Integration and Management of Multiple Radio Access Technologies in Converged Wireless Networks

Joachim Sachs

Berlin 2009

Integration and Management of Multiple Radio Access Technologies in Converged Wireless Networks

vorgelegt von
Diplom-Ingenieur
Joachim Sachs

von der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
– Dr.-Ing. –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzende: Prof. Dr. Anja Feldmann
Berichter: Prof. Dr. Adam Wolisz
Berichterin: Prof. Dr. Carmelita Görg

Tag der wissenschaftlichen Aussprache: 2.3.2009

Berlin 2009

D 83

Abstract

Wireless communication networks enable customers to use communication services – like telephony or Internet services – at anytime and anywhere. A multitude of varying radio access technologies has been developed or is currently being developed. Those access technologies have different characteristics: they differ in their radio coverage, their spectral efficiency, cell capacity, and peak data rate; they can support different services; some support user mobility; they differ in their complexity and costs. For a network operator different radio access technologies may best be suited in different parts of the network, depending on the radio environment, the expected traffic pattern, and the anticipated services. The dynamic selection of access technology according to the communication requirements and network characteristics enable end users to be *always best connected*. Therefore, it is desirable to allow various access technologies to be combined in a common network and to allow interworking of different networks to be possible, even when they deploy different access technologies. The integration of multiple heterogeneous access technologies into a common network is referred to as *network convergence*. In this work we investigate technical solutions for integrating and managing different access technologies in a converged network and derive suitable network architectures. We define three types of functionality that are required for multi-access management: *access monitoring, access selection* and *access handover*.

Access selection is the functionality that evaluates different accesses in order to determine which one is best suited for a certain communication session. This decision can be based on a number of parameters, like the characteristics of the radio link, the availability of resources in the different access systems, the requirements of the service, or policies of the user or network operator, which may depend on the business arrangement that the user has agreed to with the network operator. Access selection is performed by one or more *multi-radio resource management* functions which are typically located in multiple network nodes. Based on system simulations we evaluate the gain that can be achieved by access selection for different types of radio access technologies, radio network layouts, user distributions, and access selection algorithms. We show that the capacity gain of the system depends strongly on the considered scenario; in most cases significant gain can be achieved by access selection. We furthermore develop a model of a multi-access system that allows the system capacity to be derived analytically.

Access monitoring it is the process of deriving useful information about accesses; it is a prerequisite for access selection. We present generic access abstractions that provide information about different accesses in a generic and comparable form. A *generic link layer* derives such abstracted information from access technology specific parameters. The generic abstractions provide information about the performance, as well as the resource situation for each access. The process of access monitoring depends on the connectivity state of a user terminal. Some types of information about an access system can be derived by making local measurements and observing access system information that is broadcasted in beacon signals; other types of information require that first a connection is established to the access system and the user terminal is attached to the network. We present different solutions on how suitable information about access systems can be obtained by the user terminal in different connectivity states. We develop a new network attachment mechanism that allows embedding access system information in the connectivity setup and network attachment procedure. We analyse this integrated network attachment procedure in a scenario where users evaluate

surrounding WLAN networks; we show that the amount of signalling overhead and the time spent to evaluate the WLAN network can be significantly reduced.

Once the available accesses are evaluated and a best suited access has been determined an *access handover* may be required, that redirects the service data flow of a session to the new access. We present several different solutions how an access handover can be performed. *Generic link layer* functions enable to establish a required communication context quickly in the new access system in order to avoid data distortion and service degradation. We evaluate and compare several access handover schemes based on alternative forms of context transfer for services based on the transmission control protocol (TCP). We show that context transfer schemes can provide a significant performance gain during access handover; this gain depends on the link layer characteristics of the old and new access system.

Acknowledgement

I thank my advisor Prof. Adam Wolisz for his guidance and Prof. Carmelita Görg for accepting to be evaluator of my dissertation.

This work was conducted while being at Ericsson Research. I want to express my gratitude to Fiona Williams, Norbert Niebert and Michael Meyer for their support.

Several people have contributed to this work by inspiring discussions and fruitful collaboration. I want to mention particularly my Ericsson colleagues Per Magnusson, Mikael Prytz, Johan Lundsjö, Teemu Rinta-aho, Göran Selander, Henning Wiemann, Michael Meyer Mathias Cramby, Anders Furuskär, Jonas Pettersson, Arne Simonsson, Oya Yilmaz, Gabor Fodor, Anh Tu Tran and Birinder Singh Khurana. Parts of this work were developed within the public-funded IPonAir and Ambient Networks projects. I am grateful for many valuable meetings and discussions – in particularly within the Ambient Networks work package on multi-access.

Special thanks go to my family and friends who have spared my company on many occasions when I has committed to this work. Above all, I thank my wife Katharina and my son Benno for their patience, love and support.

Contents

Chapter 1. Introduction

1.1 Heterogeneous Access Technologies and Cooperating Networks

Fixed and wireless communication has seen a tremendous growth over the last 15 years. The number of Internet users has grown from a few million in the mid 1990s to over 1.1 billion in June 2007 [ISTAT07]. A similar development has happened in mobile communications which began to take up in the early 1990s. The number of mobile users has exceeded the number of fixed telephony lines (see Figure 1.1); according to [ERI07] the number of world wide mobile phone subscriptions has reached 3 billion in summer 2007, ahead of earlier market estimates.

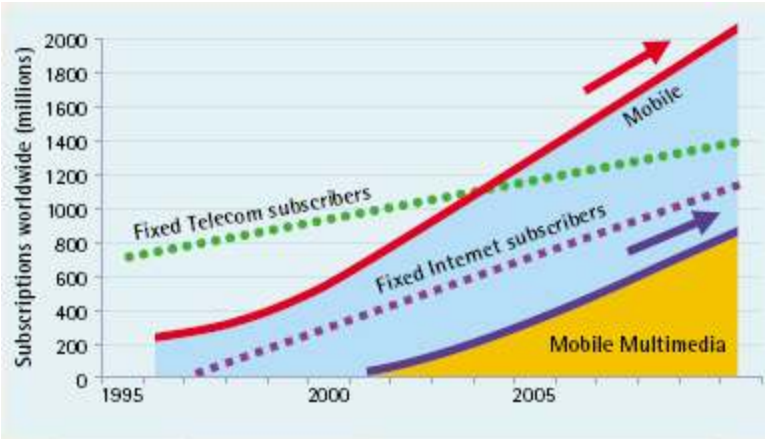


Figure 1.1: Users for fixed and mobile communications (source [UMTSF03]).

The dynamic *expansion* of the communications market is accompanied by *increasing diversity* of the technologies deployed in communication networks with a steady pace of new technological development. As a result, we have today an immense *heterogeneity* of communication technologies. Where in the past communication systems have been "vertically integrated" – one type of communication system was used for one specific service which was provided by a specific networking technology – we are today in a process of transformation called *convergence*. This process is depicted in Figure 1.2. Different networking technologies are integrated into a common communication platform and become independent of the type of service. New services are being developed and can be flexibly combined, where each service can comprise a number of different media elements. For example, photos can be shared as part of a telephony session. Services can use the ubiquitous and ambient communication system via a common service platform. This transformation process includes a convergence of different markets: the telecommunications market that originates from telephony services, the information technology market with computing and data services, and the media market with its origins in the broadcasting world. One recent example of a bundled service offering is referred to as "triple play"; it combines telephony, Internet and television services coming from the same provider. Data transfer is provided to the services by a ubiquitous communication platform. It has a feature set that is independent of a particular service, and that can comprise a variety of networking technologies. It is often referred to as *next-generation network*. In particular, the access provided to a user can be based on different

access technologies. These can be different fixed access technologies or different mobile and wireless access technologies[1]. The transformation process also includes device convergence. This means that communication devices support a variety of services and can support a multitude of access and networking technologies. For example, mobile devices can combine mobile phone functions with a mobile computing environment and also comprise a video and photo camera. Such a device can communicate via a number of fixed or wireless access technologies. Device convergence is not limited to integrated communication devices. The evolution of short-range communication technologies allows several complementary devices to interconnect into a common user network, which is often referred to as *personal area network* (PAN).

The convergence which takes place on a technological side leads at the same time to a disintegration of the market place. Business roles that are inherent to an integrated value chain need to be re-considered in a converged market. New business constellations can develop.

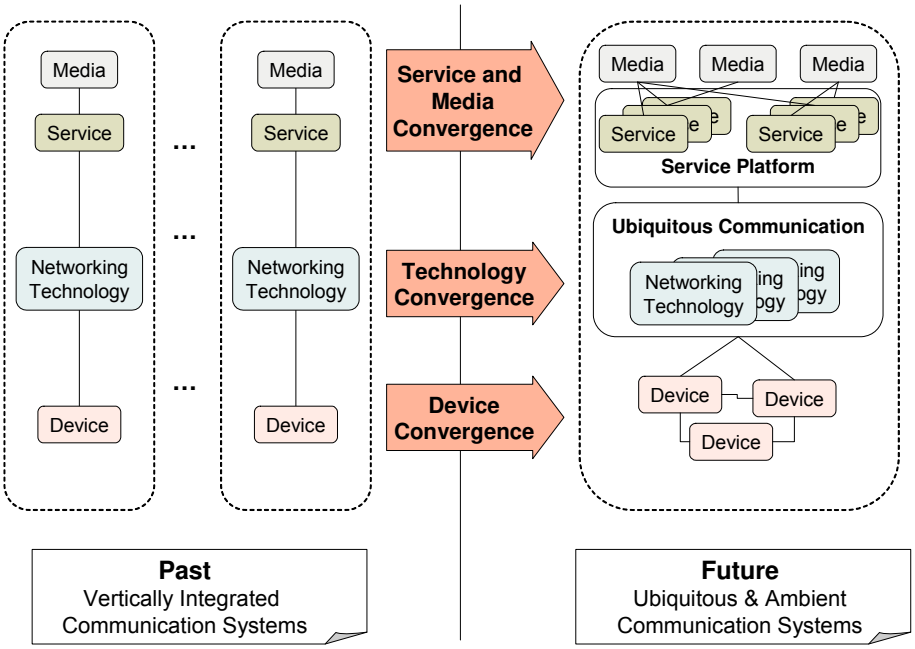


Figure 1.2: Convergence of integrated communication systems to ubiquitous and ambient communication systems.

We focus in this work on the access management functionality of the converged communication system. Although our concepts can be largely applied to fixed and wireless access systems, we limit ourselves to wireless access systems. Already a large number of radio access technologies (RATs) are deployed in today's networks. Wide-area cellular communication systems of the second generation (2G) are the *global system for mobile communications* (GSM), with the extension of *enhanced data rates for GSM evolution* (EDGE) and support for packet-switched communications via *(enhanced) general packet radio service* ((E)GPRS). A third generation (3G) mobile communication systems is the *universal mobile telecommunications system* (UMTS) based on *wide-band code division multiple access* (WCDMA) which operates in a different frequency band and provides higher data rates. Its evolution is *high-speed packet access* (HSPA) that provides higher spectral

[1] The combination of mobile and wireless access technologies is also known as *fixed-mobile convergence*.

efficiency and peak data rates. These standards have been specified within the *3rd generation partnership project* (3GPP). The next 3G evolution step *long-term evolution*[2] (LTE) is currently under development and it is based on *orthogonal frequency division multiple access* (OFDMA). All these RATs are operated in licensed frequency bands with wide-area coverage and good support for user mobility. In addition, wireless local-area access networks (WLANs) exist, which are operated in unlicensed spectrum and provide high peak data rates in small cell areas. They are often installed by private persons or organisations. The most prominent WLAN system is the IEEE 802.11 system, standardised by the *institute of electrical and electronics engineers* (IEEE) [WMB06]. 802.11 systems exist in different flavours 802.11a, 802.11b, 802.11g, of which a and b/g operate in different frequency bands. A new version 802.11n is currently being specified, which provides increased peak-data rates. Another set of RATs are wireless regional area networks also standardised by IEEE [IEEE802.16] [WMB06], typically denoted as WiMAX[3]. The IEEE 802.16 standard provides fixed wireless access without support for mobility; the version IEEE 802.16e also supports mobility. Standardisation for an evolved version IEEE 802.16m has started. A further standard that is under development is IEEE 802.20 [IEEE802.20] [GKTCW08], which targets support for high mobility of mobile devices. The heterogeneity of access technologies increases constantly, and mobile devices, like laptops or smartphones, integrate a growing number of access technologies. This development requires an increasing effort in managing the complexity of integrating these heterogeneous access technologies into converged multi-access systems. All these RATs differ in their characteristics with respect to 1) the frequency band they use, 2) whether the frequency band is licensed or un-licensed, 3) the support for user mobility, 4) the peak data rate, capacity and spectral efficiency, 5) the support for quality of service (QoS), and 6) the complexity and costs. There is no single radio access technology that is always best suited. Therefore, in the long-run different access technologies will co-exist. Depending on the scenario and its constraints the choice of access technology has to be evaluated. For network operators, a key driver motivating multi-access systems is the provisioning of transmission capacity within a targeted deployment area at the lowest deployment cost. The differing characteristics of access technologies, e.g. with respect to cell capacity, coverage range and service support, can be exploited in optimising the network topology in an integrated multi-access network. Radio access points of different access technologies can be located where it is most efficient for the desired deployment. Under the prerequisite that mobile devices can switch the connectivity between different access technologies, it is not required to provide full coverage with all access technologies. Load balancing between different access systems allows reducing the margin of spare capacity in each access system, and enables more cost efficient network deployment. The multi-access system extends the service coverage of one access technology to areas where other access technologies are available. Lowered deployment costs may be viewed as mainly an operator benefit, but on a competitive market it will ultimately benefit end users with reduced prices. In an environment where connectivity is provided by different business entities new cooperation models emerge. Different actors (business entities) target smaller network segments – at reduced financial costs – and achieve by cooperation a large footprint. For end users, the benefit of multi-access systems is not only reduced costs and a choice of selecting access providers but also an increased service experience due to the fact that multi-access provides a continuous service area out of heterogeneous access networks. The user can thus exploit the full access capabilities whenever they are available. End users are *always best*

[2] It is also called Evolved UMTS Radio Access (E-UTRA)
[3] Worldwide Interoperability for Microwave Access

connected [GJ03] with an *always best experience*. This scenery provides the framework for this work.

1.2 Problem Formulation

Several questions arise in converged wireless multi-radio access networks. In this work we address the following key research problems.

Problem 1: What is a suitable system architecture for a converged wireless network that builds on generic multi-access management functions?

For managing a combination of heterogeneous access technologies a common abstraction model is required that generically describes resources and communication objects. Common multi-access management functions operate on these abstracted objects to control the communication flow and thereby effect a technology specific realisation of the control. A multi-access system architecture needs to integrate these common multi-access management functions and the technology specific functions, and to provide suitable reference points to the technology specific realisation.

Problem 2: What gain can be achieved by jointly managing different access technologies in a converged wireless network?

We want to understand the feasibility of wireless convergence. What is the quantitative gain that access selection can bring to users and network operators? What management functions and algorithms are required; how shall users and their services be allocated to the different access systems? We also want to understand which system parameters and conditions influence the gain.

Problem 3: What system information is required to efficiently manage multiple access technologies?

For making any access selection decision, the multi-access management functions need to learn about the alternative options. It is necessary to understand what information is required, how and with what effort it can be retrieved. Because of the heterogeneity of access technologies a further problem arises: what are meaningful and generic metrics that describe different access technologies in a comparable way?

Problem 4: How can an ongoing communication session be efficiently transferred between different access technologies?

Flexible management of multiple access technologies comprises the functionality to re-allocate user sessions between different access systems. It is desirable that such an access handover provides seamless service continuity to the user. We want to understand the complexity to carry out efficient access handover and to which extent the performance of ongoing data sessions is influenced.

1.3 Research Contribution

In order to enable the integration of heterogeneous networking technologies into a common system architecture (cf. *research problem 1*) a suitable connectivity abstraction is required. We present a generic model for heterogeneous communication elements and define functional blocks for control and management of the communication elements. We present and compare different realisations of a converged multi-access system architecture and propose guidelines how to choose and combine different architecture realisations[4].

To understand what gain can be achieved by access selection (cf. *research problem 2*) we have developed a business model comprising different business entities. By investigating the interests of involved actors, we define utility models that allow formulating *access selection* as an optimisation problem of a utility function[5]. In our further evaluation we restrict the focus on scenarios where all business actors cooperate. We classify access selection algorithms according to their input parameters and dynamics[6] and develop a taxonomy to describe different types of access selection gain. We derive by which system parameters the gain is mostly affected[7]. We provide quantitative capacity gains of access selection for a wide range of scenarios and derive conditions when access selection is feasible and how heterogeneous access networks should be deployed[8]. In order to facilitate analytical capacity investigations of access selection gains, we present a new analytical model that overcomes shortcomings of the typically used single server model[9].

Before any sophisticated form of access selection can be performed information about the capabilities of the access network as well as the access connection are required (cf. *research problem 3*). We identify what information describes the suitability of a network and access from a user and operator perspective. We compare different options on how this information can be distributed and provided to user networks. The overhead and performance of those procedures is evaluated and compared in a WLAN scenario for which we have derived appropriate evaluation models[10]. We derive generic access abstractions to overcome the difficulty of deriving comparable measures for different access technologies[11].

When access selection has determined the use of a new access for a data session an access handover of the old communication path via the old access to a new communication path via the new access has to be performed. We investigate for such an access handover its implication on the service performance (cf. *research problem 4*). For the investigation of seamless and efficient access handover schemes we examine context transfer schemes; in a first step we analyse the communication context that exists within the network infrastructure and the mobile user network[12]. We present and compare all possible alternative methods to achieve continuous connectivity and seamless handover during an access handover[13]. In

---

[4] Published in [BCJLM+04] [BMBF04] [MLSW04] [NSAMS+04] [SMACK+04] [KKLBB+05] [SABGJ+06] [KADGP+07] [KADGP+07] [SM07] [SPG07].
[5] Published in [SPG07] [SM07] [TPSPS+07] [GABBE+07].
[6] Published in [JSRJ06] [SPG07].
[7] Published in [KAABB+05] [SM07].
[8] Published in [SPG07].
[9] Published in [Sac06a] [Sac06b].
[10] Published in [RAQS07] [RCMMS+07] [QRS07].
[11] Published in [MGSCA07] [SADGK+07] [SADGK+08].
[12] Published in [SWLM04] [DABKK+05] [SKM06].
[13] Published in [SWMWL04] [DABKK+05] [KAACD+05] [SKLMR+06].

contrast to other related work we investigate in particular the influence of separate processing and queuing of data at the network layer and the link layer. We derive conditions in which cases access handover is seamless and when it causes data distortion that may affect the service performance. We analyse in depth the interactions that access handover can cause for services based on the *transmission control protocol* (TCP)[14]. We present a generic link layer which provides flexible and configurable link layer functionality based on a toolbox of functions that can adapt to a variety of transmission modes and access technologies[15]. The generic link layer can, on one hand, overcome known limitations of statically configured link layers when a wide range of data rates is provided by the physical layer. On the other hand, it is suitable for dynamic software-defined and software reconfigurable communication systems.

This work has been partly performed within the collaborative projects IPonAir [IPONAIR] and Ambient Networks [AN] [NSAMS+04] [NSZH07]. Significant contributions have been provided to the reports and concepts developed in these projects[16]. This work has resulted in a significant number of publications and also a number of filed patent applications.

1.4 Outline

Chapter 2 provides a background about wireless communication systems and prior work on the integration of heterogeneous access systems. In Chapter 3 we present the system model and terminology for a multi-access system that integrates different access technologies. The main multi-access management functions and their relationship are explained. We present and discuss different options on how a multi-radio access system architecture can be realised. In Chapter 4 we describe access selection, its objectives and different types of algorithms. We present the capacity of a multi-access system for different access selection algorithms and system parameters. Chapter 5 discusses how information about the capabilities of access networks can be obtained to enable effective access selection. The performance and overhead is investigated for a WLAN scenario. In Chapter 6 we investigate the access handover that follows after an access selection decision. We discuss the requirements for an access handover and present different realisation options. We investigate the performance that access handover has on ongoing data services. The work is summarised in Chapter 7 and an outlook on future work is presented. Supplementary material to some chapters is provided in the appendices.

[14] Published in [SKM06].
[15] Published in [SS02] [WWRFa02] [WWRFb02] [Sac03a] [Sac03b] [SWMWL04] [KAACD+05].
[16] The reports are listed in the bibliography. Reports with contributions of the author are indicated.

Chapter 2. Background on Wireless Communication Systems

2.1 Types of Wireless Communication Systems

The technological fundamentals for wireless communication have been laid between the end of the 19th century until the middle of the 20th century (see e.g. [Wal02]). This development has lead to the deployment of large-scale wireless communication systems starting around the 1970s. Wireless communication systems can be distinguished into satellite-based and terrestrial wireless communication systems. In satellite-based systems the communication path between communication parties is routed via satellites; in terrestrial wireless communication systems the network infrastructure is located on the ground. Terrestrial wireless communication systems can be categorised according to their objectives as shown in Figure 2.1.

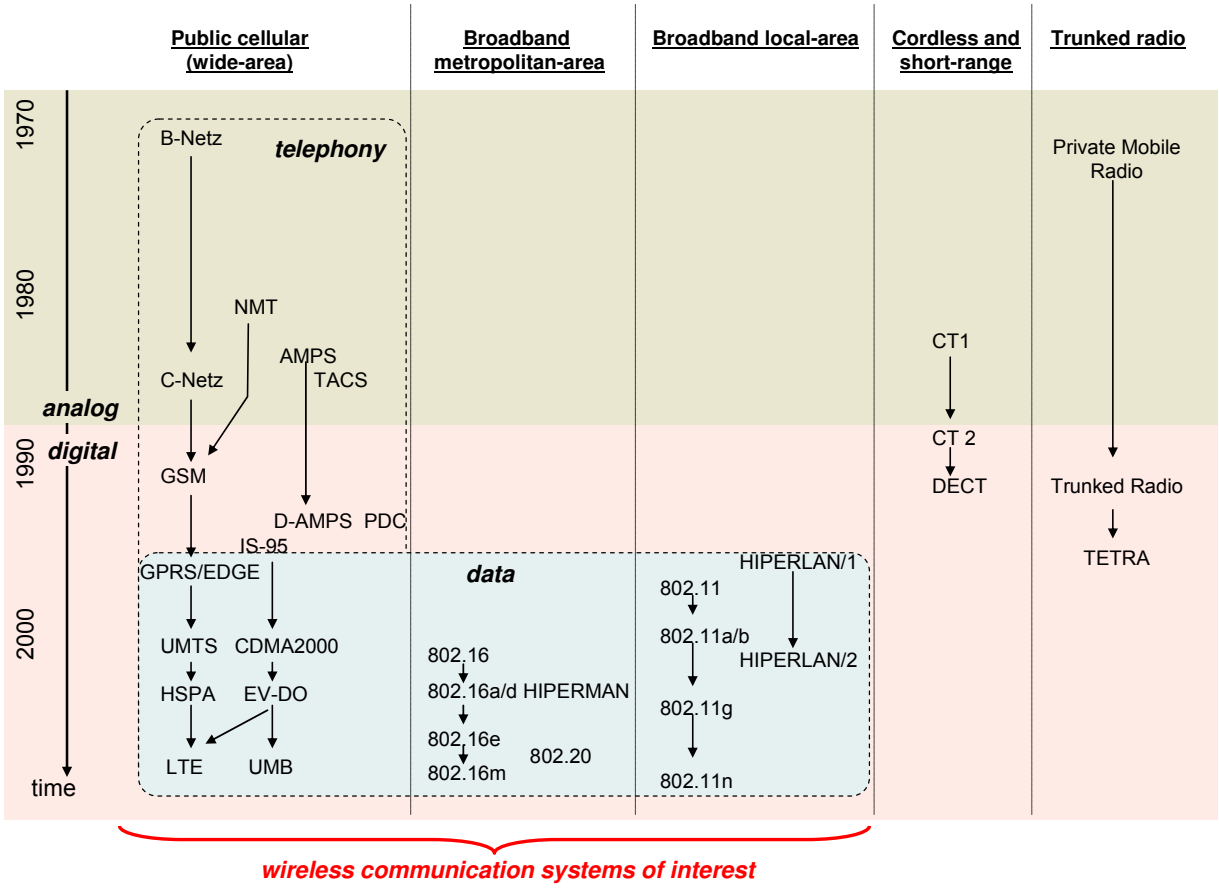


Figure 2.1: Categories of terrestrial wireless communication systems.

Public wide-area cellular networks have been developed to provide mobile telephony services to users at any location and while they are moving. The telephony service was based on a circuit-switched call setup and routing. Later support for data services has been added. A shift in technology came with the transition from analogue to digital cellular networks in the early 1990s. A further significant technology step was the addition of a packet-switched

infrastructure in 1997 which supports better various types of data services, for example Internet services. With the advancement of packet-switched transmission towards higher efficiency, higher reliability and better support to guarantee service quality, a migration of telephony services from circuit-switched to packet-switched transmission has started; eventually this will lead to the disappearance of the circuit-switched transmission infrastructure. The capabilities of the radio access technologies in cellular networks have changed tremendously. In the early cellular networks the radio access technologies supported low data rate telephony and data services for mobile users. Today's cellular radio access technologies can support services with data rates up to approximately 16 Mb/s and up to 100 Mb/s and beyond in the coming years [Wal02] [DPSB07] [HT06].

Broadband wireless local area networks (WLANs) have been developed to provide high data rate connectivity to computers, laptops and workstations as replacement to wired local area network connectivity. WLAN standards have been standardised in parallel in ETSI and in IEEE since the 1990s. In ETSI standardisation has resulted in the standards HIPERLAN/1 and HIPERLAN/2 [Wal02]; in IEEE a number of 802.11 standards have been developed [Wal02] [IEEE802.11] [IEEE802.11a] [IEEE802.11b1] [IEEE802.11g]. WLAN provides high data rates ranging from few Mb/s in early standards to several 100s of Mb/s in currently ongoing standard enhancements. WLAN radio cells have only a limited coverage range and provide limited support for mobility and quality of service. Enhancements to improve these limitations are currently being standardised. In the market 802.11 WLANs have gained a dominant position over HIPERLAN due to their lower complexity. However, several concepts of the more sophisticated HIPERLAN systems have been introduced into enhancements of the 802.11 standard.

Broadband wireless metropolitan (or regional) area networks (WMAN) have been developed to provide data connectivity to households without a necessity to install cables into every building. They have a larger coverage than WLAN systems. WMAN systems have been standardised by ETSI as HIPERMAN and by IEEE as IEEE 802.16 (also called WiMAX) [WMB06]. Originally they have been developed for stationary end devices without power limitations due to battery operation. Mobility support for handheld devices has been added in the IEEE 802.16e (Mobile WiMAX). A wireless system for highly mobile users is currently standardised in IEEE 802.20 [GKTCW08].

Other types of wireless communication systems are trunked radio systems, cordless telephony systems, and short-range communication systems; these systems are beyond the scope of this work. Trunked radio systems are communication systems limited to closed user groups, such as public safety personnel. Cordless telephony systems extend fixed telephony systems to wireless handsets. An overview of trunked radio systems and cordless telephony systems is provided in [Wal02]. Short-range communication systems provide connectivity in close proximity, and are used for wireless inter-connection of multiple devices within a personal or body area network. They provide flexibility for communication end systems by allowing to distribute end-system functionality onto different components; for example, a laptop, a mobile phone and a separate radio modem can be wirelessly inter-connected and jointly form an end system. Short-range wireless communication systems are standardised in IEEE 802.15 [WMB06].

2.2 Wide-Area Cellular Mobile Networks

2.2.1 Network Architecture

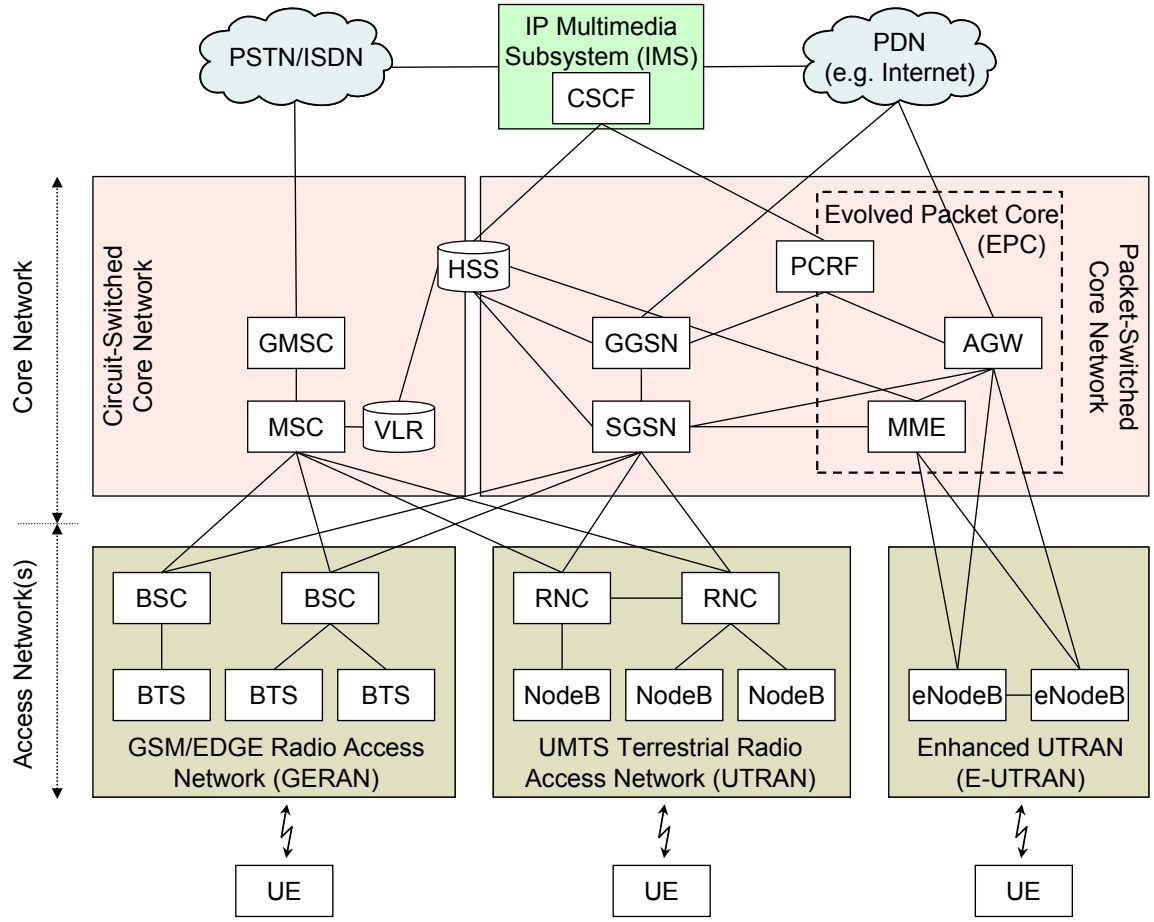


Figure 2.2: Network architecture of a 3GPP core network and GERAN, UTRAN, E-UTRAN radio access networks.

Cellular mobile networks provide telephony and data services to mobile users with wide-area service coverage. The network architecture comprises a core network and one or more access networks. 3GPP defines three different types of radio access networks that can provide connectivity between the user equipment (UE) and the core network: the GPRS/EDGE radio access network (GERAN), the UMTS terrestrial radio access network (UTRAN) and a new Enhanced UTRAN (E-UTRAN). A simplified network architecture for 3GPP networks is depicted in Figure 2.2; a more detailed description can be found in [3GPP23.060] [3GPP23.401] [3GPP23.402] [3GPP23.203]. The core network can be furthermore divided into a circuit-switched domain and a packet-switched domain. The circuit-switched domain provides PSTN/ISDN services like telephony to the end user. The switching functionality is provided by the (gateway) mobile switching centre ((G)MSC) and a visited location register (VLR) that stores information about roaming users. A Home Subscriber Server (HSS) contains the user profile and subscription information. For further information see e.g. [Wal02]. With the introduction of GPRS a packet switched core network domain was added, which provides data services to private or public packet data networks (PDNs) via a Gateway GPRS Support Node (GGSN). A Serving GPRS Support Node (SGSN) is responsible for e.g.

mobility management, authentication and authorisation of users connected via the GERAN. A policy control and charging resource function (PCRF) provides charging and policy control for different data services. A detailed overview of packet switched core network functions is given in [BW97] [BVE99] [3GPP23.060]. Currently the new LTE radio interface and E-UTRAN are developed for which new evolved packet core (EPC) functions are introduced into the packet switched core network, which take over the role of SGSN and GGSN. The first is an access gateway (AGW) which acts as mobility anchor and a mobility management entity (MME) for E-UTRAN users. The 3GPP system architecture evolution is currently being standardised; further information is provided in [3GPP23882] [3GPP23401] [3GPP23402] [DPSB07] [SO08]. The packet switched core network is connected to the IP Multimedia Subsystem (IMS), which provides service control functions for packet-switched services, like IP-based multimedia telephony, group communication or messaging services (see e.g. [CG04]). IMS services can be initiated via the end user or peer service entities in other networks; IMS provides also gateway functionality to bridge circuit-switched telephony services (e.g. ISDN/PSTN telephony) with IP-based telephony. Service usage is performed according to the user profile stored in the HSS and the transmission functionality and charging is enforced via the PCRF (see [3GPP23.203]).

2.2.2 Protocol Stack & Transmission

The packet-switched core network provides transmission for data applications running over the Internet Protocol (IP) as depicted in Figure 2.3. The transmission of user data through the core network is based on the GPRS Tunnelling Protocol (GTP) running through the IP-based core network over the User Datagram Protocol (UDP). The GTP tunnel serves two purposes. Firstly, it provides a mobility tunnel between the core network edge (i.e. GGSN or AGW) to the radio access network; if the point of attachment of the user to the core network changes, e.g. due to user mobility, the GTP tunnel is redirected to the new RAN point of attachment [3GPP23.060]. Secondly, GTP allows to provide separate transmission tunnels for different types of user data; this enables to provide quality of service independently for different application types as described in [3GPP23.107] [3GPP23.207] [3GPP23.203] [LEWL06].

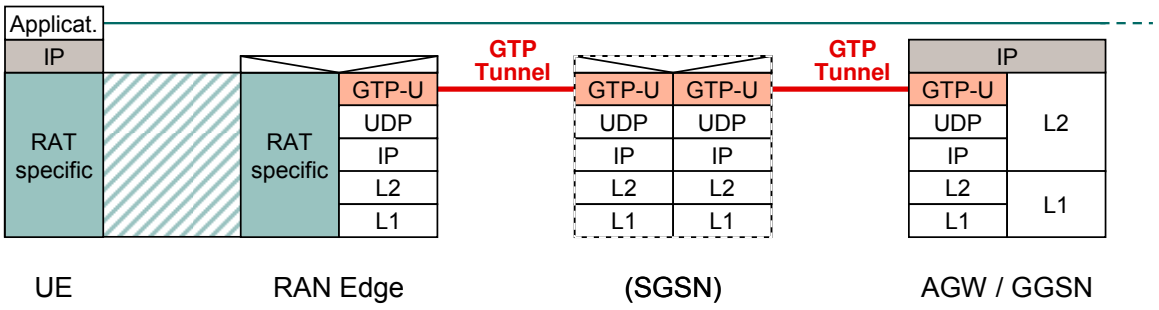


Figure 2.3: User plane transmission through the packet-switched core network.

The GTP tunnel from the core network stretches up to the edge of the radio access network, where each GTP tunnel is mapped to a radio access bearer configured to the same QoS class as the corresponding GTP tunnel. The transmission in the core network depends on the type of radio access network. GERAN contains two types of nodes: the base station controller (BSC) and the base transceiver station (BTS) at the antenna as shown in Figure 2.2. The RAN transmission in GERAN is shown in Figure 2.4 and Figure 2.5. For GERAN two alternative transmission modes exist. In the original so-called *Gb mode* (Figure 2.4) the GERAN radio

protocols are partly located in the SGSN in the core network. The physical transmission over the radio interface is performed with frequency-division duplex on carriers with 200 kHz channel bandwidth, which are in Europe typically in the 900 and 1800 MHz frequency band. On every channel transmission is performed according to time division multiple access (TDMA) with a TDMA frame being divided into 8 time slots of 4.615 ms, which can be used for either circuits-switched telephony or packet switched services. With EDGE higher order modulation, link adaptation and incrementally redundant channel coding is introduced, which allows to achieve higher data rates at good radio channel conditions. The Medium Access Control (MAC) protocol schedules different users and radio access bearers to the channel resources. The Radio Link Control (RLC) protocol performs segmentation of higher layer datagrams into packets suitable for radio transmission and performs error recovery by automatic repeat request (ARQ). Those protocols are terminated in the base station subsystem (BSS) which contains the base transceiver station (BTS) and the base station controller (BSC). The BSS is connected to the SGSN via the Base Station Subsystem GPRS Protocol (BSSGP) that performs primarily flow control between data buffers in the BSS and SGSN. The logical link control (LLC) protocol also performs ARQ to recover from losses that occur during handovers between different BSSs. The Subnetwork Dependent Convergence Protocol (SNDCP) performs header and data compression for different IP data flows. More information on GERAN can be found in [3GPP23.060] [3GPP43.064] [BW97] [BVE99] [FMMO99] [FNO99].

A second realisation of the transmission in GERAN – denoted as *Iu mode* – has been introduced in 3GPP release 5 (see Figure 2.5). In GERAN Iu mode all radio specific functionality is moved from the SGSN to the BSS in order to have a cleaner separation of core network and radio access network functionality. The treatment of mobility between different BSSs is handled by GTP procedures; IP header compression is provided by the Packet Data Convergence Protocol (PDCP). The functional architecture of GERAN in Iu mode is based on retro-fitting the functional design of the UTRAN, as depicted in Figure 2.6, onto the older GERAN. A detailed description can be found in [3GPP43.051] [MST01].

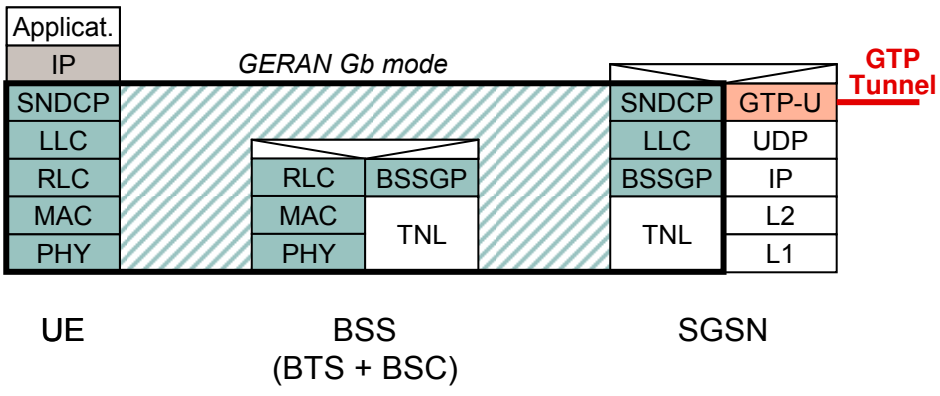


Figure 2.4: GERAN user plane transmission in Gb mode.

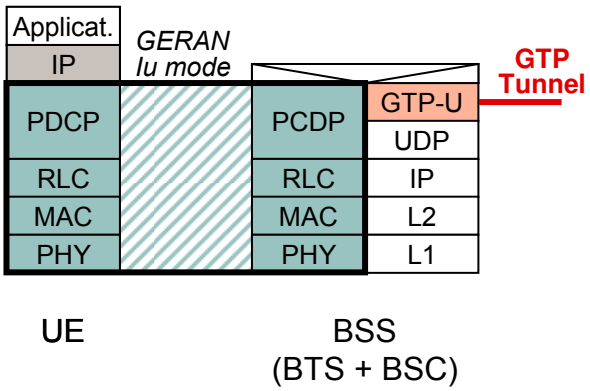


Figure 2.5: GERAN user plane transmission in Iu mode.

UTRAN consists of radio network controller (RNC) and base station (denoted as NodeB) nodes as shown in Figure 2.2. The physical transmission is performed with frequency division duplex on 5 MHz carriers located in the 2 GHz frequency band within 10 ms radio frames. Channelization is performed by Wideband Code Division Multiple Access (WCDMA) with a chip rate of 3.84 Mcps; variable physical layer transmission rates can be achieved by orthogonal channelization codes with variable spreading factors. UTRAN uses a tight frequency reuse of one, with all neighbouring cells using the same carrier. The CDMA properties allow applying soft-handover by sending the same information via multiple radio cells or NodeBs. The multiple signals transmitted via the different radio cells can then be combined, which provides a macro-diversity gain that improves link performance at the cell edges. On the other hand it requires an interface between RNCs so that signals that are transmitted via different NodeBs controlled by different RNCs can be combined. The transmission via UTRAN is depicted in Figure 2.6 for UMTS dedicated channels (Figure 2.6 (a)) and high-speed packet access (HSPA) (Figure 2.6 (b)). IP datagrams of each GTP tunnel are mapped to a separate UMTS radio bearer. The packet data convergence protocol performs IP header compression. In the RLC protocol PDCP datagrams are concatenated and segmented into RLC PDUs according to the radio block size. RLC performs error recovery by ARQ. RLC PDUs of different radio bearers are scheduled by MAC and are transported via a transport network layer (TNL) between the RNC and the NodeB. Physical radio transmission takes place between the NodeB and the UE; the physical layer part located in the RNC is responsible for macro-diversity combining.

The performance of UTRAN has been improved by introducing the High-Speed Packet Access (HSPA), which consists of High-Speed Downlink Packet Access (HSDPA) and an enhanced uplink, also known as High-Speed Uplink Packet Access (HSUPA). HSPA provides higher peak data rates, lower delays and higher capacity compared to UMTS dedicated channels. For HSPA new functionality is primarily added to the NodeB and the UE (see Figure 2.6 (b)). The improvements of HSDPA are achieved by reducing the transmission time intervals of radio frames from 10 ms to 2 ms and adding fast channel-dependent scheduling in the NodeB. All HSDPA users are sharing a common transport channel that is shared by multiple users in a time-multiplexed fashion. Furthermore, higher order modulation is used to allow higher peak data rates. Finally, hybrid ARQ with soft combining is applied between the UE and the NodeB in order to reduce the block error probability. For HSUPA also shorter radio frames of 2 ms and hybrid ARQ with soft combining are used. Furthermore, a new uplink scheduling framework is applied in order to limit interference between different UEs.

More information on UTRAN can be found in [3GPP25.301] [3GPP25.401] [HT00] and on HSPA in [3GPP25.308] [3GPP25.309] [HT06] [DPSB06].

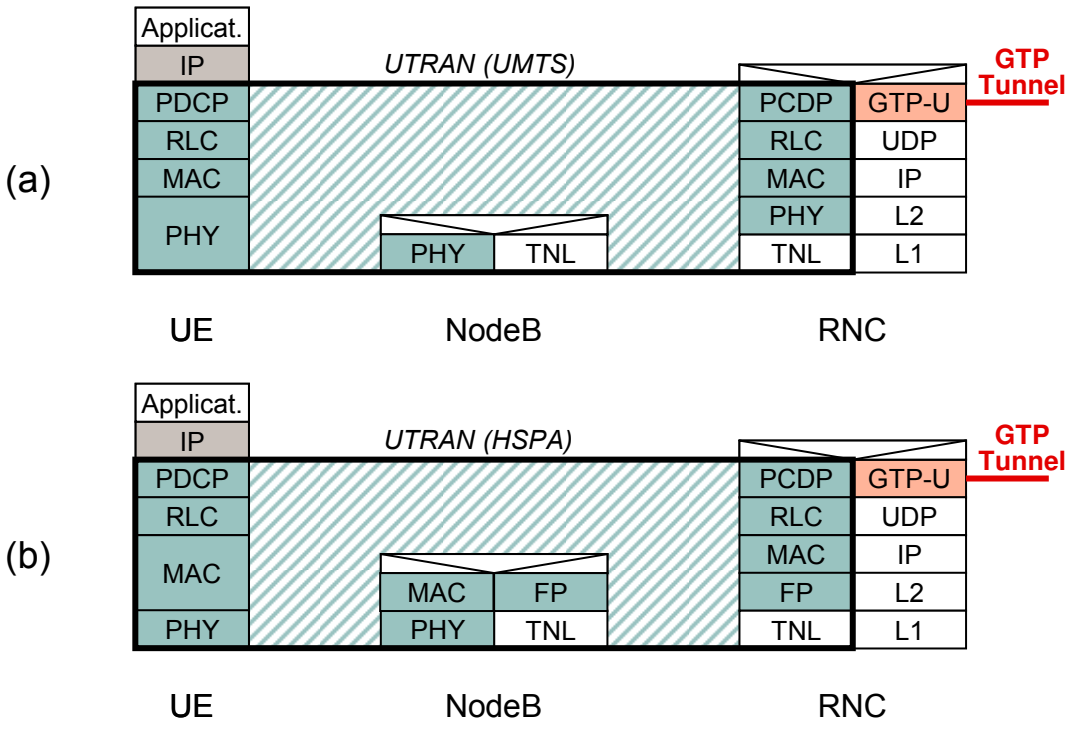


Figure 2.6: UTRAN user plane transmission via (a) UMTS dedicated channels and via (b) HSPA.

3GPP is currently specifying the E-UTRAN based on the LTE radio interface. E-UTRAN simplifies the radio network architecture by placing all RAN specific functionality into a single radio node, the evolved NodeB (eNodeB), as depicted in Figure 2.2. E-UTRAN introduces a new radio interface based on orthogonal frequency division multiple multiplexing (OFDM) in the downlink and single-carrier frequency division multiplexing (SC-FDMA) in the uplink. It supports both frequency division and time division duplexing and can operate on channel bandwidths from 1 MHz to up to 100 MHz. E-UTRAN also uses channel-dependent scheduling, rate adaptation and hybrid ARQ. It introduces inter-cell interference coordination and smart antenna concepts. The protocol stack of E-UTRAN (see Figure 2.7) is in principle similar to UTRAN; however radio protocol parameters are adapted to the new physical layer. More information on E-UTRAN can be found in [3GPP36.300] [DPSB06].

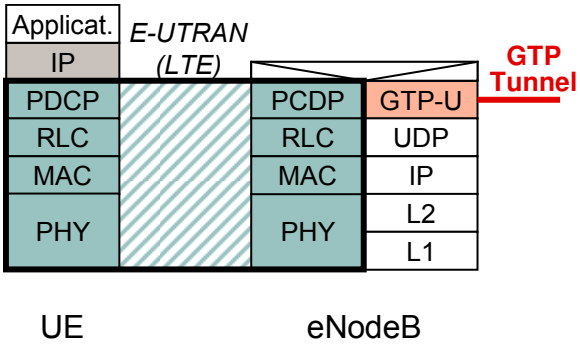


Figure 2.7: E-UTRAN user plane transmission.

Other wide-area cellular mobile networks exist, for example those standardised in 3GPP2 (IS-95, CDMA2000, UMB). Those networks do not differ fundamentally from 3GPP networks. Also WiMAX networks have a comparable architecture and transmission as 3GPP networks [Sch06]. Significant work on novel radio access concepts has been performed within the WINNER project [WPBS04] [Moh05] [PHDSP+06]; this work has partially influenced the standardisation of LTE.

2.3 Wireless Local-Area Networks

Wireless local area networks have been standardised by the *broadband radio access network* working group of ETSI (i.e. HIPERLAN), as well as in IEEE (i.e. 802.11). However, the IEEE 802.11 standard dominates the market of WLAN products. Figure 2.8 depicts the WLAN network architecture according to IEEE 802.11 [IEEE802.11] [WMB06]. Different WLAN stations (STA) using the same coordination function form a *basic service set* (BSS). In infrastructure-mode[17] one WLAN station acts as *access point* (AP), which connects the BSS to a distribution system. Different BSSs that are connected via a distribution system can form an *extended service set* (ESS). The BSS corresponds to a WLAN radio cell and is identified by the BSSID, which corresponds to the MAC address of the access point. The ESS is identified by and ESSID, which is a 32 character network name that corresponds to the common SSID configured in each AP of the ESS. A *gateway* (GW), also called portal, can provide connectivity of the ESS to external packet data networks.

[17] IEEE 802.11 also supports an ad-hoc mode, which is not further considered in this work. For more information see [WMB06] [IEEE802.11].

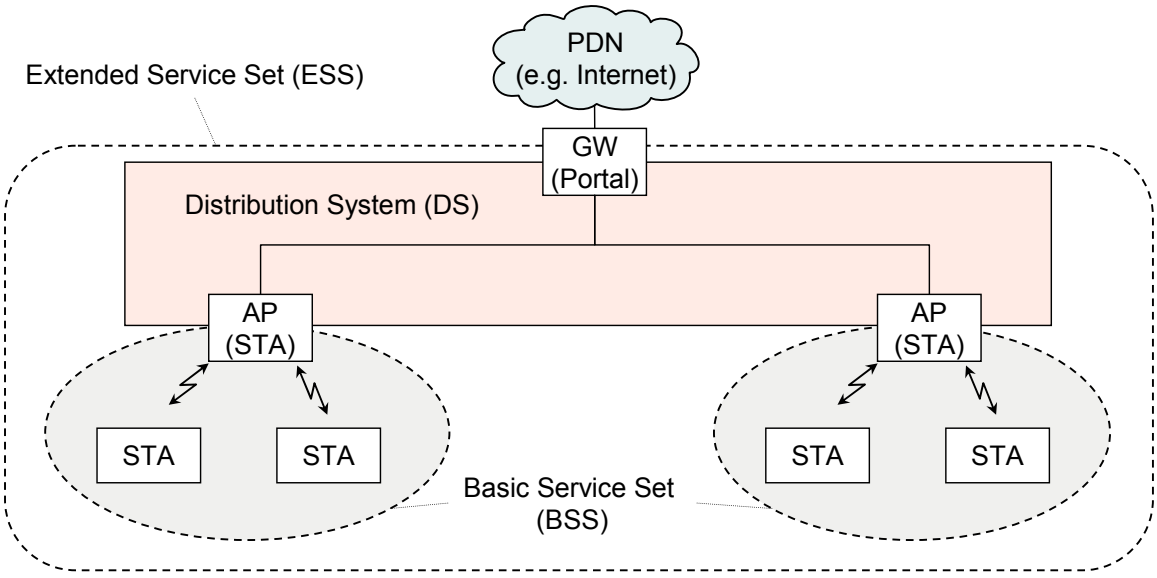


Figure 2.8: Network architecture of WLAN network.

The data transmission in a WLAN network is depicted in Figure 2.9. The application data is transmitted via end-to-end IP connectivity. The WLAN standard specifies MAC and physical layer procedures for data transmission between the station and the access point. For the physical layer different options exist; most relevant are versions 802.11a [IEEE802.11a], 802.11b [IEEE802.11b1] and 802.11g [IEEE802.11g]. All these version transmit on carrier bandwidth of 22 MHz; 802.11a uses the 5 GHz band and 802.11b and 802.11g uses the 2.4 GHz band. 802.11b uses direct sequence spread spectrum, whereas 802.11a and 802.11g use orthogonal frequency division multiplexing. The access to the transmission medium is provided to the access point and different mobile stations by the MAC layer coordination function. Two types of coordination functions are defined. The *point coordination function* (PCF) provides contention free access by centrally (at the access point) coordinated polling the different stations for transmission of data frames. The contention free period of PCF operation alternates with contention based access according to the *distributed coordination function* (DCF). With DCF all stations try to access the wireless medium independently according to a *carrier-sense multiple access with collision avoidance* (CSMA/CA) scheme. As PCF is today not implemented in practical WLAN systems DCF constitutes the dominant WLAN coordination function. A new *hybrid coordination function* has been added in [IEEE802.11e] which is not further considered in this work.

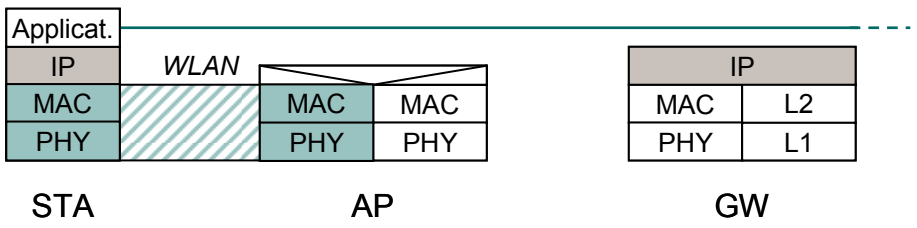


Figure 2.9: WLAN transmission.

Practical realisations of WLAN networks can often deviate from the reference architecture provided in the IEEE 802.11 standard. One reason is that in a larger WLAN network it is desirable to configure, manage and use multiple access points in a coordinated fashion

[RFC3990]. A standardised approach in this direction, in contrast to today's proprietary solutions, is pursued in the *control and provisioning of wireless access points* (CAPWAP) workgroup of the Internet Engineering Task Force. The functionality of the 802.11 MAC is thereby distributed between and an *access controller* (AC) and a *wireless termination point* (WTP), where the AC manages a collection of access points. The CAPWAP architecture is depicted in Figure 2.10.

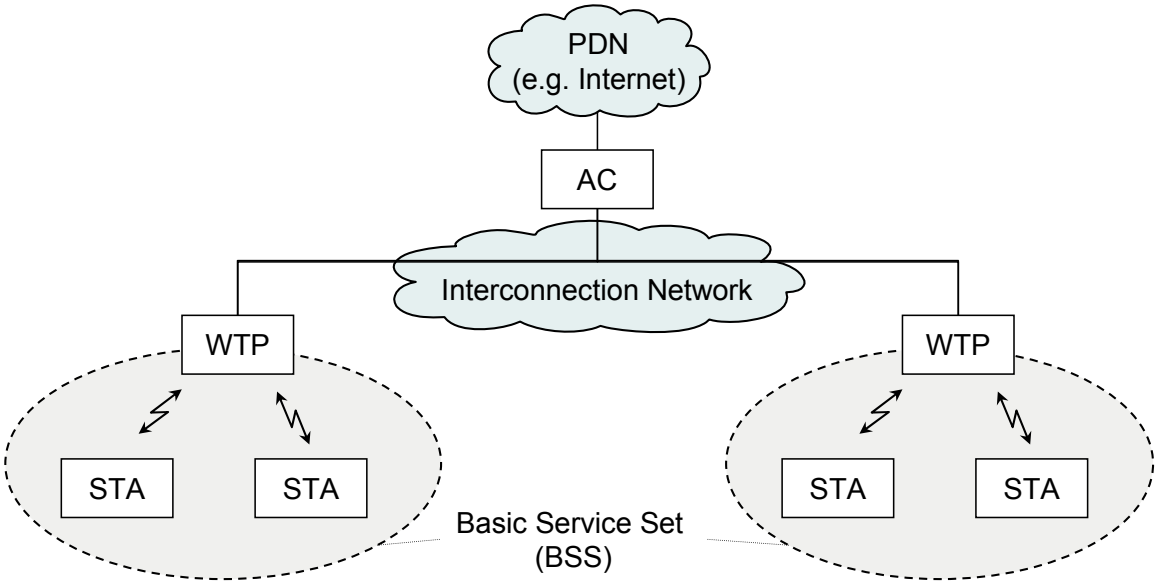


Figure 2.10: WLAN network according to the CAPWAP architecture with WLAN functionality distributed between an access controller (AC) and a collection of wireless termination points (WTP).

Figure 2.11 shows the data transmission according to the CAPWAP architecture. The 802.11 MAC functionality is split between the WTP and the AC, which are connected via the CAPWAP protocol [ID-CAPa] [ID-CAPb] that runs over an IP connection. Different MAC distribution options exist. In case of CAPWAP with *split MAC* a significant part of MAC functionality is located in the AC and the WTP mainly performs tasks such as beacon generation and power management. In contrast, for CAPWAP with *local MAC* most MAC functionality is located in the WTP and the AC performs only limited functionality like security key management.

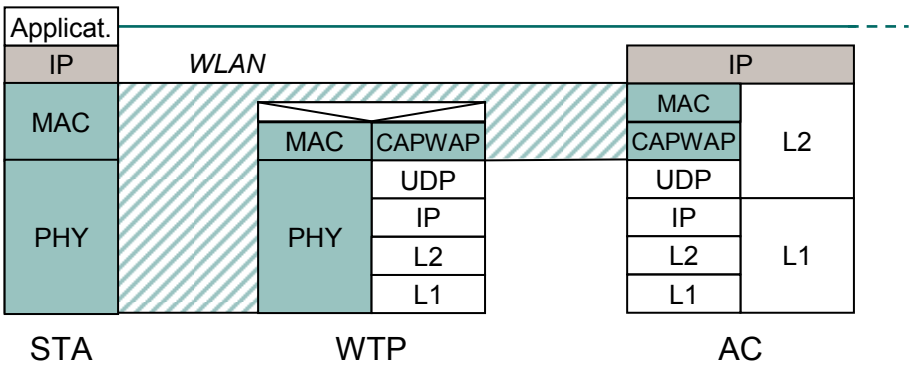


Figure 2.11: WLAN transmission according to the CAPWAP architecture.

2.4 Integration of Different Access Technologies

The multitude of access technologies has sparked investigations on integration and interworking of different access technologies. A first need in this direction became evident with the introduction of UTRAN access networks, for which a large reuse of already deployed network functions was intended [HJJ02]. This resulted in the architecture depicted in Figure 2.2 where the UTRAN and GERAN networks share the same core network functions. As UTRAN did initially only provide coverage in limited areas, while existing GERAN access networks provided almost full coverage, functionality was required to allow efficient handover between GERAN and UTRAN networks at the UTRAN coverage edges [FHTW04]. Three mechanisms have been introduced to allow efficient handover between the GERAN and UTRAN, as indicated in Figure 2.12. The core network provides mobility management functions for inter-system handover. The different RANs support the user equipment with neighbour cell information of other RANs and provide measurement opportunities during ongoing data transmission. Finally, the different RANs provide interworking functions for handover preparation and optimisation by forwarding data between RANs to avoid packet losses [ABHMM+03]. The same principles are currently also introduced for E-UTRAN [SO08].

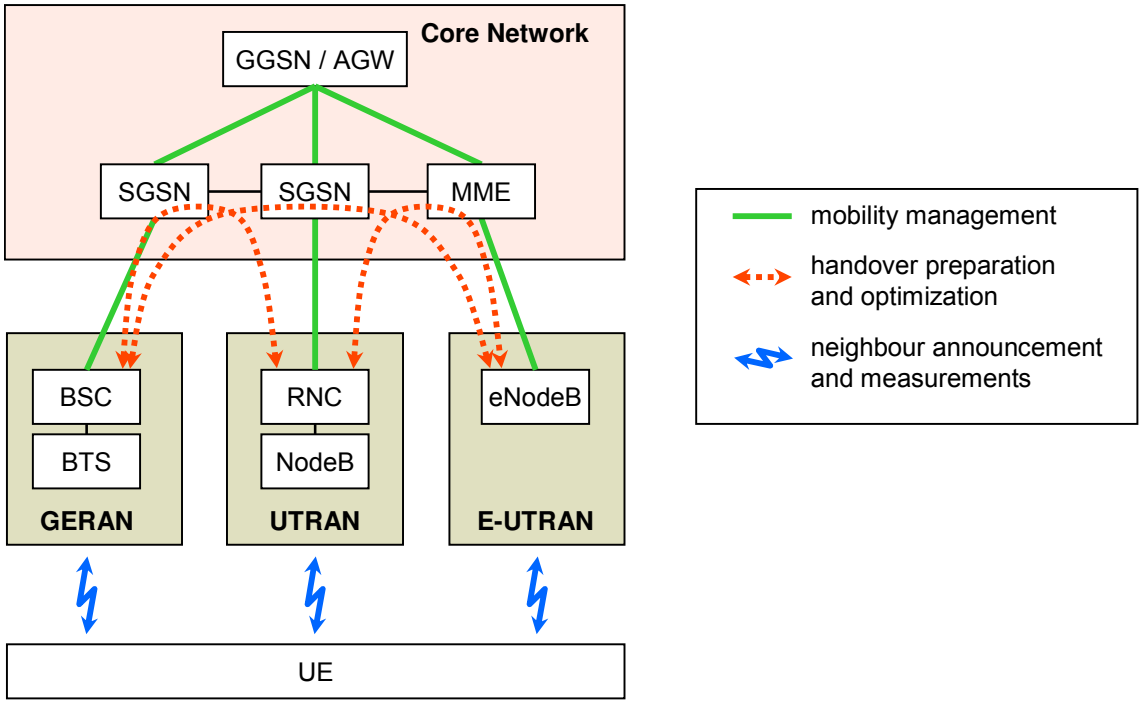


Figure 2.12: Interworking between different cellular access networks.

Interworking of wide area cellular networks and wireless local area networks is considered as a beneficial type of system integration; while cellular networks provide good coverage and service availability WLANs provide locally high capacity and possibly at low access costs. Two options of loose and tight coupling (or interworking) for the interconnection of HIPERLAN/2 and UMTS have been investigated in [ETSI957], and similarly for CDMA2000 and WLAN in [BCHLM+03]; [SFA03] has extended the interworking options for WLAN and UMTS resulting in the following solutions:

- *Open coupling*: both the cellular and the WLAN network are independent and provide separately access to the Internet. No optimised coordination is possible of which access to use for a mobile terminal.

- *Loose coupling*: the cellular and the WLAN network use a common subscriber database and coordinate the authentication procedure. Optimised handover is not possible.

- *Tight coupling*: the WLAN network is connected to the cellular core network (the SGSN) via an interworking unit - similar to how UTRAN is connected to the core network. This option enables optimised handover between WLAN and UMTS.

- *Integration*: the WLAN network is connected to the radio network controller of UTRAN and appears as a special type of radio cell. This option also enables optimised handover.

For the interworking with 3GPP and WLAN networks, 3GPP has defined six different scenarios with increasingly strict performance requirements, as listed in Table 2-1. Scenario 1 only enables common billing and customer care for a user of a WLAN and 3GPP network; the networks operate independently. In scenario 2 the authentication, authorisation and accounting mechanisms of the 3GPP network are used to provide access control to WLAN networks; otherwise the networks remain independent and provide their own set of services. Scenario 3 furthermore enables a terminal to use packet-switched services of the 3GPP network when connected via WLAN. Scenario 4 enhances scenario 3 by providing service continuity when the terminal changes between the 3GPP and the WLAN access; in addition scenario 5 adds that service continuity is seamless via optimised handover procedures. Scenario 6 enables a user to access circuit-switched services (such as telephony) while connected via the WLAN network.

Table 2-1 : 3GPP-WLAN interworking scenarios as defined by [3GPP22.934].

Scenarios: Service and operational Capabilities:	Scenario 1: Common Billing and Customer Care	Scenario 2: 3GPP system based Access Control and Charging	Scenario 3: Access to 3GPP system PS based services	Scenario 4: Service continuity	Scenario 5: Seamless services	Scenario 6: Access to 3GPP system CS based Services
Common billing	X	X	X	X	X	X
Common customer care	X	X	X	X	X	X
3GPP system based Access Control		X	X	X	X	X
3GPP system based Access Charging		X	X	X	X	X
Access to 3GPP system PS based services from WLAN			X	X	X	X
Service Continuity				X	X	X
Seamless Service Continuity					X	X
Access to 3GPP system CS based Services with seamless mobility						X

3GPP has specified mechanisms for interworking between WLAN and 3GPP networks. One mechanism is denoted as *interworking WLAN* (IWLAN), which allows terminals to connect to the packet switched core network via an encrypted IPsec tunnel [AHP03] [BGK02] [KH03] [3GPP22.234] [3GPP23.234] [3GPP22.234] [RFC4301]. The first version of IWLAN in 3GPP release 6 supports scenarios 1-3; scenarios 4-5 are currently being added. Another mechanism of WLAN interworking named *unlicensed mobile access* (UMA) was developed by a group of network operators and equipment vendors and later adopted by 3GPP under the term *generic access network* (GAN) [3GPP43.318] [BHNVO05]. It provides a new node, the generic access network controller (GANC), which acts as GERAN base station towards the 3GPP core network and provides an encrypted IPsec tunnel towards the mobile terminal. Data is transmitted in the same way between the GANC and the mobile terminal as between a GERAN base station and the mobile terminal; however all data frames that are for GERAN transmitted via the GSM/EDGE radio interface are instead tunnelled through an IPsec tunnelled over the WLAN access network. GAN allows to use circuit-switched telephony over WLAN in addition to packet-switched services; it can be considered as a form of *integration* according to [SFA03] and supports scenarios 1-6 of [3GPP22.934]. A generic approach to allow multiple access networks to provide connectivity to a common 3GPP core network, and allowing inter-system mobility with service continuity, is currently being defined in the 3GPP system architecture evolution [3GPP23.401] [3GPP23.402] [SO08].

One approach to provide access to Internet services to users via different access networks can be described as *wireless overlay networking*. In this case some functionality is located in a network beyond the access network and provides gatewaying and handover functionality. It contains the mobility anchor for the mobile device. This approach has been proposed by [KB96], and has later been investigated in several research projects with mobility management based on some form of mobile IP (e.g. DRiVE, OverDRIVE and BRAIN [TLVK02] [TMLW02] [AHPST+01]). The approach of a mobile IP based common core network for multiple access networks has been proposed in e.g. [WMH02]. In all these cases a mobile device is in charge of detecting and selecting the access network.

Chapter 3. System Description

3.1 System Model and Functional Entities

In this section we define the system architecture and functions for our considered multi-access system. By definition, a multi-access system is based on different communication technologies, which use their own terminology. We introduce generic connectivity abstractions in order to provide a uniform terminology applicable to all different communication technologies. These connectivity abstractions are largely based on those that we have developed within the Ambient Networks project [AN D1-5] [AN D2-4] [AN D2C1] [AN D7A2a]. The terminology used for functions is largely identical as in Ambient Networks. Since the overall scope of this work is covering only parts of the scope investigated within Ambient Networks, we have simplified the connectivity abstractions and functions where applicable.

We consider a communication system that enables communication between two end-systems, which we denote as the *user end-system* and its corresponding *peer end-system*. These end-systems are located in a *user network* (UN) and a *peer network* (PN) respectively. Each end-system contains a *service application function*; a communication session runs between service application functions of these end-systems via a *bearer*. This bearer is provided by *transport functions* at the bearer endpoints (BEPs). A bearer is transported via *flows*. Flows are connectivity elements that are provided by a certain communication technology; the physical realisation of a flow differs in different technologies. A flow requires three features:

- A **flow identifier** from a (technology specific) flow identifier name space. Flow identifiers can be either a single identification object, like a label in Multi-Protocol Label Switching (MPLS); alternatively, they can be a set of locators[18] of the flow endpoints (FEPs), like source and destination addresses in Internet Protocol (IP). The flow identifier can also contain a flow classifier, which characterises the required treatment of the flow along the communication path and in the flow endpoints.

- A **routing mechanism** that is based on the flow identifier.

- A mechanism to associate **service requirements** with a flow.

A flow is a unidirectional connectivity element. Some technologies automatically establish bi-directional connectivity by means of a pair of flows. A flow is limited to the domain where a common flow identifier name space is used[19]. At the boundary between different name space domains, flows are terminated and mapped to a corresponding flow in the other domain. By that, end-to-end connectivity for the bearer is provided by a series of flows, as shown in Figure 3.1 and Figure 3.2. Boundaries of the flow identifier name space typically correspond with technology boundaries. For example, one end-system can be located in an IPv4 network, and the corresponding end-system is located in an IPv6 network. In this case an IPv4 flow establishes the connectivity in the IPv4 domain, an IPv6 flow establishes the connectivity in

[18] A locator is the identifier of the location within the routing topology.
[19] In Ambient Networks this is often referred to as a *locator domain*.

the IPv6 domain, and at the domain boundary these flows are mapped to each other. A flow can span over network boundaries if they share the same flow identifier name space. A flow can have multiple end-points, thus establishing a point-to-multipoint connectivity for multicast services. However, in this work we consider only point-to-point flows and point-to-point services. Bearer management functions configure flows to the requirements of the bearer in a flow-technology specific manner. There are two alternatives for bearer management. In local bearer management (Figure 3.1) flows are only configured within the end-systems. These bearers are bound to a flow by a transport function in the end-system; the flow-setup application programming interface (API) provides some means for the configuration of the flow e.g. for QoS parameters. An example for such a bearer is an end-to-end transport protocol connection (e.g. TCP), which is bound to an IP flow in the end-systems; by setting the *diffserv codepoint* (DSCP) [RFC2474] in the IP header the requirements of the bearer can be specified. In case that multiple consecutive flows are used, the flow requirements are mapped at the flow end-points between the consecutive flows. An alternative approach is distributed bearer management as depicted in Figure 3.2. In this case bearer management functions are not only located at the bearer endpoints, but also within at least some of the intermediate networks. Bearer management signalling is used between the bearer management functions within the end systems and bearer intermediaries[20]. Common bearer management signalling protocols are the *session initiation protocol* (SIP) [RFC3261] and the *session description protocol* (SDP) [RFC4566], which describe and negotiate the requirements and parameters of multimedia applications, such as voice or video telephony. A bearer intermediary is then a SIP function located in the network, e.g. in the *IP multimedia subsystem* (IMS) [CG04]. Examples for other bearer management protocols are the *resource reservation protocol* (RSVP) [RFC2205] and the *next steps in signalling* (NSIS) framework [RFC4080]. The bearer intermediary can control directly the configuration of flows in the network.

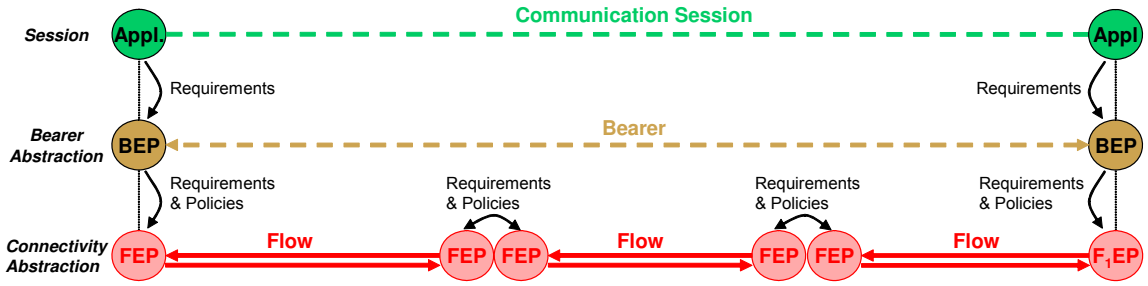


Figure 3.1: Relationship of bearer and flow connectivity elements with local bearer management.

[20] In IETF terminology a bearer intermediary is typically denoted as a middlebox, see e.g. [RFC3234] [RFC3303].

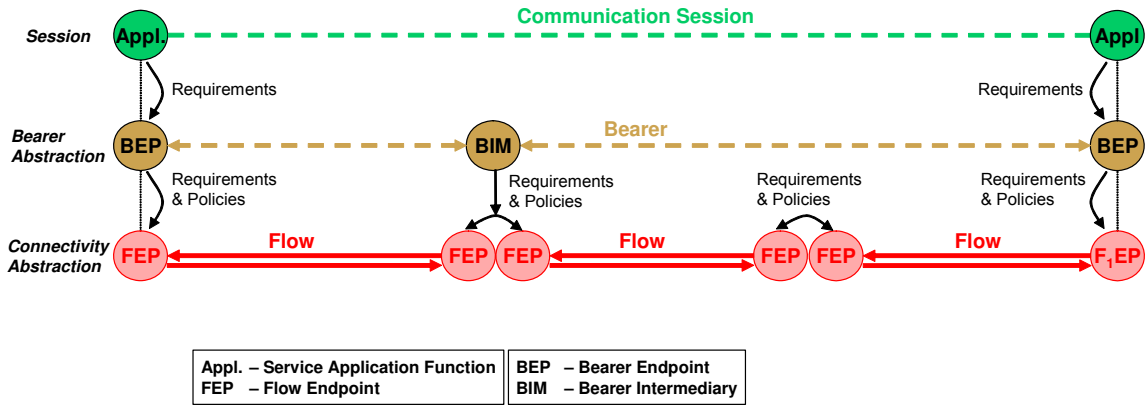


Figure 3.2: Relationship of bearer and flow connectivity elements with distributed bearer management.

In real networks flows exist on different hierarchical levels and can be nested, as shown in Figure 3.3. Different hierarchy levels can be caused by protocol layering, tunnelling and technology convergence. For example, an IP-level flow can be transported in a certain sub-network via MPLS[21] [RFC3032]: the complete MPLS domain is seen from the IP-flow level as a single link; within the MPLS domain data routing is based on the MPLS-flow which transports the IP-flow. Another example is, when a tunnel is established: the flow which is passed through the tunnel perceives the tunnel as a single link. There can be several reasons for establishing tunnels: firstly, a tunnel can be used to bridge different sub-networks which belong together logically. For example, a private corporate network can be distributed over multiple locations and still operate as a single locator domain. The tunnel that connects the different sites at the same time provides data security by encryption, and thus establishes a virtual private network (VPN). Secondly, a tunnel can be used to support mobility. *Mobile IP* (MIP) [RFC3344] [RFC3775] and the *GPRS tunnelling protocol* (GTP) [3GPP29.060] are the most prominent examples of mobility protocols, where a higher-level flow (e.g. an end-to-end IP flow) is regionally transmitted by being encapsulated within the flow provided for mobility. Thirdly, a tunnel can provide technology convergence by enhancing a communication layer with additional features on the overlay level. One example is *generic access network* (GAN) [3GPP43.318] that provides circuit-switched mobile telephony services to a mobile terminal via a packet-switched IP network, which by itself is incapable of providing these services. The requirements of a flow at one level need to be mapped to corresponding requirements of the flows that are used at the next lower level, as indicated in Figure 3.3. There can be several levels of flows, depending on the specific networking scenario.

[21] MPLS could be considered as a Quality-of-Service-enhanced *underlay* of IP.

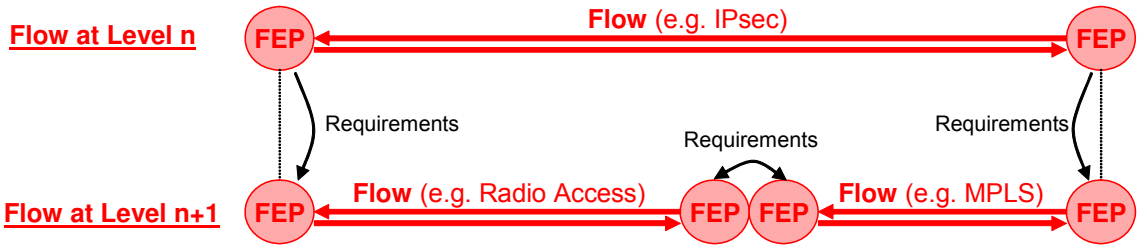


Figure 3.3: Nested flows on different hierarchy levels.

The mapping of flows – either on the same flow level or between different flow levels – is handled in *forwarding points* (FP). A FP contains a forwarding engine, which performs data routing by mapping incoming flows to outgoing flows, and a forwarding state, which contains the rule set that governs how the flow mapping of the forwarding engine has to be made. The forwarding state can be static, as in the case of VPNs, or change dynamically, as in the case of mobility controlled flows. The forwarding state in the forwarding point is controlled via some forwarding control, which is realised by a flow-specific communication technology. For example, the setup procedure of a VPN, or the mobility management update procedures of GTP or MIP are forwarding control commands.

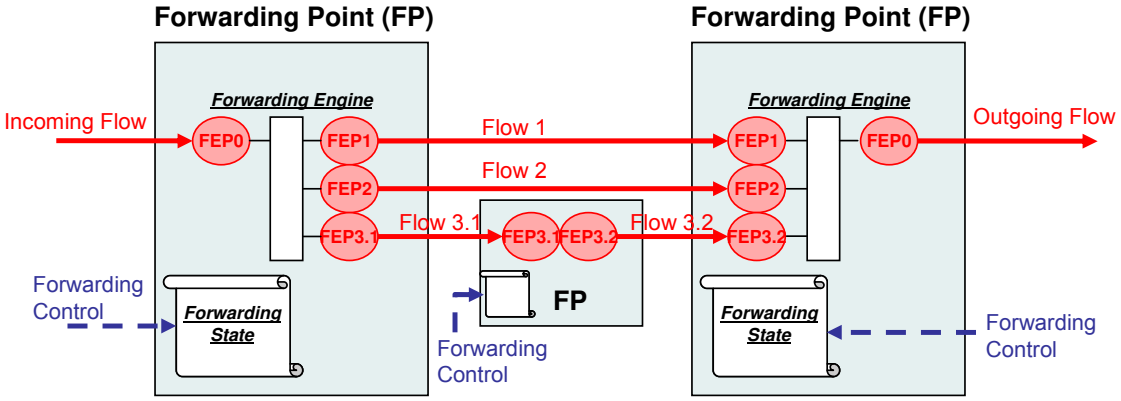
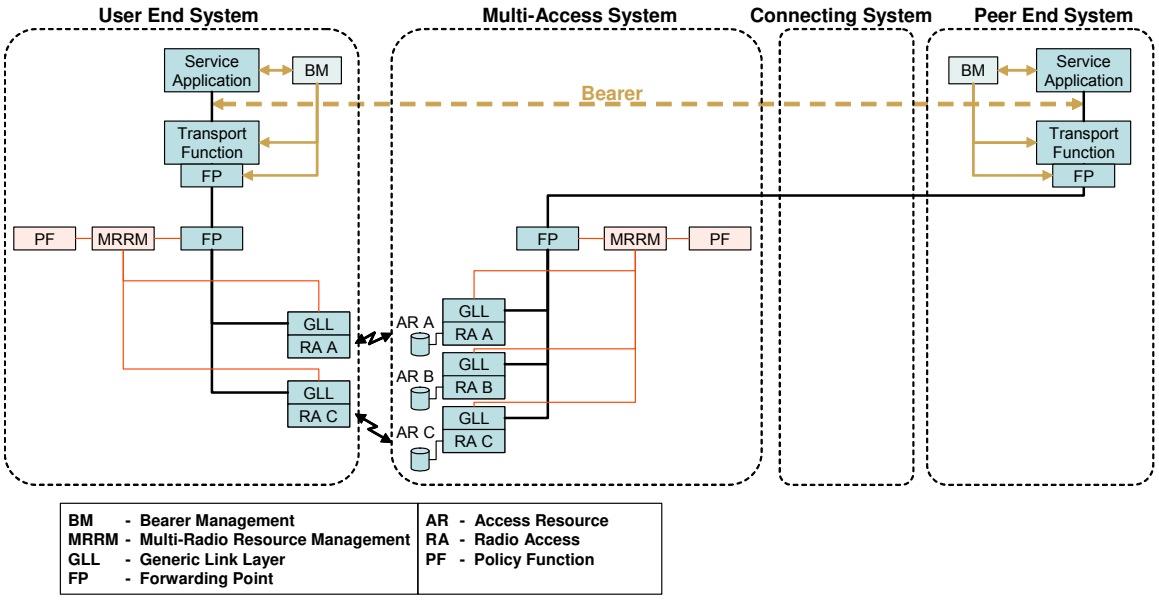


Figure 3.4: Forwarding Points (FP) at flow endpoints.

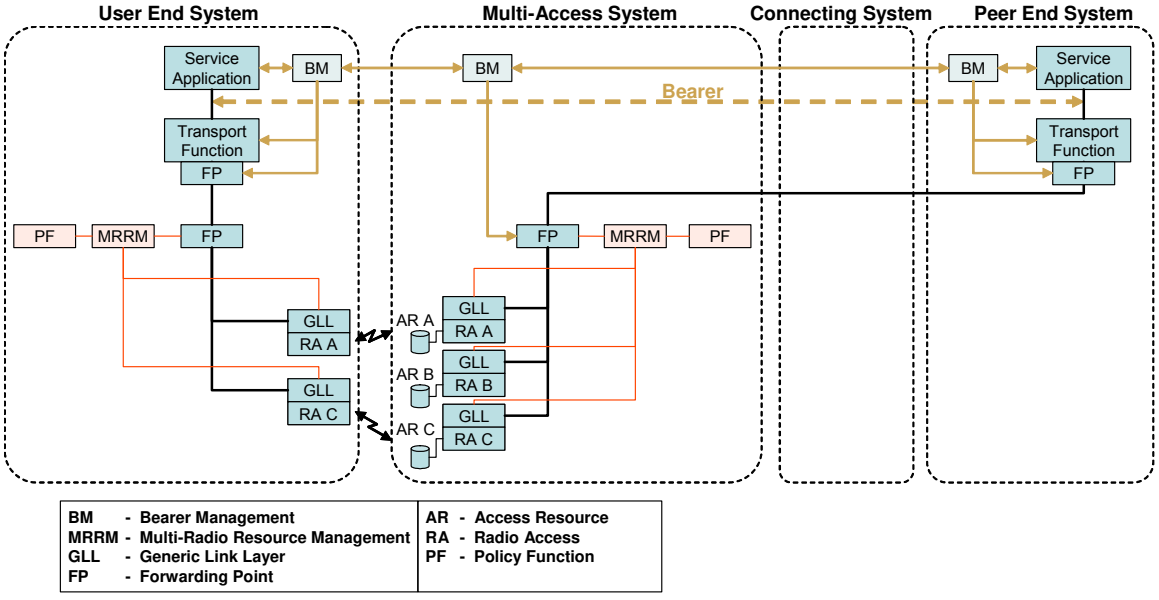
The functions performed by a FP are similar to those of a network layer intersystem gateway according to the Open System Interconnection (OSI) reference model [Hal92]. However, the OSI model – with its seven layer protocol structure – assumes the incoming and outgoing connectivity of the gateway to be link layer connections. We assume the FP to operate independent from particular protocol layers. For example, the flows 1-3 in Figure 3.4 can be provided by different communication protocols at different layers. Also the flows 3.1 and 3.2 can be at different layers.

Figure 3.5 shows the reference architecture used in this work. A service application in the user end system has a communication session, transferred via a bearer, with a corresponding service application in a peer end system. The end systems provide transport functions to the bearer, and map it to an end-to-end flow via a FP. The transport function and FP are configured by the *bearer management* (BM) function to match the service requirements and service mapping policies. Bearer management can be located either in the end systems only (Figure 3.5 (a)). Alternatively, it can be distributed over several communication systems

(Figure 3.5 (b)) with a signalling connection between the bearer management entities. The user end system is wirelessly connected to a multi-access system, which provides connectivity to the peer end system, possibly via some further connecting system. The connectivity between the user end system and the multi-access system is made up by a number of *radio access* (RA) options, which are based on *access resources* (ARs). The radio accesses and access resources are controlled by a *generic link layer* (GLL). A *multi-radio resource management* (MRRM) function monitors the properties and characteristics of the radio accesses, and steers how the service data flow is routed to the radio accesses. We denote as *service data flow (SDF)*, the flows which enter the multi-access system and for which the access connectivity is provided. This is done by controlling one or more FPs that map the service data flow to one of the radio accesses. The flow which transports the service data flow through the multi-access system we denote as *access flow*. Access selection thus consists of determining the access flow to which a service data flow is bound. The user end system and the multi-access system contain *policy functions* (PF), which contain rules of how radio accesses can be used by the user end system. The user end system functions are located on one or more interconnected devices under the authority of the end user; we also refer to the user end system as the *user network* (UN). As we see later, the functions in the multi-access system can be located in nodes belonging to different authorities, so the multi-access system can consist of multiple networks.



(a)

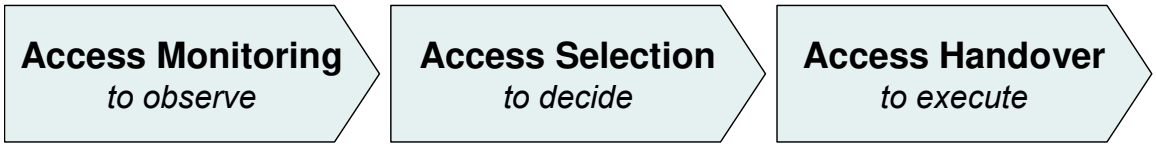


(b)

Figure 3.5: Reference system architecture with (a) local and (b) distributed bearer management.

It has to be noted, that the Ambient Networks terminology for connectivity abstractions used in this thesis differs from terminology used elsewhere. In [SPG07] [SO08] we have provided a description of multi-access connectivity according to the terminology used in 3GPP.

3.2 Multi-Radio Access Management



The key functionality in a multi-access system is *access selection*: how should the multi-access system choose the best one given the availability of a number of accesses that can be used for the transport of a service data flow. The access selection function is performed by *multi-radio resource management*. However, access selection is only part of a larger access management process. Before access selection can be performed, it is required to learn which accesses are available for each user network. Furthermore, information needs to be collected in order to determine the suitability of every access for a data session. Since different access technologies have different metrics for describing the access performance or availability of resources, a common abstraction of suitability has to be found which allows comparing information for different accesses. We denote *access monitoring* as the process to collect necessary information for access selection. Access monitoring is performed by the *generic link layer*. Once the access selection function has made the decision to change the active access, it is required that the service data flow is redirected from one access flow to another one. For this the binding of the service data flow to access flow(s) – in short *SDF binding* – has to be modified in the corresponding forwarding points. This process we call *access handover*. Access flows can be based on different types of access technologies and mobility protocols. Consequently, access flows are identified by different descriptors, for example, the locators of the access flow endpoints. Depending on the technology also the handover procedures can vary by which the SDF binding is changed. A forwarding point may need to support multiple handover procedures for different access flows. Multi-radio resource management steers a handover execution function to perform the access handover according to the appropriate handover procedure. We call this handover execution function *handover and locator management* (HOLM), according to terminology used in Ambient Networks [AN D1-5] [AN D2C1] [AN D7A2a] [AN D20B2]. HOLM keeps track of the locators of an access flow and the handover procedure/protocol that is required to update the SDF binding. It thus contains a toolbox of handover tools, from which the appropriate tool is selected depending on the source and target access flow. The handover toolbox can also comprise handover optimisation tools. Such tools can support context transfer and data forwarding for seamless and lossless access handover, or network mobility for moving networks. Depending on the capability of the source and target access flows, such optimisation tools can be used to increase the access handover performance. Handover optimisation is provided by the generic link layer. When a handover command is received by MRRM to perform a handover between two access flows, HOLM determines the suitable mobility protocol for the access flows. It furthermore determines what handover optimisation tools can be applied, e.g. for context transfer. For the handover execution, the SDF binding is updated in the corresponding forwarding points with the appropriate handover protocol, like e.g. GTP, MIP or Proxy MIP (PMIP).

In summary, one can say *access monitoring* is the process that provides sufficient information to the *access selection* process. *Access handover* is the process that realises access selection decisions.

3.3 Access Sets

For the management of multiple accesses we define different *access sets*, as shown in Figure 3.6. These access sets are used by the MRRM entities in the user network and the multi-access system for access monitoring, access selection and access handover. The *detected access set* (DAS) contains all access links that are detected by the user network, including those to which it is already connected. An element is included when a new access system is detected. An element is removed from the detected set when the connection to the access is lost. The *validated access set* (VAS) contains all accesses of the DAS, which are validated by the policy functions according to installed policies. For example, certain networks can be barred for usage for the user network. Once a service is invoked and a service data flow is setup, it has to be decided which access to use for that service data flow. A *candidate access set* (CAS) defines all those accesses with capabilities that match the requirements of the service data flow and that comply to service-specific policies. The access selection decision is then to select the best suited access from the CAS. The selected access is included in the *active access set* (AAS), and it is the access to which the service data flow is bound during the access handover. It is, in general, possible to split a service data flow onto several accesses, so the AAS can contain multiple elements. However, typically the AAS contains only a single access. These four access sets are used for basic multi-access management. Two additional sets are further used to support access monitoring. The *expected access set* (EAS) contains the accesses that a user network is expected to be able to connect to. It is determined from the user position – given geographically or by the radio cell(s) it is currently connected to – and multi-access neighbour list information. A multi-access database containing neighbour cell relationships can be dynamically maintained based on user network measurements, or by network configuration. From the EAS a *scanning access set* (SAS) is determined, which contains accesses that the terminal is directed to scan for. It is determined based on, e.g. user network capabilities about which accesses can be used by the user network. The SAS provides hints to the user network to scan for new accesses. These hints contain, for example, the type of RAT, the carrier frequency and the provider name. The EAS and SAS are in particular useful if a user network is restricted in its capabilities, e.g. due to battery limitations or if measurements and connectivity for multiple accesses cannot be performed simultaneously.

For each user network there exists one detected, validated, expected and scanning access set; in contrast, a user network can have multiple candidate and active access sets – one for every service data flow. The detected and the scanning access set are used for access monitoring; the SAS supports access discovery and the DAS contains the result of access monitoring. Access selection determines the validated, candidate and active access sets. It also determines the expected access set, e.g. if no suitable candidate set can be determined.

The access sets have been developed within Ambient Networks work on multi-radio access [AN D2-4] [AN D2C1] [SPG07] [JSRJ06] [TPSPS+07].

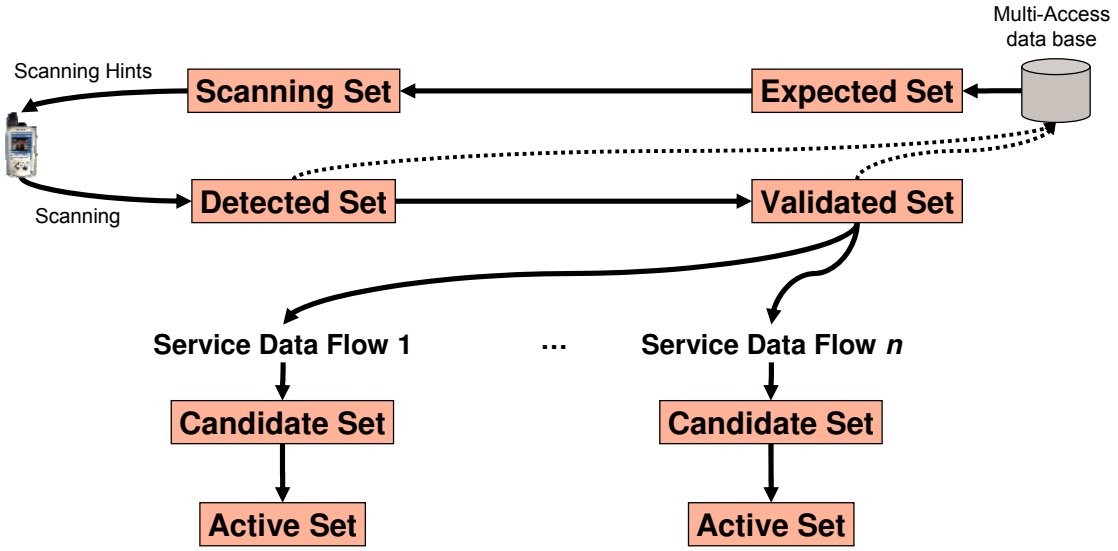


Figure 3.6: Access Sets used for multi-access management.

3.4 Evaluation Criteria

In order to verify the benefits of the concepts developed in this work evaluation metrics are required. In the following we describe the evaluation metrics that we use for different aspects of multi-access management.

Access Selection

The gain of access selection can be described by quantifying the improvement to the multi-access system by being able to allocate service data flows flexibly to different accesses. This improvement can be described from a user and from a system perspective. The user throughput is a measure that describes the service performance that an individual user perceives for a service data flow. From a system perspective, the gain of access selection is the total number of users or service data flows that can be served in the multi-access system at a given level of quality. The user and the system metrics are related: access selection can increase both the user throughput and the system capacity. This gain can be exploited in different ways. For a given traffic load and network deployment users can perceive a higher data rate and service performance when selecting the best access – the beneficiary is mainly the end user. Alternatively, the increase in efficiency by access selection can be used to either admit more users into the system, or to reduce the number of radio cells in the network, which results in a gain in system capacity (in number of served users) or in reduced network deployment costs. In this case, the main beneficiary of the access selection gain is the access provider; for the end user this gain is not directly noticeable. We consider system capacity as the primary metric for assessing the gain of access selection. Access selection is described and evaluated in Chapter 4.

Access Monitoring

The goal of access monitoring is to provide sufficient information to the access selection process to evaluate and select available accesses. This information must be able to quantify the suitability of an access for a service data flow. In order to evaluate and compare the

information provided for different accesses, such information must be comparable. Access monitoring is successful, if it provides abstract information for any access that is meaningful for access selection according to the access selection strategy. A second aspect that describes the performance of access monitoring, is the amount of signalling and delay that is required in order to provide useful information to the access selection function. Access monitoring is described and evaluated in Chapter 5.

Access Handover

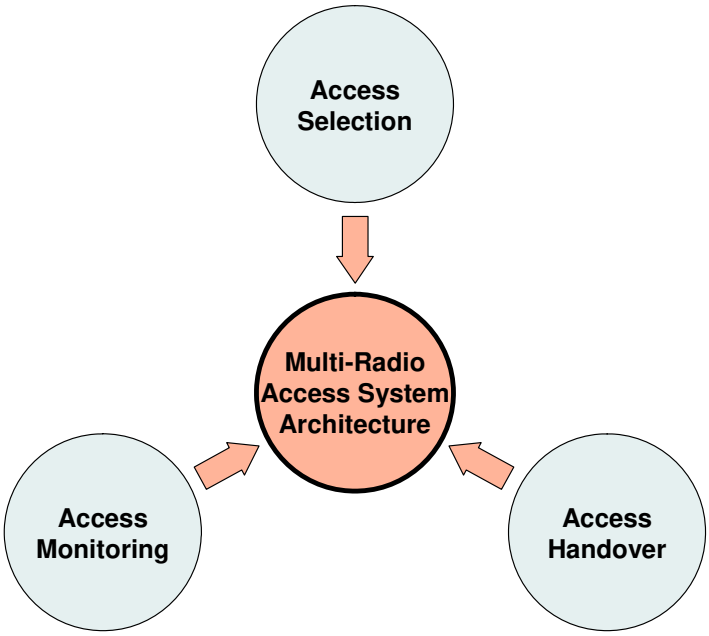
Access handover has to redirect the service data flow from one access to another one according to the access selection decision. The performance of access handover can be described by the amount of distortion that it causes to the service data flow. This distortion can be described in terms of handover interruption delay, the amount of data loss, the amount of data duplication and re-ordering. Different applications have different sensitivity towards such distortion. Furthermore, applications typically use an end-to-end bearer that is provided by some transport protocol. These transport protocols, like TCP, can also be affected by data distortion at access handover. In the case of TCP, data distortion affects the transmission rate that TCP can provide to an application. A suitable metric to evaluate access handover performance is therefore to determine the data rate perceived by the application above the transport protocols. For this we define an object bit rate as the data rate at which an application data object is transferred to the receiving application end-point. Access handover is described and evaluated in Chapter 6.

