switch to direct ad-hoc links whenever they are close enough to each other. Using the widely available GSM to signal data session requests to the peering node allows switching resource demanding IP connection down, if no data sessions are ongoing. Our simulations showed that this concept of having broadband connection on demand has the potential to considerably reduce the power consumption and optimize the networking resource utilization of heterogeneous networks. The combination of both features, namely the ability to handover ongoing infrastructure-based data sessions to direct ad-hoc links and the possibility to switch unused IP interfaces to sleep mode, is reducing the energy consumption up to 80% in certain scenarios. The introduction of ad-hoc links to increase the average session throughput whenever communicating nodes come close enough to each other, might gain further importance if the trend towards flatrate tariffing for mobile data communication is going on. If the revenue is limited due to flatrate price models, the operators will be interested in cutting down the cost of each transferred byte. Hence, switching over to ad-hoc links and liberate expensive infrastructure might be an interesting way to increase the profit. The simulation results presented in Chapter 5 estimate a potential increase of network efficiency of up to 40% when enabling the ad-hoc and the broadband on-demand feature offered by our architecture. Depending on the further trend of pricing for the data access, the management of heterogeneous IP sessions might become more important than delivering the actual data. Increasing heterogeneity also increases the complexity of the handling of these various technologies, which in turn may further shift the value proposition towards systems enabling seamless and convenient networking.

In Chapter 6, we introduced the system architecture and protocol definition required to implement the signaling framework defined in Chapter 4 and 5. The proposed protocol is message based to allow flexible selection of the underlying signaling network used to exchange the protocol data units. Consequently it is possible to apply our protocol on any type of network. The architecture includes therefore dedicated adaptation modules to transform the protocol messages according to the different network specific characteristics. To achieve the same flexibility on the data plane, we introduced dedicated modules handling the different networking interfaces. The implemented prototype provides a simple graphical user interface to handle heterogeneous data communication. The session peer can be easily selected from an address book, allowing the connection to be initiated in a very intuitive way. The system is exchanging information on the available communication technologies as well as on the required configuration and security related parameters to establish the ad-hoc link. Performance evaluations showed that the system performs quite well. The establishment of connections take about 15 seconds, whereby nearly 90% of the time is spent by the cellular network to transfer the signaling messages between the nodes. Connection establishment times of about 15 seconds are not acceptable for commercial use, which makes the utilization of SMS or USSD questionable. The application of our signaling framework on faster communication channels like GPRS might considerably increase the attractiveness in terms of connection establishment time, but decrease the benefit introduced in terms of power and resource management.

The domain of heterogeneous networking is a broad subject. We could only address a small part of it with our work presented in this thesis. The main conclusions from the work performed within this thesis can be summarized as follows. We focussed on the abstraction of data sessions to hide the underlying variety of networking component, motivated by the simplicity of mobile voice communication. We also tackled the issues of bootstrapping secured direct communication between neighboring nodes by proposing a cellular assisted approach to exchange required configuration and security parameters. With the help of our own simulator we estimated the potential benefits of our concept in terms of throughput, power consumption, network resource management, and network efficiency of a heterogeneous networking environment including infrastructure-based and direct ad-hoc links. The implementation of a prototype was used to prove our concept, the designed architecture, and protocol.

All our efforts were limited to communications involving two communication parties only. Even though the presented framework should be extendable to cope with multi-party communication, the level of complexity might considerably increase due to the decentralized signaling framework.

7.2 Future Work

In this section, we focus on the possible future work to further explore the potential of integrating ad-hoc and direct node-to-node connections with existing proposals of seamless access to heterogeneous and infrastructure-based communication networks. The concept of SMACS presented in this thesis could be extended to cope also beyond the boundaries of PANs. It is highly probable that persons will have more than one PAN in the future. Networks within the car, the home, the office, and even wearable computing devices may form a sort of personal area network. The proposed abstraction of communication endpoints to persons and the dynamic assignment of actual communication devices could be extended to all nodes belonging to any PAN of a person. The system should hence be able to dynamically create the link between a logical communication session and any physical device of one of the PANs belonging to the logical session entity. In terms of system design, this extension would require relay functionality on the supernode, allowing the CCM of the destination PAN to handle the actual physical connection.

Being a completely decentralized system, SMACS is not yet capable of handling efficiently multi-party connections. Communication sessions are explicitly handled by the end-nodes and therefore the establishment and maintenance of multiple connections to interconnect groups of nodes becomes an issue. One possibility to enable efficient management of such group communication might be to introduce a centralized component taking care of the different connections. The integration of the CAHN messages into the SIP signaling framework might be a promising way to combine the group management features of SIP with the heterogeneous network link handling capabilities of our SMACS/CAHN concept. A lot of further research efforts have to be done to optimize this interworking.

The integration of infrastructure-less communication links into existing proposals focusing on the optimized network resource management might reveal some challenging questions. Beside the technical concerns there might also be economical issues entailed by such an attempt to control free resources. Potential increase of the overall networking capacity might justify operator assistance when configuring infrastructure-less communication technologies offering node-to-node communication for free. To cope with centralized resource management, the decentralized concept presented in this thesis has to be extended to offer proper interfaces allowing the network operator to learn about the networking environment of the node and its actual networking needs.

The simulations done to estimate the potential improvements of our SMACS concepts in terms of throughput, network resource efficiency, and power management have been done with rather simple state of the mobility models. Random way point and reference point group mobility are the mostly used mobility models but are also not considered to perfectly reflecting the mobility pattern found in the real world. Buildings, streets, railways, and special attraction points are extremely influencing mobility behavior of users and therefore nodes. But also social communities have to taken into account when focusing on direct node-to-node communication like we did in the scope of this thesis. Collaborating nodes are likely to stay close to each other, when moving during their journey. Friends exchanging large movie files are probably seeing each others more frequently than strangers do. In consideration of the fact that social communities are highly influencing the probability of being within vicinity of direct communication, the obtained results may even further valorize our concept.

In our simulations, we modeled the data sessions as unicast streams sent from the source to the destination. The streams were defined to be greedy, taking the most available bandwidth. This guaranteed a maximum efficiency of the allocated network resources, for both, infrastructure-based and direct links. It might be interesting to simulate application specific data models instead. Even though, most Internet applications behave greedy, there are applications that require only a specific range of bandwidth like VoIP or audio streaming. Especially, when studying handover decision algorithms in heterogeneous networks, such applications may strongly influence the choice of network.

For every handover a new event was created by our event-driven simulator. This resulted in homogeneous sub-sessions, having a certain bandwidth available, depending on the underlying technology and its current utilization. We did not take higher layer's characteristics into account when changing available bandwidth. The transferred data rate changed immediately according to the bandwidth provided by the communication technology, which is not realistic. TCP, and hence also applications, react with a certain inertia on changing networking conditions. Primarily for frequently changing conditions, this might result in considerably different throughput. To get more realistic simulation results, one may think about delegating the sub-sessions to other simulators providing more accurate lower layer modeling. Nevertheless, the smoothing of the passages between the sub-sessions remains challenging.