Multi-Radio Access System Architecture

All functionality to manage multiple access technologies needs to be embedded in a system architecture. Such a system architecture can be evaluated according to how well access selection, access monitoring and access handover is supported. For example, an architecture can prohibit certain information to be available and thus disable a certain type of access selection algorithm. On the other hand, a system architecture needs to be flexible to be adapted to different usage scenarios, e.g. where different functions are provided by different administrative authorities. It also needs to be sufficiently scalable to support large networks with many users. Assuming that a large number of access networks and access technologies already exist today, another evaluation criterion is to what extent existing technologies can be embedded into such a multi-radio access system architecture, and what amount of changes are required for existing technologies. The system architecture is described and discussed in Section 3.5.

3.5 Multi-Radio Access System Architecture

3.5.1 Introduction



The different components for the management of heterogeneous access technologies are *access selection*, *access monitoring* and *access handover*. Before investigating those functions individually in later chapters, we discuss how these functions can be embedded in a common *multi-radio access system architecture*. In this section we derive requirements on a multi-access system architecture[22]. We derive functional components for the multi-access system architecture. We present and analyse different alternatives of mapping the functional components onto a network architecture.

3.5.2 Objectives and Requirements

All system components that support and enable the integration and interworking of different radio access technologies need to be integrated into a common multi-radio access system architecture. Such an architecture needs to fulfil several requirements.

Sophisticated Access Selection

An architecture needs to support sophisticated access selection algorithms. It must be possible to select the access in accordance with network and user policies. Furthermore, access selection has to provide the access connectivity which best matches the service requirements. In order to optimise the efficiency of used radio resources and increase the total system capacity, the load and availability of resources of the access system needs to be considered for access selection to facilitate load management.

---

[22] Many people – at Ericsson Research, as well as within the *multi-access* workpackage in *Ambient Networks* – have contributed to the architecture model

Efficient Support for Access Handover and Access Monitoring

For sophisticated access selection, sufficient information about the existence, performance and characteristics of accesses needs to be available. The multi-radio access system architecture needs to provide functionality and interfaces to provide such information about the individual access systems.

When an access selection decision results in an access handover, the multi-radio access system architecture needs to provide functions that enable the handover execution without excessive degradation of the service performance.

Flexibility to Support Different Business Scenarios

In a multi-access system it should be possible that different business actors provide different components of the multi-access management functionality. Therefore, multi-access management functions need to enable cooperation between different business entities and should be based on open interfaces.

Scalability

A multi-access system needs to be scalable. This requires a certain degree of distribution of functionality within the system architecture, to avoid that all information needs to be collected and processed in a single entity.

Evolvability of Existing Architectures

A large number of access technologies are already deployed in today's networks, and need to become part of the multi-access system architecture. Therefore, it is required that existing network architectures can be evolved into the desired multi-access system architecture.

3.5.3 Multi-Access Functional Reference Architecture

The multi-access architecture can be described based on a reference architecture which specifies the functional entities. The functional reference architecture is depicted in Figure 3.7 and contains three main functional entities. The central multi-access control entity is *Multi Radio Resource Management (MRRM)*. It collects information about available access flows for each user network and allocates one or more of these to a service data flow. It thus performs access selection and related multi-access functions like admission control and load management. For the assessment of the different access flows it determines utilities based on policies, resource and performance metrics. A result of access selection is typically either to stay with the current access or that a handover towards another access shall be executed. The *Generic Link Layer – Interface and Context Transfer* ($GLL_{I\text{-}CT}$) provides a generic interface towards MRRM and support functionality for transmission over an access link. Based on certain rules and thresholds (event filtering and classifications) it reports link events (triggers) to MRRM. In case that a flow is handed over between different GLL entities, it supports link layer context transfer.

The *Multi-Access Anchor (MAA)* which is a forwarding point that maps service data flows to access flows. It is the entity where handovers are executed. Each MAA entity needs to store the active mapping (i.e. SDF binding) of service data flows to access flows. Note that there can be multiple MAAs which are then typically structured in a hierarchical manner. In

addition the MAA may be combined with the *GLL context anchor* functionality (GLL_{CA}) to support link layer context transfers where copies of data packets are kept in the GLL_{CA} until the $GLL_{I\text{-}CT}$ entities signal that the packets have been successfully transmitted.

The multi-radio access entities can be implemented in different ways to suit different networking scenarios. While the MRRM is often depicted as a single box, it has to be stressed that it generally comprises multiple physical entities that are typically located in different nodes. Figure 3.7 shows an example of a physical implementation of multi-access with a distributed MRRM. The following MRRM entities are included in this architecture:

- $MRRM_{ASF}$: The access selection function which is the master MRRM entity responsible for deciding on the best-suited access flow for a service data flow.

- $MRRM_{ANF}$: The access network control function which configures measurements in the access network via the $GLL_{I\text{-}CT}$ and monitors access network related parameters like e.g. cell-load.

- $MRRM_{CMF}$: The connection management function which monitors the performance of access flows for the user network.

The separation of MRRM functionality is in particular useful, in case that the MAA and the $MRRM_{ASF}$ are centralised in the network hierarchy, and need to manage a very large number of users and radio cells. For scalability reasons, it is then advantageous if the $MRRM_{ASF}$ receives only limited information. For example, the $MRRM_{ASF}$ does not need to know the exact link quality for every access link of all user networks, but it is only informed by $MRRM_{CMF}$ if a link quality becomes critical or a new link is discovered.

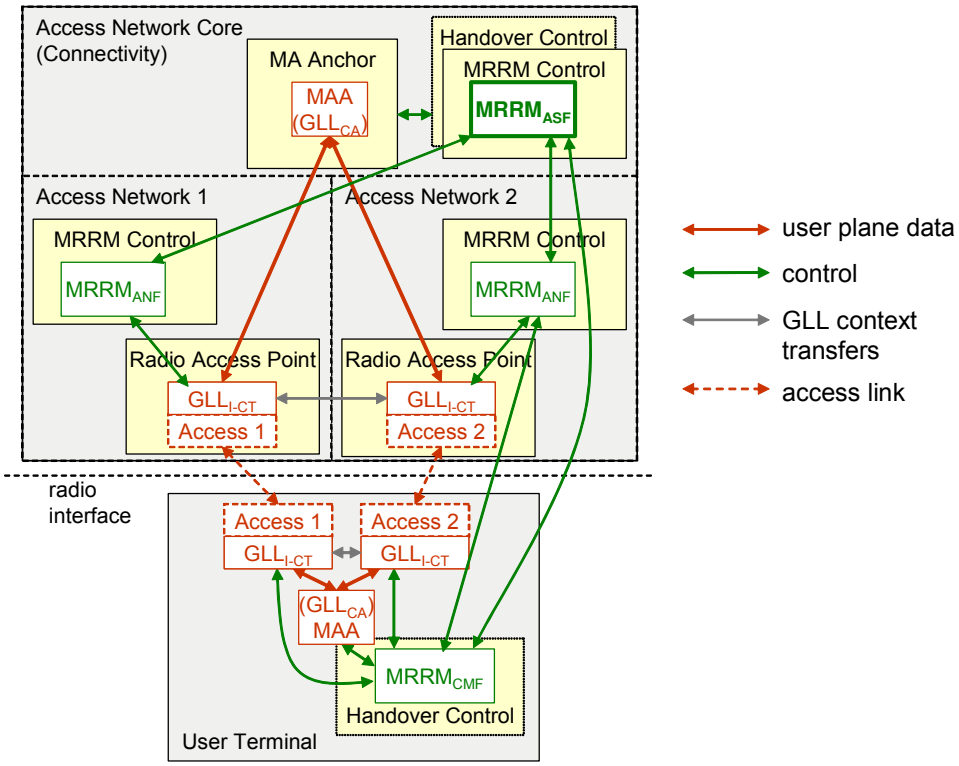


Figure 3.7 : Functional multi-radio access reference architecture.

The initial access selection – when the user network tries to establish first connectivity to access networks – takes always place in the user network. The selection can be steered or supported by policies and network capability information that have been installed in the user network. Once the user network is connected, it reports service data flow requirements to the access network as well as measurements on the actual link performance. Additional discovered accesses are also reported to the access network. Based on this information and information collected in the access network, access selection takes places.

The multi-access architecture depicted in Figure 3.7 can be applied in different business scenarios, which is later investigated in Section 4.4. For example, different access networks can belong to different business actors. Depending on the type of cooperation the distribution of functionality may vary. For example, an untrusted access network may not comprise MRRM and GLL functionality; instead the multi-access functionality is only located in the access network core and the user network. The distribution of functionality between network domains of different network operators can be based on pre-established business agreements (e.g. roaming agreements) or establishing dynamic cooperation agreements based on network composition [NSAMS+04] [AN D7-A2] [KPJS07] [3GPP22.980]. In particular two business interfaces need to be supported from a multi-access system architecture, as shown in Figure 3.8. The *roaming interface* provides connectivity to home services for users that cannot directly connect to the home network, but which can only connect to a visited (core) network. The *access roaming interface* provides users access to a (home or visited) core network from an access network that only provides access connectivity.

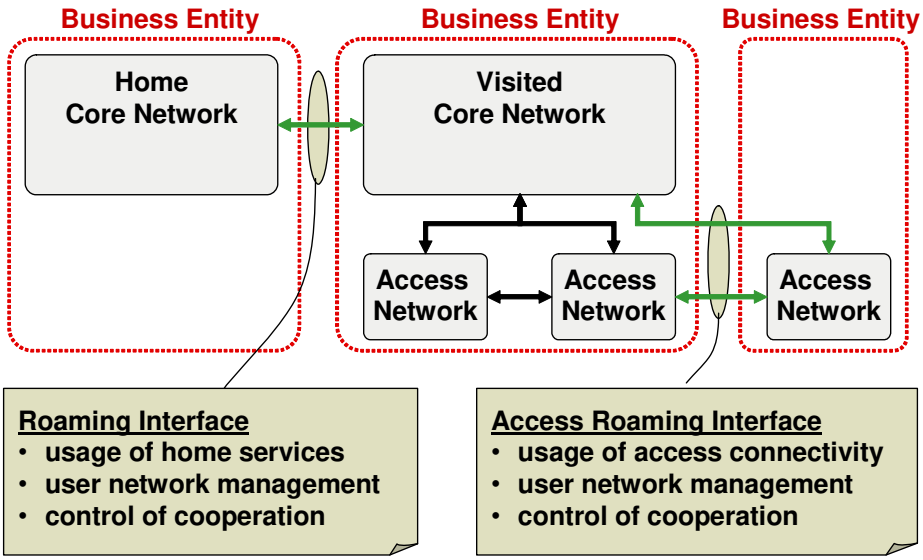


Figure 3.8 : Business interfaces in a multi-radio access system architecture.

3.5.4 Realisations of a Multi-Radio Access System Architecture

The functional components of the reference architecture can be located in different nodes of the system architecture. These differences lead to different levels of integration. In this section we present three alternative realisations of a multi-radio system architecture and discuss to what extent the requirements are met.

3.5.4.1 Integrated Multi-Radio Access Network

A system architecture that integrates different radio access technologies is depicted in Figure 3.9. Different radio access technologies are connected to a common radio bearer gateway (RBG). The radio bearers of the different radio access technologies are all terminated in the same radio bearer gateway. This means that the RBG also contains the Generic Link Layer functionality and acts as multi-access anchor[23]. The radio resources of different radio access technology are managed jointly by coordination of radio-specific radio resource management (RRM) functions. The multi-radio resource management function can be either part of the radio bearer gateway; it can also be located separately, e.g. to have a common RRM function for multiple radio bearer gateways. The radio bearer gateway is connected to a multi-access core network, which contains access gateways (AGW) which provide connectivity to external data networks, like the Internet, a public-switched telephony network (PSTN) or other service networks. The core network additionally maintains information about user-related policies (e.g. subscription), location and reachability.

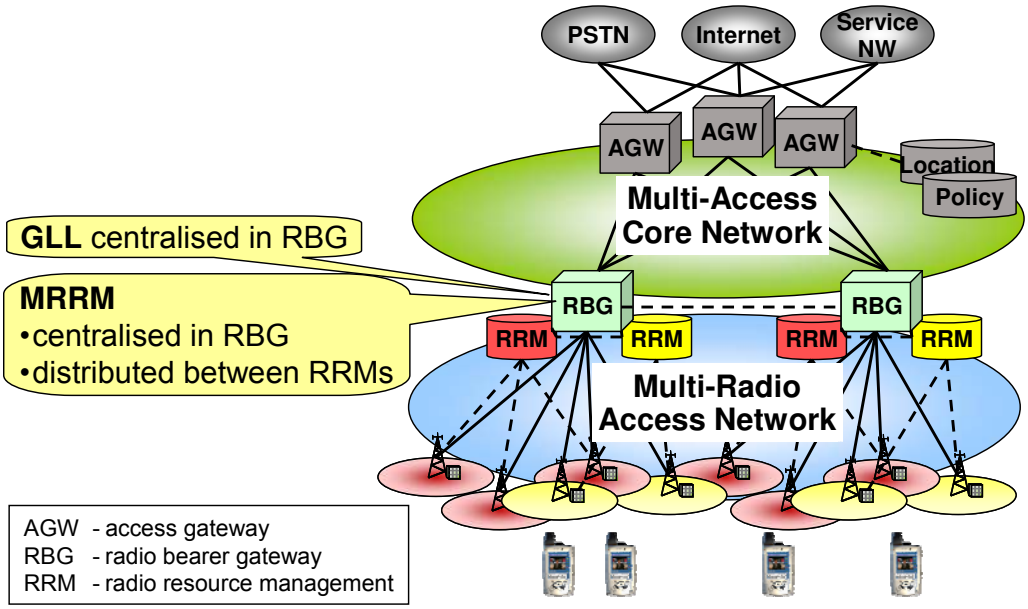


Figure 3.9 : Integration of different RATs within a common *multi-radio access network*.

An integrated multi-radio access network allows for sophisticated access selection algorithms. The MRRM access selection function is co-located or close to access-specific radio resource management functions. This provides sufficient information about the performance and resource situation of alternative accesses. Information about access-related policies can be obtained from the core network at the initial setup of the radio access flow. Also access handover can be well supported by this architecture. The multi-access anchor is located in the radio bearer gateway, where also the generic link layer and most of the communication context resides. Therefore, an access handover can be performed easily and seamless for the service. Some difficulties only appear for users that are located at the border region between two radio bearer gateways. In this case access selection and access handover requires some additional coordination between RBGs. An integrated multi-radio access network is scalable,

[23] As later described in section 6.5, the Generic Link Layer functionality can be either realised as *multi-radio generic link layer* or as *generic link layer interworking*.

since all functionality related to multi-access management is handled locally in the network topology. The scope of MRRM and the RBG is restricted to a local geographic area; all transmission via different access technologies is jointly managed and coordinated. Centralised network functions in the core network are only involved for the exchange of policies and registration of the access in use. The complexity of integrating different radio access technologies can vary. In case that new common functionality in the RBG is for different RATs all theses RATs need to be modified accordingly[24]. In other cases the data transmission remains according to the different RAT procedures and is complemented with new interworking functions[25]. For the access selection and access handover new interworking functions between the different access technologies are required, as well as a common interface towards the common core network. A disadvantage of the integrated multi-radio access network architecture is to support different business scenarios. The common multi-radio bearer gateway integrates parts of the different radio access technologies. This makes it difficult to have separate business entities provide the different radio access technologies. One way how a business separation can be achieved is by providing a tunnel from the radio bearer gateway to the user network via the access provided by another business entity. This approach is similar to the generic access network (GAN) [3GPP43.318] and interworking WLAN [3GPP23.234] defined by 3GPP. However, the tunnel approach of those solutions make access selection based on radio resource information difficult. A possible solution to support a business interface between a radio access point and a radio bearer gateway that allows transferring access resource and performance information across this interface. Such an interface could be provided based on the control and provisioning of wireless access points protocol (CAPWAP) [ID-CAPWAP]. CAPWAP provides the configuration, management of multiple WLAN access points from a centralised access controller. This protocol could be extended to support the resource management information for multi-radio resource management, as well as comprise business cooperation management functions to provide an *access roaming interface*. Furthermore, the protocol could be extended to support other access technologies than WLAN. With those extension the RBG can include the GLL context anchor (GLL_{CA}) and multi-radio resource management, and the radio access points can comprise the GLL interworking (GLL_{ICT}) function.

3.5.4.2 Integrated Multi-Access Core Network

Another system architecture that integrates different radio access technologies is depicted in Figure 3.10. For every radio access technologies a separate radio access network (RAN) is provided, which contains radio access points, radio bearer gateways and radio resource management functions, which may be integrated with the radio bearer gateway. The radio bearer gateways connect the radio access networks to the common multi-access core network. In this case the common multi-access anchor is located within the core network at an access gateway. Multi-radio resource management functionality is achieved by having a multi-radio resource management entity located within the multi-access core network. The RBGs in each RAN contain the generic link layer functionality, as well as the MRRM access network control function ($MRRM_{ANF}$). All different radio access networks need to support the common core network functionality. This includes functions for authentication, authorisation and accounting, mobility management, policy control and Quality of Service management. A harmonisation of this functionality is required for all radio access networks.

---

[24] An example is the multi-radio generic link layer with multi-radio segmentation described in section 6.5.2.2.
[25] An example is generic link layer interworking described in section 6.5.3.

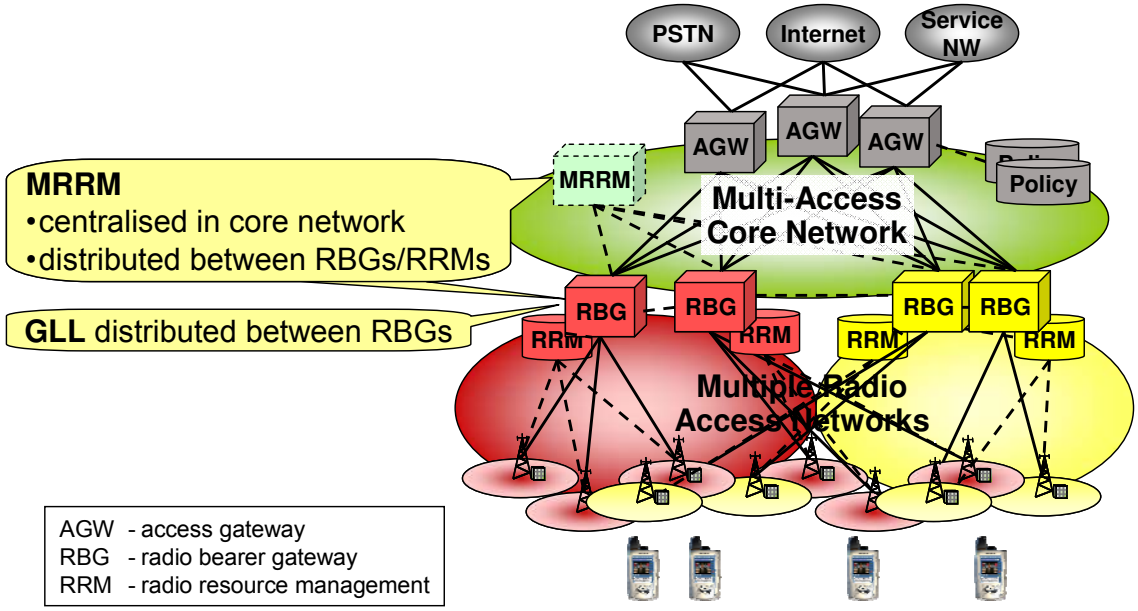


Figure 3.10 : Integration of different RANs within a common *multi-access core network*.

Management of different accesses via different radio access networks provides some difficulties regarding scalability. Access selection functions become part of the centralised core network functionality. If sophisticated access selection functions – based on the changing availability of resources and changing link performance – are performed in a centralised location, the amount of information signalled into the core network becomes large, in particular if timely reporting of access-related information is desired. This imposes some restrictions on the access selection algorithms that can be supported. Access selection algorithms can only operate on changes of parameters on low time scales. The $MRRM_{ANF}$ functions located within the radio access networks, as well as the $MRRM_{CMF}$ function in the user networks need to filter radio resource information and signal this information to the centralised $MRRM_{ASF}$ function only when critical thresholds are reached. If an access handover is to be executed, it requires that the connectivity is redirected from an access flow provided by one RAN to an access flow provided by another RAN. Core network mobility functions are always involved in the access handover. In order to make such a handover seamless to the service, the procedures need to be well coordinated between the core network functions (MRRM and AGW) and the old and new RAN functions (GLL and MRRM). Also some context transfer or bicasting function is required (see later in Section 6.5.3). Little changes are demanded from existing radio access technologies, which can evolve independently within their RAN. On the other hand, for a higher sophistication of access selection and access handover, the procedures between the different RANs and between the RANs and the core network become increasingly complex. An advantage of the separation of RANs in this architecture is that different RANs can be easily operated by different business entities.

3.5.4.3 Hybrid Multi-Radio Access Network Architecture

The previous discussion has shown that both architecture alternatives, based on a multi-radio access network or separate radio access networks, have advantages and disadvantages. The advantage of the integrated multi-radio access architecture is that in a scalable manner it allows a higher sophistication of access selection algorithms and access handover. At the

same time, multiple radio access networks enable easier cooperation of different business players to provide separate radio access networks that are combined into a common multi-access system. The advantages of both architectures can be combined in a hybrid architecture, which connects multiple radio access networks to a common core network, where some of the radio access networks can be integrated multi-radio access networks. A hybrid architecture is depicted in Figure 3.11.

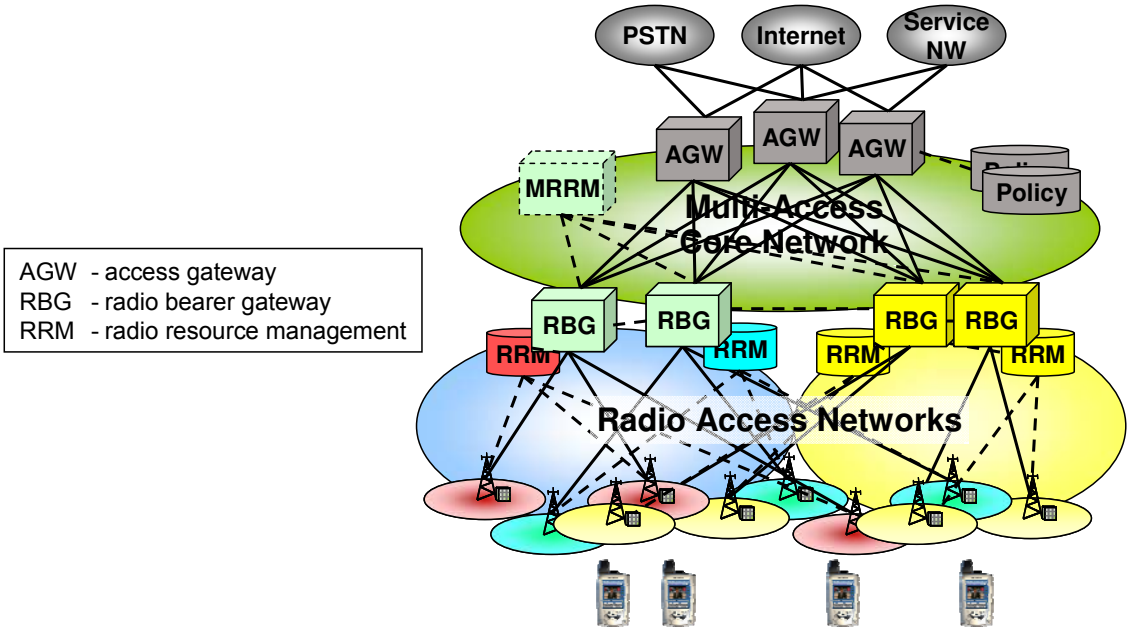


Figure 3.11 : Integration of different RANs with *hybrid multi-access network*.

Access systems can be integrated into the multi-access system architecture in two different ways. They can either be *tightly integrated* by common multi-radio RAN functions, or they can be integrated by common core network functions. The most feasible integration depends on two aspects: the practical viability of harmonising different access technologies, and the business operation of the multi-access system. The integration can be performed according to the following guidelines as summarised in Table 3-1:

1. Different access systems are only feasible for *tight integration* within a multi-radio RAN if the technical development of the access technologies can be harmonised with acceptable effort. This requires that sufficient coordination between the standard developments of the access systems occurs[26]. It also requires that an evolution of an access technology is viable. For example, if an access technology is already widely deployed on the market with little potential of improvement compared to the development of new access technologies, it may be impractical to enhance the access technology with functionality for tight integration.

2. If different access systems are operated by different network operators, the level of cooperation between the operators determines if tight integration is feasible.

[26] In most practical cases this limits the feasibility of tight integration to access systems developed within the same standardisation forum or when different standardisation fora agree to develop a common solution.

Table 3-1 : Loose vs. tight Integration.

	Viable harmonisation between different access technologies • common standardisation forum	**Harmonisation not viable** • different standardisation fora • widely deployed access technology with little further evolution
Tight cooperation between operators of different access systems	Tight integration	Loose Integration
Loose cooperation between operators of different access systems	Loose Integration	Loose Integration

3.5.5 Summary

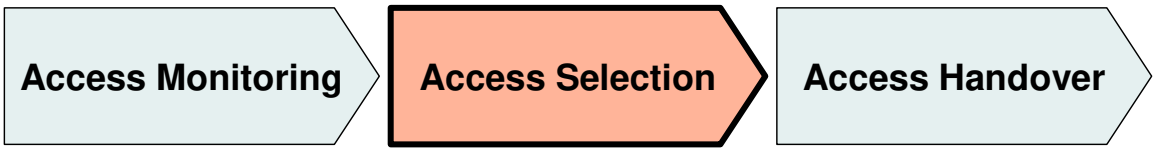
In order to build a multi-radio access system all required functionality needs to be embedded in a common system architecture. We have developed a functional reference architecture to realise multi-access functions for *access monitoring, access selection* and *access handover*. A multi-radio access system architecture needs to be scalable, must be able to support different access technologies including those that are already deployed, and should support different business scenarios of cooperation between operators. There are different alternatives how a system architecture can be realised. We have investigated and compared three different alternatives of a multi-radio access system architecture, as summarised in Table 3-2. An *integrated multi-radio radio access network* provides tight integration of different access technologies. It provides a scalable architecture with good support for multi-access management functions. While it largely re-uses core network functions – like AAA, mobility management, QoS and policy management – it requires adaptation of radio access network functions for multi-access management. On the other hand it does not easily support loose cooperation between multiple operators. Loose integration is provided by an *integrated multi-access core network* to which multiple independent radio access networks are connected. It provides little extra functionality of the individual access technologies, but requires a harmonisation of the core network functionality. While integration via a multi-access core network supports flexible business scenarios, it supports only limited features of multi-access management. A good trade-off is a hybrid approach, where different access technologies can be integrated either tightly via the RAN or loosely via the core network. The mode of integration option can be adapted to the business scenario. Sophisticated multi-access management only needs to be supported by access technologies for which this is feasible and the required harmonisation in standardisation is viable.

Table 3-2 : Realisations of a *multi-radio access system architecture*.

	Integrated Multi-Radio RAN (tight integration)	**Integrated Multi-Access Core Network (loose integration)**	**Hybrid Multi-Radio RAN and Multi-Radio Core Network (loose and tight integration)**
Access Selection	Good	Poor	• Good for tightly-integrated access systems • Poor for loosely-integrated access systems
Access Monitoring & Access Handover	Good	Poor	• Good for tightly-integrated access systems • Poor for loosely-integrated access systems
Scalability	Good	Poor	Good
Evolvability	• Poor for existing access systems • Medium for new access systems • Good for core network functions	• Poor for core network functions • Good for access systems	• Poor for core network functions • Medium for new access systems • Good for existing access systems
Business Flexibility	• Good for *roaming* • Medium for *access roaming*	• Good for *roaming* • Good for *access roaming*	• Good for *roaming* • Good for *access roaming*

Chapter 4. Access Selection and Multi-Access System Capacity

4.1 Introduction



Access selection is the decision process for determining the best access connection(s) among the available ones for the service data flows of a user network. The objective is to achieve a gain – for the service and / or the network – compared to a static access allocation. In this chapter we first investigate access selection, which is the key functionality in a multi-access system. In subsequent chapters we then describe the supporting functions of *access monitoring* and *access handover*, which complement *access selection*. In this chapter we discuss different access selection strategies and utility functions, which determine the value of the access allocation. The objectives of access selection depend on the business scenario and which business entity provides which function within the multi-access system. We discuss possible business scenarios. The utility for an access allocation depends in any case largely on the transmission properties, for example, the link quality of an access connection. Therefore, we discuss the transmission characteristics and develop a taxonomy of the gain achievable by access selection[27]. We derive numeric results for the capacity of multi-access networks for different network layouts and algorithms, which are obtained from stochastic simulations[28]. Furthermore, we develop an analytical model based on stochastic knapsacks, which can be used for an analytical evaluation of network capacity[29].

4.2 Related Work

Although there is a considerable amount of work developing multi-access systems, comparatively less related work can be found on the algorithms and performance of access selection. One reason is that a multi-access system requires a typically standardised architecture framework that defines the interactions of different system components. In contrast, access selection algorithms can be implementation specific without a need of harmonisation between industry players. Furthermore, the need for access selection is only slowly arising: the industry has only just started to build simple multi-access systems, for example, by enabling to connect to a cellular core network via WLAN; so far only very few mobile devices support the corresponding capabilities. Another reason is that until recently access selection algorithms were very simple. They assumed the performance of one access technology to be significantly superior to another access technology so that the access selection simplifies to "whenever the superior access technology is available, use it." With the

---

[27] The taxonomy and the evaluation of gains have been jointly done with Per Magnusson.
[28] The simulator has been developed by multiple people at Ericsson Research, in particular Anders Furuskär, Jonas Pettersson, Arne Simonsson, Oya Yilmaz and Per Magnusson. The simulations have been performed by Per Magnusson.
[29] Gabor Fodor has pointed to stochastic knapsacks as a modelling approach.

technical advances made for several different access technologies such obvious choices do not exist anymore and smarter logic is therefore required. In this work we present a framework for access selection logic that can target different objectives.

In general a trunking gain can be achieved when the resources of multiple access systems are bundled and jointly allocated. [THH02] have investigated some basic trunking gain in a combined GERAN-UTRAN system with arbitrary system allocation. It shows small trunking gain for real-time and large trunking gain for non-real-time services. However, the work makes very simplified assumptions for radio cells and does not consider the capacity characteristics of a radio network. We develop in this work a new model based on stochastic knapsacks, which overcomes those limitations. [BJL05] demonstrate a trunking gain when GERAN, UTRAN and WLAN radio resources are bundled for the service allocation compared to individual allocation. As access selection principle a priority order is assumed, where a lower priority access system is allocated if higher priority systems are not available or overloaded. Furuskär et al. have investigated service-based access selection in mixed-service scenarios [Fur02] [FZ05]. The allocation of services to access systems is based on a service-efficiency of the different access systems, which is derived from the capacity region. For a mixture of speech and best-effort services a significant capacity gain is shown in a combined GSM/EDGE and UMTS multi-access system. Koo et al. [KFZK04] have derived analytically the trunking gain of service-based access selection, however, based on very simplistic models of the radio capacity. In [FFL04] [Fur03] service-based access selection is extended by adding the resource costs in the access selection algorithms. An additional gain is achievable; however, it requires that resource costs can be well estimated. A different approach to access selection is investigated in [KKD05] [BL06] [KDKK06]. Fast selection between different access systems is performed depending on the instantaneous radio link quality. This access selection can provide diversity against multi-path fading. Substantial gain is demonstrated if measurement delays are negligible. Yilmaz et al. [YFPS05] [Yil04] investigate link-quality-based, as well as load-based access selection algorithms for different radio network deployment and traffic load scenarios. Their work is the basis for our investigation in Section 4.5.3. By access selection the efficiency of the access systems can be increased. As we will show, this increase in efficiency can provide a higher capacity. Alternatively, the increase in efficiency can be used to reduce the network deployment costs while maintaining the capacity. The savings in deployment costs have been investigated in [CKPHM+05] [FAJ05] [Joh05] [JF05] [JFKZ04] [PKCBK06]. While previous work has focused on evaluating particular scenarios we develop in this work a general taxonomy for the gain that can be achieved by access selection; we describe what kind of gain can be achieved depending on the access selection algorithm and the network scenario.

The objectives of access selection depend on the distribution of multi-access functionality among different system actors (e.g. network operators and users) and their business relationships. The business relationship in multi-access systems and resulting consequences for access selection have been investigated in [GJ03] and within the Ambient Networks project [Vik05] [HJM04] [MJ06] [MMK06] [MPSL06] [RHM06] [CKPHM+05]. Access selection based on dynamic market mechanisms – with a focus on a multitude of independent wireless local area networks – has been investigated by Blomgren and Hultell [BH07a] [BH07b] [Hul07]. We develop in this work a generalized model of business scenarios and investigate the influence of business relationships on access selection decisions.

4.3 Access Selection Principles

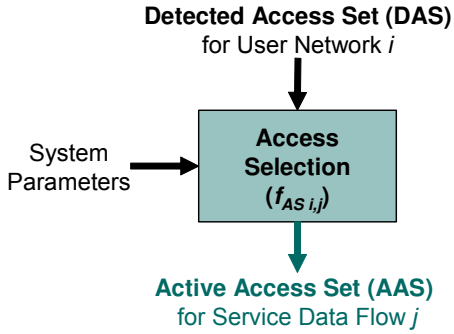


Figure 4.1: Access selection function for the service data flow *j* of a user network *i*.

An access selection function can be defined, as the function which determines for a service data flow j of a user network i the active access set of radio accesses which are used for communication out of the detected access set of radio accesses. This relationship is depicted in Figure 4.1. In a formal way, an access selection function, f_{AS}, can be expressed as the function that maximises a global utility function, u_G, over the set, A, of possible allocations of RAs to sessions:

$$f_{AS} = \arg\max_{a \in A} u_G(x, a) \quad , \tag{4.1}$$

with:

X : parameter space of system state,

x : system variable specifying all system parameters, $x \in X$,

A : set of possible allocations of sessions to radio accesses,

a : allocation of sessions to radio accesses.

As we will show later, the global utility is a combination of the utilities for different system components. The access selection function depends on several system parameters x within the parameter space X. Note, that the global utility function u_G as formally defined in eq. (4.1), describes the allocation of *all sessions* of *all user networks* within the considered system. This reflects the fact that there can be interactions between access allocations for different user networks, for example, when they compete for the same system resources. In a realistic realisation, access selection would be performed in a distributed manner for every user network individually (as shown in Figure 4.1). The interaction between the access allocation algorithms for different user networks is then reflected by these algorithms using and affecting the same system variables. For example, if a user network is assigned to one access technology, a nearby other user network will see this allocation as a change of the system variable describing the available resources.

Access selection is typically triggered at specific events, like

- *User network bootstrapping*: the UN is powered-on and detects available access technologies and networks.

- *Service data flow setup*: the user starts a new data session and the connectivity for the service data flow is established.

- *Changes in service data flow configuration*: the characteristics of the data sessions change and the service data flow requirements change accordingly.

- *Service data flow termination*: a data session and the corresponding service data flow are terminated.

- *Changes in radio link quality of an access connection*: the radio link quality and resulting link performance changes, e.g. due to movement of the user.

- *Changes in the resource situation of an access resource*: the load level of a certain access resource changes, which triggers a load management event.

- *Detection of a new access*: a new access becomes available and becomes a new candidate access.

- *Loss of connectivity for one access*: the number of available accesses is reduced, and possible the currently active access connection becomes disconnected.

- *Access handover rate exceeds a certain threshold*: the rate at which handovers are performed becomes too large.

- *Changes in system parameters*: changes in tariffs, authorisation policies, user preferences are examples for events that can trigger access selection.

4.4 Business Scenarios and Objectives for Access Selection

In a multi-access environment connectivity can be provided by different business entities. For example, different network operators can provide access connectivity to a user. In order to understand the motivation and objectives for performing access selection we have to understand the involved business players and their interests and relationships. In this section we define business roles and discuss how these roles are provided by business entities in different business scenarios. We discuss objectives and utilities for the different business roles, and derive a combined multi-objective system utility that is used for access selection.

4.4.1 Business Roles

In the service provisioning chain that provides a service to the end user different roles can be distinguished. Next we discuss the elementary roles of a communication chain, as depicted in Figure 4.2, in the context of enabling a user to choose between different accesses. In a later step we discuss how these different business roles can be taken by different business entities in varying business scenarios.

The *content provider* provides the content transmitted in a communication session. Content can be either stored or live-generated. Examples for stored data are music download, mobile-TV, video-on-demand, or general data exchange, when data is uploaded to or downloaded from a remote storage (messaging, file access, blogs, WWW-access, data backup). For live-

generated content, the communication peers provide the content directly, as for example in voice or video telephony.

The *service provider* provides the communication service to the end user. It provides content in a useful format for the end user and manages the communication sessions. One example are telephony service providers that offer audio-/video telephony, telephone conferences, or group communication (e.g. push-over-cellular), which can be enriched with messaging, and file sharing. Another example is the provisioning of mobile-TV, video or audio services to the end user in suitable formats. The service provider has a direct business relationship with the end user and it typically also acts as re-seller of the content stemming from third party content providers.

The *connectivity provider* provides connectivity for the end user to external networks, like the Internet, corporate networks or telephony networks. Furthermore, it provides reachability for the user, that a communication session can be established when initiated from communication peers. It coordinates the different accesses available to the user and enables dynamically changing the access in use.

The *access provider* provides the wireless access connectivity to a user. For different types of access technologies the coverage of the access service differs. It can be on local, regional or national level.

The *access broker* facilitates the business cooperation of access providers with connectivity providers. Connectivity providers typical operate on national level, or at least covering a larger region. Access providers can, in contrast, operate in small local areas only, in particular if based on local area radio access technologies. Access brokers bundle a number of access providers into a business relationship with the connectivity provider, and thus circumvent a large number of peered business relationships between access providers and connectivity providers.

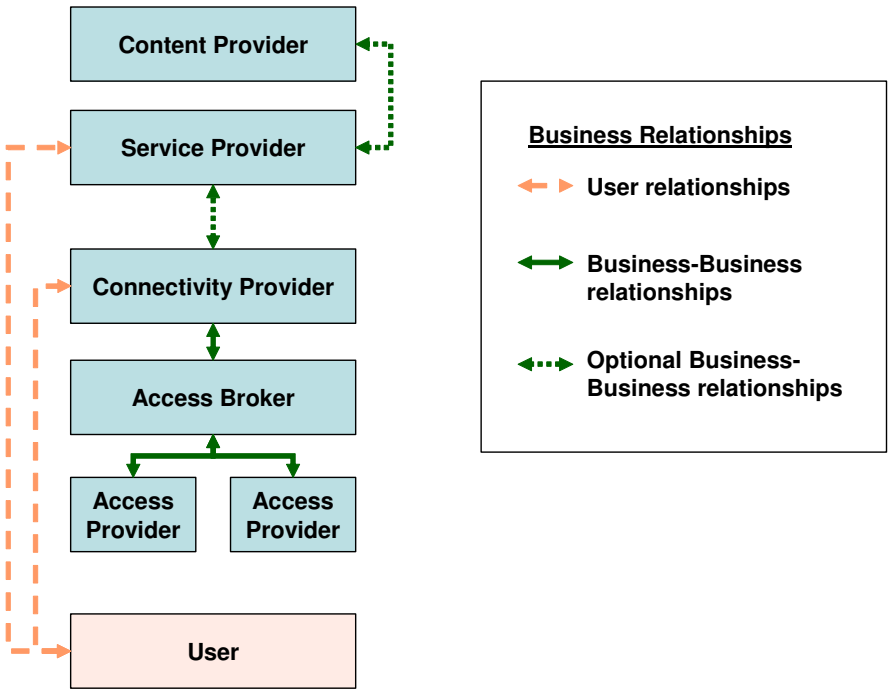


Figure 4.2: Business roles and business relationships

The user pays for the received services in monetary form. For the communication connectivity, the user compensates the connectivity provider. This compensation can be either per service, per amount of resources used during the session, it can be a flat-rate compensation package, or a combination of all these types. The connectivity provider distributes a part of the income to the access broker, which in turn provides compensation to the access providers. For the communication service, the user compensates the service provider. If the content of the communication service is not for free, the service provider compensates the content provider. These business-to-business compensation schemes can be based on the real contributions of different entities in the provisioning of the end user service, or based on a compensation package. If a business relationship exists between the service provider and the connectivity provider, a discount on the costs can be provided to the user for the bundled communication and connectivity services.

If different business entities provide complementary functionality they can reduce the costs of their operation. For example, if different access providers supply access in different regions or have only a low capacity, they can provide access with wider coverage and increased capacity by cooperation. Thereby each access provider can save investment costs, which is in particular beneficial for new access providers with high start-up investment. On the other hand, every business relationship is associated with a certain overhead. Firstly, some technical functionality is required to enable the cooperation; and secondly, the setup of business relationships involves certain costs.

In traditional cellular networks the mobile network operator combines and integrates all these roles. With the disintegration of the communications market these roles can be separated and new business entities can emerge, which provide some of these different roles. The disintegration of the market has been studied in [Vik05] [MMK06] [MJ06] [RHM06] [MPSL06]. To manage the increasing heterogeneity and resulting complexity, convergence of networking technology is required. To facilitate the diversity in business scenarios, dynamic cooperation of networking domains is a pre-requisite. This dynamic cooperation is also denoted as *composition*, which is a key concept developed in Ambient Networks [NSAMS+04] [KPJS07] [AN D7A2a] and also pursued in standardisation [3GPP22.980]. Business relationships can be diverse and complex, covering the complete range from competitive to cooperative scenarios. In this work we limit ourselves to cooperative scenarios. The business scenarios considered in our multi-access setting comprise a single operator with multiple access technologies integrated into a common packet core network. In addition, there can be different networks each comprising one or more access technologies. The cooperation can be similar to today's roaming scenarios between "equal" operators, but can also include cooperation with small access providers. An access may be untrusted, without direct cooperation between the access provider and the home connectivity/service provider. The evolved packet core network architecture that is currently developed in 3GPP [3GPP23.882] [3GPP23.401] [3GPP23.402] allows a realisation of several of those business scenarios. An overview of different business scenarios is presented in Annex A.

4.4.2 Utility-Based Access Selection

A utility is a measure to describe the satisfaction, value, profit or preference for a certain situation; it is a concept which is commonly used in econometrics and mathematical optimisation. We use it to describe the value of an access allocation within the access selection process. Utilities can be determined for different objectives and different entities. Some objectives can be opposed. An obvious example is the desire of a user to select the

access with the best perceived performance, while at the same time choosing the access with the lowest cost; typically both objectives lead to different results. For this reason it is required that a conflict resolution between different objectives is found. A larger value of the utility denotes that a situation or decision is favourable, whereas a utility of zero declares a situation or decision as unfavourable. A utility-based approach to access selection enables us to develop a system-level view of what different entities and interests need to be considered within the access selection process.

4.4.2.1 General Utilities

4.4.2.1.1 Service Utility

Access selection is performed to provide the best-suited access for a communication service; therefore, the degree of service satisfaction is a key factor in the access selection process. A service has a certain number of requirements. The *service utility* specifies quantitatively to what extent an access allocation fulfils these service requirements. In other words it describes the perceived performance of the access. Services put requirements on the communication system for the following metrics:

- Effective data rate, and variation of data rate,
- Transmission delay, and delay variation,
- Reliability of data transfer,
- Security of data transfer.

A service utility u_S can be defined for every access allocation a for the communication service of service data flow s. It is a multi-dimensional variable consisting of different components that quantify how well the different service requirements are met:

$$u_S(x,a,s) = p_s \cdot u_r(\tilde{r}_a(x),s) \cdot u_d\left(\tilde{d}_a(x),s\right) \cdot u_q(\tilde{q}_a(x),s) \cdot u_{se}(\tilde{s}_a(x),s), \qquad (4.2)$$

with:

p_S : priority for the service s,

$u_r(\tilde{r}_a(x),s)$: utility for an expected rate $r_a(x)$ and rate variation of access allocation a,

$u_d\left(\tilde{d}_a(x),s\right)$: utility for an expected transmission delay $d_a(x)$ and delay variation of access allocation a,

$u_q(\tilde{q}_a(x),s)$: utility for an expected reliability $q_a(x)$ of access a,

$u_{se}(\tilde{s}_a(x),s)$: utility for a security level $s_a(x)$ of access a.

x : system variable specifying all system parameters.

As the requirements differ between different services, consequently the service utility is also service dependent. Elastic services have the characteristics that they can use any capacity that

is given to them. The higher the data rate that is provided to such a service the higher is the perceived performance or service utility. Examples of elastic services are web browsing, e-mail synchronisation, the upload or download of photos, audio and video files. The rate dependence of the service utility is depicted in Figure 4.3 (a). It has a concave shape according to the economic law of diminishing marginal utility. It furthermore has a lower bound r_{min} below which the utility is zero. Conversely, discrete services have connectivity requirements given in discrete steps. For example, audio/video/speech conferencing or streaming services have data encoded with certain average rates and typically have certain delay requirements. An access that provides appropriate data rates and delays can support the service. If the data rate provided by the access increases further, it only brings additional value to the service if the next discrete step is reached, e.g. when the video conference can switch to a video signal with higher resolution. A special case of a discrete service is a telephony service, which has only a single discrete rate. Thus the requirement becomes binary, the service is supported if a sufficient data rate is achieved and a tolerable delay is not exceeded. A higher data rate does not add additional value for the service. The rate dependence of the service utility for a discrete service is given in Figure 4.3 (b). It consists of several sigmoid components around the discrete rates r_i. For a speech service there is only a single discrete rate, as depicted in Figure 4.3 (c).

The detailed analysis of the characteristics of the service utility for different types of services is beyond the scope of this work. Some further information can be found in [SPG07] [Oru07] [AN-SO07].

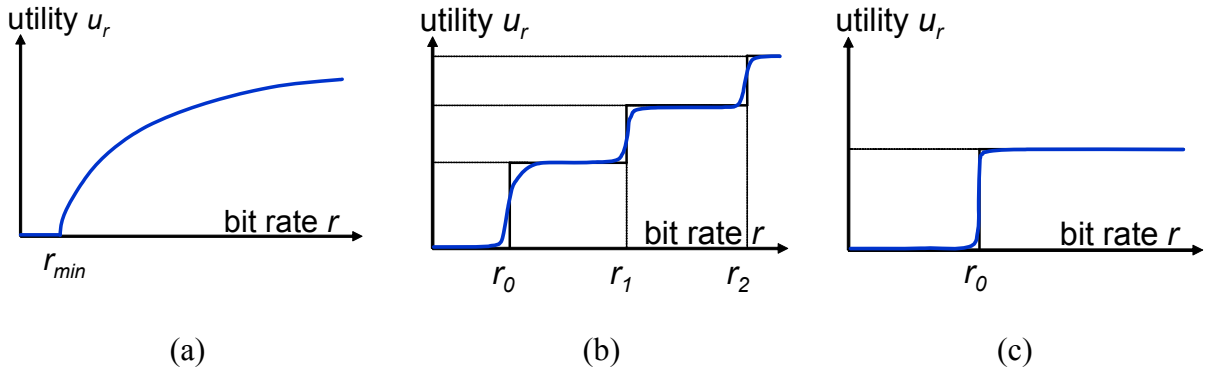


Figure 4.3: Utility of a service data flow depending on the data rate of the access flow for (a) elastic traffic, (b) discrete services, (c) speech telephony.

4.4.2.1.2 Resource Utility

An access flow is provided by different connectivity systems that provide the necessary transmission resources. For the allocation of a communication service to an access, certain resources are required. The *resource utility* describes a preference for the allocation of a service data flow to a specific access based on:

- the available resources associated with this access,
- the occupied resources associated with this access,

- the resource efficiency of using the resources associated with this access for the service data flow.

In general, these resources can be divided into access resources, connectivity resources and user resources as depicted in Figure 4.4. Access resources provide the direct connectivity for the end user to the multi-access infrastructure, i.e. to the point of attachment. Access resources can provide either fixed connectivity, like DSL or fibre, or they provide wireless connectivity. Connectivity resources connect the point of attachment to the multi-access anchor, which serves as mobility anchor and provides connectivity to external networks. The access anchor contains the service data flow filters and bindings.

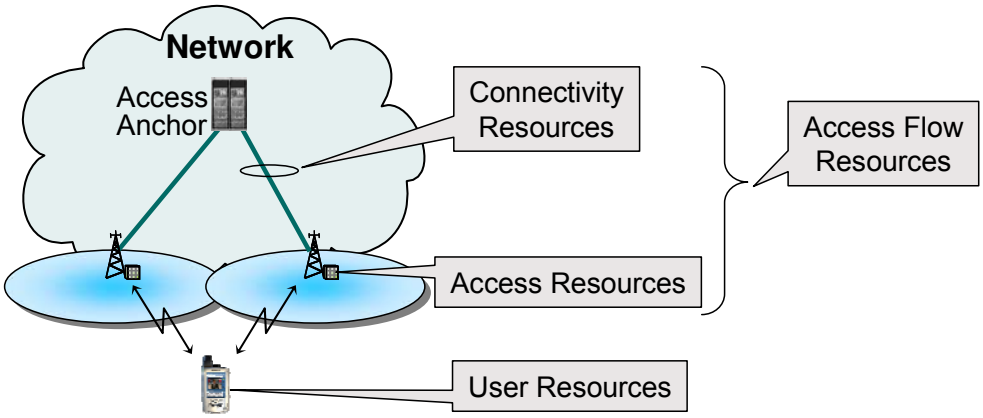


Figure 4.4: Resources used in the allocation of a service data flow to an access.

An abstract description of an access resource (AR) is necessary to capture the resource structure of different access technologies, in part to know the current capabilities (available resources) for handling service data flows, and in part to support operation when there is no service data flow active but an access system has been detected (for example, through reception of a beacon). The AR is a resource on which an access link can be established. This access link is part of the connectivity provided for the access flow. The access flow can span further than the access link (for example, to an anchor node) and may use other (non-access) connectivity resources for the remaining (non-access) connectivity. In wireless networks the AR corresponds to the radio resources of a radio cell, where the radio resources are allocated to active access links using some multiple access scheme (for example, TDMA, FDMA, CDMA, SDMA, or some combination thereof). In a fixed network the AR can correspond to, for example, the resources (transport, ports) of a DSL Aggregator/Multiplexing (DSLAM) node. At the setup of an access flow for a particular service data flow, admission control can be performed and access resources can be reserved. If resource reservation and admission control is used or not is dependent on the type of access technology and the service requirements associated with the service data flow. For best-effort service data flows typically no reservation is performed. In the case of service data flows with minimum quality of service requirements, cellular access technologies apply admission control and resource reservations [LEWL06]. For many other access technologies, e.g. WLAN, no dedicated resources are reserved. For evaluating the suitability of an access flow for a service data flow, the load and availability of access resources are important parameters. When different alternative access flows with sufficient access flow performance exist, the resource situation of the different access resources can be used to balance the load between the access systems.

An access resource utility $u_R(s)$ for a service s can be expressed as

$$u_{Ri}(s) = p_{Ri}(s) \cdot R_i \cdot L_i \cdot E_i(s) \quad , \tag{4.3}$$

with:

$p_{Ri}(s)$: preference value of access resource i for service s,

R_i : function of the amount of resources of access resource i,

L_i : function of the amount of occupied access resources (load) of access resource i,

$E_i(s)$: function of the resource efficiency of transmitting service s via access resource i.

Although the access resources constitute in most cases the bottleneck of an access flow, there exist other constellations when rather the connectivity resources are the bottleneck. An obvious example is a WLAN access point with a net peak data rate of approximately 27 Mb/s, which is connected to the access anchor via a DSL line with 6 Mb/s peak data rate. Even if the capacity of the connectivity network is larger than the capacity of an access resource, congestion can occur due to traffic aggregation of service data flows from a possibly large number of access resources. In these cases the limitations of the connectivity resources determine the suitability of the access flow. It is not trivial to evaluate the performance and resource situation of the connectivity resources. In some case, it may be a single connectivity link that connects the point of attachment of the access resource to the multi-access anchor of the access flow. In other cases, the point of attachment and the anchor can be connected via a connectivity path through a complete connectivity network. In this case traffic engineering methods can be applied in the connectivity network to avoid resource limitations. In the case of cooperating networks, the point of attachment can be even located in a visited network with the anchor being located in the home network; the connectivity path then leads through two connectivity networks connected via a third interconnection network. This plethora of options makes a general description of the resource situation and the performance provided by connectivity resources to a service data flow difficult. We have developed a practical solution in [GABBE+07], where a bottleneck of the connectivity resources is determined by a constraint value. This constraint represents all resources beyond the access resource jointly. It can be obtained, either dynamically by network management functions, or more static by operation and maintenance procedures. In some cases it may not be possible to detect a bottleneck of the connectivity path at all, in which case access selection is based on incomplete information. This constraint is a simple descriptor for the resource utility of connectivity resources. Similarly, there may be some limitations in user network resources. User network resources are, for example, the battery energy. In case that the remaining energy is low, a constraint can be determined for every access technology depending on its energy efficiency. In case that different batteries are used for different access technologies, a constraint can be determined depending on the available battery energy per access technology. This constraint is a descriptor of the user network resource utility.

The total resource utility $u_R(a,s)$ for an allocation a of a service data flow s to an access flow, can be described as the product of resource utilities for all resources i that are part of that access flow

$$u_R(x,a) = \prod_i u_{Ri,s}(x,a) \quad . \tag{4.4}$$

4.4.2.1.3 Business Utility

The provision and usage of access and connectivity resources is based on business relationships between the end user and the resource providers. The user provides monetary compensation for the resource usage, which is distributed among the involved resource owners. An allocation of a communication service to an access thus involves costs and revenues for the involved business entities.

For the end user, a business utility $u_{B\,u,s}$ can be defined which describes the cost efficiency of using an access for a service s. It relates the performance provided for the service, as well as the user resources required to the costs associated with the usage of an access allocation a:

$$u_{B\,u}(s,a,x) = \frac{p_{Bs}(s) \cdot p_{Ba}(s,a)}{C_s(a,s)}, \tag{4.5}$$

with:

$C_s(a,s)$: costs for using access a for service s,

p_{Bs} : priority value of service s,

$p_{Ba}(s,a)$: preference value for the access provider of a for service s.

The priority value p_{Bs} enables the user to choose the acceptable cost level per service. The priority value $p_{Ba}(s,a)$ specifies the preference for the access provider; it can be based on the trust that the user has for a certain provider and the type of tariff (e.g. if the tariff depends on the amount of usage).

Similarly, for all network business entities i that provide access and connectivity resources, a business utility $u_{B\,i}$ can be defined which describes the revenue of allocating service s to access a:

$$u_{Bi}(s,a,x) = p_{Bs}(s,a) \cdot R_s(s,a), \tag{4.6}$$

with:

$R_s(s,a)$: revenue for allocating access a to service s,

$p_{Bs}(s,a)$: priority value for allocating the service s of user to access a.

The priority value p_{Bs} defines the priority level that a user has for a certain access, which can depend on the type of business relationship; for example, if the user has a gold, silver or bronze type of subscription. Also loyal users, or users with a high amount of usage can receive an increased priority value.

Every business entity has the interest to maximise the accumulated value or revenue. As many users and service data flows share the same resources, the access allocation also comprises the selection for which users and service data flows resources are allocated. Consequently, service data flows that generate larger revenue are preferably allocated to the resource than service data flows that provide lower revenue. The business utility and the resulting access selection strategy depend on the compensation and tariffing scheme for the usage of access and connectivity resources. Next we briefly discuss with which access selection scheme the profit of a connectivity supplier is maximised, depending on the tariffing scheme.

In a *flat rate pricing* scheme, a fixed price is paid for usage of resources, independent of the amount of used resources. For a resource provider, the profit can be maximised if the number of flat-rate users is maximised, which means that the available resources should be distributed to as many paying users as possible. Every served user contributes equally to the revenue of the provider. As a consequence, in high load situations the service performance for every user is reduced to the minimum level. The minimum level is the lowest level before users start choosing alternative resource providers.

In *time-based pricing schemes*, the profit of the resource provider is proportional to the total minutes of resource usage. The revenue is maximised when the available resources remain unused for the shortest time, and resources are shared by as many users as possible. The users with lower data rate due to poor radio link quality contribute more to the profit of the resource provider, than users with high data rate. There is even an incentive for the resource provider at low load to reduce the achievable data rate for users in order to increase the duration of usage.

The revenue in *volume-based pricing schemes* is equivalent to the total amount of transmitted bits. The objective of access selection is to provide resources to users such, that the total system capacity is maximised. Users which achieve higher service performance contribute larger to the profit of the provider.

In *resource-usage-based pricing*, the revenue is constant per required amount of resources. A user with poor radio link performance has to pay a proportionally higher price than a user with good radio link performance in order to achieve the same level of service performance. This is analogous to *congestion pricing* where a user is charged according to its contribution to the congestion level of the resource.

In all of the above pricing schemes, users can be differentiated according to a priority class. For example, different tariffs and priorities can be used depending on if a user prefers a gold, silver or bronze tariff scheme. Thus users willing to pay a higher price per allocated resource generate higher revenue for the resource provider. An alternative pricing scheme for user differentiation is *auction-based resource pricing*. In this case the revenue per allocated resource is determined dynamically in a bidding process. Furthermore, the congestion level of the resource can be included in the auction according to a supply-versus-demand market mechanism.

Finally, pricing schemes can be *service-based*, which means that the price depends on the type of service. In this case, the profit of the resource provider is to allocate resources to services that achieve the highest revenue per allocated resource. Note, that a low-price service for a user with good radio link performance can provide higher revenue profit than a high-price service of a user with bad radio link performance, due to the more efficient usage of radio resources. Service-based pricing can be combined with the previous pricing schemes, e.g. the time, volume or resource tariffs can be made service dependent.

Different pricing schemes have different complexity, depending on the complexity for the setting of tariffs and accounting. Flat-rate pricing has the lowest complexity as the tariffs are static and no metering is required for accounting. Time-, volume- and service-based pricing have also static tariffs, but require additional complexity of metering for accounting. For resource-usage-based pricing; in addition, the tariffs need to be dynamically adapted. The highest complexity has auction-based pricing, where tariffs are also dynamic; furthermore, a resolution process with user interactions is required in order to determine the appropriate tariff. In the remainder of the work we do not further consider different pricing schemes.

Instead, we assume a pricing scheme that maximises the revenue when the highest system capacity is achieved.

The total business utility of the multi-access system is a combination of the business utilities of all involved resource providers. If different resource providers belong to different business entities the business relationships between these entities determine how the individual business utilities need to be combined. In the simple case that all resources belong to the same business entities, the total business utilities of all resource providers u_{Bn} is the sum of business utilities per resource:

$$u_{Bn}(s,a,x) = \sum_i u_{Bi}(s,a,x) \quad . \tag{4.7}$$

4.4.2.2 Combined Utility Functions

The different utilities derived in the previous section – service utility, resource utility, and business utility – are combined to a global utility u_G for access selection by means of a suitable operation:

$$u_G(x,a) = u_S(x,a) \circ u_R(x,a) \circ u_{Bu}(x,a) \circ u_{Bn}(x,a), \tag{4.8}$$

with:

$u_S(x,a)$: service utility,

$u_R(x,a)$: user utility,

$u_{Bu}(x,a)$: business utility of the user,

$u_{Bn}(x,a)$: business utility of the multi-access system (network).

Figure 4.5 shows the system model with the different utilities. The service utility reflects the performance of an access with respect to the service requirements. The resource utilities describe the efficiency of usage for allocating a certain access to a service. The business utility of the user describes the costs implied with the usage of an access; the business utility of the network describes the revenue obtained by a user allocation.

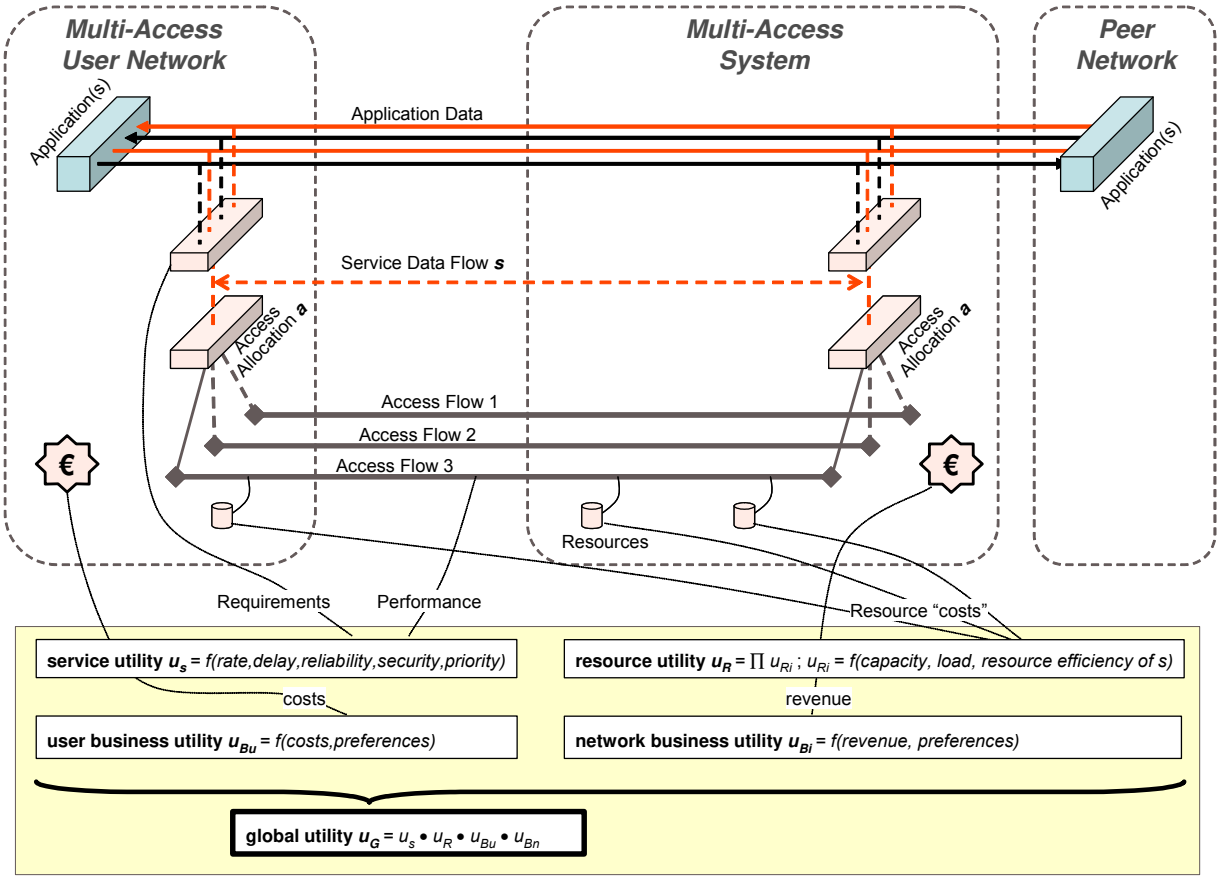


Figure 4.5: Different utilities combined to a global utility in a multi-access system.

The utility-based approach for access selection provides a multi-access system model that defines the impact of different system parameters on the access selection decision. It provides a mathematical description of the access selection problem. The ideal access selection solution for the multi-access system is found for the combination of allocations of user service data flows to accesses that maximise the global utility. The solution is limited by a number of boundary conditions, e.g., that the total number of allocations cannot exceed the capacity of all system resources. These conditions are implicitly included in the definition of the utility functions. For example, the resource utility tends towards zero if the service allocations to a resource reach its capacity. Also policy-like boundary conditions are represented in the utility. For example, the acceptable cost level for using an access is set in the definition of user business utility – for different user classes (e.g. business and private users) the utility can decrease at different cost levels. Similarly, security requirements are included by setting an appropriate minimum security level for the service utility.

For many realistic situations it may not be possible to describe all utilities explicitly in a simple mathematical form, like e.g. business utilities in a complex business setting. Instead, simplifications have to be made for specific scenarios. The utility model depicted in Figure 4.5 still helps to understand the relationship of parameters and indicate what assumptions are implied by a certain simplification.

4.5 Access Selection Performance

4.5.1 Access Selection Algorithms

The utility-based approach to access selection described in the previous section comprises the complete problem space of access selection. All relevant system parameters are collected and used to determine a global utility which is optimised. In practical realisations a distributed approach is desirable where the solution space is limited to feasible regions. Such an approach is discussed in this section.

In order to achieve the objectives of maximising the system utility access selection among the available accesses is performed for every user network. In the following we split the access selection function into two different sub-functions. This is motivated by the fact that the overall access selection function is a quite complex and large optimisation problem. A separation allows an easier modelling of access selection and a simplified two-step realisation of the optimisation problem. We separate access selection into *policy based access control* and *dynamic access selection*. The differentiating factor between these two sub-functions is, firstly, the type of system parameters used for the algorithm, and secondly, the time dynamics of these parameters. Static access control is based on system parameters that vary only rarely within the time frame of an ongoing session. These parameters can be considered as static, or semi-static for the rare cases that they change during a session lifetime. The type of access parameters are related to user or network policies. In contrast, dynamic access selection is based on performance related system parameters that typically vary dynamically during the lifetime of a session.

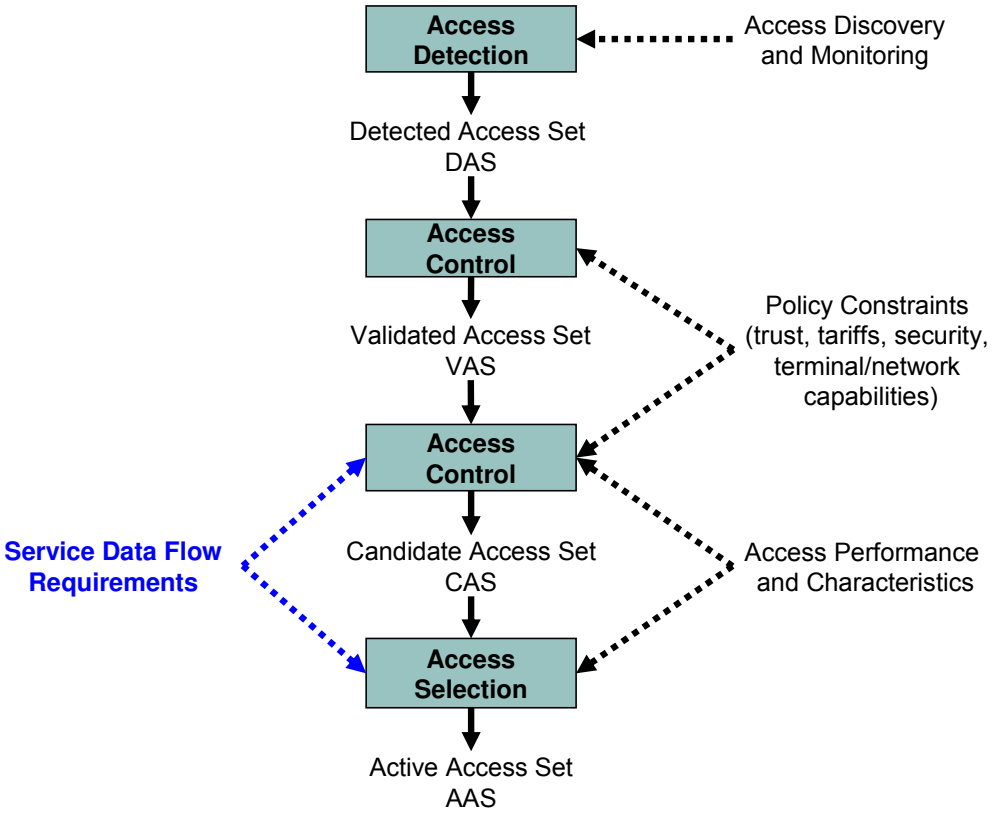


Figure 4.6: Access sets and two-tier access selection function for the sessions of a user network.

Figure 4.6 shows how these access selection functions relate to the *access sets* that are introduced in Section 3.3:

The *detected access set* (DAS) contains the access links detected by the access detection function of a user network, which performs access monitoring and scanning of available access resources. The elements of the detected access set are validated by the access control function according to access control policies of the user network and in the access network. Access control policies determine if a user network is authorised to use the access resources, and if an access network is acceptable by the user network. These decisions can be based on a number of policies. For example, it is validated if a sufficient trust level exists between the user network and an access network. Furthermore, the compensation scheme and tariff of using the access resource is validated, and if the minimum level of security provided by the access network is sufficient according to the user requirements. The access control function also validates if the access network provides the required capabilities which are also supported by the network. Such capabilities can be the used methods for authentication, security, mobility and communication protocols (e.g. IP version). Access links in the detected access set which fulfil the policy requirements are included in the *validated access set* (VAS). For a data service, the service requirements are specified for the service data flow. For every data service flow a *candidate access set* (CAS) is determined; the candidate access set contains those elements from the validated access set which provide sufficient performance to fulfil the service requirements and which comply with service-specific policies. The service-specific policies specify the acceptable costs for the user for that particular service, as well as service-specific security requirements. For example, for a mobile TV service encrypted transmission may not be required, conversely, for a telephony service it may be a requirement. Service-specific policies also determine if a user is authorised for the use of that particular service via the access network. Within the candidate access set dynamic access selection is performed. From the candidate access set the accesses which provide the best ratio of service performance and costs are chosen into the *active access set* (AAS), provided that the load level of the corresponding access resources is acceptable. While it is possible to transmit a service data flow via multiple access connections, typically only a single access connection is used (i.e. the AAS contains only a single element). Different service data flows have different requirements. As a result, there exist separate CAS and AAS for every service data flow, whereas the DAS and VAS is common for the user network.

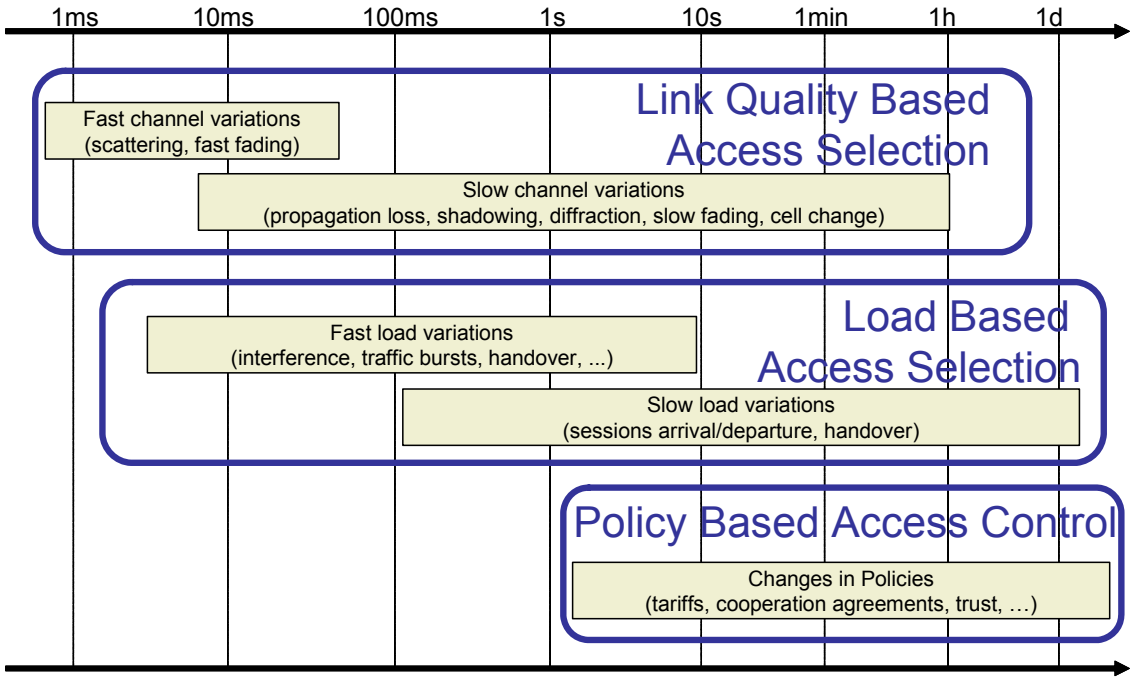


Figure 4.7: Time scales and classification of access selection.

Dynamic access selection reacts to changes caused by variations of the characteristics of the access links. Dynamic access selection can be categorised into two types as shown in Figure 4.7. *Link quality based access selection* reacts to variations of the access link quality. This can happen on different time scales. Fast variations are caused by rapid changes of the transmission channels, for example caused by fast fading due to multipath propagation. Slow link quality variations are mainly owing to user mobility and the consequent changes in channel path loss due to propagation loss, shadowing, diffraction. They can also be caused by access handovers, when the access link is switched to another radio access point. *Load based access selection* reacts to changes in the availability of access resources or changes in the demand for access resources, which can again be owed to changes in radio link quality. Fast load variations can be caused by changes in the interference level, traffic bursts or access handover between different cells. Slow load variations are effected by variations in the user density and traffic demand within a geographic region. *Policy based access control* reacts to changes in user or network policies, the dynamics are much lower compared to changes in link quality or load. Changes in policies can be time-varying tariffs (e.g. higher costs during busy hour), changes in the subscription or composition agreement, or changed user preferences (preferred networks, or security requirements).

The expected benefits of dynamic access selection are depicted in Figure 4.8. Compared to a static access allocation, dynamic access selection can improve the experienced service performance by selecting access connections with better link performance. The flexible usage of access resources of different access technologies can also increase the system capacity due to more efficient usage of available resources. If in addition load management is included in the access selection algorithm an additional gain in capacity and service performance can be achieved. However this effect becomes only visible at high system load. In the following sections we analyse the access selection gain in detail and present quantitative values.

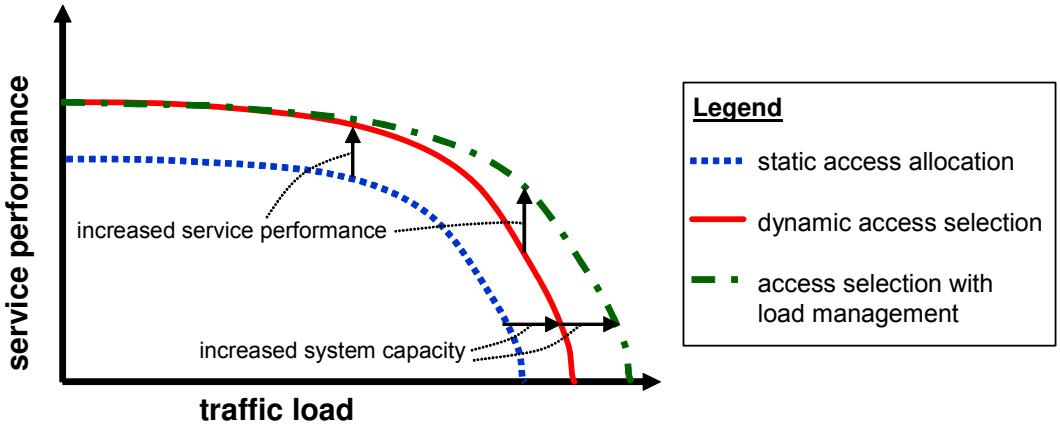


Figure 4.8: Benefits of dynamic access selection.

4.5.2 Taxonomy of Access Selection Gain

Next we want to analyse how, and in which situations a gain can be achieved from access selection. Firstly, we discuss the key radio transmission characteristics required to understand how access selection is used and present reference radio cell layouts of a multi-radio access system. Secondly, we categorise access selection gain into different types, and discuss to what extent they contribute to the overall gain, depending on the radio cell layouts and traffic load.

4.5.2.1 Performance and Resource Characteristics of Radio Transmission

For access selection it is required to quantify the performance of different access connections that are available for a UN. The performance of an access link is determined mainly by three aspects (see Annex B). Firstly, it depends on the quality of the received radio signal, which is largely determined by the location of the UN with respect to the location of the radio access point and the resulting radio propagation path loss. Secondly, it depends on the characteristics of the RAT like channel bandwidth, modulation and coding scheme. Thirdly, it depends on the load in the radio cell and the portion of resources available for the UN.

4.5.2.1.1 Spatial Distribution of Capacity and Resource Requirements

The radio propagation characteristics and the link quality dependent channel capacity lead to an inhomogeneous spatial cell capacity distribution (see Annex B). Figure 4.9 depicts this behaviour: a user close to the radio access point perceives a large radio link capacity; a user at the cell edge perceives a small radio link capacity[30]. It can be clearly seen that the capacity depends on the distance d of a user network from the radio access point and the resulting SINR. Note, that this is a pivotal property for evaluating the access selection gain, which has been omitted in related work by Koo et al. [KFZK04] and Tölli et al. [THH02]. We define the *capacity unit c* as the amount of capacity that is required for a service to be fulfilled at a certain QoS level. For example, this could be a minimum required data rate.

[30] For simplicity, we assume that the SINR decreases for increasing distance d from the radio access point. This assumption is not generally true due to local shadowing effects and irregular cell patterns; however, we ignore shadowing in order to simplify the discussion.

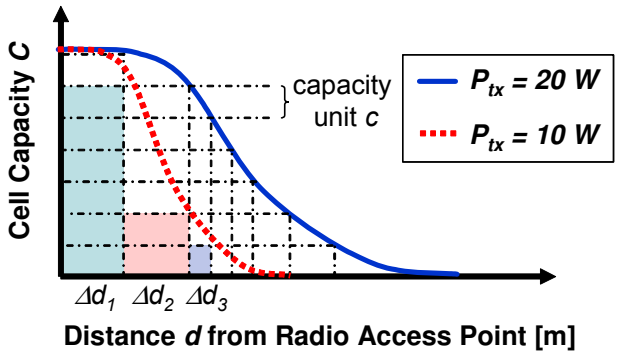


Figure 4.9: Cell capacity distribution within the cell.

Also the amounts of radio resources that are needed to provide a capacity unit *c* depend on the distance of the users from the radio access point. The main radio resource shared in the downlink[31] of a radio access system is the available transmission power. In case of multiple active users within a radio cell, these users share the common transmission resources in a time-, frequency- or code-division multiple access fashion. In the example of Figure 4.9 the cell capacity is depicted in multiples of capacity units. Hereby the capacity is shown for two different power levels at the radio access point. It can be seen that a transmit power of 10 W provides a capacity of one capacity unit in the area Δd_3, a capacity of two capacity units in area Δd_2, or a capacity of six capacity units in the area Δd_1 respectively. Consequently, a service request in area Δd_3 is six times more costly in terms of required radio resources than the same service request in area Δd_1. A user in an unfavourable position requires more radio resources than a user in a better location. According to the example of Figure 4.9, with a certain transmit power (i.e. radio resource) six capacity units c (equivalent to a data rate) can be provided in area Δd_1, but only one capacity unit can be provided in Δd_3.

The total cell capacity depends on the distribution of these users within the cell. The highest total cell capacity can be achieved by preferring users in favourable conditions (link-quality dependent scheduling) at the disadvantage of those users in poor locations (see e.g. [TKSG04] for a HSPA scenario). This results in a reduced fairness. For a uniform spatial user distribution, the number of users located closer to the cell edge is larger than in the cell centre due to the larger area as depicted in Figure 4.10. It is not sufficient, to only consider the number of service requests that arrive within a certain radio cell, also the location of the service request is significant.

[31] The analysis of the uplink is similar to the downlink case. However, the resources are the available time slots and sub-carriers in TDMA, FDMA schemes and the interference head-room in CDMA schemes.

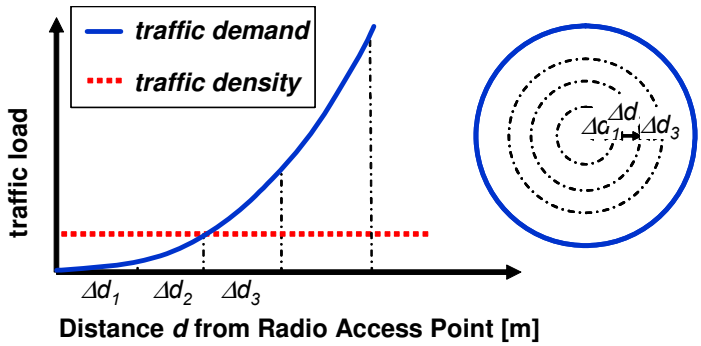


Figure 4.10: Spatial traffic load distribution for uniformly distributed traffic demand.

The usage of resources by different users which are uniformly distributed over the cell area is depicted in Figure 4.11. See Section B.3 in Annex B for a derivation of the capacity distribution and the required resources.

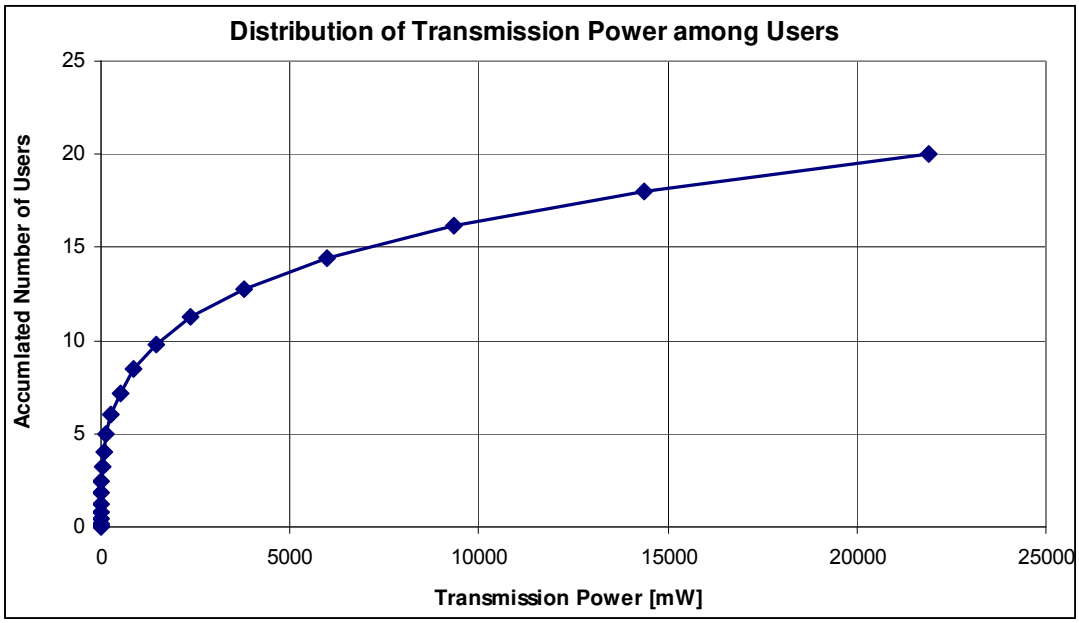


Figure 4.11: Distribution of transmission power among users within a radio cell (see Annex B).

4.5.2.1.2 Load-dependent link performance

The capacity of a radio channel in dependence of the radio link quality (cf. eq. (B.5)) describes the link capacity or performance, when resources are allocated to this radio link. In multi-user systems the system resources, e.g. the transmission power, are shared between the users. The radio link performance for every user thus depends on the system load. Two effects lead to a reduced link performance: *multi-user resource distribution* and *load-dependent interference*.

Multi-user resource distribution

In multi-user systems, the transmission resources are distributed among the different users according to a multiple access scheme. Multiple access schemes can be time, frequency or code division multiple access (TDMA, FDMA, CDMA respectively), where the users are separated in time, frequency or via different signal sequences (i.e. codes). Two types of resource allocation schemes can be distinguished: *centralised* and *distributed resource allocation*.

In *centralised resource allocation*, a single resource management entity distributes the resources to different users. A scheduling discipline governs the rules for how resources are distributed. With round-robin scheduling, resources are evenly shared between the users; other scheduling schemes can prioritise users according to a preference value (e.g. based on the subscription) or according to the service requirements. Also users with better radio link performance can be preferred (i.e. channel-dependent scheduling), which increases the efficiency of resource utilisation and leads to higher system capacity at the cost of reduced fairness between users. In general, centralised resource allocation achieves high efficiency of resource utilisation; for downlink transmission no scheduling overhead is introduced; for uplink transmission the overhead consists of signalling of resource requests from the users to the scheduler, and signalling of resource assignments back from the scheduler to the users. The load-dependent scaling of the link performance for a specific user depends on the proportion of resources that are available for this user. It thus depends on the number of users in the system, and the amount of resources they request and are assigned to. Let us assume an example system with centralised round-robin scheduler. As soon as the sum of service requests λ_n of all users N exceeds the system capacity according to the users link capacities C_n, the effective data rate R_i of a user i decreases with load according to:

$$R_i = \min\left\{ \lambda_i \;;\; C_i(SINR_i) \;;\; \frac{C_i(SINR_i)}{\sum_{n=1}^{N} \frac{\lambda_n}{C_n(SINR_n)}} \right\}. \qquad (4.9)$$

This scaling of link performance depending on load is depicted in Figure 4.12. The relative resource demand of a user n is λ_n/C_n. The cell capacity is reached when the sum of demanded resources reaches the available resources. When the total traffic demand is lower than the available capacity, every user request is satisfied and some capacity remains unused. The maximum data rate that can be assigned to a single user is then upper bounded by the users link capacity C_i – provided that the user has sufficient data to send (i.e. $\lambda_n > C_n$). When the total traffic demand surpasses the available capacity, the achievable link performance is scaled down according to the proportion that the demand exceeds the capacity. Centralised resource allocation is used in most cellular access technologies, such as GSM, (E)GPRS, UMTS, HSPA, and LTE. The 802.11 WLAN standard also defines a centralised resource allocation mode (i.e. point-coordination function, PCF); however, it is currently not used in deployed WLAN systems.

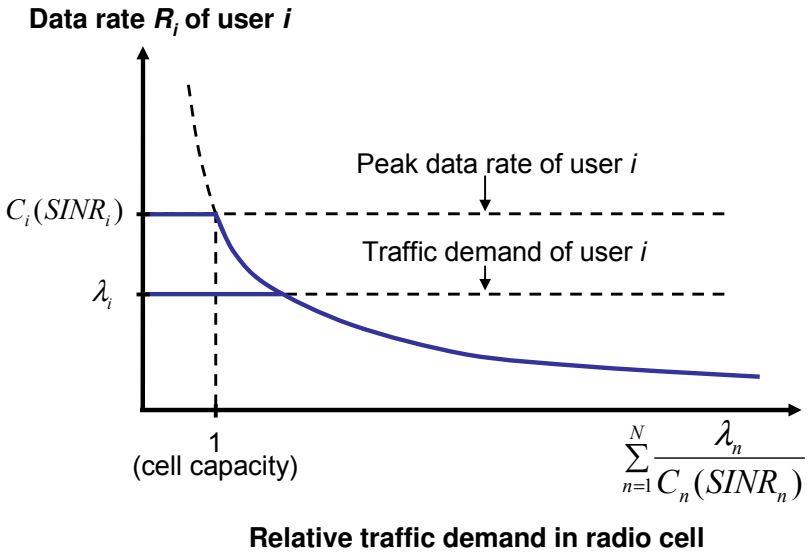


Figure 4.12: Load-dependent link performance for centralised round-robin scheduling.

In *decentralised resource allocation* the different users contend for channel resources in a probabilistic fashion. The most common decentralised resource allocation scheme in wireless communications is the distributed coordination function (DCF) of the IEEE 802.11 WLAN system. It is based on *carrier-sense multiple access with collision avoidance* (CSMA/CA) [WMB06]. Users sense the channel and only try to access the channel when it has been observed as idle for a certain time. In order to avoid collisions of simultaneous channel access by multiple users, every user waits for a random backoff time whenever the channel is idle before they try to access the channel. In case of collisions of simultaneous channel access, the backoff time is exponentially increased. As shown by Bianchi [Bia00] and Heusse et al. [HRBD03] the overhead added by channel contention increases with a larger number of users due to higher collision probabilities. This leads to a non-linear link performance reduction per user at increasing load. Heusse et al. [HRBD03] have derived the transmission time $T_i(N)$ of a data packet of user i with N active users in the cell as

$$T_i(N) = t_{tr}(C_i) + t_{ov} + t_{cont}(N), \tag{4.10}$$

with

$t_{tr}(C_i)$: transmission time of the packet depending on link capacity C_i,

t_{ov} : constant overhead per packet,

$t_{cont}(N)$: load dependent overhead due to contention of N users.

The constant overhead per packet t_{ov} comprises the transmission time of an acknowledgement, and the channel idle time for carrier sensing for the data packet and the acknowledgement. The contention overhead $t_{cont}(N)$ can be approximated in case of saturation[32] as

[32] Saturation means that the total traffic demand is larger than the cell capacity, so there are always users that want to transmit data.

$$t_{cont}(N) \cong t_{slot} \cdot \frac{CW_{\min}}{2} \cdot \frac{2 - \left(1 - 1/CW_{\min}\right)^{N-1}}{2 \cdot N} \quad (4.11)$$

with

t_{slot} : constant slot time of the contention window,

CW_{min} : minimum contention window size,

N : number of active users in the cell.

The achievable rate of a user *i* in saturation can thus be expressed as

$$R_i = \frac{C_i(SINR_i)}{N} \cdot \frac{t_{tr}(C_i)}{t_{tr}(C_i) + t_{ov} + t_{cont}(N)} \quad (4.12)$$

However, this achievable rate per user can only be obtained, if all users have the same radio link quality *SINR*. If users are distributed over the radio cell and thus have different radio link qualities another effect takes place, which is known as the *performance anomaly* of WLAN [HRBD03]: in WLAN resources are not equally shared between different users, instead different users have equal probability of access to the channel. As a result, users with lower link performance $C_i(SINR_i)$ occupy the channel for a proportionally longer time than users with higher link performance. Heusse et al. [HRBD03] have shown that the effective data rate of all users in the cell, is limited to the effective rate of the user with the lowest link performance. From that the effective rate of every user can be approximated as:

$$R_i = \frac{\min_{j \in N}\left(C_j(SINR_j)\right)}{N} \cdot \frac{t_{tr}\left(\min_{j \in N}\left(C_j(SINR_j)\right)\right)}{t_{tr}\left(\min_{j \in N}\left(C_j(SINR_j)\right)\right) + t_{ov} + t_{cont}(N)}, \quad (4.13)$$

with

$\min_{j \in N}\left(C_j(SINR_j)\right)$: link performance of user with lowest *SINR*,

$t_{tr}\left(\min_{j \in N}\left(C_j(SINR_j)\right)\right)$: transmission time of the user with the lowest *SINR*.

Consequently, the achievable data rate of one user does not depend on its own link performance, instead it depends on the performance of another user – the one with lowest link performance. The achievable data rate does therefore not only depend on the number of users in the cell, but also on where users are located within the radio cell and what radio link quality they experience. The contention overhead depends only on the number of active users. It increases non-linearly with the number of users per cell.

Load-dependent interference

An increase in system load has a further adverse effect on link performance. With a larger number of users also the mutual interference between the transmitted signals of different users increases. Interference can be distinguished into intra-cell interference between users of the

same cell, and inter-cell interference between users in different cells that use the same or overlapping frequency bands. Some radio technologies allow to eliminate intra-cell interference, e.g. by interference cancellation or sufficient guard intervals between different signals. In WLAN systems with small cell sizes, interference only plays a role if cells are closely located to each other or overlapping. An increase in interference leads to a decreased SINR. For very densely deployed WLAN systems inter-cell interference is perceived as a merging of multiple cells: users in both cells are contending for the same channel. As a result, the channel may not be sensed as idle due to transmission in the neighbouring cell. Furthermore, the collision probability increases as users in the neighbouring cells also access the same radio channel.

4.5.2.2 Example Multi-Radio Network Layouts

From the previous sections it has become evident, that the location of radio cells, the location of users and the resulting radio link quality plays a significant role for the capacity of multi-access systems. Therefore it is important to consider how multi-access systems are deployed. In heterogeneous access networks different radio access systems are deployed which are typically operated in different frequency bands. Thus, there is no inter-system interference between the radio access systems, and each radio access system can be planned and deployed independently. From a geographic perspective these different access systems form an overlay of radio cells; at many locations a mobile user network can access multiple access systems. For the investigation of system capacity and the performance of access selection algorithms, there are two key aspects to consider:

- the cell layout of the different radio access systems,
- the user distribution in the area, and thus the geographic distribution of traffic demand.

The cell layout determines how the radio capacity of the radio access systems is geographically distributed. We present three cell layouts for two radio access systems in Figure 4.13, which are later used as reference in the capacity evaluation. The first two cell layouts (a) and (b) assume that both radio access systems are wide area radio systems. In case (a) we consider a co-located cell layout for both radio access systems, where e.g. common base station sites are used. In case (b), denoted as non-co-located cell layout, we assume that the cell centres of one radio access system are at maximum distance between the cell centres of the other radio access system. These two cases are extremes; realistic cell plans will often be somewhere in-between. For the third case (c) we consider a wide-area radio access system combined with a local-area or hotspot radio access system. The cell size of the local area system is smaller than the wide area cell. We assume one hotspot cell per wide area cell, which can be located anywhere between the cell centre and the cell edge of the wide-area cell. For simplicity we only consider regular cell structures.

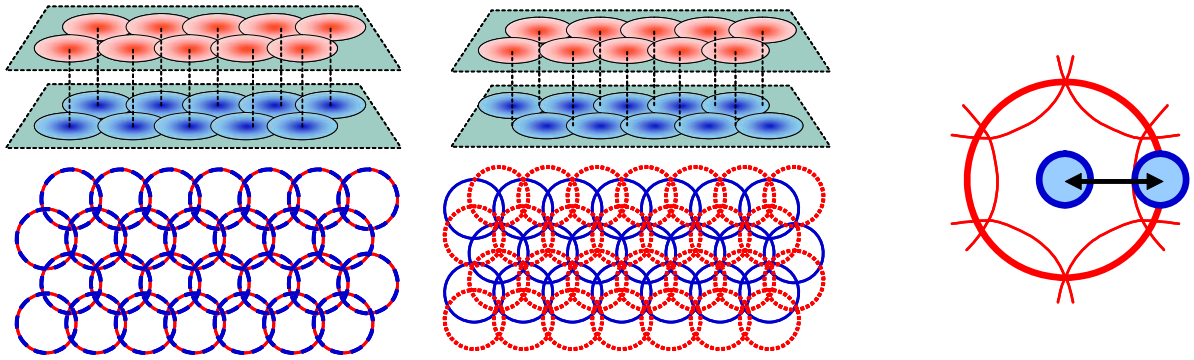


Figure 4.13: Example cell layouts for heterogeneous access networks.

The radio capacity at different locations for an overlay of two radio systems is depicted for co-located radio cells in Figure 4.14 and for non-co-located radio cells in Figure 4.15.

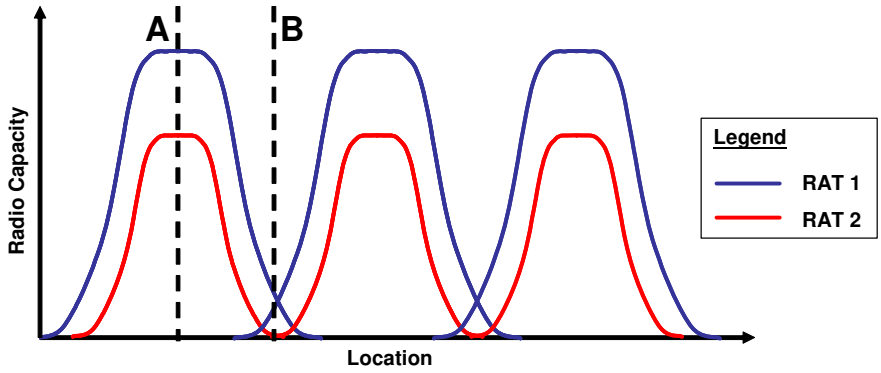


Figure 4.14: Capacity distribution for co-located RATs.

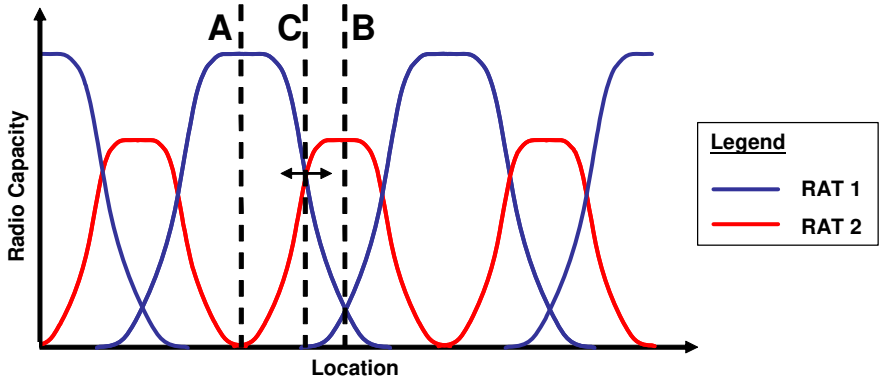


Figure 4.15: Capacity distribution for non-co-located RATs.

Users can be distributed arbitrarily over the system area. For the numeric evaluation we consider two different cases. The first case considers a homogenous distribution of users in the complete area. The second case assumes that hotspot areas exist where users are located with higher probability.

4.5.2.3 Types of Access Selection Gain

Flexible allocation of users to access systems can, on one hand, provide a gain for the end user in form of higher service quality, and on the other hand, provide a gain for the system in form of higher system capacity. In the following we discuss three different effects that can contribute to these gains.

4.5.2.3.1 Spatial Transmission Diversity

As discussed previously, the capacity provided by a radio system is unevenly distributed over the geographic area; it depends on the location of radio access points and the radio propagation characteristics as depicted in Figure 4.14 and Figure 4.15. As *spatial transmission diversity gain*, we denote the benefit that can be achieved by a user having the choice of access selection based on his location and the resulting difference in path loss (L_D and L_S in eq. (B.7) in Annex B) between the different radio systems. Ideally, the user network selects the radio access point or radio cell with the best radio link performance depending on path loss and RAT characteristics. The gain from spatial diversity is large in geographic locations, where the difference in performance of the different radio cells is large; it is small if the difference in performance is low. Spatial transmission diversity is also referred to as macro-diversity.

From Figure 4.14 we can see that the spatial diversity gain is not so large in case of co-located access systems; one access provides high capacity at the same locations as the other access system. In the example of Figure 4.14 a user at both locations A and B would choose RAT 1, which provides a better performance than RAT 2. At location A a better performance is perceived compared to location B. Only when the load in RAT 1 would become high and load management is applied, users would be allocated to RAT 2.

Conversely, for non-co-located access systems as depicted in Figure 4.15, one access system can provide high capacity at locations where the other access system has low capacity. A user at location A would profit from the good performance of RAT 1, and a user at location B would profit from the good performance of RAT 2. The overall gain depends also on the geographic distribution of traffic demand. If access selection considers the load in individual radio cells, it is possible to adapt the allocation regions flexibly; in Figure 4.15 this means that the location C, where the transition of users being allocated to RAT 1 or RAT 2 takes place, can be moved closer to either direction A or B, depending on the load in these regions. This allows further increasing the system capacity, and reducing the amount of unused capacity in areas with low traffic load.

4.5.2.3.2 Stochastic Transmission Diversity

The instantaneous channel quality of a radio signal can fluctuate by up to 30 dB in SINR due to multi-path propagation (L_M in eq. (B.7) in Annex B), as shown in Figure 4.16. These fast channel variations are statistically independent for different users and for different access systems. The statistic independence implies that the probability is lower that both signals have unfavourable values at the same moment than the probability of each random variable by itself having an unfavourable value. For fast fading an unfavourable value corresponds to destructive interference of the different multi-path components of the signal leading to large channel attenuation. In radio communication systems, the fast fading variation is independent

for different radio technologies, even if both radio links are provided by the same base station. The reason is that the radio propagation characteristics (cf. Section B.2 in Annex B) depend on the carrier frequency of the different signals. Consequently, a gain – denoted as *stochastic transmission diversity gain* – can be achieved when the transmission can be dynamically switched between the different access systems to the one with better channel conditions. This type of diversity is also referred to as micro-diversity. Theoretical capacity gains in the order of 40% have been shown by Koudouridis, Karimi et al. [KKD05] [KDKK06] for two RATs by dynamically selecting the RAT with the better link quality.

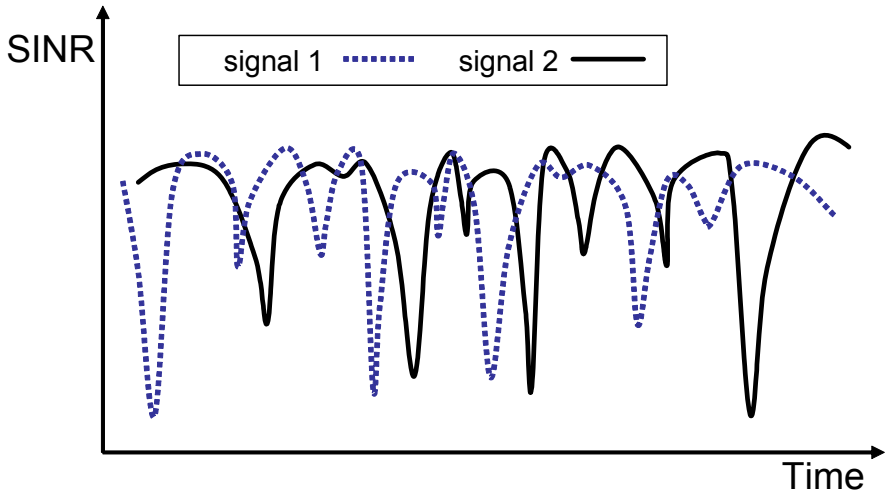


Figure 4.16: Two signals with independent fast fading.

It has to be considered that the channel variations are in the order of milliseconds. Consequently, very fast access selection is required to provide this gain. It has been shown [AN R2-8] that delays in channel measurements and signalling can drastically reduce the gain and therefore a tightly integrated multi-access network architecture is required. Furthermore, since the multipath component L_M is only an additive term in the total path loss, in many locations the radio link performance is dominated by the distance dependent attenuation L_D and the shadowing component L_S (cf. eq. (B.7)), as shown in Figure 4.17. Therefore, a gain of stochastic transmission diversity is limited to certain regions in the system as indicated in Figure 4.18. In all other locations – where one access system significantly outperforms the others despite independent channel variations – no stochastic diversity gain can be achieved.

The gain of stochastic transmission diversity also depends on the load in the system. Most access systems deploy multi-user scheduling where channel resources are only allocated to users with good channel properties. This can be seen from Figure 4.16, if we assume that the different signals belong to different users within the same radio cell. This approach is referred to as *multi-user diversity*. Therefore, at medium to high system load channel variations are already exploited in each access system independently by multi-user diversity; little additional stochastic transmission diversity gain is achievable by multi-access allocation [KKD05] [KAABB+05].

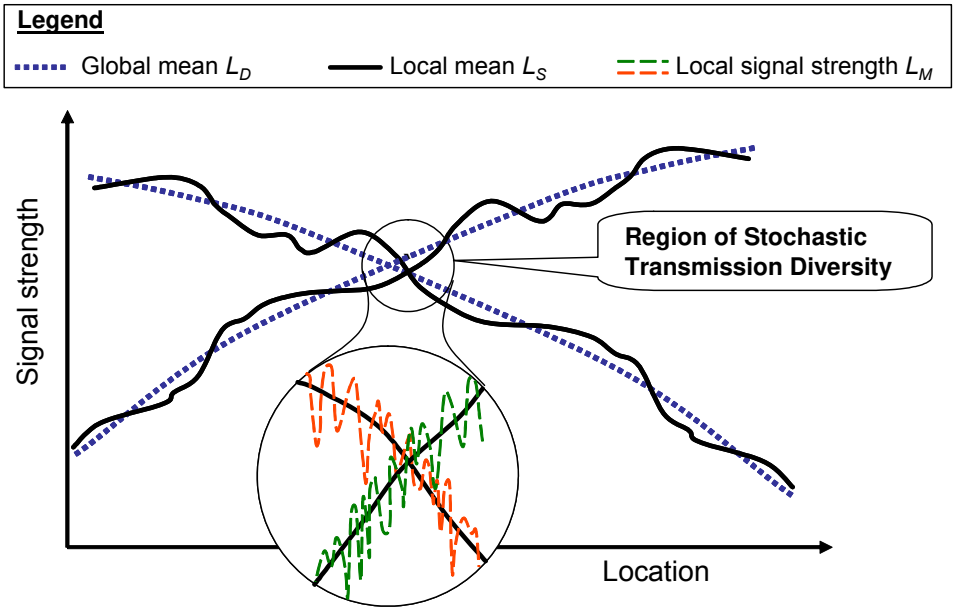


Figure 4.17: Region of stochastic transmission diversity.

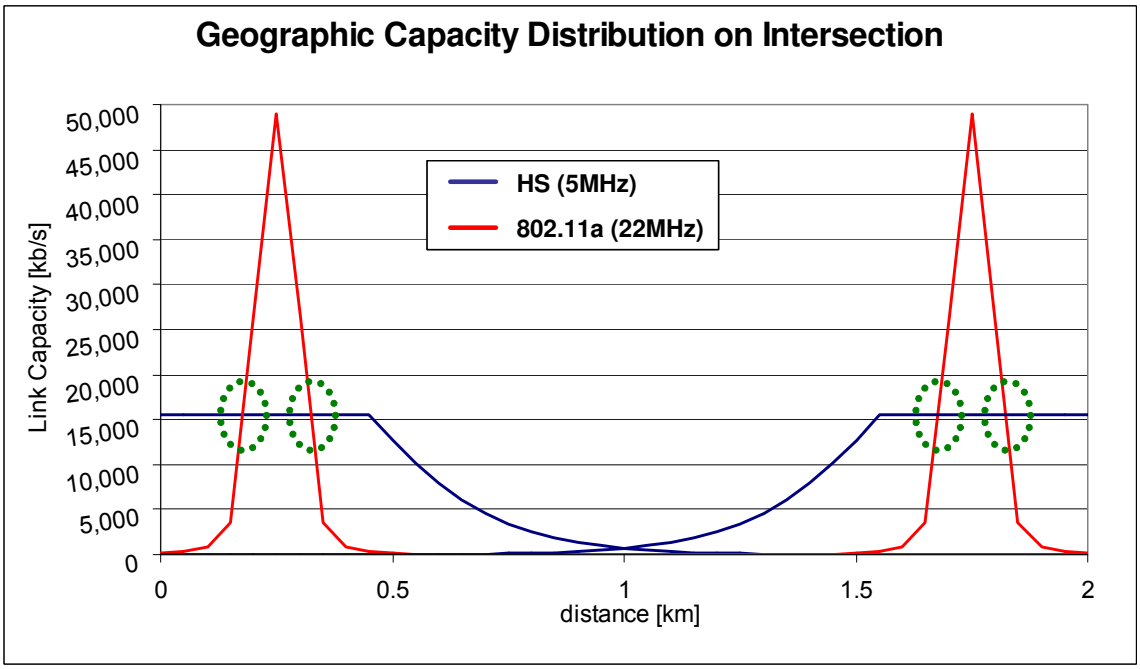


Figure 4.18: Regions of stochastic transmission diversity are marked with dotted circles for an overlay of idealised HS and 802.11a.

4.5.2.3.3 Trunking Gain

According to teletraffic theory a gain can be achieved if resources of multiple systems are combined and stochastic service requests can be served out of the combined resource pool. Figure 4.19 shows a static allocation of users to access systems without access selection as a single-server model. User service requests are either a-priori allocated to access system A or access system B. A resource manager (or server) in each access system distributes the

resources of the system among the service requests. The joint allocation of system resources A and B among all users, as represented in the single-server model of Figure 4.20, corresponds to load-based dynamic access selection. When different service requests arrive with independent stochastic behaviour, the joined allocation of the combined resources results in a larger system capacity. A larger total number of service requests can be served at the same quality of service level, e.g. blocking probability or mean queuing delay, compared to the service requests being dedicated to one system only. This is the basic concept of trunking and it has been largely investigated within teletraffic theory. The achievable gain is called *trunking gain* or *statistical multiplexing gain*. The gain stems from the reduced probability that the total number of active service requests exhaust the available resources in combined resource pools.

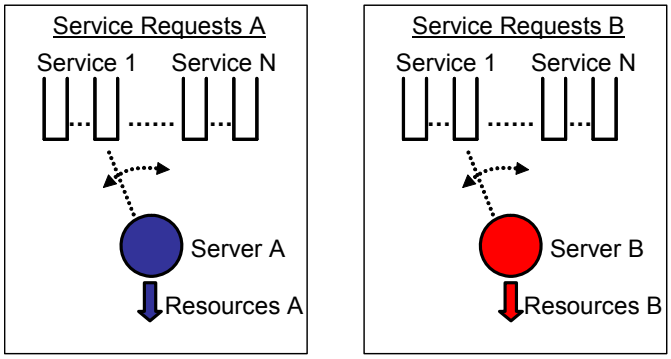


Figure 4.19: Static allocation of user service requests to either access system A or access system B.

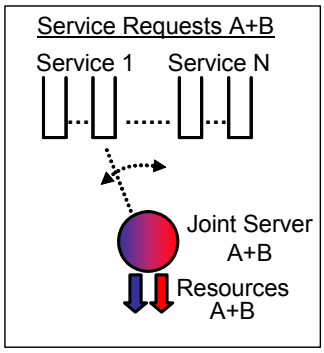


Figure 4.20: Allocation of user service requests to either of the access system A and B.

The trunking gain for the single-server model can be analytically derived by the Erlang formula. A single-server multi-access model has been used by Tölli et al. [THH02] and Koo et al. [KFZK04] to study the trunking gain in a multi-radio access system. However, the single-server model is insufficient to derive reliable quantitative results of trunking gain in a multi-radio access system. The reason is that radio characteristics and the resulting geographic capacity distribution are not considered in this model; it is rather assumed that the cell capacity is uniformly distributed over the cell area as shown in Figure 4.21. Thus one key property of radio systems is missing, which is that the cell capacity is different if service requests arrive in the cell centre or at the cell edge. Furthermore, the model does not distinguish where in the radio cell service requests occur. As we have seen in Section 4.5.2.1 (cf. Figure 4.10), typically more users arrive at the cell edge, where they require more

resources and where less capacity is available. The inhomogeneous availability of capacity can be represented by multiple servers for different regions of a cell with different capacity, as shown in Figure 4.22. However, these different servers in a cell are coupled; depending on the resource allocation of the server of one cell region, more or less resources can be assigned by servers in other cell regions. This makes an analytical evaluation difficult. Even if it does not allow to provide easily quantitative results for the trunking gain, it helps at this point to understand the trunking gain qualitatively. Later in Section 4.6 we present a new model based on stochastic knapsacks, which allows the investigation of the cell capacity with non-uniform capacity distribution. This new model overcomes the limitation of the single-server model, or the model of coupled servers.

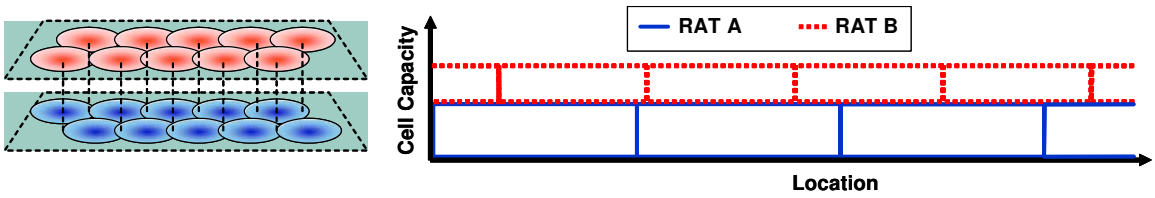


Figure 4.21: Single-server multi-radio cell capacity model with uniform distribution of cell capacity and resource consumption.

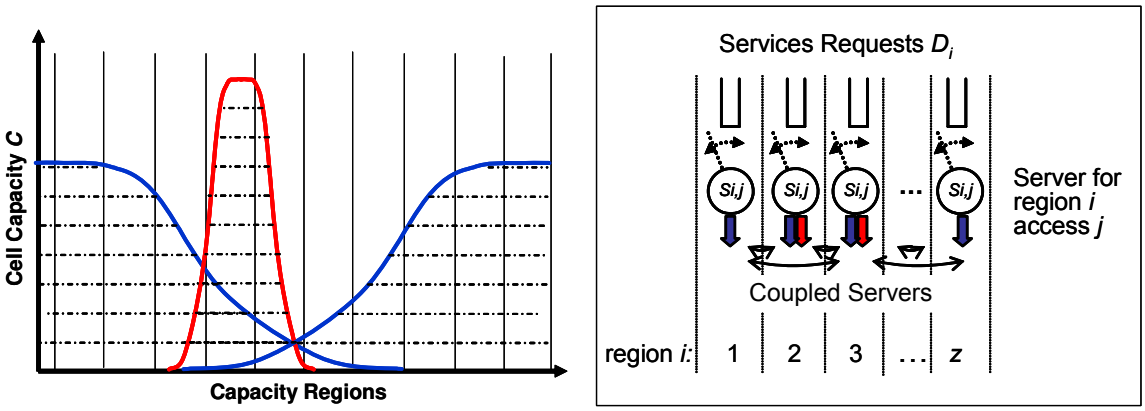


Figure 4.22: Multi-radio access model with multiple coupled servers.

In order to understand the influence of the trunking gain for different multi-radio cell layouts, we use the model of separate capacity servers for different cell regions, according to Figure 4.22. We resolve the coupling of the resources of different capacity regions by the restriction that service requests are only allowed to arrive in one region. Figure 4.23 shows this model for the co-located (a) and non-co-located (b) case, where we consider three capacity regions A, B and C. For co-located wide-area cells, the joint resource pool of the combined radio capacity of RAT 1 and RAT 2 in every region is significantly larger than the individual resource pools. Consequently, the trunking gain is large. For non-co-located wide area cells in regions A and C one RAT has a much larger capacity than the other RAT, so the joint resource pool is not significantly larger than the individual resource pools of the RAT with larger capacity. Therefore, the trunking gain is small. Only in the intermediate region B a trunking gain can be found. If we consider a wide-area system combined with a local-area system, the trunking gain is typically small for two reasons: firstly, a gain can only be found in a small fraction of the total area; secondly, at most locations one of the systems clearly outperforms the other one, so the joint resource pool is dominated by one system.

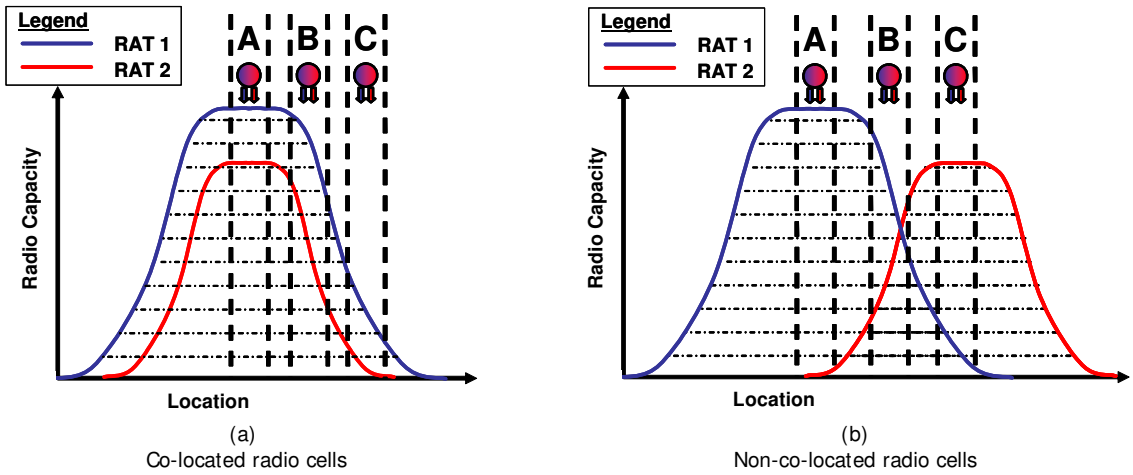


Figure 4.23: Model for the capacity in different cell regions – represented by separate servers A and B – for an overlay of co-located (a) and non-co-located radio cells.

4.5.2.4 Conclusion on Types of Access Selection Gain

The observations from the investigation of the different types of access selection gain is summarised in Table 4-1. One conclusion is that the dominating type of access selection gain in a multi-access system depends on the considered scenario, in particular the cell layout and the traffic load distribution. A second observation is that different types of gain do not simply add up to an overall gain; instead, we see rather one type of gain or another one. In cell layouts with a large trunking gain, the gain of spatial diversity is small and vice versa. The gain of stochastic diversity is only visible at low traffic load, when no multi-user diversity is exploited within the individual access systems.

The numeric evaluation in Section 4.5.3 is investigates the spatial transmission diversity gain and trunking gain.

Table 4-1 : Taxonomy of access selection gain.

Access Selection Gain	Co-located wide area cells (a)	Non-co-located wide area cells (b)	Wide area + hotspot (c)
Spatial Diversity	Low (without load management or at low load), Medium (with load management)	Medium (without load management), High (with load management)	High
Stochastic Diversity	Medium (at low load), Low (at high load)	Medium (at low load), Low (at high load)	Medium (at low load), Low (at high load)
Trunking Gain	High/medium	Low	Low

4.5.3 Numeric Evaluation of Access Selection Gain

In this section we investigate the gain of access selection in a multi-access system by means of simulations for a wide range of scenarios.

4.5.3.1 Objectives and Approach

The goal of this investigation is to study the downlink capacity and user perceived QoS in terms of data rate in a multi-radio access system for different scenarios. These studies extend prior work by Yilmaz et al. [YFPS05]. We consider two wide-area radio access technologies: the 3GPP High-Speed Downlink Packet Access (HS) [HT06] and a Future RAT (FRAT) with increased capacity, which could e.g. be based on 3GPP Long-Term Evolution [EFKMP+06], 3GPP evolved HS [3GPP25.913] or the WINNER radio interface [PHDSP+06]. As local-area radio access technology we regard IEEE 802.11a and 802.11b [IEEE802.11] [IEEE802.11a] [IEEE802.11b1].

We investigate two different access selection algorithms. The first algorithm, denoted as rate-based access selection (*AS(rate)*), selects for every user the RAT which provides the largest expected data rate *R* depending on the *SINR*:

$$\text{selected access } j = \underset{i \in \{RAT1, RAT2\}}{\arg\max} \{R^{(i)}(SINR^{(i)})\}, \tag{4.14}$$

The second algorithm, denoted as rate-and-load-based access selection (*AS(rate+load)*), in addition considers the load in the different RATs. The load is considered as the load-dependent scaling of the expected rate as discussed in Section 4.5.2.1.2. For every user the RAT is selected which provides the highest achievable data *R* at the given load after load scaling:

$$\text{selected access } j = \underset{i \in \{RAT1, RAT2\}}{\arg\max} \{R^{(i)}(SINR^{(i)}, load^{(i)})\}, \tag{4.15}$$

Access selection determines to what extent users are allocated to the different RATs at different load levels as shown in Figure 4.24. For the evaluation of the service performance we determine the distribution of user bit rate perceived by all users in the system. The distribution is given as a mean value and the 10- and 90-percentile (see Figure 4.25). Figure 4.24 and Figure 4.25 show an example with a combined FRAT and a HS radio access system; the access systems are non-co-located according to Figure 4.13 (b) and users are uniformly distributed over the system area. The total traffic load is given by the sum of the average traffic load per HS cell and the average traffic load per FRAT cell. For a total traffic load up to 8 Mb/s per combined FRAT and HS cell there is no significant difference between the rate-based (*AS(rate)*) and rate-and-load-based (*AS(rate+load)*) access selection scheme: the majority of users are allocated to FRAT which has the better link performance. Only few users – those located at the cell edge of FRAT and in the cell centre of HS – are allocated to HS. When the total traffic load increases further, a difference between the two access selection algorithms becomes noticeable. The rate-based scheme continues to allocate the majority of users to FRAT, which has the larger nominal link performance for most users. However, due to the load-dependent behaviour as described in 4.5.2.1.2, the effective bit rate provided by FRAT to a user is significantly lower than the nominal rate, and it is often even lower than the effective bit rate of the lightly loaded HS cells. If this is the case, the rate-and-load-based algorithm allocates the user to HS instead of FRAT. This can be seen by the increasing number of users allocated to HS at high load in Figure 4.24. As a result a higher user bit rate is achieved at high load. The rate-based access selection algorithm shows the gain

of spatial transmission diversity. In addition, the rate-and-load-based access selection shows also a trunking gain, since users can be allocated to either of the access systems (i.e. the trunked resources) depending on the resource availability within these access systems.

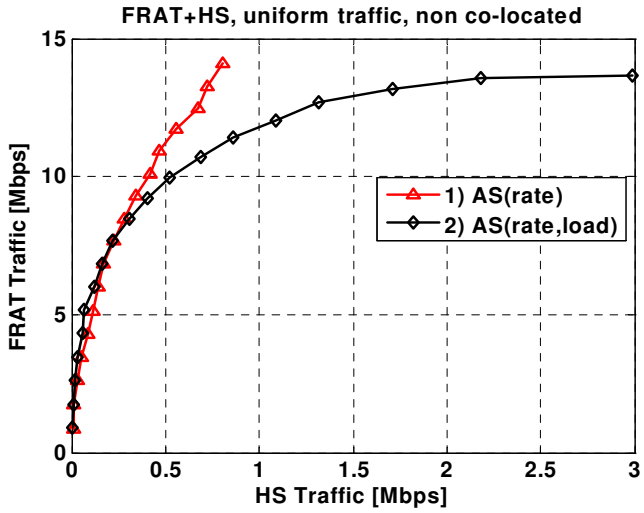


Figure 4.24: Allocation of users to FRAT and HS for increasing traffic load.

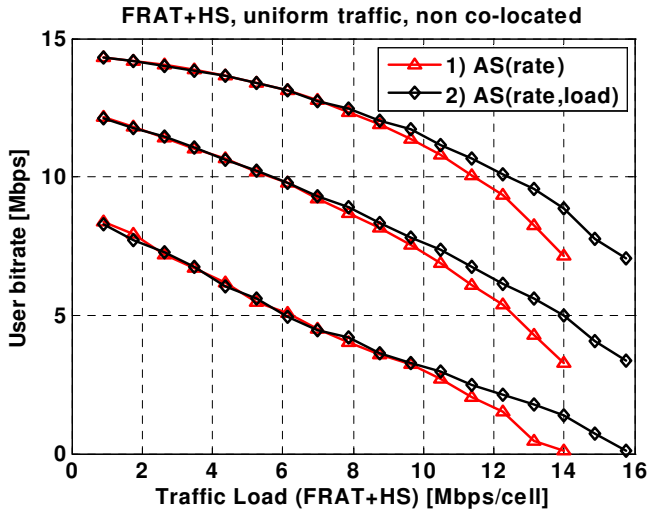


Figure 4.25: User bit rate distribution for non-co-located HS and FRAT (mean, 10- and 90-percentile).

Since we investigate a large number of scenarios, we limit the results in the following solely to the capacity of the multi-radio access system. We define the capacity of the system as the number of users that can be supported such that 90% of the users have a user bit rate of at least 500 kb/s. The results quantify the spatial diversity gain and trunking gain of access selection. We limit our study to downlink transmission.

We focus on elastic services like file download. Furthermore, we assume that users remain stationary, i.e. user mobility is not considered. Our investigation is based on a static Monte-Carlo simulator of a multi-cell environment, which is developed in MATLAB®. Results are taken from 100 simulation runs with different seeds of the random number generators. We consider an overlay of two radio access system, with each radio access system having a

regular cell layout of 7 radio cells including wrap-around to avoid border effects. The cell plan has an inter-site distance of 866 m and the two radio access systems can be shifted with respect to each other, so that we can study the two reference cell layouts for co-located and non-co-located systems according to Figure 4.13 (a) and (b). By configuring one radio access systems as a local area radio access with smaller cell size, we derive an overlay of a wide-area and a local area system as depicted in Figure 4.13 (c). We define hotspot areas around the cell centres of one access system. With a certain hotspot probability users are assigned to a hotspot area. Hotspot users are located according to a two-dimensional Gaussian distribution around the centre such that 95% are located within 100 m radius from the hotspot centre. The remaining users are uniformly distributed over the simulation area. In the following we consider two cases. Firstly, all users are uniformly distributed; secondly, 90% of the users are allocated to a hotspot. For the interference calculations it is assumed that each user generates an average traffic load of 100 kb/s. For each user the *SINR* values for the different RATs are calculated and the corresponding bit rates are determined from *SINR*-to-rate mapping tables. The bit rate for each user is furthermore adapted by a load scaling factor, which represents the portion of resources allocated to the user. It depends on the RAT channel allocation method, e.g. linear scaling for centralised channel allocation and non-linear scaling for distributed channel access. In our investigation we consider three different types of RATs. HS uses a *SINR*-to-rate mapping table derived from link level simulations with a peak bit rate of 7.2 Mb/s. The maximum transmission power is 20 W with 5 MHz carrier bandwidth. For FRAT we assume a wider carrier bandwidth of 20 MHz with 80 W transmit power and use the HS *SINR*-to-rate mapping table scaled by a factor of 4. Furthermore, we neglect all intra-cell interference in the *SINR* calculation, as we assume either an OFDM multiple access scheme or a CDMA scheme with advanced receivers. WLAN IEEE 802.11 uses 22 MHz bandwidth with 1 W transmit power for 802.11a [ETSI893] and 100 mW for 802.11b [ETSI328]. In addition, we assume non-overlapping cells and thus ignore co-channel interference. For each WLAN user the nominal link rate and respective channel occupation time is determined based on the *SINR*; a channel occupation time is also assigned to the TCP acknowledgment packets in the uplink. For the multiple access scheme a channel contention overhead is assumed according to a geometrical distribution for the number of trials. The resulting data rate per user is determined according to the approach of Heusse et al. [HRBD03] (see also Section 4.5.2.1.2). For HS and FRAT an urban radio propagation model with path loss exponent of 3.52 and a shadow fading standard deviation of 8 dB is assumed. For WLAN we assume a line-of-sight propagation model up to 60 m and then a constant attenuation of 0.3 dB/m with 3 dB shadow fading standard deviation as in [YFPS05].

In the following sections we first investigate an overlay of two wide area radio access systems, and then consider the overlay of a wide-area and a local-area radio access system.

4.5.3.2 Overlay of Different Wide-Area Radio Access Systems

In total we investigate eight scenarios of overlaid wide-area radio access systems. The wide area systems FRAT can be combined with HS or another FRAT system, the traffic distribution can be uniform or hotspot-centred, and the cell layout can be co-located or non-co-located. For all these cases we compare rate-based access selection with rate-and-load-based access selection. Figure 4.26 shows the capacity for all the different scenarios. Obviously, the system capacity is always higher for two FRAT systems, compared to a combined FRAT and HS system due to the higher FRAT capacity. The increase is between 16% and a factor of 2.2 for the different scenarios.

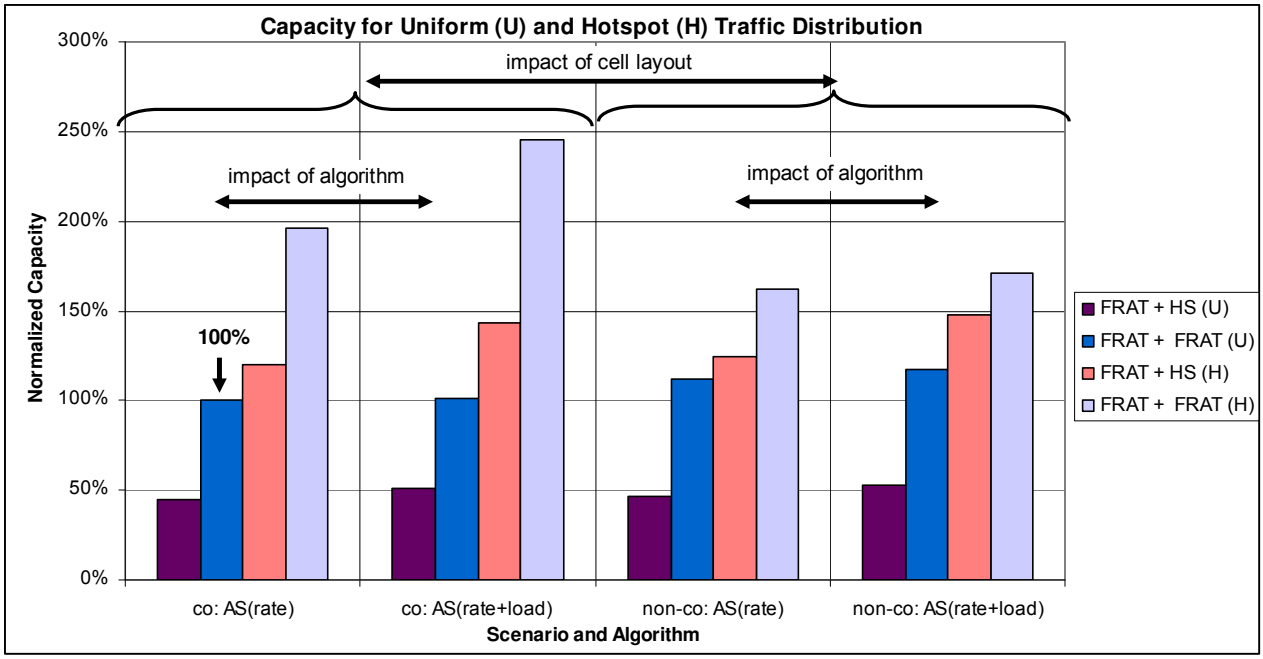


Figure 4.26: System capacity for two combined wide-area access systems, normalised to the capacity of two FRAT systems with co-located cells, uniform traffic load and rate-based access selection.

Our studies show that the system capacity increases for cell layouts where the geographic capacity distribution matches better to the geographic traffic load distribution. For a homogeneous traffic load distribution, a non-co-located cell layout increases the capacity compared to co-located cells as it will take advantage of the spatial diversity, i.e. one access system covers the "weak" areas at the cell edges of the other access system. For two combined FRAT access systems, this leads to a capacity gain of 12%-16%. For a FRAT access system combined with HS, the capacity gain is only at 2%-3% when changing from co-located to non-co-located cells, since HS contributes only to a small amount to the total capacity. When we investigate the case of a hotspot traffic distribution, we see a capacity increase by 17%-30% if two FRAT access systems are co-located compared to non-co-located cells. The reason is that now both access systems have their cell centre – and thus capacity peak – located at the hotspot. However, for a combined FRAT and HS access system the capacity decreases by 3%-4% in the same case, which means that the capacity is slightly larger when the HS cells are not located at the hotspot. A closer investigation of this scenario reveals that another effect takes place. Only users at the very cell edge of FRAT are allocated to the HS access system, thereby the FRAT system is offloaded from the most costly users (with highest propagation path loss). This enables FRAT to allocate more resources to the hotspot users which have good channel conditions. In total this adds more FRAT capacity to the hotspot than what the HS cells could have provided by also being located in the hotspot centre.

Another conclusion from the results is that the rate-and-load-based access selection algorithm always outperforms the rate-based algorithm. Depending on the scenario a capacity gain of up to 24% is achieved due to that considering load allows optimising the spatial diversity gain. Only in the case of two access systems of the same type with uniform traffic distribution there is hardly a gain when considering load. The reasons is that in this case both access systems

perform equally well, therefore terminals automatically perform load management by randomly choosing one of the access systems.

4.5.3.3 Overlay of Wide-Area and Local-Area Radio Access Systems

The system capacity for a combined wide-area and local-area radio access systems is presented for a HS system in combination with IEEE 802.11b in Figure 4.27. A FRAT system combined with either 802.11a or 802.11b is shown in Figure 4.28. As in the previous case, capacity gains can be achieved if the geographic capacity distribution matches the traffic load distribution. WLAN is characterised by having a high system capacity within a small area. Therefore, for a uniform traffic load distribution, a large part of the WLAN capacity remains unused. Only few users can profit from the WLAN capacity, and the total system capacity is not significantly increased. The capacity increase of adding a WLAN system for uniform traffic remains always below 11%.

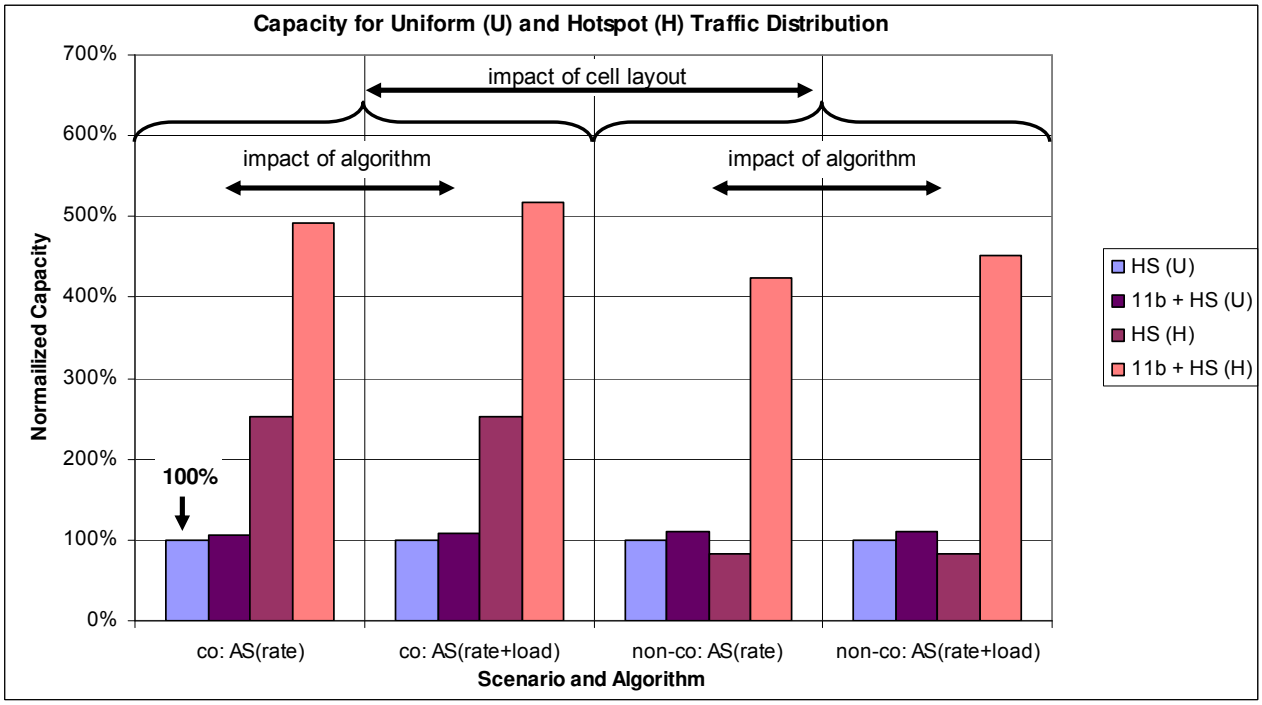


Figure 4.27: System capacity for combined wide-area HS and local-area 802.11b access systems normalised to capacity of the HS access system with uniform traffic load.

Another scenario is when the traffic is accumulated in a hotspot. If the hotspot is centred around the wide-area cell centres, the capacity of wide-area-only systems compared to uniform traffic increases by a factor of 2.5 for HS (see "HS(H)" in Figure 4.27) and a factor of 2.7 for FRAT (see "FRAT(H)" in Figure 4.28). If the hotspot is non-co-located, i.e. it is placed at the edge of the wide-area cells, the capacity of the wide-area-only system compared to uniform traffic decreases by approximately 20% for both HS and FRAT. With traffic mainly located at a hotspot, the additional capacity provided by WLAN cells becomes by far more significant. In this case the placement of WLAN cells at the hotspots significantly increases the total system capacity. For the case of HS in Figure 4.27, 802.11b cells at the hotpots increase the capacity (compared to the HS-only capacity with uniform traffic) in the range of 252% to 493% for co-located cells. An additional increase of 24% is achieved if also

rate-and-load-based access selection is used. This gain is if even larger, if the hotspot is not co-located with the wide-area cells. In this case the capacity increases in the range of 82% to 424% for rate-based and up to 452% for rate-and-load-based access selection. For the case of FRAT in Figure 4.28 the addition of a WLAN system does not increase the capacity if the hotspot is co-located with the FRAT cell and rate-based access selection is used. The reason is that the FRAT system provides better link performance for the users than the co-located WLAN system; consequently all users are assigned to the FRAT system. However, if rate-and-load-based access selection is used, the capacity increases in the range of 273% (compared to FRAT-only with uniform traffic) to 306% for 802.11b and up to 339% for 802.11a. The load management of the access selection algorithm allocates more users to the WLAN cells and thus off-loads the FRAT cell, so that the FRAT can serve more users outside the WLAN coverage area. If the hotspot is located at the edge of the FRAT cell, the capacity increases by adding a WLAN cell at the hotspot and using rate-based access selection by 77% for FRAT-only to 99% for 802.11b and 125% for 802.11a. Rate-and-load based access selection brings an additional gain with a resulting capacity of 127% for 802.11b and 146% for 802.11a.

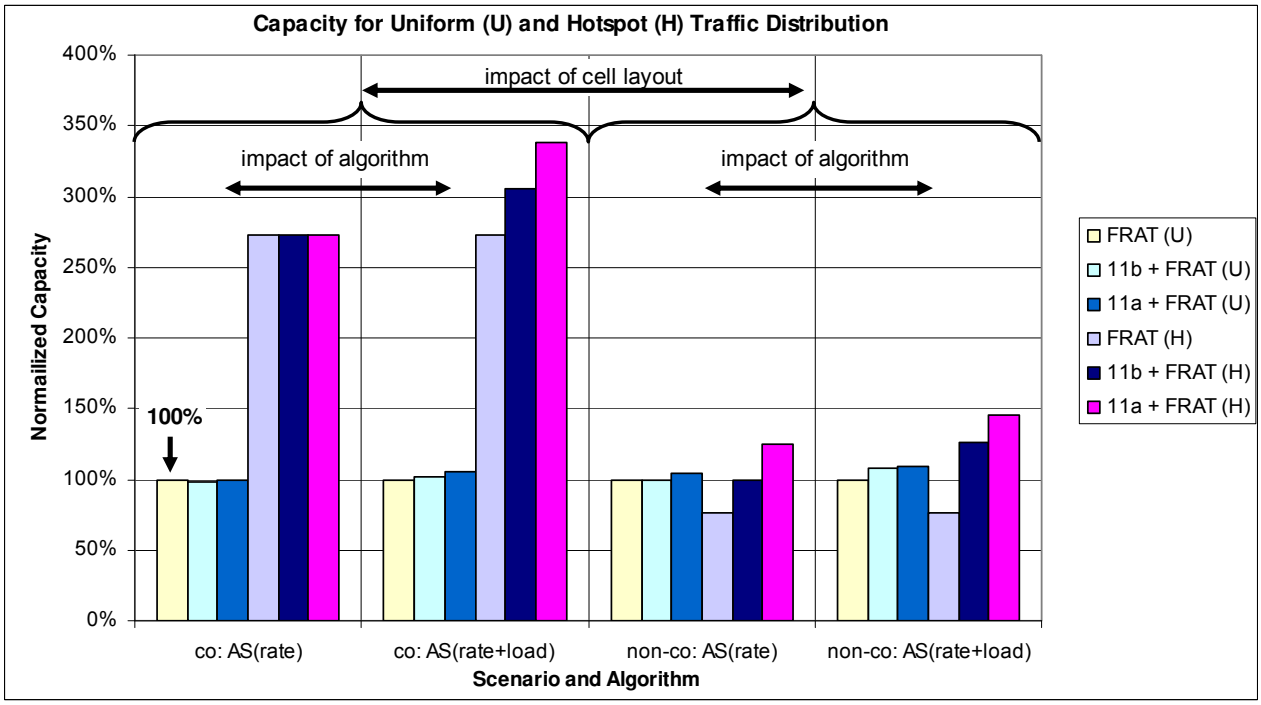


Figure 4.28: System capacity for combined wide-area FRAT and local-area 802.11a and 802.11b access systems normalised to capacity of the FRAT access system with uniform traffic load.

4.5.3.4 Conclusion

The capacity gain for an overlay of two different wide-area access systems and for an overlay of a wide-area and a local-area access system is evaluated for different cell deployment scenarios and traffic load distributions. Two access algorithms are investigated, one is based on radio link performance only, while the other one also considers the load in radio cells. We show that access selection in a multi-radio access system can significantly increase system capacity. The gain is largest when the cell deployment matches best to the traffic load distribution. For uniform distributed traffic, little gain can be achieved by adding a WLAN

cell to every cell of a wide-area access system. However, when the traffic is accumulated in a hotspot, and WLAN cells are located at these hotspots, a significant capacity gain can be found. Furthermore, the capacity can be increased by up to 27% if load in the radio cells is considered in the access selection algorithm.

4.6 Analytical Model for (Multi-)Radio Access Network Capacity Evaluation

4.6.1 Motivation

The capacity gain of access selection in multi-radio access networks has been investigated analytically by Koo et al. [KFZK04]. The model used by them represents the radio cell capacity as a single-server system, which is a common model for investigating fixed communication systems. As already described in Section 4.5.2.3.3, the single-server model of a multi-radio access system according to Figure 4.21 implies the following assumptions:

- Each radio cell provides a certain constant capacity over the whole cell area,
- Each service request has access to both radio access subsystem, so full radio coverage of the radio access subsystems is assumed,
- The service requests are uniformly distributed over the area.

As discussed in Section 4.5.2.3.3 and in Sections B.1, B.2 of Annex B, these assumptions are invalid for a radio cell. Consequently, the single-server model is very inaccurate and inappropriate for investigation of wireless system. It assumes that all users arriving in a cell require the same amount of resources to experience a certain service, independent from their location within the cell. The single-server model does not consider that available resources may be sufficient to serve user if they arrive at one location, while the same users may be blocked if they arrive at another location within the cell. It is also not suited to investigate situations when the geographic user arrival distribution is non-uniform, or when a non-regular radio cell layout is assumed. It is thus meaningless to compare the results obtained therein with our simulations. The single-server model has also been used in other multi-radio access capacity evaluation studies by Tölli et al. [THH02] and Fodor et al. [FFL04]. Therefore, we conclude that it is common to use the single-server model for wireless networks despite its inaccuracies. The objective of this section is to derive a new analytical model, which overcomes the limitations of the single-server model, and can be used for realistic analytical investigations of multi-radio access capacity.

4.6.2 Requirements and Approach

In order to meet the characteristics of radio transmission in a wireless network, a radio cell capacity model needs to include the following characteristics (see Sections B.1, B.2 in Annex B):

- The radio link quality and capacity increases when a user is located closer to the cell centre.

- A larger amount of radio resources is required to obtain a certain data rate when a user is further away from the radio cell centre.
- The total cell capacity depends on the distribution of users to different areas within the radio cells.
- Typically, there are more users located in areas with high resource consumption (i.e. at the cell edge) than in areas with low resource consumption (i.e. near the cell centre).

Our approach is to develop a model of a single-radio cell which takes all the above characteristics into consideration. It is desirable, to find a modelling approach based on well-understood models as used in teletraffic theory. In a second step, we extend the model so that multi-radio access networks with an overlay of different radio cells can be investigated. We also want to see how different access allocation strategies can be described in the model. Finally, we want to discuss for what scenarios the model is suited, and also pinpoint limitations of the model.

4.6.3 The Multi-Class Stochastic Knapsack

A stochastic knapsack is a multi-service loss model which generalises the Erlang loss system. As we show, it is well suited as basis to model the capacity of a radio cell. A stochastic knapsack is depicted in Figure 4.29. It is characterised by objects of K classes entering the knapsack of capacity C; elements of class k enter with a random arrival rate λ_k, occupy b_k resources and leave the knapsack after a random holding time with mean $1/\mu_k$ [Ros95]. There are n_k class-k objects in the knapsack. Then the total number of used resources R is:

$$R = \sum_{k=1}^{K} b_k \cdot n_k \leq C \qquad (4.16)$$

Stochastic knapsacks have been extensively studied and a profound description is provided by Ross [Ros95]. Several metrics of the stochastic knapsack can be calculated from its underlying Markov process for Poisson object arrivals. An extension for burst arrivals has been developed by Sarangan et al. [SGGA05]. These metrics include:

- the equilibrium distribution $\pi(n_1, n_2, \ldots, n_K)$ giving the probability of the knapsack being in any state $(n_1, n_2, \ldots, n_K)$ within the state space,
- the average throughput T of objects entering the stochastic knapsack,
- the utilisation U of the stochastic knapsack,
- the blocking probability B_k per service class for different admission control schemes, like peak rate admission, admission based on effective bandwidths or service differentiation.

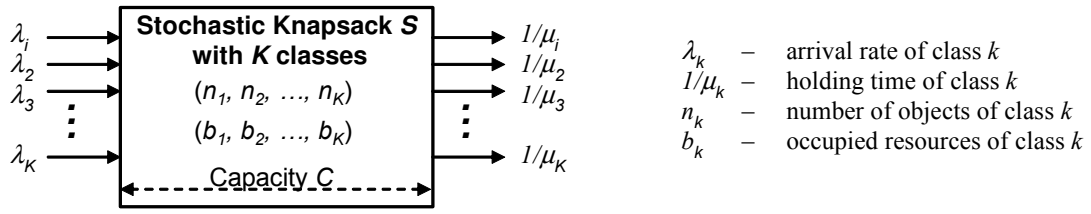


Figure 4.29: Stochastic Knapsack with *K* classes.

4.6.4 Modelling Capacity of a Radio Access Network

4.6.4.1 Radio Cell as a Stochastic Knapsack

We now show that the characteristics of a radio access network can be sufficiently modelled as a separate stochastic knapsack for every radio cell. The cell area is divided into area elements a_k which are characterised by similar radio link capacity and radio resource consumption. A size of an area element is denoted as s_k. The occupied resources b_k represent the radio resources that are required to provide a capacity unit c to the service request in that area element (cf. Figure 4.4). It depends on the path loss, noise and interference in a_k. With eqs. (B.9) and (B.5) in Annex B b_k can be derived as:

$$b_k := P_{Tx,k} = \left(2^{\frac{c}{B}} - 1\right) \cdot \Delta SINR \cdot L_P(d_k) \cdot \left(I(d_k) + N\right) \tag{4.17}$$

For constant interference, a path loss model according to eq. (B.8) – converted from dB scale to linear scale and applied to eq. (4.17) – and without shadow fading, a_k is a ring and b_k only depends on the distance d_k from the cell centre (cf. Figure 4.9 and Figure 4.10):

$$b_k = \left(2^{\frac{c}{B}} - 1\right) \cdot \Delta SINR \cdot \left(I + N\right) \cdot 10^{\alpha} \cdot {d_k}^{\beta} \tag{4.18}$$

The maximum range of coverage $d_{k,max}$ for a service with requirement c is reached when d_k is increased until b_k approaches C.

The holding time $1/\mu$ is the same for all area elements k, as we consider only one type of service. The arrival rate λ_k is proportional to the area size s_k and the user density distribution η_k in a_k. Let us assume again that a_k is a ring of width Δk. For a uniform user density distribution η with λ_0 as an arrival rate within a normalised area unit, λ_k depends on the size of the ring and becomes:

$$\lambda_k = \lambda_0 \cdot \eta \cdot \pi \cdot \left(2 \cdot d_k \cdot \Delta_k + {\Delta_k}^2\right) \tag{4.19}$$

4.6.4.2 Arbitrary Propagation Path Loss and User Traffic Distribution

The model in Section 4.6.4.1 is described bearing in mind a radio path loss that is equal for all directions. Furthermore, shadow fading is neglected. As a consequence, the radio cell is a

circle and the area elements a_k of equal capacity and resource consumption are rings. However, these assumptions are only made in order to provide a comprehensible illustration. For eq. (4.18) we have assumed a propagation behaviour according to eq. (B.8). Any other propagation formula can be applied to eq. (4.17). This model can just as well be applied for any arbitrary radio propagation scenario. For example, a map of path loss, interference and *SINR* could be used. In this case a_k would be the sum of all areas in which similar path loss and *SINR* is found. In this case b_k does not directly depend on the distance d_k. Instead the known path loss and *SINR* values would be directly applied in eq. (B.9) to determine b_k as:

$$b_k := P_{Tx,k} = \left(2^{\frac{c}{B}} - 1 \right) \cdot \Delta SINR \cdot L_P(d_k) \cdot \left(I(d_k) + N \right) \cdot SINR(d_k) \tag{4.20}$$

The model is also not limited to uniform traffic distribution. For an arbitrary traffic distribution, the arrival rate λ_k is determined according to the traffic arrival in all areas that are combined to a_k.

4.6.4.3 Multi-Service Traffic Requests

So far we have only considered a single type of service. The model can easily be extended to multiple services classes. Different services are characterised by requiring different capacity units c to fulfil the service requirement. Consequently, they occupy differing amount of resources b_k according to eq. (4.17). Therefore, for every area element a_k multiple classes need to be defined, which may have different traffic distribution, service arrival rate λ and holding time $1/\mu$. These parameters are determined as for the single service case described in the previous sections. If the radio cell is divided into L area elements and we have M service classes, then the number of classes K of the stochastic knapsack is $K=L \cdot M$.

4.6.5 Stochastic Knapsack Model for Multi-Radio Access Networks

4.6.5.1 Overlay of Radio Cells

So far we have modelled a single radio cell as a stochastic knapsack. In this section we extend this model to a multi-radio access network, where we have an overlay of radio cells. Figure 4.30 and Figure 4.31 show an overlay of different radio cells of two radio access systems RA1 and RA2. Note that the radio cells do not need to be co-located, but they can be arbitrarily located with respect to each other. Each of the radio cells can be described as a separate stochastic knapsack. For simplicity we limit the number of radio cells to two, without loss of generality. We extend the notation to identify the different knapsacks: $x^{(a)}$, $a \in \{1,2\}$ describes if a knapsack parameter x refers to either knapsack 1 or 2. In case that traffic arrives independently for each radio cell and no re-allocation of traffic to the other cell is possible, the system metrics for the combined radio cells are:

$$U^{(1 \cup 2)} = \frac{U^{(1)} \cdot C^{(1)} + U^{(2)} \cdot C^{(2)}}{C^{(1)} + C^{(2)}}, \tag{4.21}$$

$$T^{(1 \cup 2)} = T^{(1)} + T^{(2)}, \tag{4.22}$$

$$B_k^{(1\cup 2)} = \frac{B_k^{(1)} \cdot \lambda_k^{(1)} + B_k^{(2)} \cdot \lambda_k^{(2)}}{\lambda_k^{(1)} + \lambda_k^{(2)}} \quad . \tag{4.23}$$

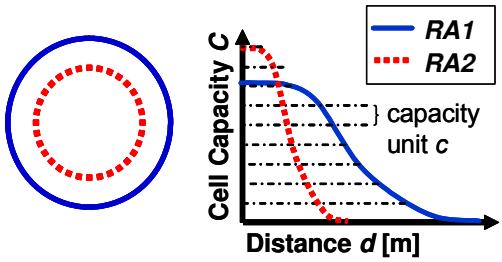


Figure 4.30: Overlay of co-located radio cells.

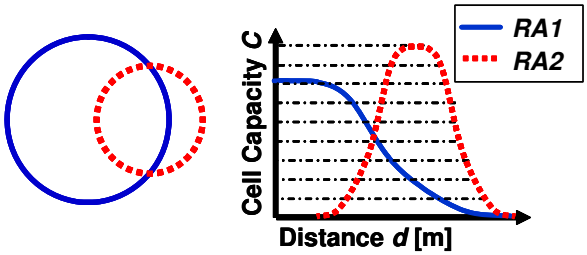


Figure 4.31: Overlay of arbitrarily located radio cells.

4.6.5.2 Multi-Radio Access Allocation

In a multi-radio access network an arriving traffic request in the overlapping area of the two radio cells can be freely allocated to either of the radio cells according to an access allocation strategy. Recall that the amount of required resources $b_k^{(\cdot)}$ to satisfy the request can differ significantly between the radio cells. Access selection introduces a coupling of the two stochastic knapsacks. This is problematic, since the analytical evaluation of a stochastic knapsack requires knowledge of the arrival rates λ_k. In this section we determine the arrival rates of the different knapsacks for a number of access allocation strategies and thereby decouple the knapsacks.

Let us first define for the radio cells new area elements $\tilde{a}_i^{(\cdot)}$, which are subsets of the area elements $a_i^{(\cdot)}$ and exclude the areas of joint coverage (see Figure 4.32):

$$\tilde{a}_k^{(1)} := a_k^{(1)} \setminus a_i^{(2)}, \text{ for all } i \text{ of } S^{(2)} \quad \text{and} \quad \tilde{a}_k^{(2)} := a_k^{(2)} \setminus a_i^{(1)}, \text{ for all } i \text{ of } S^{(1)} \tag{4.24}$$

We furthermore define the new joint area elements $\tilde{a}_{i,j}^{(1\cap 2)}$ as the overlapping part of area element $a_i^{(1)}$ and area element $a_j^{(2)}$.

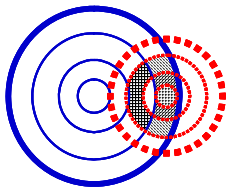


Figure 4.32: Two radio cells with non-overlapping areas $\tilde{a}_k^{(1)}$, $\tilde{a}_k^{(2)}$ (un-shaded) and overlapping areas $\tilde{a}_{i,j}^{(1\cap 2)}$ (shaded).

The goal of the access allocation strategy is to separate the arrivals $\lambda_{i,j}^{(1\cap 2)}$ in all joint area elements $\tilde{a}_{i,j}^{(1\cap 2)}$ into arrivals for the different knapsacks. With allocation probability $p_{i,j}^{(1)}$ the arrivals $\lambda_{i,j}^{(1\cap 2)}$ are allocated to $a_i^{(1)}$ and increase $\lambda_i^{(1)}$; with allocation probability $p_{i,j}^{(2)}$ the arrivals $\lambda_{i,j}^{(1\cap 2)}$ are allocated to $a_i^{(2)}$ and increase $\lambda_i^{(2)}$:

$$\lambda_i^{(1)} := p_{i,j}^{(1)} \cdot \lambda_{i,j}^{(1\cap 2)} \quad , \tag{4.25}$$

$$\lambda_i^{(2)} := p_{i,j}^{(2)} \cdot \lambda_{i,j}^{(1\cap 2)} \quad , \tag{4.26}$$

with $p_{i,j}^{(1)} + p_{i,j}^{(2)} = 1$.

4.6.5.3 Random or Priority Based Access Allocation

One type of access allocation is that an arrival is randomly allocated to one of the two radio cells according to constant allocation probabilities $p_{i,j}^{(1)}$, $p_{i,j}^{(2)}$. An example of random allocation is when users have different terminal capabilities which support only one radio access system.

A similar allocation strategy is when arrivals are allocated to a radio access system depending on priorities. These priorities are determined by user or network preferences and policies. Also in this case the allocation probabilities are constant.

4.6.5.4 Link Quality and Link Capacity Based Access Allocation

Some access selection strategies allocate all traffic requests in a joint area element to only one access based on a certain condition:

$$p_{i,j}^{(1)} = \begin{cases} 1, & \text{for } condition \quad (\rightarrow p_{i,j}^{(2)} = 0) \\ 0, & \text{else} \qquad\qquad\quad (\rightarrow p_{i,j}^{(2)} = 1) \end{cases} . \tag{4.27}$$

For an access allocation strategy that selects the radio access system which has the best link quality, the condition in eq. (4.27) is:

$$condition: \quad SINR_i^{(1)} > SINR_j^{(2)} \quad . \tag{4.28}$$

While the previous access allocation strategy considers radio link conditions, it does not consider different characteristics of the individual radio access system (see Figure B.7 and the relationship described in eq. (B.5)). A strategy which does this by considering the achievable link capacity looks like:

$$condition: \quad C_{link,i}^{(1)} > C_{link,j}^{(2)} . \tag{4.29}$$

4.6.5.5 Resource Cost Based Access Allocation

While the access allocation strategies in Section 4.2.2. aim at maximising the performance for a service request, another approach is to minimise the amount of resources that are required to serve the service request. In this case the condition in eq. (4.28) is:

$$condition: \quad b_i^{(1)} < b_j^{(2)} . \tag{4.30}$$

It shall be noted that the allocation strategies in eqs. (4.28)-(4.30) are related and mostly lead to the same result according to the relationship expressed by eq (4.17). In a real system they differ in what parameter is measured as input to the access allocation process.

4.6.5.6 Load Based Access Allocation

An important class of access allocation algorithms allocates user requests to a radio access system depending on the load (or utilisation) and capacity of the radio access systems, as e.g. discussed in [AN-M+06]. One algorithm allocates a service request to the access system with the largest number of available resources, that is: *maximise* $\{C^{(i)}\text{-}E[R^{(i)}]\}$, with $R^{(i)}$ as defined in eq. (4.16). Another algorithm allocates a service request to the access system with the best efficiency of resource usage, or lowest relative resource cost, which means *minimising* $\{ b_k^{(i)}/(C^{(i)}\text{-}E[R^{(i)}]) \}$. These algorithms can be approximated with the following allocation probabilities:

$$\text{max available resources}: \qquad p_{i,j}^{(1)} = \frac{Z^{(1)} / Z^{(2)}}{1 + Z^{(1)} / Z^{(2)}} \quad \text{and} \quad p_{i,j}^{(2)} = \frac{Z^{(2)} / Z^{(1)}}{1 + Z^{(2)} / Z^{(1)}}, \tag{4.31}$$

$$\text{min relative costs}: \qquad p_{i,j}^{(1)} = \frac{\frac{Z^{(1)} \cdot b_{i,j}^{(2)}}{Z^{(2)} \cdot b_{i,j}^{(1)}}}{1 + \frac{Z^{(1)} \cdot b_{i,j}^{(2)}}{Z^{(2)} \cdot b_{i,j}^{(1)}}} \quad \text{and} \quad p_{i,j}^{(2)} = \frac{\frac{Z^{(2)} \cdot b_{i,j}^{(1)}}{Z^{(1)} \cdot b_{i,j}^{(2)}}}{1 + \frac{Z^{(2)} \cdot b_{i,j}^{(1)}}{Z^{(1)} \cdot b_{i,j}^{(2)}}} . \tag{4.32}$$

where

$$Z^{(i)} = C^{(i)} - E\left[R^{(i)}\right] . \tag{4.33}$$

Load based access allocation algorithms have a general problem: they cannot be easily applied to the stochastic knapsack model, where utilisation and system state are determined in

equilibrium. The traffic allocation decision is based on the mean equilibrium utilisation of the two access systems, so the utilisation must be known prior to access allocation. On the other hand, the utilisation can only be determined for each system, when the arrival rate for each system is known, thus after access allocation. This problem can be solved iteratively: in each iteration an approximation of $E[R^{(i)}]$ is determined, from which $p_{i,j}^{(.)}$ for the next iteration is derived according to eq. (4.31) or eq. (4.32). The last iteration is reached, when the allocation probabilities $p_{i,j}^{(.)}$ converge, that is, the difference of $p_{i,j}^{(.)}$ in successive iterations is smaller than a pre-determined threshold. For the first iteration an initial value $\hat{p}_{i,j}^{(.)}$ of the allocation probabilities has to be determined such that the allocation probabilities converge in a few iteration steps. If the size of all joint area elements $\tilde{a}_{i,j}^{(1\cap 2)}$ is small compared to the complete system area, two options of appropriate initial values exist. Firstly, if the arrival rate in the overlapping area is small, it can be neglected:

$$\hat{p}_{i,j}^{(1)} = \hat{p}_{i,j}^{(2)} = 0, \qquad \text{for small } \lambda_{i,j}^{(1\cap 2)} . \tag{4.34}$$

Secondly, if the arrival rate in the overlapping area is large, it is duplicated:

$$\hat{p}_{i,j}^{(1)} = \hat{p}_{i,j}^{(2)} = 1, \qquad \text{for large } \lambda_{i,j}^{(1\cap 2)} . \tag{4.35}$$

If the size of all joint area elements $\tilde{a}_{i,j}^{(1\cap 2)}$ spans over a large portion of the complete system area, the initial allocation probabilities are best chosen according to the ratio of the capacities of the different access systems:

$$\hat{p}_{i,j}^{(1)} = \frac{C^{(1)}}{C^{(1)} + C^{(2)}} \qquad \text{and} \qquad \hat{p}_{i,j}^{(2)} = \frac{C^{(2)}}{C^{(1)} + C^{(2)}} . \tag{4.36}$$

4.6.6 Validity and Limitations of the Stochastic Knapsack Model

The multi-class stochastic knapsack is well suited to model two key aspects of radio access networks: firstly, the capacity is not uniformly distributed over the area, and secondly, the amount of radio resources required for a certain service depends on the radio path loss between sender and receiver, and thus on the location of a user within a radio cell. In order to determine the classes, the geographic area elements a_k with similar resource requirements $b_k \pm \varepsilon$ (thus similar path loss and *SINR*) need to be determined. The granularity ε largely determines the number of classes that are required and the precision of the model. For every class the arrival rate has to be determined based on the session arrival rate within the particular area element a_k. For the description of the classes the geographic path loss, *SINR* and traffic load distributions must be known. For an idealised omni-directional path loss and uniform user distribution they can be determined according to eqs. (4.17) and (4.19). But other arbitrary distributions can also be used, e.g. based on measurements or other propagation models. Determining these classes for different area elements may be a large computational effort.

The model can be used to consider different types of service classes. The number of classes for the knapsack is then the product of the number of area elements and the number of service classes. The model allows determining the capacity of a radio access network. As we have shown, it can also be used to determine the capacity of multi-radio access networks with an overlay of radio cells. The number of classes increases then since the combined area elements

in the overlapping region need to be determined. We have shown how different access allocation algorithms can be modelled. Analytical results can be obtained for Poisson-like traffic, and approximations exist for bursty arrival processes.

However, the presented model has also several limitations. Some of them are not easy to overcome. A comparison with a detailed simulation-based evaluation is desirable, which allows quantifying the errors introduced by these restrictions. The main limitations are:

- It is assumed that a service request always requires a fixed capacity unit *c*. But in real systems a large portion of traffic is elastic, meaning that a larger amount of capacity units is used when available. Therefore, elastic traffic can only be investigated approximately at high traffic load, when each service only obtains its minimum amount of resources *c*.

- The model assumes constant interference. In a realistic radio access system, the amount of interference within a radio cell, as well as between different radio cells increases with increasing traffic load. This cannot be easily modelled as it would dynamically change the area elements, and thus the classes and arrival rates of the knapsack.

- The model assumes stationary users within a radio cell. With user mobility a user would pass through different area elements. As a consequence, an ongoing service would jump between classes during its holding time in the knapsack. This is a major limitation for a realistic capacity evaluation. For mobile users a mobility margin needs to be included in the system capacity.

Some extensions to the model are also possible. So far it is assumed that a service request in a multi-access network is only served by a single radio cell. In some radio access systems, it is possible that a service requested is handled by multiple radio cells at the same time (so-called soft-handover). Also in a multi-radio access network it is possible to use multiple radio access systems in parallel. For this it is required that a traffic request is split into several sub-requests that are to be handled by different knapsacks. In this case also the gain of macro-diversity should be considered, which implies that the total resources required to be served by multiple radio cells is smaller compared to when the service is served by only a single radio cell.

4.6.7 Conclusion

In this section we have presented a new model of a radio access system based on a multi-service stochastic knapsack. It models the geographic distribution of radio link capacity and traffic distribution, which are key aspects for system capacity in a realistic radio access network. Based on the known properties of the stochastic knapsack, the system capacity can be evaluated, including equilibrium distribution for admitted users, throughput, utilisation, as well as blocking probabilities. We have also presented an extended model to evaluate the capacity in multi-radio access networks with an overlay of different radio cells. Different access allocation algorithms are described. We discuss for what scenarios the model is well suited, and also pinpoint limitations of the model. It is left for future work to evaluate the system capacity of reference scenarios numerically and compare them with capacity figures obtained from detailed radio access network simulations.

4.7 Summary

Access selection is a key function in a multi-access system architecture that comprises an overlay of different access technologies. In a multi-access system different roles can be distinguished, which lead to different business scenarios depending on how these roles are allocated to different business entities. We have developed a utility-based system model, in which the utility of an access allocation is defined for different system roles. This utility-based system model allows formulating access selection as an optimisation problem of a global system utility.

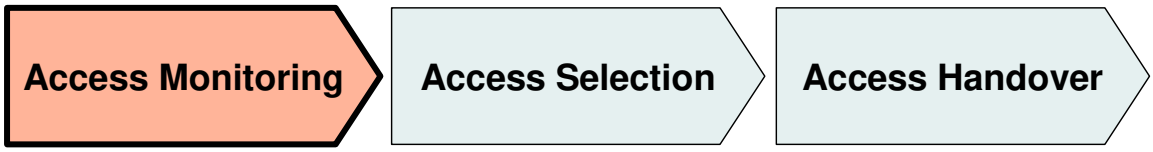
Several different access selection algorithms can be distinguished depending on their objective. Based on different dynamics of those objectives we proposed a two-stage approach of realising access selection. In a first step policy-based access selection is performed, which limits the possible access allocation options to those permitted by network and user preferences. In a second step, access selection is dynamically selecting the best suited access depending on the changing link quality and load for the remaining access allocation options.

It is important to understand what gain can be achieved by access selection in a multi-access system. In heterogeneous multi-radio access networks the network capacity is geographically distributed in a non-uniform manner. This distribution depends on the location of radio access points, the characteristics of the RATs and the radio propagation conditions. We classify access selection gain into three types: trunking gain, gain due to spatial transmission diversity, and gain due to stochastic transmission diversity. We argue that the total capacity gain of access selection depends largely on the network deployment and that the spatial transmission diversity gain and trunking are the most significant components. In a simulation environment we have investigated the capacity gain of access selection for an overlay of two access systems for different cell layouts and traffic load distributions. We considered an overlay of two wide-area wireless networks, as well as a combination of a wide-area and a local-area wireless network. The study shows that access selection can bring significant increase in capacity compared, in particular when load-based access selection is used. The gain is largest in cell layouts that match the traffic load distribution. WLAN systems can only provide little capacity gain when the traffic load is uniformly distributed over the system area. Conversely, if users are mainly centred around hotspot areas WLAN systems in such areas can largely increase the multi-access system capacity.

The analytical evaluation of access selection gain has been based on rudimentary radio cell capacity models in the past. These models neglect the geographic distribution of radio link capacity and traffic distribution, which are key aspects for system capacity in a realistic radio access network. We present a new model of a radio access system based on a multi-service stochastic knapsack. It allows evaluating the capacity in multi-radio access networks with an overlay of different radio cells for access allocation algorithms

Chapter 5. Access Monitoring

5.1 Introduction



An access network provides access and connectivity services to a user network. Each access network can comprise multiple access technologies, as depicted in Figure 5.1. Different networks can differ in the access and connectivity services that they provide. Before access selection can take place, the suitable access networks and access technologies for a user network need to be determined. Once an access network has been selected the access properties need to be monitored. We denote this process of discovering and monitoring access performance and access network characteristics as *access monitoring*.

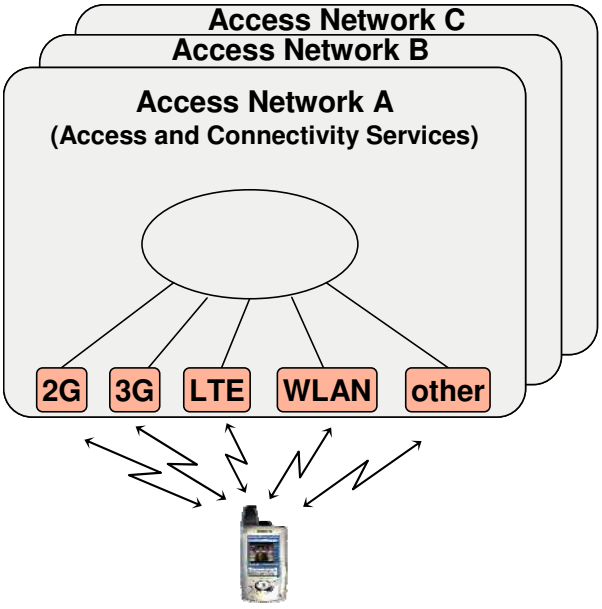


Figure 5.1: Access Networks with multiple access technologies.

In a practical scenario, there can be limitations for a user network concerning its capability to monitor or connect to multiple accesses at the same time. These limitations can stem from the implementation of the radio modems. For example, a device based on software defined/reconfigurable radio design [Tut99] has only a single configurable radio front-end; most access technology specific operations are realised by software. Such a device can only connect to a single radio access technology at a time. Before changing to another access, the radio front-end needs to be reconfigured to the new carrier frequency and carrier bandwidth, and the software modules must be reconfigured for the new access functions. The access functions include the coding and modulation scheme, multi-antenna configuration and algorithms, as well as radio protocol functions, like medium access control and scheduling,

segmentation and automatic repeat request, ciphering and header compression. As a consequence, the user network first needs to disconnect from one access before it can connect to the new access. Already for making measurements on other accesses, the radio front-end must be temporarily reconfigured. But also terminals with multiple separate implementations of radio modems face limitations. For example, due to interference between RATs in close frequency bands, it may not be possible to connect to two such RATs simultaneously. Also measurements in the terminal of one RAT can be hampered if data is transmitted simultaneously on another RAT in a close frequency band. Finally, even if no further restrictions on simultaneous usage of different RATs remain, simultaneous connectivity and RAT measurements require substantial battery resources; wise usage of measurements and connectivity via multiple RATs is required for battery-powered mobile devices.

For the management of different accesses, a common interface to the different access technologies is required. The generic link layer provides this interface, as depicted in Figure 5.2. The generic link layer provides an abstraction of access technology specific information, which is used by multi-radio resource management for access selection. It furthermore supports generic procedures for mobility management and network advertisement and discovery. In this chapter we focus on two functions. Firstly, we describe generic measures for the access performance and the resource usage characterisation of different access technologies[33]. Secondly, we derive how new accesses can be discovered and what information describes the capabilities of the access during the attachment process[34]. Next we evaluate the delay and signalling overhead for different connectivity setup, advertisement and attachment schemes in a WLAN scenario[35].

[33] The access performance and load abstractions have been jointly done with Per Magnusson, Mikael Prytz and Teemu Rinta-aho. These abstractions have been contributed to the *Ambient Networks* work package on *multi-access*, where they were further refined.

[34] Advertisement and discovery of network/access capabilities and connectivity setup and attachment have been largely developed in cooperation with the *Ambient Networks* task force on *network attachment* based on earlier work of the *Ambient Networks security theme*. In particular Göran Selander and Teemu Rinta-aho are to be mentioned who contributed to this work in several discussions.

[35] The performance evaluation has been performed together with Anh Tu Tran and it extends [Tra07].

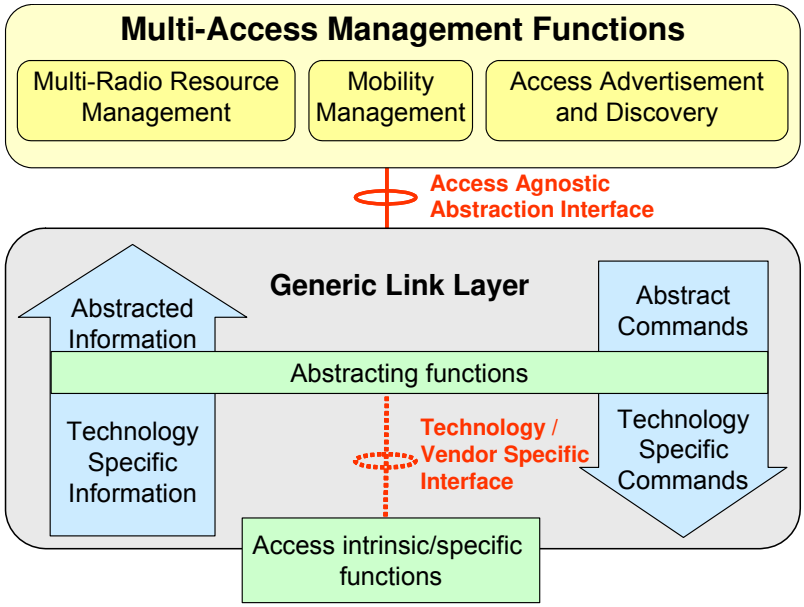


Figure 5.2: Generic abstractions of access specific functions and characteristics to support multi-access management.

5.2 Related Work

An abstraction of link performance that is applicable for different transmission technologies has been defined in relation to IETF protocols at several occasions. The objective is, for example, to accelerate higher layer mobility management or provide indications to transport protocols [FG00a] [FKG01] [APFPG03] [AGP04] [ID-L2ABST] [IEEE802.21] [RFC4907]. The link performance is basically abstracted to the two states "link up" and "link down", indicating whether a link is useful (up) or unuseful (down). Access technology specific threshold of the link quality can be defined to define the link up/down states. As indicated in [RFC4907] the lack of clear definition of these states leaves ambiguity. IEEE 802.21 extends the abstraction to include two new states "link going down" and "link going up", which indicate that a state transition between "up" and "down" is expected to happen [IEEE802.21]. In cellular networks link performance descriptors remain access-specific. In order to support inter-system handover, specific translations tables are used between equivalent link performance measures of different technologies [3GPP25.304]. Load performance reports are used for abstracting the load in 2G and 3G cellular systems as a relative load measure between 0%-100% [3GPP25.881]. However, as indicated in [3GPP25.881] this load measure is insufficient for decent load balancing due to the different capacities of different access systems. We develop in this work new generic abstractions of link performance and load levels in access systems, which allow a better description and comparison of access system and therefore can be used for more sophisticated access selection algorithms.

The problem of determining network capabilities has been described in IETF [RFC5113] [IETF-EAP] and in IEEE 802.11u [IEEE802.11u1] [IEEE802.11u2]. IEEE 802.21 [IEEE802.21] proposes a media independent information service which is similar to our approach of advertisement. In 3GPP specifications, advertisements are already defined and included in *system information blocks* [3GPP25.331] [3GPP24.008]. These can indicate other cooperating (roaming) networks, and neighbour list information about other access

technologies. In this work we discuss extensively what access and network information may be relevant for access selection; furthermore, we present and evaluate different options of how such information can be retrieved.

5.3 Generic Access Abstraction

For access selection, the suitability of each access for a given service data flow is evaluated to determine the best possible choice. The requirements of the service are defined in a *service specification*. For a multi-radio resource management function it is desired that it does not need to understand access technology specific characteristics. This requires that quantitative measures are available that allow to compare the characteristics of the different access technologies. The measurement values typically used in different access technologies are not comparable; therefore, generic descriptors are required. Two types of information describe the suitability of an access: the performance provided by that access (i.e. *access performance abstraction*) and the availability and usage of resources (i.e. *access resource abstraction*).

5.3.1 Service Specification

It is necessary that the access selection function knows the service requirements, and that this service specification is sufficient for the access selection to judge the suitability of an access system. A service specification, as depicted in Figure 5.3, contains the service requirements as discussed in Section 4.4.2.1.1 separately for downlink and uplink direction. In addition the service specification contains the *service type* and a *service priority*. The *service type* classifies the service according to its requirements in terms of reliability, delay and rate. Alternative service types could be the UMTS QoS classes as defined in [3GPP23.107]. In some cases an access provider restricts the usage of access resources for certain service types. For example, the amount of resources that may be used for best-effort data services may be limited to 70% of the available resources while the remaining resources are exclusively reserved for conversational services like telephony. The service type is used in order to determine to what extent an access system is useable for the particular service. Moreover, a priority value describes the priority level of a user service to obtain access to the transmission resources. This priority level is typically determined in the service level agreement between the end user and the access provider. For example, some users may subscribe to a cheap best-effort telephony service with low priority, while other users subscribe to a more costly premium telephony service that guarantees preferred service. Service priority levels could, for example, be classified as “bronze”, “silver” or “gold.” The service priority is part of the service specification; it allows a resource management function to determine whether the requirements of a particular service can be met at a given traffic load and traffic mix.

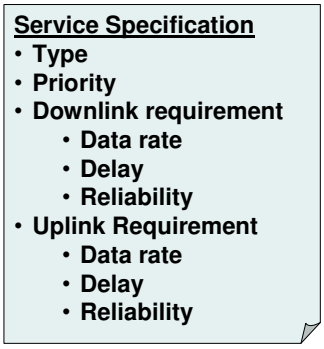


Figure 5.3: Service specification.

5.3.2 Generic Access Performance Abstraction

A typical way to characterise the performance of an access, is to describe the link quality of the radio link. Radio link quality measurements are already used within each access technology for handover or cell selection. However, these descriptors are insufficient, as they are access-specific and differ between different accesses technologies. It is instead necessary, to abstract the radio link quality into a radio link performance, which describes the access link capabilities with respect to the same parameters in which the service requirements are expressed. Furthermore, the link performance abstraction needs to include other performance measures required by a service. The performance requirements of a service are related to the following criteria:

- Error-Sensitivity: The sensitivity of an application towards residual bit errors or packet errors.
- Delay-Sensitivity: The requirements of an application on transmission delay and variation.
- Rate-Sensitivity: The requirements of an application on obtained data rate.
- Connection reliability: The sensitivity of an application towards connectivity interruption or disconnection.

The abstract link information is derived by the GLL and it is provided to multi-radio resource management for access selection. Suitable performance abstraction values matching the service requirements are listed below (see Figure 5.4):

- Link rate (instantaneous value, expected average, minimum, maximum, expected variation)
- Delay (minimum, maximum, expected average, expected variation)
- Residual Bit Error Rate (BER) (minimum, maximum, expected average, expected variation)
- Residual Packet Error Rate (PER) (minimum, maximum, expected average, expected variation)
- Connection reliability (handover degradation factor and grade of coverage)

The next sections describe how these parameters depend on the access technology characteristics.

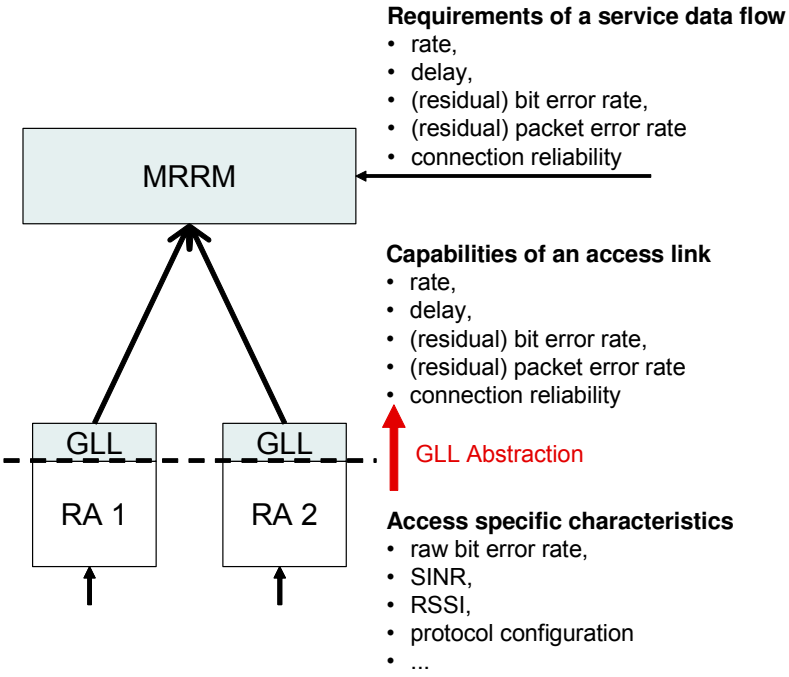


Figure 5.4: Link performance abstraction.

5.3.2.1 Transmission Reliability

Applications have a certain requirement on the reliability of the data transmission. This requirement describes the amount of bit errors and packet errors, which the application can tolerate. In general, a data stream can be corrupted when being transmitted. In particular in wireless transmission channels it is common that transmission errors occur. Link layer technologies can apply a number of methods to reduce the amount of transmission errors that are introduced:

- *Error detection* is a method to determine if transmission errors have occurred. The most common way of error detection is to compute a *cyclic redundancy check* (CRC) or a *checksum* which is added to the transmitted datagram. Then a receiver can also compute the CRC for the received datagram and compare it with the original CRC. If no discrepancy is detected, the likelihood of the datagram being corrupted is smaller than a certain residual bit error probability, which depends on the length and type of the CRC code.
- *Forward error correction* (FEC) is a method to encode a datagram in a way that a certain amount of bit errors can be corrected at the decoder of the receiver. It depends on the type of FEC code, how many bit errors can be corrected.
- *Backward error correction*, or *automatic repeat request* (ARQ), is a method to correct errors by (partly) retransmitting corrupted datagrams. ARQ can be configured with different levels of persistence which determines the level of reliability. A fully-reliable ARQ procedure achieves error-free transmission up to the level at which errors become undetectable[36].

[36] This depends on how the error detection sensitivity is configured. For most practical error detection methods it can be assumed that all errors are detectable.

Typically a link layer performs all three operations, error detection and forward and/or backward error correction[37]. The residual bit error and packet error rate depends on the configuration of the transmission and protocol parameters.
The abstracted description of transmission reliability of an access are the residual bit error and the residual packet error probabilities. In case that these error probabilities are configurable, the range of configurable error probabilities can be provided in the abstraction.

5.3.2.2 Connection Reliability

Transmission errors can also be introduced by handovers. When a user network changes the access resource that it connects to, data in-flight or stored in buffers can be lost. The connection reliability can be characterised by the amount of loss introduced by handover:

- *Handover cost* depends on how a handover is performed. The amount of loss depends on the length of the handover interruption. This depends on the fact whether *break-before-make* (BBM) or *make-before-break* (MBB) handover is supported. It furthermore depends on what kind of handover optimisation scheme is used. Handover can be more or less lossless by bi-casting or context transfer (see Chapter 6).
- *Handover probability* denotes how frequently a handover is expected to happen. It depends on the average cell size (which depends on the access technology), and the user mobility. The average cell size can be determined by the access network provider. The handover probability can be derived for some reference user velocities.

These two terms can be combined into a general *handover degradation factor*, which is the product of handover probability and a normalised handover cost value.
In addition, in a mobile environment there is a risk that the connectivity is dropped in situations when the user network moves into a coverage hole.

- *Grade of Coverage* denotes the reliability of an access system to provide coverage within a certain area. The probability to be affected by coverage holes increases with higher user mobility. The grade of coverage can be determined by an access network depending on the cell layout. It can be described for different reference area sizes.

The connection reliability can be abstracted by the grade of coverage and the handover degradation factor.

5.3.2.3 Transmission Delay

The transmission delay of an access system depends on the network structure and the radio access characteristics. The network structure, i.e. topology of network nodes and type of links connecting these nodes, is deployed by a network operator according to his network planning objectives. Delays resulting from the network structure remain fairly constant and can be determined by operation and maintenance functions. The transmission delay of the radio link depends mainly on the radio technology. It depends, for example, on the radio transmission time intervals, according to which data is scheduled, and, the interleaving depth. The block error rate together with the link layer round-trip time impact the transmission delay; they determine to what extent data transmission is delayed by link layer retransmissions (see e.g. [PM01]). The transmission delay is determined by the QoS principles of the access system. In an access system with multiple users, the transmission delay depends on the amount of resources that is allocated to each user. QoS-enhanced access systems group service data flows into QoS classes, for which different reliability on transmission delay can be provided,

[37] ARQ can also be combined with FEC in a common procedure, which is called hybrid-ARQ (HARQ).

as described in [LEWL06]. With admission control, a maximum transmission delay can be guaranteed for real-time service data flows, whereas for best-effort data flows the delay is theoretically unbound. Therefore, for an access system the abstract values for transmission delay are given for the supported QoS classes.

5.3.2.4 Transmission Rate

The transmission rate of the access system depends on the radio link quality, the carrier bandwidth and the radio transmission characteristics of the access technology. The radio link quality is determined by different types of measures for different access technologies, like for example received signal strength indicator (RSSI), a channel quality indicator (CQI) or signal-to-noise-and-interference ratio (SINR). Within each specific access technology, the radio link quality can be translated into an expected data rate. Hereby vendor specific implementations also need to be considered. For example, a device with an advanced radio receiver can achieve a higher data rate at the same signal strength than a device with simple receiver. In the simplest form a data rate could be derived from the modified Shannon formula in eq. (B.5) with access- and implementation-specific setting of the parameters $\Delta SINR$ and R_{max}. Rate variations depend on the load in the system and the QoS support. In a radio system with QoS-support, the rate variation depends on the QoS classes. For example, for a QoS-class with guaranteed minimum bit rate the data rate has a lower bound, which is not the case for a best-effort QoS-class. Some RATs use admission control to limit the amount of service data flows that are allowed to be served by a resource, thereby ensuring that a minimum rate can be maintained for some or all QoS-classes.

5.3.3 Generic Access Resource Abstraction

Different RATs have widely different mechanisms for using and sharing their available resources among users. For example, resources can be divided into time slots, frequency sub-carriers or codes; they can be allocated in a deterministic fashion or statistically, using contention-based schemes. The notions of the amount of total resource occupied (the load level), and the amount of resources that a particular user session occupies differs in different RATs. For MRRM to exploit radio resource information from heterogeneous RATs in its operation (e.g. for load management), it is necessary to have a mechanism that can derive comparable, relevant measures.

Some notion of the scope of a particular resource is necessary. For wireless accesses the geographical coverage is of particular interest. For the resource abstraction the resources in the multi-access system are divided into Access Resource Areas (ARA), which can differ for different RATs (see Figure 5.5). Typically an ARA corresponds to a cell area, but it could also be a smaller or larger unit. For one RAT in the multi-access system the resources in each cell area could be visible to MRRM, whereas for another RAT MRRM may only see the aggregated resources over multiple cell areas. In the latter case the aggregate of cell areas is seen as one ARA. Possibly MRRM can see the resources in individual cell areas even if they are managed by access-specific RRM. MRRM is assumed to be aware of all ARAs for all RATs in the multi-access system. MRRM needs to know which ARAs a user network is connected to. This is achieved via abstracted link quality/link performance reports (cf. Section 5.3.2), which indicate to which ARA a user network is connected or can connect.

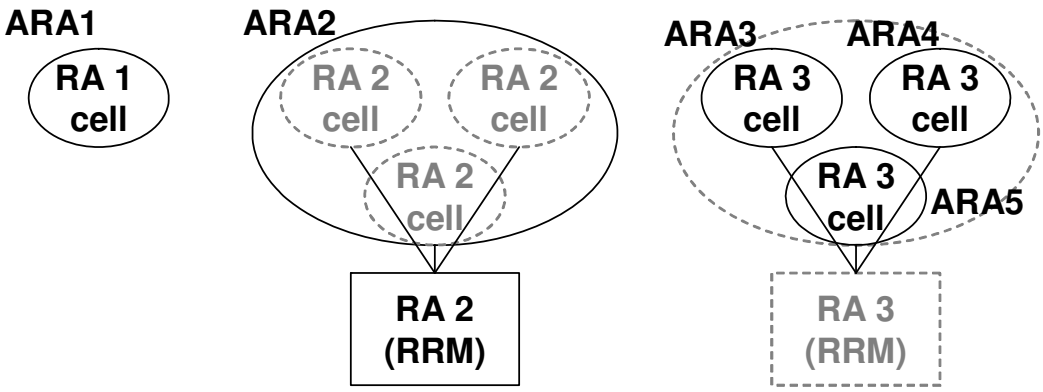


Figure 5.5: Different types of access resource areas (ARA) for access resources of different radio accesses (RA).