The design presented in this thesis offers functionality to assist multi-hop connection management. The signaling framework introduced can be used to learn about the networking environment of the peer but also of intermediate nodes. The cellular assistance might further help to overcome shortcomings of existing multi-hop routing protocols. Having the cellular network as an umbrella covering all nodes eases lot of things in the domain of security and also location management, which are often basic requirements to successfully handle billing issues. Especially when thinking of rewarding schemes in multi-hop networking, this spanning signaling channel might provide a basic element to increase the level of trust when being dependent on collaboration. There is still a lot of work to extend our framework to multi-hop networking. Being designed to handle sessions management between end-nodes, our system has to face severe limitations when connection establishment requires multiple nodes to be involved, as it is the case for multi-hop links.

Last, but not least, the complete implementation of the SMACS/CAHN architecture is an important goal of our future work to expose any errors and inconsistencies in the design.

List of Abbreviations

ABC	Always Best Connected
ADSL	Asymmetric Digital Subscriber Line
AES	Advance Encryption Standard
AH	Authentication Header
AP	Access Point
API	Application Programming Interface
AR	Access Router
ARPU	Average Revenue Per User
BSS	Basic Service Set
BTS	Base Transceiver Station
CAHN	Cellular Assisted Heterogeneous Networking
CAN	Cellular Aware Node
CCM	CAHN Communication Module
CDMA	Code Division Multiple Access
CRC	Cyclic Redundancy Code
CSD	Circuit Switched Data
DNS	Domain Name Service
DOI	Domain of Interpretation
DSL	Digital Subscriber Line
EDGE	Enhanced Data for GSM Evolution
ENUM	E.164 Number Mapping
ESP	Encapsulating Security Protocol
FTP	File Transfer Protocol
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
GUI	Graphical User Interface
HIP	Host Identity Protocol
HNS	Heterogeneous Network Simulator
HTTP	Hypertext Transfer Protocol
ICMP	Internet Control Message Protocol
IEEE	Institute of Electrical & Electronics Engineers
IETF	Internet Engineering Task Force
IKE	Internet Key Exchange
IP	Internet Protocol
LAN	Local Area Network
LS	Logical Signaling
LSS	Logical Signaling Session
LU	Location Update
MAC	Media Access Control

MIP	Mobile IP
NAI	Network Access Identifier
NAT	Network Address Translation
NAV	Network Allocation Vector
NCAN	Non Cellular Aware Node
NFC	Near Field Communication
OFDM	Orthogonal Frequency Division Multiplexing
PAN	Personal Area Network
PDP	Packet Data Protocol
PIN	Personal Identification Number
PKI	Public Key Infrastructure
PPP	Point-to-Point Protocol
PSS	Physical Signaling Session
RAN	Radio Access Network
RFC	Request For Comment
RPGM	Reference Point Group Mobility Model
RSVP	Resource Reservation Protocol
RWP	Random Way Point Mobility Model
SA	Security Association
SDP	Service Discovery Protocol
SID	Session Identifier
SIM	Subscriber Identification Module
SIP	Session Initiation Protocol
SMS	Short Message Service
SMSC	Short Message Service Center
SNW	Sub Network
SPI	Security Parameter Index
SSID	Service Set Identifier
SSL	Secure Sockets Layer
TCP	Transmission Control Protocol
TDMA	Time Division Multiple Access
UA	User Agent
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
URI	Uniform Resource Identifier
USSD	Unstructured Supplementary Services Data
UWB	Ultra-Wide Bandwidth
VPN	Virtual Private Network
WEP	Wired Equivalent Privacy
WLAN	Wireless Local Area Network
WPAN	Wireless Personal Area Network
XML	eXtensible Markup Language

Bibliography

[1] (2000) 3GPP, Unstructured Supplementary Service Data, 4.90. The 3rd Generation Partnership Project (3GPP). Last visited 02.12.2005. [Online]. Available: http://www.3gpp.org/ftp/Specs/archive/04_series/04.90/

[2] (2004) 3GPP. The 3rd Generation Partnership Project (3GPP). Last visited 02.12.2005. [Online]. Available: http://www.3gpp.org/Default.htm

[3] (2004) 3GPP, Short Message Service, 3.40. The 3rd Generation Partnership Project (3GPP). Last visited 02.12.2005. [Online]. Available: http://www.3gpp.org/ftp/Specs/archive/03_series/03.40/

[4] (2004) 3GPP, Short Message Service, 3.41. The 3rd Generation Partnership Project (3GPP). Last visited 02.12.2005. [Online]. Available: http://www.3gpp.org/ftp/Specs/archive/03_series/03.41/

[5] G. N. Agglou and R. Tafazolli, "On the relaying capability of next-generation GSM celluar networks," *IEEE Personal Communications*, vol. 8, no. 1, pp. 40–47, february 2001.

[6] B. Ahlgren, L. Eggert, B. Ohlman, and A. Schieder, "Ambient networks: Bridging heterogeneous network domains," in *Proceedings of the 16th IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, september 2005, pp. 11–15.

[7] "Public key cryptography for the financial service industry: The elliptic curve digital signature algorithm. ANSI X9.62," American National Standards Institute, january 1998.

[8] J. G. Andrew T. Campbell and A. G. Valkó, "An overview of cellular IP," in *WCNC 1999 - IEEE Wireless Communications and Networking Conference, no. 1*, 1999, pp. 606–610.

[9] G. Appenzeller, K. Lai, P. Maniatis, M. Roussopoulos, E. Swierk, X. Zhao, and M. Baker, "The Mobile People Architecture, Tech. Rep. CSL-TR-99-777, 1999.

[10] A. Balachandran, G. M. Voelker, and P. Bahl, "Wireless hotspots: current challenges and future directions," in *WMASH '03: Proceedings of the 1st ACM international workshop on Wireless mobile applications and services on WLAN hotspots.* New York, NY, USA: ACM Press, 2003, pp. 1–9.

[11] L. Becchetti, F. D. Priscoli, T. Inzerilli, P. Mhnen, and L. Muoz, "Enhancing IP service provision over heterogeneous wireless networks: A path toward 4G," *IEEE Communications Magazine*, vol. 39, no. 8, pp. 74–81, august 2001.

[12] R. Berezdivin, R. Breinig, and R. Topp, "Next-generation wireless communications concepts and technologies," *IEEE Communications Magazine*, vol. 40, no. 3, pp. 108–116, march 2002.

[13] B. Bhargava, X. Wu, Y. Lu, and W. Wang, "Integrating heterogeneous wireless technologies: a cellular aided mobile ad hoc network (CAMA)," *Mob. Netw. Appl.*, vol. 9, no. 4, pp. 393–408, 2004.

[14] C. Bisdikian, "An overview of the Bluetooth wireless technology," *IEEE Communications Magazine*, vol. 39, no. 12, pp. 86–94, december 2001.

[15] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Communications of the ACM*, vol. 13, no. 7, pp. 422–426, 1970.

[16] Bluetooth core specification v1.2. Bluetooth Special Interest Group (SIG). Last visited 02.12.2005. [Online]. Available: https://www.bluetooth.org/spec/

[17] (2000-2005) BlueZ, Official Linux Bluetooth Protocol Stack. The BlueZ Project. Last visited 02.12.2005. [Online]. Available: http://www.bluez.org/

[18] E. A. Brewer, R. H. Katz, Y. Chawathe, S. D. Gribble, T. Hodes, G. Nguyen, M. Stemm, T. Henderson, E. Amir, H. Balakrishnan, A. Fox, V. N. Padmanabhan, and S. Seshan, "A network architecture for heterogeneous mobile computing," *IEEE Personal Communications Magazine*, vol. 5, no. 5, pp. 8–24, 1998.

[19] M. Buddhikot, A. Hari, K. Singh, and S. Miller, "MobileNAT: a new technique for mobility across heterogeneous address spaces," in *WMASH '03: Proceedings of the 1st ACM international workshop on Wireless mobile applications and services on WLAN hotspots.* New York, NY, USA: ACM Press, 2003, pp. 75–84.

[20] M. Calisti, T. Lozza, and D. Greenwood, "An agent-based middleware for adaptive roaming in wireless networks, AAMAS workshop on agents for ubiquitous computing, new york," July 2004.

[21] A. Calvagna and G. D. Modica, "A user-centric analysis of vertical handovers," in *WMASH '04: Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots.* New York, NY, USA: ACM Press, 2004, pp. 137–146.

[22] N. Cam-Winget, R. Housley, D. Wagner, and J. Walker, "Security flaws in 802.11 data link protocols," *Communications of the ACM*, vol. 46, no. 5, pp. 35–39, May 2003.

[23] T. Camp, J. Boleng, and V. Davies, “A survey of mobility models for ad hoc network research,” *Wireless Communications & Mobile Computing (WCMC): Special Issue on Mobile Ad Hoc Networking: Research, Trends, and Applications*, vol. 2, no. 5, pp. 483–502, Aug. 2002.

[24] A. T. Campbell, J. Gomez, S. Kim, C.-Y. W. András G. Valkó, Z. R. Turányi, and A. G. Valkó, “Design, implementation, and evaluation of cellular IP,” *IEEE Personal Communications Magazine*, vol. 7, no. 4, pp. 42–49, august 2000.

[25] A. T. Campbell, J. Gomez, S. Kim, C.-Y. Wan, Z. R. Turányi, and A. G. Valkó, “Comparison of IP micromobility protocols,” *IEEE Communications Magazine*, vol. 9, no. 1, pp. 72–82, february 2002.

[26] C. Castelluccia, “Extending Mobile IP with adaptive individual paging: a performance analysis,” *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, no. 2, pp. 14–26, 2001.

[27] C. Castelluccia and P. Mutaf, “An adaptive per-host IP paging architecture,” *SIGCOMM Comput. Commun. Rev.*, vol. 31, no. 5, pp. 48–56, 2001.

[28] I. Castineyra, N. Chiappa, and M. Steenstrup, “The Nimrod Routing Architecture, rfc 1992,” Internet Engineering Task Force (IETF), august 1996, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc1992.txt

[29] D. Cavalcanti, D. Agrawal, C. Cordeiro, B. Xie, and A. Kumar, “Issues in integrating cellular networks, WLANs, and MANETs: A futuristic heterogeneous wireless network,” *IEEE Communications Magazine*, vol. 12, no. 3, pp. 30–41, june 2005.

[30] K. Chakraborty, A. Misra, S. Das, A. McAuley, A. Dutta, and S. K. Das, “Implementation and performance evaluation of TeleMIP,” in *Proceedings of ICC 2001 - IEEE International Conference on Communications, no. 1*, june 2001, pp. 2488–2493.

[31] D. Cheriton and M. Gritter, “TRIAD: A scalable deployable NAT-based Internet Architecture, Stanford Computer Science Technical Report,” january 2000.

[32] T. Choi, L. Kim, J. Nah, and J. Song, “Combinatorial Mobile IP: a new efficient mobility management using minimized paging and local registration in Mobile IP environments,” *Wirel. Netw.*, vol. 10, no. 3, pp. 311–321, 2004.

[33] D. Clark, R. Braden, A. Falk, and V. Pingali, “FARA: reorganizing the addressing architecture,” in *FDNA ’03: Proceedings of the ACM SIGCOMM workshop on Future directions in network architecture.* New York, NY, USA: ACM Press, 2003, pp. 313–321.

[34] G. Cristache, K. David, and M. Hildebrand, “Aspects for the integration of ad-hoc and cellular networks,” in *Proceedings of the 3rd Scandinavian Workshop on Wireless Ad-hoc Networks*, Stockholm, May 2003.

[35] R. Daniel, “Resolution of URIs using the DNS, rfc 2168,” Internet Engineering Task Force (IETF), june 1997, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc2168.txt

[36] M. Danzeisen, “Secure Mobile IP Communication, Diploma thesis, University of Bern, mai 2001.”

[37] M. Danzeisen and T. Braun, “Access of Mobile IP users to firewall protected VPNs,” in *Proceedings of WLAN/GIWS, Workshop Mobile Communication over Wireless LAN at Informatik 2001, Vienna, September 26-29*, 2001, pp. 562–567.

[38] ——, “Secure Mobile IP Communcation,” in *Proceedings of 26th Annual IEEE Conference on Local Computer Networks (LCN'2001), Tampa, USA, Nov 15-16*, 2001.

[39] M. Danzeisen, T. Braun, D. Rodellar, and S. Winiker, “Heterogeneous networking establishment assisted by cellular operators,” in *Proceedings of MWCN, the fifth IFIP TC6 International Conference on Mobile and Wireless Communications Network, Singapore*, october 2003.

[40] ——, “Implementation of a cellular framework for spontaneous network,” in *Proceedings of IEEE WCNC, New Orleans, CD-ROM ISBN 0-7803-8967-0*, march 2005.

[41] ——, “Heterogeneous communication enabled by cellular operators,” *IEEE Vehicular Technology Society VTS*, feb 2006.

[42] M. Danzeisen, T. Braun, I. Steiner, and R.Rodellar, “On the benefits of heterogeneous networking and how cellular mobile operators can help,” in *ICPP Workshops, IEEE WSNET05, Oslo, Norway.* IEEE Computer Society, june 2005, pp. 366–371.

[43] M. Danzeisen and J. Linder, “Method and system for Mobile IP nodes in heterogeneous networks, EP 1 271 896 A2, 24 april 2002, priority data: 18 june 2001, Swisscom Mobile AG.”

[44] ——, “Method and system for Mobile IP nodes in heterogeneous networks, US 2002/0194385 A1, 19 december 2002, foreign application priority data: 18 june 2001 (EP) 01 810 594.0, Swisscom Mobile AG.”

[45] ——, “Method and system for Mobile IP nodes in heterogeneous networks, WO 02/103978 A2, 27 december 2002, priority data: 18 june 2001, Swisscom Mobile AG.”

[46] M. Danzeisen, D. Rodellar, S. Winiker, and T. Braun, “Heterogeneous Networking facilitated by Cellular Networks,” in *Comtec 03/2004*, pp. 18–21.

[47] S. Das, A. Misra, P. Agrawal, and S. K. Das, “TeleMIP: Telecommunications enhanced mobile ip architecture for fast intradomain mobility,” *IEEE Personal Communications Magazine*, vol. 7, no. 4, pp. 50–58, august 2000.