5.3.3.1 Generic Access Resource Metrics

The generic link layer abstracts and computes resource levels in each ARA according to a relative resource level. Depending on what is the limiting resource(s) in the RAT, it can be computed, for example, as the relative number of (time) slots or codes, the relative amount of (downlink) power, the relative occupied bandwidth, the average collision ratio; also combinations of these, such as power and slots, can be used to determine the relative resource level. The resource abstraction is performed by GLL on-demand from MRRM, for example during an admission control process or to get up-to-date information prior to a load management decision. The computation can also be performed continuously by GLL or when RAT-specific events occur; in this case the values could optionally be communicated to MRRM only when certain thresholds have been reached in order to reduce signalling load.
The following resource levels are computed per ARA, as depicted in Figure 5.6:

- r_{min} is the current minimum required relative amount of resources for all active service data flows served in the ARA.

- r_{occ} is the current occupied relative amount of resources for all active service data flows in the ARA. This can be larger than r_{min} whenever extra non-guaranteed resources are provided for service data flows that can benefit from it (i.e. elastic flows). It is assumed that all of the extra resources (r_{occ}-r_{min}) can be reclaimed, either instantaneously or after some delay, and used for other users.

- r_{max} is the current maximum relative amount of resources that can be used in the ARA. The "headroom" (1-r_{max}) is the margin required to cope with changing resource usage of active users in the ARA due to, for example, user mobility. If the time to free up extra, non-guaranteed resources is very small or zero, r_{occ} resources could be allowed to grow beyond the r_{max} limit as these can be freed up to give room for increasing resource usage of active users.

The following resource levels are derived from the above:

- $\delta_{avail} = r_{max} - r_{min}$ is the relative amount of currently available resources, including resources that can be (instantaneously) freed up if they are currently used to provide extra quality for some (or all) users.

- $\delta_{free} = r_{max} - r_{occ}$ is the relative amount of currently free resources.

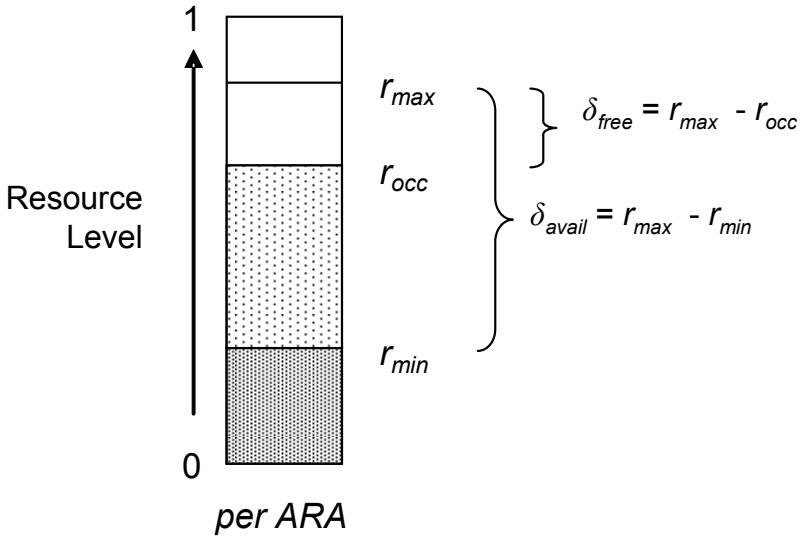


Figure 5.6: Relative resource levels in each *access resource area* (ARA).

In addition to the general resource levels, further (abstracted) resource values are derived. Consider an MRRM multi-access control decision (access selection, admission control, load management) for a service data flow *i* requesting quality $Q_{i,req}$ in an area covered by at least one ARA. $Q_{i,req}$ is defined by a service specification (cf. 5.3.1), which contains the service requirements on minimum bit rate and maximum delay and error tolerance, as well as how much additional quality (typically extra bit rate) is beneficial for the service. For each relevant ARA, MRRM transmits $Q_{i,req}$ to the associated GLL and requests a RAT-specific mapping of $Q_{i,req}$ to generic resource measures. The GLL retrieves resource information from the underlying RAT-specific entities, computes the resource levels r_{min}, r_{occ}, and r_{max}, and finally computes the following generic MRRM resource measures per service data flow:

- $Q_{i,offered}$ is the offered quality such as offered bit rate, maximum delay and additional quality (typically extra bit rate). This is typically only reported back to MRRM when $Q_{i,offered}$ differs from $Q_{i,req}$.

- $q_{i,min}$ is the relative RAT-specific (instantaneous) resource usage if service data flow *i* was "served" in the ARA with minimum quality requirements. For example, if there are 10 slots in the ARA and one slot is required then $q_{i,min}=0,1$. Note that if $q_{i,min} > 1$ then it is not possible to meet the minimum quality requirements in the ARA because the minimum amount of requested resources exceeds the available resources and the request is typically rejected.

- $q_{i,extra}$ is similar to $q_{i,min}$, but where service data flow *i* is given some extra quality. So $q_{i,extra} \geq q_{i,min}$ as it contains additional spare resources.

- $\sigma_{i,min} = q_{i,min} / \delta_{avail}$ is the relative resource efficiency of serving service data flow *i* in the ARA with minimum quality requirements. For example, if $\delta_{avail} = 0.4$ and $q_{i,min} = 0,1$ then $\sigma_{i,min} = 0,25$. This indicates that the service request requires 25% of the remaining capacity

δ_{avail}. Note that if $\sigma_{i,min} > 1$ then the amount of resources requested exceeds the available resources and such a request is typically rejected.

- $\sigma_{i,extra} = q_{i,extra} / \delta_{occ}$ is as $\sigma_{i,min}$, but where service data flow i receives extra quality beyond $q_{i,min}$.

- α_i is the service availability which describes the number of services requests of type i that can still be supported by the access resource. Note that it is related to $1/\sigma_{i,min}$. However typically the amount of required resources depends non-linearly on the load level. The non-linear relationship depends on the access technology. In the determination of α_i also the resource usage policy of the access provider needs to be considered. For example, an access provider can limit the amount of resources that may be occupied by a certain service type. In this case the service availability does not depend on the overall remaining resources, but on the available budget of service specific resources.

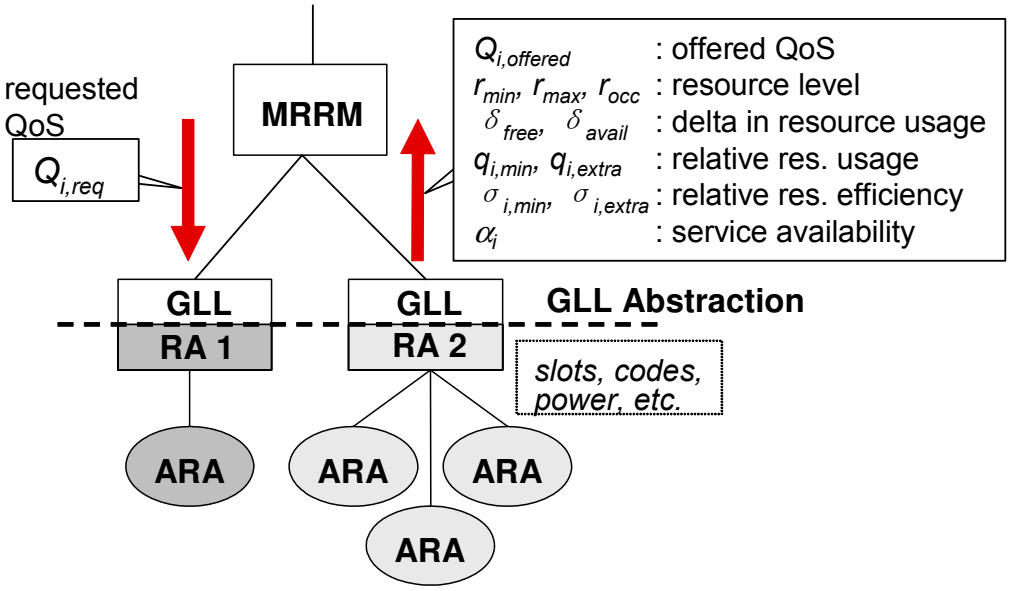


Figure 5.7: GLL resource abstraction of RAT-specific resources.

Note that all measures apart from $Q_{i,req}$ and $Q_{i,offered}$ are given as relative measures, as shown in Figure 5.6. The (absolute) service request $Q_{i,req}$ is received at MRRM from an application function. MRRM forwards $Q_{i,req}$ to one or more GLLs, which convert it into access-specific relative resource requirements q_i. The other relative measures are calculated and reported back to MRRM. GLL is responsible for the conversion between absolute and relative resource measures.

5.3.3.2 Access Resource Structures and Combined Access Resource Metrics

In Section 5.3.3.1 it is presented how the status of a particular access resource can be described in abstract metrics, which enables generically to evaluate the suitability of an access resource to be used for a specific data service. In a realistic access system an access resource consists of several components. All of these need to be considered in the evaluation of an access system. Most notably an access resource has an uplink and a downlink component. A service flow can only be served if the resources in both uplink and downlink are sufficient for

the corresponding service requirements. The uplink and downlink resource components can be independent; this is the case for access technologies using frequency division duplex. For time division duplex systems the partitioning of the access resource into uplink and downlink components can be dynamically adapted.

The main component of the access resource is the *physical resource*; in a wireless communication system it corresponds to the transmit power and the interference head room. An access resource is said to be suitable for a service only if sufficient transmit power is available at the given interference and noise level so that a link budget (i.e. SINR) is achievable that provides sufficient link performance. However, the usage of the physical access resource requires the availability of a further resource component, which is denoted as *channelization resource*. This channelization resource provides a service data flow access to the physical resource; it can be considered as a multiplexing identifier within a limited name space. The channelization resource depends on the access technology. In circuit-switched time-division multiple access based access technologies (e.g. GSM/EDGE) the channelization resources corresponds to the time slots; in code division multiple access based access technologies (e.g. UMTS) the channelization resource is equivalent to the channelization codes; in access technologies based on dynamic multiplexing (e.g. EGPRS, LTE, WiMAX) the channelization resource corresponds to the logical connection identifiers that can be assigned.

Depending on the structure of an access network and the traffic load, the access resource can be either limited by the physical resource or the channelization resource. Let us consider a network with a large number of users with data intensive services. Users in the same cell or neighbouring cells cause a significant amount of interference to each other; when the interference exceeds a certain load level not enough transmit power remains available to achieve a sufficiently high link budget. This scenario is limited by the physical resource. In another example, a large number of users with low data rate services like voice-over-IP or chat are assumed. In this scenario the low traffic volume causes little interference and sufficient transmission power is available. However, when exceeding a certain number of active users the available channelization resources are exhausted and no new services can be admitted any more. In this case the system is limited by the channelization resource.

For multiple components of the access resource a combined resource abstraction has to be determined, as depicted in Figure 5.8. The combined resource abstraction is determined either as the maximum (for r_{min}, r_{occ}, q_i, σ_i) or as the the minimum (for α_i, δ_{avail}, δ_{free}, r_{max}) of the resource metrics of the different resource components.

For a service request $Q_{i,req}$ multiple suitable service realisation options can exist. For example, a service request for an elastic service type has a minimum data rate requirement but can benefit from obtaining a higher throughput. For such a service multiple realisation options of different data rate $\{Q_{i,1}, Q_{i,2}, \ldots\}$ can be provided by GLL with the corresponding resource abstractions, as shown in Figure 5.8.

	Downlink	**Uplink**
Service Request	Rate/Delay/Reliability: $Q_{i,req}$	Rate/Delay/Reliability : $Q_{i,req}$
Resource component 1 (e.g. physical resource)	❖ Resource Status $\{r_{max}, r_{min}, r_{occ}, \delta_{free}, \delta_{avail}\}^{(1)}$ ❖ Service realisation options ➢ $\{Q_{i,1}, q_{i,1}, \sigma_1, \alpha_1\}^{(1)}$ ➢ $\{Q_{i,2}, q_{i,2}, \sigma_2, \alpha_2\}^{(1)}$ ➢ …	❖ Resource Status $\{r_{max}, r_{min}, r_{occ}, \delta_{free}, \delta_{avail}\}^{(2)}$ ❖ Service realisation options ➢ $\{Q_{i,1}, q_{i,1}, \sigma_1, \alpha_1\}^{(2)}$ ➢ $\{Q_{i,2}, q_{i,2}, \sigma_2, \alpha_2\}^{(2)}$ ➢ …
Resource component 2 (e.g. channelization resource)	❖ Resource Status $\{r_{max}, r_{min}, r_{occ}, \delta_{free}, \delta_{avail}\}^{(3)}$ ❖ Service realisation options ➢ $\{Q_{i,1}, q_{i,1}, \sigma_1, \alpha_1\}^{(3)}$ ➢ $\{Q_{i,2}, q_{i,2}, \sigma_2, \alpha_2\}^{(3)}$ ➢ …	❖ Resource Status $\{r_{max}, r_{min}, r_{occ}, \delta_{free}, \delta_{avail}\}^{(4)}$ ❖ Service realisation options ➢ $\{Q_{i,1}, q_{i,1}, \sigma_1, \alpha_1\}^{(4)}$ ➢ $\{Q_{i,2}, q_{i,2}, \sigma_2, \alpha_2\}^{(4)}$ ➢ …
…	…	…
Combined resource abstraction	❖ Resource Status $\{\min(r_{max}^{(x)}), \max(r_{min}^{(x)}), \max(r_{occ}^{(x)}), \min(\delta_{free}^{(x)}), \min(\delta_{avail}^{(x)})\}$ ❖ Service cost options ➢ $\{Q_{i,1}, \max(q_{i,1}^{(x)}), \max(\sigma_1^{(x)}), \min(\alpha_1^{(x)})\}$ ➢ $\{Q_{i,2}, \max(q_{i,2}^{(x)}), \max(\sigma_2^{(x)}), \min(\alpha_2^{(x)})\}$ ➢ …	

Figure 5.8: Combination of access resource metrics.

5.3.4 Access Selection Based on Generic Abstractions

Generic abstractions provide the parameters which are required for access selection (see Section 4.5.1). Figure 5.9 depicts how the abstractions are used in the access selection process. In the access discovery process access performance abstractions are used in order to determine which accesses provide a minimum threshold performance in order to be included in the set of detected accesses and be further considered for access selection. In a first policy-based access control process the detected accesses are validated by comparing typically static access parameters with a set of access usage policies. The criteria for validating accesses mainly consist of user and network preferences (e.g. allowed networks and RATs and access usage tariffs) and the radio access configuration settings including terminal/access network capabilities, supported QoS classes, security support.

When a service data flow is initiated the service specification defines the requirements of the service and the dependency of the service from performance parameters. From the set of validated accesses a subset (i.e. the candidate access set) is determined which contain all accesses that meet the minimum service requirements. In a dynamic access selection process the best access out of the candidate access set is selected. The dynamic access selection process typically operates on changes of parameters in the range of minutes to seconds, or even milliseconds. The dynamic access selection can be of two types: *access-performance*

based or *resource-based*. For this dynamic access selection the access performance and resource abstractions are required.

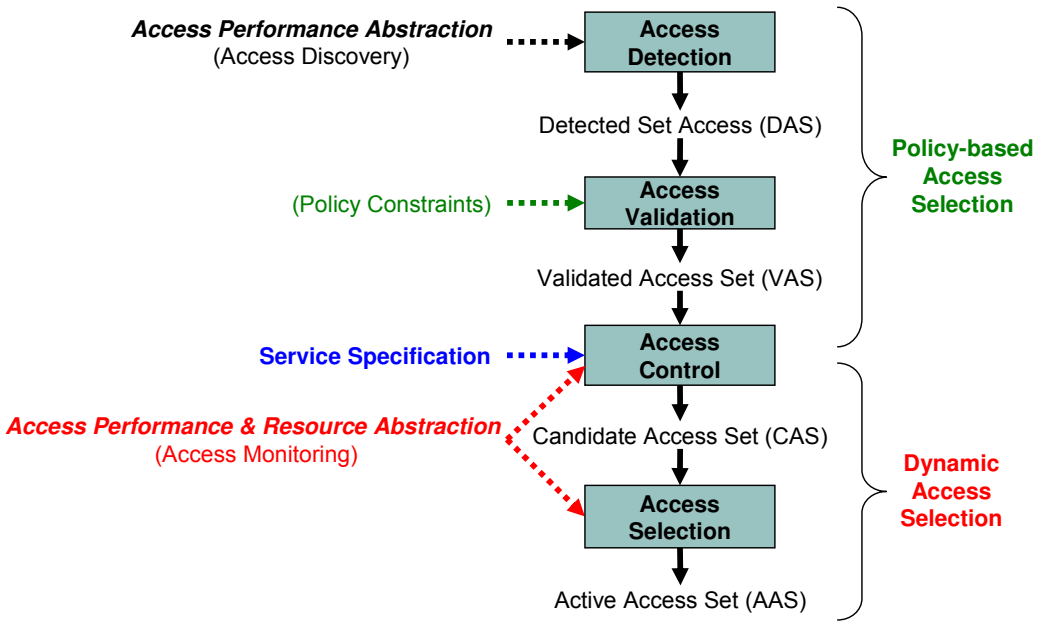


Figure 5.9: Generic abstractions for access selection.

5.3.4.1 Access-performance Based Access Selection

The service specification describes the dependency and sensitivity of the service data flow of the access performance parameters. For all accesses of the candidate set, a value is determined that characterises the suitability of each access for the service data flow (i.e. the service utility in Section 4.4.2.1.1). The generic access performance abstractions provide the performance parameters which are needed to determine the suitability. Depending on the sensitivity of the service data flow on different performance parameters, thresholds can be set for the GLL when changes in access performance are reported.

5.3.4.2 Resource Based Access Selection

Resource based access selection complements access-performance based access selection in order to achieve an optimised usage of the available radio resources and to avoid dropped sessions due to congestion. This is achieved by assigning a weight to the elements of the candidate access set based on the resource situation and thus re-sorting the CAS according to a combination of link performance and resource weight. In [3GPP25.881] a relative load measure (similar to r_{occ}) is defined. However, it is also noted in [3GPP25.881] that this measure is insufficient for load balancing due to varying cell capacities of different RATs. This problem is overcome by adding a relative resource impact (σ_i) in the abstraction, which allows comparing radio cells with different capacity.

Once computed, the MRRM resource measures and possibly the resource levels are signalled from GLL to MRRM, which then uses the values to determine access selection weights (i.e. the resource utility in Section 4.4.2.1.2).

An example of MRRM decisions is the following: For admission control a first step is to check whether $\sigma_{j,min} \leq 1$ for the ARA; if not then the user/session j cannot be admitted there. Initial access selection or load management can be based on many different algorithms. Examples can be:

- Choosing the RAT/ARA with maximum amount of available resources δ_{avail} for a user session:

$$i = \arg\max_j \{\delta_{avail}(ARA_j)\}, \qquad \text{for all } j : \sigma_{j,\min} \leq 1,$$

- Choosing the RAT/ARA with maximum amount of free resources δ_{free}:

$$i = \arg\max_j \{\delta_{free}(ARA_j)\} \qquad \text{for all } j : \sigma_{j,\min} \leq 1,$$

- Choosing the RAT/ARA with minimum amount of relative required resources q_i:

$$i = \arg\min_j \{q_j(ARA_j)\} \qquad \text{for all } j : \sigma_{j,\min} \leq 1,$$

- Choosing the RAT/ARA with the maximum service availability α_i:

$$i = \arg\max_j \{\alpha_j(ARA_j)\} \qquad \text{for all } j : \sigma_{j,\min} \leq 1,$$

- Choosing the RAT/ARA with minimum resource usage efficiency $\sigma_{i,min}$, that is:

$$i = \arg\max_j \{\sigma_{j,\min}(ARA_j)\} \qquad \text{for all } j : \sigma_{j,\min} \leq 1,$$

- Choosing the RAT/ARA with minimum resource usage efficiency $\sigma_{i,extra}$, that is:

$$i = \arg\max_j \{\sigma_{j,extra}(ARA_j)\} \qquad \text{for all } j : \sigma_{j,\min} \leq 1.$$

  Note that if the current load is high in all ARAs, i.e. $\sigma_{j,extra} > 1$ for all RATs, then alternatively the RAT / ARA could be selected which minimises $\sigma_{j,min}$.

Many other MRRM decision algorithms can be considered based on the MRRM resource measures above. The key purpose of the measures is to provide sufficient, comparable radio information on current radio resource state and resource usage efficiency for various heterogeneous RATs for effective MRRM operation.

Note that in some cases additional interactive negotiation between GLL(s) and MRRM can be performed, when GLLs provide additional information about the best service performance that can be provided (without guarantees). For this a translation from relative measures back to an absolute measure $Q_{i,offered}$ is required, as exemplified below:

- MRRM gets a request for (absolute) resources $Q_{i,req}$. E.g. $Q_{i,req}$ is a request for 150 kb/s service with a certain transmission delay limit and required reliability level.

- MRRM passes the request $Q_{i,req}$ on to GLLs, where it is translated to relative resources $q_{i,min}$.

- GLLs reply to MRRM relative load values for load balancing to determine the relative resource costs $q_{i,min}$.

- If spare access resources are available, it may be beneficial to know what absolute service level $Q_{i,offered}$ could be provided to the service by the access (i.e. what maximum Q can be provided by $q_{i,extra}>q_{i,min}$). This requires that GLL would not only make a translation $Q_{i,req}\leftrightarrow q_{i,min}$ but also $q_{i,extra}\leftrightarrow Q_{i,offered}$.

- Then MRRM would get the information from GLL: "Your request $Q_{i,req}$ can be handled, it costs the relative resources $q_{i,min}$. But the access could even support the service request at level $Q_{i,offered}$, which would cost the resources $q_{i,extra}$."

So far, only abstracted values of resource availability and resource costs have been considered. In a realistic scenario, these resource abstractions can be weighted in the MRRM access selection decision according to a priority of different RATs/ARAs. In this case the generic resource metrics need to be adapted by a priority weight factor.

The priorities can be set for several reasons:

- To reflect the operator or terminal priorities of RAT/ARA usage.

- Some RATs/ARAs may be provided by other cooperating operators. In this case additional roaming/cooperation charges may exist for the usage of those RATs/ARAs.

- Some RATs/ARAs may have less efficient operation, e.g. they require more signalling for handover or AAA signalling.

- Some RATs/ARAs may provide less security.

5.4 Access Discovery, Capability Detection and Attachment

For access selection among the available accesses of a user network the corresponding capabilities need to be detected. These capabilities comprise the capabilities of the access as such, as well as the capabilities of the network to which the access belongs. The capability detection depends on the connectivity and attachment states of the user network with respect to the accesses and access network. The *access connectivity established state* refers to the state in which the user has established an access connection to the access network. The *network attachment established state* refers to the state when a user network has achieved mutual authentication with the access network, and the usage of network services has been authorised. The usage of network services is determined by a cooperation agreement between the user network and the access network. There can be different types of cooperation agreements. In mobile networks, the cooperation agreement is typically based on long-term subscriptions; necessary authentication credentials and policy information concerning network service usage are maintained within the user network in a tamper-proof subscriber identity module. In private or corporate networks, authentication credentials and policy rules are often statically configured. With the advent of open local and regional area access networks, there is often no pre-established business relationship between user network and

these access networks. In the Ambient Networks project a framework has been developed for establishing cooperation agreements on-the-fly, in the so-called *composition* process. The trust relationship is established on public-private key pairs, which can be self-generated and allow opportunistic trust relationships. The rules and policies of the cooperation agreement are negotiated within the composition process and lead to a *composition agreement*. Composition agreements are not only established between user networks and access networks. Access networks can also form composition agreements in order to provide joint access services to user networks. Composition agreements can be maintained for a longer time. For example, once a user network has established a composition agreement with an access network, it can reuse this agreement when it attaches to the access network again at a later time. Pre-configured and subscription based cooperation agreements can be seen as a special form of composition agreements. More information concerning network composition can be found in [AN D1-5] [AN D7A2a] [AN D3G1A] [KPJS07] [3GPP22.980].

5.4.1 Network and Access Information

Different types of information are useful to assess the capabilities of an access. A user network can only consider an access network as suitable if the access network capabilities match to the user network capabilities. Network and access information can be grouped into different categories. Some information is related to the network capabilities, other information is related to the access capabilities; this includes the information about access performance and access resource situation according to Section 5.3. The different categories of information are depicted in Figure 5.10.

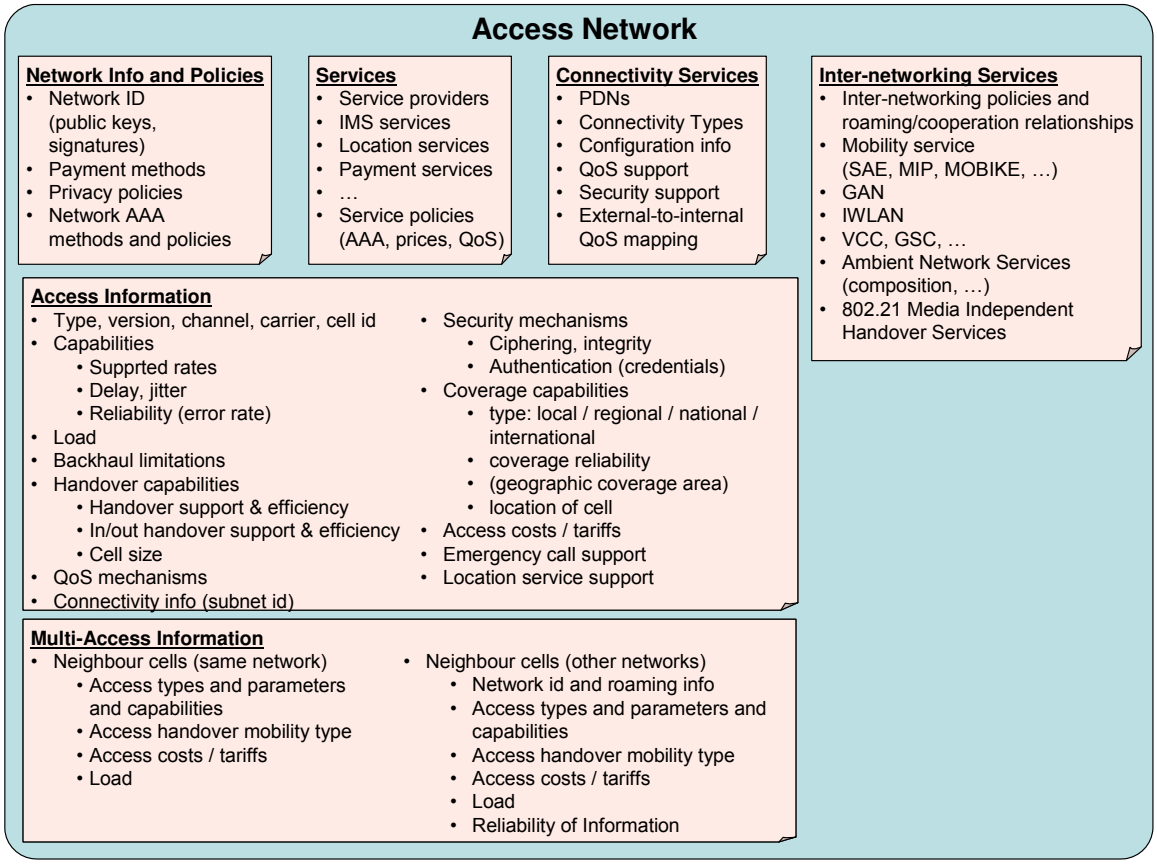


Figure 5.10: Network and access information.

Network Information and Policies

An access network is identified by a network identifier. In addition, a public key and digital signatures allow validating information related to the network and are useful to establish trust relationships for the negotiation of cooperation schemes. Other network information to validate the suitability of an access network are privacy policies about how the information of users are protected. Payment methods describe how the usage of network services can be compensated by a user. Network authentication, authorisation and accounting schemes describe what authentication methods are used, what credentials are used for authentication and what network usage policies apply.

Service Information

The service information describes the services which are accessible via the access network. This can be a list of service providers with which service agreements exist, e.g. for mobile-TV services. Other services are IP Multimedia Subsystem (IMS) services like multimedia telephony or push-to-talk [CG04]. In addition, the access network can provide positioning and location services. Other services can be payment services, when external user payments can be directly authenticated and authorised via the access network. Service usage policies and prices express the rules and conditions of the services, and QoS information describes the supported and required QoS for a service.

Connectivity Service Information

Connectivity information specifies to which packet data networks (PDNs) the access network provides connectivity. A packet data network can be the public Internet; it can also be a private or corporate network, or another type of dedicated network. The connectivity type provides information about the type of network connectivity that is provided, like the IP version. For example, this enables an IPv6 mobile device to only connect to IPv6-capable networks. Additional configuration information enables an easy auto-configuration of the user network, when it connects to the access network. This information can contain pointers to *domain name system* (DNS) servers or auto-configuration methods to be used. QoS support information describes which QoS mechanisms are used and what QoS classes are defined. Additional information can specify how the QoS classes and mechanisms of the access network are mapped to QoS classes and mechanisms of external networks. Security support information expresses the security mechanisms that are provided.

Inter-networking Service Information

Inter-networking service information describes how the services of the access network can be accessed, and how the connectivity can be maintained when changing the accesses. This information contains a list of cooperation relationship with other access providers, like roaming partners. An important interworking service is to provide a multi-access mobility anchor that enables to change the access while maintaining connectivity to the network services. Different types of mobility anchors and corresponding mobility mechanism exist, like the ones defined for the 3GPP system architecture evolution [3GPP23.882] [3GPP23.401] [3GPP23.402]. Other mobility mechanism are based on versions of *mobile IP* (MIP) [RFC3344] [RFC3775] [MIP4] [MIP6], MOBIKE (IKEv2 Mobility and Multihoming Protocol) [RFC4555], *proxy mobile IP* [RFC5213], etc. Another inter-networking service is the *generic access network* (GAN) service, which provides connectivity to both circuit-switched and packet-switched services of a 3GPP access network via any IP based access

system [3GPP43.318]. Interworking WLAN (IWLAN) is another option to access 3GPP packet-switched services via a WLAN or similar access system [3GPP23.234]. Voice-call continuity (VCC) [KVB06] and multimedia session continuity [3GPP23.893] are inter-networking services that allow a transition from packet-switched to circuit-switched transmission, for example, when a user network is leaving the coverage area of an access with sufficient packet-switched transmission capabilities while an access with sufficient circuit-switched transmission capabilities is available. Other inter-working services that can be supported by an access network are ambient network services [AN D2C1] [AN D7A2a] or media independent handover services [IEEE802.21]. Ambient network services enable to establish dynamic cooperation agreements between different access networks, as well as between the user network and access networks. They furthermore allow negotiating dynamically the policies, roles and algorithms for access selection. Media independent handover services allow the transport and notification of access-related events.

Access Information

Access information describes the type of access that is available and the version, e.g. LTE according to 3GPP release 8 and 9, or WLAN according to 802.11 versions a, b and g. The access is furthermore specified by the carrier frequency and the cell identifier. For an access the capabilities can be described in terms of supported data rates, delay and reliability, as well as the quality of service mechanisms. Other information elements are the load of the radio cell, and if the access is limited by a backhaul connection with lower capacity than the radio connection (e.g. a WLAN access point connected via DSL). The reliability of an access also depends on the handover capabilities of the access technology, for example, to what extent lossless or seamless handover is supported within this access technology, and the amount of handovers that can be expected depending on the average cell size. The coverage capabilities describe if an access type has local, regional, national or international coverage. Furthermore, the grade of coverage indicates to what extent black spots without coverage need to be anticipated. Information about the subnet identifier of an access can indicate if higher layer mobility procedures need to be involved, like IP mobility into a new subnet. Other information of access capabilities is if location services or emergency calls are supported. Last not least, the access costs and tariffs characterise the suitability of an access for a user.

Multi-Access Information

Multi-access information relates different accesses to each other. It describes for a user being connected to one access, what other accesses he is likely to be able to connect to. An example is a neighbour cell list, which lists all other accesses with overlapping coverage. For each alternative access, the complete access information described above may be available. Furthermore, the handover mobility type and performance describes what mobility procedure is required to change to the alternative access. Alternative accesses can be grouped into those which belong to the same access provider, and those which are provided by other cooperating access providers. For the accesses of other access providers, the network identity specifies the network owner and roaming information describes the policies of the cooperation. Neighbour cells are further characterised by the reliability that the information is up-to-date. Multi-access information can be provided per geographic area, e.g. as a location service [3GPP23.271]. Multi-access information is currently being standardised for the *access network discovery and selection function* in 3GPP [3GPP23.402] and for the *media-independent information service* in IEEE [IEEE802.21].

5.4.2 Network Advertisements and Capability Retrieval

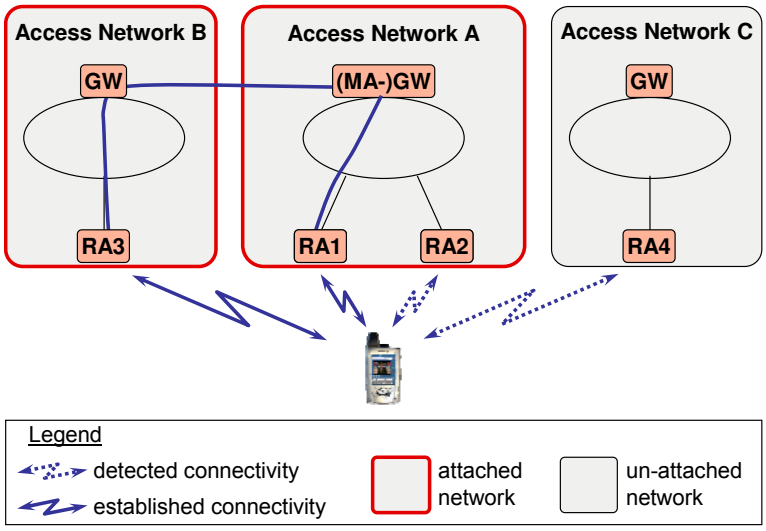


Figure 5.11: User network with different connectivity and attachment states.

Next we discuss how network and access information can be obtained, depending on the connectivity and attachment states of a user network. Figure 5.11 exemplifies different access connectivity and network attachment states of a user network. The user network has established access connectivity via access RA1 and RA3 and has attached to both access networks A and B. The external connectivity is provided via a gateway (GW), and the multi-access gateway (MA-GW) of access network A serves as multi-access anchor. In addition, the user network observes accesses RA2 and RA4 without access connectivity being established. RA2 belongs to access network A, to which the user network is already attached; RA4 belongs to access network C, to which the user network is not attached. Information about network capabilities can be determined for every network that the user network has attached to. In case that a user network detects and identifies a network for which it has already an existing composition agreement, the user network can determine the capabilities based on the locally stored information.

Access capabilities can be determined for every access to which the user network has established access connectivity. A subset of access capabilities can already be determined by the user network, when the access has been discovered. The access performance can be estimated from the type of access and its general characteristics, as well as the measured access link quality. However, without access connectivity information the resource situation and load of an access cannot be determined.

Network and Access Capability Retrieval

When a user network has already established connectivity via one access, it can exploit the connectivity to retrieve information about other accesses. This requires that network and access capabilities can be requested from an access network. We will discuss this process in depth for the scenario depicted in Figure 5.12. A user network has established connectivity via access RA1 to access network A, and it discovers accesses RA2 and RA5. Via the existing connectivity it requests the capabilities of these accesses. RA2 also belongs to access network A; access network A can determine the network capabilities and fetch the access characteristics (e.g. resource availability) of RA2 and provide this information back to the

user network. Access RA5 belongs to another access network C; the user network can directly request capability information from access network C via the connectivity provided by access network A. If we assume that access networks A and C are cooperating, they can share their capabilities. In this case, access network A can directly provide information back to the user network about the capabilities of access network C. However, for scalability reasons it is unlikely that dynamically changing capabilities are shared in this manner. For example, it cannot be expected that access network C provides information about the resource situation of RA5 to access network A. In this scenario we assume that the user network is also within the coverage of accesses RA3 and RA4. However, these accesses have not been detected by the user network. If we assume that access network A maintains multi-access information, it can provide information back to the user network about the availability of those accesses, including information about the frequency band and radio configuration parameters. This enables the user network to start scanning for those accesses if desirable. Since RA4 belongs to another access network, the announcement of its availability via RA-neighbour information is only realistic in case that access networks A and B are cooperating.

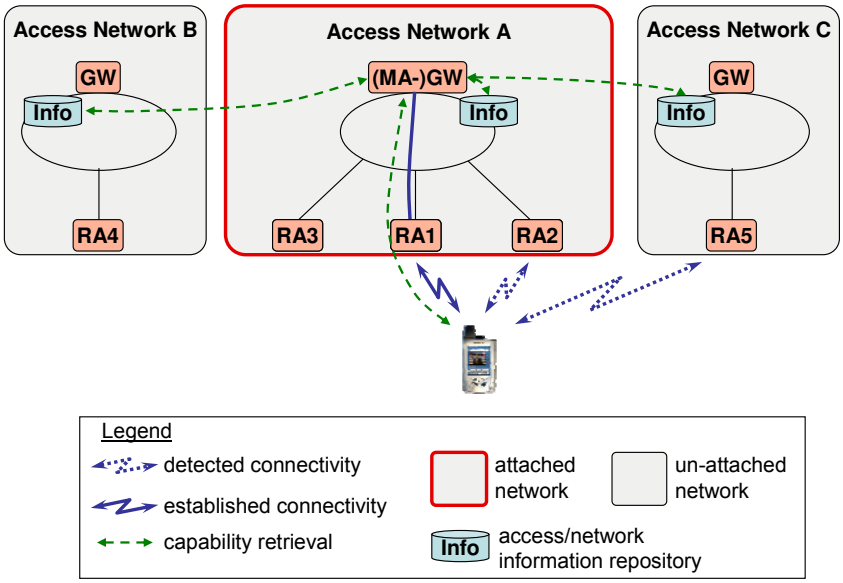


Figure 5.12: Network and access information retrieval.

A special case is when the network and access capability repository is located in a separate network, as shown in Figure 5.13. The information contained in the access capability repository can be provided by the different access networks. It can also be provided by user networks directly. If a user network has found an access network and has determined its capabilities during attachment, it can store this information in the access information repository.

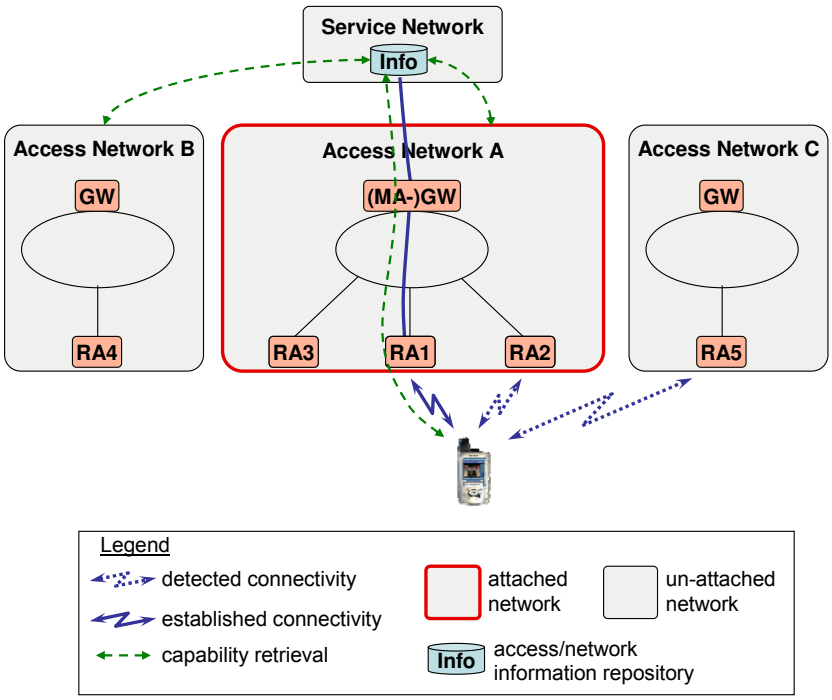


Figure 5.13: Network and access information retrieval with centralised capability repository.

Broadcasted Network and Access Advertisements

An alternative form of providing access and network capabilities to user networks is to add them in advertisements, as shown in Figure 5.14. These advertisements can be distributed via system broadcast distribution to all users within a certain geographic area (typically the coverage area of the access network or a subset thereof). Advertisements via broadcast distribution require little action from the user networks; it is only required that they listen to the broadcast distribution channels. On the other hand, broadcast information increases the system overhead of the access resources. Broadcast information is transmitted without knowledge if receivers for the information are available. Furthermore, broadcast information must be transmitted in sufficient quality within the complete broadcast coverage range[38]. For a wireless system this implies that the information is coded and modulated to be receivable at the cell edge, which requires a robust transmission format with a large overhead. The overhead depends on the frequency at which advertisements are transmitted. In order to limit the amount of overhead, capability information can be grouped according to priorities, and the different priorities can be transmitted at different rates. If the broadcast rate of recurrence is known to the user networks, they can remain in a battery efficient state by only listening to broadcast distribution channels at the correct time intervals.

In addition to an access network providing information about its own capabilities in advertisements, it is also possible to provide access information about cooperating access networks. This is denoted as proxy advertisements. In the example of Figure 5.14 access network C sends advertisements for its own capabilities. Access network A provides in addition to its own capabilities also proxy advertisements for access network C.

[38] Note that particularly user networks with poor radio quality may benefit from searching for alternatives.

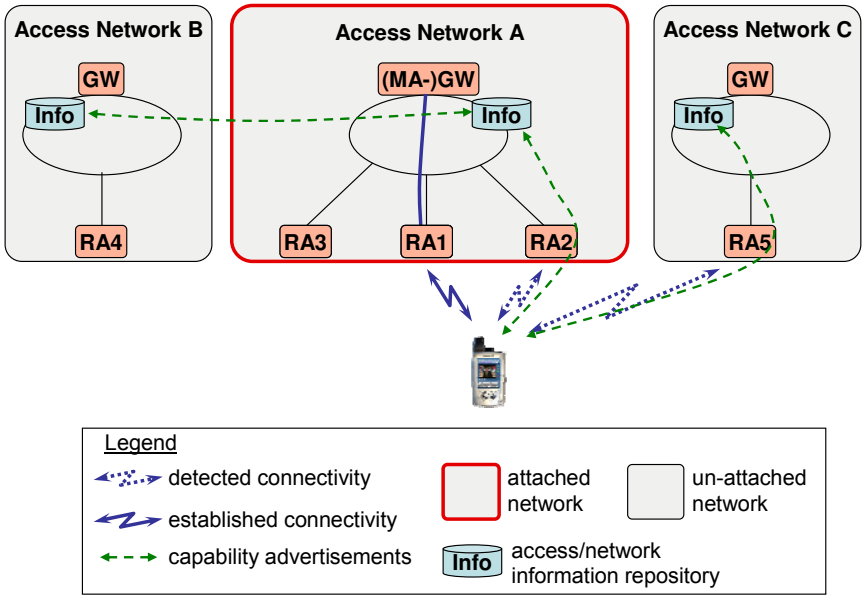


Figure 5.14: Network and access advertisements.

Different types of access networks can be distinguished. Wide-area cellular operators provide access services on regional, national or international level. Users have typically long-term agreements to such network providers, e.g. based on a subscription, and have an established trust relationship. These operators provide access services typically in a known and licensed frequency band, and user networks have pre-installed knowledge how to establish connectivity to these networks. Typically, user networks are constantly attached to their home wide-area network, to be reachable for incoming service requests, like telephony calls. In contrast, local area access providers have a local scope, they provide access services in an unlicensed frequency band and, depending on the location of the user, a plethora of different local networks is available. The trust relationship to these local access networks is typically established on-the-fly, and access network capabilities are provided as described above. In order to increase the trust into access and network advertisements of local access networks that are yet unknown to a user, a trusted third party can confirm their trustworthiness. Such a trusted third party can be a wide-area access provider that has cooperation agreements with a number of local access providers. The trusted third party can produce an assertion for capabilities and services that are advertised by the local access provider. The assertion contains the identities of the assertion provider and the local access provider, as well as the scope (which capabilities/services are assured) and time validity of the assertion. The assertion is signed with the private key of the trusted third party. These assertions could be restricted to be valid only within a certain geographic area, e.g. given as cell identifiers of corresponding (trusted) wide-area access provider, or in a generic geographic area (e.g. as described in [3GPP23.032], which is used for location services in 3G networks [3GPP23.271] [3GPP 25.305]).

5.4.3 Network Attachment and Network/Access Advertisements

Before communication via a local access network is possible, several procedures need to be performed. These procedures include the setup of the access connectivity, the mutual authentication and authorisation between the user network and the access network, the configuration of IP connectivity and the registration of the local connectivity for a mobility

protocol like *mobile IP*. Figure 5.15 shows an example derived in [AETHP06] [AN D7-2] for network attachment to a WLAN network running IPv6 and using *mobile IPv6* as mobility protocol. The procedure contains more than 20 messages being exchanged between the user network and access network, in order to perform:

- WLAN association and connectivity establishment,
- Authentication and authorisation via the 802.1X procedure [IEEE802.1X],
- Establishment of a WLAN security association via the 802.11i procedure [IEEE 802.11i],
- IPv6 connectivity setup via router solicitation/advertisement [RFC2461], secure neighbour discovery (SEND) [RFC3971], multicast listener discovery (MLD) [RFC3810] and duplicate address detection (DAD) [RFC4429],
- *Mobile IP* binding update and acknowledgement messages [RFC3775].

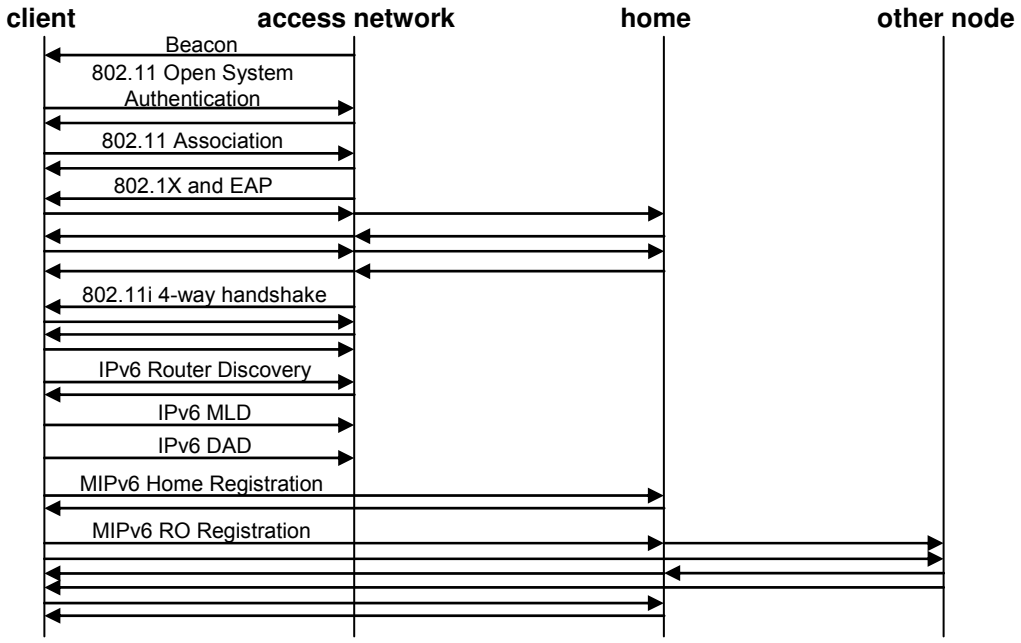


Figure 5.15: Network attachment in WLAN for IPv6 [AN D7-2].

In case that multiple networks are available, network attachment has to be performed multiple times to all networks before the best suited network can be selected. This approach introduces a large overhead and delay. It is largely caused by separate procedures being performed independently from each other. A new *quick network attachment protocol* (QNAP) has been proposed for WLAN in [AETHP06], which can substantially reduce the amount of messages by integrating different procedures in a cross-layer approach. Based on QNAP we have proposed the *ambient network attachment protocol* (ANAP) which generalises network attachment to a wider range of networking scenarios and access technologies and enables to embed access and network advertisements into the attachment and connectivity setup procedure [RCMMS+07] [RAQS07] [AN D7A2a] [AN D7A2b] (see Figure 5.16). ANAP is extending the *HIP base exchange* (I1, R1, I2, R2) used by the *host identity protocol* (HIP) to setup a security association [ID-HIP]. For some access technologies it may be possible to embed/piggyback the ANAP message exchange into the access specific connection setup procedure.

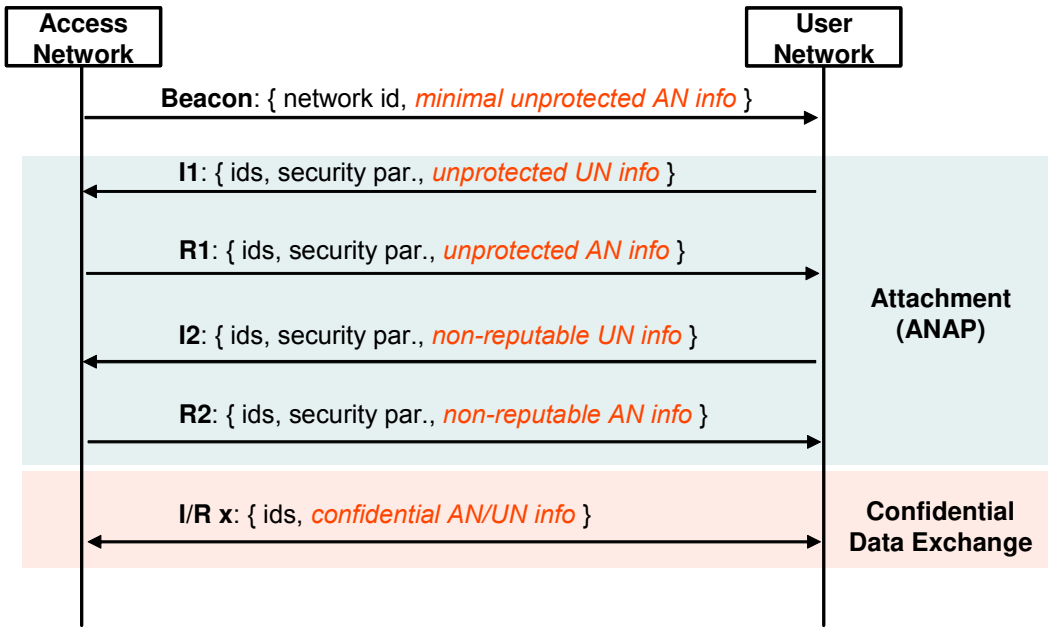


Figure 5.16: ***Ambient network attachment protocol*** **(ANAP) indicating the confidentiality level of network advertisements.**

ANAP comprises a 4-way handshake, in which a security association is established between the two networks. As in QNAP, this handshake comprises configuration parameters for the setup of the connectivity. Within the procedure also different types of network advertisements can be embedded. Depending on when advertisement information is transmitted in the ANAP procedure, different security levels can be differentiated. Advertisements provided in radio broadcast messages are unprotected and can only be trusted if assertions are included that are provided by trusted parties. Broadcast advertisements, however, increase the general system overhead and reduce available radio capacity. Within the first ANAP handshake (I1/R1 in Figure 5.16) advertisements can be included with the same trust and security level as in broadcast messages. However, these advertisements are not broadcasted but only transmitted to the attaching user network. These advertisement can include general networking capabilities, e.g. if connectivity to external networks is provided and what IP version is supported. After the first ANAP handshake the networks are mutually authenticated and the further information provided is non-reputable. Advertisements included within the second ANAP handshake (I2/R2 in Figure 5.16) is suitable to provide cost or QoS information about the access network. After this second ANAP handshake security associations are established and any information can be transmitted encrypted, providing for confidential and non-reputable advertisements and access negotiation.

A generalised advertisement scheme for different access technologies is depicted in Figure 5.17. Different function entities (FEs), like security, compensation or network configuration, provide information elements (IE) to be advertised. These are collected by a network advertisement and discovery (NAD) function and forwarded to multi-radio resource management, which forwards them to generic link layer entities. GLL integrates the ambient network advertisements with access specific signalling procedures. For example, some information elements can be integrated into access technology specific beacons, like WLAN beacons or UMTS system information blocks transmitted on broadcast channels. For other access technologies that do not allow integration of advertisement within access specific procedures, advertisements are exchanged in a separate procedure once access connectivity is

established. For other networking functions involved in the advertisement and discovery procedure the differences between access technologies is transparent, as it is hidden to them by GLL. At the receiving side, GLL receives advertisement information elements and forwards them via MRRM to NAD, where they are dispatched to the corresponding receiving functional entities.

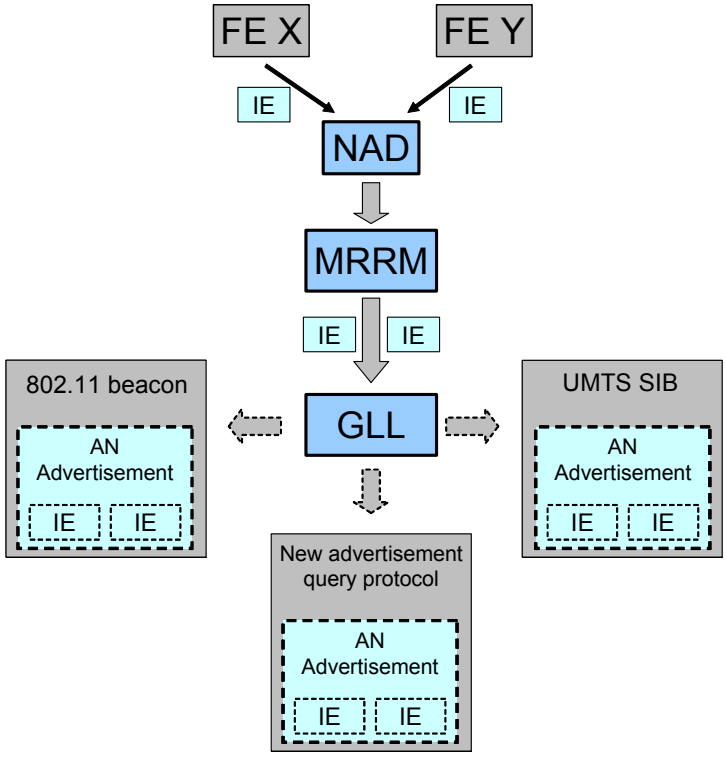


Figure 5.17: Coordination of ambient network advertisements.

The procedure of access network advertisements with respect to connectivity setup and network attachment depends on the level of support provided by a specific access technology. Figure 5.18 – Figure 5.20 show three options of access technology support for advertisement and attachment:

- **Access Technology without Advertisement/Attachment Support** (Figure 5.18)

  In case that the access technology does not support advertisement and attachment, access specific connectivity needs to be established in a first step. The availability of an access is announced via access-specific beacons, which contain only access-specific information. After the access connectivity has been established the advertisement procedure is triggered. The access ambient network advertises its capabilities to the user ambient network; which in turn decides if the access network is a suitable candidate. In this case the attachment procedure is initiated. During the ANAP procedures further network information can be included with increased security level (cf. Figure 5.16).

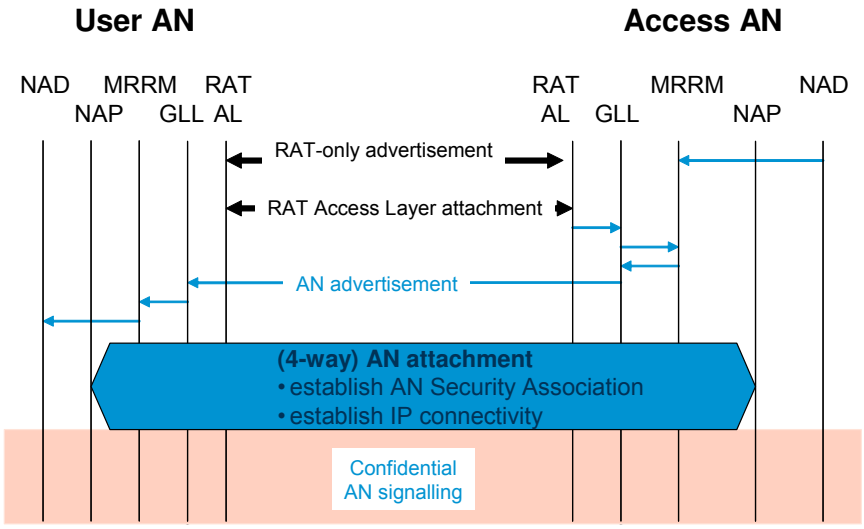


Figure 5.18: ***Ambient network advertisements*** **and** ***ambient network attachment*** **for an access technology without advertisement/attachment support.**

- **Access Technology with some Advertisement/Attachment Support** (Figure 5.19)

  The evaluation of access network capabilities can be improved if the access technology supports access network advertisements. Some information elements can already be embedded within the access technology specific beacon signal. This enables a user network to determine the suitability of an access network before establishing connectivity. If the access network is considered as suitable, the access-specific connectivity is established. After that the ambient network attachment procedure is started in which more network information is retrieved.

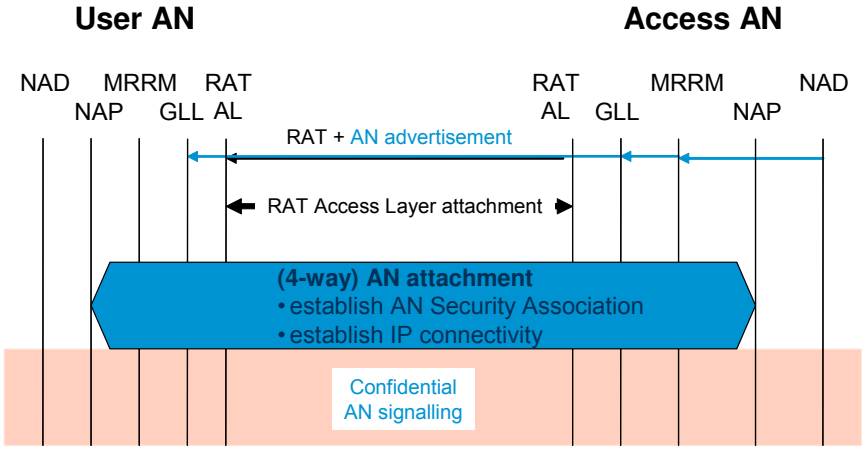


Figure 5.19: ***Ambient network advertisements*** **and** ***ambient network attachment*** **for an access technology with some advertisement/attachment support.**

- **Access Technology with Integrated Advertisement/Attachment** (Figure 5.20)

  Access advertisements are included within the access-layer beacon. If it is decided to establish connectivity to the access network, the first steps of the ambient network attachment procedure are included within the access layer connectivity setup.

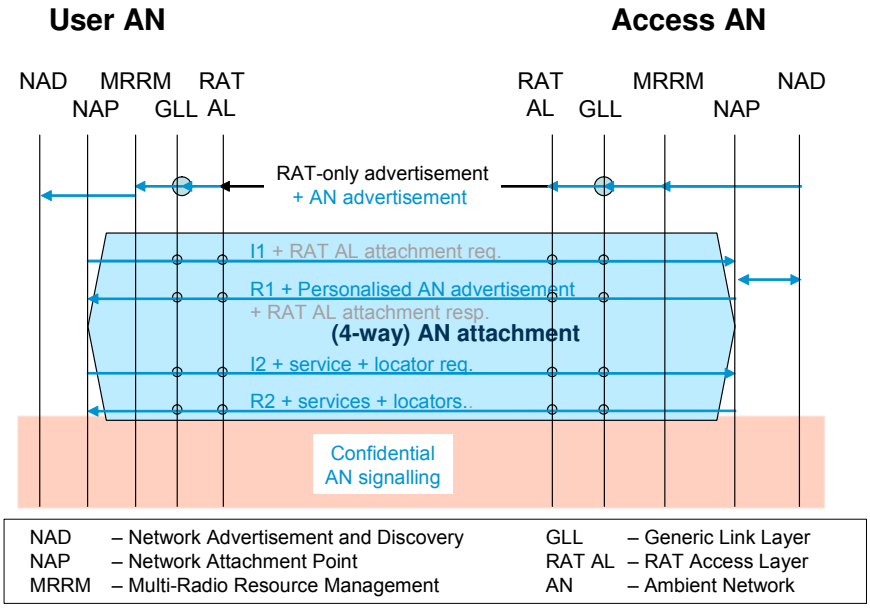


Figure 5.20: ***Ambient network advertisements*** **and** ***ambient network attachment*** **for an access technology with integrated advertisement/attachment.**

5.4.4 Evaluation of Access Discovery and Attachment

5.4.4.1 Objectives and Approach

Next we want to investigate the performance and costs of access discovery, access advertisement and access attachment. The performance is described by the delay introduced by a certain procedure; the costs are described by the number of bits that need to be transmitted in a procedure.

We focus on wireless-local area systems in our evaluation. We assume that wide-area cellular systems have full coverage; a user network is expected to be always connected to a wide-area system, e.g. to remain reachable. Wireless local-area networks are mostly deployed in an unplanned fashion within the unlicensed frequency bands. Every network has a small coverage area. At some locations no WLAN network is available; at other locations a multitude of WLAN networks are present. Since WLAN networks can provide high peak data rates they may be interesting candidate networks for end users. For this reason user networks want to discover WLAN networks and evaluate if those are suitable – in terms of access performance, as well as with respect to network policies and cost of usage. This evaluation procedure consists of three parts: firstly, the physical connectivity to the access network needs to be established; secondly, the capabilities of the access network need to be received via access advertisements; thirdly, the user network needs to attach to the access network and establish a security association. We compare two different procedures:

- **Independent connectivity setup, advertisement, and attachment (CSAA)** perform the different parts independently, as it is common in today's WLAN networks.

- **Integrated connectivity setup, advertisement, and attachment** combine the three parts into a common procedure as described in Section 5.4.3.

For WLAN we consider IEEE 802.11b with *distributed coordination function* (DCF) [IEEE802.11]. We neglect all propagation and processing delay in our evaluation.

The discovery of available access networks is independent from the performance of advertisement and discovery. It depends on how many access networks are available, what carrier frequencies they use, and what kind of scanning is used by the user network. In contrast, the performance of the different connectivity setup, advertisement and attachment procedures depend on the corresponding procedure itself and the load in the network. For this reason we evaluate the performance of these different procedures independently as shown in Figure 5.21.

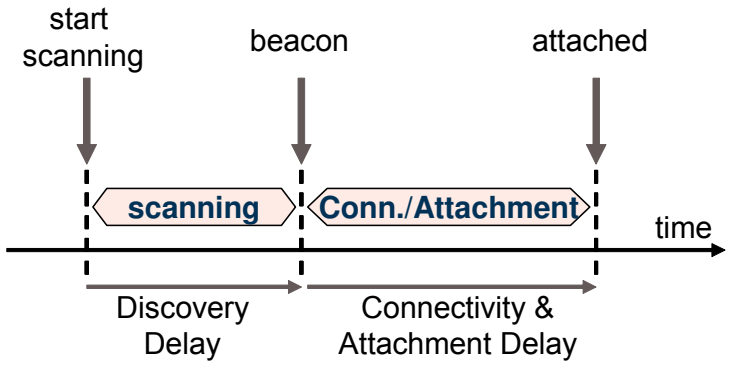


Figure 5.21: ***Discovery*** **and** ***connectivity setup, advertisement & attachment*****.**

5.4.4.2 Comparison Access Discovery and Attachment Options

5.4.4.2.1 Independent Connectivity Setup, Advertisement and Attachment

The procedure with independent connectivity setup, advertisement and attachment (CSAA) is shown in Figure 5.22. The sizes of the different messages (excluding 802.11 specific headers) are indicated. The WLAN network is discovered by a WLAN beacon [IEEE802.11]; the beacon contains a WLAN network identifier (SSID) plus an indication of supported data rates. Firstly, WLAN connectivity is established, consisting of an open system authentication procedure and an association procedure. Next, a new advertisement procedure provides the user network with network capabilities of the WLAN network. We assume four different network services being announced. Finally, the user network attaches to the WLAN network, setups network layer connectivity and establishes a security association with the 4-way handshake of the *ambient networks attachment protocol*. The individual messages, their content, and their length are described in Section C.1of Annex C.

This procedure is characterised by minimal modification of the existing WLAN functionality. The beacon and the access connectivity setup are according to the IEEE 802.11 standard. The advertisement is a new procedure which allows providing network information within WLAN management frames prior to the attachment.

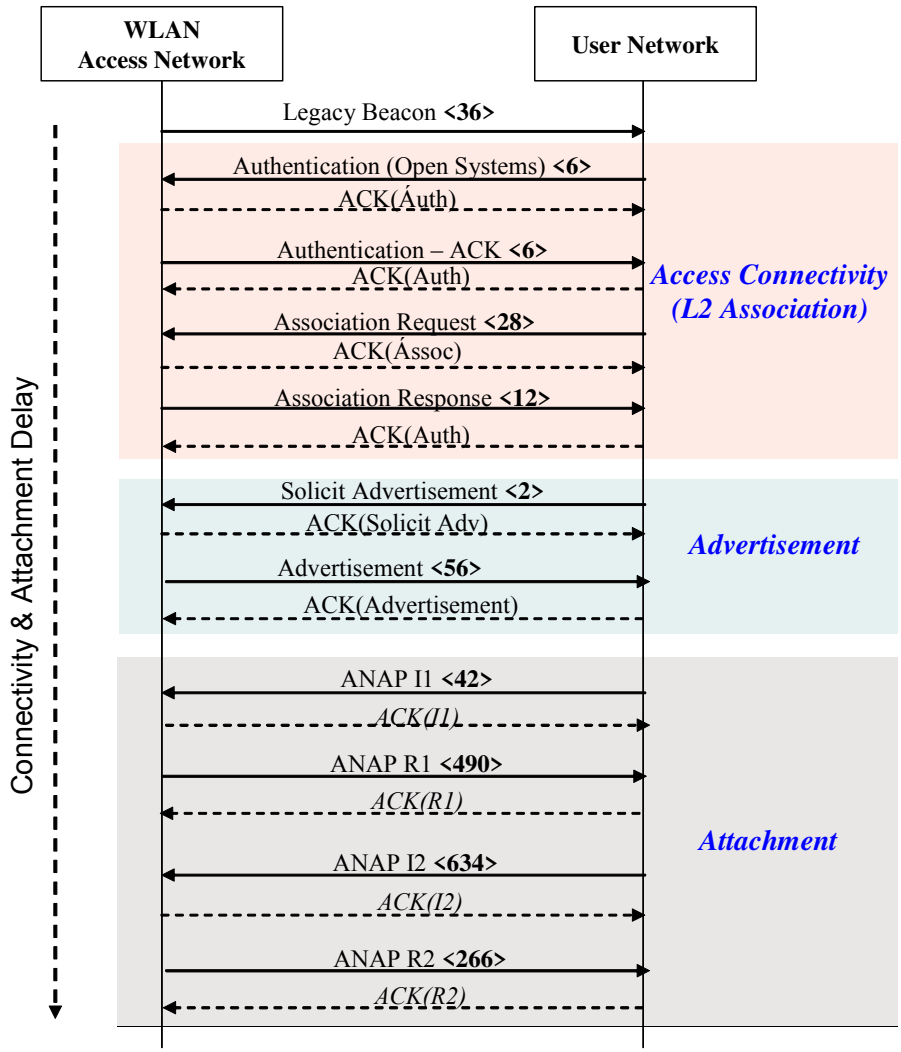


Figure 5.22: Independent *connectivity setup, advertisement and attachment*.
(<x> indicates the message sizes in byte – excluding WLAN PHY, MAC and LLC headers.)

5.4.4.2.2 Integrated Connectivity Setup, Advertisement and Attachment

The integrated procedure follows the principles as described in Section 5.4.3. The procedures of different protocol functions and protocol layers are integrated in order to shorten the signalling procedure and avoid duplication of functionality, like authentication on link layer followed by network layer authentication in the ANAP procedure. The integrated procedure is depicted in Figure 5.23. This procedure differs in several aspects from the previous procedure. Firstly, the network service advertisement is already included in the WLAN beacon. As a result a user network can already discover the network capabilities before contacting the WLAN network and establishing connectivity. Secondly, the association procedure already contains the first handshake of the ANAP attachment procedure. Thirdly, since ANAP comprises mutual authentication for the setup of a security association, this authentication is re-used for access authentication and the WLAN open system authentication can be omitted. The individual messages, their content and their length is described in Section C.1 of Annex C.

This procedure minimises the number of handshakes that are required for network attachment. It also provides network capabilities at an early stage, which allows a user network to determine the suitability of the access before engaging in expensive signalling procedures. This procedure requires from the WLAN standard that additional information can be included in the beacon frames. The IEEE 802.11 *beacon management frame* already contains a *capability information field*; however only few bits are unused (reserved) in this field[39]. Alternatively, it is possible to add a new *network capability information field* to the beacon management frame. Such a field needs to be type-length-value encoded and can be of arbitrary format with a maximum size of 255 byte. Such a beacon extension has been defined for the security extensions 802.11i of 802.11 [IEEE802.11i][40]. For the integration of the WLAN association and the ANAP handshake, the corresponding WLAN management frames would need to be extendable, e.g. by a flexible extension field.

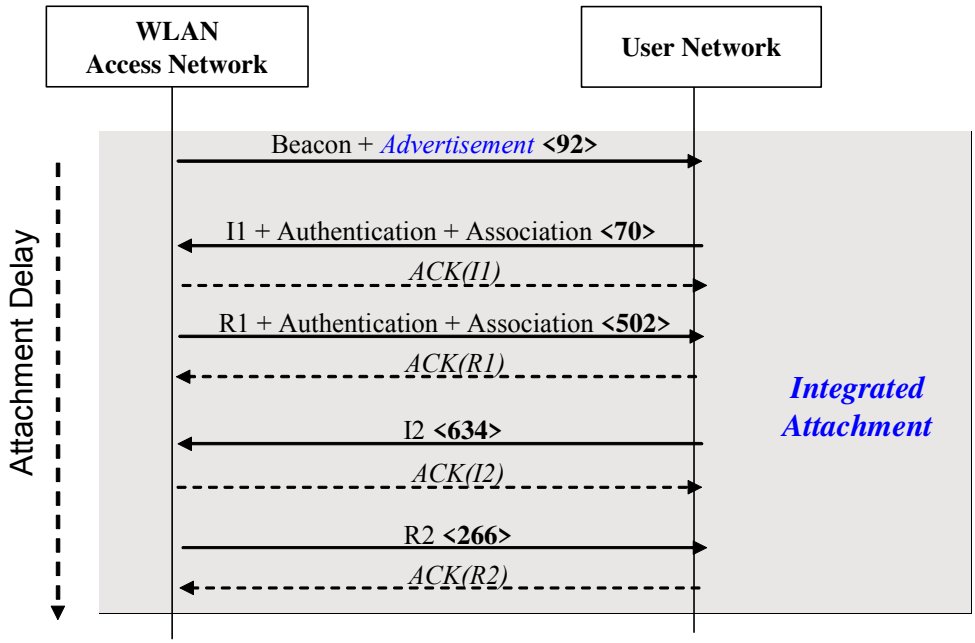


Figure 5.23: Integrated *connectivity setup, advertisement and attachment*.
(<x> indicates the message sizes in byte – excluding WLAN PHY, MAC and LLC headers.)

5.4.4.3 Signalling Costs for Connectivity Setup, Advertisement and Attachment

A metric to determine the efficiency of a procedure is the total amount of data that needs to be transmitted. For a procedure consisting of multiple messages the total amount of transmitted data is the sum of the message sizes L_i of the different messages

$$L_{procedure} = \sum_{i=1}^{n} L_i \quad (5.1)$$

[39] The number of reserved fields are 11, 8, 6, 1 bits respectively for 802.11 versions a, b, g, e.
[40] Since the establishment of a security association is already embedded in ANAP, the 802.11i extension could be omitted and security keys could instead be derived from the keys of the ANAP security association.

Figure 5.24 shows a comparison of the independent and integrated connectivity setup, advertisement, and attachment schemes. We distinguish what amount of bytes are transmitted as part of the different stages – beacon transmission and CSAA. The extended beacon of *integrated CSAA* increases the size of the beacon by 88%. At the same time the amount of transmitted bytes for advertisements and discovery is reduced. Figure 5.25 clear shows that the amount of data required for the ANAP protocol remains constant for the two procedures. The reduction of transmitted bytes stems from the reduced MAC overhead due to a smaller number of totally transmitted MAC frames.

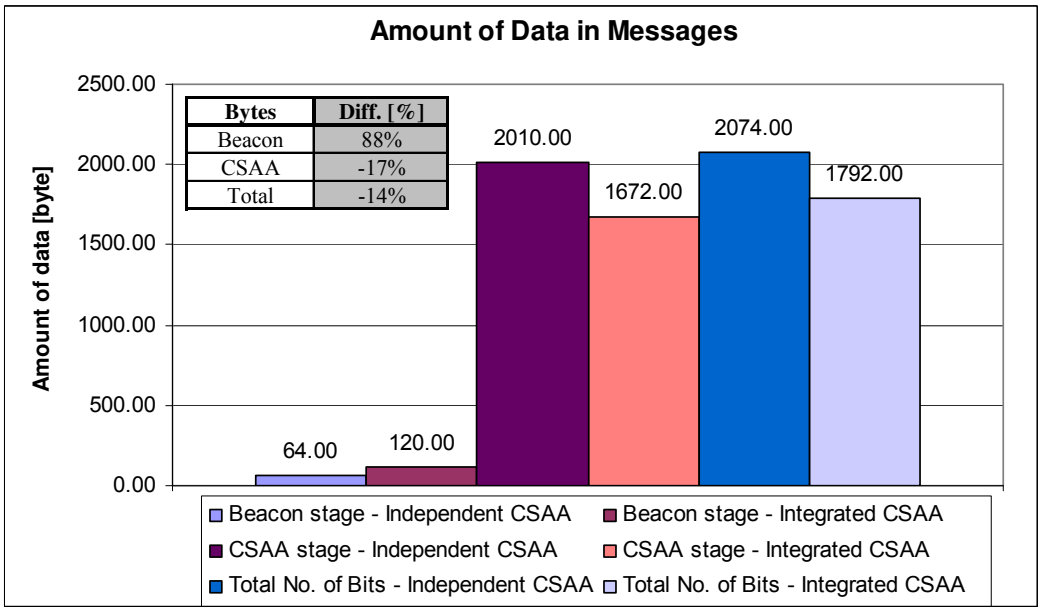


Figure 5.24: Transmitted bytes for different stages.

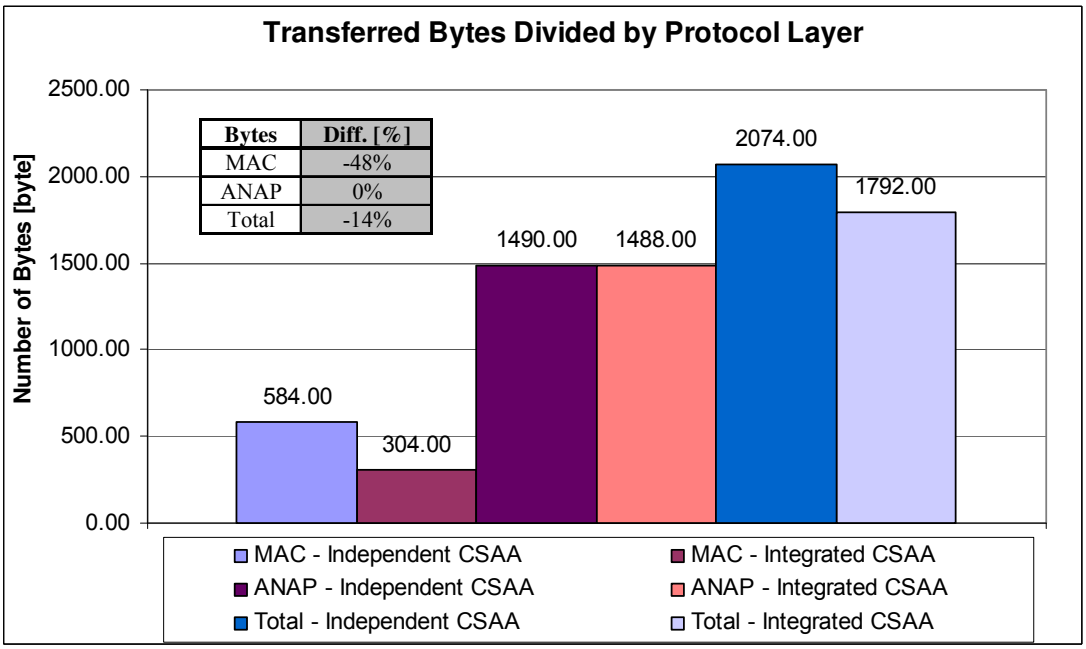


Figure 5.25: Message sizes for different protocols.

5.4.4.4 Access Discovery Delay

A user network has to determine which WLAN networks are available and on what frequency they operate before the access can be evaluated. The WLAN standards can operate in the 2.4 GHz[41], as well as in the 5 GHz[42] frequency band. In Europe[43] 13 channels are available for the 2.4 GHz band [IEEE802.11b2] as depicted in Figure 5.26 and 19 channels in the 5 GHz band [IEEE802.11h]. The channels in the 2.4 GHz band have a channel spacing of 5 MHz. Given a signal bandwidth of 22 MHz these channels are overlapping. Only three channels have sufficiently low adjacent channel interference that they can be considered as non-overlapping; in the 5 GHz band all channels are non-overlapping. The discovery procedure of a WLAN network is denoted as *scanning*. The IEEE 802.11 standard defines two ways of scanning [IEEE802.11]: *passive scanning* and *active scanning*.

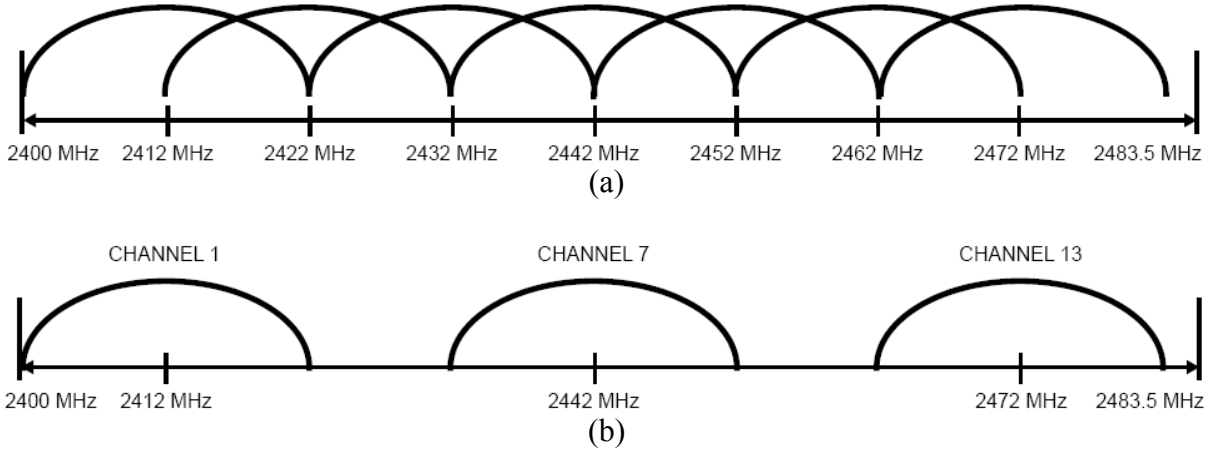


Figure 5.26: WLAN ÎEEE 802.11b configuration in Europe with 13 overlapping channel. (a) depicts channels {1, 3, 5, 7, 9, 11, 13} and (b) depicts a typical configuration of the 3 non-overlapping channels {1, 7, 13}[IEEE802.11b1].

Passive Scanning

A WLAN access point sends out periodic beacon broadcast signals; a user network can detect a WLAN network by searching for beacon signals on the different channels. A beacon signal is transmitted at the most robust transmission mode (i.e. lowest data rate) so that it can also be received by devices located at the cell edge. If a beacon signal is corrupted, e.g. due to collision with another transmission, it is not retransmitted; it is only repeated after the next beacon interval.

A user network searching for available networks scans over different channels according to a configured channel list, which depends on the regulatory domain where the device is operated. The user network scans each channel for up to a *ChannelTime* duration to detect all available access points before switching to the next channel. The *ChannelTime* value should be larger than the beacon interval. The beacon interval is not a-priori known by the UN and

---

[41] 802.11b [IEEE802.11b1] [IEEE802.11b2] and 802.11g [IEEE802.11g]
[42] 802.11a [IEEE802.11a] [IEEE802.11h]
[43] In the USA 11 WLAN channels are available in the 2.4 GHz band [IEEE802.11b1] [IEEE802.11b2] and 12 channels in the 5 GHz band [IEEE802.11a]. The spectrum is regulated by the Federal Communications Commission in [FCC15].

can vary between access points. The beacon interval is included within the beacon itself. The beacon interval between successive beacon signals is a parameter that strongly affects the delay of network discovery for passive scanning. If a beacon is transmitted too frequently it consumes a large amount of transmission resources; if it is sent infrequently it takes a long time before an access point can be discovered. [VK04] and [PR04] have investigated the beacon interval and conclude that a beacon interval of 50-60 ms is a good trade-off. The proportion of WLAN capacity required for beacon transmission according to [VK04] ranges from 32% at 10 ms, via 8% at 50 ms to 3% at 100 ms beacon interval. The typical beacon interval used in WLAN products is 100 ms [VK04] [PR04]. According to [VK04] a good estimation of the discovery delay for passively scanning N channels is $N \cdot 100\ ms$. A way to reduce the discovery delay is to reduce the number of channels that are scanned. Due to the adjacent channel interference of overlapping channels there exist pre-dominant configurations of which channels are used. In Europe such a configuration with three non-overlapping channels is depicted in Figure 5.26. [PR04] describe a *fast scanning* procedure where the list of channels to be scanned is reduced to typical channel configurations. A further enhancement is to provide a UN with explicit knowledge about surrounding access points and their channel configurations via neighbour graph information [PR04] [MSA03].

Active Scanning

A method for faster discovering available access networks is *active scanning* [IEEE802.11]. In this procedure the user network solicits the discovery of access points by broadcasting a probe-request. All access points receiving the probe-request reply with a probe-response, as shown in Figure 5.27. An active scanning cycle is again performed for all channels in the channel list. After sending the probe-request, the UN senses the channel for at least a time *MinChannelTime* to determine if the channel is being used. If no activity is detected during that time the channel is assumed as not used and the UN starts scanning the next channel. If traffic is detected instead, the UN waits for at least a duration *MaxChannelTime*, which gives all access points sufficient time to reply even when contending for the channel to send the response. The discovery delay $T_{discovery}$ for actively scanning M channels is thus bound by (cf. [MSA03]):

$$M \cdot MinChannelTime \leq T_{discovery} \leq M \cdot MaxChannelTime\ . \qquad (5.2)$$

As for passive scanning, the scanning procedure can be accelerated by scanning only a subset of WLAN channels (i.e. *fast scanning*). This subset can either be the most commonly used channels or channels that are indicated as neighbour cells.

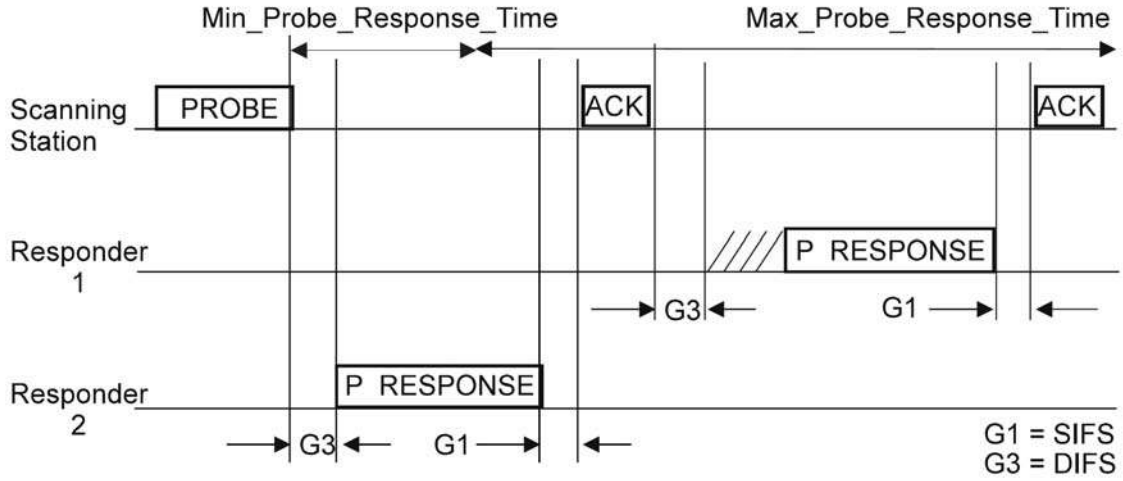


Figure 5.27: Active scanning in IEEE 802.11 [IEEE802.11].

The IEEE 802.11 standard does not specify values for *MaxChannelTime* and *MinChannelTime*. However Velayos et al. [VK04] derive suitable values. *MinChannelTime* should be set large enough so that an AP has time to respond when the channel is otherwise idle. The AP has for an idle channel a minimum congestion window CW_{min}, so that we derive

$$MinChannelTime \geq t_{DIFS} + (CW_{min} \cdot t_{slot}), \tag{5.3}$$

with

t_{DIFS} : DCF interframe space,

t_{slot} : slot length of the physical layer.

Using the values defined for 802.11b[44] [IEEE802.11b1] and further considering that the value must be a multiple of a time unit of 1.024 ms, a suitable value for *MinChannelTime* is 1.024 ms. For *MaxChannelTime* it is required to allow every AP to transmit a probe-response which depends strongly on the cell load. Velayos et al. [VK04] determine a value by simulations for 10 active hosts, leading to a *MaxChannelTime* of 10.24 ms. A similar value is found by Mishra et al. [MSA03] in measurements[45].

Velayos et al. [VK04] consider that a probe-request should always be transmitted twice, to reduce the risk of collision, since a broadcast message is not protected by retransmission. They thus derive the time periods needed to scan a used channel (T_u) or an empty channel (T_e) as:

$$T_u = 2 \cdot T_d + MaxChannelTime, \tag{5.4}$$

[44] *DIFS* = 50µs; CW_{min} = 31; t_{slot} = 20µs

[45] Mishra et al. [MSA03] derive *MinChannelTime* and *MaxChannelTime* from probe-response delay measurements (not limited to idle channels) and derive suitable values of *MinChannelTime* = 6.5ms and *MaxChannelTime* = 11ms.

$$T_e = 2 \cdot T_d + MinChannelTime, \tag{5.5}$$

where T_d is the probe-request transmission time. The total discovery delay $T_{scanning}$ thus depends on the number of used (i.e. carrying traffic) channels u and the number of unused (empty) channels e:

$$T_{scanning} = u \cdot T_u + e \cdot T_e, \quad \text{with } u + e = M, \tag{5.6}$$

The transmission delay T_d depends on the load in the WLAN cell. Velayos et al. [VK04] derive the transmission delay for various load values by simulations.

Figure 5.28 shows the access discovery delay for active and passive scanning of N channels. According to the simulation results of Velayos et al. [VK04] we assume a transmission delay T_d of 20 ms in eqs. (5.4) and (5.5), which corresponds to a load of 5 active mobile terminals in a WLAN cell. For active scanning we assume that all channels are used. It is thus an upper bound; for a portion of channels being empty the active scanning delay would be smaller. The graph shows that scanning can take a considerable amount of time. Passive scanning of three channels takes approximately 300 ms and of 13 channels even 1.3 s. Active scanning provides a considerable gain in delay compared to passive scanning; the scanning delay can be reduced by at least 50%.

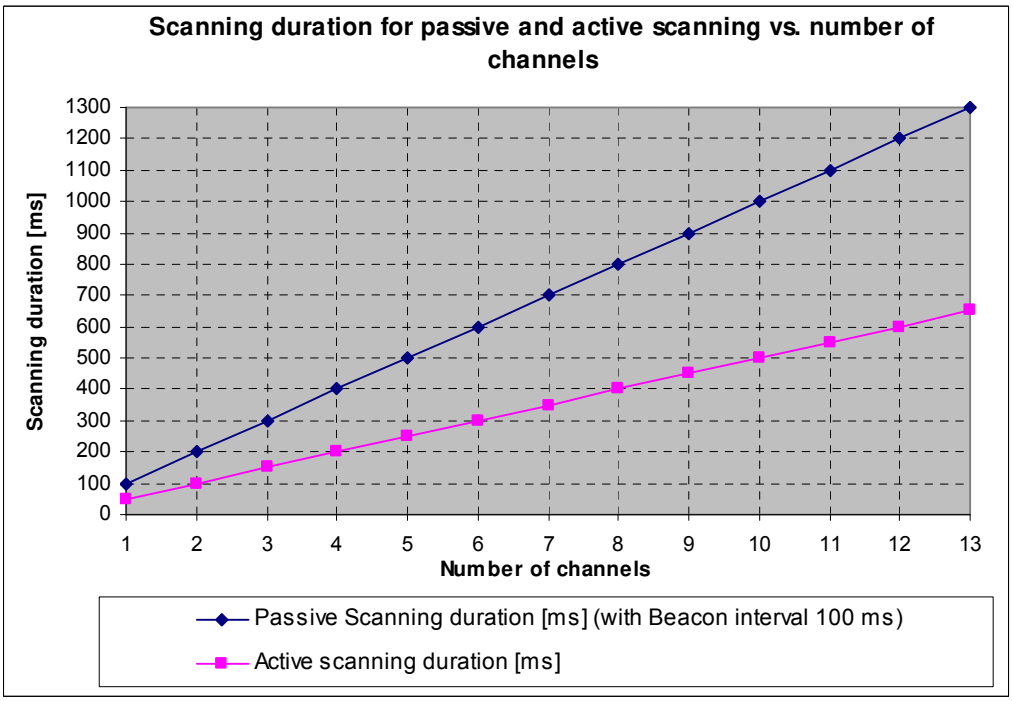


Figure 5.28: Channel discovery time (scanning duration) for active and passive scanning.

5.4.4.5 Connectivity Setup, Advertisement and Attachment Delay

We want to determine the delay of the different procedures for connectivity setup, access advertisement and attachment. It is required to determine the effective transmission times for

the different messages within a procedure. Our approach for determining the delay in 802.11 networks is largely based on the models developed by Heusse et al. [HRBD03].

The transmission delay of a message i of size L_i via 802.11 comprises three portions, as depicted in Figure 5.29. The data transmission time $t_{tr,i}$ corresponds to the time needed for transmitting the data bits and it depends on the used data rate. Furthermore, there is a delay t_{ov} caused by the overhead included in the 802.11 transmission scheme. This overhead is constant and it comprises the MAC layer acknowledgement (ACK), the physical layer preamble for the data frame and the ACK, and the MAC specific interframe spaces (i.e. DIFS and SIFS). In addition, there is the contention delay t_{cont} of the *carrier sense multiple access* (CSMA) scheme that is used in 802.11 DCF (distributed coordination function) mode and is a random backoff delay. The contention delay depends on the number of active users N within the WLAN cell. Finally, there is a delay when the channel is occupied by other users $t_{busy}(N)$ and the transmission is therefore deferred.

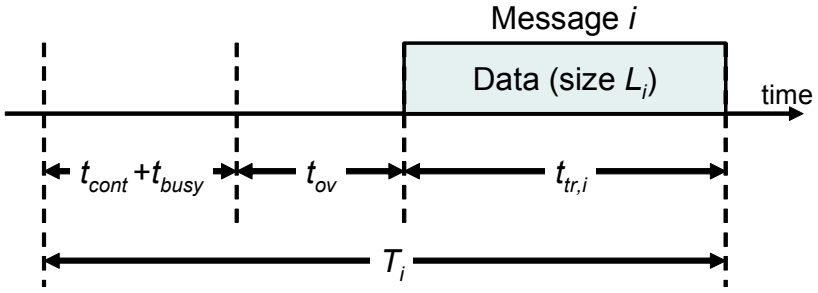


Figure 5.29: Transmission delay for a message of size *L*.

The transmission delay of a message i is then

$$T_i = t_{tr,i} + t_{ov} + t_{cont}(N) + t_{busy}(N), \tag{5.7}$$

where

$$t_{tr,i} = \frac{L_i}{rate}, \tag{5.8}$$

$$t_{ov} = t_{DIFS} + t_{Phy\,\mathrm{Pr}eamble} + t_{SIFS} + t_{Phy\,\mathrm{Pr}eamble} + t_{ACK}, \tag{5.9}$$

$$t_{cont}(N) \cong t_{slot} \cdot \frac{CW_{\min}}{2} \cdot \frac{1 + P_c(N)}{2 \cdot N}, \tag{5.10}$$

with

L_i : length of data message (on MAC level),

rate : physical layer data rate according to the link adaptation mode,

$t_{PhyPreamble}$: length of the physical layer preamble [46],

t_{DIFS} : DCF interframe space,

t_{SIFS} : short interframe space,

t_{ACK} : transmission time of the MAC acknowledgement,

t_{slot} : slot length of the physical layer,

CW_{min} : minimum congestion window size of DCF,

N : number of active users in the cell,

$P_c(N)$: proportion of collisions experienced for each successfully acknowledged MAC frame ($0 \le P_c(N) < 1$).

The term for $t_{cont}(N)$ in eq. (5.10) is an approximation which is based on the assumption that a host always senses a busy channel when it attempts to transmit data [HRBD03]. It furthermore assumes that multiple successive collisions are negligible. The proportion of experienced collisions is then (see [HRBD03])

$$P_c(N) = 1 - \left(1 - \frac{1}{CW_{\min}}\right)^{N-1} . \tag{5.11}$$

When multiple users are active they transmit in turn on the channel. While one user is active the transmission of other users is deferred according to the carrier-sensing principle of CSMA. In the long run CSMA provides a fair opportunity for every user to access the channel. The time period when the channel is used by N users can then be described as the sum of the channel usage time of all users plus the average time spent in collision. We are interested in one particular user, the one which is running the connectivity setup and attachment procedures. We additionally simplify that the other *N-1* users occupy the channel for an equal amount of time. Then we find the relationship (see [HRBD03])

$$\begin{aligned} T_{N-users} &= \sum_{i=1}^{N} T_{user-i} + T_{collisions} \\ &\cong T_{own} + (N-1)\cdot T_{other} + P_c(N)\cdot t_{jam}\cdot N \end{aligned} \tag{5.12}$$

where

$$t_{jam} = \frac{2}{N}\cdot T_{own} + \left(1 - \frac{2}{N}\right)\cdot T_{other} \tag{5.13}$$

with

N : number of active users in cell,

$T_{N\text{-}users}$: channel occupancy of N users,

$T_{user\text{-}i}$: channel occupancy of user i,

$T_{collisions}$: total time of collisions,

---

[46] We consider the physical preamble to contain both the PLCP (*physical layer convergence protocol*) header and the PLCP preamble.