[48] C. de Waal. (2002-2005) BonnMotion - a mobility scenario generation and analysis tool. University of Bonn. Last visited 02.12.2005. [Online]. Available: http://www.cs.uni-bonn.de/IV/BonnMotion/

[49] O. Dousse, P. Thiran, and M. Hasler, "Connectivity in ad-hoc and hybrid networks," in *Proceedings IEEE Infocom*, New York, June 2002.

[50] R. Droms, "Dynamic host configuration protocol, rfc 2131," Internet Engineering Task Force (IETF), march 1997, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc2131.txt

[51] A. Dutta, O. Altintas, H. Schulzrinne, and W. Chen, "Multimedia SIP sessions in a mobile heterogeneous access environment," in *Proceedings 3G Wireless*, San Francisco, USA, may 2002.

[52] A. Dutta, T. Zhang, S. Madhani, K. Taniuchi, K. Fujimoto, Y. Katsube, Y. Ohba, and H. Schulzrinne, "Secure universal mobility for wireless internet," in *WMASH '04: Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*. New York, NY, USA: ACM Press, 2004, pp. 71–80.

[53] D. Eastlake, "Secure Domain Names System dynamic update, rfc 2137," Internet Engineering Task Force (IETF), april 1997, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc2137.txt

[54] R. Elz and R. Bush, "Clarifications to the DNS specification, rfc 2181," Internet Engineering Task Force (IETF), july 1997, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc2181.txt

[55] K. M. et al., "On the required features and system capacity of basic access network in the MIRAI," in *WPMC*, september 2001, pp. 1199–1204.

[56] X.-X. W. et al., "MADF: Mobile-Assisted Data Forwarding for wireless data network," *J. Commun. and Network*, 2002.

[57] EURESCOM Project P1013-FIT-MIP. First steps towards umts: Mobile IP services, a European testbed. EURESCOM. Last visited 02.12.2005. [Online]. Available: http://www.eurescom.de/public/projects/P1000-series/p1013/default.asp

[58] S. M. Faccin, P. Lalwaney, and B. Patil, "IP multimedia services: Analysis of Mobile IP and SIP interactions in 3G networks," *IEEE Communications Magazine*, vol. 42, no. 1, pp. 113–120, january 2004.

[59] P. Faltstrom, "Electronic numbering and DNS, rfc 2916," Internet Engineering Task Force (IETF), september 2000, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc2916.txt

[60] L. M. Feeney and M. Nilsson, "Investigating the energy consumption of a wireless network interface in an ad hoc networking environment," in *Proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '01)*, Anchorage, AK, USA, Apr. 2001, pp. 1548–1557.

[61] D. Funato, K. Yasuda, and H. Tokuda, “TCPR: TCP mobility support for continuous operation,” in *Proceedings of IEEE International Conference on Network Protocols 97*, 1997, pp. 229–236.

[62] M. Gastpar and M. Vetterli, “The capacity of wireless networks: The relay case,” 2002.

[63] V. Gazis, N. Alonistioti, and L. Merakos, “Toward a generic always best connected capability in integrated WLAN/UMTS cellular mobile networks (and beyond),” *IEEE Wireless Communications Magazine*, vol. 12, pp. 20–29, june 2005.

[64] I. N. Golem, “Mobile Unlimited, Unifying WLAN, UMTS, and GPRS,” 2004, last visited 02.12.2005. [Online]. Available: http://www.golem.de/0406/31554.html

[65] C. Guo, Z. Guo, Q. Zhang, and W. Zhu, “A seamless and proactive end-to-end moblility solution for roaming accross heterogeneous wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 22, pp. 834–848, june 2004.

[66] E. Gustafsson and A. Jonsson, “Always Best Connected,” *IEEE Wireless Communications Magazine*, vol. 10, pp. 49–55, february 2003.

[67] C. T. Hager and S. F. Midkiff, “An analysis of Bluetooth security vulnerabilities,” in *Proceedings of IEEE WCNC03*, 2003, pp. 1825–1831.

[68] ——, “Demonstrating vulnerabilities in bluetooth security,” in *Proceedings of IEEE GLOBECOM03*, 2003, pp. 1420–1424.

[69] M. Handley and V. Jacobson, “Session Description Protocol (sip), rfc 2327,” Internet Engineering Task Force (IETF), april 1998, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc2327.txt

[70] D. Harkins and D. Carrel, “Internet key exchange, rfc 2409,” Internet Engineering Task Force (IETF), november 1998, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc2409.txt

[71] D. Harrington, R. Presuhn, and B. Wijnen, “Simple Network Management Protocol (snmp), rfc 3411,” Internet Engineering Task Force (IETF), december 2002, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc3411.txt

[72] H. Haverinen and J. Salowey, “Extensible Authentication Protocol Method for GSM Subscriber Identity Modules (EAP-SIM), Internet Draft (work in progress),” Internet Engineering Task Force (IETF), december 2004, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/internet-drafts/draft-haverinen-pppext-eap-sim-16.txt

[73] H. Haverinen, J. Mikkonen, and T. Takamki, “Cellular access control and charging for mobile operator wireless local area networks,” *IEEE Wireless Communications Magazine*, vol. 9, no. 6, pp. 52–60, december 2002.

[74] M. Heissenbuettel, “Routing and broadcasting in ad-hoc networks, PhD thesis, University of Bern, june 2005.”

[75] T. Henderson, “End-Host Mobility and Multihoming with the Host Identity Protocol, Internet Draft (work in progress),” Internet Engineering Task Force (IETF), june 2005, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/internet-drafts/draft-ietf-hip-mm-02.txt

[76] T. R. Henderson, “Host mobility for IP networks: A comparison,” *IEEE Network*, vol. 17, no. 6, pp. 18–26, november 2003.

[77] T. R. Henderson, J. M. Ahrenholz, and J. H. Kim, “Experience with the host identity protocol for secure host mobility and multihoming,” in *Proceedings of IEEE WCNC03*, 2003, pp. 2120–2125.

[78] R. Housley, W. Polk, W. Ford, and D. Solo, “Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile, rfc 3280,” Internet Engineering Task Force (IETF), april 2002, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc3280.txt

[79] H.-Y. Hsieh and R. Sivakumar, “Performance comparison of cellular and multi-hop wireless networks: a quantitative study,” in *SIGMETRICS ’01: Proceedings of the 2001 ACM SIGMETRICS international conference on Measurement and modeling of computer systems.* New York, NY, USA: ACM Press, 2001, pp. 113–122.

[80] ——, “A hybrid network model for cellular wireless packet data networks,” in *GLOBECOM 2002 - IEEE Global Telecommunications Conference, vol. 21, no. 1*, november 2002, pp. 971–975.

[81] ——, “On using the ad-hoc network model in cellular packet data networks,” in *MobiHoc ’02: Proceedings of the 3rd ACM international symposium on Mobile ad hoc networking & computing.* ACM Press, 2002, pp. 36–47.

[82] ——, “Towards a hybrid network model for wireless packet data networks,” in *ISCC ’02: Proceedings of the Seventh International Symposium on Computers and Communications (ISCC’02).* Washington, DC, USA: IEEE Computer Society, 2002, p. 264.

[83] R. Hsieh and A. Seneviratne, “A comparison of mechanisms for improving Mobile IP handoff latency for end-to-end TCP,” in *MobiCom ’03: Proceedings of the 9th annual international conference on Mobile computing and networking.* New York, NY, USA: ACM Press, 2003, pp. 29–41.

[84] G. Huston, “To NAT or IPv6 -That is the question,” december 2000, last visited 02.12.2005. [Online]. Available: http://www.potaroo.net/ispcol/2001-01/2001-01-ipv6.pdf

[85] H. Y. Hwang, S. J. Kwon, Y. W. Chung, and D. K. Sung, “A mobility management scheme for reducing power consumption in IP-based wireless networks,” in *Proceedings of GLOBECOM 2002 - IEEE Global Telecommunications Conference, vol. 21, no. 1*, november 2002, pp. 2986–2990.

[86] (2005) Ambient Networks Project. Information Society Technology (IST). Last visited 02.12.2005. [Online]. Available: http://www.ambient-networks.org/

[87] M. Inoue, K. Mahmud, H. Murakami, and M. Hasegawa, "MIRAI: A solution to seamless access in heterogeneous wireless networks," in *Proceedings of the IEEE ICC*, 2003, pp. 1033–37.

[88] M. Inoue, K. Mahmud, H. Murakami, M. Hasegawa, and H. Morikawa, "Design and implementation of out-of-band signaling for seamless handover in wireless overlay networks," in *Proceedings of the IEEE ICC*, june 2004, pp. 3932–3936.

[89] ——, "Novel out-of-band signaling for seamless interworking between heterogeneous networks," *IEEE Wireless Communications Magazine*, vol. 11, no. 2, pp. 56–63, april 2004.

[90] IEEE 802.11 standard for local and metropolitan area networks. The Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ, USA. Last visited 02.12.2005. [Online]. Available: http://www.ieee802.org/11/

[91] IEEE 802.11a standard for local and metropolitan area networks. The Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ, USA. Last visited 02.12.2005. [Online]. Available: http://grouper.ieee.org/groups/802/11/index.html

[92] IEEE 802.11b standard for local and metropolitan area networks. The Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ, USA. Last visited 02.12.2005. [Online]. Available: http://grouper.ieee.org/groups/802/11/index.html

[93] IEEE 802.11e standard for local and metropolitan area networks. The Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ, USA. Last visited 02.12.2005. [Online]. Available: http://grouper.ieee.org/groups/802/11/index.html

[94] IEEE 802.15 working group for wireless personal area networks (WPAN). The Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ, USA. Last visited 02.12.2005. [Online]. Available: http://www.ieee802.org/15/

[95] IEEE 802.16 standard for local and metropolitan area networks. The Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ, USA. Last visited 8.11.2005. [Online]. Available: http://www.ieee802.org/16/published.html

[96] IEEE 802.21 working group for Media Independent Handover (MIH). The Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ, USA. Last visited 02.12.2005. [Online]. Available: http://grouper.ieee.org/groups/802/21/

[97] (2005) The International Public Telecommunication Numbering Plan, E.164. Interantional Telecommunication Union. Last visited 02.12.2005. [Online]. Available: http://www.itu.int/ITU-T/

[98] (2005) Signaling System 7, SS7. Interantional Telecommunication Union. Last visited 02.12.2005. [Online]. Available: http://www.itu.int/ITU-T/

[99] "IP Routing for Wireless/Mobile Hosts (mobileip)," Internet Engineering Task Force (IETF), 2003, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/html.charters/mip4-charter.html

[100] "Layer 3 Virtual Private Networks (l3vpn)," Internet Engineering Task Force (IETF), 2004, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/html.charters/l3vpn-charter.html

[101] (2004, Nov.) Mobile Ad-hoc Networks (manet) Working Group. Internet Engineering Task Force (IETF). Last visited 02.12.2005. [Online]. Available: http://www.ietf.org/html.charters/manet-charter.html

[102] "Mobility for IPv6 (mipv6)," Internet Engineering Task Force (IETF), 2005, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/html.charters/mip6-charter.html

[103] V. Jacobson, "Congestion avoidance and control," in *Proceedings of ACM SIGCOMM88*, august 1988, pp. 314–329.

[104] M. Jakobsson and S. Wetzel, "Security Weaknesses in Bluetooth," *Lecture Notes in Computer Science*, vol. 2020, pp. 176+, 2001.

[105] A. D. Joseph, J. A. Tauber, and M. F. Kaashoek, "Mobile computing with the rover toolkit," *IEEE Transactions on Computers*, vol. 46, no. 3, pp. 337–352, 1997.

[106] J.-W. Jung, R. Mudumbai, D. Montgomery, and H.-K. Kahng, "Performance evaluation of two layered mobility management using Mobile IP and Session Initiation Protocol," in *GLOBECOM 2003 - IEEE Global Telecommunications Conference, vol. 22, no. 1, Dec 2003, pp. 1190 - 1194*, 2003, pp. 1190 – 1194.

[107] C. Kaufman, "Internet Key Exchange v2, draft-ietf-ipsec-ikev2-17.txt," Internet Engineering Task Force (IETF), september 2004, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/internet-drafts/draft-ietf-ipsec-ikev2-17.txt

[108] W. Kellerer, H.-J. Vgel, and K.-E. Steinberg, "A communication gateway for infrastructure-independent 4G wireless access," *IEEE Communications Magazine*, vol. 40, no. 3, pp. 126–131, march 2002.

[109] J. Kempf, "Dormant Mode Host Alerting (IP Paging) Problem Statement, rfc 3132," Internet Engineering Task Force (IETF), june 2001, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc3132.txt

[110] J. Kempf, C. Castelluccia, P. Mutaf, N. Nakajima, Y. Ohba, R. Ramjee, Y. Saifullah, B. Sarikaya, and X. Xu, "Requirements and Functional Architecture for an IP Host Alerting Protocol, rfc 3154," Internet Engineering Task Force (IETF), august 2001, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc3154.txt

[111] J. Kempf and P. Mutaf, "IP paging considered unnecessary: Mobile IPv6 and IP paging for dormant mode location update in macrocellular and

hotspot networks," in *Proceedings of WCNC 2003 - IEEE Wireless Communications and Networking Conference, vol. 4, no. 1*, march 2003, pp. 1032–1036.

[112] S. Kent and R. Atkinson, "Authentication Header (AH), rfc 2402," Internet Engineering Task Force (IETF), november 1998, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc2402.txt

[113] ——, "IP Encapsulating Security Payload (ESP), rfc 2406," Internet Engineering Task Force (IETF), november 1998, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc2406.txt

[114] S.-E. Kim and J. A. Copeland, "TCP for seamless vertical handoff in hybrid mobile data networks," in *Proceedings of IEEE GLOBECOM03*, vol. 22, no. 1, 2003, pp. 661–665.

[115] R. Kravets, C. Carter, and L. Magalhaes, "A cooperative approach to user mobility," *SIGCOMM Comput. Commun. Rev.*, vol. 31, no. 5, pp. 57–69, 2001.

[116] T. T. Kwon, M. Gerla, S. Das, and S. Das, "Mobility management for VoIP service: Mobile IP vs. SIP," *IEEE Wireless Communications Magazine*, vol. 9, no. 5, pp. 66–75, october 2002.