T_{own} : channel usage of the user connecting and attaching to the network,

T_{other} : channel usage of any of the other *N-1* users in the cell,

t_{jam} : channel occupancy of frames subject to collisions,

$P_c(N)$: proportion of collisions experienced per transmitted frame (see eq. (5.11)).

The time period $t_{busy}(N)$ in eq. (5.7) in which the channel is sensed on average as busy from the perspective of a particular user – and in which that user defers transmission attempts – is

$$t_{busy}(N) = T_{N-users} - T_{own} \quad (5.14)$$

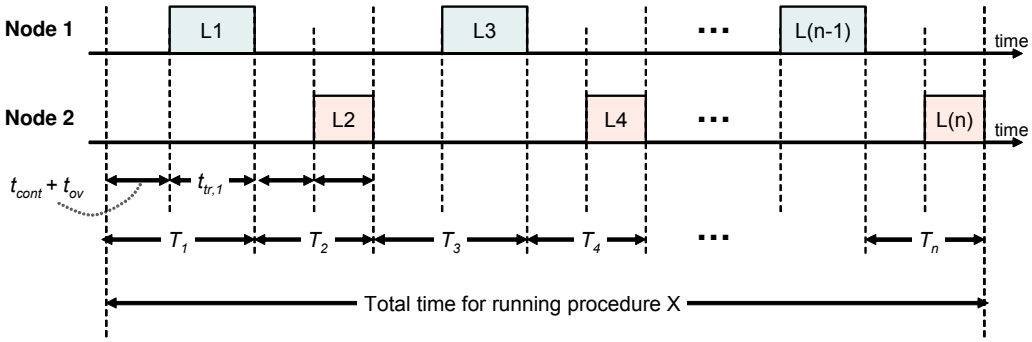


Figure 5.30: Transmission delay for a procedure consisting of n messages.

When running a signalling procedure consisting of multiple messages as shown in Figure 5.30, the total delay of the signalling procedure is the sum of the delay of the different messages:

$$T_{procedure} = \sum_{i=1}^{n} T_i \quad (5.15)$$

where the signalling delay T_i is determined independently for every message i according to eq. (5.7).

We compare now the two procedures for connectivity setup and attachment for an IEEE 802.11b WLAN system: the independent connectivity setup, advertisement, and attachment (*independent CSAA*) in Figure 5.22 and the integrated connectivity setup, advertisement, and attachment (*integrated CSAA*) in Figure 5.23. The WLAN parameters and the content of the different messages of those procedures are given in Annex C.

Transmission Delay without Contention and Busy Channel

The total transmission delay in eq. (5.7) comprises terms t_{busy} and t_{cont} that depend on the activity of other hosts in the cell, as well as terms that only depend on the CSAA procedure itself. In order to understand the influence of these different terms independently we first investigate the transmission delay without other users. For this we neglect delays due to contention or deferral caused by a busy channel (i.e. only the delay values $t_{tr,i} + t_{ov}$ in eq. (5.7) are considered). We consider the connecting host to have a data rate of 1 Mb/s.

The total transmission time of the different stages – broadcast of the beacon and CSAA – is depicted in Figure 5.31. The extended beacon of *integrated CSAA* increases the transmission time of the beacon by 59%; at the same time the delay for the CSAA procedure is reduced by 28%, resulting in a reduction on 25% for the combined beacon and CSAA stages.

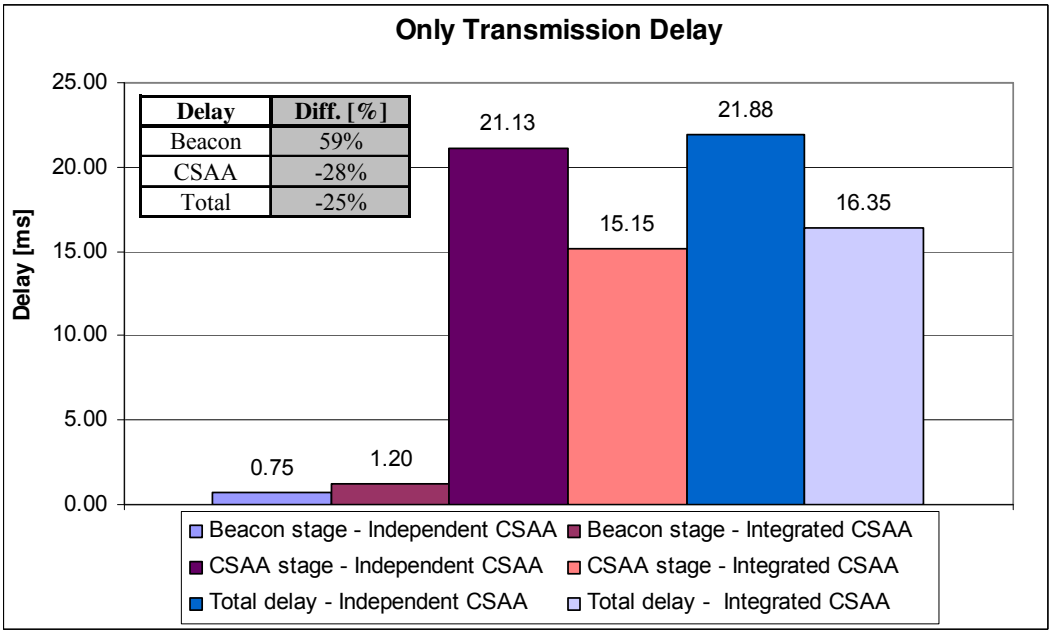


Figure 5.31: Transmission delay for different stages (excluding backoff and busy channel times).

The proportion of transmission time that is caused by different protocol layers (see Annex C) is shown in Figure 5.32. It can be seen that the *integrated CSAA* procedure reduces the overhead of the physical and MAC layer by 62% and 48% respectively. This gain is due to the number of reduced messages that are sent. The delay contribution of the ANAP protocol for advertisements and attachment is the same for both procedures.

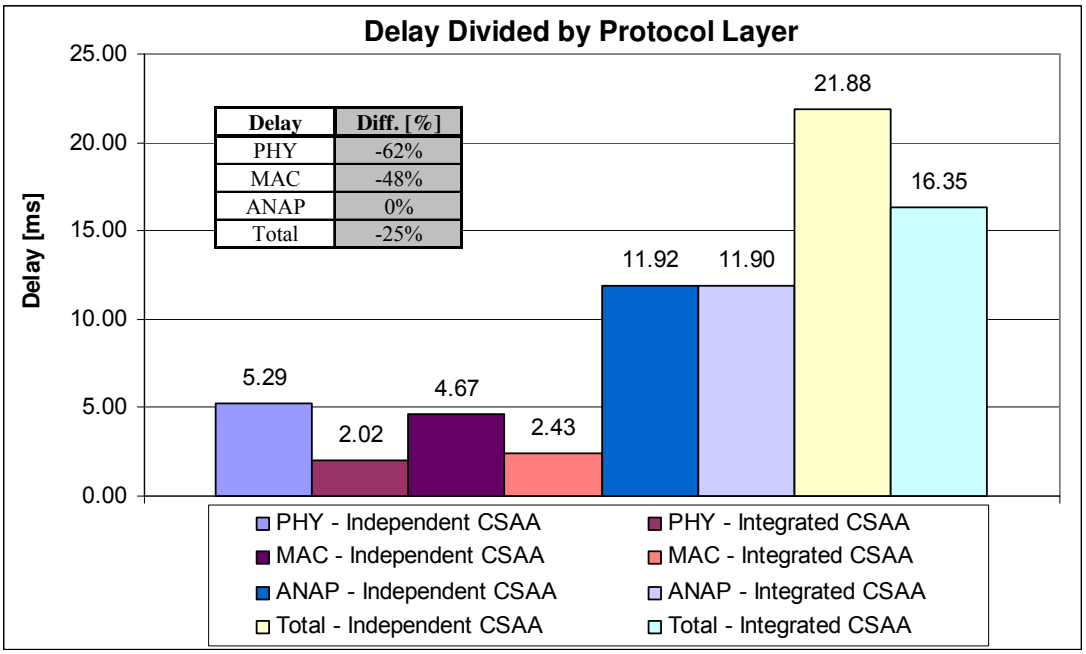


Figure 5.32: Contribution of different protocol layers to the transmission delay for (excluding backoff and busy channel times).

Transmission Delay with Contention and Busy Channel

We now include the delay components of backoff and deferral due to busy channel (i.e. t_{busy} and t_{cont} in eq. (5.7)) in the investigation. The attaching host competes with *N-1* other hosts for access to the WLAN channel. We assume that these other hosts have a physical layer data rate of 5.5 Mb/s and transmit typical Internet traffic with packet sizes of 1500 byte.

Figure 5.33 shows the time that it takes to transfer the beacon for different load levels. The extended beacon for the *integrated CSAA* has a larger beacon size (MAC PDU size) of 120 byte compared to the 64 byte beacon of *independent CSAA* due to the embedded advertisement. Therefore the beacon broadcast takes approx. 0.5 ms longer for the *integrated CSAA*. If there is no other active host in the cell, the extended beacon doubles the beacon transmission time. At higher load the contention becomes quickly the dominating portion of the beacon transmission delay. As a result, at loads of 5, 10 and 20 active hosts the increase of beacon transmission delay for the extended beacon reduces to 11%, 5% and 2% respectively. We note that an extended beacon can significantly increase the transmission delay and thus the capacity that is occupied by the beacon. However, this is only the case at low load when capacity is not an issue. At high load when capacity becomes critical the relative capacity overhead of the beacon reduces to only a few percent. Figure 5.34 shows the delay of the two CSAA options depending on load excluding the beacon transmission time. The *integrated CSAA* clearly reduces the signalling delay compared to *independent CSAA*; the reduction ranges from 40% at low load to almost 60% at high load. In Figure 5.35 the combination of the total CSAA is depicted, including the beacon transmission. Although *integrated CSAA* has a longer beacon time, it still reduces the total CSAA procedure between 27% at low load and 52% at high load. The exact values of the beacon and CSAA delay are summarised in Table 5-1.

The *integrated CSAA* reduces the number of messages that are exchanged. Since every message comprises a certain MAC and PHY overhead the overhead can be reduced by including the required information in fewer messages. The total amount of data transmitted for *integrated CSAA* is 14% lower than for *independent CSAA* (cf. Figure 5.25). However, the total gain of *integrated CSAA* is by far larger than what can be saved in number of bits. Every contention event, i.e. every independent signalling message, leads to a capacity loss caused by the contention process. The reduced number of messages reduces this contention loss. The overall gain of *integrated CSAA* is at low load equally achieved by reducing the amount of transmitted bits and the number of contention events. At medium to high load the reduced contention loss contributes stronger to the gain than the reduction in amount of data.

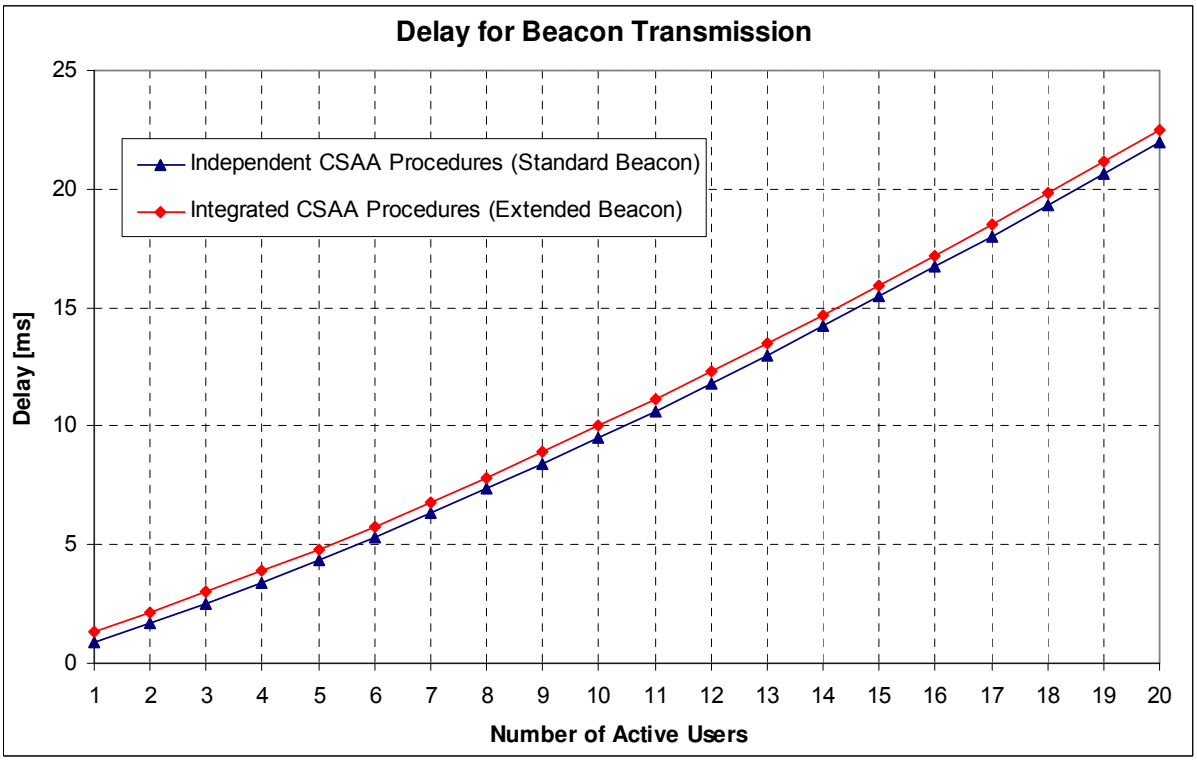


Figure 5.33: Beacon transmission delay for independent and integrated procedures of connectivity setup and attachment.

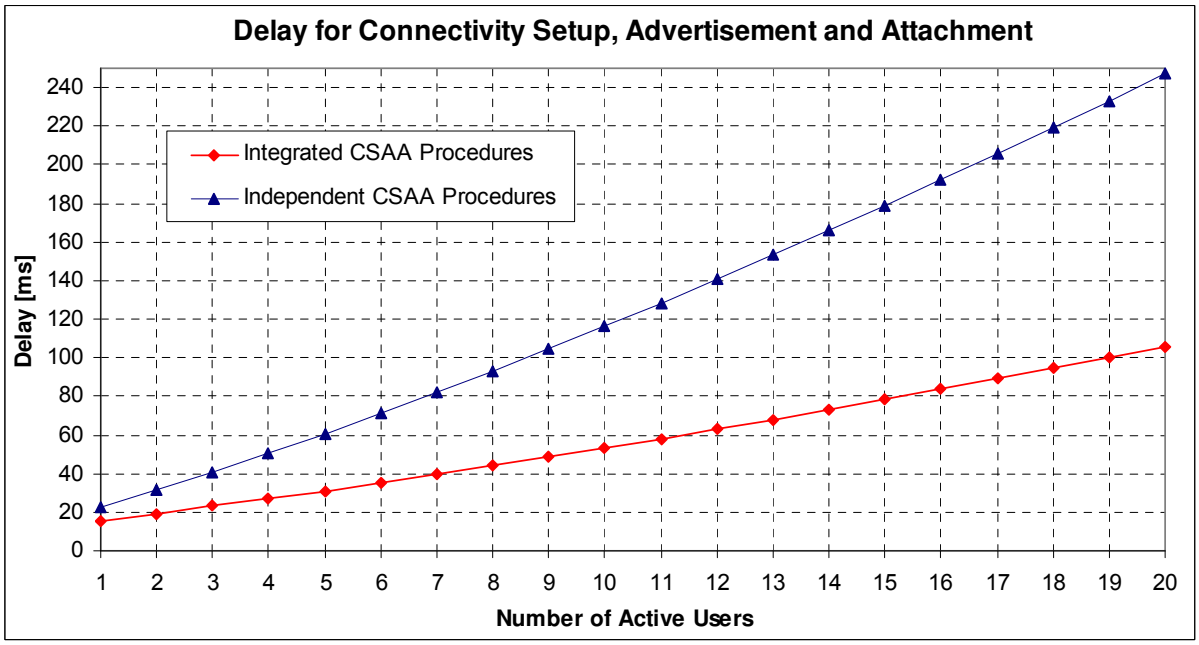


Figure 5.34: Transmission delay for independent and integrated connectivity setup and attachment.

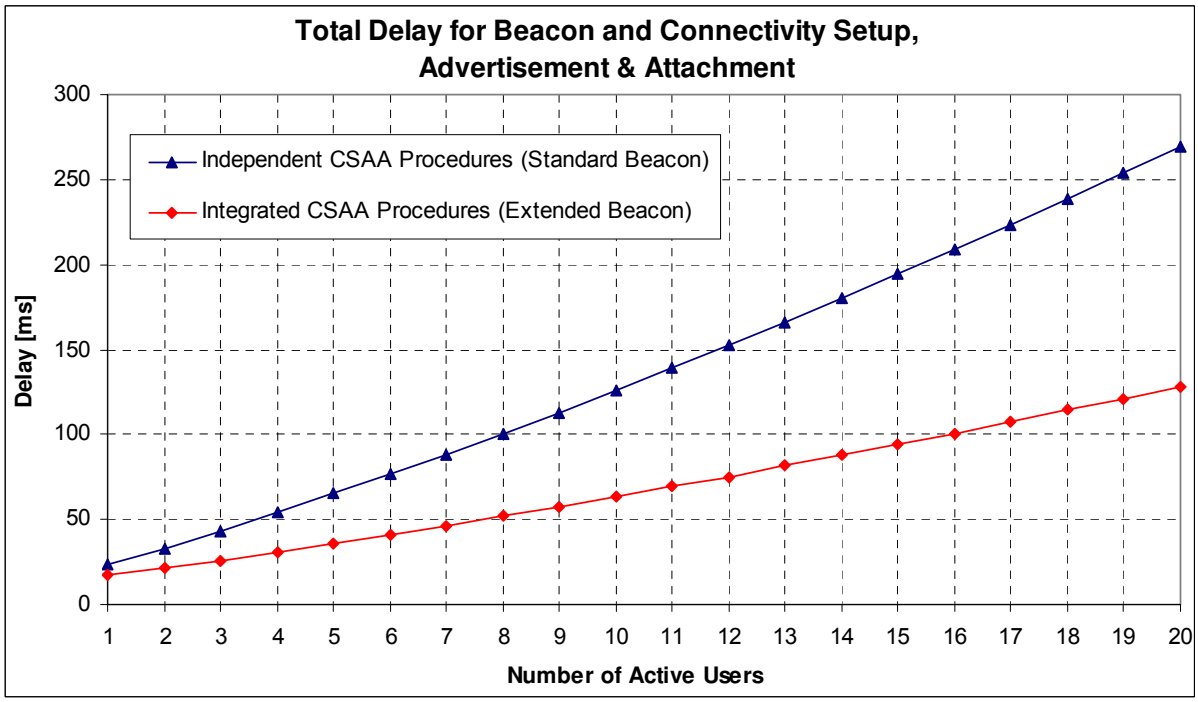


Figure 5.35: Total delay (beacon + procedures) for independent and integrated connectivity setup and attachment.

Table 5-1 : Comparison of independent (Figure 5.22) and integrated (and Figure 5.23) connectivity setup for WLAN 802.11b.

Number of Active Users (N)	Time Spent for Beacon Transmission [ms]			Time Spent for Connectivity Setup, Advertisement & Attachment [ms]			Total Delay [ms}		
	Independent Procedures	Integrated Procedure	Diff. [%]	Independent Procedures	Integrated Procedure	Diff. [%]	Independent Procedures	Integrated Procedure	Diff. [%]
1	0.91	1.36	49%	22.68	15.77	-30%	23.59	17.13	-27%
2	1.70	2.15	27%	31.56	19.32	-39%	33.26	21.48	-35%
3	2.54	3.00	18%	40.86	23.05	-44%	43.40	26.04	-40%
4	3.42	3.88	13%	50.57	26.93	-47%	54.00	30.81	-43%
5	4.35	4.81	11%	60.67	30.97	-49%	65.02	35.78	-45%
6	5.31	5.78	9%	71.14	35.16	-51%	76.45	40.93	-46%
7	6.31	6.78	7%	81.95	39.48	-52%	88.26	46.26	-48%
8	7.35	7.82	6%	93.09	43.94	-53%	100.44	51.76	-48%
9	8.42	8.89	6%	104.54	48.52	-54%	112.96	57.41	-49%
10	9.52	10.00	5%	116.29	53.22	-54%	125.82	63.22	-50%
11	10.65	11.13	4%	128.32	58.03	-55%	138.98	69.16	-50%
12	11.82	12.30	4%	140.62	62.95	-55%	152.44	75.25	-51%
13	13.01	13.49	4%	153.17	67.97	-56%	166.17	81.46	-51%
14	14.22	14.71	3%	165.95	73.08	-56%	180.17	87.79	-51%
15	15.46	15.95	3%	178.96	78.29	-56%	194.42	94.23	-52%
16	16.72	17.21	3%	192.18	83.57	-57%	208.91	100.79	-52%
17	18.01	18.50	3%	205.61	88.94	-57%	223.61	107.44	-52%
18	19.31	19.81	3%	219.22	94.39	-57%	238.53	114.20	-52%
19	20.64	21.13	2%	233.01	99.90	-57%	253.65	121.04	-52%
20	21.98	22.48	2%	246.97	105.49	-57%	268.95	127.97	-52%

5.4.5 Frequency of Access Discovery

In Section 5.4.4 we have investigated the difference in performance between varying options of advertising access network capabilities and attaching to an access network. In this section we examine how often it is desirable to validate the capabilities of an access network. Let us re-emphasise that a mobile user network is battery operated. Measuring and attaching to access networks consumes a considerable amount of battery energy. We have shown in [Tra07] [AN D21C3] that constantly scanning for WLAN networks requires approximately the same amount of power as constantly receiving data at 600 kb/s. The same study has revealed that the battery consumption varies only by a few percent between the different discovery and attachment schemes described in Section 5.4.4.2. This stresses that a search for WLAN access networks should be limited to situations when it is reasonable.

We assume again that a user network is always connected to a wide-area wireless network, e.g. to remain reachable for incoming data sessions. WLAN access networks are evaluated and used whether they either improve the service performance or whether they reduce the costs for a data session. In order to understand how often a search for WLAN networks is desirable we look at the data rate that is achievable in a WLAN cell depending on the distance from the access point when link adaptation selects the suitable physical layer transmission mode according to the link quality, as depicted in Figure 5.36.

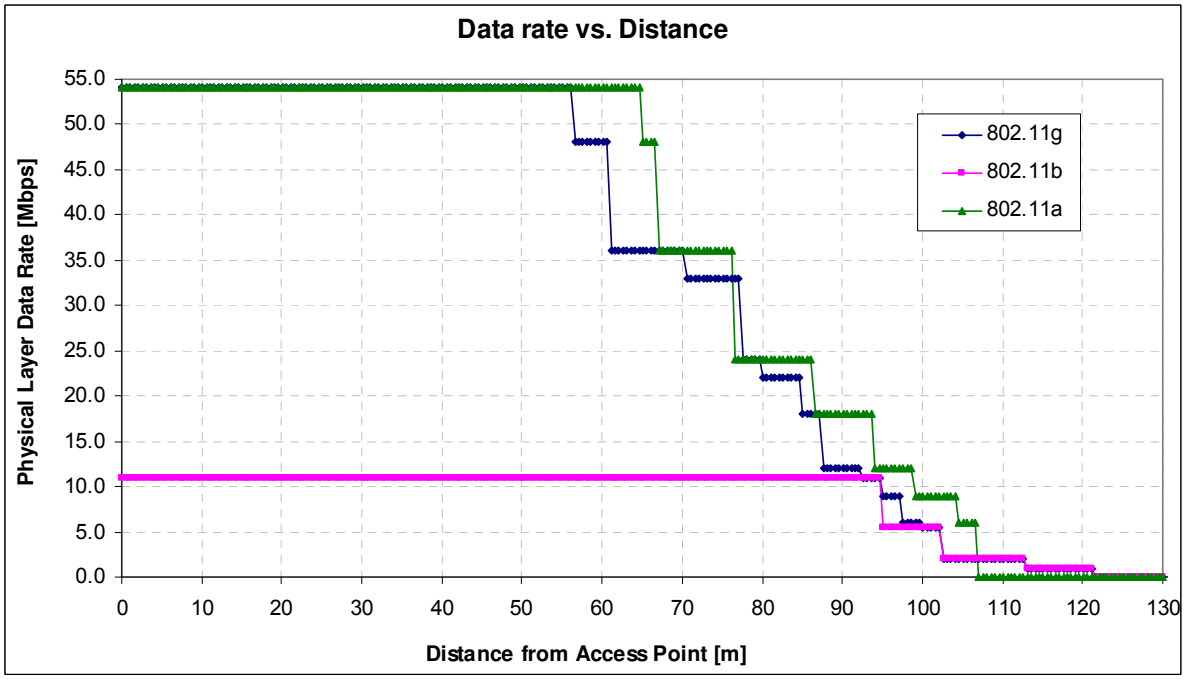


Figure 5.36: WLAN data rates versus distance between user network and access point for a propagation scenario and WLAN parameters as described in Section C.3 of Annex C.

If we ignore shadow fading, the regions of different achievable data rates within a 802.11b WLAN network are as shown in Figure 5.37. We consider that a user is moving through an area containing one or more WLAN cells. We declare it as a considerable event whenever the achievable data rate in the WLAN cell changes by a factor of two. Such a change should be significant enough for a user network to re-evaluate if the capabilities of a network may be sufficient to attach or handover an ongoing session to WLAN.

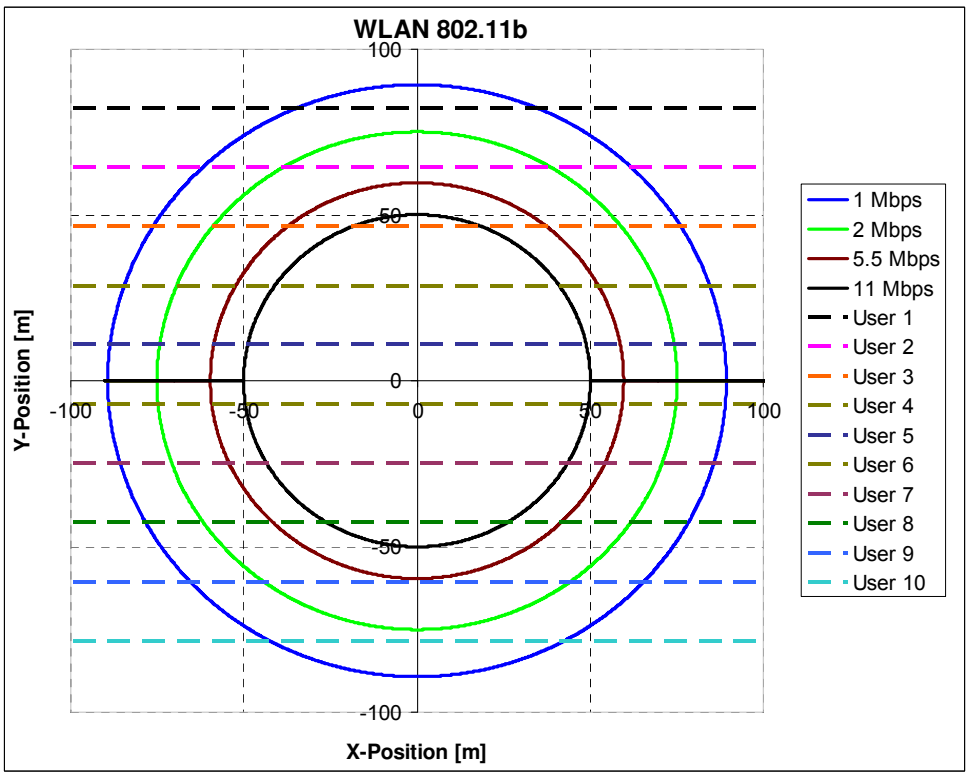


Figure 5.37: Achievable data rates in a 802.11b WLAN cell (neglecting shadow fading) with trajectories of users moving through the cell.

Figure 5.37 shows the coverage within a WLAN radio cell for the different achievable data rates. Ten trajectories of how users can move through the WLAN cell are indicated. With a geometric analysis we derive the distances that a user has to move along the different trajectories such that the WLAN data rate changes by a factor of two. Table 5-2 shows the difference for the varying IEEE 802.11 versions, showing the minimum, maximum and average distances for the different trajectories according to Figure 5.37.

Table 5-2 : Distances between points in a WLAN in IEEE 802.11 b/g/a cell where the achievable data rate changes by a factor of two.

WLAN 802.11b			
Data rate	Min. distance [m]	Max. distance [m]	Average distance [m]
1 Mb/s	14.3	84.3	23.8
2 Mb/s	15.5	88.1	27.0
5.5 Mb/s	9.7	18.8	12.7
11 Mb/s	37.2	99.2	75.8
WLAN 802.11a			
Data rate	Min. distance [m]	Max. distance [m]	Average distance [m]
6 Mb/s	10.4	68.1	18.0
12 Mb/s	13.9	68.3	22.9
24 Mb/s	17.0	55.6	25.9
54 Mb/s	27.3	47.9	38.9
WLAN 802.11g			
Data rate	Min. distance [m]	Max. distance [m]	Average distance [m]
1 Mb/s	14.3	84.3	23.8
2 Mb/s	18.9	88.1	30.6
6 Mb/s	9.1	23.1	13.2
12 Mb/s	12.0	39.7	16.8
24 Mb/s	14.9	50.3	25.6
54 Mb/s	36.3	39.7	38.0

From Table 5-2 we derive that whenever a user moves approximately 20 m we expect that the achievable data rate of a WLAN network has changed by a factor of two. Note that if realistic shadowing would also be considered, this distance would further decrease. Table 5-3 shows the scanning intervals for WLAN of a user network depending on its velocity. We can see that a stationary user network or one that moves very slowly should re-evaluate / scan for WLAN networks approximately once every minute. A user moving at pedestrian velocity should re-evaluate / scan for WLAN networks roughly every 10-20 seconds. User networks moving in vehicles within a city should already re-evaluate / scan for WLAN networks every 1-2 seconds. A user network moving beyond 50 km/h should disable WLAN due to a too high scanning frequency. In order to save battery resources and to avoid unnecessary load in surrounding WLAN networks, a user network should determine autonomously when a validation of WLAN networks is reasonable, i.e. whenever it moves 20 m. For this it is desirable that a user network can determine its velocity, e.g. based on GPS positioning, and adapt its scanning and discovery procedures accordingly.

Table 5-3 : Scanning intervals depending on user velocity.

User velocity [km/h]	802.11a /b/g	
	Avergage distance **d [m]**	Scanning interval **t [s]**
1	20.0	**72.0**
3	20.0	**24.0**
6	20.0	**12.0**
30	20.0	**2.4**
50	20.0	**1.4**
80	20.0	**0.9**
120	20.0	**0.6**
150	20.0	**0.5**
180	20.0	**0.4**

5.4.6 Conclusion of Evaluation

We have investigated the discovery, connectivity setup, advertisement and attachment (CSAA) in a WLAN scenario. Two different CSAA procedures have been evaluated: *independent CSAA* performs connectivity setup independently from access network advertisement and attachment; *integrated CSAA* combines all these functions within a common procedure. We have developed models, to investigate the signalling overhead and delay of these procedures.

Although the *integrated CSAA* scheme increases the size of the WLAN beacon by 88%, the total amount of data transmitted during the CSAA procedure is reduced by 14%. The delay for performing CSAA can be separated into two phases. The discovery phase denotes the time to determine which WLAN networks exist by scanning the WLAN channels. It does not dependent on the CSAA type; however, it depends on the number of WLAN channels that need to be scanned. For the 2.4 GHz (IEEE 802.11b/g) there are 13 channels available in Europe, of which three are non-overlapping. The discovery phase also depends on the scanning mode used by the mobile device. In passive scanning mode the discovery delay for scanning 1, 3 and 13 channels takes approximately 100 ms, 300 ms and 1300 ms respectively. By active scanning this delay can be reduced by approximately 50%. The second phase of the delay is caused by the CSAA signalling and it depends on the load within the WLAN cell. The signalling delay for *independent CSAA* ranges from 24 ms for 1 active user in the cell to 270 ms for 20 active users in the cell. For *integrated CSAA* the signalling ranges from 17 ms for 1 active user to 128 ms for 20 active users. *Integrated CSAA* can thus reduce the signalling delay in the range of 30%-40% at very low load; at higher load (more than 6 users) the gain approaches 50%.

An important question relating to network advertisement and attachment is how frequently a user is expected to search for other networks. We have defined that whenever the data rate of the WLAN link would change by a factor of two a mobile user network should scan for alternative networks. Based on typical radio propagation assumptions (but neglecting shadow fading) we derived that for 802.11a/b/g a user network should scan for available WLAN networks approximately every 20 m. The scanning frequency thus depends on the velocity of the user. For nomadic user (with velocity up to 1 km/h) a user network should perform WLAN discovery approximately once per minute. At pedestrian velocity (i.e. 6 km/h) the user

network should perform WLAN discovery around every 10 s. At vehicular velocities (i.e. 30-50 km/h) WLAN discovery should be performed every 1-2 s.

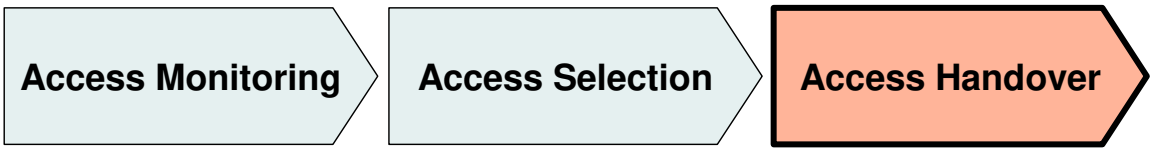
5.5 Summary

Access monitoring is the process of discovering and monitoring access properties and access network characteristics. It is required to obtain relevant information to determine a basis of candidate accesses for access selection. This information must be comparable between different access technologies. We have defined generic access abstractions that are derived from access technology specific metrics. Generic abstractions define the performance of different accesses, as well as the resource situation. Generic performance abstractions describe the suitability of an access with respect to the requirements of a service data flow. These requirements include data rate, delay and transmission reliability. The generic resource abstractions determine at what load level an access is operated and to what extent a new service data flow impacts the load level.

In addition to the performance of the access itself, also the capabilities of an access network determine whether it is suitable for a user network. We have determined which network characteristics are important for determining whether an access network is suited as candidate access network. These characteristics include network policies and security capabilities, supported services, networking capabilities, and support for multi-access management. Providing network and access information of an access network to a user network we denote as *access advertisements*. We have described different options of how access advertisements can be provided to the user network. Before a user network can make use of an access network it has to attach to it. Based on previous work we have developed an *ambient network attachment protocol* (ANAP) that performs network attachment and sets up a security association at the same time. Since some access advertisement information elements require that a certain security and trust level is established, access advertisement depends on network attachment. We have investigated different options on how access discovery, connectivity setup, access advertisements and network attachment (CSAA) can be combined. An integrated approach combines the functionality into a common procedure, whereas an independent approach performs the different procedures in independent procedures. We have evaluated the overhead and the delay of the two approaches in a WLAN scenario and have demonstrated that an integrated approach of connectivity setup, access advertisements and attachment can improve the performance significantly.

Chapter 6. Access Handover

6.1 Introduction



Access handover is the process to switch a service data flow from one access to another one. Handover between radio cells is already a key feature of today's mobile communication systems. It is caused by user mobility: a user leaves the coverage area of one cell and enters another cell. Neighbouring radio cells typically overlap only to a small amount in order to allow smooth handover. In multi-access networks there is a geographic overlay of radio cells of different RATs; multiple radio cells are available at the same location. The access technology and radio cell is selected dynamically for each user to improve user performance and to increase system capacity. Consequently, (inter-system) access handover is not primarily the consequence of a terminal leaving the cell coverage; it is rather the outcome of an optimisation process – *access selection* – which allocates users to RATs based on system parameters like cell load, link performance and transmission costs. In this chapter we discuss how an access handover can be realised and investigate its performance. In Section 6.2 we introduce the objectives and requirements, followed by related work. Section 6.4 discusses the communication context that exists for a communication session in the network and the user terminal; and which is affected by access handover. In Section 6.5 we describe different options of how access handover can be realised between different access technologies. The general approach is based on the concept of a *generic link layer*[47]; the alternative solutions can be grouped into solutions based on a *multi-radio generic link layer* and those based on *generic link layer interworking*. In Section 6.6 we will investigate the performance of several access handover schemes[48].

6.2 Objective and Requirements

The objective of access handover is to enforce an access selection from one access technology to another one in a timely and efficient manner. Two types of requirements can be defined for access handover. Firstly, the performance of access handover must be sufficient. It is expected that access handover can support session continuity without significant degradation of perceived service quality. If the performance degradation of access handover is too large, any gain that can be potentially achieved with access selection cannot be properly exploited. Secondly, the effort and complexity of applying an access handover scheme for different access technologies must be manageable.

---

[47] The work on a generic link layer concept has been influenced by many discussions, in particular with Henning Wiemann, Michael Meyer, Gabor Fodor, Johan Lundsjö, Mathias Cramby, Mikael Prytz, as well as several people from the *multi-access* workpackage within *Ambient Networks*.

[48] The performance evaluation has been largely performed by Birinder Singh Khurana. It is based on [Khu05].

Performance Requirements

Access handover causes an access handover delay and / or data distortion. The access handover delay is the time from the moment that the access handover decision is made (i.e. the access selection decision) until the moment when the first data is transmitted via the new access. Data distortion is the amount of distortion that is caused to the service data flow by the access handover. This distortion can be caused by loss of data, re-ordering of data or duplication of data. The impact of data distortion on the service performance depends on the service requirements. For example, some applications can cope with data loss and interruption times, while others are more sensitive. In addition, a bearer used by a service uses a transport protocol, like the *transmission control procotol* (TCP) or the *user datagram protocol* (UDP). The performance of these transport protocols can be adversely affected by data distortion of the data flow. Transport protocols can thereby either amplify or conceal data distortion for the application.

Deployment Requirements

Access handover schemes need to be applicable to existing and new access technologies that are expected to be integrated into multi-access networks. The effort to migrate existing architectures and the associated amount of standardisation has to remain feasible. Furthermore, access handover schemes should be applicable in different business and architecture scenarios of multi-access (cf. Section 4.4).

6.3 Related Work

Handovers are performed with mobility management protocols that redirect the communication path in the network. The mobility protocol used in mobile radio networks for packet-switched transmission is the *GPRS tunnelling protocol* [3GPP29.060] [3GPP23.060]. Another series of mobility management protocols for IP services are developed by the IETF. These comprise *mobile IP* [RFC3344] [RFC3775], *hierarchical mobile IP* [RFC4140], IP mobility for IPsec [RFC4555] and network mobility [RFC3963], as well as network based versions like *proxy mobile IP* [RFC5213] [ID-NETLMM]. Further extensions allow data to be transmitted simultaneously via multiple access systems to *multihomed* user networks [ID-MCOA] [ID-NOMAD] [ID-NOMAD6] [ID-POLIM] [FKGBT04] [FUGZ03] [KFKTG03]. Several additions have been proposed to reduce the delay and data loss at handover for IP mobility schemes. [PW99] proposes to buffer data in the network in mobility agents and forward the buffered data along with new data after a new mobile IP location registration. Some solutions add multicast transmission to the old and new access point during the handover procedures [FW00] [FKW03] [ID-BMIP] [BA04] [FKW07], others (denoted as *fast mobile IP*) forward data from the old to the new access router [RFC4988] [RFC4068]; *seamless mobile IP* [HGS03] proposes a combination of data forwarding and bicasting. In contrast to our work, data forwarding only considers IP packet queues and not data that is already in process of being transmitted by the link layer of the radio access system. *Context transfer*[49] has been proposed for IP mobility schemes in order to setup a communication

[49] Most work considers the re-location of data between radio access points as not being part of context transfer and refers to it instead as *data forwarding* or *data transfer*. We later use the term context transfer more widely: we define the user-plane data of a service data flow as part of its communication context that can be included in the context transfer. Data forwarding thus becomes a part of context transfer.

context already prior to the handover in a target access system [RFC3374] [RFC4066] [RFC4067], similar solutions have been identified for WLAN systems [IEEE802.11f] [HSK05]. It has been proposed in [FG00a] [FKG01] [APFPG03] [AGP04] [ID-L2ABST] [IEEE802.21] to trigger a handover procedure based on link layer information in order to reduce handover delay; this is similar to the *generic link layer* reports (based on generic access abstractions) described in Chapter 5, which lead to an access selection decision of *multi-radio resource management* that may result in a handover execution. The 3GPP handover procedure of *SRNS relocation* integrates context transfer, data forwarding and handover triggering [3GPP23.060] [3GPP25.936] [SO08], bicasting has originally also been considered [3GPP25.936]. A framework with a mobility toolbox that allows multiple mobility protocols to be combined is proposed in [AN D20B2] [SAEG07]; similarly 3GPP is specifying a variety of mobility management protocols in its evolved system architecture [SO08]. A comparison of different IP mobility schemes is made in [HS03] [YW06]. An overview of standardisation activities around mobility management is provided in [EWKKW06] and specifically for 3GPP networks in [SO08]. In contrast to previous work we focus in this work on the communication context that exists in an access network at different locations; this includes the data buffered in IP packet queues, as well as context contained in link layer functions. We investigate the interaction of different handover procedures on this communication context and develop new methods to reduce the impact of the handover on service performance.

TCP is used as transport layer protocol for reliable Internet services; approximately 90% of Internet traffic is based on TCP according to [IM04]. TCP performs window-based end-to-end congestion control to ensure that the total amount of traffic does not overload network nodes. It is well known that the achievable TCP throughput is limited by the amount of packet loss even for infinite link capacities [MSM97]. The transmission characteristics of radio access networks can interact with TCP congestion control such that the available radio capacities cannot be fully utilised by end-to-end services. In particular in wireless networks with large bandwidth-delay-product[50] TCP is slow in recovering from congestion events or adapting to changes in the pipe capacity of the network [SM01] [DZ01] [SRBW01] [CR02] [MSH03] [RFC3819]. Delay spikes or excessive packet re-ordering, e.g. at a handover, can also lead to spurious TCP timeouts or TCP fast retransmissions [LK00] [SRBW01] [Gur01] [HFW02] [HF03] [GL03]. Packet losses that can occur in radio access networks due to transmission errors or handover losses are interpreted by TCP as congestion indication and result in performance degradation [CI95] [BSW95] [PW99] [FG00b] [SRBW01]. Different approaches address these problems: by changing TCP to identify and recover spurious timeouts or retransmissions [LK00] [Gur01] [GL03], applying proxy solutions for TCP transmission [BSW95] [SRBW01] [MSH03], adaptively regulating the transmission for TCP acknowledgements [CR02] or introducing cross-layer indications from mobile IP to TCP to temporarily disable TCP timer calculations [FG00b]. Other approaches propose to make the handover procedure more reliable and faster by adding data forwarding, multicasting, context transfer and/or handover triggering to mobility management protocols, as described above. The handover performance for TCP is investigated e.g. in [PW99] [HFW02] [HF03] [HS03]. In this work we also develop methods for improving the access handover and thus provide improved performance for services based on TCP or other transport protocols. In particular, we develop novel link layer functions for optimizing the performance of access handover; we evaluate those for TCP-based services in different scenarios.

[50] The product of end-to-end round-trip time and bottleneck data rate of the transmission path.

6.4 Communication Context Management at Access Handover

For the transfer of a service data flow an associated communication context is established and maintained in several network nodes. When the access for a service data flow is changed at access handover, the service data flow subsequently follows a different communication path. The communication context associated with the service data flow has to be either re-established within nodes of the new communication path, or some context can be transferred from nodes of the old communication path to the corresponding new nodes on the new path. Within this work we are mainly concerned with the context maintained in the access system. In this section we first describe the communication context associated with a service data flow; this is followed by a discussion about communication context management at access handover.

6.4.1 Communication Context

The communication context in the access domain can be categorised into control-plane context and user-plane context, as shown in Figure 6.1 for a link layer transmitter.

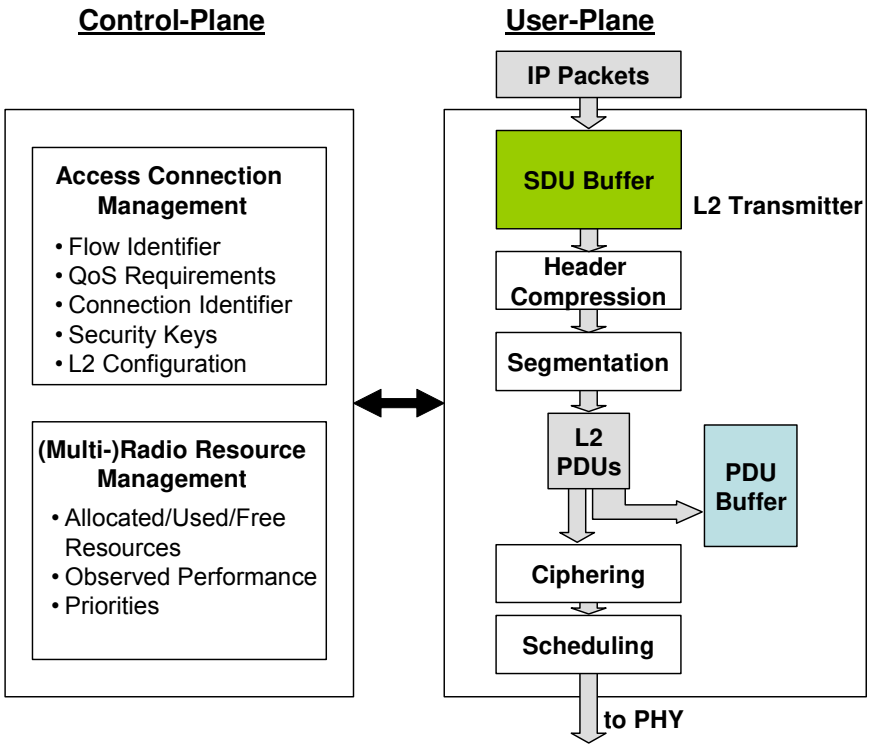


Figure 6.1: Communication context in an access node

The control-plane functions configure the user-plane functions and monitor the transmission performance. *Access connection management* maintains the flow identifier of a data flow and the corresponding service requirements (e.g. QoS). Furthermore it maintains the identifier of the access connection, to which the flow is mapped. It stores the security keys and link layer configuration parameters. These can be parameters for active queue management, header compression, segmentation sizes, and ARQ parameters and timers. *Radio resource management* monitors the used and available resources, the performance achieved by the

access connection for every flow, and priorities between flows. For the setup of the control-plane context, an interaction is required with AAA and policy control functions.

The user-plane functions receive IP packets that are stored in the *service data unit* (SDU) *buffer*. After *header compression* they are segmented into link layer *packet data units* (PDU) and obtain a sequence number[51]. Transmitted PDUs are stored in the *PDU buffer* for retransmission (ARQ), *ciphered* and *scheduled* for transmission by the physical layer. For the SDU buffer active queue management counters and parameters are maintained. Header compression contains a separate compression context. For the PDU buffer an ARQ state is maintained about the successful and unsuccessful transmitted PDUs, as well as different retransmission timers. The PDU buffers at the transmitter and receiver are depicted in Figure 6.2; they contain the corresponding ARQ windows. The lower edge of the transmission window marks the PDU with the lowest sequence number that has been transmitted but for which no sequence number has been received (denoted as VT(A) in [3GPP25.322]); when a status report is received from the receiver the lower edge is moved forward to the first PDU that is not positively acknowledged. Negatively acknowledged PDUs are retransmitted starting with the one at the lower window edge. The upper edge of the transmission window is referring to the highest transmitted PDU (denoted as VT(S) in [3GPP25.322]); it is moved forward with every newly transmitted PDU. The highest PDU that can be sent is bound by the maximum transmission window size, which is limited according to the sequence number space (see e.g. [Tan96]). The transmitter ARQ window between VT(A) and VT(S) can contain PDUs that are either acknowledged or negatively acknowledged, as well as PDUs that have been transmitted but for which no receiver feedback has yet been received. At the receiver, the lower edge of the ARQ window is marked by the first missing PDU that has not yet been received correctly (denoted as VR(R) in [3GPP25.322]). Whenever the PDU with the sequence number equal to VR(R) is received the lower window edge is advanced to the first missing PDU. If possible, SDUs are reassembled and delivered to the higher layer and the PDUs are purged from the PDU buffer. Note that the receiver can still have some PDUs in the buffer that are below the lower window edge; this is the case when PDUs are still missing for a SDU to be reassembled. The upper edge of the receive window is marked by the PDU with the highest sequence number that has been received (denoted as VR(H) in [3GPP25.322]). All sequence numbers that are received outside the maximum receive window (i.e. above VR(R) + receive window size) are discarded by the receiver. The receiver transmits status reports to the transmitter; status reports are triggered either by a request from the transmitter or according to receiver specific mechanisms (see e.g. [3GPP25.322]). Whenever a status report is sent it reports over the sequence number range from VR(R) to VR(H); all PDUs below VR(R) are automatically acknowledged since VR(R) acts as cumulative acknowledgement. As shown in Figure 6.2, the ARQ windows of the transmitter and the receiver are not fully synchronised. The transmitter window is larger: at the lower edge this is caused by data that is already correctly received by the receiver, for which however no status report has been received yet at the sender. At the upper edge it is due to data that has already been sent by the transmitter, but which has not yet been correctly received by the receiver.

[51] The sequence number space is limited by the length of the sequence number field. A modulo operation is applied to the sequence numbers so that the next higher sequence number after the maximum sequence number becomes zero. When we refer to “higher” or “lower” sequence numbers this refers to the sequence number order prior to the modulo operation.

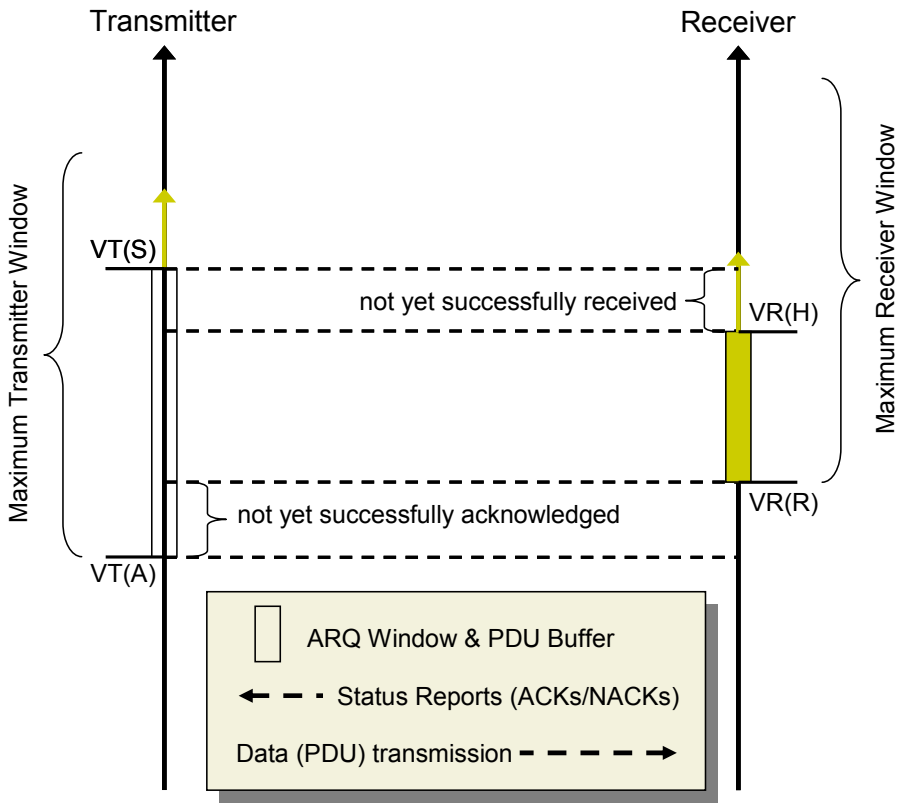


Figure 6.2: ARQ windows and PDU buffers at link layer transmitter and receiver.

6.4.2 Context Management Procedures

The performance of access handover is determined by the access handover delay and the amount of distortion that it causes for the service.

Access handover consists of one mandatory and two optional parts. The optional parts provide optimisation of the access handover by transferring (parts of) the communication context from the old to the new access system (see Table 6-1).

Update of forwarding state (mandatory)

The mandatory functionality of access handover is to update the forwarding state in the corresponding forwarding points (FPs). Thereby the traffic of a service data flow, or at least parts of it, is redirected to the new access system.

Access handover preparation (optional)

In order to accelerate the access handover procedure and increase the success probability of the access handover, the access handover can be prepared prior to updating the forwarding state. Access handover preparation can include the establishment of a communication context in the new access system. This can include assigning connection identifiers, establishing a security context for authentication, ciphering and integrity protection, establishing policy and QoS rules, authorise resource usage, and possibly reserve resources for the new connection. The communication context of the source system may be transferred and used as basis for the new context. The access handover preparation by means of context transfer allows to faster

establish a communication context in the target access and thereby reduce the access handover delay, as indicated in Table 6-1. It also allows to detect that the access handover will fail (e.g. due to lack of resources in the target system) before the access handover takes place; the access handover decision can then still be abandoned.

Access handover optimisation (optional)

The role of access handover optimisation is to avoid or minimise the distortion of the service. Distortion of the service is, on one hand, caused by a perceivable access handover delay; on the other hand it is caused by a distortion of the data stream. The distortion can be caused by loss, re-ordering or duplication of data. Ways to reduce the distortion are context transfer procedures for data forwarding, or bi-casting of data from the forwarding point to both the old and the new radio access points. This also allows to reduce the access handover delay since data is quickly available for transmission at the new radio access points. Access handover optimisation is the key functionality of access handover investigated in this work.

Table 6-1 : Access handover improvements.

	Access Handover Delay	Data Distortion
Access Handover Preparation	Pre-establishment of connection state	-
Access Handover Optimisation	Accelerate access handover by data forwarding.	Reduce data loss and possibly also data duplication and re-ordering.

6.5 Methods for Access Handover

6.5.1 Generic Link Layer Concept

An access handover is performed between different radio access connections. The *generic link layer* is a concept that shall facilitate the management of and the interworking between different access systems, and support efficient and seamless access handover. The objective is to define generic parts of the transmission functions for different access systems, and common interfaces to multi-access management functions, as shown in Figure 6.3. Functionality of different link layers is very similar; this similarity can be exploited to derive common functionality for different RATs. For example, link layer functionality that is similar for multiple RATs is:

- Queue management and queuing of higher layer datagrams,
- Data compression, e.g. of higher layer headers or payload,
- Segmentation and reassembly of higher layer datagrams,
- Error recovery by retransmission of erroneous data (ARQ),
- Ciphering of data,

- QoS monitoring of the link performance,
- Prioritisation among flows for the same user.

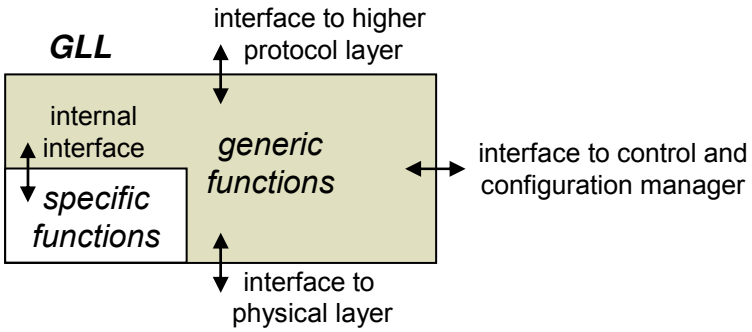


Figure 6.3: *Generic link layer* interfaces.

Some access systems may still require access-specific functions. These need to be integrated with the generic part of the generic link layer via appropriate interfaces as shown in Figure 6.3.

Different functions of the generic link layer are depicted in Figure 6.4. Depending on the configuration, functions can be either access-generic or access-specific.

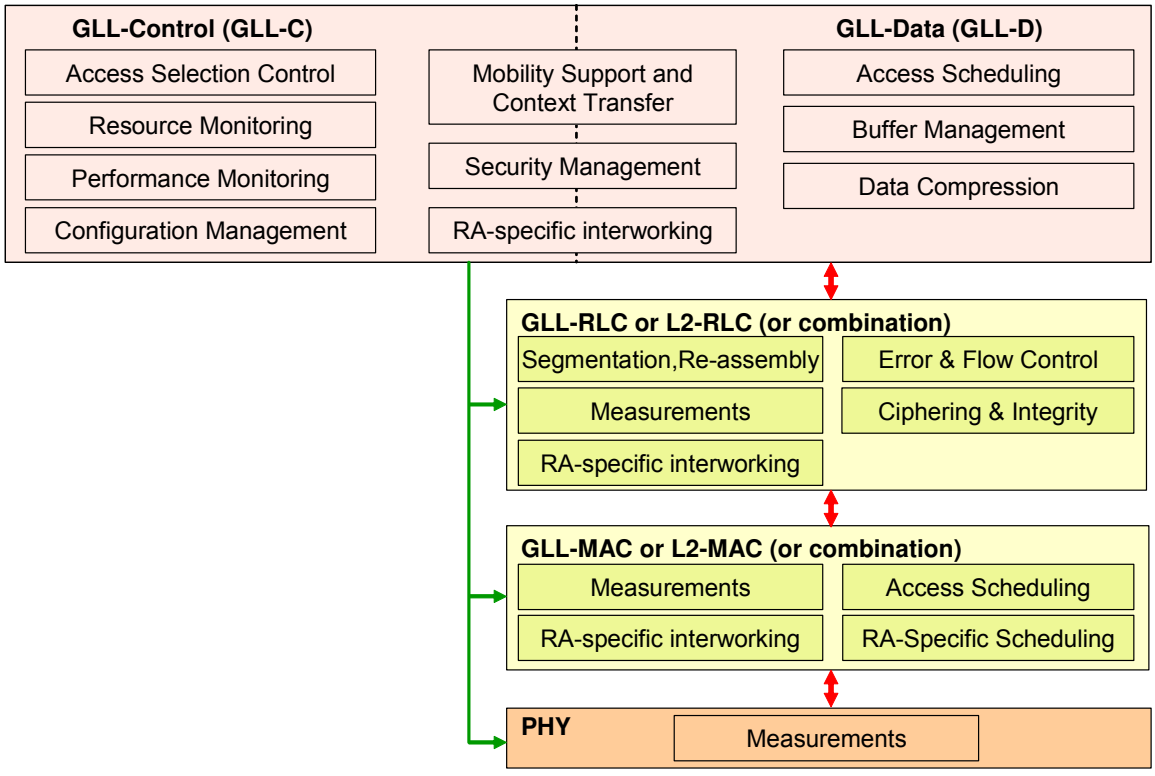


Figure 6.4: Functional composition of the generic link layer.

We have identified different relevant realisations of the generic link layer corresponding to different levels of integration as shown in Figure 6.5. Towards the left in the figure the

amount of common functionality provided by the GLL for different RATs increases. At the same time the complexity increases. Towards the right in the figure an increasing amount of access specific functions are provided independently by different radio specific link layers, and the GLL provides in the extreme only a common interface.

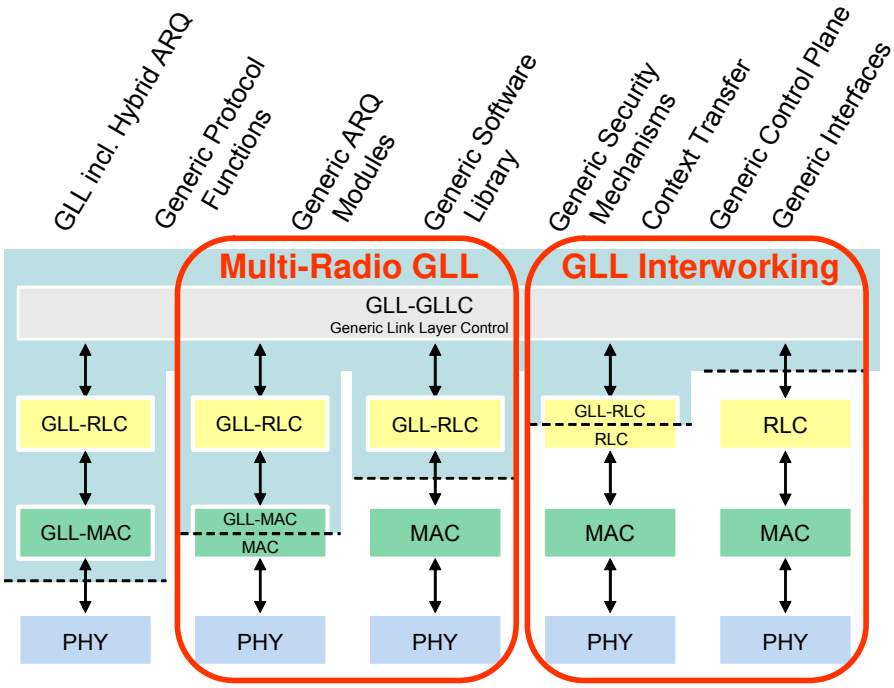


Figure 6.5: Options for a *generic link layer*.

We investigate two realisations of GLL in detail: a *multi-radio generic link layer* (MR-GLL) that tightly integrates different access technologies, and *generic link layer interworking* (GLL-IW) that enables efficient coordination between different RATs. The different realisations of GLL lead to a number of different access handover optimisation schemes. An overview of the optimisation schemes is depicted in Figure 6.6 and is discussed in the following sections.

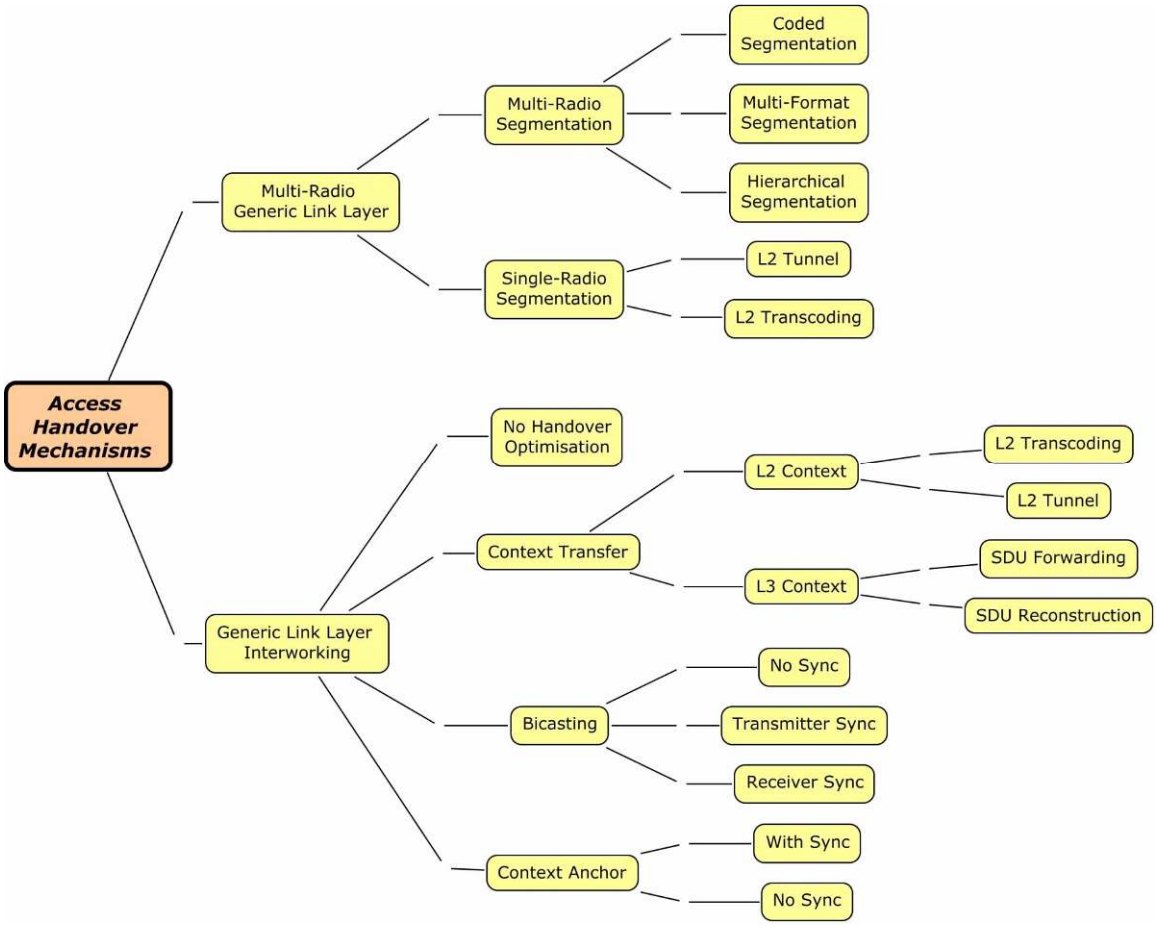


Figure 6.6: Access handover mechanisms for different options of *generic link layer* realisation.

6.5.2 Multi-Radio Generic Link Layer

6.5.2.1 Design Principle and Functionality

The principle of the multi-radio generic link layer is to provide as much common link layer functionality for different radio access technologies as feasible. It shall maintain common link layer operation while changing between different accesses. The communication context is maintained during the access handover, as shown in Figure 6.7. This prohibits packet duplication or loss at access handover. Even simultaneous transmission via multiple access technologies can be supported. All changes of underlying access technology are managed by mere reconfiguration of the link layer functions. The multi-radio generic link layer is a stepping stone towards a software-defined, software-reconfigurable radio realisation. It allows smooth reconfiguration of link layer functions during ongoing data sessions.

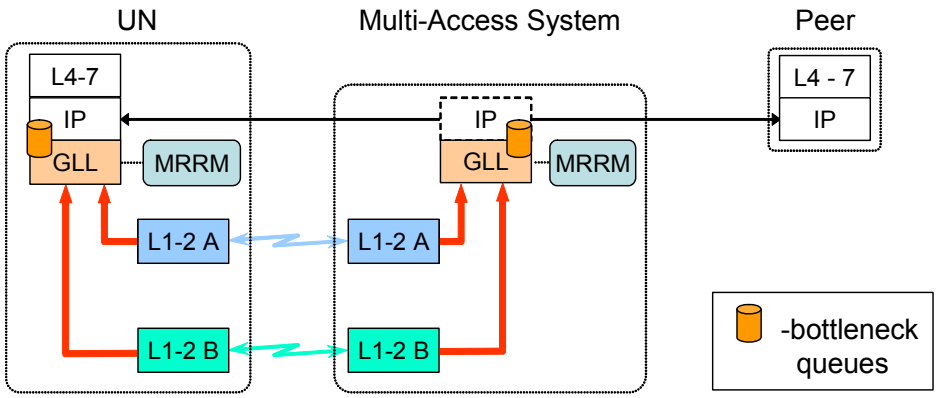


Figure 6.7: Multi-Access based on a *multi-radio generic link layer*.

The benefit of a multi-radio generic link layer is a re-use of existing functionality for different access technologies. Most functions of today's radio link layers are similar, so a complete new development has to redefine standard functionality. To give an example, a function performed by each radio link layer is automatic repeat request (ARQ) to recover from transmission errors by retransmitting erroneous data blocks. However different flavours of ARQ are used in various radio link layers, which differ only in details although the same functionality is performed. The radio link control protocol of UTRAN has generalised the ARQ functionality by specifying an ARQ toolbox, which can be configured in various ways [3GPP25.322][52]. This allows configuring the UTRAN RLC ARQ function very similarly to ARQ functions used in other link layer protocols. This kind of generalisation should be expanded to other link layer functions to obtain a generic link layer toolbox. The toolbox shall include generic methods for buffer management, priority handling and scheduling, data segmentation and also link security handling. Such a multi-radio generic link layer eases the evolution of access technologies, and reduces design costs and standardisation effort in development of new radio technologies. Only limited new functionality needs to be developed and included for new access technologies. The un-interrupted use of common link layer entities during access handover avoids the need for context transfer. It also avoids the necessity of new interfaces and procedures to other networking functions. For example, existing procedures can be re-used without any changes in the further networking infrastructure, e.g., for authentication, authorisation, charging, QoS and policy management, and mobility management. Furthermore, a common interface to (multi-)radio resource management and configuration management functions is used for different access technologies.

In order to deploy the generic link layer in a particular radio access network, it has to be configured to the underlying link by setting protocol parameters and timers accordingly. A suitable configuration of the generic link layer toolbox has to be determined by a configuration management function. Some link parameters like e.g. the link round-trip time can also be measured by the generic link layer and some protocol parameters and timers can

[52] In 1999 two alternative solutions for RLC were proposed in the 3GPP standardisation, which basically performed the same functionality. As none of the proposals could be proven to outperform the other one no decision could be derived. This deadlock could be overcome when we proposed the RLC toolbox. It was accepted instead of the other two alternative link layer proposals. The flexibility and configurability of the RLC toolbox enable it to behave similar as the other proposals by appropriate configuration.

be adopted accordingly. The only functionality required by the generic link layer in addition to standard link layer functions, is support for lossless reconfiguration at access handover.

There may remain specific radio link layer functions, which are only applicable in the context of a particular radio physical layer. It may not be feasible to extent the generalisation of link layer functions to those specific cases. This can include e.g. new ciphering algorithms, or functions related to unequal error protection that may be supported only by some radio access technologies. Also scheduling mechanism should remain radio technology specific; they depend on how the channel resources are partitioned in the time, frequency, code or space domains, which is very specific to each radio access technology. Access-specific functions need to be embedded via an internal interface, and activated depending on the selected access technology.

Access handover can imply that the node containing the multi-radio generic link layer entities is changed. This can happen, when the access handover is initiated by user mobility into a different area[53]. In this case, a re-location of the generic link layer entity is required. The procedure for this GLL re-location is similar to handover procedures applied within certain RATs (i.e. intra-system access handover). The multi-radio generic link layer makes the same method applicable, even if the access handover implies a change of access technology (i.e. inter-system access handover). The steps used in the re-location procedure are, as shown in Figure 6.8:

- At access handover the existing GLL communication context is transferred to the new point of execution. The new point of execution is the node in which the GLL is located after access handover.

- The generic link layer entities are reconfigured according to the new radio access technology. This may imply that new segmentation sizes are used according to the slot formats of the new radio access technology. The GLL receiver maintains the old communication context. This is required to reconstruct higher layer datagrams that have not been completely transmitted before access handover.

- The outstanding part of higher layer data from the old transmission context is transmitted from the reconfigured GLL transmitter via the new radio connection. The receiver reconstructs the data from the old receiver context.

- After all outstanding data of the old GLL transmission context has been delivered, the transmission continues in normal operation via the new radio connection.

- This reconfiguration procedure enables lossless inter-system access handover while minimizing the amount of data transmitted over the radio link.

[53] Note, that this relocation can be avoided in case of maintaining a link layer anchor and redirecting the radio access connection to a remote radio access point. This approach is used in UTRAN where the link layer remains located in the *serving RNC* even if the radio access point (NodeB) is controlled by a *drift RNC*.

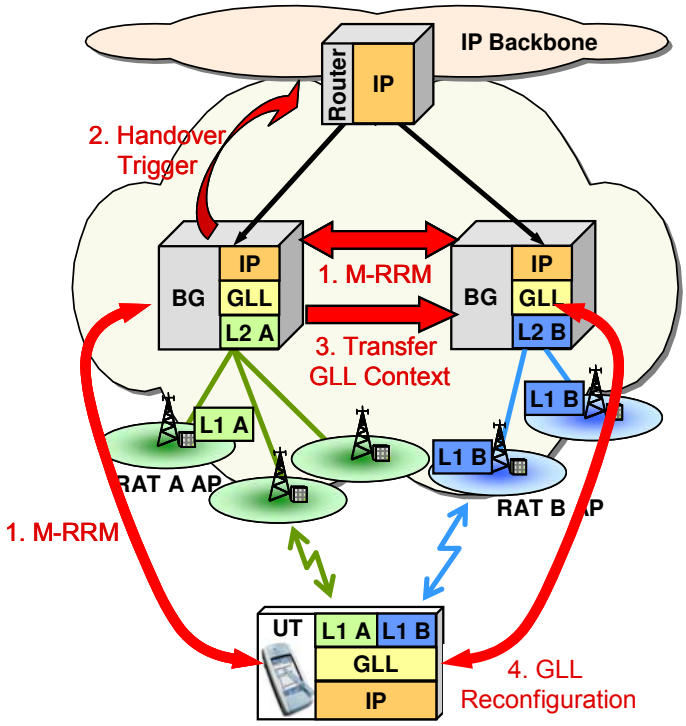


Figure 6.8: Re-location of the *multi-radio generic link layer* context at access handover.

The communication context described in the above procedure contains all GLL protocol data containing higher layer data, which has not been transmitted successfully when the access handover occurs. For the GLL transmitter this includes all un-segmented data and all segmented data blocks not yet acknowledged by the receiver. For the GLL receiver the transmission context contains all data blocks which have been received but could not be used for reconstruction of higher layer datagrams because of still missing data blocks. The header compression context also needs to be maintained in order to allow proper header reconstruction.

One key requirement for a multi-radio generic link layer is to provide flexible segmentation formats and sequence numbering schemes. This enables the segmentation and the ARQ process to be adapted to the transport blocks provided by different access technologies according to the specific resource partitioning schemes. A static segmentation scheme, as e.g. provided by UTRAN RLC [3GPP25.322] cannot easily adapt the segmentation sizes which are best suited for the underlying radio technology. The combination of static segmentation with a fixed sequence number space also limits the link layer ARQ procedure in adapting to radio link characteristics like round-trip time and data rate. With the addition of HSPA to UTRAN, including dynamic access handover between HSPA and 3G radio transmission, this has led to the situation that the segmentation formats cannot be chosen to be suited for both transmission modes. The transmission performance can become limited by the ARQ behaviour due to stalling of the ARQ window, thus the available data rate of HSPA cannot be exploited. This has been described by [IHITU+04].

We have developed different methods of flexible segmentation for the multi-radio generic link layer, as shown in Figure 6.9. These methods can be grouped into *multi-radio segmentation*, which allows simultaneous transmission via multiple access technologies, and *single-radio segmentation*, enabling sequentially switching between access technologies.

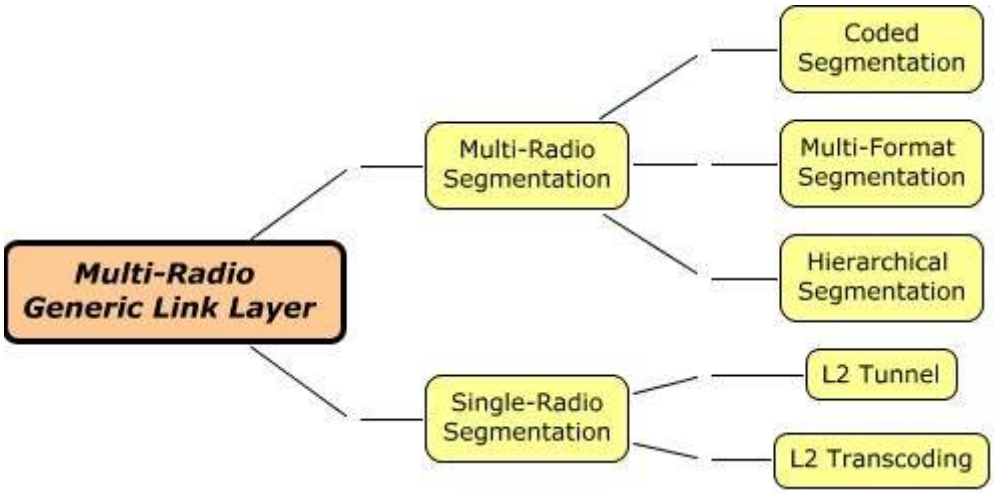


Figure 6.9: Access handover schemes for the *multi-radio generic link layer*.

6.5.2.2 Generic Link Layer with Multi-Radio Segmentation

A generic link layer with multi-radio segmentation uses common segmentation for different access technologies. As shown in Figure 6.10, the entire GLL functionality above the access-specific MAC functionality is shared for the two access technologies A and B. Access scheduling between the different RATs is performed in the GLL-MAC layer. Optionally, access-specific security algorithms can be used in the GLL-RLC layer, allowing for a separation of the security domains of different RATs, as proposed in [AN D7-1].

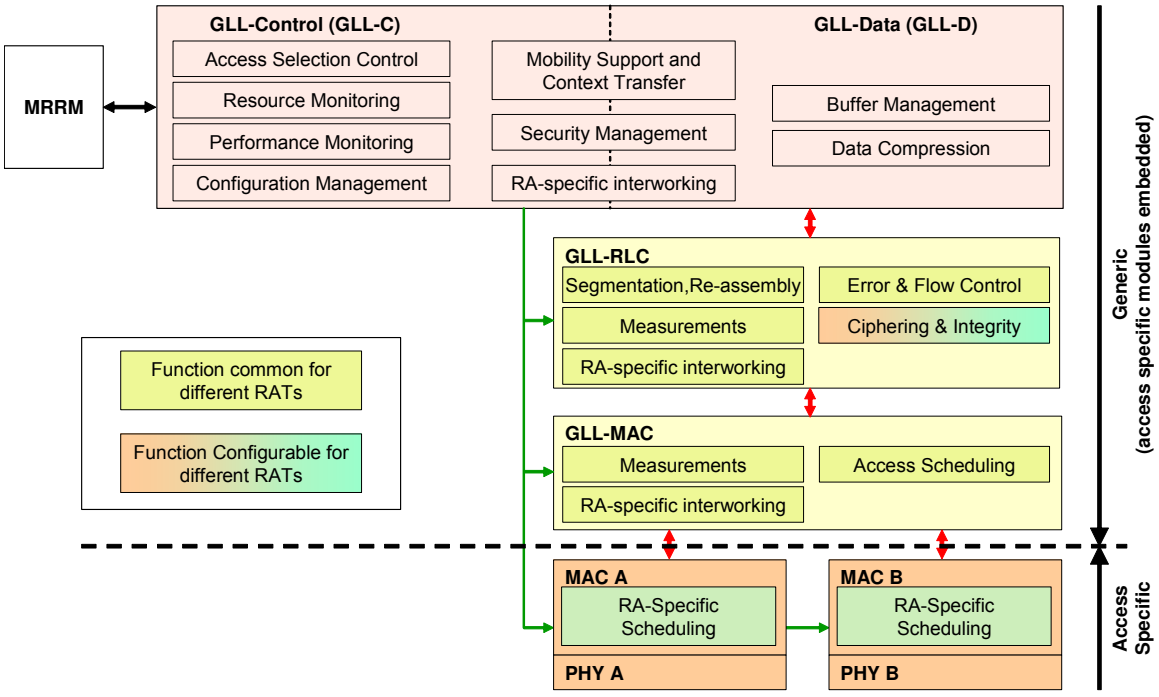


Figure 6.10: Functional model of the *generic link layer* with *multi-radio segmentation*.

Generic Segmentation and Multi-Radio ARQ

One challenge for a reconfigurable multi-radio generic link layer is to adapt the data transmission to the specific transmission modes of the underlying radio access technologies. A key role plays the segmentation of data. Data segmentation has to perform two tasks. Firstly, it has to partition data into protocol data units of a size that is suited for the transmission format of the radio link. This size is typically imposed by the physical time slot structure and the coding and modulation scheme of the access technology. Secondly, every segmented data unit needs to be assigned a sequence number. This sequence number is a unique identifier that is required, to reassemble at the receiver the original SDUs out of received PDUs. It is also used to enable selective retransmission of PDUs that have been lost or corrupted during transmission. For the MR-GLL the format of the radio link can change dynamically according to the radio access technology that is used. A problem arises if PDUs need to be transmitted in a transmission format, which differs from the transmission format for which the PDUs have been originally segmented and for which a sequence number has been assigned. We refer to this functionality as *multi-radio ARQ* [AN R2-4]. A typical example is when a PDU has been segmented and transmitted according to a first transmission format, but the PDU is corrupted during transmission. At a later time the PDU is scheduled for retransmission, while the transmission format has already changed.

Generic Link Layer Protocol Formats

For the transmission of protocol information between peer protocol entities certain message formats of the packet data units need to be defined. Typical protocol messages contain several fields, of which some have a generic and others have a radio specific meaning. Examples for generic fields are those related to segmentation and reassembly of higher layer datagrams, while specific fields depend e.g. on the radio access specific addressing schemes. A common protocol field in a PDU is e.g. a sequence number, however the required length of the sequence number field depends on the rate of the wireless link and the segmentation size. Therefore protocol fields need to be specified in a generic way such that they can be configured in a specific way (e.g. length of fields). Some specific fields also can be added with a meaning only within a certain radio access technology.

The specific configuration of protocol fields needs to be defined at the establishment or reconfiguration of a pair of link layer entities. The configuration can be either clarified in a connection set-up procedure or alternatively by an according link layer configuration from the control and configuration management functions within the radio access network.

6.5.2.2.1 Flexible Segmentation Functions

We have defined three different segmentation functions for the multi-radio generic link layer (cf. Figure 6.9), which fulfil the requirements of flexible segmentation.

Coded Segmentation

The principle of *coded segmentation* is depicted in Figure 6.11. Data SDUs are segmented into a first PDU format, which is ideally suited for one of the underlying RATs. When (re-)transmissions are performed in one RAT with a transport block size incompatible with the original PDU size, the PDUs need to be re-formatted to match the new transport block size. These re-formatted PDUs are denoted as PDU' (see Figure 6.11).

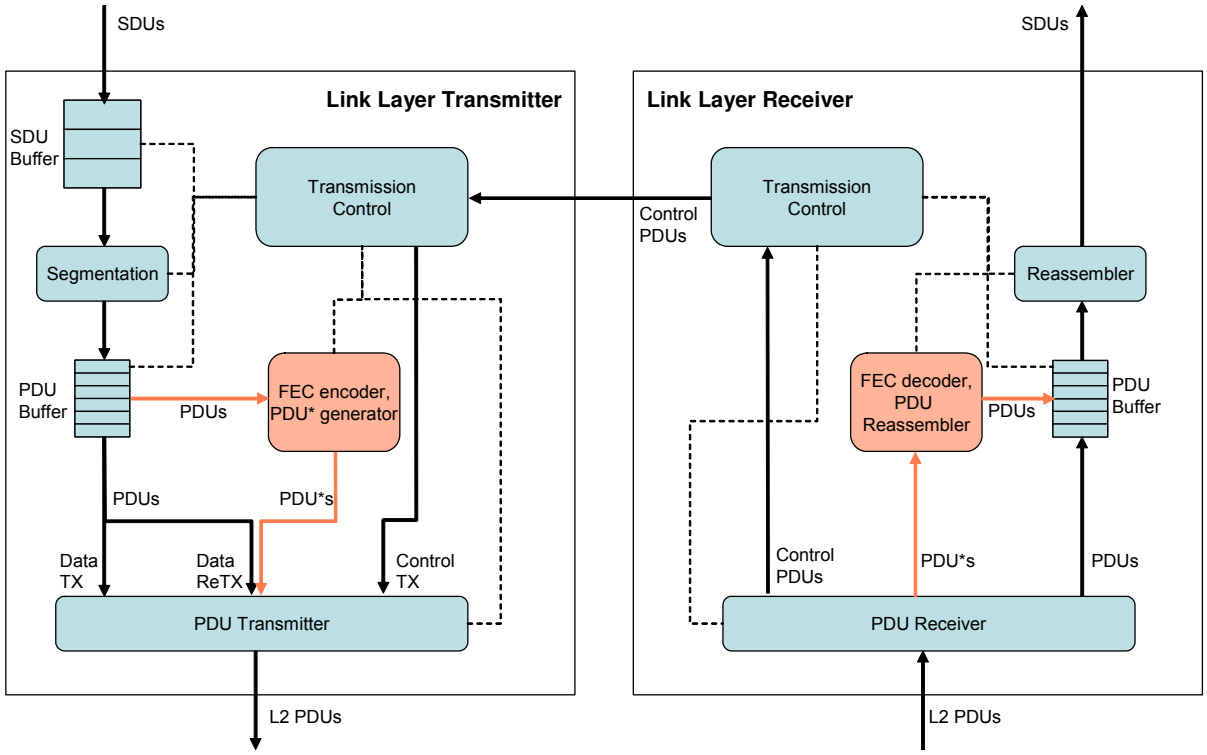


Figure 6.11: Coded segmentation.

For the generation of the PDU's we propose a forward-error correction coding (FEC) based on rateless erasure-codes (also known as fountain codes[54] [Mac05]). The procedure is depicted in Figure 6.12. The original PDU is encoded into a large byte stream consisting of a multitude of encoding units. From this byte stream different PDU's can be generated in a size as desired by the current access technology. The receiver is able to reconstruct the original PDU as soon as it has received enough encoding units out of the encoded PDU. It is not significant which encoding units are included in PDU'. In order for the receiver to be able to reconstruct the original PDU, it needs the following information together with the a PDU' derived from PDU: the FEC codec algorithm (marked as (1) in Figure 6.12), the sequence number of the original PDU (marked as (2)), the length of the PDU (marked as (3)), the pointer to the encoding unit of the encoded PDU stream (denoted as *enc(PDU)*) that is included in PDU' (marked as (4)), the length of PDU' (marked as (5)). The codec is typically known to the receiver, as well as the size of the received PDU'. The PDU size can also be known in case of fixed size segmentation. In addition to the sequence number of a PDU, two to three parameters need to be included in the PDU' header as overhead, to enable the reconstruction of the original PDU. For coded segmentation, the first segmentation from SDU to PDU is not required; the SDU could be directly used for generating encoded PDUs. This would be the preferred choice when the generic link layer is simultaneously transmitting via multiple RATs. If RATs are used sequentially, the encoding scheme would only be required at access handover, or when a RAT is used with a different segmentation size.

[54] Fountain codes are e.g. LT codes [Lub02], raptor codes [Sho06] or online codes [MM03]. They are used to transfer data over networks with packet losses, in particular for multicast transmission [BLMR98] [3GPP26.346] [ETSI472] [RFC3453] [RFC5053] or transmission with unidirectional transport [RFC3926].

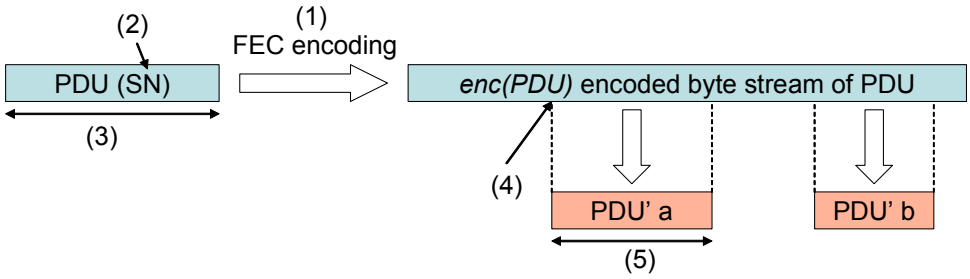


Figure 6.12: PDU generation by forward error correction coding of PDU with sequence number SN.

Hierarchical Segmentation

In the case of a hierarchical segmentation, the segmentation scheme can be adapted to the transport block sizes of the access technology, as depicted in Figure 6.13.

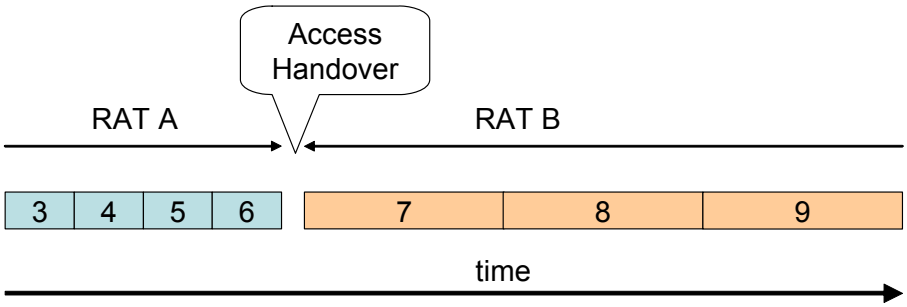


Figure 6.13: Flexible segmentation scheme.

A problem only occurs in case of ARQ, when data has been corrupted and the access handover occurs before the retransmission. In this case, a re-segmentation is not possible after access handover. Hierarchical segmentation enables the capability to embed PDUs segmented in one format in the transmission of another format. This is shown in Figure 6.14 and in Figure 6.15: in case of Figure 6.14 two corrupted PDUs 4, 5 that are segmented according to RAT A, are embedded in a single PDU with frame format according to RAT B for retransmission. In case of Figure 6.15 a corrupted PDU 9 of type B is fragmented, and retransmitted in three separate PDUs of type A. Thus a retransmission of PDUs after access handover is possible, by concatenation and fragmentation PDUs into a *retransmission block*. A retransmission block can thus contain multiple PDUs or parts of PDUs. This requires a modified header for the retransmission block, as depicted in Figure 6.16. In order to allow status reporting also for correctly received or corrupted PDU fragments after retransmission, an extended status report that can report on fragments of PDUs is required, as shown in Figure 6.17. This type of re-segmentation has recently been adopted for 3GPP LTE when the radio link quality and data rate decrease.

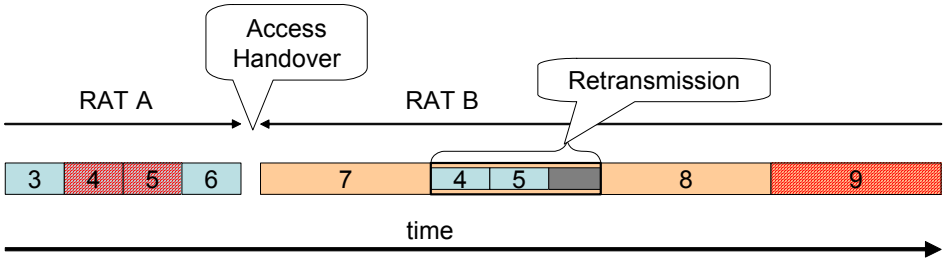


Figure 6.14: Hierarchical segmentation with concatenated (a) and fragmented (b) transmission.

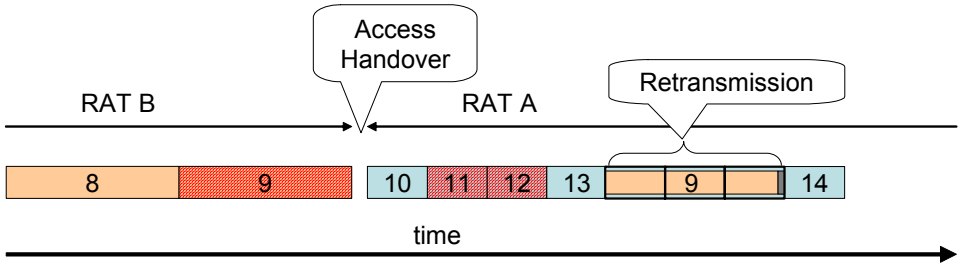


Figure 6.15: Hierarchical segmentation with concatenated (a) and fragmented (b) transmission.

Basic Header	Sequence Number	Other	F	E
Fragmentation Header	Fragment Size			
Concatenation Header	PDU Offset			

Figure 6.16: Extended PDU header for retransmissions with *hierarchical segmentation*.

Status Report	Standard status report (e.g. bitmap)		E
Status Extension	Seq. Number	Ack/Nack	E
	First byte	Last byte	

Figure 6.17: Extended status report for *hierarchical segmentation*.

Multi-Format Segmentation

Another solution for segmentation, denoted as *multi-format segmentation*, is to segment SDUs multiple times, into separate PDUs of different format for all supported RATs. These are stored in separate PDU buffers, as shown in Figure 6.18.

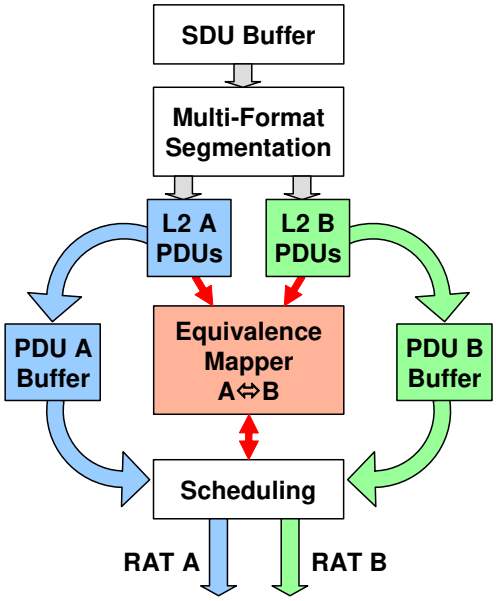


Figure 6.18: ***Multi-format segmentation*.**

In addition, an equivalence mapping is maintained, which describes how data contained in PDUs of one format corresponds to data contained in PDUs of the other format. This is depicted in Figure 6.19. If a retransmission of PDUs of type A is required after an access handover to a RAT of type B, the equivalent PDUs of type B corresponding to those of type A will be retransmitted. For example, let us consider PDUs 7-8 are not received correctly before an access handover from RAT A to RAT B. It will then be sufficient to retransmit the equivalent PDU 3 of type B instead, which contains the same original data.