[117] B. Landfeldt, T. Larsson, Y. Ismailov, and A. Seneviratne, "SLM, a framework for session layer mobility management, international conference on computer communications and networks (ICCCN)," in *Proceedings of ICCN99*, 1999, pp. 452–456.

[118] E. J. Latvakoski and P. Pääkkönen, "Remote interaction with networked appliances attached in a mobile personal area network," in *Proceedings of ICC 2003 - IEEE International Conference on Communications, vol. 26, no. 1*, 2003, pp. 769–773.

[119] J. Latvakoski, D. Pakkala, and P. Pääkkönen, "A communication architecture for spontaneous systems," *IEEE Communications Magazine*, vol. 11, no. 3, pp. 36–42, june 2004.

[120] B. A. Laura Marie Feeney and A. Westerlund, "Spontaneous networking: An application-oriented approach to ad hoc networking," *IEEE Communications Magazine*, vol. 39, no. 6, pp. 176–181, june 2001.

[121] Y.-D. J. Lin and Y.-C. Hsu, "Multihop cellular: A new architecture for wireless communications," in *INFOCOM*, 2000, pp. 1273–1282.

[122] (2005) Merling U630 Wireless PC Card Modem. Lucent Technologies. Last visited 02.12.2005. [Online]. Available: http://www.lucent.com/livelink/090094038003841f_Brochure_datasheet.pdf

[123] (2005) Skype technologies. Lucent Technologies S.A. Last visited 02.12.2005. [Online]. Available: http://www.skype.com

[124] H. Luo, R. Ramjee, P. Sinha, L. E. Li, and S. Lu, “UCAN: a unified cellular and ad-hoc network architecture,” in *MobiCom ’03: Proceedings of the 9th annual international conference on Mobile computing and networking.* New York, NY, USA: ACM Press, 2003, pp. 353–367.

[125] E. Maghsoodi, “Design and implementation of WLAN support for Cellular Assisted Heterogeneous Networking, Diploma thesis, University of Bern,” november 2004.

[126] K. Mahmud, G. Wu, Y. Hase, and M. Mizuno, “Using variable rate transmission for capacity improvement of basic access network in MIRAI,” in *Proceedings of Wireless Personal Multimedia Conference, vol. 3*, september 2001, pp. 1199–1204.

[127] K. Mahmud, G. Wu, M. Inoue, and M. Mizuno, “Mobility management by basic access network in MIRAI architecture for heterogeneous wireless systems,” in *Proceedings of GLOBECOM 2002 - IEEE Global Telecommunications Conference, vol. 21, no. 1*, november 2002, pp. 1718–1722.

[128] D. A. Maltz and P. Bhagwat, “MSOCKS: An architecture for transport layer mobility,” in *INFOCOM (3)*, 1998, pp. 1037–1045.

[129] P. Maniatis, M. Roussopoulos, E. Swierk, K. Lai, G. Appenzeller, X. Zhao, and M. Baker, “The Mobile People Architecture,” *ACM Mobile Computing and Communications Review*, july 1999.

[130] A. N. Marco Carli and A. R. Picci, “Mobile IP and Cellular IP integration for inter access networks handoff,” in *Proceedings of ICC 2001 - IEEE International Conference on Communications, no. 1*, june 2001, pp. 2467–2471.

[131] A. McAuley and R. Morera, “Name and address decoupling in support of dynamic networks,” in *Proceedings IEEE MILCOM02, vol. 21, no. 1*, 2002, pp. 969–974.

[132] J. McNair and F. Zhu, “Vertical handoffs in fourth-generation multinetwork environments,” *IEEE Communications Magazine*, vol. 11, no. 3, pp. 8–15, june 2004.

[133] M. Mealling and R. Daniel, “Naming Authority Pointer, rfc 2915,” Internet Engineering Task Force (IETF), september 2000, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc2915.txt

[134] D. L. Mills, “Network time protocol (version 3) specification, implementation and analysis, rfc 1305,” Internet Engineering Task Force (IETF), march 1992, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc1305.txt

[135] P. Mockapetris, “Domain Names, rfc 1035,” Internet Engineering Task Force (IETF), november 1987, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc1035.txt

[136] G. Montenegro, “IP Mobility Support for IPv4, rfc 3024,” Internet Engineering Task Force (IETF), january 2001, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc3024.txt

[137] G. Montenegro and C. Castelluccia, “Statistically Unique and Cryptographically Verifiable SUCV identifiers and addresses, in NDSS'02, february,” 2002.

[138] R. Moskowitz and P. Nikander, “Host Identity Protocol Architecture, Internet Draft (work in progress),” Internet Engineering Task Force (IETF), august 2005, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/internet-drafts/draft-ietf-hip-base-04.txt

[139] R. Moskowitz, P. Nikander, P. Jokela, and T. Henderson, “Host Identity Protocol, Internet Draft (work in progress),” Internet Engineering Task Force (IETF), june 2005, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/internet-drafts/draft-ietf-hip-base-03.txt

[140] P. Mutaf and C. Castelluccia, “Hash-based paging and location update using bloom filters: a paging algorithm that is best suitable for IPv6,” *Mob. Netw. Appl.*, vol. 9, no. 6, pp. 627–631, 2004.

[141] N. Nakajima, A. Dutta, S. Das, and H. Schulzrinne, “Handoff delay analysis and measurement for SIP based mobility in IPv6,” in *ICC 2003 - IEEE International Conference on Communications, vol. 26, no. 1*, may 2003, pp. 1085–1089.

[142] N. Niebert, A. Schieder, H. Abramowicz, G. Malmgren, J. Sachs, U. Horn, C. Prehofer, and H. Karl, “Ambient networks: An architecture for communication networks beyond 3G,” *IEEE Communications Magazine*, vol. 11, no. 2, pp. 14–22, april 2004.

[143] (2005) Open Mobile Alliance (OMA). Wireless Application Protocol (WAP). OMA. Last visited 02.12.2005. [Online]. Available: http://www.openmobilealliance.org/tech/affiliates/wap/wapindex.html

[144] A. W. O'Neill and G. Tsirtsis, “Edge mobility architecture: Routeing and hand-off,” *BT Technology Journal*, vol. 19, no. 1, pp. 114–126, 2001.

[145] (2005, Nov.) OPNET. OPNET Technologies, Inc. Last visited 02.12.2005. [Online]. Available: http://www.opnet.com/home.html

[146] Open Service Gateway Initiative, OSGi. Last visited 02.12.2005. [Online]. Available: http://www.osgi.org

[147] K. Pahlavan, P. Krishnamurthy, A. Hatami, M. Ylianttila, J.-P. Makela, R. Pichna, and J. Vallstrm, “Handoff in hybrid mobile data networks,” *IEEE Personal Communications Magazine*, vol. 7, no. 2, pp. 34–47, april 2000.

[148] S. Panagiotakis and A. Alonistioti, “Intelligent service mediation for supporting advanced location and mobility-aware service provisioning in reconfigurable mobile networks,” *IEEE Wireless Communications Magazine*, vol. 9, no. 5, pp. 28–38, december 2002.