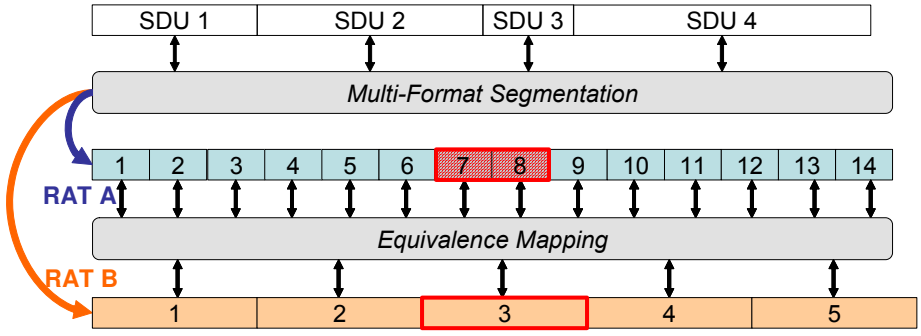


Figure 6.19: Equivalence mapping for ***multi-format segmentation*.**

6.5.2.2.2 Access Selection and Access Handover

The approach of access selection with a MR-GLL extends the concept of channel switching used in UTRAN. In UTRAN, different transport channel types are defined [3GPP25.301] [SWW00], as shown in Figure 6.20. The random access channel (RACH) and forward access channel (FACH) do not require the reservation of physical channel resources. They are used for terminals with low transmission activity. The dedicated channel (DCH) uses dedicated channel resources and provides QoS support. The downlink shared channel (DSCH) provides link layer data flows of several users to be transmitted via a common physical channel;

additionally each user has dedicated channels for radio resource management. In HSPA, users also share a common channel; in addition *hybrid ARQ* (HARQ) is used for better error recovery, and link adaptation for dynamic resource management and high peak rates. For an ongoing link layer connection, the underlying transport channel can be dynamically changed. For example, when moving from a radio cell with HSPA capabilities into a cell without HSPA, a dedicated channel is used instead. In principle it is even possible to transmit a link connection via multiple transport channels simultaneously. The radio protocol state and functionality of the PDCP, RLC and MAC-d layers remain operational, whenever the transport channel is changed, e.g. due to changes of cells or for radio resource management purposes. Also the SDU and PDU buffers are continuously used; no buffer re-location is required. In short, no context transfer is required, since all relevant context remains at the same point of execution.

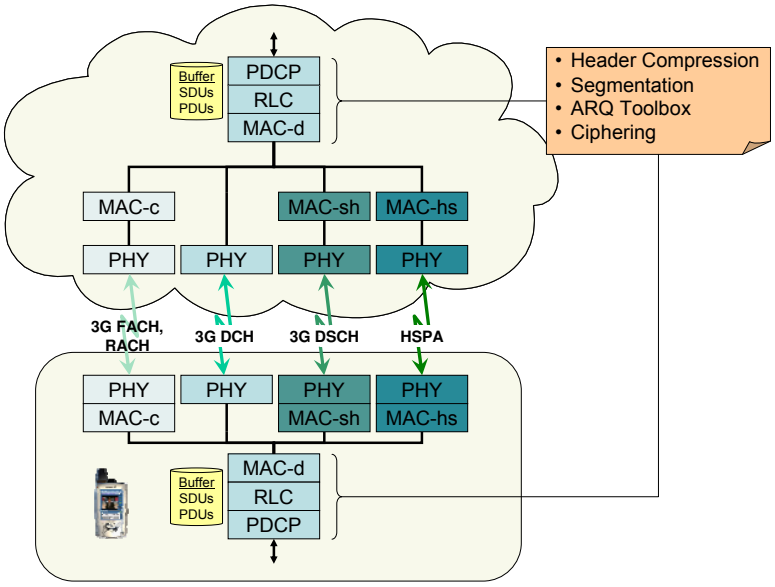


Figure 6.20: UTRAN link layer for multiple transport channels.

The MR-GLL applies the same principle as in UTRAN for the multi-access context, as shown in Figure 6.21. The GLL performs all protocol functions for different types of underlying radio channels. This can be either different UTRAN transport channels or completely different radio technologies.

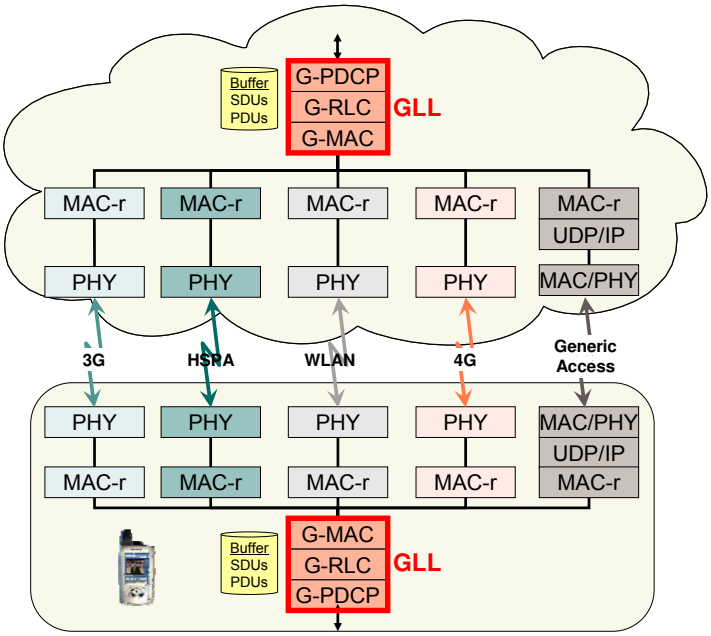


Figure 6.21: *Multi-radio GLL* for several radio technologies.

The functionality of the MR-GLL extends the functionality of the UTRAN link layer. In order to cope with the characteristics of the different access technologies it provides flexible segmentation as presented in Section 6.5.2.2. It also adapts the multi-radio ARQ functionality to the different link characteristics, like adapting ARQ protocol timers to different round-trip times. The RAT-specific MAC entities (MAC-r) perform the scheduling according to the radio transmission properties and transmission timing of the corresponding physical layer. This scheduling can be based on radio-specific characteristics, e.g. depending on instantaneous radio link quality (i.e. channel-dependent scheduling). Thus dynamic access selection happens on short time-scales in an asynchronous manner; each MAC-r distributes resources according to its own timing. A second level of access selection is controlled by MRRM on a lower time scale. MRRM interacts with RAT-specific radio resource management (RRM) to monitor radio resources for the different RATs and the overall link performance. It defines the RATs that can be used by a terminal, by defining the *GLL active set*[55], that is, the MAC-r entities that are bound to the GLL-MAC entity for that terminal. Note, that there can be a multitude of GLL-PDCP and GLL-RLC entities for each user, and there can be multiple users. For every user there is only on GLL-MAC entity that can be bound to multiple MAC-r entities, depending on which RATs are available for each terminal. Also RATs that do not support the MAC-r functionality (e.g. legacy RATs) can be included into this architecture, as shown on the right of Figure 6.21. This *generic access* can be integrated, by defining a generic MAC-r that connects a generic RAT by means of UDP/IP tunnelling, similar to 3GPP generic access network [3GPP43.318].

[55] The *GLL active set* is a subset of the *MRRM active access set* described in section 3.3. MRRM assigns the *GLL active set* to a GLL entity [AN R2-4] [AN D2-4]. The scheduling of data to any access from the *GLL active set* is performed autonomously by GLL without explicit control of MRRM.

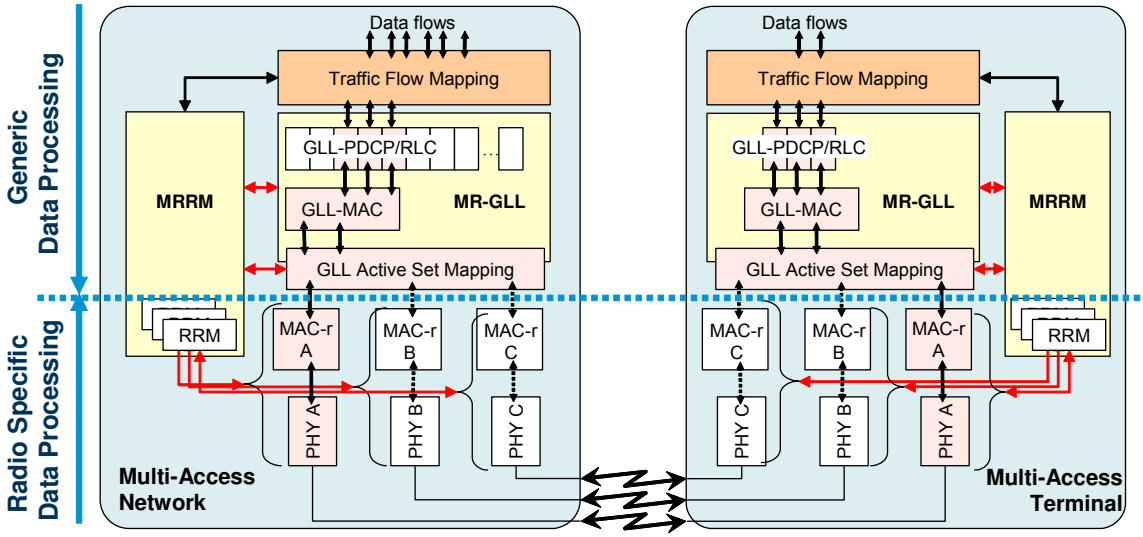


Figure 6.22: Transmission via *multi-radio GLL* between a multi-access network and one multi-access terminal.

The transmission between a network and a terminal with MR-GLL is shown in Figure 6.22. The interaction between the different entities is depicted in Figure 6.23. A MAC-r determines the available data rate for each user, and sends buffer status requests to all GLL-MAC entities bound to it. The buffer status requests also contain segmentation sizes, which may depend on the physical layer data rate. These requests are forwarded to GLL-RLC entities which respond with their demand. When the demand is forwarded back to MAC-r it also contains QoS attributes (like priorities), according to which MAC-r decides to which GLL-MAC resources are assigned. GLL-MAC prioritises among the different GLL-RLC entities associated to it for the user. In the next step data is transmitted. The slower access selection is performed by MRRM, which controls which MAC-r entities are bound to which GLL-MAC entities, e.g. based on the load in the access systems.

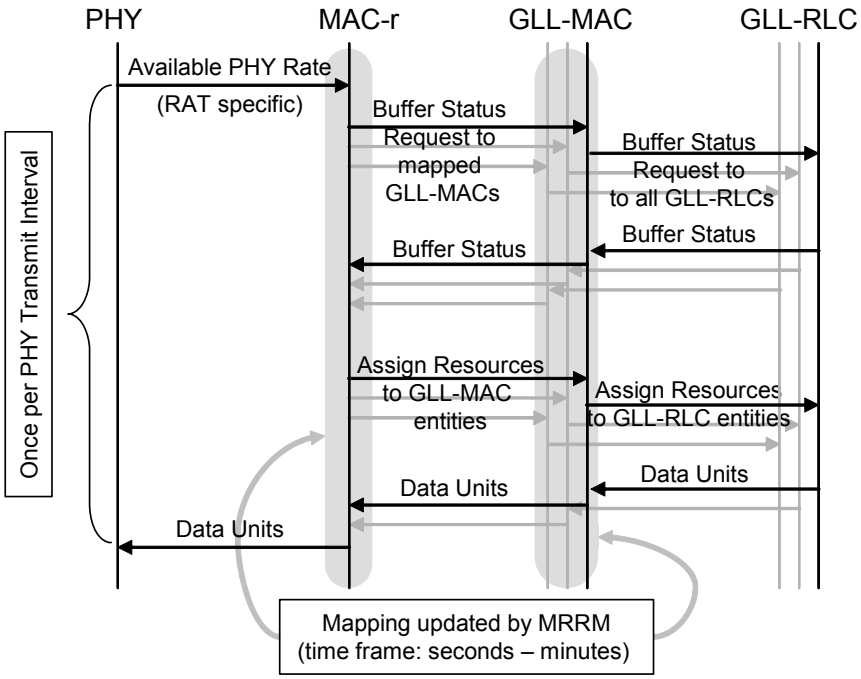


Figure 6.23: Multi-access scheduling sequence for *multi-radio generic link layer*.

6.5.2.3 Generic Link Layer with Single-Radio Segmentation

An alternative to a generic *multi-radio segmentation* is to maintain access-specific segmentation functions and introduce a reconfiguration for the access handover. We denote this as *single-radio segmentation*. As shown in Figure 6.24, access specific segmentation and security schemes are used. In addition, a reconfiguration function between the different access-specific functions supports seamless access handover. Single-radio segmentation is useful when the transmission of a service data flow uses only one access at a time. At those times, only the access-specific segmentation schemes are used. Merely during the access handover, the reconfiguration functionality is invoked.

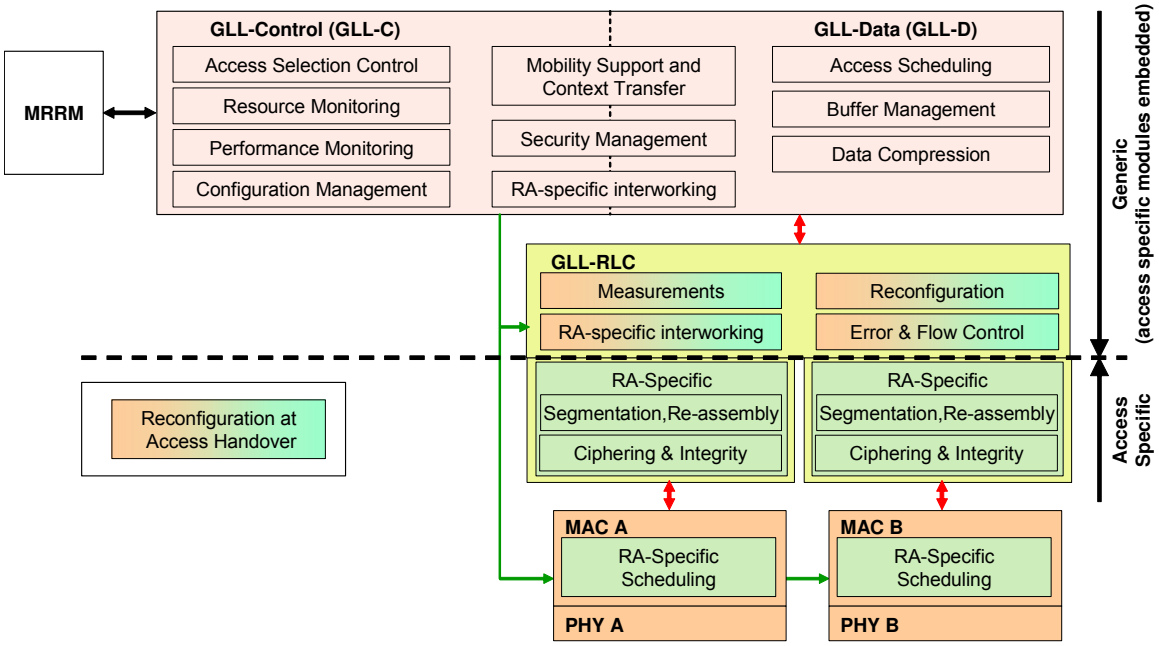


Figure 6.24: Functional model of the *multi-radio generic link layer* for single-radio segmentation.

6.5.2.3.1 Flexible Segmentation Interworking Functions

If the communication context shall be maintained during an access handover, it is required that the access-specific state is converted for the new access. We have developed two different conversion schemes.

Layer 2 Tunnel

Layer 2 Tunnel is a lossless context conversion procedure, which involves both the transmitter as well as the receiver, as depicted in Figure 6.25. The complete SDU buffer is transferred to the new transmission function. In addition, PDUs from the old transmission function are moved to the new transmitter; however, only those PDUs that have not been acknowledged by the receiving peer via the ARQ process are transferred. So only the necessary part of the PDU buffer is forwarded. These PDUs are included in a context container, which is transferred to the new link layer entity. This container is transmitted like any other SDU to the receiver side. The new link layer receiver forwards the context container to the old link layer receiver entity. There the old PDUs are extracted and added to the receiver PDU buffer. Now the PDUs can be reassembled to SDUs which are delivered to the higher layer. In short, the context transfer procedures at sender and receiver establish a virtual tunnel from the old transmitter link layer entity to the old receiver link layer entity – via the new link layer connection. The encapsulation of data during the access handover is shown in Figure 6.26. Before the access handover, IP packets are transmitted via the old link layer and physical layer. During the access handover reconfiguration old layer 2 (L2) PDUs (containing IP packets) are transmitted within new L2 PDUs. After the reconfiguration IP packets are directly transmitted via the new link layer.

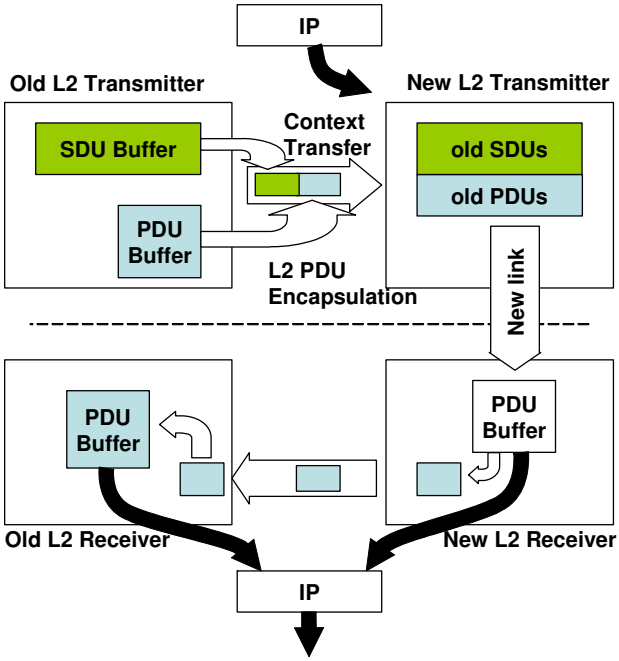


Figure 6.25: Generic link layer tunnelling (Layer 2 Tunnel).

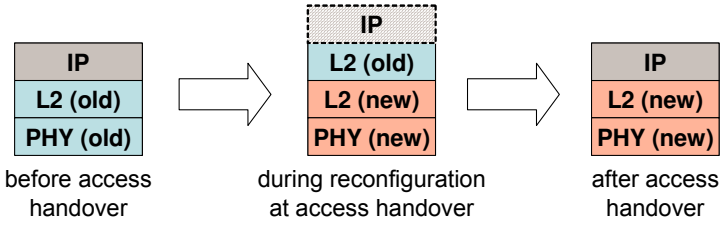


Figure 6.26: Encapsulation during access handover (Layer 2 Tunnel).

Layer 2 Transcoding

Layer 2 Transcoding is also a lossless context conversion procedure, which involves transmitter and receiver. As the segmentation formats of the old (A) and new (B) link layer are known, it is possible to transcode the segmented PDUs of format A into PDUs of format B, as shown in Figure 6.27. Similarly, the ARQ status is transcoded, so that it is known which of the new PDUs B have been received correctly (as PDUs of type A) and which not. This procedure needs to be performed at both the transmitter and receiver. Furthermore, it is required that the transcoding at sender and receiver are synchronised and use a common reference point.

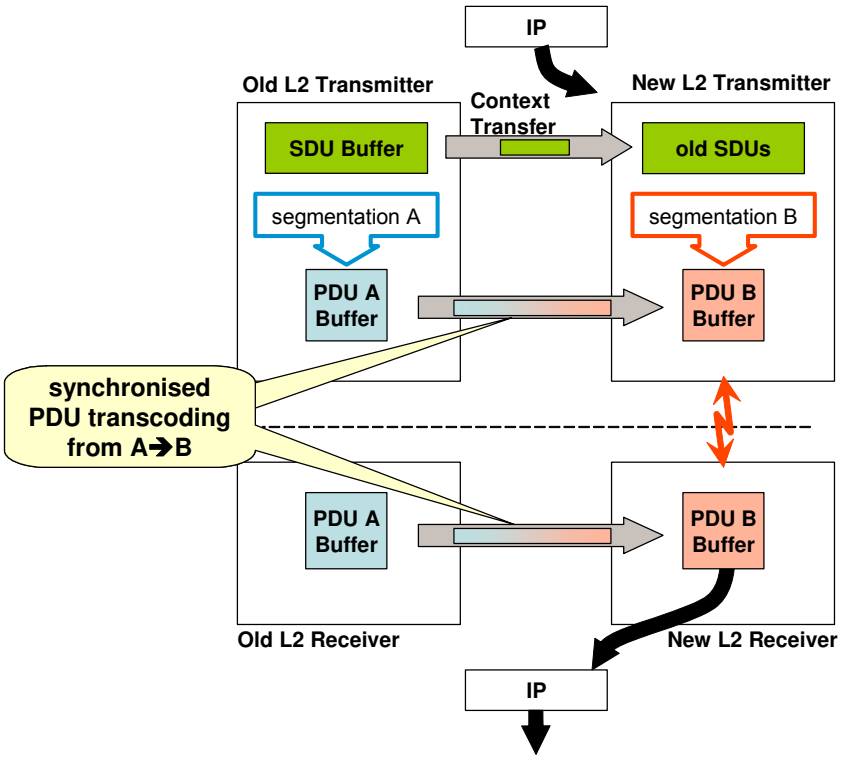


Figure 6.27: Layer 2 Transcoding from a segmentation format A to a segmentation format B at access handover.

6.5.2.3.2 Access Selection and Access Handover

The access handover procedure for MR-GLL with single-radio segmentation is depicted in Figure 6.28. MRRM gathers radio information of the different access technologies and commands MR-GLL to perform access handover. The link layer transmission contains a generic part for both accesses, which comprises the SDU buffer and possibly header compression state. The remaining link layer functions are access-specific, i.e. segmentation, ciphering and scheduling are performed according to the access technology specifications. In addition, a GLL reconfiguration function is used to support the access handover. The RLC communication context for one access is thereby converted to a corresponding RLC context of the other access. After the reconfiguration transmission continues according to the new access.

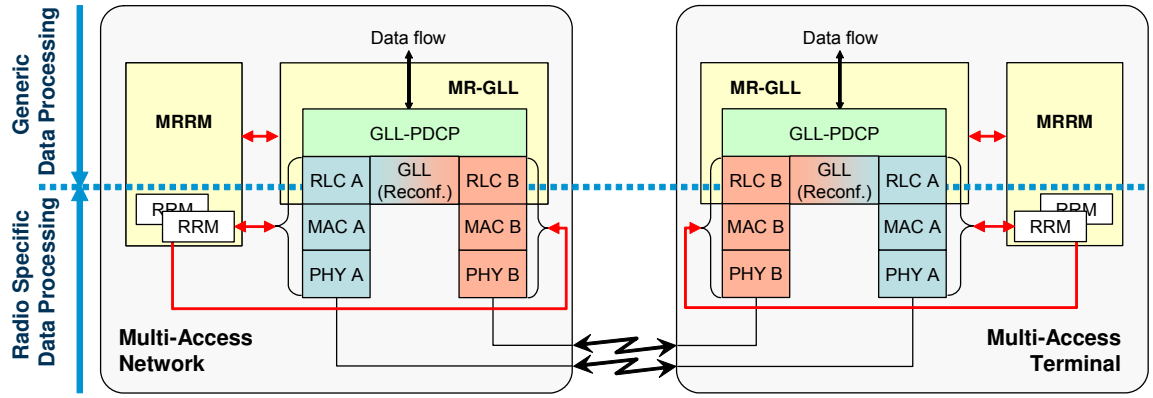


Figure 6.28: Access handover between two access technologies A and B with single-radio segmentation.

6.5.2.4 Discussion

The multi-radio generic link layer generalises the link layer functionality of different access technologies in a generic link layer toolbox, and thereby supports the transmission over different access technologies. The MR-GLL maintains a common communication context independent of the RAT in use, and thus enables seamless and efficient access handover. It allows a close integration of access selection functionality (performed by MRRM) and access handover. It is well suited for a centralised multi-radio access architecture, where a common anchor includes the MR-GLL and is connected to a multitude of radio access points of different RATs. Once a MR-GLL is in place, it will simplify the evolution of RATs. The flexibility of the MR-GLL toolbox will provide sufficient flexibility to be used for newly developed RATs. In case that new functionality was required, an extension of the GLL toolbox could be standardised. The MR-GLL will also reduce development costs, as all extensions can be based on a common GLL software package.

A disadvantage of the MR-GLL is that it requires adaptation of the link layer functionality of different RATs. This implies that the link layers of different RATs need to be replaced by the MR-GLL. This seems feasible from a technical point of view, since the GLL largely embeds in a generalised form functionality that is already available in all link layers[56]. However, from a practical point there are severe limitations. Firstly, a large amount of legacy radio equipment is already deployed and cannot be easily exchanged. This is in particular the case for mobile terminals, where it is outside the control of e.g. large operators, when end users upgrade their devices. If existing RATs are evolved to support the MR-GLL, it will be required that both MR-GLL are supported, as well as the legacy link layer protocols. Another practical difficulty is that different RATs are standardised in different standardisation fora. It will be required, that the standardised MR-GLL is compatible between such RATs. Consequently, an agreement and harmonisation of standardisation activities is required across different standardisation fora. This implies a major effort and very wide industry support. Another risk

[56] As shown for the generic access in Figure 6.21, legacy RATs can be integrated into the MR-GLL architecture by establishing tunnels over the legacy protocols. However, this is not desirable as a general solution. The problem is that the access-specific scheduling is split into the new MAC-r functionality and the legacy MAC. This separation makes it difficult for MAC-r to obtain sufficient information of the radio resource characteristics for efficient scheduling.

is that new RATs may pose new requirements onto the link layer, which are currently not foreseen. This reduces the gain of the MR-GLL concept. We anticipate, that new RAT requirements will only affect the access-specific MAC-r part of the GLL architecture; but this cannot be proven. A further disadvantage of the MR-GLL is that it puts some limitations on the possible business realisations. The MR-GLL assumes a tight, centralised integration of different RATs. This makes it unfeasible for a scenario, where different business entities operate different radio access networks. The MR-GLL imposes technical restrictions on the possibility of business "tussles" that are desirable (see Clark [CWSB02] [CWSB05]).

The disadvantages of the MR-GLL make it unsuitable as a general solution for multi-radio access. Still, the MR-GLL is feasible for limited number of specific RATs and scenarios. The MR-GLL approach has been adopted for the currently developed WINNER radio interface [PHDSP+06] [WPBS04] [Moh05]. WINNER defines different radio transmission modes with different characteristics. In order to allow interworking between these different modes efficiently, a link layer protocol architecture has been defined that includes configurable generic protocol functions, which can be extended with mode-specific functionality. Also in 3GPP some aspects of the MR-GLL are introduced into the UTRAN radio layer protocols. For the evolution of HSPA [3GPP25.899], the radio protocols need to be able to cope with an increasing range of transmission characteristics of the UTRAN transport channels. This requires an increasing flexibility of link layer segmentation function. Flexible link layer segmentation have been adopted in 3GPP for evolved HSPA [R2-070036] [R2-072766] [DPSB07] [PWSTG+07]. The MR-GLL is generally interesting for software-defined and software-reconfigurable radio (SDR/SRR) systems [WWRFa02] [WWRFb02].

6.5.3 Generic Link Layer Interworking

6.5.3.1 Design Principle and Functionality

The objective of generic link layer interworking (GLL-IW) is to provide interworking functionality of a link layer for access handover. The difference to MR-GLL with single-radio segmentation is, that GLL-IW does not modify and harmonise the link layer functions for different RATs. Instead, GLL-IW provides adaptation and interworking with generic networking functions (see Figure 6.29), like mobility management and security functions based on existing link layer functionality. GLL-IW supports mobility management or security functions by providing a communication context, which enables a more efficient access handover procedure. A security context of the old access is used to establish the security context faster at the new access. Similarly, the data context is maintained and transferred to reduce the amount of data distortion. We focus on the management of the data context. The principle of GLL-IW is to minimise the amount of modification and new functionality that needs to be provided by a specific link layer.

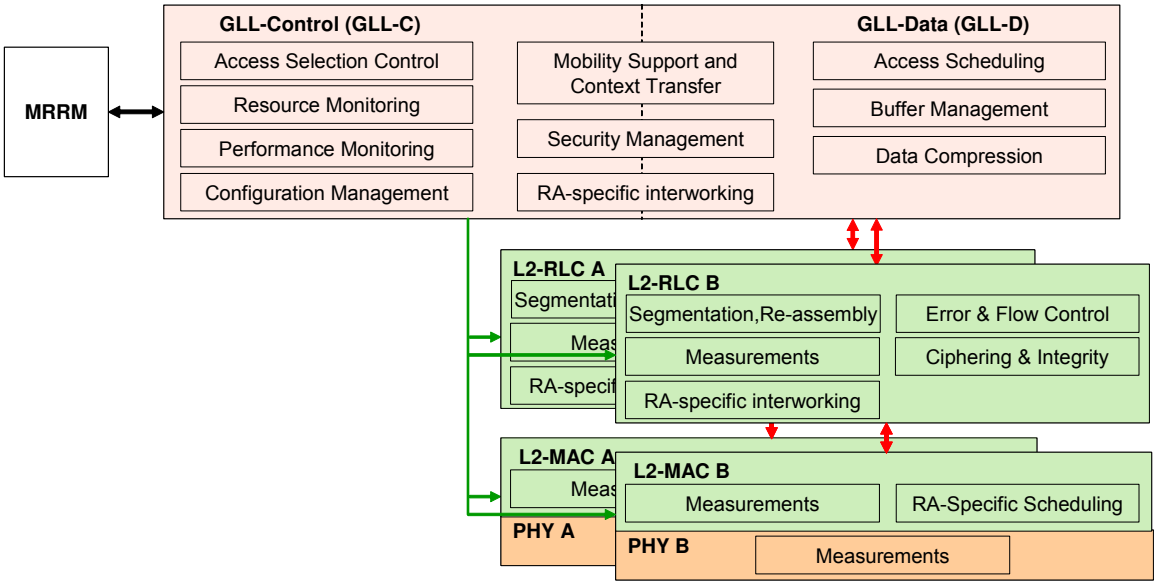


Figure 6.29: Functional model of *generic link layer interworking*.

We classify access handover with generic link layer interworking into three types, as shown in Figure 6.30. Access handover with context transfer moves part of the communication context from one link layer to the new link layer. Access handover with bicasting performs access handover by duplicating the data that is transmitted via multiple entities; GLL-IW is used to synchronise the different data streams. Access handover with a context anchor maintains a common communication context within a common anchor node; GLL-IW is used to update the context in the context anchor. In addition we consider a reference case without any interworking functionality for handover optimisation.

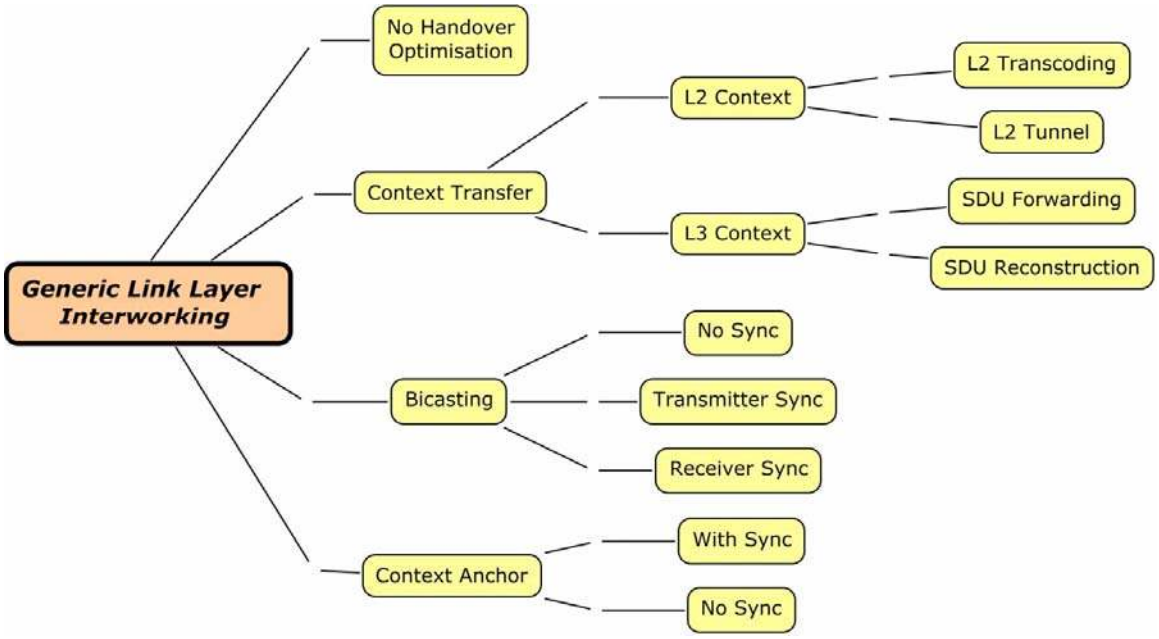


Figure 6.30: Access handover schemes for *generic link layer interworking*.

6.5.3.2 Access Handover without Access Handover Optimisation

In case that no forwarding of a data context occurs, the data buffered in the link layer of the old access is lost at access handover, as shown in Figure 6.31. We denote this case as *no context transfer*, and it is use as reference for the other mechanisms.

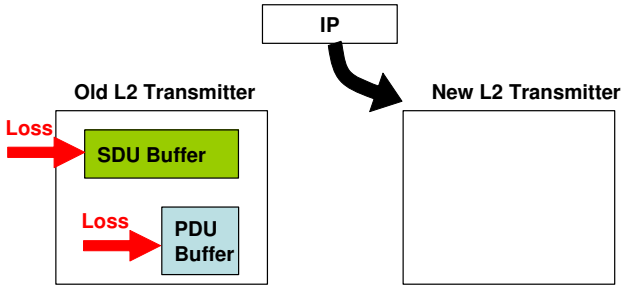


Figure 6.31: Data loss at access handover without forwarding of data context.

6.5.3.3 Access Handover with Context Transfer

In a context transfer procedure some communication context from the old access is transferred to the new access. Context transfer can be part of a mobility management protocol (e.g. SRNS relocation [3GPP25.832], or data forwarding in FMIP [RFC4068] [RFC4988]), or it can be a separate procedure (e.g. [RFC4066] [RFC4067]). The GLL-IW function has to support context transfer by building the appropriate context. Such a context can be a master security key, which is used in the new access to derive access-specific session keys; it can be an authorisation token for fast re-authentication at the new access; it can be a resource request to validate a candidate new access. During the access handover execution, the main task of

context transfer is to move remaining data from the old access to the new access, in order to reduce the amount of data distortion.

There are two alternatives for GLL-IW to construct a data context at the old access. First, the data context can comprise all SDUs (i.e. IP packets) buffered at the old link layer. This mechanism we denote as *SDU context transfer*. No particular processing of the SDUs is required, they are embedded in context datagrams[57] and then transferred to the new link layer entity. A consequence of this mechanism compared to *no context transfer* is, that the amount of data loss is reduced: only data in the PDU buffer is lost, and all data in the SDU buffer is forwarded.

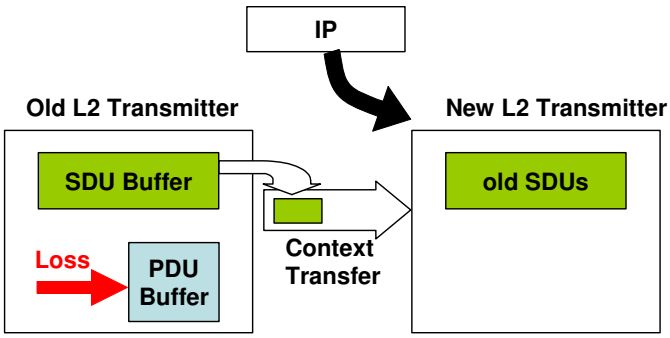


Figure 6.32: *SDU context transfer* at access handover.

An alternative context transfer mechanism, shortly referred to as *SDU Reconstruction*, extends the previous *SDU context transfer* mechanism. In order to make the access handover lossless, all data that is already segmented and stored in the PDU buffer is first reconstructed back to SDUs. The data context that is forwarded to the new link layer entity comprises both the reconstructed SDUs, as well as the SDUs from the SDU buffer. This mechanism is lossless, since every SDU that has not been confirmed to be successfully received via the link layer ARQ process is included in the data context. The mechanism is schematically depicted in Figure 6.33.

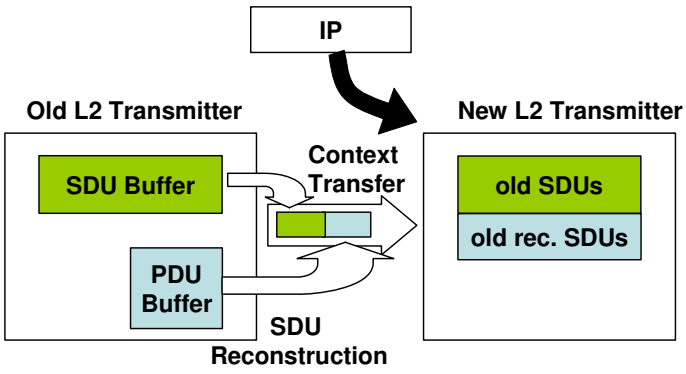


Figure 6.33: *SDU reconstruction* and context transfer at access handover.

[57] Context datagrams can be transported within a forwarding tunnel, e.g. IP-in-IP encapsulation, or within a specific context transfer protocol.

Both *SDU context transfer* and *SDU Reconstruction* we refer to as layer 3 (L3) context transfer (cf. Figure 6.30).

Context transfer can also include a layer 2 (L2) context, when it is combined with the multi-radio generic link layer as described in Section 6.5.2. This is the case when the old and new MR-GLL entities are located in different network nodes. For L2 context transfer, the relevant link layer state (buffer and/or ARQ state) is transferred, as indicated in the previous section in Figure 6.25 and Figure 6.27.

A schematic overview of context transfer is given in Figure 6.34. The access selection function controls during access handover if the forwarding point is directing traffic towards RAT A or RAT B. It triggers the context transfer function. GLL provides the context to be included in the context transfer procedure. The same procedure is performed for uplink and downlink transmission. In the downlink, context transfer is typically performed between different access nodes. For the uplink, both access technologies are either embedded into the same device, or they can be on different hardware devices of the user network. The main difference between uplink and downlink is that different delays are involved.

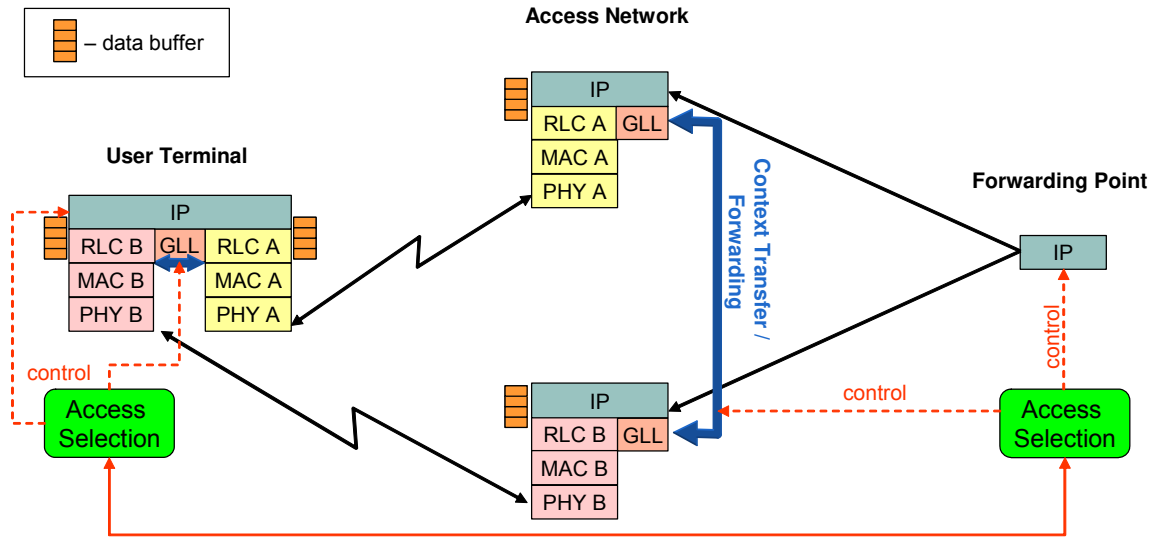


Figure 6.34: ***Generic link layer interworking*** **with context transfer.**

The performance of access handover depends of the timing of the different access handover steps. Figure 6.35 shows the transmission before access handover. Data is forwarded by the forwarding point to the active access link layer. There data is buffered and transmitted. Regular status reports from the receiver update the ARQ status and successfully transmitted data is removed from the buffer.

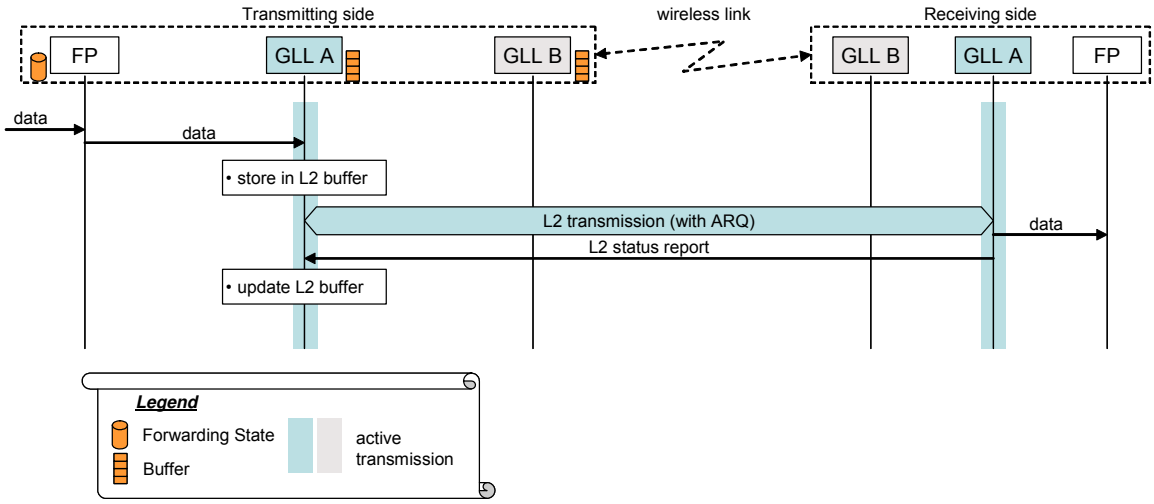


Figure 6.35: Operation of *generic link layer interworking* with context transfer before access handover.

Without context transfer, an access handover can lead to packet loss combined with either packet re-ordering or interrupted transmission. Figure 6.36 shows the case when the transmission via the new access starts before the old access is disconnected (i.e. *make-before-break* (MBB) access handover). When the forwarding point receives the access handover request, it updates the forwarding state and forwards new data to the new access (denoted as GLL B). Access B starts transmission; the access handover delay is the time period from the access handover request to the moment when the first data via the new access is received. In the meantime data buffered in the old access (GLL A) is still transmitted. At some time, the old access is disconnected and potentially remaining data is discarded. While data is transmitted via both the old and new access simultaneously, re-ordering of data occurs.

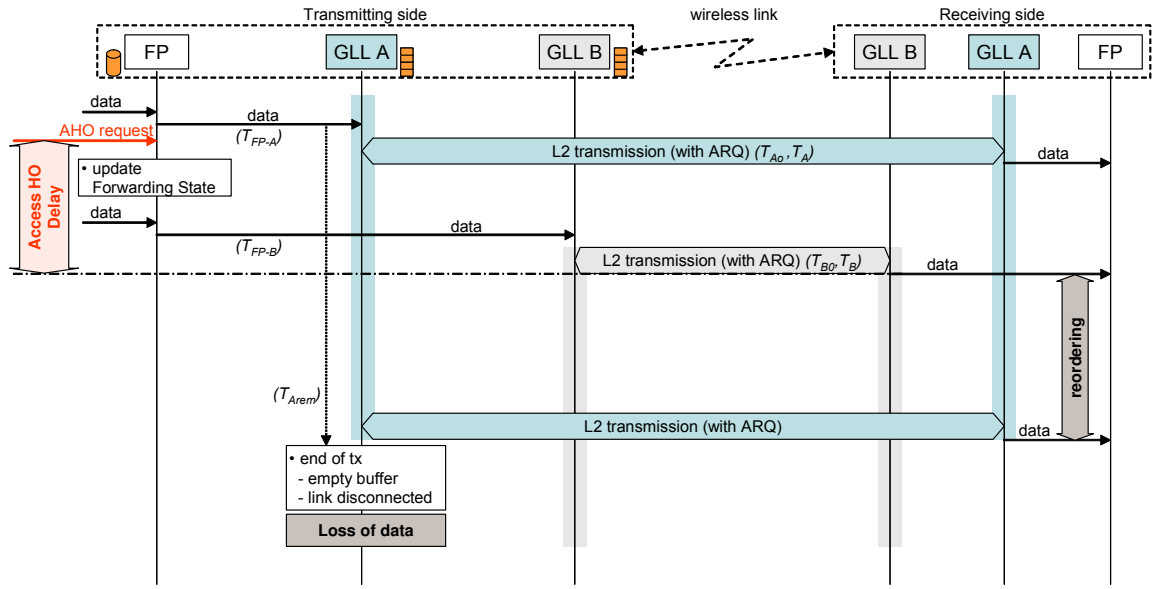


Figure 6.36: Basic access handover without context transfer leading to re-ordering and packet loss.

Figure 6.37 shows the same case, however the transmission of the old access is first terminated before the new access becomes active (i.e. *break-before-make* (BBM) access handover). The termination of the old access can lead to a data loss. Furthermore, there can be an access handover interruption when no data is transmitted.

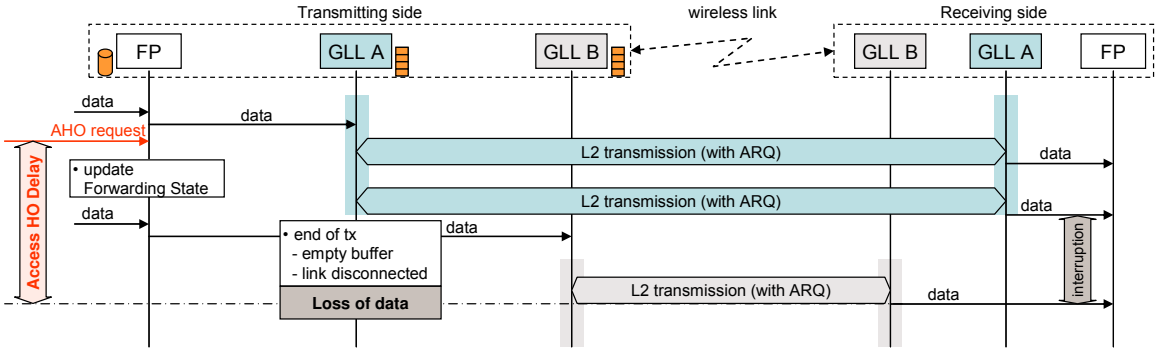


Figure 6.37: Basic access handover without context transfer leading to interruption and packet loss.

Let us quantify the amount of data distortion and investigate conditions for the occurrence of data distortion. We define:

T_{A0}, T_{B0} : Transmission latency on link A/B for the transmission of a hypothetic packet of size 0,

$T_{FP\text{-}A}$, $T_{FP\text{-}B}$: Time to forward a packet / send a message from the FP to GLL A/B,

T_{Aup}, T_{Bup} : Time to establish L2 connection for access A/B,

T_A, T_B : Transmission time of a packet on link A/B,

T_{Arem} : Time to transmit the remaining queue of GLL A after the *access switch request*,

T_{Amax} : Maximum time that link A remains active after the *access switch request*,

Q_A : Queue size at GLL A at time of the *access handover request,*

s_B : packet size of first packet on link B,

α_A, α_B : L2 transmission expansion factor[58] due to L2 error recovery for links A, B,

r_A, r_B : data rates on links A, B,

D_L : Amount of packet loss at access handover,

D_R : Amount of packet re-ordering at access handover[59],

D_{Ro} : Offset of packet re-ordering at access handover[60],

[58] The L2 transmission expansion factor reflects that the transmission delay of data does not only depend on the link rate and latency, but also on the ARQ error recovery of the link layer which delays the transmission time until data is successfully received at the receiver (see e.g. [PM01]).
[59] The amount of re-ordering is the number of bytes that are received out-of-sequence.
[60] The offset of re-ordering is the difference in bytes that packets on the new access overtake packets still being transmitted via the old access.

D_D : Amount of packet duplication at access handover,

T_I : Time of interruption at access handover.

For the basic transmission mode without context transfer we can quantify the amount of packet loss, access handover interruption and duplication with the conditions that any of them occurs.

Packet re-ordering occurs under the condition:

$$\min(T_{Arem}, T_{Amax}) > T_{FP-B} + T_{Bup} + T_B \quad . \tag{6.1}$$

With

$$T_{Arem} = \alpha_A \cdot \left(\frac{Q_A}{r_A} \right) + T_{A0} \quad , \tag{6.2}$$

$$T_B = \alpha_B \cdot \left(\frac{s_B}{r_B} \right) + T_{B0} \quad , \tag{6.3}$$

the amount of re-ordering in byte corresponds to

$$\begin{aligned} D_R &= \left(\min(T_{Arem}, T_{Amax}) - T_B - T_{Bup} - T_{FP-B} \right) \cdot r_B \\ &= \min\left(\left(\alpha_A \cdot Q_A \cdot \frac{r_B}{r_A} + \alpha_A \cdot T_{A0} \cdot r_B \right), (T_{Amax} \cdot r_B) \right) \\ &\quad - \alpha_B \cdot s_B - \alpha_B \cdot T_{B0} \cdot r_B - T_{Bup} \cdot r_B - T_{FP-B} \cdot r_B \end{aligned} \tag{6.4}$$

The re-ordering offset is

$$D_{Ro} = Q_A - \frac{r_A}{\alpha_A} \cdot \left(T_{FP-B} + T_{Bup} + T_B \right) \quad . \tag{6.5}$$

Access handover interruption occurs under the condition:

$$\min(T_{Arem}, T_{Amax}) < T_{FP-B} + T_{Bup} + T_B \quad , \tag{6.6}$$

and the interruption delay is :

$$T_I = T_{FP-B} + T_{Bup} + T_3 - \min(T_{Arem}, T_{Amax}) \quad . \tag{6.7}$$

Packet loss occurs under the condition:

$$T_{Amax} < T_{Arem}, \tag{6.8}$$

and the amount of loss in byte corresponds to:

$$D_L = Q_A - T_{Amax} \cdot \frac{r_A}{\alpha_A}. \tag{6.9}$$

The conditions for different types of data distortion are listed in Table 6-2.

Table 6-2 : Data distortion at basic access handover without context transfer.

Basic Access Handover	**Packet Loss**	**Packet Re-ordering**	**Packet Duplication**	**Access Handover Interruption**
$T_{Amax} > T_{Arem} > T_{FP-B} + T_{Bup} + T_B$ (make-before-break)	-	Yes ($D_R>0$) Eq. (6.4)	-	-
$T_{Arem} > T_{Amax} > T_{FP-B} + T_{Bup} + T_B$ (make-before-break)	Yes ($D_L>0$) Eq. (6.9)	Yes ($D_R>0$) Eq. (6.4)	-	-
$T_{Arem} < T_{Amax} < T_{FP-B} + T_{Bup} + T_B$ (make-before-break or break-before-make)	-	-	-	Yes ($T_I>0$) Eq. (6.7)
$(T_{Amax} < T_{Arem})$ $\wedge (T_{Amax} < T_{FP-B} + T_{Bup} + T_B)$ (make-before-break or break-before-make)	Yes ($D_L>0$) Eq. (6.9)	-	-	Yes ($T_I>0$) Eq. (6.7)

The performance of access handover can be improved by forwarding data via a context transfer procedure, as shown in Figure 6.38. When the forwarding point receives the access handover request, it updates its forwarding state, and immediately notifies the old and the new GLL entity about the access handover. The new GLL entity stores all new data that arrives. The old GLL entity stops the transmission of data and requests a status report from the receiver in order to know what part of the buffered PDUs have already arrived at the receiver. It then creates a forwarding context with all outstanding data, which is forwarded to the new GLL entity. The new GLL entity first transmits data forwarded via the context transfer procedure, and then continuous with the transmission of new already stored data or data arriving from the forwarding point.

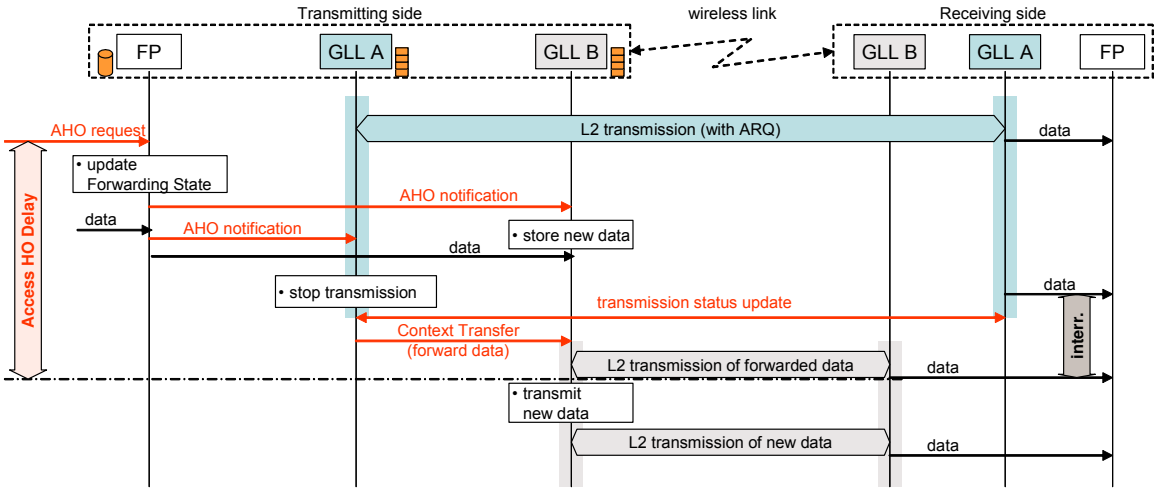


Figure 6.38: Optimised access handover with context transfer.

Depending on the timing of the different procedures, there can be an interruption time due to the access handover as indicated in Figure 6.38. This interruption time is:

$$T_I = \max\left(\left(T_{FP-A} + T_{AL2sync} + T_{CT}\right), \left(T_{FP-B} + T_{Bup}\right)\right) + T_B - T_{FP-A} \quad (6.10)$$

where:

$T_{AL2sync}$: Time for the L2 *transmission status update* procedure,

T_{CT} : Time to forward context from GLL A to GLL B.

It is a special case of access handover when simultaneous transmission of a service data flow via multiple accesses is supported. Then the service data flow is split into multiple sub-flows, which are individually forwarded to the different accesses. In this case, MRRM determines what amount of data is to be sent via the different accesses. For that it receives rate reports by the GLL entities of the different accesses to determine a suitable splitting ratio, as shown in Figure 6.39. Since different access technologies have different transmission delays, and due to varying queuing delay, the data received at the receiver via the different accesses is disordered. Therefore, this scheme is only advisable for service data flows that are robust to re-ordering.

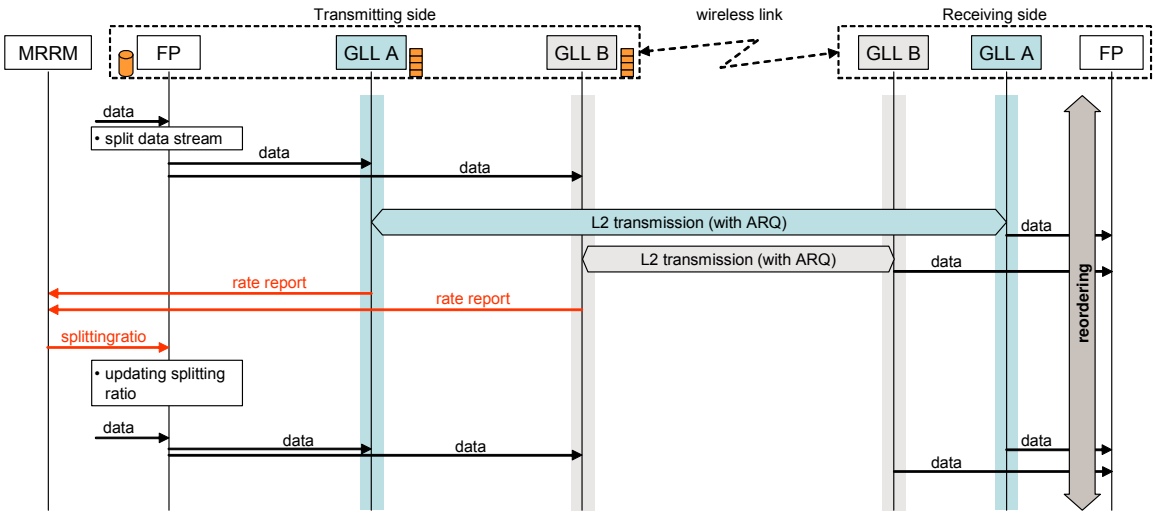


Figure 6.39: Simultaneous multi-access transmission

6.5.3.4 Access Handover with Bicasting

As an alternative to context transfer, access handover can be performed with bicasting. This scheme is depicted in Figure 6.40. When the access handover request is received at the forwarding point, the forwarding point establishes a forwarding state for data forwarding to both the old and new access. Thus all incoming data is duplicated. The data is independently transmitted via the old and new access. This reduces the access handover delay and amount of data loss, at the cost of some data duplication. The bicasting state is terminated when access handover is completed. This can be either achieved by a timeout for the bicasting or a feedback from the old link layer, when it has terminated the transmission.

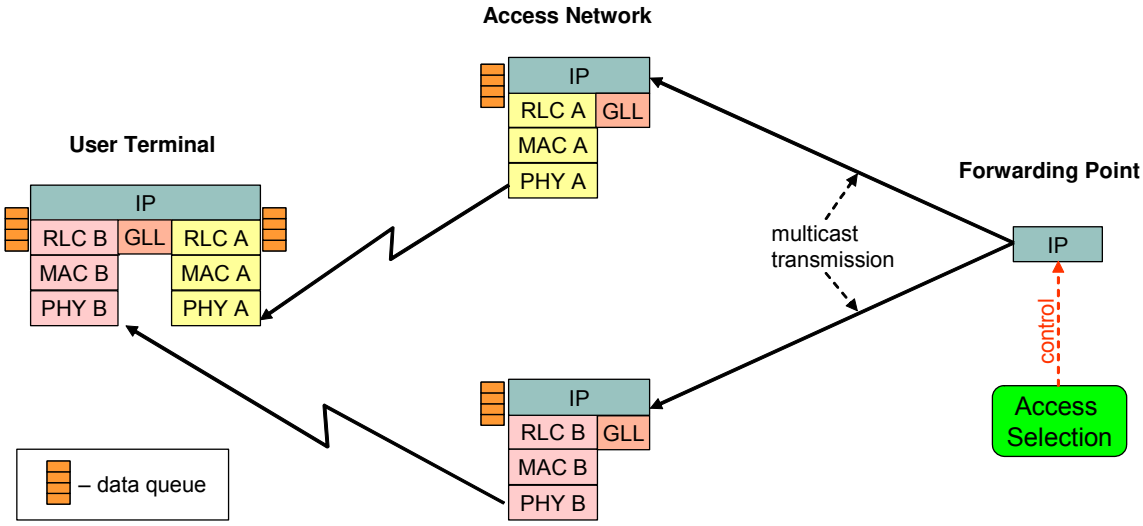


Figure 6.40: Access handover with bicasting

Access handover with bicasting leads to packet duplication. Figure 6.41 shows the case where data is transmitted in parallel via the old and the new access. While both accesses are active, packets are received out-of-sequence at the receiver. Figure 6.42 depicts the case, when the

transmission via the old access terminates before transmission on the new access starts. As a consequence there is an access handover interruption. The first packets received via the new access are duplications of the last packets received via the old access.

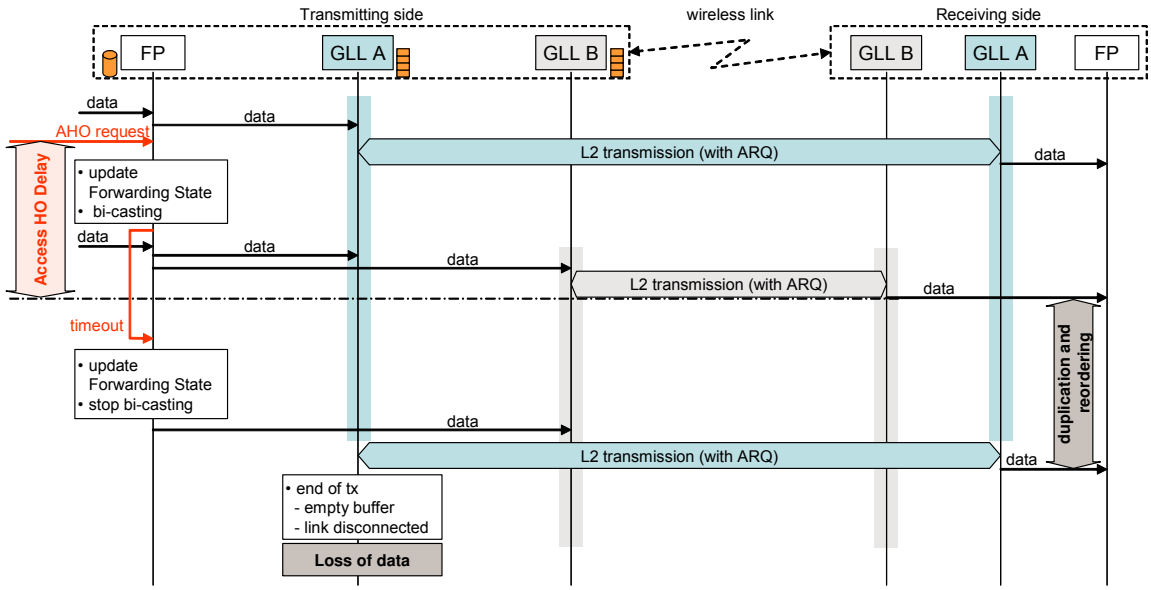


Figure 6.41: Access handover with basic bicasting ("no sync"), leading to re-ordering, duplication and loss.

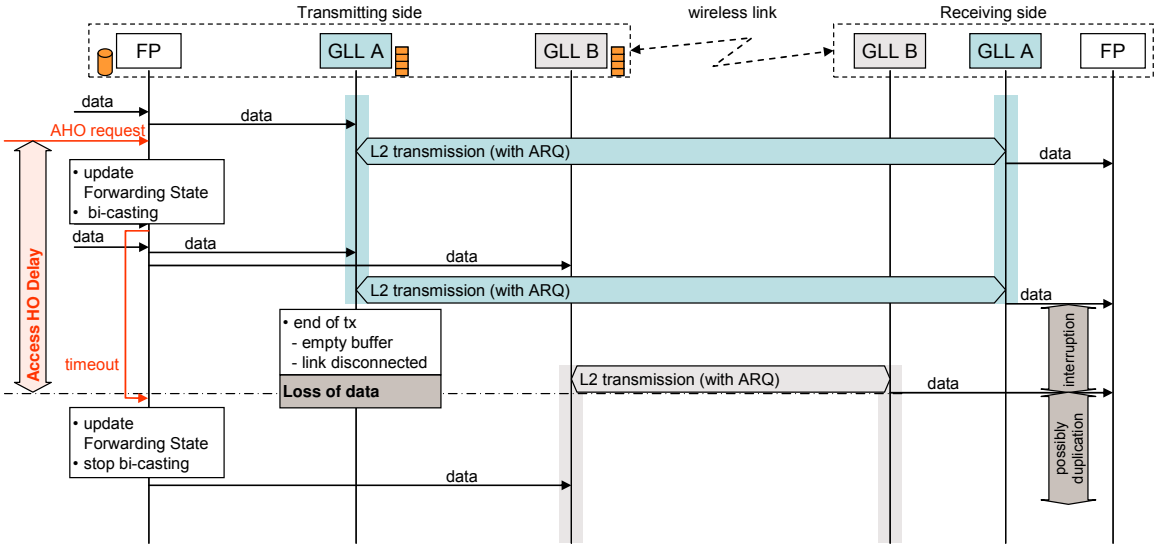


Figure 6.42: Basic bicasting access handover leading to interruption and packet duplication.

Packet re-ordering occurs under the condition:

$$\min(T_{Arem}, T_{Amax}) > T_{FP-B} + T_{Bup} + T_B, \tag{6.11}$$

with a re-ordering offset of

$$D_{Ro} = \max\left(Q_A - \frac{r_A}{\alpha_A} \cdot \left(T_{FP-B} + T_{Bup} + T_B\right), 0 \right). \quad (6.12)$$

Access handover interruption occurs under the condition:

$$T_{Amax} < T_{FP-B} + T_{Bup} + T_B, \quad (6.13)$$

and the interruption delay is :

$$T_I = \max\left(T_{FP-B} + T_{Bup} + T_3 - T_{Amax}, 0\right). \quad (6.14)$$

The amount of duplicated data due to bi-casting is

$$D_D = \max\left(T_{Amax} \cdot \frac{r_A}{\alpha_A} - Q_A, 0 \right). \quad (6.15)$$

For access handover with bicasting, two identical data streams are generated, which are in the network transmitted independently. Data distortion stems from the different transmission characteristics along the two paths, e.g. different queue sizes, transmission delay, data rate. A way to reduce the impact of data distortion is to re-synchronise the two bicasted data streams again. This can be done either at the transmitter or at the receiver. The transmitter-side synchronisation is depicted in Figure 6.43. When the forwarding point starts transmission, it notifies the old and new GLL entities about the start of bicasting, and marks the first duplicated packet. The new link layer does not yet start transmission of data, instead it stores received data until it is notified from the old link layer. At the moment that the transmission via the old link layer is terminated, a context is created and forwarded to the new link layer. This context indicates which of the bicasted data packets have been successfully transmitted via the old link layer. The new link layer can remove those packets from the buffer, in order to avoid packet duplication. It can also happen that the transmission via the old access is stopped before the first bicasted packet is transmitted, e.g. when the old link layer has a queue of outstanding data. In this case, the context transfer can forward all data packets, which are not yet successfully transmitted up to the marked first bicasted packet.

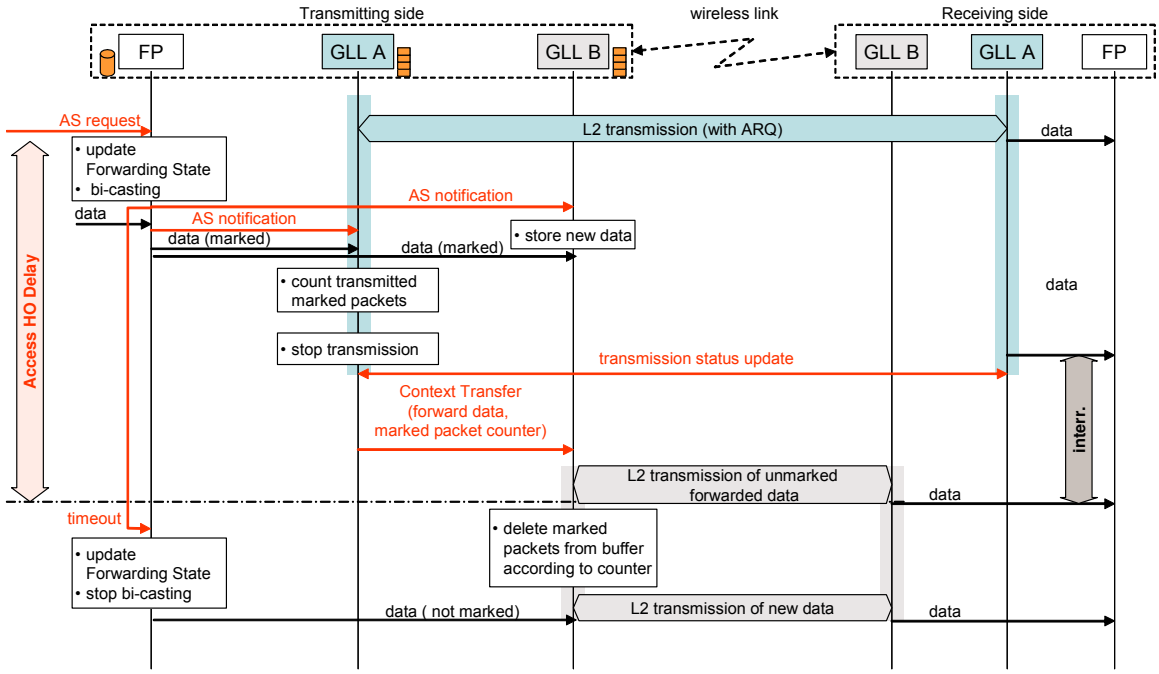


Figure 6.43: Access handover with lossless bicasting including context transfer for transmitter synchronisation ("transmitter sync").

Receiver-side synchronisation of the access handover is depicted in Figure 6.44. Again, the GLL entities keep track of which bicasted data has still been received via the old access. Bicasted data is marked as duplicated. The old GLL receiver counts the correctly received bicasted packets. When the new access is setup, the new GLL receiver validates how many packets had still been received by the old GLL receiver, and notifies the new GLL transmitter. The new GLL transmitter discards all packets that have already been received, before it starts transmission. Thereby, packet duplication is avoided.

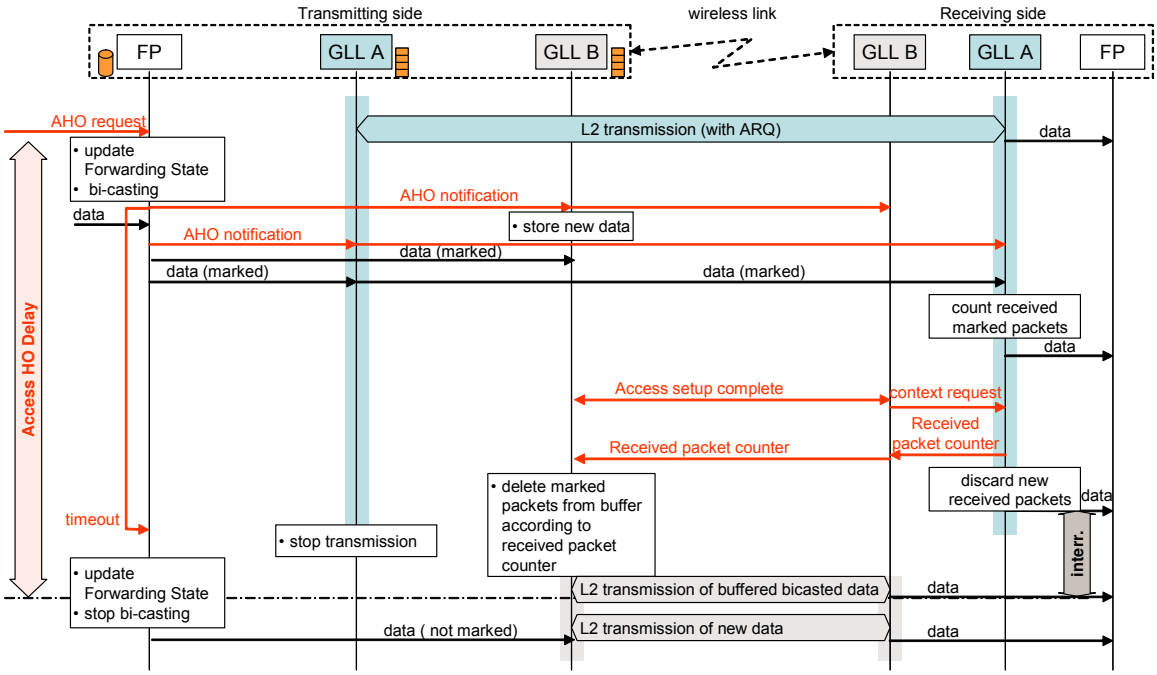


Figure 6.44: Access handover with lossless bicasting including context transfer for receiver synchronisation ("receiver sync").

A further method to avoid duplication is to use the basic bicasting scheme of Figure 6.41 and Figure 6.42 and in addition perform duplicate detection and filtering at the receiver. It requires the receiver to be able to identify every packet. In case that transmitted packets contain sequence number, duplications could be detected based on a sequence number; or a hash value that is calculated from (parts of) the received data packet. In case of a sequence number also re-ordered packets can be put back into the correct order. This solutions is however inefficient, since duplicates are only filtered out after having been transmitted over the radio interface.

6.5.3.5 Access Handover with Context Anchor

Another way to manage access handover is based on a common context anchor for the different accesses. This is depicted in Figure 6.45. The communication context is centrally maintained in a context anchor, which is located at the forwarding point. Access handover is performed by redirecting the communication path at the context anchor. The communication context and data located at one access is lost at the access handover. The context anchor maintains sufficient context information to quickly re-establish a new context for the new access and make the access handover lossless.

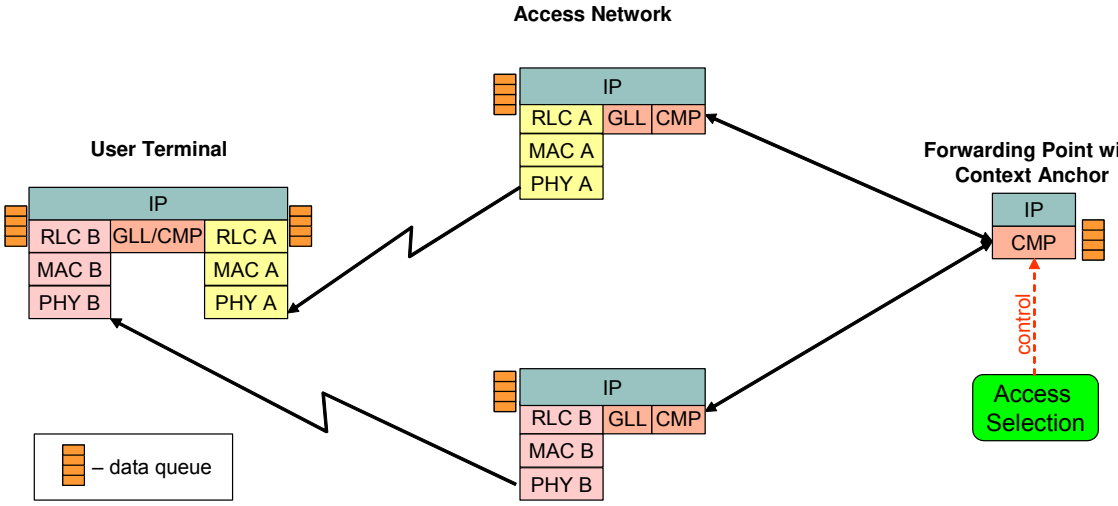


Figure 6.45: ***Generic link layer interworking*** **with context anchor and context management protocol (CMP).**

Figure 6.46 shows the operation of the GLL-IW before access handover. The forwarding points adds a sequence number to data packets forwarded to access A, and it also keeps a local copy of the data. Data is transmitted via the access and status reports are received about successful transmission. The GLL notifies the forwarding point about successful transmission and these packets are purged from the buffer at the forwarding point accordingly. These transmission reports can also be used for flow control, to manage the amount of data stored in the buffers at the access.

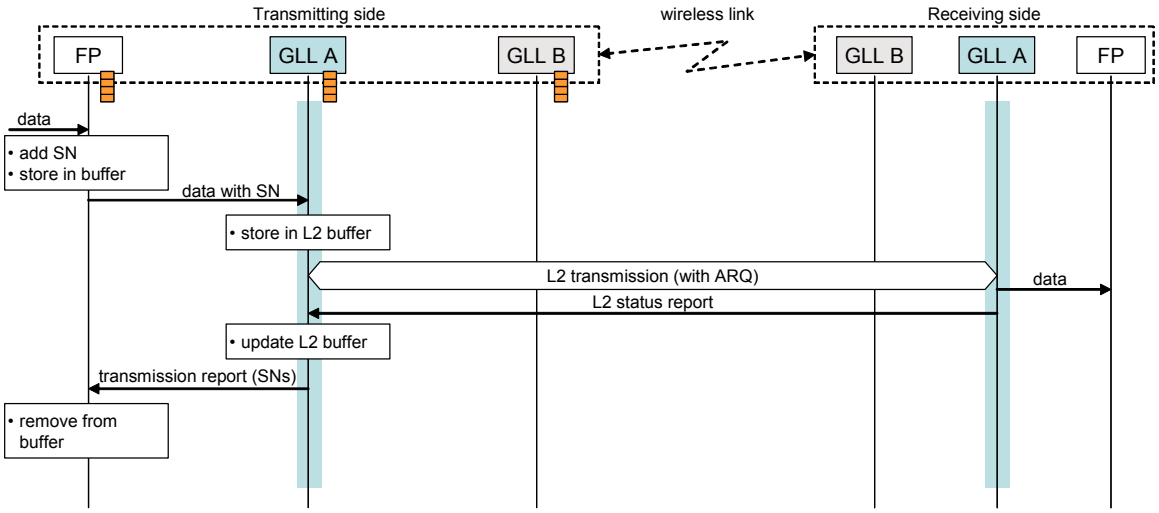


Figure 6.46: Operation of GLL with context anchor before access handover.

In case of an access handover, as depicted in Figure 6.47, the forwarding point stops forwarding data to the old access and notifies the old and new GLL of the access handover. The new link layer receives information (e.g. security keys) to establish the communication context and setup connectivity. After updating the forwarding state, the forwarding point forwards data to the new access. In the procedure, shown in Figure 6.47, the old access continues transmitting data in the GLL buffer, which leads to duplication of all data that is still queued at the old access and which is also forwarded to the new access. The duplication can also lead to packet re-ordering.

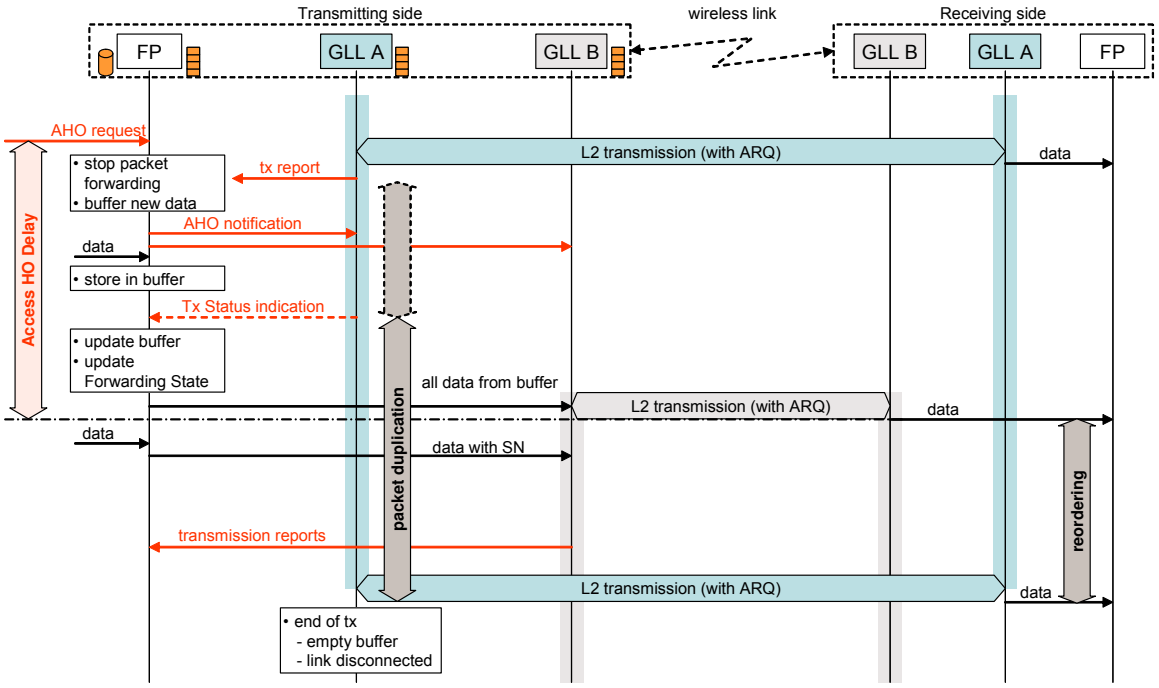


Figure 6.47: Basic access handover of GLL with context anchor resulting in data duplication and re-ordering.

The packet duplication is similar to the case of bicasting; some data is transmitted from the forwarding point to both accesses. The amount of duplicated data is as in eq. (6.15), and the condition of packet re-ordering and the re-ordering offset is as in eqs. (6.11) and (6.12) respectively. In case that the buffer of the old access is empty, or the access is early disconnected, access handover interruption occurs according to eqs. (6.13) and (6.14).

One way to avoid the data distortion of packet duplication and re-ordering is to correct it at the receiver side. This requires that the sequence number added at the transmitting forwarding point is transmitted all the way to the receiving forwarding point. The receiving forwarding point can then store data, bring it back into order and discard duplicates. Another way to avoid packet duplication is to synchronise the data transmitted via the old access with the data transmitted via the new access. This procedure is shown in Figure 6.48. When the old access receives the access handover notification, it stops transmission and synchronises with the receiver the successfully transmitted data. It notifies the forwarding point about successful transmission, and the forwarding point only forwards not-yet-transmitted data to the new access. In this case we obtain an access handover interruption without duplication or re-ordering.

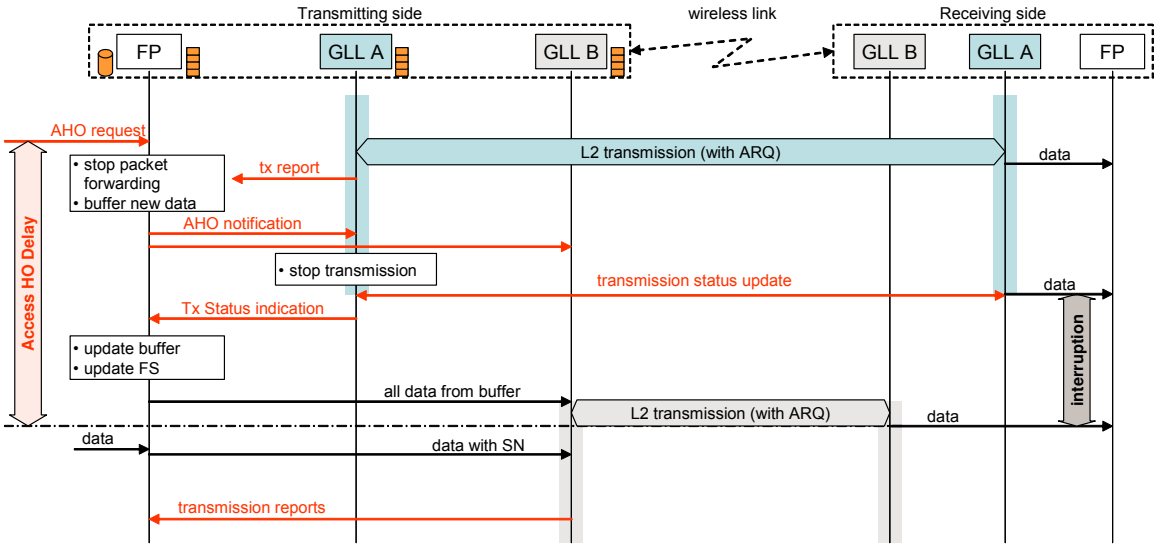


Figure 6.48: Lossless access handover of GLL with context anchor.

Access handover with context anchor can be used for simultaneous transmission of data via both links, as shown in Figure 6.49. The forwarding point splits the service data flow and sends one part to access A and the other part to access B. The transmission reports from the GLL entities can serve two different purposes. As in the case of transmission via a single access, transmission reports indicate successfully transmitted data, which is purged from the forwarding point buffer. In addition, the transmission reports indicate the transmission rates of the two accesses to the forwarding point. The splitting ratio of data transmitted via the different accesses can be thus adapted to the perceived ratio of transmission rates. Simultaneous transmission leads to re-ordering of packets at the receiver due to different transmission delays of the different accesses.

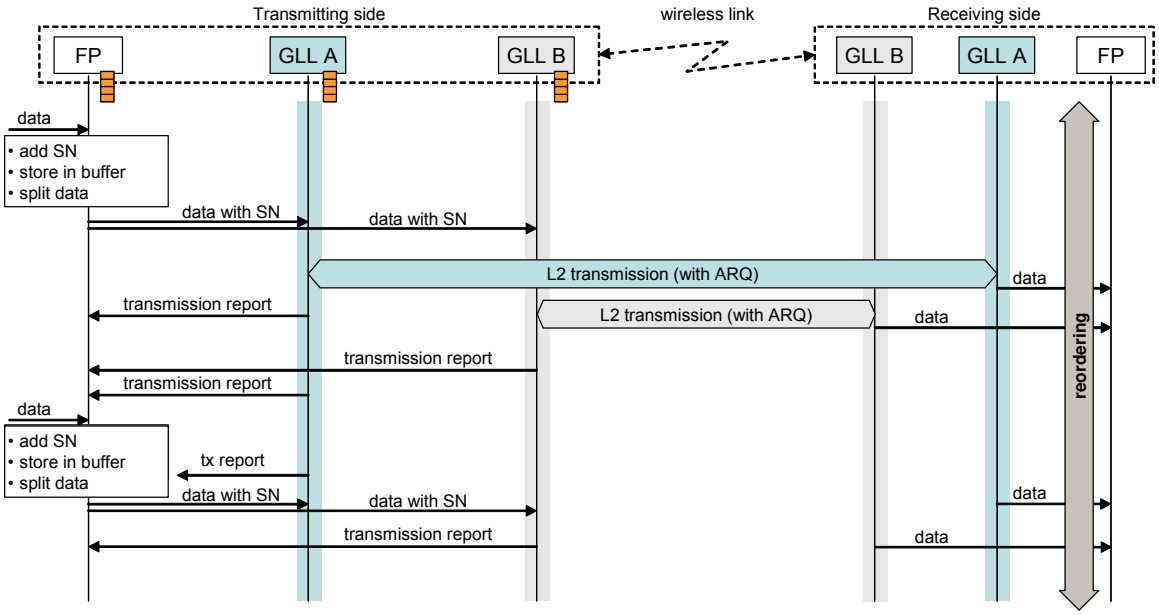


Figure 6.49: Parallel transmission with re-ordering.

6.5.3.6 Discussion

Generic link layer interworking adds interworking functionality to the link layer of different access technologies. This interworking functionality provides support for re-establishing or maintaining a communication context at access handover. The access technology specific transmission procedures remain largely untouched by the new interworking functionality. This makes generic link layer interworking particularly applicable for the integration of existing access technologies into a multi-access system. Only limited standardisation effort is required to add the interworking functionality to existing RATs. Generic link layer interworking is useful for decentralised multi-access architectures including separate technology-specific access networks. It facilitates that different access networks are operated by different business entities, thus providing flexibility in deploying new business models and business relationships.

With generic link layer interworking, access handover can lead to data distortion due to packet loss, packet duplication, packet re-ordering and access handover interruption times. This distortion can be reduced or even removed with synchronisation procedures, which coordinate the transmission via multiple accesses at access handover. These procedures require interactions between the different generic link layer entities and / or forwarding points. This comes at the costs of signalling overhead in the network and increased delays for access handover.

6.6 Access Handover Performance Evaluation

In this section we investigate the performance of different access handover schemes by means of simulations. In particular, we investigate the impact of the access handover on services using the transport layer protocol TCP, which accommodates approximately 90% of Internet services [IM04]. TCP is a connection oriented reliable transport protocol, which operates end-to-end and performs error recovery for lost packets, flow control to avoid overload of the receiver, congestion control to avoid congestion in the network. The transmission is based on a sliding ARQ window (denoted as congestion window) and the receiver reports received packets in cumulative acknowledgements [Ste94] [Tan96] [RFC793] [RFC813]. The congestion window size is dynamically adapted to the congestion state of the network [RFC2581] [RFC3042]; thereby the transmission rate of the TCP sender and the amount of data in flight is adapted to the pipe capacity of the end-to-end path. Packet losses can be detected at the TCP receiver by gaps in the received sequence numbers. The receiver reacts by sending duplicate acknowledgements[61] back to the TCP sender; this leads to a *TCP fast recovery* and results in a halving of the congestion window. Packet losses can also be detected by a TCP timeout[62] and lead to the congestion window being reset to one TCP segment. Successful transmission of packets lead to an exponential increase of the congestion window size over time in the slow start phase, which occurs at the beginning of the TCP connection and after timeouts; otherwise the congestion window increases linearly in time in the congestion avoidance phase. Note, several TCP versions exist that can differ largely in their performance of error recovery. TCP selective acknowledgements (SACK) allow a TCP connection to recover in a single round trip time from multiple losses without any unnecessary retransmissions [RFC2018] [RFC3517]. SACK is the most sophisticated TCP variant for error recovery; it can be found in many operating systems nowadays and it is increasingly deployed [PB04]. Therefore, we consider only TCP SACK in this evaluation.

6.6.1 Evaluation Scenario and Performance Metrics

We consider a file download application running over TCP SACK. We are interested in the interactions of data distortion caused by access handover and the congestion control algorithm of TCP. As access handover schemes we investigate three context transfer schemes, which we compare with a scheme without access handover optimisation (see Figure 6.50). Concerning the data distortion, the properties of access handover with *Layer 2 Tunnel* are similar to those of *Layer 2 Transcoding* and the multi-radio GLL schemes (assuming that transmission is not using multiple accesses simultaneously). Therefore, we assume that the TCP performance for all these schemes (marked in light blue in Figure 6.50) is similar. Among those schemes we investigate only *Layer 2 Tunnel* in the simulations. Two other schemes that we evaluate are *SDU context transfer* and *SDU reconstruction*. The latter has similar properties in data distortion as transmission with a *context anchor*, so we would expect similar results.

[61] Multiple acknowledgements indicate the same sequence number up to which data has been received correctly.
[62] The timeout value of the TCP retransmission timer is dynamically adapted to estimations of the end-to-end round-trip time [RFC2988].

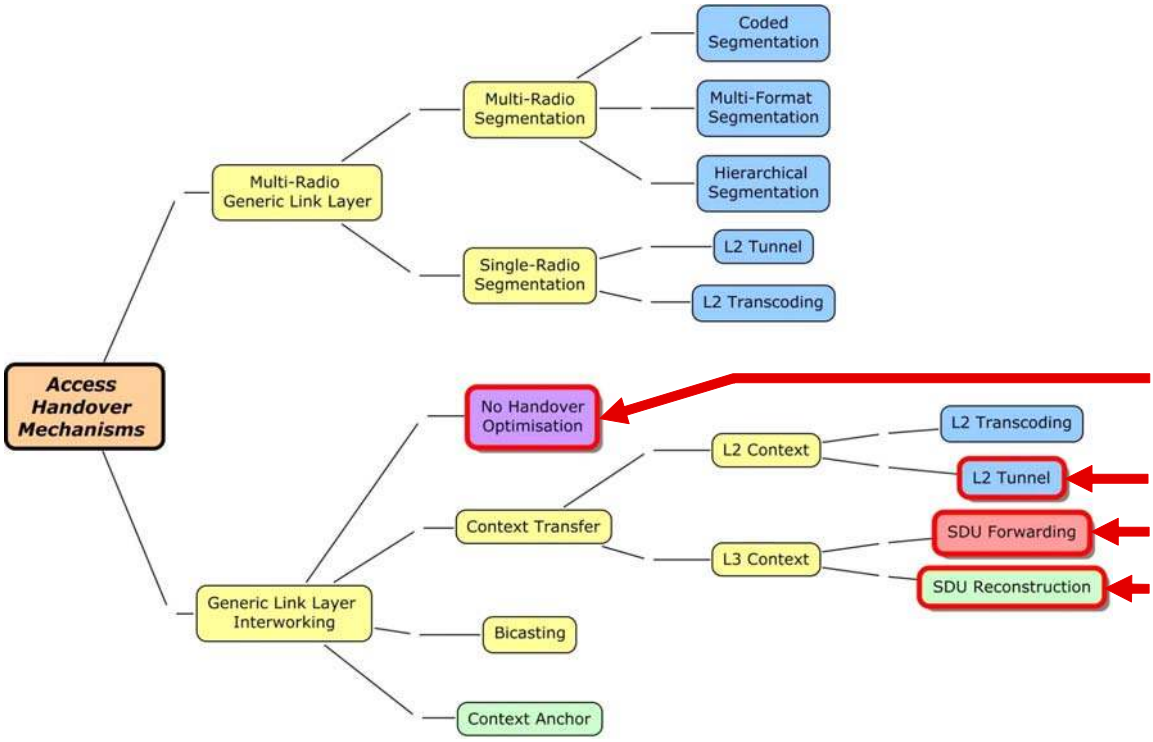


Figure 6.50: Evaluated access handover schemes.

We consider a scenario where a switch between different accesses is triggered by some access selection algorithms during an ongoing TCP session. The radio link characteristics are configurable, so that we can investigate access handover between different types of access technologies.

As performance metrics for the data session we define the *object bit rate* (OBR) as the throughput perceived end-to-end at the application layer above TCP for the transmission of a file of a certain size:

$$OBR\,[\text{kb/s}] = \frac{\text{file size}\,[\text{kbits}]}{\text{transmission time}\,[\text{s}]}. \qquad (6.16)$$

We further define the *normalised object bit rate* as the object bit rate when access handover is performed divided by the object bit rate without access handover. In order to investigate the statistical significance of the results we perform 31 simulation runs with different seeds of the random number generators, unless mentioned otherwise. According to the central limit theorem, the samples can be assumed to follow a normal distribution [Jai91]. We derive the 95% confidence intervals.

We structure our evaluation into three studies. In a first investigation (Section 6.6.3) we analyse the impact of the different access handover schemes on TCP performance. For that we consider an access handover between two accesses with the same characteristics. In a second investigation (Section 6.6.4) we investigate the influence of radio link parameters on the TCP performance. Again we consider an access handover between two accesses with the same

characteristics. In a third investigation (Section 6.6.5) we investigate the access handover between accesses with different characteristics.

6.6.2 Simulation Model and Parameters

For the evaluation of access handover with different context transfer mechanisms we have developed an event-driven simulator as depicted in Figure 6.51.