[149] C. Perkins, “Minimal Encapsulation within IP, rfc 2004,” Internet Engineering Task Force (IETF), october 1996, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc2004.txt

[150] ——, "IP Mobility Support, rfc 3344," Internet Engineering Task Force (IETF), january 2002, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc3344.txt

[151] C. Perkins and D. B. Johnson, "Internet Draft: Route Optimization in Mobile IP," Internet Engineering Task Force (IETF), work in progress, september 2001.

[152] C. Politis, K. A. Chew, and R. Tafazolli, "Multilayer mobility management for All-IP networks: Pure SIP vs. hybrid SIP/Mobile IP," in *IEEE VTC Spring Conference, Jeju, Korea, 21-24th April 2003*, april 2003.

[153] C. Qiao and H. Wu, "iCAR: an integrated cellular and ad-hoc relay system," 2000.

[154] X. Qu, J. X. Yu, and R. P. Brent, "A mobile TCP socket," Canberra 0200 ACT, Australia, Tech. Rep. TR-CS-97-08, 1997.

[155] E. R. Braden, "Requirements for Internet Hosts - application and support, rfc 1123," Internet Engineering Task Force (IETF), october 1989, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc1123.txt

[156] P. F. Ramakrishna, "IPNL: A NAT-extended internet architecture," in *SIGCOMM '01: Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications.* New York, NY, USA: ACM Press, 2001, pp. 69–80.

[157] R. Ramanathan, "Mobility Support for Nimrod : Challenges and Solution Approaches, rfc 2103," Internet Engineering Task Force (IETF), february 1997, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc2103.txt

[158] R. Ramjee, L. Li, T. L. Porta, and S. Kasera, "IP paging service for mobile hosts," in *MobiCom '01: Proceedings of the 7th annual international conference on Mobile computing and networking.* New York, NY, USA: ACM Press, 2001, pp. 332–345.

[159] R. Ramjee, L. Li, T. F. L. Porta, and S. K. Kasera, "IP paging service for mobile hosts," in *Mobile Computing and Networking*, 2001, pp. 332–345.

[160] R. Ramjee, L. Li, T. la Porta, and S. Kasera, "IP paging service for mobile hosts," *Wirel. Netw.*, vol. 8, no. 5, pp. 427–441, 2002.

[161] R. Ramjee, K. Varadhan, L. Salgarelli, S. R. Thuel, S.-Y. Wang, and T. L. Porta, "HAWAII: A domain-based approach for supporting mobility in wide-area wireless networks," *IEEE/ACM Transactions on Networking*, vol. 10, no. 3, pp. 396–410, 2002.

[162] (2005, august) The Fedora Project. Red Hat Enterprise Linux. Last visited 02.12.2005. [Online]. Available: http://fedora.redhat.com/about/

[163] P. Reinbold and O. Bonaventure, "IP micro-mobility protocols," *IEEE Communications Surveys I& Tutorials*, vol. 5, no. 1, pp. 40–57, 2003.

[164] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "Session Initiation Protocol (sip), rfc 3261," Internet Engineering Task Force (IETF), june 2002, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc3261.txt

[165] M. Roussopoulos, P. Maniatis, E. Swierk, K. Lai, G. Appenzeller, and M. Baker, "Person-level routing in the Mobile People Architecture," in *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, october 1999.

[166] D. Saha, A. Mukherjee, I. S. Misra, and M. Chakraborty, "Mobility support in IP: A survey of related protocols," *IEEE Network*, vol. 18, no. 6, Nov. 2004.

[167] J. H. Saltzer, D. P. Reed, and D. D. Clark, "End-To-End arguments in system design," *ACM Transactions on Computer Systems*, vol. 2, no. 4, pp. 277–288, Nov. 1984.

[168] A. Sanmateu, F. Paint, L. Morand, S. Tessier, P. Fouquart, A. Sollund, and E. Bustos, "Seamless mobility accross IP networks using mobile IP," *Computer Networks*, vol. 40, pp. 181–190, 2002.

[169] A. Saulnier, "Heterogeneous and spontaneous VPN networking using SIP, Diploma thesis, EURECOM," september 2005.

[170] (2004, Nov.) Qualnet. Scalable Network Technologies (SNT). Last visited 02.12.2005. [Online]. Available: http://www.qualnet.com/

[171] H. Schulzrinne and E. Wedlund, "Application-layer mobility using SIP," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 4, no. 3, pp. 47–57, 2000.

[172] P. D. Silva and H. Sirisena, "A mobility management protocol for IP-based cellular networks," *IEEE Wireless Communications Magazine*, vol. 9, no. 3, pp. 31–37, june 2002.

[173] W. Simpson, "Point-to-Point Protocol, rfc 1661," Internet Engineering Task Force (IETF), july 1994, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc1661.txt

[174] A. C. Snoeren and H. Balakrishnan, "An end-to-end approach to host mobility," in *Proceedings of MOBICOM*, 2000, pp. 155–166.

[175] A. C. Snoeren, H. Balakrishnan, and M. F. Kaashoek, "Reconsidering internet mobility," in *Proc. 8th Workshop on Hot Topics in Operating Systems (HotOS-VIII)*, 2001.

[176] P. Srisuresh and K. Egevang, "Network Address Translation, rfc 3022," Internet Engineering Task Force (IETF), january 2001, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc3022.txt

[177] I. Steiner, "Synergy of peer-to-peer and infrastructure based networks. Diploma thesis, University of Bern," june 2005.

[178] I. Stoica, D. Adkins, S. Zhuang, S. Shenker, and S. Surana, "Internet indirection infrastructure," in *SIGCOMM '02: Proceedings of the 2002 conference on Applications, Technologies, Architectures, and Protocols for Computer Communications.* New York, NY, USA: ACM Press, 2002, pp. 73–86.

[179] J. Tourrilhes. (1996, Oct.) Linux Programmers Manual, iwconfig(8). Last visited 02.12.2005. [Online]. Available: http://linuxcommand.org/man_pages/iwconfig8.html

[180] C. Tschudin, H. Lundgren, and H. Gulbrandsen, "Active routing for ad-hoc networks," *IEEE Communications Magazine*, vol. 38, no. 4, pp. 122–127, april 2000.

[181] Z. Turányi and A. Valkó, "4+4: Expanding the Internet Address Space without IPv6, Ericsson internal report," august 1999.

[182] Z. Turányi, A. Valkó, and A. T. Campbell, "4+4: an architecture for evolving the internet address space back toward transparency," *SIGCOMM Comput. Commun. Rev.*, vol. 33, no. 5, pp. 43–54, 2003.

[183] (2005) Ultra Lab, University of Southern Clifornia. University of Southern California. Last visited 02.12.2005. [Online]. Available: http://ultra.usc.edu/New_Site/publications.html

[184] The Network Simulator - ns-2. University of Southern California, Information Science Institute ISI. Last visited 02.12.2005. [Online]. Available: http://www.isi.edu/nsnam/ns/

[185] F. N. van Kempen, A. Cox, P. Blundell, and A. Kleen. (2000, Aug.) Linux Programmers Manual, ifconfig(8). Last visited 02.12.2005. [Online]. Available: http://linuxcommand.org/man_pages/ifconfig8.html

[186] E. Vanem, S. Svaet, and F. Paint, "Effects on multiple access alternatives in heterogeneous wireless networks," in *Proceedings of IEEE WCNC 2003*, 2003.