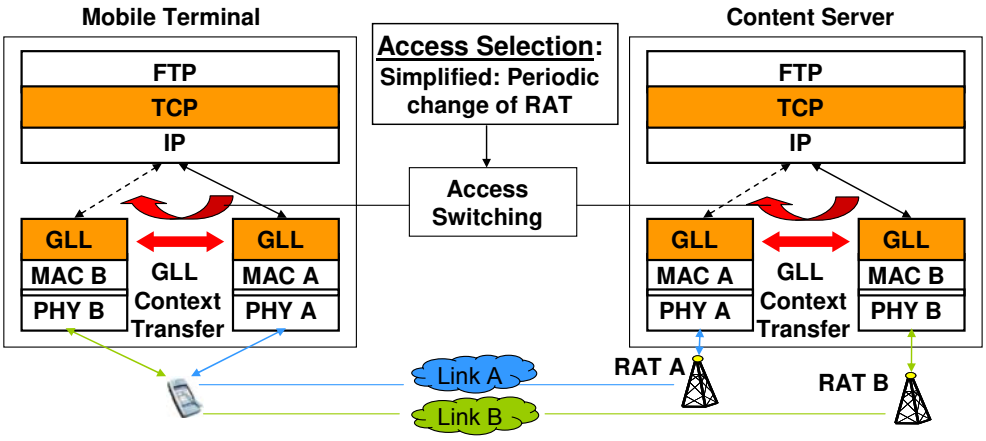


Figure 6.51: Simulation environment.

We consider a traffic scenario with bulk data transfer of a large file. The file size is chosen to be very large, so that many access handover events occur during the file download. As performance metric the object bit rate is calculated, which is the size of the file divided by the total download time. As transport protocol TCP SACK has been implemented according to [RFC2018] and [RFC3517], which also includes *limited transmit* [RFC3042]; the TCP initial window size is set according to [RFC2414]. The TCP configuration parameters are listed in Table 6-3.

Table 6-3 : TCP settings in simulation environment.

Parameter	Value
Maximum Sender Congestion Window [byte]	4 * 65.535 byte
TCP Version	SACK + Limited Transmit
Maximum Segment Size [byte]	1460 byte
Header Size [byte]	20 byte
Initial Congestion Window Size [byte]	3 * 1460 byte

The RATs are modelled by means of a configurable generic link layer (GLL), which is largely based on the UMTS radio link control protocol [3GPP25.322]. The GLL contains SDU and PDU buffers, segmentation and reassembly functions, configurable ARQ functionality with in-order delivery. To support access handover, the four context transfer schemes have been implemented. The configuration of the ARQ toolbox is according to Table 6-4.

Table 6-4 : The "standard" RLC toolbox configuration.

Transmitter	**Value**
Transmitter ARQ Window Size [#PDU]	2047
Parameter *Poll on Last PDU in Buffer*	True
Poll Timer (ms)	RTT[63]*3
Parameter *Poll every X PDU*	6
Parameter *Poll every X SDU*	1
Parameter *Poll on Last PDU in retransmission buffer*	True
Parameter *Poll Window* [%]	70%
Timer *Poll Prohibit* [ms]	RTT*1.2
PDU Size [byte]	160 byte
Receiver	**Value**
Receiver ARQ Window Size [#PDU]	2047
Timer *Status Prohibit* [ms]	RTT*1.1
PDU Size [byte]	160 byte

The medium access control (MAC) schedules PDUs for the radio transmission. The radio link is configured for a certain data rate, a transmission time interval and a transmission delay. As error model a simple block error model is used with a configurable block error rate (BLER). The Table 6-5 shows a typical configuration for the parameters of the MAC and physical layer, which is also used for the "standard" configuration.

Table 6-5 : MAC and physical layer configuration.

Parameter	**Value**
Transmission time interval (*TTI*) [ms]	10 ms
Block error rate (*BLER*) [%]	0%-15% (default 5%)
Round-trip time (*RTT*) [ms]	50 ms
Radio link rate (*r*) [Mbps]	5.25 Mb/s

The SDU buffer size is adapted to the TCP bandwidth-delay-product[64] (BDP). The bandwidth-delay-product describes the amount of data that TCP needs to have in-flight in order to fully utilise the available capacity on the end-to-end path (see e.g. [Ste94]). The bandwidth-delay-product is

$$BDP = r_{bottleneck} \cdot RTT_{e2e}, \tag{6.17}$$

where $r_{bottleneck}$ is the rate of the bottleneck link, and RTT_{e2e} is the end-to-end round-trip time without queuing. Most wireless systems use dedicated buffers per user resulting in a low amount of statistical multiplexing between multiple TCP flows within a queue. The size of the bottleneck queue should be somewhat larger than the bandwidth delay product as a good trade-off to achieve good link utilisation and avoid excessive over-buffering [SLMP03a] [SLMP03b] [LLB05] [ES03]. Assuming that the link layer round-trip time is the most

[63] Round-trip time of the link layer.
[64] In this case "bandwidth" refers to the (bottleneck) data rate of the end-to-end path and not to a frequency range.

significant contributor to the RTT_{e2e} and accommodate for some link layer retransmissions, we set the SDU buffer size to

$$\text{SDU buffer size} = 10 \cdot r \cdot RTT \quad (6.18)$$

A large number of access selection algorithms exist, which can operate on several time scales, ranging from fractions of seconds to minutes (see Section 4.5.1). Within the simulator access selection is simply modelled to be periodic with a random offset, and the *access selection frequency* is a simulation parameter. In the figures the *access selection period* (ASP) is depicted, which is the inverse of the access selection frequency. When an access handover is triggered, the IP service data flow is immediately redirected from one link layer entity to another link layer entity. At the same time the context transfer procedure is triggered. We do not consider any delay for the context transfer procedure.

6.6.3 TCP Performance for Different Access Handover Schemes

For the evaluation of the access handover performance both radio links, the old one and the new one, are configured with the parameters as given in Table 6-6. We consider a file transfer of 100 Mbyte of data and access handover is performed every 10 s. The object bit rate of the transmission thus comprises a large number of handover events.

Table 6-6: "Standard" configuration MAC and PHY parameters.

	r [Mb/s]	BLER [%]	RTT [ms]	PDUSize [byte]	BDP [kbyte]
Link A	5,25	5	50	160	262,5
Link B	5,25	5	50	160	262,5

Figure 6.52 presents the object bit rate of the different context transfer schemes for different access selection periods. It can be clearly seen that the influence of access handover events on performance decreases for large access selection periods. If access handovers occur less frequent than once per minute, the overall effect of access handover on the average performance is low (less than 10%). What can also be seen is that the difference in performance between the context transfer schemes becomes larger at higher access selection frequencies. *No context transfer* always shows the poorest performance followed by *SDU context transfer* and *SDU Reconstruction* respectively. *Layer 2 Tunnel* always performs best with hardly any performance degradation. Even for access handovers occurring every 5s, the performance is only 2% lower than without any access handover. It seems remarkable that *SDU Reconstruction* is so strongly affected by access handover, although it provides a lossless access handover. A closer investigation reveals that the performance degradation is caused by packet duplication as will be described in Section 6.6.3.3. These packet duplications lead the TCP receiver to generate duplicate acknowledgements, which are interpreted by the TCP sender as congestion. As a reaction the TCP transmission rate is halved.

In the following sections we explain the effects leading to the transmission performance shown in Figure 6.52.

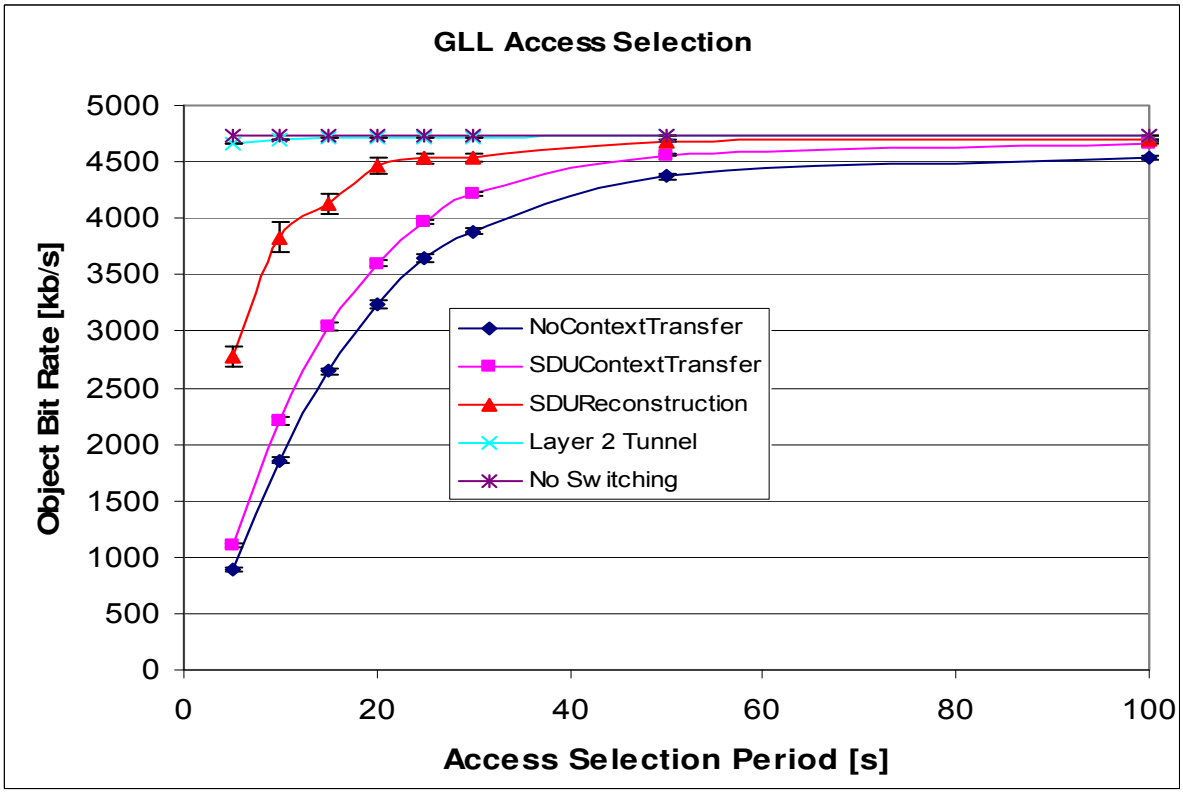


Figure 6.52: Performance of access handover for different access selection periods.

6.6.3.1 No Context Transfer

The access handover scheme *no context transfer* does not use any handover optimisation procedure; it is described in Section 6.5.3.2. At handover all data buffered at the link layer of the old access is lost. This packet loss triggers the TCP congestion control algorithm, and the TCP sender reduces the rate. As a result, TCP does not utilise the available capacity after the handover until TCP congestion avoidance has increased the data rate again. The data rate at which TCP is sending data is represented in the TCP congestion window at the sender. Whenever the congestion window is lower than the bandwidth-delay-product, the path capacity is not fully utilised. Figure 6.53 depicts the TCP congestion window over the file transfer time for one simulation run with a total of 42 access handovers. After every access handover the TCP congestion window decreases due to the detected packet loss. In some cases the window size decreases to one segment size of 1460 bytes. Then the packet loss leads to a TCP timeout, which leads to *TCP slow start*. In the other cases, the window size halves and TCP recovers from the loss by *TCP fast recovery*. TCP fast recovery can recover from the packet loss more quickly.

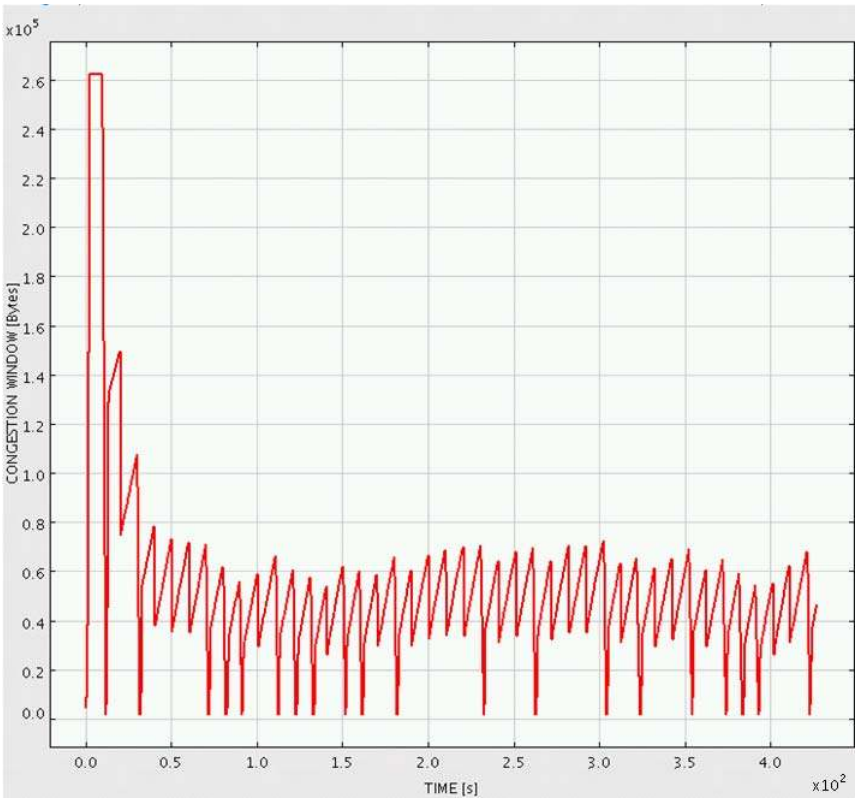


Figure 6.53 : TCP congestion window for *no context transfer*

In this example 20 out of the 42 access handover lead to TCP slow start, as shown in Table 6-7.

Table 6-7 : TCP loss recovery algorithm occurrence for *no context transfer*.

Loss Recovery Algorithm	Occurrence in 42 access handovers [%]
Fast Recovery	52%
Slow Start	48%

Figure 6.54 shows the TCP trace during the access handover at time *t = 140.684 s* leading to TCP fast recovery. The access handover is marked by a vertical black line. Before access handover the TCP sender has the maximum number of segments outstanding, so it requires new acknowledgements before it can advance its congestion window (*cwnd*) and send new data. The TCP receiver has already sent out a burst of ACKs, marked as 1), however these ACKs are not received at the sender as they are lost at the access handover. The TCP segments (marked as 3)) are also lost during the access handover. The TCP sender remains idle until a TCP acknowledgement (marked as 2)), is sent at *t = 140,8805 s*. This ACK is triggered by the expiry of the *TCP delayed ACK timer*, which is set to 200 ms as in typical TCP implementation. This ACK is received via the new access and cumulatively acknowledges all received TCP segments, thus including the information of the lost ACKs. The TCP sender can then advance its congestion window and transmit new data. When the new data segments are received at the TCP receiver, they trigger *duplicate acknowledgements* (DUPACKs). These DUPACKs signal to the TCP sender that the data segments (marked as 3)) have been lost. After reception of the third DUPACK, the TCP sender halves its

congestion window and starts retransmitting the lost segments. The fast recovery is only triggered after receiving three DUPACKs. This is the case because the reception of ACK 2) at *t = 140,8805 s* triggers the transmission of more than three new TCP segments. The access handover reduces the performance of the file transfer in two ways. Firstly, it leads to an idle time until the ACK 2) is transmitted; secondly, it reduces the congestion window which reduces the TCP transmission rate.

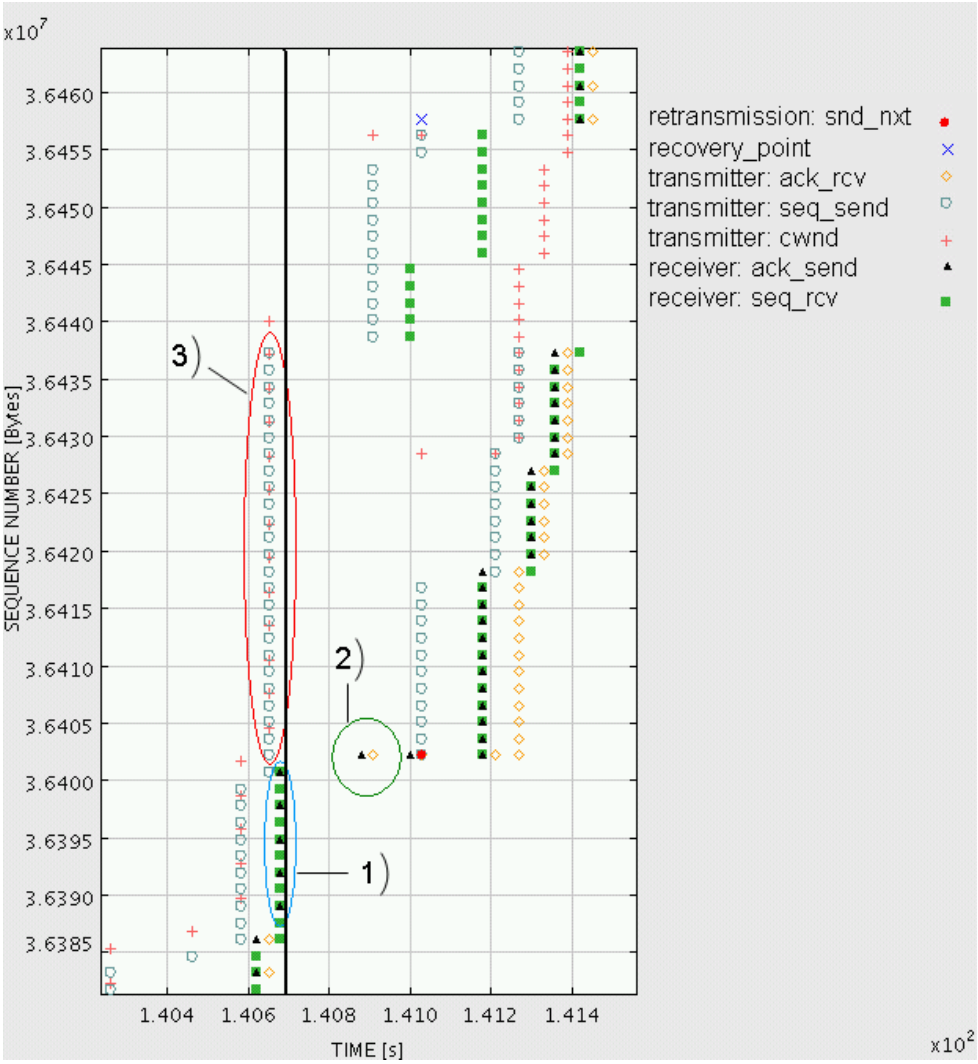


Figure 6.54 : TCP fast recovery at access handover (marked with line) with *no context transfer*.

Figure 6.55 shows another TCP trace during the access handover at time *t = 131.056 s*. In this case the *delayed ACK timer* is not running as an ACK for two received segments have been sent just before the handover. All outstanding data is lost at the access handover and no ACKs are generated that can trigger the TCP fast recovery. TCP slow start is triggered by the expiry of the *TCP retransmission timer* at time *t = 133.026*. The congestion window is set to 1460 bytes, i.e. the TCP maximum segment size, and then TCP start to retransmit all lost segments. The TCP timeout leads to an idle time of more than 2 s, and results in a link underutilisation until the congestion window has reached the pipe capacity again.

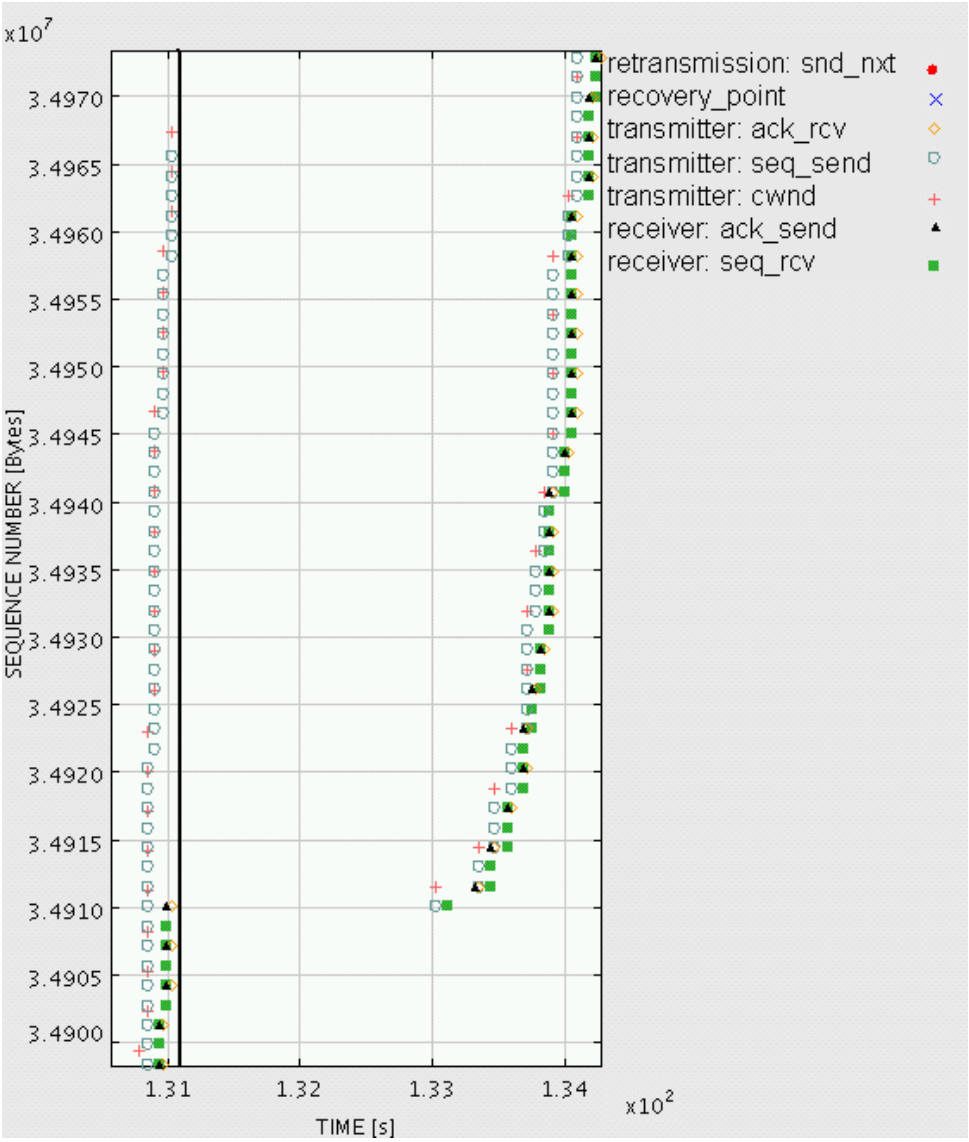


Figure 6.55 : TCP slow start after access handover (marked with line) with *no context transfer*.

As shown in Table 6-8 (cf. Figure 6.52) the performance degradation of access handover increases with lower access selection periods, i.e. larger access handover frequencies.

Table 6-8 : Access Handover frequency and OBR for varying access selection periods (ASP).

ASP [s]	OBR [kb/s]	Confidence Interval 95%	Average Amount of Access Handovers
5	891	[870;913]	90
10	1853	[1825;1882]	21
15	2645	[2618;2672]	10
20	3241	[3201;3282]	6
25	3646	[3613;3681]	4
30	3887	[3854;3920]	3
50	4369	[4346;4393]	2
100	4542	[4526;4559]	1

As described above, both TCP fast recovery and TCP slow start contribute to the performance degradation, however to a different extent. The proportion of fast recovery and slow start are equal. This is shown in Table 6-7 for a single simulation run; it is also confirmed for a large number of simulation runs as shown in Figure 6.56. This can be explained with the *TCP delayed acknowledgement* policy [RFC813] [RFC1122] [RFC2581], which aims at sending a TCP ACK only for every second received TCP segment. When a first TCP segment is received, no TCP ACK is sent. Instead the TCP delayed ACK timer is started. When a second TCP segment is received before the delayed ACK timer expires, an ACK is sent for the two segments, otherwise an ACK is sent for the single segment at timer expiry. The timer is upper bound by 500 ms [RFC2581] and a typical value is 200 ms. In our scenario data segments and ACKs are lost at access handover. A new ACK after the access handover can trigger the transmission of new TCP segments and corresponding ACKs; this can trigger the TCP fast recovery. Without delayed ACK, TCP is forced into a timeout. As the delayed ACK timer is started for every second TCP segment, the probability of having a TCP ACK timer running is 50% when the access handover occurs. This explains the equal probability of TCP fast recovery and TCP slow start for the access handover without context transfer, where all queued at the bottleneck link is lost.

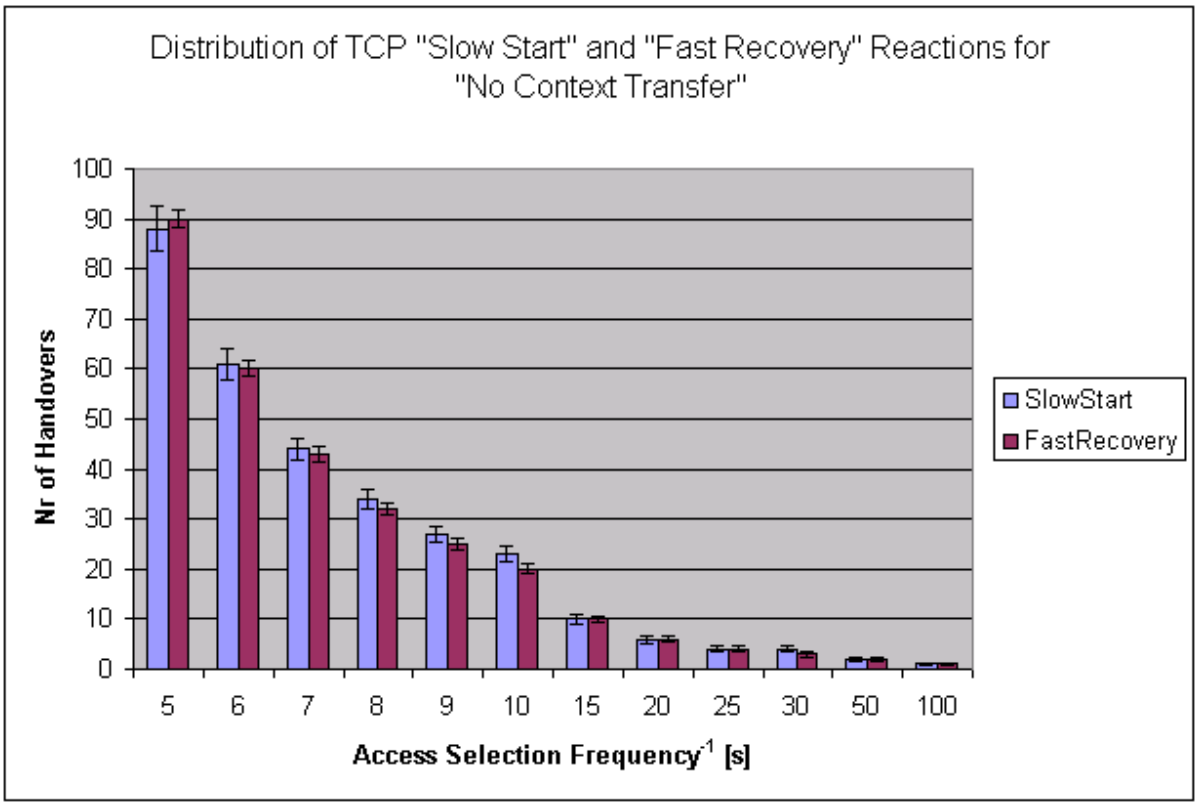


Figure 6.56 : TCP loss recovery reactions at access handover for varying access selection periods.

To understand the impact of TCP slow start and TCP fast recovery on the file transfer performance, we investigate the handover transmission delay. We define the handover transmission delay as the time period from the moment of access handover until the time when the TCP sender transmits the first segment after the access handover. The average handover transmission delay is depicted for TCP slow start in Figure 6.57 and for TCP fast recovery in Figure 6.58. The delay does not depend on the access selection period. A TCP

slow start leads to an average idle time of 1.8 s, whereas fast recovery leads to an idle time of approximately 0.19 s. TCP slow start consequently leads to ten times larger idle times than TCP fast recovery. The access handover delay does not depend on the access selection frequency.

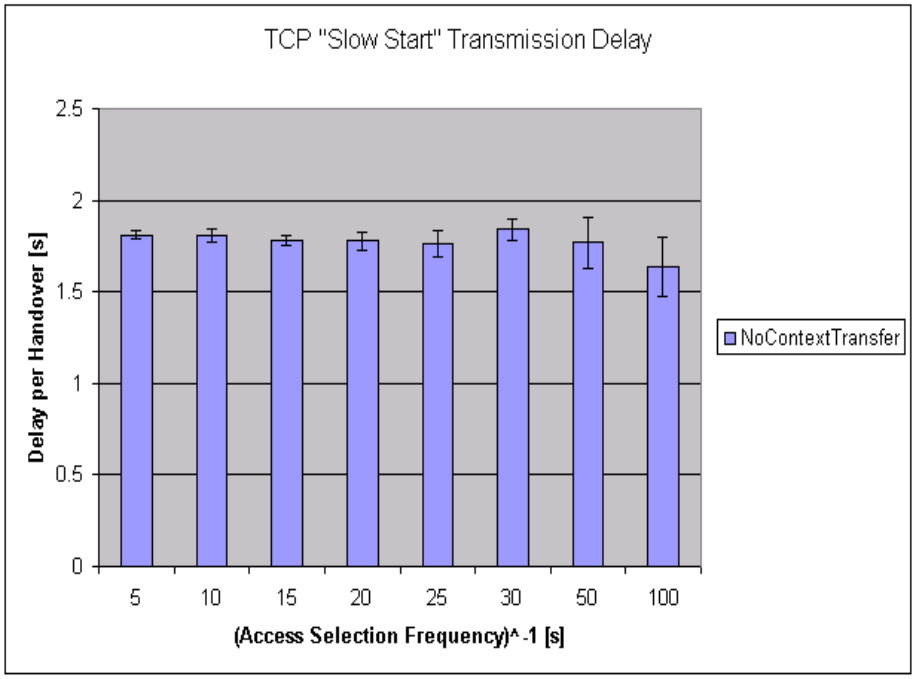


Figure 6.57 : Access handover delay with TCP slow start.

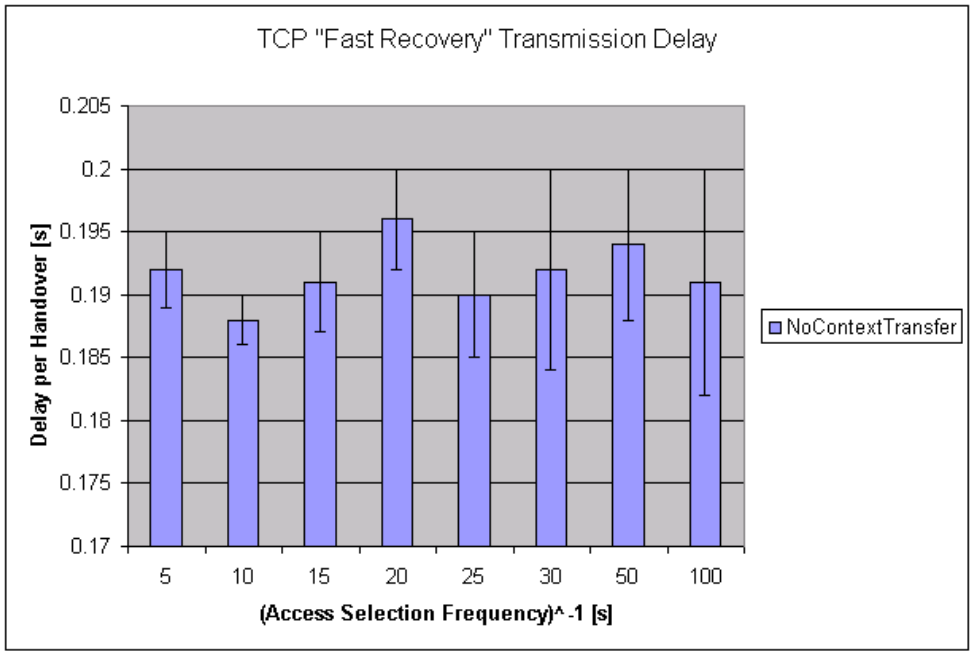


Figure 6.58 : Access handover delay with fast recovery.

6.6.3.2 SDU Context Transfer

Access handover with SDU context transfer is described in Section 6.5.3.3. During the handover all SDUs from the old link layer SDU buffer are transferred to the new link layer, as shown in Figure 6.32. The development of the TCP congestion window is depicted in Figure 6.59 for one simulation run. Similar to the case without context transfer (cf. Figure 6.53), every access handover leads to a packet loss and triggers TCP congestion control. In difference to the case without context transfer, more frequently TCP fast recovery is triggered instead of TCP slow start. In 80% of 35 access handovers TCP fast recovery is triggered, as summarised in Table 6-9. As a consequence, with SDU context transfer the file transmission time reduces from 428 s to 358 s, thus increasing the object bit rate from 1890 kb/s to 2230 kb/s.

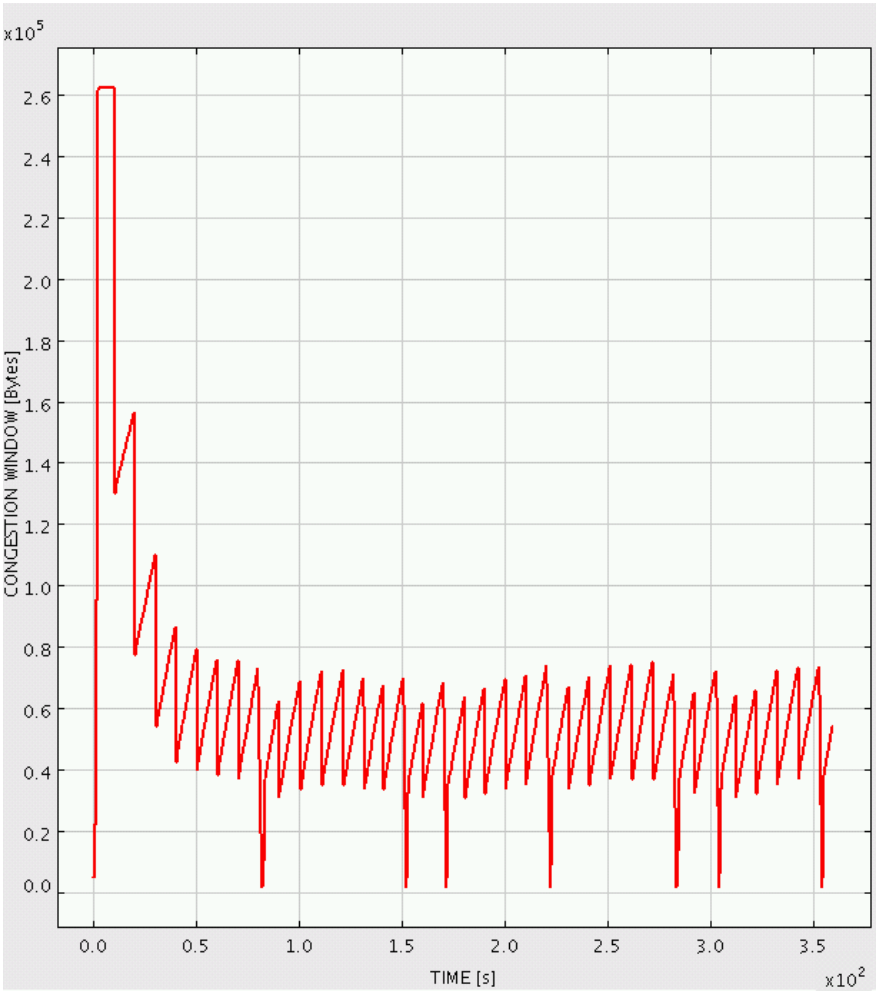


Figure 6.59 : TCP server congestion window for *SDU context transfer*.

Table 6-9 : TCP loss recovery algorithm occurrence for *SDU context transfer*.

Loss Recovery Algorithm	Occurrence in 35 access handovers [%]
Fast Recovery	80%
Slow Start	20%

The TCP trace depicted in Figure 6.60 illustrates the TCP reaction to *SDU context transfer* for the access handover at time $t = 39.972\ s$. At the time of access handover a large number of TCP segments are in flight. The TCP segments marked as 1) are stored in the SDU buffer, whereas those marked by 2) are stored in the PDU buffer and are being transmitted by the link layer. All 45 TCP segments 2) are lost at access handover. The 16 TCP segments from the SDU buffer are transferred to the new link layer, and are subsequently transmitted via the new access. These segments arrive out-of-sequence and thus trigger TCP DUPACKs. The TCP sender reacts by halving its congestion window and by retransmitting missing segments. When all lost segments are retransmitted the sender continues sending new segments. The benefit of SDU context transfer is that the TCP data flow continues immediately after the access handover. It is essential for TCP to have data in flight, such that the resulting ACKs can trigger the error recovery. It is sufficient if the SDU buffer contains more than three TCP segments which lead to enough DUPACKs for TCP to perform fast recovery.

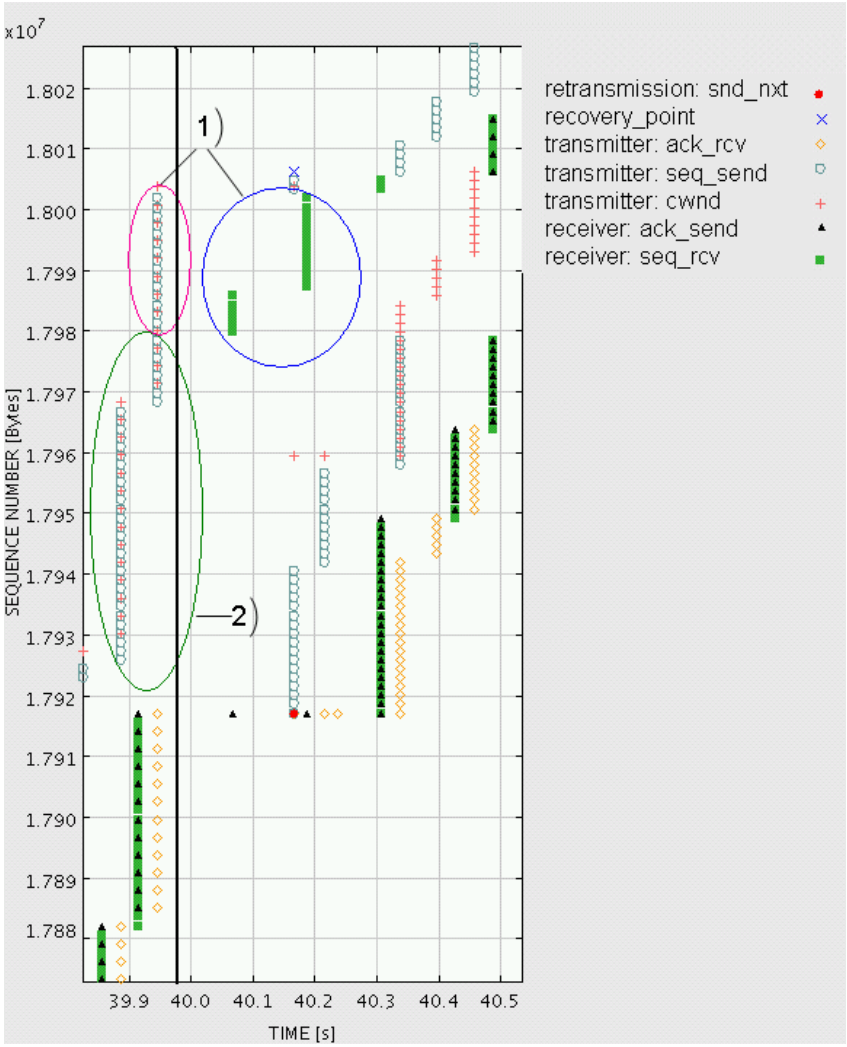


Figure 6.60 : Server TCP trace of fast recovery during *SDU context transfer* (access handover marked with line).

Most of the data which is in flight from the server's TCP point of view is located in the RLC buffer and is discarded.

Table 6-10 summarises the distribution of all the segments that are in flight over the SDU buffer and the RLC buffer in the old transmitting RLC entity just at the time access handover is performed.

Table 6-10 : Distribution of buffered data in a "standard" configuration and ASP = 10 s.

RLC Buffer Type	Distribution buffered data [%]
SDU Buffer	26%
PDU Buffer	74%

The object bit rate of a file transfer with *SDU context transfer* is given in Table 6-11 (cf. Figure 6.52).

Table 6-11 : Number of access handovers and OBR for varying access selection periods.

ASP [s]	OBR [kbit/s]	Confidence Interval 95%	Average Amount of Access Handovers
5	1106	[1085;1127]	71
10	2207	[2178;2236]	18
15	3043	[3015;3070]	8
20	3601	[3580;3621]	5
25	3966	[3949;3983]	4
30	4220	[4205;4236]	3
50	4557	[4550;4564]	2
100	4543	[4526;4559]	1

It can be seen that the performance of *SDU context transfer* is better compared to not using context transfer. The gain is in the range of 4% for very large access selection periods (100s) to up to 24% at very low access selection periods (5s). The performance gain can be explained as follows. The relocation of data by SDU context transfer re-starts the data flow immediately after the access handover *within* the network. It is not required to only rely on the TCP endpoints to trigger the re-start of the TCP transmission procedure (i.e. expiry of the delayed ACK timer in the TCP receiver, or expiry of the retransmission timer in the TCP sender). Instead, the transmission of the relocated buffer leads to a series of TCP DUPACKs, which trigger TCP fast recovery. This reduces the access handover delay. Figure 6.61 shows the time to restart TCP transmission after the handover leading to TCP fast recovery. The delay is reduced from around 190 ms to approximately 100 ms. When the TCP recovery is recovered in the TCP sender by slow start, the access handover delay between the two schemes are the same as depicted in Figure 6.62.

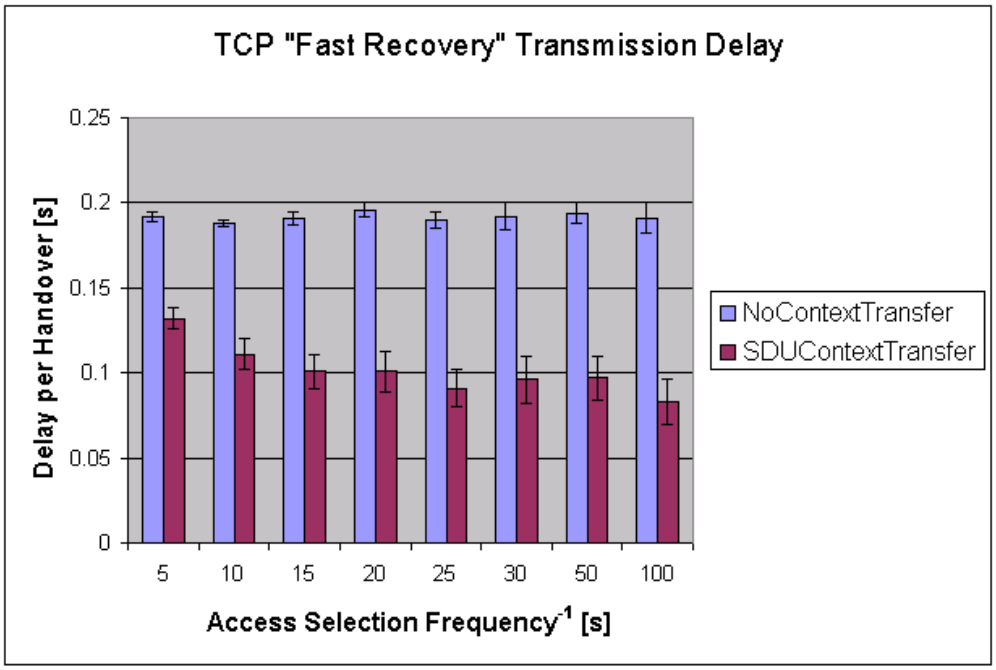


Figure 6.61 : Access handover delay with fast recovery TCP response.

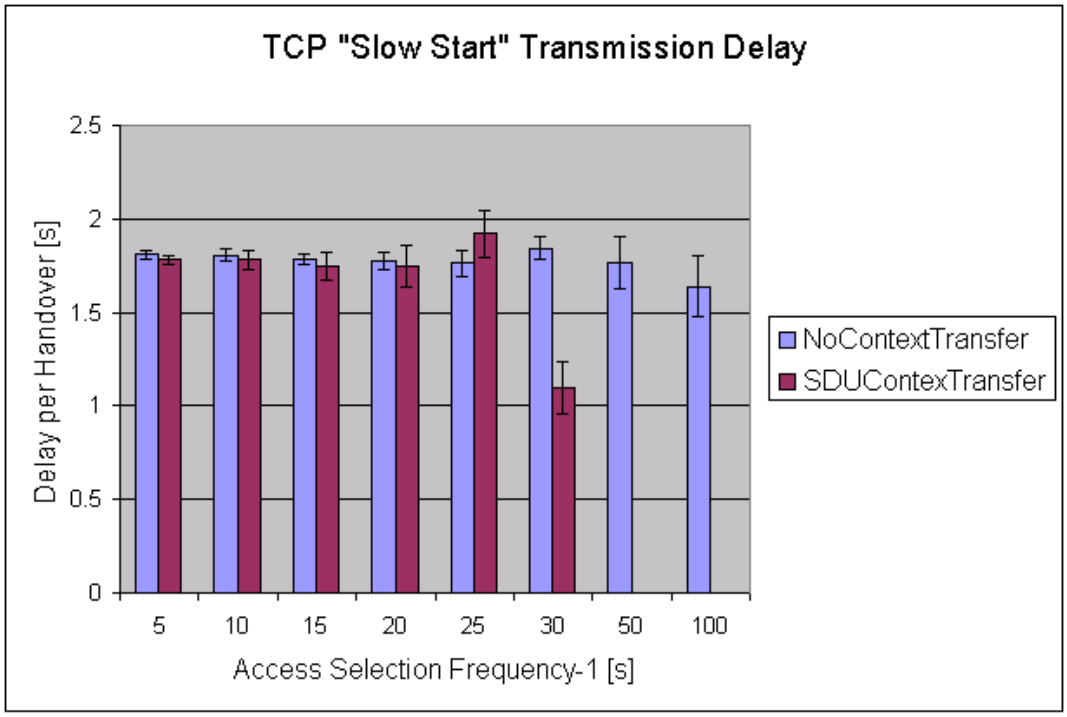


Figure 6.62 : Access handover delays with slow start TCP response.

The fact that some in-flight TCP segments are preserved with SDU context transfer results in more outstanding segments and ACKs that can trigger TCP fast recovery. This is depicted in Figure 6.63. TCP fast recovery is the predominant reaction to access handover. Only at very low access selection periods of 5 s, a significant proportion of TCP slow start events occur. At such frequent handovers, TCP does not reach a stationary congestion avoidance phase before

the next access handover occurs. As a result, the radio link often remains idle and on average the SDU buffer contains less data which can trigger fast recovery after the access handover.

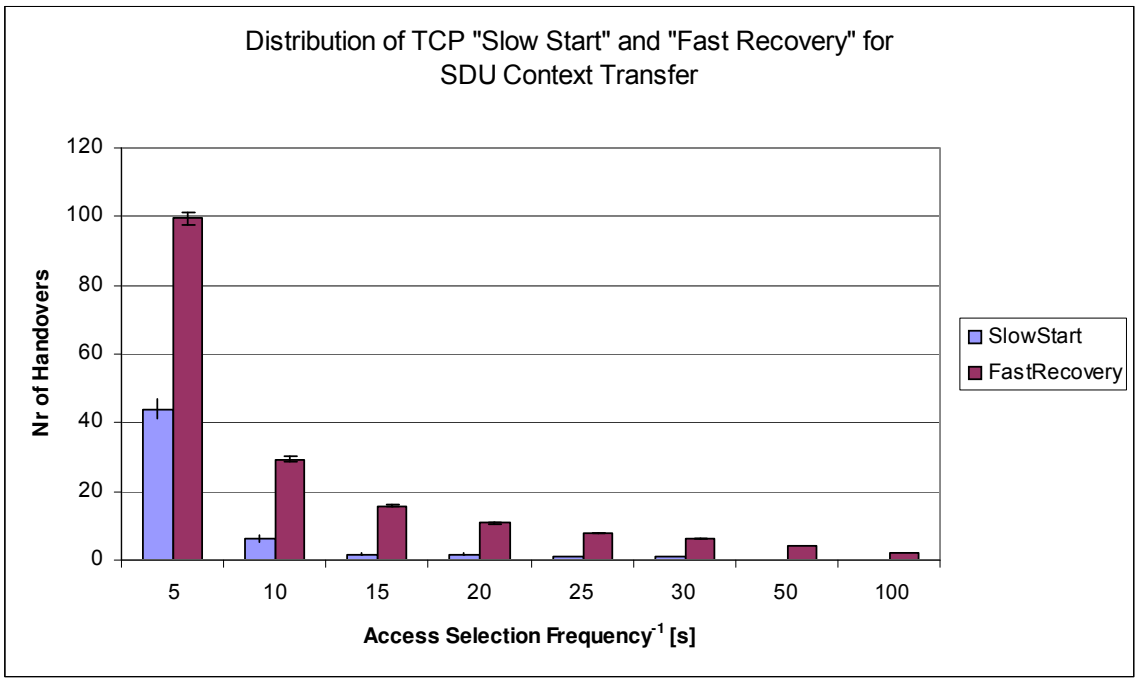


Figure 6.63 : Distribution of TCP loss recovery algorithms for varying switching frequencies.

6.6.3.3 SDU Reconstruction and SDU Context Transfer

Access handover with SDU reconstruction first restores all outstanding SDUs from the PDU buffer before the SDU context transfer is performed. As discussed in Section 6.5.3.3, this access handover scheme is lossless, in difference to the previous two schemes *SDU context transfer* and *no context transfer*. Figure 6.64 shows the TCP congestion control window for one simulation run for an access selection period of 10 s. Totally 20 access handovers occur; 9 of those trigger TCP fast recovery without any TCP slow start (see Table 6-12). The file transfer lasts for 203.25 s, which is equivalent to an object bit rate of 3939 kb/s.

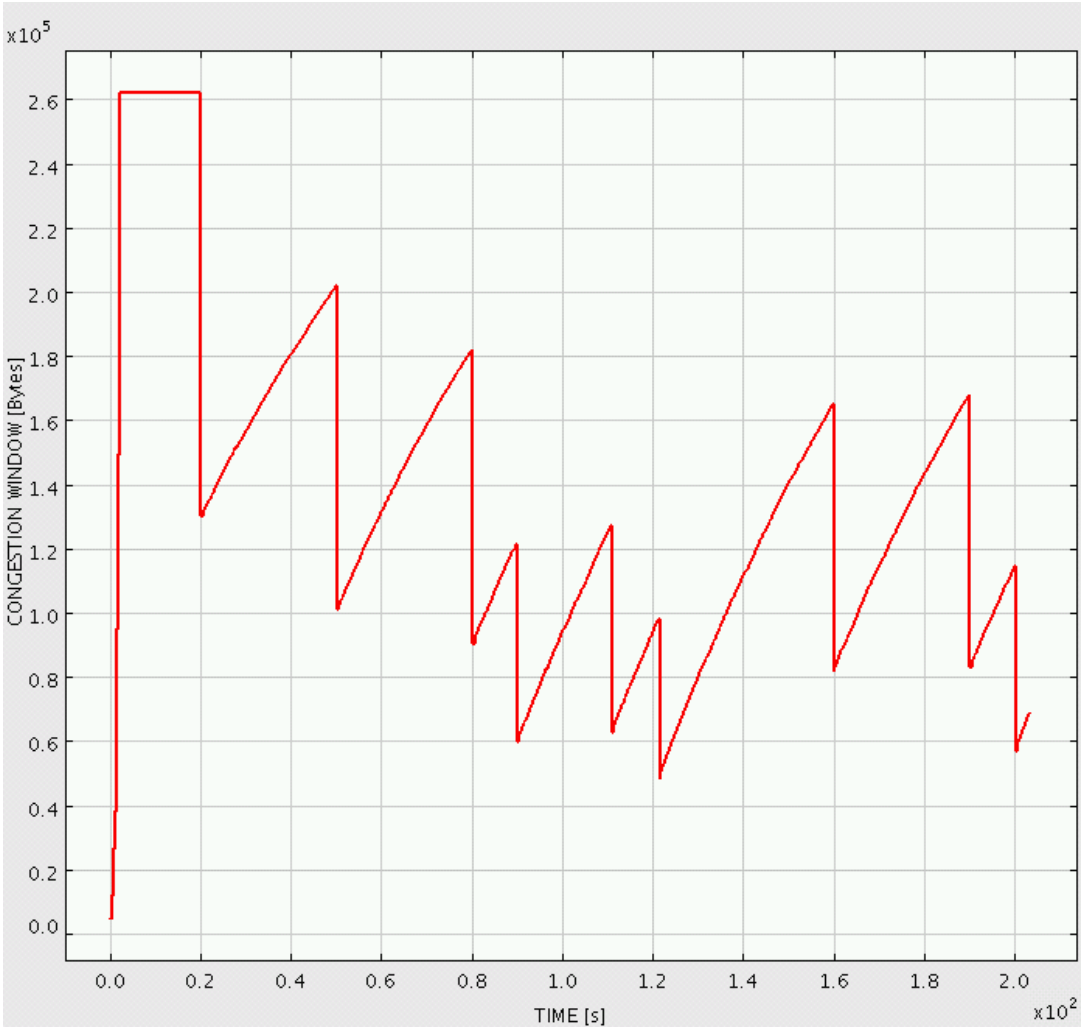


Figure 6.64 : TCP congestion window for *SDU reconstruction*.

Table 6-12 : TCP loss recovery algorithm occurrence for *SDU reconstruction*.

Loss Recovery Algorithm	Occurrence in 20 access handovers [%]
Fast Recovery	45%
No TCP Loss Recovery performed	55%

It may appear counterintuitive that TCP fast recovery is triggered, although the access handover itself is lossless. The explanation can be derived from the TCP trace, which is presented Figure 6.65 for the access handover at time 80.09 s. The segments marked with 1) are all TCP segments which are stored in the SDU buffer at time of handover. We can see from the trace, that some segments that have already been received by the TCP receiver (marked by 2)) are received a second time after the handover. Consequently, these segments are duplicated. The TCP reaction to duplicate segments is to drop them and send a normal cumulative ACK for all correctly received segments. When the first bunch of duplicated TCP segments is received (marked as 3)), a series of DUPACKs (4)) are then sent. These DUPACKs trigger TCP fast recovery, although no data has been lost. The duplication of data leads to a degradation of TCP performance.

Figure 6.65 : TCP trace of fast recovery for *SDU reconstruction* (access handover marked with line).

The duplication of packets can be explained by looking at the RLC trace just before the access handover, which is shown in Figure 6.66. At the moment of access handover the RLC sender and receiver are out of synchronisation. The RLC receiver has successfully received a number of TCP segments and advanced its receiver ARQ window by increasing VR(R) (cf. Figure 6.2). However, the RLC transmitter has not received a status report which notifies it about the correct reception of the data. The PDU is marked with 1) up to which the RLC transmitter knows that all PDUs have been received successfully by the receiving RLC. At time 80.075 s a new status report is transmitted, which is lost during the transmission. The range of the receiver ARQ window reported is marked with 2). The lower edge of the ARQ window can always differ between the sender (i.e. VT(A) in Figure 6.2) and the receiver (i.e. VR(R) in Figure 6.2) due to the delay of status reports. The loss of a status reports amplifies this discrepancy in synchronisation. At the time of the access handover 424 PDUs are outstanding

and not acknowledged (i.e. VR(R) - VT(A) in Figure 6.2), which is equivalent to approximately 45 TCP segments. These TCP segments are thus reconstructed by *SDU reconstruction* at the link layer transmitter at access handover, irrespective of the fact that they have already been received before the access handover. If more than 4 TCP segments are duplicated, TCP fast recovery is triggered.

In fact, not only TCP segments are duplicated but also TCP acknowledgement on the reverse link. However, duplicated TCP ACKs do not trigger TCP fast recovery. These ACKs do acknowledge data below the TCP transmission window and so do not count as DUPACKs [RFC2581]; these duplicated ACKs are silently discarded at the TCP sender without triggering TCP error recovery.

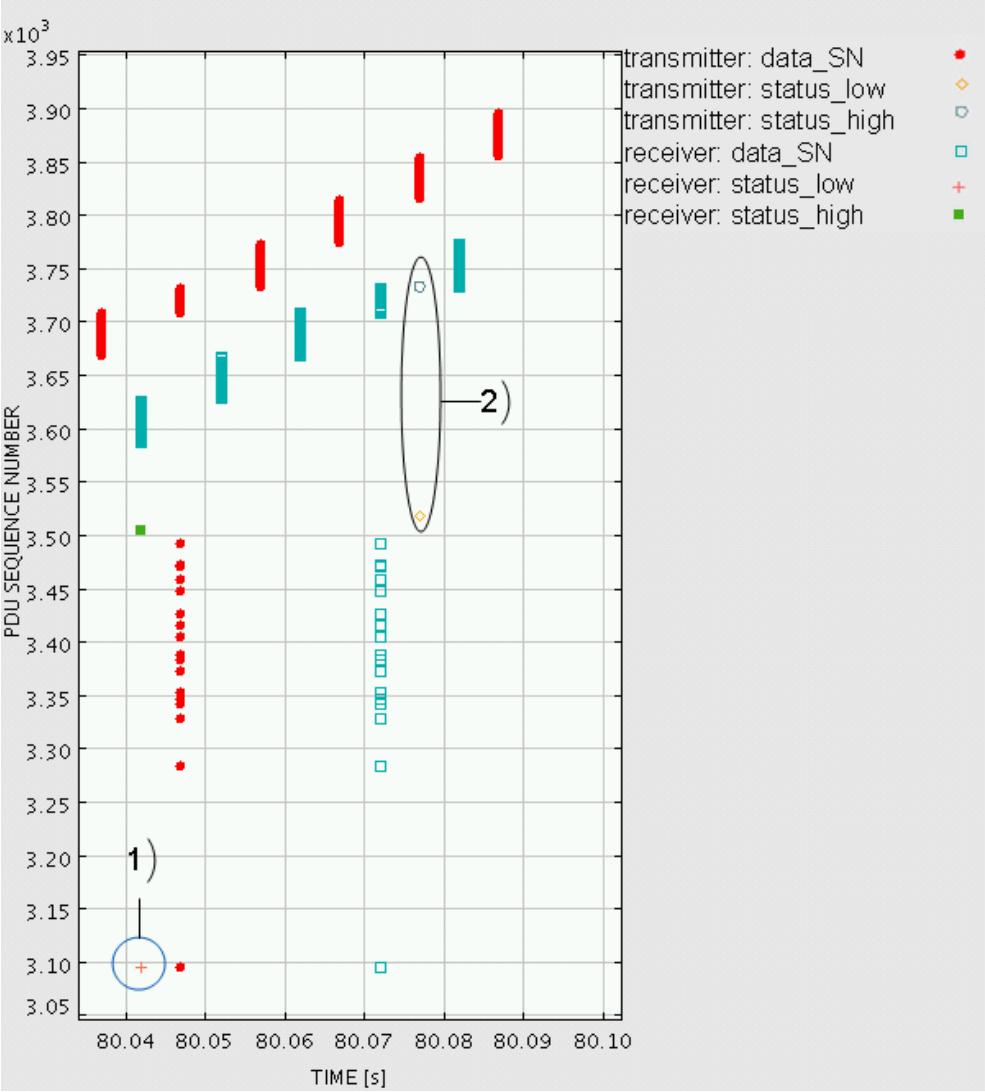


Figure 6.66 : RLC trace for downlink traffic during *SDU reconstruction*.

The amount of lack of synchronisation of link layer ARQ peers depends largely on the radio link characteristics and the link layer configuration. We will see later in Section 6.6.4, how some parameters influence the performance of SDU reconstruction. In the link layer configuration considered here, around 40% of the access handovers lead to sufficient packet

duplication to trigger TCP fast recovery, as seen in Figure 6.67. The access handover delay for the cases of TCP fast recovery is around 50 ms as shown in Figure 6.68.

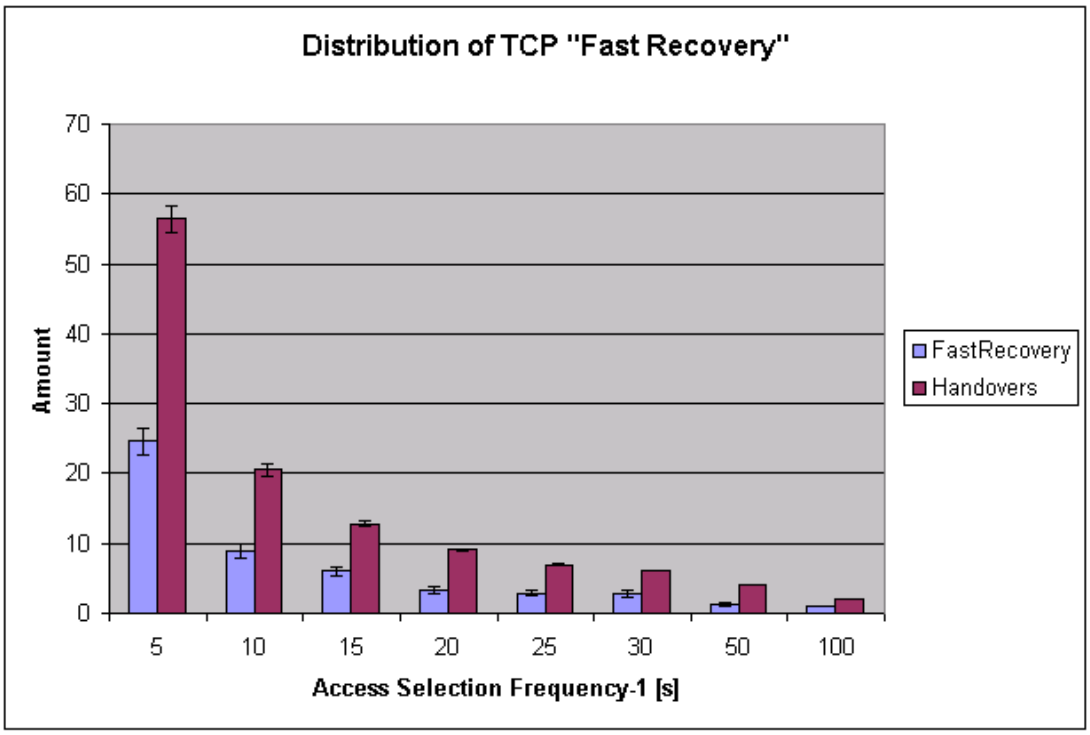


Figure 6.67 : Frequency of TCP fast recovery out of total number of handover.

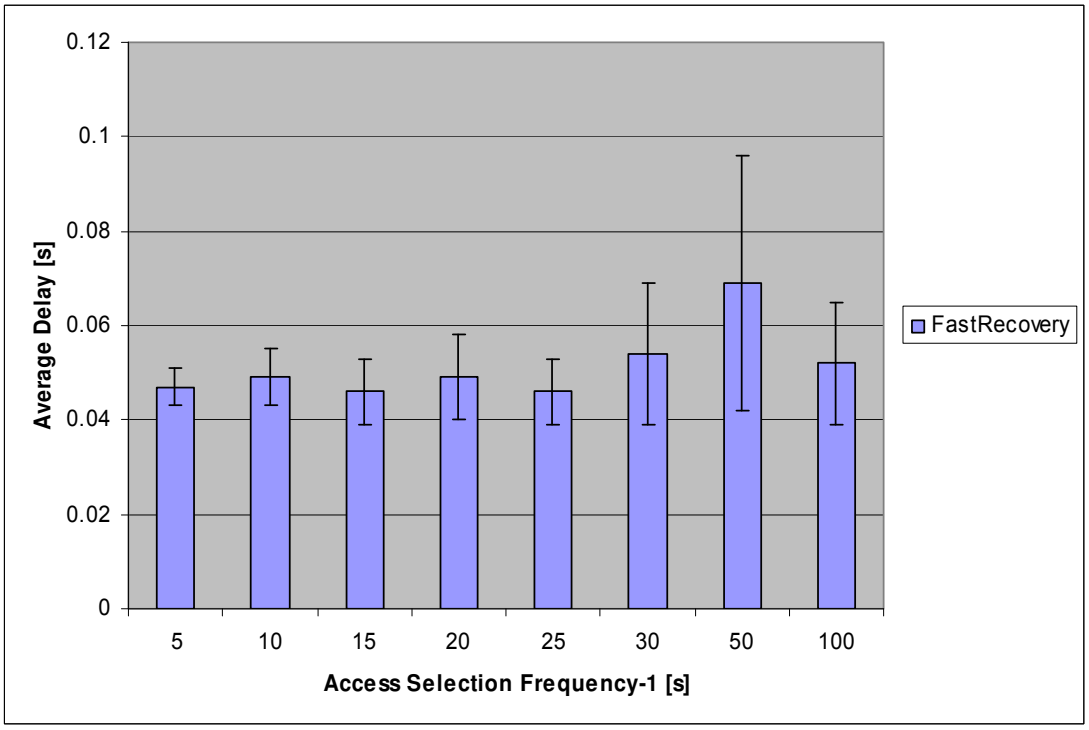


Figure 6.68 : Average Delay introduced by TCP fast recovery.

The problem of packet duplication in SDU reconstruction can be eliminated. For this it is required to synchronise the ARQ window of the link layer transmitter and receiver before performing SDU reconstruction. This is shown in Figure 6.38 as “transmission status update”. This procedure introduces an additional delay of at least one link layer round-trip time to the access handover delay. However, it then achieves a lossless handover without packet duplication.

6.6.3.4 Layer 2 Tunnelling

In access handover with *layer 2 tunnelling*, as described in Section 6.5.2.3.1, the access handover is optimised by link layer mechanisms, and context transfer is used for context exchange between the old and new link layer entities. *Layer 2 tunnelling* is a lossless handover scheme. In addition to SDU context transfer, the content of the old PDU buffer is transferred to the new link layer entities. There it is transmitted in a tunnel to the new link layer receiver, from where it is forwarded to the old link layer receiver, where IP packets are reconstructed. Any packet duplication is eliminated within the link layer. At the IP layer it is thus also perceived as a scheme without packet duplication. As discussed in Section 6.5.2.3.1, for *layer 2 tunnelling* the old and new link layer entities remain active during the access handover procedure; only when the tunnelling procedure is finished the old link layer entities are discarded. Figure 6.69 shows the TCP congestion window development with *layer 2 tunnelling* for one simulation run. We can see that the congestion window remains completely unaffected by the access handovers. In this scenario, the advertised window of the TCP receiver is slightly lower than the size of the SDU and PDU buffers. Therefore we do not see the sawtooth pattern caused by occasional overflows of the SDU buffer. In this example, the *layer 2 tunnelling* access handover achieves an object bit rate of 4697 kb/s.

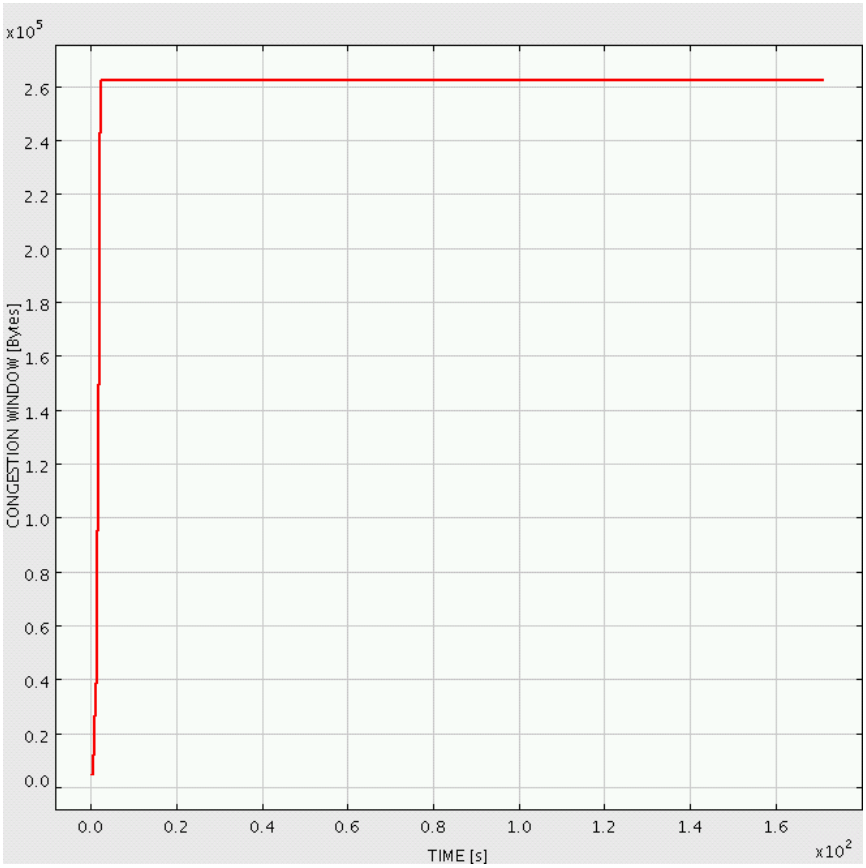


Figure 6.69 : TCP congestion window for *layer 2 tunnelling*.

In Figure 6.70 the TCP trace for the access handover at time t = 50.109 s is shown. For TCP the access handover is seamless without any disturbance, all recovery is performed at the link layer. At the time of the access handover the TCP segments marked with 1) are still in stored in the SDU buffer; the TCP segments marked with 2) are already in the PDU buffer and link layer transmission for those segments is ongoing.

Figure 6.70 : TCP Trace during *layer 2 tunnelling* (access handover marked with line).

The operation of *layer 2 tunnelling* can be seen by looking at the RLC trace of the old link layer entities for the same time interval of access handover, which is presented in Figure 6.71. The first observation is that the old RLC entities are still active after the handover, although the transmission over the corresponding physical layer has already stopped. The RLC trace continues until approximated 200 ms after the access handover. For the other access handover schemes, the old link layer entities are deleted at the access handover. There are two categories of RLC PDUs that are received after the handover. The first category comprises RLC retransmissions, which are marked by 1) in Figure 6.71. These are non-consecutive PDUs that have been requested by the status report (marked by 2)) as not correctly received. After the access handover, the RLC transmitter forwards all non-acknowledged PDUs into the L2 tunnel, including those retransmissions. The second category of RLC PDUs contains those PDUs, which have been transmitted for a first time prior to the access handover, but for which no status report has been received yet. Those PDUs are marked as 3) in Figure 6.71. Several PDUs are transmitted via the L2 tunnel after the access handover, although they have already been received correctly at the RLC receiver before the access handover. However, since no status report has been received in time, the RLC sender is not aware of it. The reception of

duplicates after the access handover is handled by the old RLC receiver by discarding them. Treatment of duplicate transmissions is a normal feature of RLC; for example, it can occur when status reports are lost or when ARQ timers are not properly adjusted to the link layer round-trip time. Consequently, this duplication is not visible to the above protocol layers like TCP and IP. If a synchronisation of the ARQ windows of the old transmitter and receiver RLC entities is performed prior to the access handover, the data duplication can be avoided.

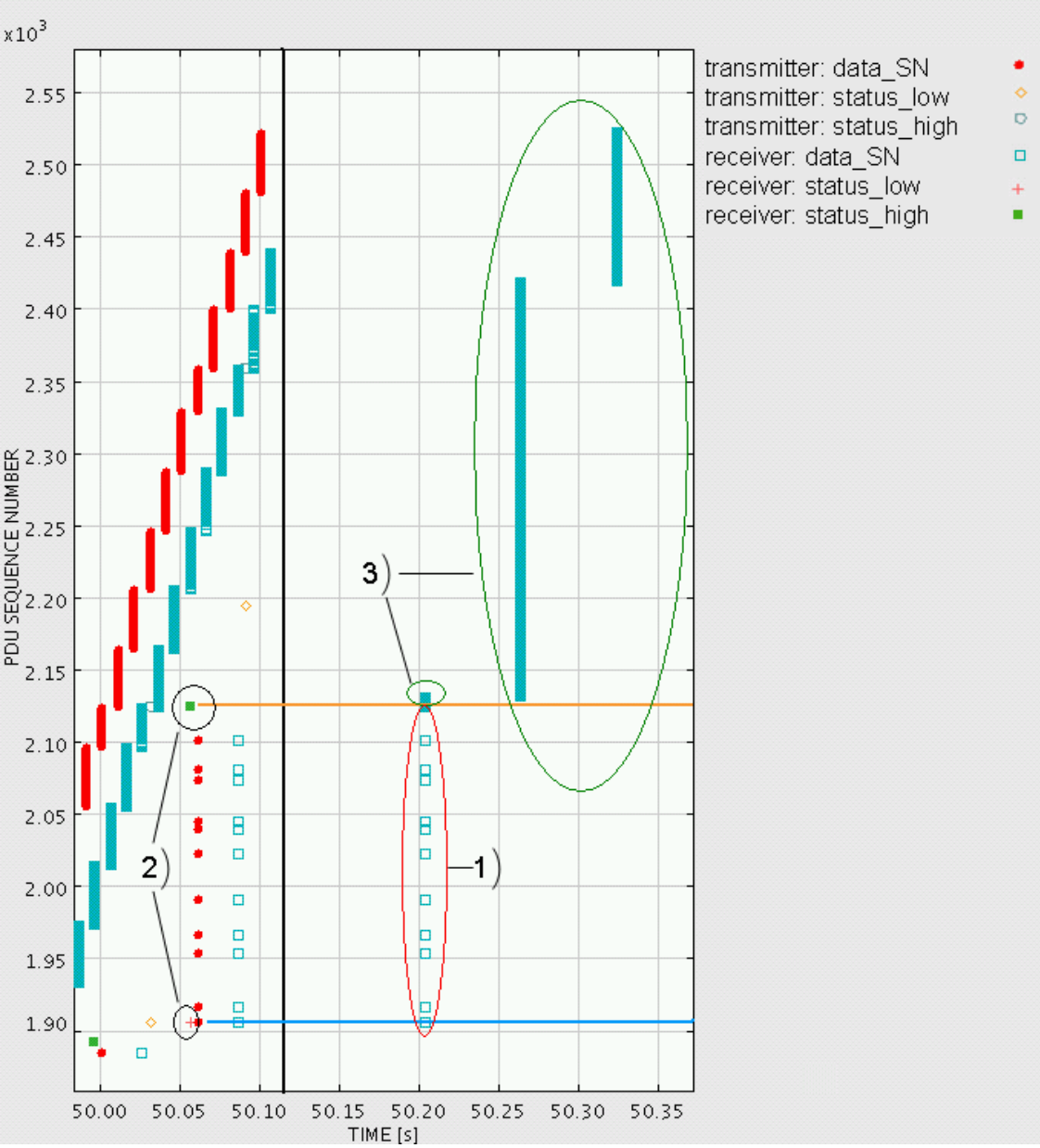


Figure 6.71 : RLC trace during *layer 2 tunnelling* (access handover marked with line).

The procedure of *layer 2 tunnelling* introduces some overhead, which is the amount of PDUs that are unnecessarily transmitted via the layer 2 tunnel. This overhead wastes some transmission resources for the new radio link. Since mainly those PDUs are transferred that have been unsuccessfully transmitted by RLC previously, the overhead of *layer 2 tunnelling* is smaller than the overhead of SDU reconstruction. Furthermore, additional overhead by TCP error recovery is avoided. Figure 6.72 shows the total amount of tunnelled PDUs, as well as

the overhead, which are those PDUs that the old RLC receiver had already received correctly. The amount of overhead depends on the radio link characteristics of the old link (data rate, PDU size, transmission time interval, RLC round-trip time, block error rate) and the configuration of the old RLC ARQ toolbox (ARQ timers and status reporting frequency). An analysis of those parameters on the performance is presented in Section 6.6.4.

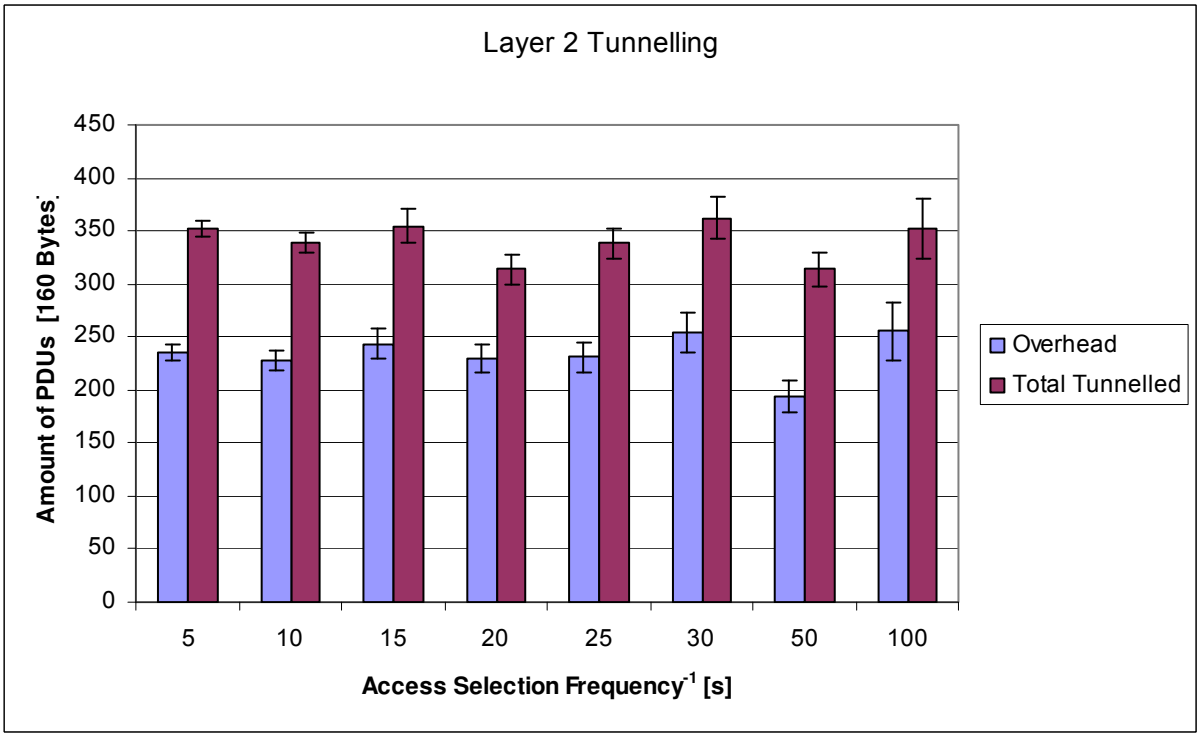


Figure 6.72 : ***Layer 2 tunnelling*** **efficiency**

6.6.3.5 Summary and Conclusion

Access handover without optimisation (denoted as *no context transfer*) leads to data loss and requires loss recovery from TCP. It typically leads to TCP timeouts with slow start; if a TCP acknowledgement is triggered after the access handover by TCP delayed ACK timer at the receiver, this ACK can restart the TCP data flow and lead to TCP fast recovery. Every access handover leads to a performance degradation of TCP throughput.

Access handover with *SDU context transfer* also leads to a loss of the data contained in the PDU buffer, and thus requires TCP loss recovery. The access handover delay is reduced compared to *no context transfer* as the transferred SDU buffer can faster re-establish the TCP data flow. More frequently TCP fast recovery is used as loss recovery mechanism and fewer TCP time-outs occur.

Access handover with *SDU reconstruction* is a lossless handover scheme. However, it can lead to packet duplication, when data is reconstructed from the PDU buffer of the link layer sender that has already been received correctly at the link layer receiver without being reported back to the sender. If the amount of duplication exceeds four TCP segments this leads to an unnecessary TCP fast recovery procedure.

Access handover with *layer 2 tunnelling* is also a lossless handover scheme. It requires a context transfer procedure between the old and new link layer entities at both the sender and receiver. It can also lead to packet duplication, however this duplication is treated by the link layer and it is not visible to higher protocol layers. It only leads to some transmission overhead. *Layer 2 tunnelling* does not lead to any unnecessary TCP error recovery reactions.

Table 6-13 summarises how the different access handover schemes are perceived by higher protocol layers like TCP.

Table 6-13: Data loss and duplication for the different access handover schemes.

	PDU Buffer	SDU Buffer	SDU duplication
No Context Transfer	loss	loss	No
SDU Context Transfer	loss	lossless	No
SDU Reconstruction	lossless	lossless	Yes
Layer 2 Tunnelling	lossless	lossless	No

The TCP performance of the different access handover schemes for different access selection periods is shown in Figure 6.52. Table 6-14 shows the relative gain of the different access handover schemes compared to non-optimised access handover (i.e. *no context transfer*). At very infrequent access handovers (every 50-100 s) the gain of the optimisation schemes is below 10%. The difference between the schemes is small. If access handovers occur more frequently (every 20-50 s) optimised access handover schemes provide substantial gain of up to 50%. The gain of the different schemes varies significantly. Where *SDU context transfer* can provide gains from 9%-11%, *SDU reconstruction* can provide gains of 17%-38%, and *layer 2 tunnelling* achieves gains of 21%-46%. For very frequent access handovers (below every 20 s) the non-optimised scheme performs very poor. Even if *SDU context transfer* can provide gains up to 24%, it also performs rather poor. *SDU reconstruction* can increase the performance by a factor of up to 3, and *layer 2 tunnelling* even by a factor of up to 5.

Table 6-14: Gain of different access handover schemes compared to non-optimised handover.

ASP [s]	NoContextTransfer	SDUContextTransfer	SDUReconstruction	L2Tunnelling
5	100%	124%	311%	522%
10	100%	119%	207%	253%
15	100%	115%	156%	178%
20	100%	111%	138%	146%
25	100%	109%	124%	129%
30	100%	109%	117%	121%
50	100%	104%	107%	108%
100	100%	103%	103%	104%

6.6.4 Influence of Different System Parameters on Access Handover Performance

In the previous section we have presented the performance of access handover schemes for a standard configuration of the radio link layer parameters. In this section we investigate which influence different system parameters have on the performance, as shown in Table 6-15. The other system parameters are as described in Section 6.6.2.

Table 6-15: Sensitivity analysis towards different link parameters.

Section	Parameter	Standard Configuration[65]	Variation
6.6.4.1	Link BLER	5%	[0, 0.01, 0.1, 1, 5, 10, 15] %
6.6.4.2	Link RTT	50 ms	[0, 10, 50, 75, 100] ms
6.6.4.3	L2 PDU size	160 byte	[80, 160, 640, 1503] byte

6.6.4.1 Influence of Block Error Rate

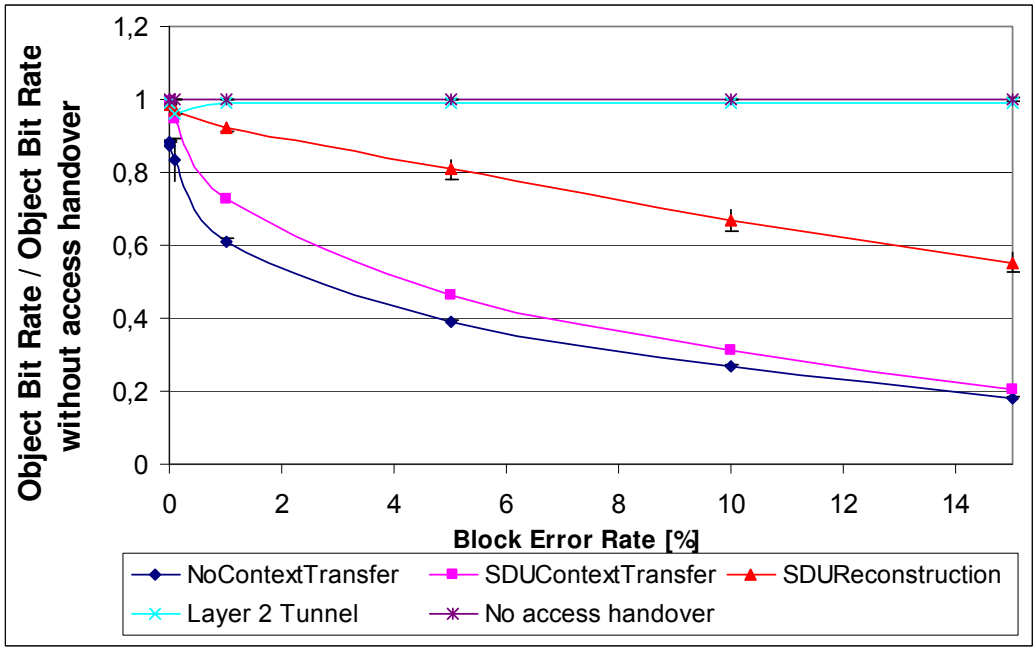


Figure 6.73 : Performance of access handover for different BLER (access selection period of 10 s).

In Figure 6.73 the influence of block error rate on performance is depicted for a fixed access selection period of 10 s. The performance is given, relative to the same scenarios without any access handover, i.e. with infinite access selection period. *Layer 2 Tunnel* shows no performance degradation independent of the BLER. For an error free channel with 0% BLER the performance loss due to access handover for *no context transfer* is 13%, and for the other three schemes it is around 1%. For increasing BLER the performance of *SDU Reconstruction*, *SDU context transfer* and *no context transfer* decreases. This can be explained by the increasing size of the RLC ARQ window with increasing BLER. An increasing ARQ window means that a larger amount of data is lost for *SDU context transfer* and *no context transfer*. Without context transfer a performance loss of 40% is found at 1% BLER, with a loss of 73% at 10% BLER and a loss of more than 80% at 15% BLER. *SDU context transfer* performs slightly better, with a loss of 27% at 1% BLER, a loss of 69% at 10% BLER and a loss of 79% at 15% BLER. For *SDU reconstruction* a larger ARQ window increases the amount of packet duplication and thus unnecessary TCP interactions. This leads to a loss of 7% at 1% BLER, a loss of 33% at 10% BLER and a loss of 44% at 15% BLER. For *Layer 2 tunnelling* the increased ARQ window has little influence, as all access handover distortion is hidden

[65] Configuration used in section 6.6.3.

from the TCP layer. The performance of *Layer 2 tunnelling* is independent of the BLER and maintains a loss of below 4%.

6.6.4.2 Round Trip Time Variation

Figure 6.74 shows the performance of the different access handover schemes when the link layer round-trip time is varied. For small link layer RTT values of 10-25 ms the performance degradation of *SDU reconstruction* and *SDU context transfer* is only 3-10% lower than without access handover. The corresponding performance loss of *no context transfer* is in the order of 10-20%. For link layer RTTs beyond 25 ms the performance loss becomes large. At 50 ms RTT the normalised performance of *no context transfer* is 39% and at 100 ms RTT only 17%. At RTTs larger than 25 ms *SDU context transfer* performs only 7-2% better than without context transfer. The reason is that with larger RTT the ARQ window size increases, and therefore the RLC buffer becomes more significant than the SDU buffer. *SDU Reconstruction* is much less sensitive to BLER and RTT compared to *no context transfer* and *SDU context transfer*. When the PDU buffer is also transferred at access handover with *SDU reconstruction*, the performance increases to 81% at 50 ms RTT and 41% at 100 ms RTT. *Layer 2 tunnelling* remains unaffected by BLER and the performance is only up to 2% below the reference case without access handover.

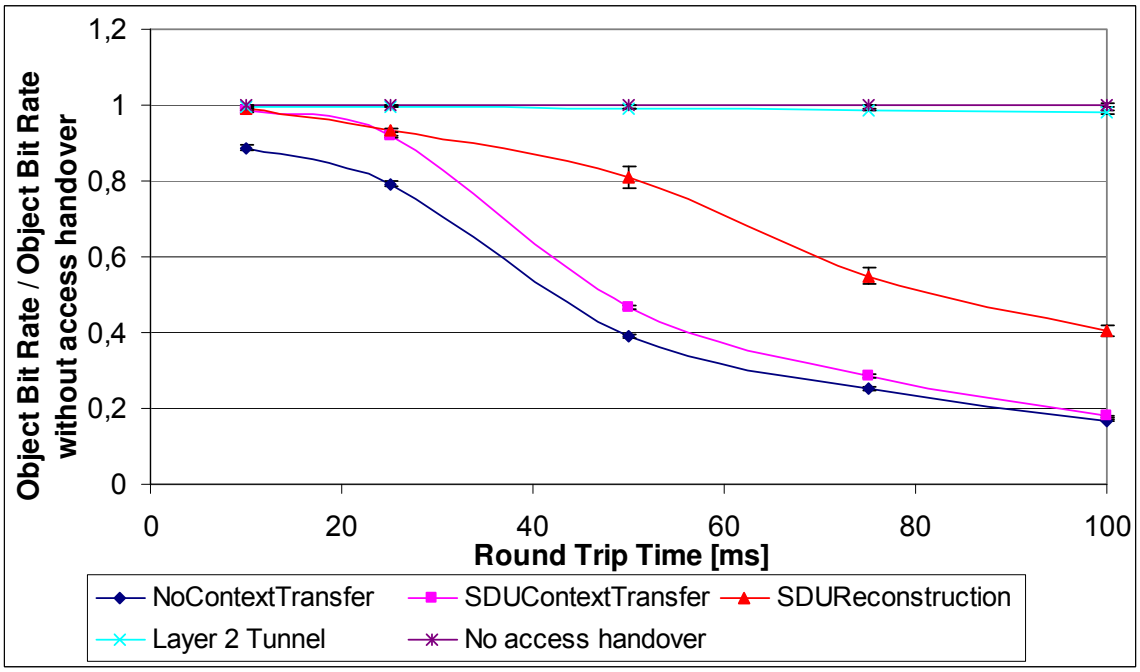


Figure 6.74 : Performance of access handover for different RTT (access selection period of 10 s).

6.6.4.3 RLC PDU Size Variation

Figure 6.75 shows the dependency of the access handover performance compared to not performing access handover, when the PDU size is varied. The PDU sizes of 80, 160 and 320 bytes are typical PDU sizes used in UMTS. A PDU size of 1503 byte is sufficient to include a full IP packet. Smaller PDU sizes have smaller amount of padding overhead due to segmentation compared to larger PDU sizes. The performance of access handover increases

with larger PDU size with our assumption of a constant block error rate which is independent of the block size[66]. At a PDU size of 80 byte *no context transfer* achieves a performance of only 35%. This increases to up to 64% for 1503 byte PDU size. *SDU context transfer* performs a bit better, with a performance of 41% at 80 byte PDU size and 75% at 1503 byte PDU size. The performance of *SDU reconstruction* is significantly better than that for the other schemes. At a PDU size of 80 byte the normalised performance is 78% and at a PDU size of 1503 it is 89%. *Layer 2 tunnelling* is not affected by the PDU size, the normalised performance is at 99% for all PDU sizes.

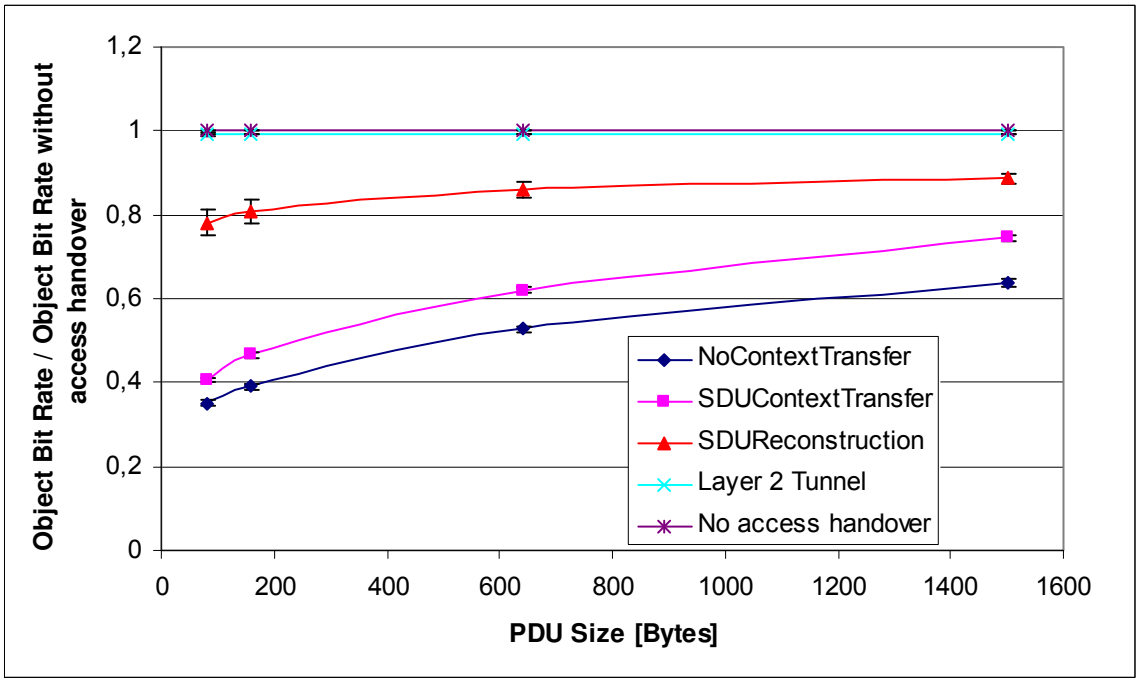


Figure 6.75 : Performance of access handover for different PDU sizes (access selection period of 10 s).

6.6.4.4 Summary

In this section we have investigated the influence of three link layer parameters, the block error rate, the round-trip time and the PDU size on the performance of access handover. The results can all be explained with the same effect: with increasing BLER, increasing RTT and decreasing PDU size the ARQ window pipe capacity of the radio connection increases [PM01]. As a consequence, the sizes of the PDU buffers grow. For the lossy access handover schemes, *no context transfer* and *SDU Reconstruction*, this automatically leads to a larger amount of packet loss at access handover. For *SDU Reconstruction* the larger PDU buffers increase the probability of packet duplication, which can also trigger TCP congestion control to reduce the transmission rate. For *layer 2 tunnelling* the increased PDU buffer has little influence. *Layer 2 tunnelling* has no interaction with TCP congestion control. The increased

[66] In real systems there is typically a dependency of PDU size and block error rate. If a radio link has a certain radio link quality this results in a certain bit error probability. Larger PDU sizes therefore lead to larger block error rates. However, if a further layer of hybrid ARQ is used, like in HSPA and LTE, the BLER is almost independent of PDU size.

buffer size only requires that some more PDUs need to be tunnelled; this effect is negligible for the complete file transfer.