[187] P. Vixie, S. Thomson, Y. Rekhter, and J. Bound, "Dynamic updates in the domain name system (DNS UPDATE, rfc 2136," Internet Engineering Task Force (IETF), april 1997, last visited 02.12.2005. [Online]. Available: http://www.ietf.org/rfc/rfc2136.txt

[188] H. J. Wang, B. Raman, C.-N. Chuah, R. Biswas, R. Gummadi, B. Hohlt, X. Hong, E. Kiciman, Z. Mao, J. S. Shih, L. Subramanian, B. Y. Zhao, A. D. Joseph, and R. H. Katz, "ICEBERG: An internet core network architecture for integrated communications," *IEEE Personal Communications Magazine*, vol. 7, no. 4, pp. 10–19, 2000.

[189] Q. Wang, M. A. Abu-Rgheff, and A. Akram, "Design and evaluation of an integrated mobile IP and SIP framework for advanced handoff management," in *ICC 2004 - IEEE International Conference on Communications, vol. 27, no. 1, June 2004*, 2004, pp. 3921–3925.

[190] E. Wedlund and H. Schulzrinne, "Mobility support using SIP," in *WOWMOM*, 1999, pp. 76–82.

[191] H.-Y. Wei and R. D. Gitlin, "Two-hop-relay architecture for next-generation WWAN/WLAN integration," *IEEE Wireless Communications Magazine*, vol. 11, no. 2, pp. 24–30, april 2004.

[192] S. Winiker, "Integration of Cellular Assisted Heterogeneous Networking and bluetooth service discovery protocol, Diploma thesis, University of Bern," may 2004.

[193] D. Wisely, H. Aghvami, S. L. Gwyn, T. Zahariadis, J. Manner, V. Gazis, N. Houssos, and N. Alonistioti, "Transparent ip radio access for next-generation mobile networks," *IEEE Wireless Communications Magazine*, vol. 10, no. 4, pp. 26–35, august 2003.

[194] K. D. Wong, A. Dutta, K. Young, and H. Schulzrinne, "Managing simultaneous mobility of ip hosts," in *Proceedings of the MILCOM, vol. 22, no. 1*, october 2003, pp. 785–790.

[195] W. S. V. Wong and V. C. M. Leung, "Location management for next-generation personal communications networks," *IEEE Network*, vol. 14, no. 5, pp. 18–24, september/october 2000.

[196] G. Wu, M. Mizuno, and P. J. M. Havinga, "MIRAI architecture for heterogeneous network," *IEEE Communications Magazine*, vol. 40, no. 2, pp. 126–134, february 2002.

[197] H. Wu, C. Qiao, S. De, and O. Tonguz, "Integrated cellular and ad hoc relaying systems: iCAR," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp. 2105–2115, october 2001.

[198] S.-L. Wu, T.-K. Lin, Y.-C. Tseng, and J.-P. Sheu, "Route optimization on wireless mobile ad-hoc networks," in *In The 5th Mobile Computing Workshop*, Taiwan, 1999, pp. 143–150.

[199] S.-H. G. C. Xiaoxin Wu and B. Mukherjee, "MADF: A novel approach to add an ad-hoc overlay on a fixed cellular infrastructure," in *Proceedings of WCNC 2000 - IEEE Wireless Communications and Networking Conference, no. 1,*, september 2000, pp. 549–554.

[200] J. Xie, "Paging-aided connection setup for real-time communication in mobile internet," in *Proceedings of ICC 2003 - IEEE International Conference on Communications, vol. 26, no. 1*, 2003, pp. 1858–1862.

[201] C.-H. Yeh, "ACENET: Architectures and Protocols for High Throughput, Low Power, and QoS Provisioning in Next-Generation Mobile Communications," in *Proceedings of PIMRC02*, 2002.

[202] M. Ylianttila, M. Pande, J. Mäkelä, and P. Mähönen, "Optimization scheme for mobile users performing vertical handoffs between IEEE 802.11 and GPRS/EDGE networks," in *IEEE GLOBECOM 2001, no. 1*, 2001, pp. 3439–3443.

[203] J. Yoon, M. Liu, and B. Noble, "Random Waypoint considered harmful," in *IEEE Computer and Communications Societies (INFOCOM 2003), San Francisco, USA*, 2003.

[204] Q. Zhang, C. Guo, Z. Guo, and W. Zhu, "Efficient mobility management for vertical handoff between wwan and wlan," *IEEE Communications Magazine*, vol. 41, pp. 102–108, november 2003.

[205] X. Zhang, J. Castellanos, and A. T. Campell, "Design and Performance of Mobile IP Paging," *ACM Mobile Networks and Applications (MONET), Special issue on Modeling Analysis and Simulation of Wireless and Mobile Systems*, vol. 7, no. 2, march 2002.

[206] X. Zhang, J. G. Castellanos, and A. T. Campbell, "P-MIP: paging in mobile IP," in *WOWMOM '01: Proceedings of the 4th ACM international workshop on Wireless mobile multimedia.* New York, NY, USA: ACM Press, 2001, pp. 44–54.

[207] Y. Zhang, H. Vin, L. Alvisi, W. Lee, and S. K. Dao, "Heterogeneous networking: a new survivability paradigm," in *NSPW '01: Proceedings of the 2001 workshop on New security paradigms.* New York, NY, USA: ACM Press, 2001, pp. 33–39.

[208] F. Zhu and J. McNair, "Optminizations for vertical handoff decision algorithms," in *Proceedings of the IEEE WCNC04*, 2004, pp. 867–872.

[209] S. Zhuang, K. Lai, I. Stoica, R. Katz, and S. Shenker, "Host mobility using an internet indirection infrastructure, tech. rep. ucb/csd-02-1186, computer science division, u. c. berkeley," 2002.

[210] M. Zivkovic, K. Lagerberg, and J. van Bemmel, "Secure seamless roaming over heterogeneous networks, Bell Labs Advanced Technologies EMEA The Netherlands. IST Project AlbatrOSS, february," 2004.

Acknowledgements

The work presented in this thesis was accomplished in tight collaboration with Swisscom Innovations, the Research and Development department of the incumbent telecommunication operator of Switzerland. This collaboration positively influenced the work by focusing on actual problems raised by the convergence of the telecommunication and information technology. Results of the thesis are further elaborated within Swisscom internal projects aiming at the development of new products. The proposed system architecture and protocol form the basis for an operator assisted data communication framework currently further developed by Swisscom Innovations to seamless integrate the different available networking technologies.

Curriculum Vitae

1974	Born on February, 2 in Rio de Janeiro, Brazil
1981 - 1985	Elementary School Muri, Bern, Switzerland
1985 - 1990	Secondary School Gümligen
1990 - 1994	Gymnasium Bern-Kirchenfeld, Typus E
1994 - 1996	University of Neuchâtel, Switzerland
1996 - 2001	University of Bern, Switzerland. Major in Computer Science and Minors in Mathematics and Micro-Electronics at the University of Neuchâtel
2001	M.Sc. in Computer Science, University of Bern
2001 - 2005	Ph.D. Student and Research Assistant at the Institute for Computer Science and Applied Mathematics, University of Bern
Since 2001	R&D Consultant at Swisscom Innovations, Bern