6.6.5 Performance Evaluation in a Heterogeneous RAT Environment

In the previous sections we have investigated access handover between two radio links with the same characteristics in data rate and delay. In this section we examine access handover between different radio access technologies. At an access handover the data rate and delay of the old and new radio link change. We consider three different RAT configurations, as listed in Table 6-16. The radio link model used in the simulation is based on UMTS. For the other two RATs, HSPA and a future RAT (FRAT), we use the same simulation model as for UMTS with modified parameters. FRAT could be the 3GPP Long-Term Evolution [EFKMP+06], 3GPP evolved HS [3GPP25.913] or the WINNER radio interface [PHDSP+06]. The data rate of the radio links is larger than for UMTS and the round-trip times are smaller, as shown in Table 6-16. We also increase the PDU sizes to avoid stalling of the ARQ window at high rate. HSPA deploys hybrid ARQ at the physical layer, which we also assume for FRAT. This is not implemented in the simulation model. Instead, we model this by reducing the block error rate that is visible to the RLC layer. The transmission time interval (TTI) we assume as constant. For FRAT we slightly adapt the TTI, such that an integer number of PDUs can be transmitted per TTI. Furthermore, the SDU buffer size must be sufficiently large to avoid under-utilisation of the radio link. We configure the SDU buffer size according to the RAT with the larger bandwidth-delay product.

Table 6-16: Simulator configuration for radio access technologies.

	r [Mbps]	BLER [%]	RTT [ms]	PDUSize [byte]	TTI [ms]
FRAT	40	0.01	16	1503	10.23
HSPA	4	0.1	60	160	10
UMTS	0.384	10	80	40	10

For investigation of heterogeneous access handover we consider two different scenarios. The first scenario is similar to the approach in Sections 6.6.3 and 6.6.4 for the access handover between homogeneous radio links. During the transmission of a large file (216 MB) we perform regular access handover between the two radio links. This means that we perform multiple access handovers from the radio link with the low data rate to the radio link with the high data rate (i.e. *up-switching*) and vice versa (i.e. *down-switching*). However, it is desirable to separate those two cases. We want to see how the different access handover schemes influence either up-switching or down-switching. Therefore, we modify the simulation model such that the access handover from RAT 1 to RAT 2 is always according to the configured access handover scheme; the reverse direction of access handover from RAT 2 to RAT 1 always uses the *layer 2 tunnelling* access handover. We have previously seen, that *layer 2 tunnelling* has no interaction with TCP and can be considered as an “ideal” access handover scheme. By configuring either RAT 1 or RAT 2 as the radio link with higher data rate, we can then separately investigate the access handover performance of up-switching and down-switching. This investigation is described in Section 6.6.5.1 with a focus on access handover between HSPA and FRAT. A disadvantage of this scenario is that with the increasing data rates of HSPA and FRAT only few access handover events take place during the file download, in particular at higher access selection periods.

In a second scenario we examine the transmission of a comparatively small file of 5 MB. We now consider a single access handover during the transmission of this file. This provides us with a separate investigation of up-switching and down-switching access handover. It also emphasises the influence of the access handover on the total transmission time. The access handover is performed randomly with equal probability in the time interval between 0.5 s and 1 s during the file download. This scenario is described in Section 6.6.5.2 and the UMTS, HSPA and FRAT are considered.

6.6.5.1 Bulk File Transfer

We investigate the periodic access handover between FRAT and HSPA for a bulk file download. For down-switching from the FRAT radio link to the HSPA radio we use the different access handover schemes; for up-switching, i.e. the access handover from HSPA to FRAT, always *layer 2 tunnelling* is performed. Figure 6.76 shows the performance of the different schemes. We see that there is no significant difference between the different access handover schemes. The achievable object bit rate is around 22 Mb/s, which is slightly more than the average link rate of HSPA and FRAT. This is due to the fact, that FRAT is used for a slightly longer time during the download. At higher access selection periods there are some variations in the achieved object bit rate, depending on if FRAT or HSPA is used for a longer period of time during the file transfer. Only if access handover becomes very frequent, with access selection periods below 5 s, a difference between the access handover schemes can be noted. *Layer 2 tunnelling* shows no performance degradation; *no context transfer* performs worst. *SDU reconstruction* and *SDU context transfer* show similar performance, which is between the performance of *layer 2 tunnelling* and *no context transfer*. We can summarise that the influence of the access handover scheme is not as significant as in the heterogeneous handover case investigated in Sections 6.6.3 and 6.6.4.

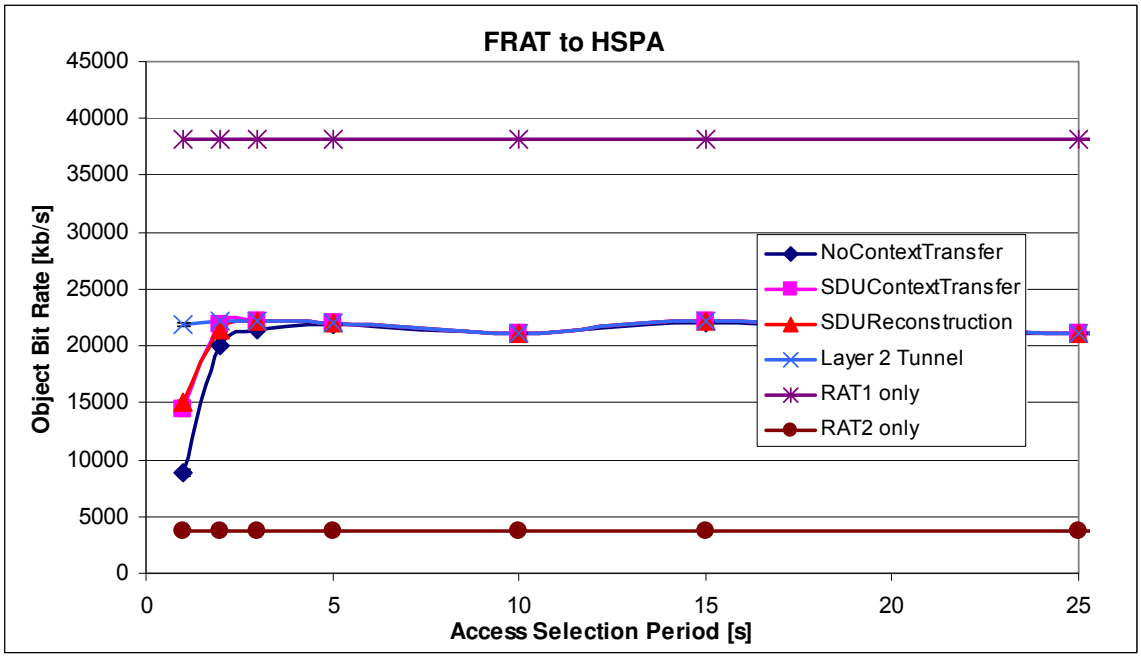


Figure 6.76 : Object bit rate for access handover (down-switching) from FRAT to HSPA.

Figure 6.77 shows the access handover performance for the access handover for up-switching from HSPA to FRAT; in the downlink *layer 2 tunnelling* is always performed. The achievable object bit rate is around 20 Mb/s; this is slightly lower than in the down-switching case, since this time HSPA is used for a slightly longer time than FRAT. The general observation is similar to the down-switching case. However, *no context transfer* always performance worse and the degradation becomes significant at already higher access selection periods of 15 s and below.

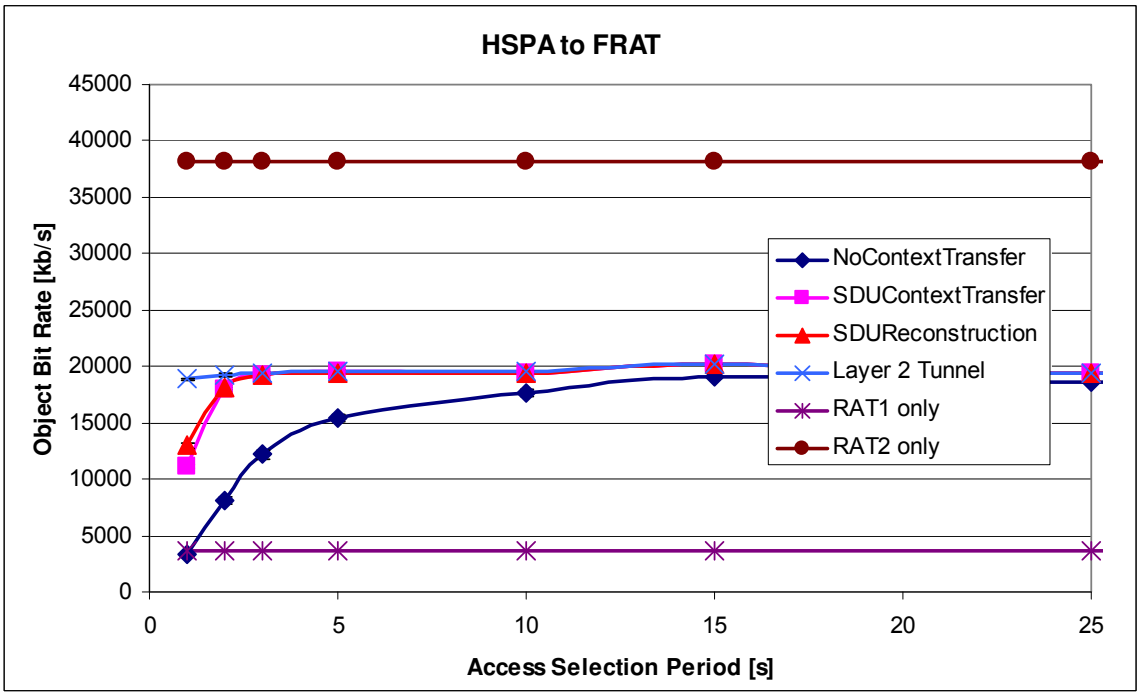


Figure 6.77 : Object bit rate for access handover (up-switching) from HSPA to FRAT.

6.6.5.2 Single Access Handover during Transfer of Small File

We now consider a scenario where we transmit a small file and perform a single access handover switch during this file transfer. The considered RATs are UMTS, HSPA and FRAT. Section 6.6.5.2.1 summarises the down-switching case, i.e. the access handover from a RAT with high data rate to another one with low data rate. In Section 6.6.5.2.2 we investigate the up-switching from the low data rate to the high data rate.

6.6.5.2.1 Access Handover to a Slower Link (down-switching)

An access handover from FRAT to HSPA is shown Figure 6.78. *SDU context transfer*, *SDU reconstruction* and *Layer 2 tunnelling* achieve an object bit rate of the file transfer which is 4% larger compared to *no context transfer*. There is no significant difference in performance between the optimised access handover schemes.

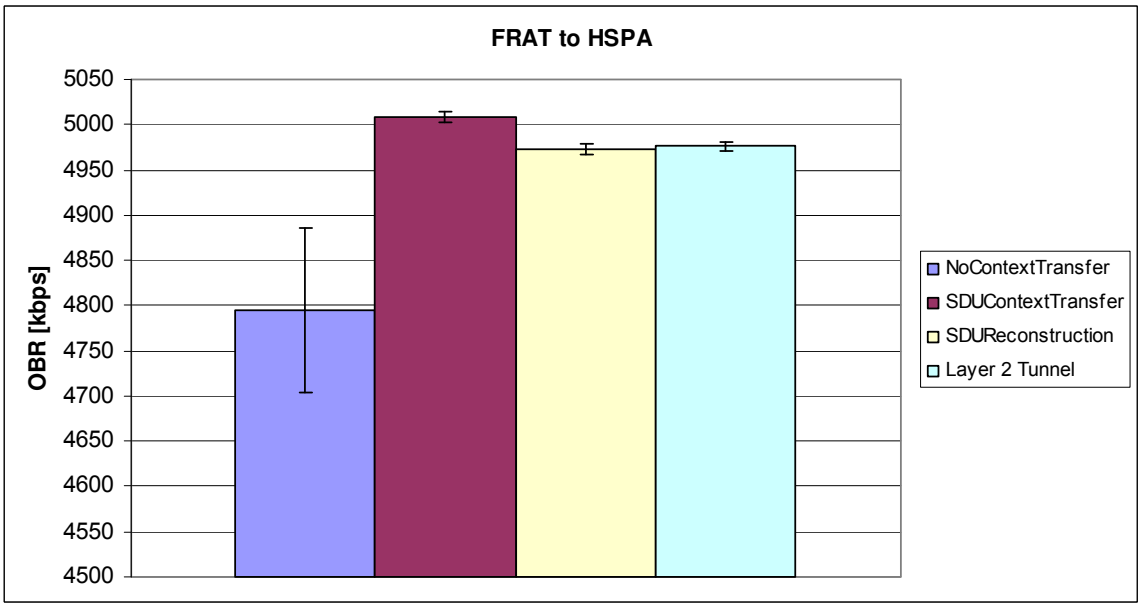


Figure 6.78 : Performance with a single access handover from FRAT to HSPA.

Figure 6.79 shows the result for an access handover from FRAT to UMTS. Again, *SDU context transfer*, *SDU reconstruction* and *Layer 2 tunnelling* show a better performance than *no context transfer*. The gain of *SDU context transfer* and *SDU reconstruction* is 4%, and the gain of *layer 2 tunnelling* 9%. There is no significant difference between *SDU context transfer* and *SDU reconstruction*.

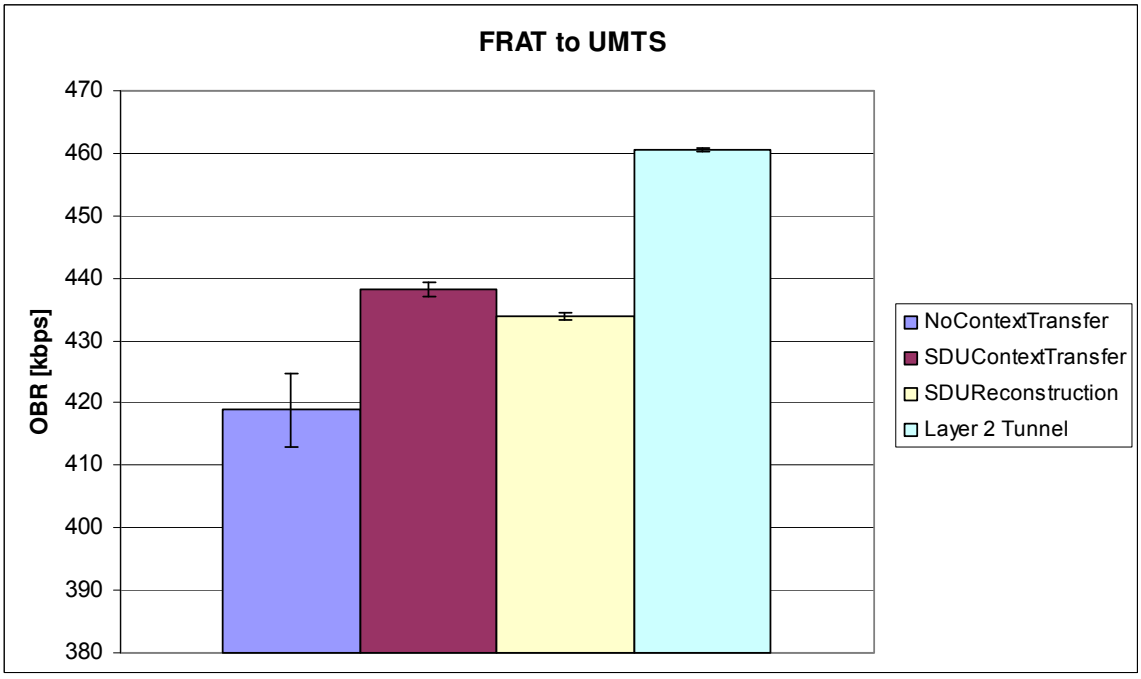


Figure 6.79 : Performance with a single access handover from FRAT to UMTS.

The access handover scenario from HSPA to UMTS is depicted in Figure 6.80. There is no significant difference in performance between the access handover schemes; *layer 2*

tunnelling has the best performance which is approximately 1% better than for the other schemes.

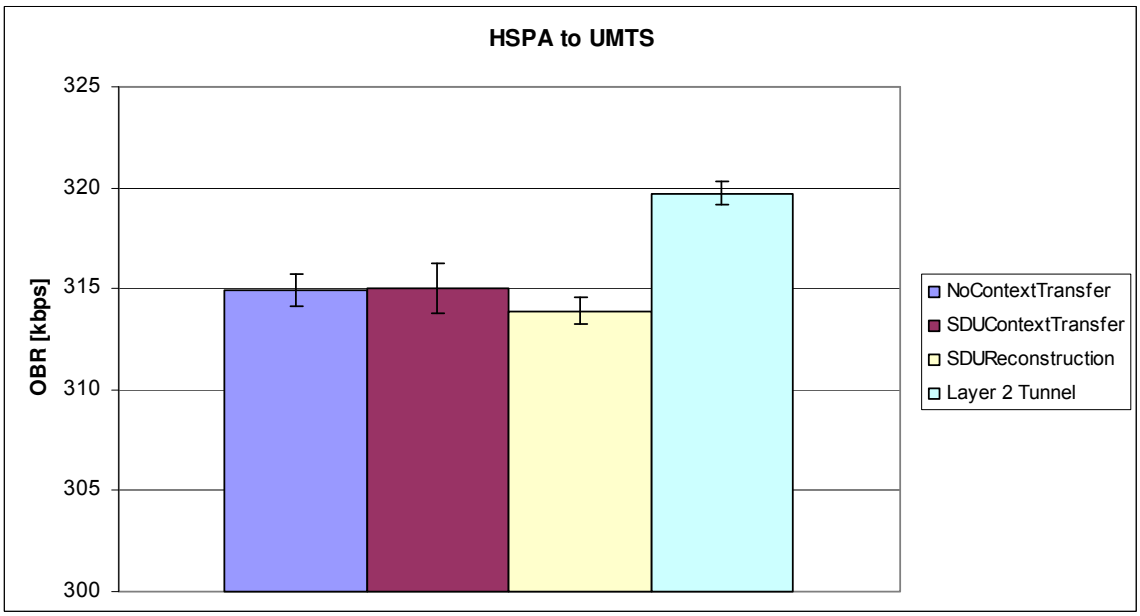


Figure 6.80 : Performance with a single access handover from HSPA to UMTS.

The results show that for TCP-based applications the down-switching at access handover, from a RAT with better performance to a RAT with lower performance, is not significantly influenced by the access handover scheme. This can be explained with the functionality of the access handover optimisation schemes. A handover without context transfer, i.e. without any optimisation, can lead to interactions with TCP congestion control, which decreases the TCP transmission rate. Access handover optimisation schemes can reduce this effect. However, when an access handover leads to a lower performance of the bottleneck link, TCP congestion control must adapt to this change in any case by reducing its transmission rate. Therefore, a handover optimisation scheme also leads to reduction of the TCP transmission rate and thus cannot provide any substantial gain.

6.6.5.2.2 Access Handover to a Faster Link (Up-switching)

In the following we investigate up-switching, which is an access handover to a new RAT with better performance. Figure 6.81 shows an access handover from UMTS to HSPA and Figure 6.82 shows an access handover from UMTS to FRAT. *SDU context transfer* improves the access handover over *no context transfer*; for UMTS to HSPA the gain is 8% and for UMTS to FRAT it is 23%. *SDU reconstruction* and *layer 2 tunnelling* outperform *SDU context transfer*; for UMTS to HSPA the gain is 28% and for UMTS to FRAT it is 62%. The difference between *layer 2 tunnelling* and *SDU reconstruction* is marginal.

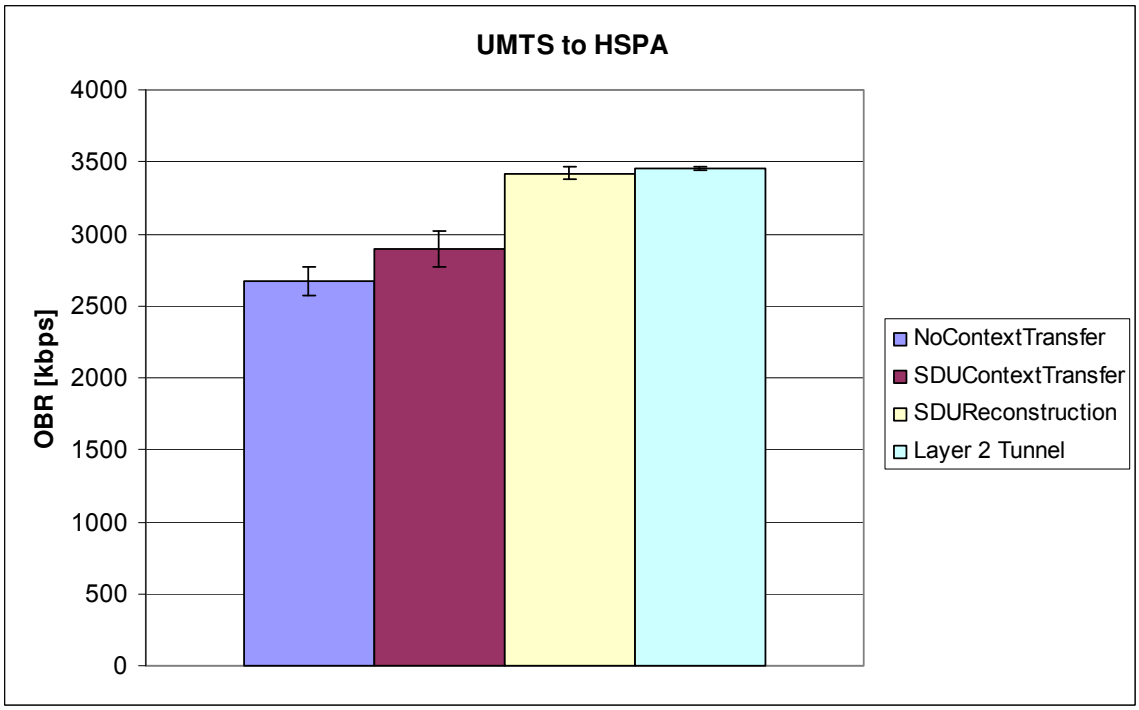


Figure 6.81 : Performance of a single access handover from UMTS to HSPA.

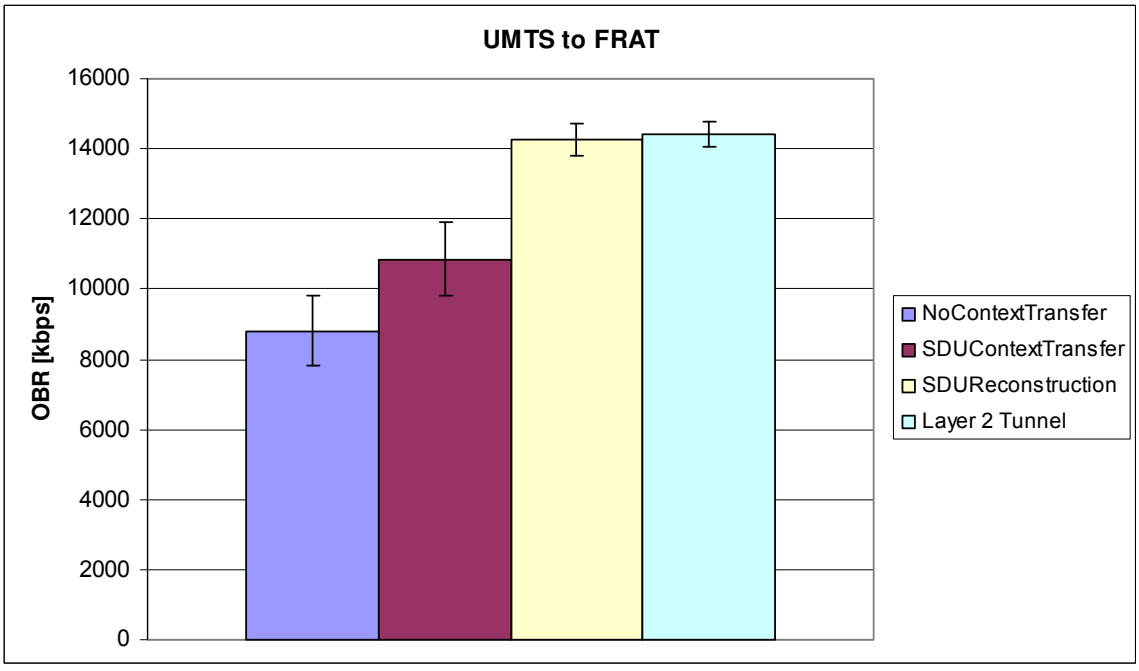


Figure 6.82 : Performance with a single access handover from UMTS to FRAT.

Figure 6.83 shows the object bit rate for an access handover from HSPA to FRAT. *SDU context transfer*, *SDU reconstruction* and *layer 2 tunnelling* achieve a performance gain of 28% over *no context transfer*. However, there is no significant difference between these access handover optimisation schemes. This can be explained with the link layer design of HSPA. The block error rate that is perceived by the RLC protocol in HSPA is two orders of

magnitude lower than for UMTS due to the usage of HARQ in the MAC layer. Therefore the size of the ARQ window remains low and the main contribution to packet losses at access handover comes from the SDU buffer rather than from the PDU buffer. The benefit of *SDU reconstruction* and *layer 2 tunnelling* over *SDU context transfer*, to transfer the PDU buffer, brings no additional gain.

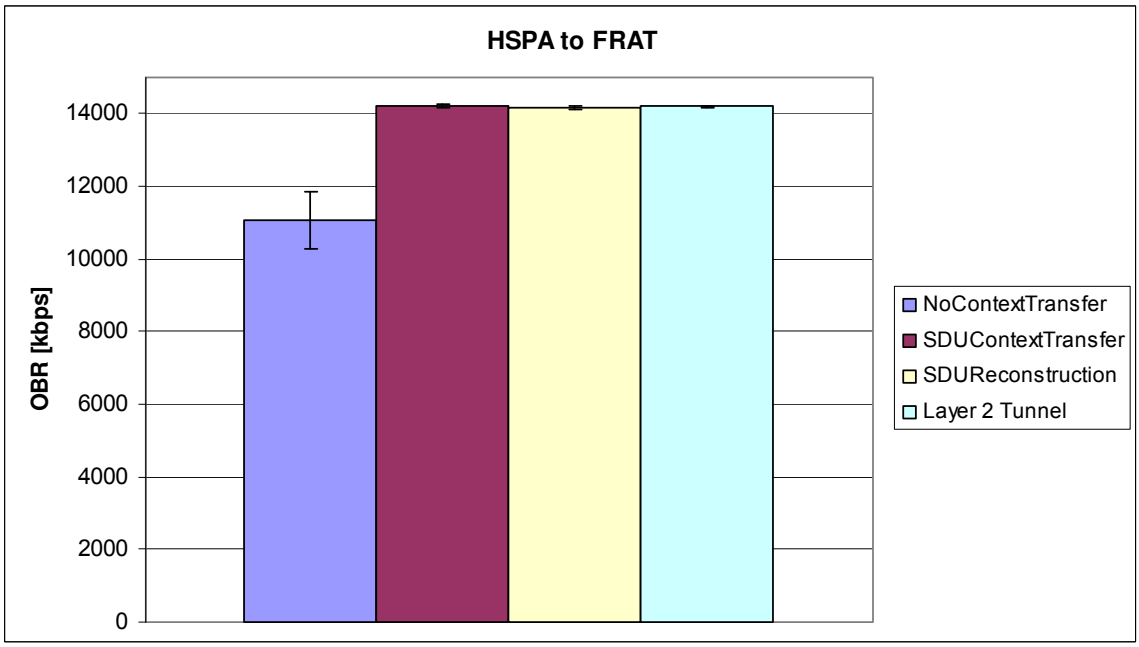


Figure 6.83 : Performance with a single access handover from HSPA to FRAT.

In contrast to the down-switching case, an access handover optimisation scheme can improve the performance for up-switching. At the access handover, TCP congestion control has to adapt to the change by increasing its transmission rate. A data loss resulting from access handover, however, decreases the TCP transmission rate. Therefore it takes longer time to adapt the TCP transmission rate to the new value. An access handover optimisation scheme can avoid the data loss, allowing TCP to adapt to the increased bottleneck data rate more quickly.

6.6.6 Conclusion

We have examined the performance of four different access handover schemes for TCP traffic based on TCP SACK. We have shown that the access handover can lead to interactions with the TCP congestion control, which can reduce the performance of file transfer. On one hand, the interactions that occur are packet loss for an access handover without context transfer or when *SDU context transfer* is used. These packet losses trigger TCP error recovery; in some cases the losses can be recovered by TCP congestion avoidance, in other cases by TCP timeout. The performance degradation caused by TCP timeout is more severe than the one caused by TCP congestion avoidance. The conservation of some buffered data by *SDU context transfer* can restart the TCP data flow more quickly, thus avoids in many cases TCP timeouts. As a result, access handover with *SDU context transfer* performs better than that of *no context transfer*. On the other hand, an access handover can lead to packet duplication when *SDU reconstruction* is used. Even if the access handover is lossless, the duplicated

packets can also trigger TCP congestion avoidance. The packet duplication is caused by an unsynchronised ARQ state in the link layer transmitter and receiver. It could be avoided if a link layer synchronisation procedure was used at the access handover prior to *SDU reconstruction*, as discussed in Section 6.5.3.3 (see Figure 6.38). The performance of access handover with *SDU reconstruction* is in general better than *SDU context transfer*. We have furthermore investigated an access handover scheme based on *layer 2 tunnelling*, which is lossless, without packet duplication and minimises the overhead of access handover. *Layer 2 tunnelling* does not lead to any TCP interactions and shows by far the best performance of the investigated access handover schemes. At the same time it is the most complex scheme, as it requires context transfer at both receiver and transmitter, as well as link layer support for access handover.

The access handover schemes have been investigated in different handover scenarios. In the homogenous handover scenario access handover was performed between two radio links with the same radio link characteristics. In this case, a clear difference in performance between the different access handover schemes can be noticed. In particular, this discrepancy becomes significant if access handovers are performed very frequently. We have investigated the influence of different radio link parameters on the access handover performance. The difference in performance for the different access handover schemes increases for increasing radio link block error rate, for increasing link layer round-trip time and for decreasing PDU sizes. The reason is that these parameters influence the ARQ window size of the link layer error recovery schemes. This inflates the pipe capacity of the link layer and thus also the sizes of the link layer buffers. This intensifies the amount of packet loss or packet duplication at access handover. For block error rates of 1% or lower and round-trip times of not more than 10 ms only access handover without context transfer still shows performance degradation. The difference in performance between the other access handover schemes diminishes. *Layer 2 tunnelling* is hardly effected by changes in link layer parameters and always performs without significant performance degradation.

In a heterogeneous handover scenario, the access handover between radio link with different link properties has been investigated. When the data rates of the different radio links differs by at least one order of magnitude, the influence of the access handover scheme is not as significant as in the homogeneous handover scenario. The reason is, that such link changes anyway require TCP to adapt its transmission rate to the new link characteristics. In this process, interactions of the access handover schemes with TCP congestion control are less prominent. In particular for down-switching from a radio link with high data rate to a radio link with low data rate, little gain can be achieved with sophisticated access handover schemes. Packet losses and duplication both lead to a reduction of the TCP throughput, which is required anyway when handing over to an access technology with lower data rate. For up-switching from a radio link with low data rate to a radio link with high data rate, sophisticated access handover schemes can significantly improve the performance. In this case an interaction with TCP congestion control that would reduce the TCP transmission rate are counter-productive to the required increase in TCP transmission rate.

6.7 Summary

A multi-access system provides a multitude of access technologies to user networks. We have shown in Section 4.5 that dynamic selection of the best suited access system for the user networks has the potential to increase the system capacity, provide flexibility in the

deployment of different access technologies and increase the achievable performance for data sessions of the end users. On the other hand, dynamic access selection leads to an increasing rate of access handovers between the different access systems. In this chapter we have investigated different access handover schemes. A key issue of access handover is to manage the communication context of ongoing data sessions. The objective is to make the access handover efficient and cause little performance degradation when changing between access systems.

We have investigated different access handover schemes. We have presented a *multi-radio generic link layer* based on a generic link layer toolbox to provide a common management of the communication context for the data transmission over different access systems. The multi-radio generic link layer can also enable the simultaneous transmission of a service data flow via multiple access systems. Our performance evaluation indicates that a multi-radio generic link layer can achieve very good access handover performance. However, there are also several drawbacks of the multi-radio generic link layer. Firstly, it requires a certain degree of harmonisation of link layer functionality for different access systems. This makes it difficult to evolve legacy access technologies into a multi-access system based on a multi-radio generic link layer. Also from a standardisation perspective, it would require that standardisation efforts in different standardisation fora are largely harmonised, which is difficult to achieve in reality. Furthermore, a multi-radio generic link layer assumes a tight integration of different access technologies into a multi-access architecture. This makes a realisation difficult, where different access systems are operated by different business entities. As an alternative to the multi-radio generic link layer, we have presented *generic link layer interworking*. For this it is required that different access technologies provide interworking functionality for managing the communication context at access handover. Access handover based on generic link layer interworking can lead to data distortion, like packet loss, packet duplication, packet re-ordering and interruptions. We have presented different access handover optimisation schemes how to reduce the amount of data distortion. Our analysis for TCP applications has shown that data distortion can severely reduce the service performance. Particularly, this is the case when access handovers occur frequently. We have also shown that the performance degradation depends on the link layer characteristics of the old and new radio link. Elevated link block error rates and link layer round-trip times increase the necessity for optimised access handover schemes. We have also shown that optimisation schemes are in particular important, when an access handover occurs between radio links with similar link characteristics, or for access handovers from a radio link with low link performance to a radio link with high link performance (up-switching). For access handovers to radio link with lower link performance the performance degradation – and thus the need for optimisation schemes – is less significant.

Chapter 7. Conclusion and Outlook

7.1 Conclusion

The wireless communications market has experienced a remarkable growth for more than a decade. A plethora of wireless communication services is still developing, which has particularly been stimulated with the addition of packet-switched transmission in wireless systems. Packet-switched transmission is the basis for Internet services, which are increasingly being used on mobile devices. The permanent increase in traffic volume requires a continuous extension of the wireless network infrastructure. Technical development has led to a variety of diverse radio access technologies with different characteristics. Some are targeting local areas, others provide wide area coverage. Radio access technologies can differ in their capacity and spectral efficiency, their support for certain services, their complexity and costs, and in their support for mobility. As a result, the best suited radio access technology can depend on the types of services that are anticipated, on the radio environment and the traffic patterns. Therefore, technical solutions are required to integrate multiple radio access technologies in a common mobile network architecture. This integration of heterogeneous access technologies into a network is referred to as *network convergence*; it allows building simpler and cheaper networks by providing a re-use of common functionality for multiple access systems. A multi-access system extends the service coverage of one access technology to areas where other access technologies are available. For network operators multi-access networks may yield better cost efficiency, which can also lead to reduced prices for end users. An end user can furthermore exploit the capabilities of different accesses by dynamically selecting the access which is best suited for the ongoing services – to be *always best connected* with an *always best experience*.

The realisation of a multi-access system imposes several technical challenges that we have investigated in this work. We have developed a functional system architecture for multi-access systems and specified functional elements that are required to manage the usage of access technologies. These elements comprise a *multi-radio resource management* function that determines which access is to be used by different users and data services – which we denote as *access selection*. Another function is the *generic link layer*, which derives generic metrics of different accesses that are required for access selection. We denote this functionality as *access monitoring*. Furthermore the generic link layer provides functionality to enable an efficient and seamless *access handover* between different access technologies. A *multi-access anchor* acts as forwarding point towards different access system; access handover is performed by re-directing the data flow at the multi-access anchor to another access via mobility management procedures. We have investigated different multi-radio access system architecture alternatives that provide all functionality required for multi-access management (i.e. access monitoring, selection, and handover). A multi-access system architecture is required to be scalable, be evolvable to integrate new access technologies, and be capable to migrate to from existing network architectures. The system architecture additionaly needs to be flexible to support different business cases; it shall enable the composition of a multi-access network by cooperation between different network operators that offer different accesses. Our investigation results in three system architecture alternatives. An *integrated multi-radio radio access network* provides tight integration of different access technologies with good and scalable support for multi-access management functions.

However, it requires modifications of access network functions of already deployed access systems and does not easily support loose cooperation between multiple operators. An *integrated multi-access core network* with multiple independent radio access networks enables a looser integration of various access technologies. It requires little extra functionality of the individual access technologies and flexibly supports different business scenarios. On the other hand, it supports only limited performance of multi-access management. A good trade-off is a hybrid approach, where different access technologies can be integrated either tightly via the RAN or loosely via the core network. The mode of integration option can be adapted to the business scenario. Sophisticated multi-access management only needs to be supported by access technologies for which this is feasible and the required harmonisation in standardisation is viable.

Access selection is the key function in a multi-access system; its functionality can be distributed to multiple entities and depends on the business scenario and on which business role one or more network operators take within the system architecture. We have developed a utility-based system model; it allows understanding access selection from the perspective of different system components and the roles that they take. Several different access selection algorithms have been described. We have investigated the performance of access selection measured as capacity of a multi-access system and service performance perceived by the end user by system simulations. We have shown that the gain of access selection depends on the scenario, in particular the traffic load distribution in the system, on the characteristics of the radio access technologies and the layout of the radio networks. We have investigated two scenarios where traffic load is either uniformly distributed in the system area or where traffic load is mainly confined to hotspot areas. We have considered a combination of either two different wide-area wireless networks as access technologies or a combination of a wide-area and a local-area wireless network; these networks can be co-located or non-co-located. We have shown that access selection can bring significant increase in capacity. This is particularly the case when access selection is based on load levels in the different access systems. The gain is largest in cell layouts that match the traffic load distribution. WLAN systems can only provide little capacity gain when the traffic load is uniformly distributed over the system area. Conversely, if users are mainly centred around hotspot areas WLAN systems in such areas can largely increase the multi-access system capacity. Furthermore we have developed a model based on stochastic knapsacks that allows investigating the performance of multi-access wireless networks analytically. Our model overcomes limitations of earlier studies.

A prerequisite for access selection is to discover and monitor the access properties and access network capabilities of different accesses in a comparable manner. We denote this functionality as *access monitoring*. We have identified which types of information about different accesses are helpful in order to enable a decision about the most suitable access. We have defined generic access abstractions for the radio link performance and the resource situation. Generic performance abstractions describe the suitability of an access with respect to the requirements of a service data flow on data rate, transmission delay, and transmission reliability. Generic resource abstractions describe the availability and the usage costs for access resources. In addition, we have investigated how access network information can be obtained for a user network, for example in access advertisements. Some information can be acquired by listening to system information that is broadcasted in beacon messages by the access network; other information can only be obtained after the user network has established a connection and has attached to the access network. We have shown that the latter requires extensive signalling. As an improvement we have developed an *ambient network attachment protocol* that integrates the access discovery, connectivity setup, access advertisements, and

access attachment in a combined procedure. We have evaluated this ambient network attachment procedure in a scenario where a user network evaluates surrounding WLAN networks for access selection; we have shown that the overhead and the delay imposed by the WLAN network evaluation can be significantly reduced compared to a legacy network evaluation procedure.

The dynamic selection of accesses requires the capability to dynamically re-allocate data connections between access systems. This *access handover* requires that a new communication context is established in the new access system or that the communication context of the old access system is transferred to and adapted for the new access system. We have developed and evaluated different link layer functions to facilitate access handovers in order to avoid data distortion for ongoing data sessions. Data distortion can result in reduced service performance; it comprises the loss or duplication of data, re-ordering of the sequence in which data is transmitted, or the interruption caused by an access handover. Access handover has been evaluated for data services that use the transmission control protocol (TCP) as a transport protocol in a number of different scenarios. One solution that we have developed is a *multi-radio generic link layer* that is based on a generic link layer toolbox to provide a common management of the communication context for the data transmission over different access systems. The multi-radio generic link layer can also enable the simultaneous transmission of a service data flow via multiple access systems. Our performance evaluation indicates that a multi-radio generic link layer can achieve very good access handover performance. However, the multi-radio generic link layer requires harmonisation of different access technologies, which may result in substantial changes to existing access technologies. It also leads to certain limitations for scenarios where a multi-access system is based on multiple access networks that are provided by different operators. We have developed an alternative solution with *generic link layer interworking*. For this it is required that different access technologies provide interworking functionality for managing the communication context at access handover. Access handover based on generic link layer interworking can lead to data distortion, like packet loss, packet duplication, packet re-ordering, and interruptions. In order to reduce the amount of data distortion we have developed different access handover optimisation schemes. Our analysis for TCP applications has shown that data distortion can severely reduce the service performance. In particular this is the case when access handovers occur frequently. We have also shown that the performance degradation depends on the link layer characteristics of the old and new radio link. Elevated link block error rates and link layer round-trip times increase the necessity for optimised access handover schemes. In addition, we have identified optimisation schemes to be particularly important when an access handover occurs between radio links with similar link characteristics, or for access handovers from a radio link with low link performance to a radio link with high link performance (up-switching). For access handovers to a radio link with lower link performance the performance degradation – and thus the need for optimisation schemes – is less significant.

7.2 Future Work

We have developed a multi-access system architecture and functionality for access monitoring, access selection, and access handover; it allows to integrate and manage different access technologies in an efficient manner, and it is flexible enough to support a variety of integration scenarios. More work is required to investigate to what extent this architecture and functionality can be included in the mobile network system architectures and radio access

technologies that are already developed and deployed. It is particularly attractive that some of our concepts can be adopted to the evolved 3GPP system architecture that targets to be a multi-access system. We have recently contributed to the definition of the *access network discovery and selection function* in the 3GPP evolved system architecture [SO08] [3GPP23.402]. While this function adopts some of our features of access monitoring, the integration of our access selection functionality may only be feasible for access technologies developed by 3GPP. For other access technologies these solutions would require more changes in the system architecture and would also lead to impractical dependencies between different standardisation fora. Therefore, we have developed a different solution of distributed policy-based access selection, which is simple and scalable. An evaluation of this policy-based access selection scheme and a comparison with the more complex solutions developed in this work is desirable.

For the evaluation of access handover we have only investigated scenarios where data services use TCP as transport protocol. Although this covers the largest part of services found in the Internet today, it is desirable to also study new services that are based on other transport protocols, for example, like mobile TV or multimedia telephony. Furthermore, a significant amount of research investigates new forms of reliable transport protocol that may eventually replace the currently dominant TCP protocol. It will be interesting to see, to what extent our conclusions for TCP also remain valid for those other protocols.

Bibliography

Research Articles and Books

[ABHMM+03] G. Alsenmyr, J. Bergström, M. Hagberg, A. Milén, W. Müller, H. Palm, H. van der Velde, P. Wallentin, F. Wallgren, "Handover between WCDMA and GSM," *Ericsson Review*, no. 1, 2003.

[AETHP06] J. Arkko, P. Eronen, H. Tschofenig, S. Heikkinen, A. Prasad, "Quick NAP - Secure and Efficient Network Access Protocol," in Proc. 6th International Workshop on Applications and Services in Wireless Networks, May 29-31, 2006.

[AGP04] S. Aust, C. Görg, C. Pampu, "Proactive Handover Decision for Mobile IP based on Link Layer Information," in Proc. 1st IFIP International Conference on Wireless and Optical Communications Networks (WOCN 2004), Muscat, Oman, June 2004.

[AHP03] K. Ahmavaara, H. Haverinen, R. Pichna, "Interworking Architecture between WLAN and 3G Systems," *IEEE Communications Magazine*, vol. 41, no. 11, November 2003.

[AHPST+01] M. Annoni, R. Hancock, T. Paila, E. Scarrone, R. Tönjes, D. Wisely, R. Mort, "Radio access networks beyond 3rd generation: A first comparison of architectures," in Proc. IST Mobile Communications Summit, Barcelona, October 1-4, 2001.

[APFPG03] S. Aust, D. Proetel, N. Fikouras, C. Pampu, C. Görg, "Policy Based Mobile IP Handoff Decision (POLIMAND) using Generic Link Layer Information," in Proc. 5th IEEE International Conference on Mobile and Wireless Communication Networks (MWCN 2003), Singapore, October 2003.

[BA04] H. Badis, K. Al Agha, "Fast and efficient vertical handoffs in wireless overlay networks," in Proc. 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), September 5-8, 2004.

[BCHLM+03] M. M. Buddhikot, G. Chandranmenon, S. Han, Y.-W. Lee, S. Miller, L. Salgarelli, "Design and Implementation of a WLAN/CDMA2000 Interworking Architecture," *IEEE Communications Magazine*, vol. 41, no. 11, November 2003.

[BCJLM+04] P. Beming, M. Cramby, N. Johansson, J. Lundsjö, G. Malmgren, J. Sachs, "Beyond 3G Radio Access Network Reference Architecture," in Proc. 59th IEEE Vehicular Technology Conference, Milan, Italy, May 17-19, 2004.

[BGK02] T. Boström, T. Goldbeck-Löwe, R. Keller, "Ericsson Mobile Operator WLAN solution," *Ericsson Review*, no. 1, 2002.

[BH07a] M. Blomgren, J. Johan Hultell, "Decentralized Market-Based Radio Resource Management in Multi-Network Environments," in Proc. IEEE Vehicular Technology Conference, Dublin, May 2007.

[BH07b] M. Blomgren, J. Hultell, "Demand Responsive Pricing in Open Wireless Access Markets," in Proceedings of IEEE Vehicular Technology Conference, Dublin, May 2007.

[BHNVO05] M. Bäckström, A. Havdrup, T. Nylander, J. Vikberg, P. Öhman, "Mobile@Home—GSM services over wireless LAN," *Ericsson Review*, no. 2, 2004.

[Bia00] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, March 2000.

[BJL05] A. Baraev, L. Jorguseski, R. Litjens, "Performance Evaluation of Radio Access Selection Procedures in Multi-Radio Access Systems," in Proc. 8th International Symposium on Wireless Personal Multimedia Communications (WPMC), Aalborg, Denmark, September 17-22, 2005.

[BL06] F. Berggren, R. Litjens, "Performance Analysis of Access Selection and Transmit Diversity in Multi-Access Networks," in Proc. 12th Annual International Conference on Mobile Computing and Networking (MobiCom), Los Angeles, USA, September 24-29, 2006.

[BLMR98] J. Byers, M. Luby, M. Mitzenmacher, A. Rege, "A digital fountain approach to reliable distribution of bulk data," in Proc. ACM SIGCOMM, Vancouver, BC, Canada, January 1998.

[BSW03] L. Berlemann, M. Siebert, B. Walke, "Software Defined Protocols Based on Generic Protocol Functions for Wired and Wireless Network," in Proc. Software Defined Radio Technical Conference, November 2003.

[BSW95] H. Balakrishnan, S. Seshan, R. H. Katz, "Improving Reliable Transport and Handoff Performance in Cellular Wireless Networks," *Wireless Networks* (WINET), vol. 1, no. 4, December 1995.

[BVE99] C. Bettstetter, H-J. Vögel, J. Eberspächer, "GSM Phase 2+ General Packet Radio Service: Architecture, Protocols and Air Interface," *IEEE Communications Surveys*, vol. 2, no. 3, 1999.

[BW97] G. Brasche, B. Walke, "Concepts, services, and protocols of the new GSM phase 2+ general packet radio service," *IEEE Communications Magazine*, vol. 35, no. 8., 1997.

[CG04] G. Camarillo, M.A. Garcia-Martin, *The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular Worlds*, John Wiley & Sons, 2004.

[CI95] R. Caceres, L. Iftode, "Improving the Performance of Reliable Transport Protocols in Mobile Computing Environments," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 5, June 1995.

[CKPHM+05] C. Cedervall, P. Karlsson, M. Prytz, J. Hultell, J. Markendahl, A. Bria, O. Rietkerk, I. Karla, "Initial Findings on Business Roles, Relations and Cost Savings Enabled by Multi-Radio Access Architecture in Ambient Networks," in Proc. 14th Wireless World Research Forum, San Diego, California, USA, July 2005.

[CR02] M.C. Chan, R. Ramjee, "TCP/IP Performance over 3G Wireless Links with Rate and Delay Variations," in Proc. International Conference on Mobile Computing and Networking (MobiCom), Atlanta, Georgia, USA, September 23-28, 2002.

[CWSB02] D. D. Clark, J. Wroclawski, K. R. Sollins, R. Braden, "Tussle in Cyberspace: Defining Tomorrow's Internet," in Proc. ACM SIGCOMM, Pittsburgh, Pennsylvania, USA, August 19-23, 2002.

[CWSB05] D. D. Clark, J. Wroclawski, K. R. Sollins, R. Braden, "Tussle in Cyberspace: Defining Tomorrow's Interne," *IEEE/ACM Transactions on Networking*, vol. 13, no. 3, June 2005.

[DABKK+05] K. Dimou, R. Agüero, M. Bortnik, R. Karimi, G. P. Koudouridis, S. Kaminski, H. Lederer, J. Sachs, "Generic Link Layer: A Solution For Multi-radio Transmission Diversity For Communication Networks beyond 3G," in Proc. 62nd IEEE Conference In Vehicular Technology, Dallas, USA, September 2005.

[DPSB07] E. Dahlman, S. Parkvall, J. Sköld, P. Beming, *3G Evolution – HSPA and LTE for Mobile Broadband*, Academic Press, 1st edition, 2007.

[DZ01] D. Dutta, Y. Zhang, "An Active Proxy based Architecture for TCP in Heterogeneous Variable Bandwidth Networks," in Proc. Global Communications Conference (Globecom), November 2001.

[EFKMP+06] H. Ekström, A. Furuskär, J. Karlsson, M. Meyer, S. Parkvall, J. Torsner, M. Wahlqvist, "Technical Solutions for the 3G Long-Term Evolution," *IEEE Communications Magazine*, vol. 44, no. 3, March 2006.

[EL03] H. Ekström, R. Ludwig, "Queue Management for TFRC-based Traffic in 3G Networks," in Proc. Workshop on Mobile and Wireless Networks, Providence, Rhode Island, USA, May 19-22, 2003.

[ES03] H. Ekström, A. Schieder, "Buffer Management for the Interactive Bearer in GERAN," in Proc. 57th IEEE Vehicular Technology Conference (VTC), April 22-25, 2003.

[EWKKW06] M. Emmelmann, S.Wiethoelter, A.Koepsel, C.Kappler, and A.Wolisz, "Moving towards Seamless Mobility -- State of the Art and Emerging Aspects in Standardization Bodies," *International Journal on Wireless Personal Communication*, vol. 43, no. 3 , November 2007, DOI 10.1007/s11277-007-9344-6.

[FAJ05] A. Furuskär, M. Almgren, K. Johansson, "An Infrastructure Cost Evaluation of Single- and Multi-Access Networks with Heterogeneous Traffic Density," in Proc. IEEE Vehicular Technology Conference (VTC), Stockholm, Sweden, May 30 - June 1, 2005.

[FFL04] G. Fodor, A. Furuskär, J. Lundsjö, “On Access Selection Techniques in Always Best Connected Networks,” in Proc. 16th ITC Specialist Seminar, Antwerp, Belgium, August 31 - September 02, 2004.

[FG00a] N. Fikouras, C. Görg, “Performance Comparison of Hinted and Advertisement Based Movement Detection Methods for Mobile IP and Hand-Offs,” in Proc. European Wireless, Dresden, Germany, September 12 - 14, 2000.

[FG00b] N. A. Fikouras, C. Görg, “Performance Analysis and Improvement of TCP during Mobile IP Hand-Offs,” in Proc. First Polish German Teletraffic Symposium (PGTS), Dresden, Germany, September 24-26, 2000.

[FHTW04] R-M. Furtenback, T. Hunte, D. Turina, U. Wahlberg, “GSM and WCDMA—Common network approach,” *Ericsson Review*, no. 2, 2004.

[FKG01] N. Fikouras, A. Könsgen, C. Görg, “Accelerating Mobile IP Hand-Offs through Link-layer Information,” in Proc. 11th GI/ITG Conference on Measuring, Modelling and Evaluation of Computer and Communication Systems, Aachen, Germany, September 2001.

[FKGBT04] N. Fikouras, K. Kuladinithi, C. Görg, C. Bormann, A. Timm-Giel, “Multiple Access Interfaces Management and Flow Mobility,” in Proc. 13th IST Mobile and Wireless Communications Summit, Lyon, France, June 2004.

[FKW03] A. Festag, H.Karl, A. Wolisz, "Classification and Evaluation of Multicast-Based Mobility Support in All-IP Cellular Networks," in Proc. Kommunikation in Verteilten Systemen (KiVS), Leipzig, Germany, February 2003.

[FKW07] A. Festag, H.Karl, and A. Wolisz, "Investigation of Multicast-Based Mobility Support in All-IP Cellular Networks," *Journal on Wireless Communications and Mobile Computing*, vol. 7, no. 3, November 2007.

[FMMO99] A. Furuskär, S. Mazur, F. Müller, H. Olofsson, “EDGE: enhanced data rates for GSM and TDMA/136 evolution,” *IEEE Personal Communications*, vol. 6, no. 3, June 1999.

[FNO99] A. Furuskär, J. Näslund, H. Olofsson, “EDGE – Enhanced data rates for GSM and TDMA/136 evolution,” *Ericsson Review*, no.1, 1999.

[FUGZ03] N.A. Fikuoras, A. Udugama, C. Görg and W. Zirwas, “Experimental Evaluation of Load Balancing for Mobile Internet Real-Time Communications,” in Proc. Sixth International Symposium on Wireless Personal Multimedia Communications (WPMC), Yokosuka, Kanagawa, Japan, October 2003.

[Fur02] A. Furuskär, “Allocation of multiple services in multi-access wireless systems,” in Proc. of IEEE Mobile and Wireless Communications Networks (MWCN), Stockholm, Sweden, November 2002.

[Fur03] A. Furuskär, “Radio resource sharing and bearer service allocation for multi-bearer service multi-access wireless networks,” Ph.D. dissertation, Radio Communications Systems Laboratory, Royal Institute of Technology, Stockholm, Sweden, May 2003.

[FW00] A. Festag, A. Wolisz, "MOMBASA: Mobility Support - A Multicast-based Approach," in Proc. European Wireless, Dresden, Germany, September 2000.

[FZ05] A. Furuskär, J. Zander, “Multiservice allocation for multiaccess wireless systems,” *Transactions on Wireless Communications*, vol. 4, no. 1, 2005.

[GABBE+07] A. Gunnar, B. Ahlgren, O. Blume, L. Burness, P. Eardley, E. Hepworth, A. Surtees, J. Sachs, "Access and Path Selection in Ambient Networks," in Proc. 16th IST Mobile & Wireless Communications Summit, Budapest, Hungary, July 1-5, 2007.

[GJ03] E. Gustafsson, A. Jonsson, “Always best connected,” *IEEE Wireless Communications*, vol. 10 , no. 1, February 2003.

[GKTCW08] A. Greenspan, M. Klerer, J. Tomcik, R. Canchi, J. Wilson, “IEEE 802.20: Mobile Broadband Wireless Access for the Twenty-First Century,“ *IEEE Communications Magazine*, vol. 46, no. 7, July 2008.

[GL03] A. Gurtov, R. Ludwig, “Responding to Spurious Timeouts in TCP,” in Proc. IEEE Conference on Computer Communications (Infocom), San Francisco, March 30 - April 3, 2003.

[Gur01] A. Gurtov, “Effect of Delays on TCP Performance,” in Proc. IFIP Personal Wireless Communications (PWC), Lappeenranta, Finland, August 8-10, 2001.

[Hal92] F. Halsall, *Data Communications, Computer Networks and Open Systems*, Addison-Wesley, Wokingham, UK, 3rd edition, 1992.

[HF03] W. Hansmann, M. Frank, "On Things to Happen During a TCP Handover," in Proc. 28th Annual IEEE International Conference on Local Computer Networks, Bonn/Königswinter, Germany, October 20-24, 2003.

[HFW02] W. Hansmann, M. Frank, M. Wolf, "Performance Analysis of TCP handover in a Wireless/Mobile Multi-Radio Environment," in Proc. IEEE Conference on Local Computer Networks (LCN), Tampa, FL. USA, November 2002.

[HGS03] R. Hsieh, Z.-G. Zhou, A. Seneviratne, “S-MIP: A Seamless Handoff Architecture for Mobile IP,” in Proc. IEEE Conference on Computer Communications (Infocom), San Francisco, March 30 - April 3, 2003.

[HJJ02] R. Heickerö, S. Jelvin, B. Josefsson, “Ericsson seamless network”, *Ericsson Review*, no. 2, 2002.

[HJM04] J. Hultell, K. Johansson, J. Markendahl, “Business models and resource management for shared wireless networks,” in Proc. IEEE Vehicular Technology Conference, Los Angeles, USA, September 26-29, 2004.

[HRBD03] M. Heusse, F. Rousseau, G. Berger-Sabbatel, A. Duda, “Performance Anomaly of 802.11b,” in Proc. Annual Joint Conference of the IEEE Computer and Communications Societies (Infocom), March 30 – April 3 2003.

[HS03] R. Hsieh, A. Seneviratne, “A Comparison of Mechanisms for Improving Mobile IP Handoff Latency for End-to-End TCP,” in Proc. Annual International Conference on Mobile Computing and Networking (Mobicom), San Diego, USA, September 14-19, 2003.

[HSK05] I. Herwono, J. Sachs, R. Keller, "Performance Improvement of Media Point Network using the Inter Access Point Protocol according to IEEE 802.11f," in Proc. 11th European Wireless 2005, Nicosia, Cyprus, April 10-13, 2005

[HT00] H. Holma, A. Toskala, *WCDMA for UMTS*, Wiley, Chichester, 2000.

[HT06] H. Holma, A. Toskala, *HSDPA/HSUPA for UMTS: High Speed Radio Access for Mobile Communications*. Wiley, Chichester, 2006.

[Hul07] J. Hultell, “Access Selection in Multi-System Architectures,” Licentiate Thesis, Royal Institute of Technology (KTH), Stockholm Sweden, ISSN 1653–6347, 2007.

[IHITU+04] H. Ishii, A. Hanaki, Y. Imamura, S. Tanaka, M Usuda, T. Nakamura, "Effects of UE capabilities on high speed downlink packet access in WCDMA systems," in Proc. 59th IEEE Vehicular Technology Conference (VTC), May 17-19, 2004.

[Jai91] R. Jain, *The Art of Computer Systems Performance Analysis*, Wiley, 1991.

[JF05] K. Johansson, A. Furuskär, “Cost Efficient Capacity Expansion Strategies Using Multi-Access Networks,” in Proc. IEEE Vehicular Technology Conference (VTC), 2005.

[JFKZ04] K. Johansson, A. Furuskär, P. Karlsson, J. Zander, “Relation Between Base Station Characteristics and Cost Structure in Cellular Systems,” in Proc. IEEE Personal, Indoor, and Mobile Communications (PIMRC), 2004.

[Joh05] K. Johansson, “Cost efficient provisioning of wireless access: infrastructure cost modelling and multi-operator resource sharing”, Licentiate Thesis, Royal Institute of Technology (KTH), Stockholm, Sweden, ISSN 1400-9137, 2005.

[JSRJ06] M. Johnsson, J. Sachs, T. Rinta-aho, T. Jokikyyny, “Ambient Networks – A Framework for Multi-Access Control in Heterogeneous Networks,” in Proc. IEEE Vehicular Technology Conference (VTC), Montreal, Canada, September 25 – 28, 2006.

[KAABB+05] G. P. Koudouridis, R. Agüero, E. Alexandri, M. Berg, A. Bria, J. Gebert, L. Jorguseski, H. R. Karimi, I. Karla, P. Karlsson, J. Lundsjö, P. Magnusson, F. Meago, M. Prytz, J. Sachs, "Feasibility Studies and Architecture for Multi-Radio Access in Ambient Networks," in Proc. 15th Wireless World Research Forum Meeting, Paris, France, December 8-9, 2005.

[KAACD+05] G.P. Koudouridis, R. Agüero, E. Alexandri, J. Choque, K. Dimou, H.R. Karimi, H. Lederer, J. Sachs, R. Sigle, "Generic Link Layer Functionality for Multi-Radio Access Networks," in Proc. 14th IST Mobile & Wireless Communications Summit, Dresden, Germany, June 19-23, 2005.

[KADGP+07] G.P. Koudouridis, R. Agüero, K. Daoud, J. Gebert, M. Prytz. T. Rinta-aho, J. Sachs, H. Tang, "Access Flow Based Multi-Radio Access Connectivity," in Proc. 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Athens, Greece, September 3-6, 2007.

[KDKK06] H. R. Karimi, K. Dimou, G. P. Koudouridis, P. Karlsson, "Switched Multi-Radio Transmission Diversity for Non-Collocated Radio Accesses," in Proc. IEEE of Vehicular Technology Conference, Melbourne, Australia, May 7 – 10, 2006.

[KFKTG03] K. Kuladinithi, N. A. Fikouras, A. Könsgen, A. Timm-Giel, C. Görg, "Enhanced Terminal Mobility through the use of Filters for Mobile IP," in Proc. IST Mobile and Wireless Communications Summit (IST Summit), Aveiro, Portugal, June 2003.

[KFZK04] I. Koo, A. Furuskär, J. Zander, K. Kim, "Erlang Capacity of Multiaccess Systems With Service-Based Access Selection," *IEEE Communication Letters*, vol.8, no. 11, November 2004.

[KH03] G.M. Koien, T. Haslestad, "Security aspects of 3G-WLAN interworking," *IEEE Communications Magazine*, vol. 41, no. 11, November 2003.

[KKD05] G. P. Koudouridis, H. R. Karimi, K. Dimou, "Switched Multi-Radio Transmission Diversity in Future Access Networks," in Proc. IEEE Vehicular Technology Conference, Dallas, USA, Sep. 25-28, 2005.

[KKLBB+05] G. P. Koudouridis, P. Karlsson, J. Lundsjö, A. Bria, M. Berg, L. Jorguseski, F. Meago, R. Agüero, J. Sachs, R. Karimi, "Multi-Radio Access in Ambient Networks," presented at IST Everest Workshop on Trends in Radio Resource Management, Barcelona, Spain, 16 November 2005.

[KP01] R. Koodli, C. E. Perkins, "Fast Handovers and Context Transfers," *ACM Computer Communication Review*; vol 31; no. 5; October 2001.

[KPJS07] C. Kappler, P. Poyhonen, M. Johnsson, S. Schmid, "Dynamic network composition for beyond 3G networks: a 3GPP viewpoint," *IEEE Network*, vol. 21; no. 1, 2007.

[KVB06] R. Keller, M. Vorwerk, C. Barschel, "Voice Call Continuity – A novel mobility scheme for voice on call control level," in Proc. 64th IEEE Vehicular Technology Conference, Montreal, Canada. September 25 – 28, 2006.

[LAABC+05] J. Lundsjö, R. Agüero, E. Alexandri, F. Berggren, C. Cedervall, K. Dimou, J. Gebert, R. Jennen, L. Jorguseski, H.R. Karimi, F. Meago, H. Tang, R. Veronesi, "A Multi-Radio Access Architecture for Ambient Networking," in Proc. 14th IST Mobile and Wireless Communications Summit, Dresden, Germany, June 19-23, 2005.

[Lee90] W. C. Y. Lee, "Estimate of Channel Capacity in Rayleigh Fading Environment," *IEEE Transactions. On Vehicular Technology*, vol. 39, no. 3, Aug. 1990.

[LEWL06] R. Ludwig, H. Ekström, P. Willars, N. Lundin, “An Evolved 3GPP QoS Concept,” in Proc. 63rd Vehicular Technology Conference Spring, Melbourne, Australia, May 7-10, 2006.

[LK00] R. Ludwig, R. H. Katz, “The Eifel algorithm: Making TCP robust against spurious retransmissions,” *ACM Computer Communication Review*, vol. 30, no. 1, January 2000.

[Lub02] M. Luby, “LT Codes,” in Proc. 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002.

[Mac05] D.J.C. MacKay, “Fountain Codes,” *IEE Proceedings Communications*, vol. 152, no. 6, December 2005.

[MBKLM+05] P. Magnusson, F. Berggren, I. Karla, R. Litjens, F. Meago, H. Tang, R. Veronesi, “Multi-Radio Resource Management for Communication Networks beyond 3G,” in Proc. IEEE Vehicular Technology Conference (VTC), Dallas, Texas, September 25-28, 2005.

[MGSCA07] F. Meago, J. Gebert, J. Sachs, J. Choque, R. Aguero, "On capacity/load-based and availability-based resource abstractions for multi-access networks,” in Proc. Workshop on Mobility, Multiaccess, and Network Management (M2NM), Sydney, Australia, October 16-19, 2007.

[MJ06] J. Markendahl, M. Johnsson, “Ambient networking and related business concepts as support for regulatory initiatives and competition,” in Proc. 5th Conference on Telecommunication and Techno-Economics (CTTE), Athens, June 8-9, 2006.

[MLSW04] P. Magnusson, J. Lundsjö, J. Sachs, P. Wallentin, "Radio Resource Management Distribution in a Beyond 3G Multi-Radio Access Architecture," in Proc. IEEE Global Telecommunications Conference (Globecom), Dallas, USA, Nov. 29 - Dec. 3, 2004.

[MM03] P. Maymounkov, D. Mazières, “Rateless Codes and Big Downloads,” in Proc. 2nd International Workshop on Peer-to-peer Systems, Berkeley, CA, USA, February 20-21, 2003.

[MMK06] J. Markendahl, M. Johnsson, P. Karlsson, “Business opportunities and regulatory issues of Ambient Networks,” in Proc. 17th European Regional ITS Conference, Amsterdam, Netherlands, August 22-24, 2006.

[Moh02] W. Mohr, “Heterogeneous Networks to Support User Needs with Major Challenges for New Wideband Access Systems,” *Wireless Personal Communications*, vol. 22, no. 2, August 2002.

[Moh03] W. Mohr, “Spectrum demand for systems beyond IMT-2000 based on data rate estimates,” *Wireless Communications and Mobile Computing*, vol. 3, no. 7, November 2003.

[Moh05] W. Mohr, “The WINNER (Wireless World Initiative New Radio) Project - Development of a Radio Interface for Systems beyond 3G,” in Proc. IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications, Berlin, Germany, September 11-14. 2005.

[MPSL06] J. Markendahl, P. Poyhoenen, O. Strandberg, J. Laganier, “Business implications of AN composition framework,” in Proc. Helsinki Mobility Roundtable, Helsinki, Finland, June 1-2, 2006.

[MSA03] A. Mishra, M. Shin, W. Arbaugh, "An empirical analysis of the IEEE 802.11 MAC layer handoff Process," *ACM SIGCOMM Computer Communication Review*, vol. 33, no. 2, April 2003.

[MSH03] M. Meyer, J. Sachs, M. Holzke, “Performance Evaluation of a TCP proxy in WCDMA networks,“ *IEEE Wireless Communications*, vol. 10, no. 5, Oct 2003.

[MSM97] M. Mathis, J. Semke, J. Mahdavi, “The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm,” *ACM Computer Communication Review*, vol. 27, no. 3, July 1997.

[NSAMS+04] N. Niebert, A. Schieder, H. Abramowicz , G. Malmgren, J. Sachs, U. Horn, C. Prehofer, H. Karl, “Ambient Networks - An Architecture for Communication Networks Beyond 3G,” *IEEE Wireless Communications*, vol. 11, no. 2, April 2004.

[NSZH07] N. Niebert, A. Schieder, J. Zander, R. Hancock, *Ambient Networks: Co-Operative Mobile Networking for the Wireless World*, John Wiley & Sons, 1st edition, 2007.

[PB04] K. Pentikousis, H. Badr, “Quantifying the Deployment of TCP Options – A Comparative Study,” *IEEE Communication Letters*, vol. 8, no. 10, October 2004.

[PHDSP+06] A. Pollard, J. von Häfen, M. Dottling, D. Schultz, R. Pabst, E Zimmerman, “WINNER – Towards Ubiquitous Wireless Access,” in Proc. 63rd IEEE Vehicular Technology Conference (VTC), Melbourne, Australia, May 7-10, 2006.

[PKCBK06] M. Prytz, P. Karlsson, C. Cedervall, A. Bria, I. Karla, "Infrastructure Cost Benefits Of Ambient Networks Multi-Radio Access," in Proc. 63rd IEEE Conference In Vehicular Technology (VTC), Melbourne, Australia, May 7-10, 2006.

[PM01] J. Peisa, M. Meyer, “Analytical Model for TCP File Transfers over UMTS,” in Proc. International Conference on Third Generation Wireless and Beyond, San Francisco, USA, May 30 – June 2, 2001.

[PR04] R. Pries, K. Heck, “Performance Comparison of Handover Mechanisms in Wireless LAN Networks,” in Proc. Australian Telecommunication Networks and Applications Conference (ATNAC), Sydney, Australia, December 8 – 10, 2004.

[PW99] C. Perkins, K.-Y. Wang, “Optimized Smooth Handoffs in Mobile IP,” in Proc. IEEE Symposium of Computers and Communications, Red Sea, Egypt, July 1999.

[PWSTG+07] J. Peisa, S Wager, M. Sâgfors, J. Torsner, B. Göransson, T. Fulghum, C. Cozzo, S. Grant, “High-speed Packet-access Evolution – Concept and Technologies,” in Proc. IEEE Vehicular Technology Conference (VTC), , Dublin, Ireland, April 22 – 25, 2007.

[QRS07] O. Queseth, T. Rinta-aho, J. Sachs, "Ambient Networks Advertisements for Attachment," presented at IST B3G Cluster Workshop on Network Detection and Heterogeneous Radio Resource Management, Brussels, Belgium, March 20, 2007.

[RAQS07] T. Rinta-aho, N. Akthar, O. Queseth, J. Sachs, "Ambient Network Advertisements," in Proc. Workshop on Mobility, Multiaccess, and Network Management (M2NM), October 16-19, Sydney, Australia.

[RCMMS+07] T. Rinta-aho, R. Campos, U. Meyer, A. Méhes, J. Sachs, G. Selander, "Ambient Network Attachment," in Proc. 16th IST Mobile & Wireless Communications Summit, Budapest, Hungary, July 1-5, 2007.

[RHM06] O. Rietkerk, G. Huitema, J. Markendahl, “Business roles enabled by Ambient Networking to provide access for anyone to any network and service,” presented at Helsinki Mobility Roundtable, Helsinki, Finland, June 1-2, 2006.

[Ros95] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*, Springer, London 1995.

[SABGJ+06] J. Sachs, R. Agüero, M. Berg, J. Gebert, L. Jorguseski, I. Karla, P. Karlsson, G.P. Koudouridis, J. Lundsjö, M. Prytz, O. Strandberg, “Migration of Existing Access Networks Towards Multi-Radio Access,” in Proc. IEEE Vehicular Technology Conference (VTC) Fall, Montreal, Canada, September 25 – 28, 2006.

[Sac00] J. Sachs, "Mobile Internet - Performance Issues Beyond the Radio Interface," in Proc. Workshop Mobile Internet 2000, Beijing, China, October 20-21, 2000.

[Sac03a] J. Sachs, "A Generic Link Layer for Integrated Multi-Radio Networks," presented at International Workshop on Multiradio Multimedia Communications, Dortmund, Germany, February 26-27, 2003.

[Sac03b] J. Sachs, "A Generic Link Layer for Future Generation Wireless Networking," in Proc. IEEE International Conference on Communications (ICC), Anchorage, AK, USA, May 11-15, 2003.

[Sac06a] J. Sachs, "A Stochastic Knapsack Model for the Capacity Evaluation of (Multi-) Radio Access Networks," in Proc. 4th Polish-German Teletraffic Symposium (PGTS), Wroclaw, Poland, September 21-22, 2006.

[Sac06b] J. Sachs, "A Stochastic Knapsack Model for the Capacity Evaluation of (Multi-) Radio Access Networks," *Systems Science*, vol 32, no 3, 2006.

[SADGK+07] J. Sachs, R. Agüero, K. Daoud, J. Gebert, G.P. Koudouridis, F. Meago, M. Prytz, T. Rinta-aho, H. Tang, "Generic Abstraction of Access Performance and Resources for Multi-Radio Access Management," in Proc. 16th IST Mobile & Wireless Communications Summit, Budapest, Hungary, July 1-5, 2007.

[SADGK+08] J. Sachs, R. Agüero, K. Daoud, J. Gebert, G.P. Koudouridis, F. Meago, M. Prytz, T. Rinta-aho, H. Tang, "Generic Abstraction of Access Performance and Resources for Multi-Radio Access Management," book chapter in I. Frigyes, J. Bito, P. Bakker (editors), *Advances in Mobile and Wireless Communications: Views of the 16th IST Mobile and Wireless Communication Summit*, Springer, 1st edition, 2008.

[SAEG07] A. Surtees, R. Aguero, J. Eisl, M. Georgiades, "Mobility Management in Ambient Networks," in Proc. Vehicular Technology Conference (VTC) Spring, Dublin, Ireland, April 23 – 25, 2007.

[Sch06] M. Schopp, "Trends in Mobile Network Architectures: 3GPP LTE, Mobile WiMAX, Next Generation Mobile Networks," presented at ITG-Fachtagung Zukunft der Netze, Bremen, Germany, November 17, 2006.

[SFA03] R. Samarasinghe, V. Friderikos, A.H. Aghvami, "Analysis of Intersystem Handover: UMTS FDD & WLAN," in Proc. London Communications Symposium, September 8-9. 2003.

[SGGA05] V. Sarangan, D. Ghosh, N. Gautam, R. Acharya, "Steady State Distribution for Stochastic Knapsack with Bursty Arrivals," *IEEE Communication Letters*, vol. 9, no. 2, 2005.

[Sha48] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[Sho06] A. Shokrollahi, "Raptor codes," *IEEE Transactions on Information Theory*, vol. 52, no. 6, June 2006.

[SKLMR+06] J. Sachs, M. Kampmann, J. Lundsjö, F. Meago, T. Rinta-aho, B. Tharon, "Continuous Connectivity in Ambient Networks," in Proc. IST Mobile and Wireless Communications Summit, Myconos, Greece, June 4-8, 2006.

[SKM06] J. Sachs, B. S. Khurana and P. Mähönen, "Evaluation of Handover Performance for TCP Traffic based on Generic Link Layer Context Transfer," in Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Helsinki, Finland, September 11-14, 2006.

[SLMP03a] M. Sågfors, R. Ludwig, M. Meyer, J. Peisa, "Queue Management for TCP Traffic over 3G Links," in Proc. IEEE Wireless Communications and Networking Conference (WCNC), New Orleans, USA, March 16-20, 2003.

[SLMP03b] M. Sågfors, R. Ludwig, M. Meyer, J. Peisa, "Buffer Management for Rate-Varying 3G Wireless Links Supporting TCP Traffic," in Proc. 57th Vehicular Technology Conference (VTC), April 22-25, 2003.

[SM01] J. Sachs, M. Meyer, "Mobile Internet - Performance Issues Beyond the Radio Interface," in Proc. 10th Aachen Symposium on Signal Theory (ASST) - Algorithms and Software for Mobile Communications, Aachen, Germany, September 20-21, 2001.

[SM07] J. Sachs, P. Magnusson, "Assessment of the Access Selection Gain in Multi-Radio Access Networks," *European Transactions on Telecommunications*, November 1, 2007, DOI 10.1002/ett.1254.

[SMACK+04] J. Sachs, L. Muñoz, R. Aguero, J. Choque, G. Koudouridis, R. Karimi, L. Jorguseski, J. Gebert, F. Meago, F. Berggren, "Future Wireless Communication based on Multi-Radio Access," in Proc. 11th meeting of the Wireless World Research Forum, Oslo, Norway, June 10-11, 2004.

[SO08] J. Sachs, M. Olsson, "Access Network Discovery and Selection in the Evolved 3GPP Multi-Access System Architecture," submitted for publication, September 2008.

[SPG07] J. Sachs, M. Prytz, J. Gebert, "Multi-Access Management in Heterogeneous Networks," *Wireless Personal Communications*, November 2007, DOI 10.1007/s11277-007-9431-8.

[SRBW01] M. Schläger, B. Rathke, S. Bodenstein, A. Wolisz, "Advocating a Remote Socket Architecture for Internet Access using Wireless LANs," *Mobile Networks & Applications*, vol. 6, no. 1, January 2001.

[SS02] J. Sachs, A. Schieder, "Generic Link Layer," in Proc. 5th Meeting of the Wireless World Research Forum, Tempe, Arizona, USA, March 7-8, 2002.

[Ste94] W.R. Stevens, *TCP/IP Illustrated, volume 1 (The Protocols)*, Addison Wesley, November 1994.

[SWLM04] J. Sachs, H. Wiemann, J. Lundsjö, P. Magnusson, “Integration of Multi-Radio Access in a Beyond 3G Network,” in Proc. 15th IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), Barcelona, Spain, September 5-8, 2004.

[SWMWL04] J. Sachs, H. Wiemann P. Magnusson, P. Wallentin, J. Lundsjö, “A Generic Link Layer in a Beyond 3G Multi-Radio Access Architecture,” in Proc. International Conference on Communications, Circuits and Systems (ICCCAS), Chengdu, China, June 27-29, 2004

[SWW00] J. Sachs, S. Wager, H. Wiemann, "Performance of Shared and Dedicated Resources in UMTS," in Proc. 2nd IEEE Wireless Communications and Networking Conference (WCNC), Chicago, IL, USA, September 23-28, 2000.

[Tan96] A.S. Tanenbaum, *Computer Networks*, Prentice Hall, 1996.

[THH02] A. Tölli, P. Hakalin, H. Holma, "Performance Evaluation of Common Radio Resource Management (CRRM)," in Proc. IEEE International Conference on Communications (ICC), New York, USA, April 28 - May 2, 2002.

[TKSG04] U. Türke, M. Koonert, R. Schelb, C. Görg, “HSDPA Performance Analysis in UMTS Radio Network Planning Simulations,” in Proc. Vehicular Technology Conference (VTC), May 17-19, 2004.

[TLVK02] R. Tönjes, T. Lohmar, M. Vorwerk, R. Keller, “Flow Control for Multi-Access Systems,” in Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communication (PIMRC), September 2002.

[TMLW02] R. Tönjes; K. Moessner, T. Lohmar, M. Wolf, “OverDRiVE - Spectrum Efficient Multicast Services to Vehicles,” in Proc. IST Mobile Communications Summit 2002, Thessaloniki, June 17-19, 2002.

[TPSPS+07] H. Tang, P. Pöyhönen, O. Strandberg, K. Pentikousis, J. Sachs, F. Meago, J. Tuononen, R. Aguero, "Paging Issues and Methods for Multiaccess," in Proc. IEEE International Conference on Communications and Networking in China (ChinaCom), Shanghai, China, August 22-24, 2007.

[Tut99] W.H.W. Tuttlebee, “Software-defined radio: Facets of a developing technology”, *IEEE Personal Communications*, vol. 6, no. 2, 1999.

[Vik05] O. Viktorsson, “Business Aspects and Requirements on Ambient Networks”, presented at 1st Symposium of the Wireless World Initiative (WWI), Brussels, Belgium, December 9-10, 2004.

[VK04] H. Velayos, G. Karlsson, “Techniques to reduce the IEEE 802.11b handoff time,” in Proc. International Conference on Communications (ICC), June 2004.

[Wal02] B. Walke, *Mobile Radio Networks: Networking, Protocols and Traffic Performance*, John Wiley & Sons, Chichester, UK, 2nd edition, 2002.

[WMB06] B. Walke, S. Mangold, L. Berlemann, *IEEE 802 Wireless Systems*, John Wiley & Sons, Chichester, UK, 1st edition, 2006.

[WMH02] G. Wu, M. Mizuno, P. J. M. Havinga, “MIRAI Architecture for Heterogeneous Network,” *IEEE Communications Magazine*, vol. 40, no. 2, February 2002.

[WPBS04] B. Walke, R. Pabst, L. Berlemann, D. Schultz, “Architecture Proposal for the winner Radio Access Network and Protocol,” in Proc. 11th Meeting of the Wireless World Research Forum, Oslo, Norway, June 2004.

[YFPS05] O. Yilmaz, A. Furuskar, J. Pettersson, A. Simonsson, “Access Selection in WCDMA and WLAN Multi-Access Networks,” in Proc. 61st Vehicular Technology Conference (VTC), Stockholm, Sweden, May 30-June 1, 2005.

Standards

[3GPP22.234] 3GPP TS 22.234, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Requirements on 3GPP system to Wireless Local Area Network (WLAN) interworking (Release 7), V7.5.0, December 2006.

[3GPP22.934] 3GPP TR 22.934, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Feasibility study on 3GPP system to Wireless Local area Network (WLAN) interworking (Release 7), V7.0.0, June 2007

[3GPP22.980] 3GPP TR 22.980, 3rd Generation Partnership Project; 3GPP Technical Specification Group Services and System Aspects; Network composition feasibility study (Release 8); V8.1.0; June 2007.

[3GPP23.002] 3GPP TS 23.002, 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects; Network architecture (Release 6), V6.8.0, June 2005.

[3GPP23.032] 3GPP TS 23.032, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Universal Geographical Area Description (GAD) (Release 7), V7.0.0, June 2006.

[3GPP23.060] 3GPP TS 23.060, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; General Packet Radio Service (GPRS); Service description, Stage 2 (Release 7), V8.0.0; March 2008.

[3GPP23.107] 3GPP TS 23.107, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Quality of Service (QoS) concept and architecture (Release 6), V6.4.0; March 2006.

[3GPP23.203] 3GPP TS 23.203, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Policy and charging control architecture (Release 7); V7.1.0; December 2006.

[3GPP23.207] 3GPP TS 23.207, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; End-to-end Quality of Service (QoS) concept and architecture (Release 6), V6.6.0; September 2005.

[3GPP23.234] 3GPP TS 23.234, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; 3GPP system to Wireless Local Area Network (WLAN) interworking; System description (Release 7); V7.4.0; December 2006.

[3GPP23.271] 3GPP TS 23.271, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Functional stage 2 description of Location Services (LCS) (Release 7), V7.9.0, September 2007.

[3GPP23.401] 3GPP TS 23.401, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (Release 8), V8.3.0, September 2008.

[3GPP23.402] 3GPP TS 23.402, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Architecture enhancements for non-3GPP accesses (Release 8), V8.3.0, September 2008.

[3GPP23.882] 3GPP TR 23.882, 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; 3GPP System Architecture Evolution: Report on Technical Options and Conclusions (Release 7); V1.6.1; November 2006.

[3GPP23.893] 3GPP TR 23.893, 3rd Generation Partnership Project; Technical Specification Group Services and Architecture; Feasibility Study on Multimedia Session Continuity; Stage 2 (Release 8), V8.0.0, June 2008.

[3GPP24.008] 3GPP TS 24.008, Technical Specification 3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; Mobile radio interface Layer 3 specification; Core network protocols; Stage 3 (Release 8), V8.3.0, September 2008.

[3GPP25.301] 3GPP TS 25.301, 3rd Generation Partnership Project; Technical Specification Group RAN; Radio interface protocol architecture, V6.3.0, June 2005.

[3GPP25.304] 3GPP TS 25.304, Technical Specification, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; User Equipment (UE) procedures in idle mode and procedures for cell reselection in connected mode (Release 7), V7.1.0, December 2006.

[3GPP25.305] 3GPP TS 25.305, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Stage 2 functional specification of User Equipment (UE) positioning in UTRAN (Release 8), December 2007.

[3GPP25.308] 3GPP TS 25.308, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; High Speed Downlink Packet Access (HSDPA); Overall description; Stage 2 (Release 8), V8.3.0, September 2008.

[3GPP25.309] 3GPP TS 25.309, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; FDD Enhanced Uplink; Overall description; Stage 2 (Release 6) ,V6.6.0, March 2006.

[3GPP25.322] 3GPP TS 25.322, 3rd Generation Partnership Project; Technical Specification Group RAN; “Radio Link Control (RLC) protocol specification”, V6.6.0, December 2005.

[3GPP25.331] 3GPP TS 25.331, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Radio Resource Control (RRC); Protocol Specification (Release 7), V7.3.0, December 2006.

[3GPP25.401] 3GPP TS 25.401, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; UTRAN overall description (Release 8), V8.1.0, September 2008.

[3GPP25.832] 3GPP TR 25.832, 3rd Generation Partnership Project; Technical Specification Group RAN; Manifestations of Handover and SRNS Relocation, V4.0.0, March 2001.

[3GPP25.881] 3GPP TR 25.881, 3rd Generation Partnership Project; Technical Specification Group RAN; Improvement of Radio Resource Management (RRM) across RNS and RNS/BSS, V5.0.0, December 2001.

[3GPP25.899] 3GPP TR 25.899, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; High Speed Download Packet Access (HSDPA) enhancements (Release 6), V6.0.0, June 2004.

[3GPP25.913] 3GPP TR 25.913, 3rd Generation Partnership Project; Technical Specification Group RAN; "Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN) (Release 7)", V7.2.0, December 2005.

[3GPP25.936] 3GPP TR 25.936, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; "Handovers for real-time services from PS domain (Release 4), V4.0.1, December 2001.

[3GPP26.346] 3GPP TS 26.346. 3GPP Technical Specification Group Services and System Aspects; Multimedia Broadcast/Multicast Service (MBMS); Protocols and codecs (Release 8); V8.0.0; September 2008.

[3GPP29.060] 3GPP TS 29.060. 3GPP Technical Specification Group Core Network and Terminals; GPRS Tunnelling Protocol (GTP) across the Gn and Gp interface (Release 7); V7.4.0; December 2006.

[3GPP36.300] 3GPP TS 36.300 V1.0.0 (2007-03), Technical Specification, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 8), March 2007.

[3GPP43.051] 3GPP TS 43.051, 3rd Generation Partnership Project; Technical Specification Group GSM/EDGE Radio Access Network; Overall description - Stage 2; (Release 7), V7.0.0, August 2007.

[3GPP43.064] 3GPP TS 43.064, 3rd Generation Partnership Project; Technical Specification 3rd Generation Partnership Project; Technical Specification Group GSM/EDGE Radio Access Network; General Packet Radio Service (GPRS); Overall description of the GPRS radio interface; Stage 2 (Release 7), V7.0.0 2008.

[3GPP43.318] 3GPP TS 43.318, 3rd Generation Partnership Project; Technical Specification Group GSM/EDGE Radio Access Network; Generic access to the A/Gb interface; Stage 2 (Release 7); V7.0.0; November 2006.

[ETSI112] ETSI TR 101 112, European Telecommunication Standards Institute, Selection procedures for the choice of radio transmission technologies of the UMTS (UMTS 30.03, version 3.2.0). Technical report, April 1998.

[ETSI328] ETSI EN 300 328, European Telecommunication Standards Institute, Candidate Harmonized European Standard, Electromagnetic compatibility and Radio spectrum Matters (ERM); Wideband transmission systems; Data transmission equipment operating in the 2,4 GHz ISM band and using wide band modulation techniques; Harmonized EN covering essential requirements under article 3.2 of the R&TTE Directive; V1.7.1; May 2006.

[ETSI472] ETSI TS 102 472, European Telecommunication Standards Institute, Digital Video Broadcasting (DVB); IP Datacast over DVB-H: Content Delivery Protocols, Technical Specification, V1.2.1, December 2006.

[ETSI893] ETSI EN 301 893, European Telecommunication Standards Institute, Candidate Harmonized European Standard, Broadband Radio Access Networks (BRAN); 5 GHz high performance RLAN; Harmonized EN covering essential requirements of article 3.2 of the R&TTE Directive; V1.3.1; March 2005.

[ETSI957] ETSI TR 101 957, European Telecommunication Standards Institute, Broadband Radio Access Networks (BRAN); HIPERLAN Type 2; Requirements and Architectures for Interworking between HIPERLAN/2 and 3rd Generation Cellular systems, V1.1.1, August 2001.

[FCC15] Federal Communications Commission (FCC); Part 15 “Radio Frequency Devices”; Section 15.247 “Operation within the bands 902 - 928 MHz, 2400 - 2483.5 MHz, and 5725 - 5850 MHz”; May 4, 2007. Available at http://www.fcc.gov/oet/info/rules/ (accesses July 2007).

[ID-BMIP] K. El Malki, H. Soliman, “Simultaneous Bindings for Mobile IPv6 Fast Handovers,” Internet Engineering Task Force, work in progress, draft-elmalki-mobileip-bicasting-v6-03.txt, May 2003.

[ID-CAPa] P. Calhoun, M. Montemurro, D. Stanley, “CAPWAP Protocol Specification,” Internet Engineering Task Force, work in progress, draft-ietf-capwap-protocol-specification-11.txt, July 10, 2008.

[ID-CAPb] P. Calhoun, M. Montemurro, D. Stanley, “CAPWAP Protocol Binding for IEEE 802.11,” Internet Engineering Task Force, work in progress, draft-ietf-capwap-protocol-binding-ieee80211-07.txt, July 10, 2008.

[ID-HIP] R. Moskowitz, P. Nikander, P. Jokela, T. Henderson, “Host Identity Protocol,” Internet Engineering Task Force, work in progress, draft-ietf-hip-base-08.txt, June 11, 2007.

[ID-L2ABST] F. Teraoka, K. Gogo, K. Mitsuya, R. Shibui, K. Mitani, “Unified L2 Abstractions for L3-Driven Fast Handover,” Internet Research Task Force, work in progress, draft-irtf-mobopts-l2-abstractions-02.txt, February 20, 2007.

[ID-MCOA] R. Wakikawa, V. Devarapalli, T. Ernst, K. Nagami, “Multiple Care-of Addresses Registration,” Internet Engineering Task Force, draft-ietf-monami6-multiplecoa-09.txt, August, 2008.

[ID-NETLMM] H. Levkowetz, G. Giaretta, K. Leung, M. Liebsch, P. Roberts, K. Nishida, H. Yokota., M. Parthasarathy, "The NetLMM Protocol," Internet Engineering Task Force, work in progress, draft-giaretta-netlmm-dt-protocol-02, October 5, 2006.

[ID-NOMAD] N. A. Fikouras, A. Udugama, K. Kuladinithi, C. Görg, W. Zirwas, "Filters for Mobile IP Bindings (NOMAD)," Internet Engineering Task Force, work in progress, draft-nomad-mobileip-filters-05.txt, October 2003.

[ID-NOMAD6] K. Kuladinithi, N. A. Fikouras, C. Görg, K. Georgios, F. Pavlidou, "Filters for Mobile IPv6 Bindings (NOMADv6)," Internet Engineering Task Force, work in progress, draft-nomadv6-mobileip-filters-03.txt, October 2005.

[ID-POLIM] S. Aust, N. A. Fikouras, C. Pampu, C. Görg, "Policy based Mobile IPv6 Handover Decision (POLIMAND)," Internet Engineering Task Force, work in progress, draft-iponiar-dna-polimand-01.txt, February 2004.

[IEEE802.1X] IEEE 802.1X, Institute of Electrical and Electronics Engineer, Port-Based Network Access Control; IEEE Std 802.1X, 2004 Edition; December 2004.

[IEEE802.11] IEEE 802.11, Institute of Electrical and Electronics Engineer, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications; ANSI/IEEE Std 802.11, 1999 Edition (R2003); June 2003.

[IEEE802.11a] IEEE 802.11, Institute of Electrical and Electronics Engineer, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications; High-speed Physical Layer in the 5 GHz Band; IEEE Std 802.11a-1999 (R2003), June 2003.

[IEEE802.11b1] IEEE 802.11, Institute of Electrical and Electronics Engineer, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications; Higher-Speed Physical Layer Extension in the 2.4 GHz Band; IEEE Std 802.11b-1999 (R2003), June 2003.

[IEEE802.11b2] IEEE 802.11, Institute of Electrical and Electronics Engineer, 802.11 Amendment 2: Higher-speed Physical Layer (PHY) extension in the 2.4 GHz band—Corrigendum1, June 2003.

[IEEE802.11e] IEEE 802.11, Institute of Electrical and Electronics Engineer, 802.11 Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements", November 2005.

[IEEE802.11f] IEEE Std 802.11F-2003, Institute of Electrical and Electronics Engineer, IEEE Trial-Use Recommended Practice for Multi-Vendor Access Point Interoperability via an Inter-Access Point Protocol (IAPP) Across Distribution Systems Supporting IEEE 802.11 Operation, July 2003.

[IEEE802.11g] IEEE Std 802.11g-2003, Institute of Electrical and Electronics Engineer, 802.11 Amendment 4: Further Higher Data Rate Extension in the 2.4 GHz Band, June 2003.

[IEEE802.11h] IEEE Std 802.11b-2003, Institute of Electrical and Electronics Engineer, 802.11 Amendment 5: Spectrum and Transmit Power Management Extensions in the 5 GHz band in Europe", October 2003.

[IEEE802.11i] IEEE Std 802.11i-2004, Institute of Electrical and Electronics Engineer, 802.11 Amendment 6: Medium Access Control (MAC) Security Enhancements, July 2004.

[IEEE802.11u1] M. Moreton, "TGu Requirements", Contribution IEEE 802.11u, IEEE 802.11-05/0822r8, November 2005. (available at http://www.iab.org/liaisons/ieee/2005-12-ieee802-liaison-report.html)

[IEEE802.11u2] Institute of Electrical and Electronics Engineer, IEEE 802.11 Task Group U, Interworking with External Networks, http://grouper.ieee.org/groups/802/11/Reports/tgu_update.htm (accessed June 2008).

[IEEE802.16] IEEE 802.16 Working Group on Broadband Wireless Access Standards, Institute of Electrical and Electronics Engineer, http://www.ieee802.org/16/ (accessed July 2007).

[IEEE802.20] IEEE 802.20 Mobile Broadband Wireless Access (MBWA), Institute of Electrical and Electronics Engineer, http://www.ieee802.org/20/ (accessed July 2007).

[IEEE802.21] IEEE 802.21, Institute of Electrical and Electronics Engineer, Draft IEEE Standard for Local and Metropolitan Area Networks: Media Independent Handover Services; D02.00; September 2006.

[IETF-EAP] Internet Engineering Task Force, charter Extensible Authentication Protocol (eap), http://www.ietf.org/html.charters/eap-charter.html (accessed July 2007).

[IETF-MIP4] Internet Engineering Task Force, charter Mobility for IPv4 (mip4), http://www.ietf.org/html.charters/mip4-charter.html (accessed July 2007).

[IETF-MIP6] Internet Engineering Task Force, charter Mobility for IPv6 (mip6), http://www.ietf.org/html.charters/mip6-charter.html (accessed July 2007).

[R2-070036] 3rd Generation Partnership Project, contribution R2-070036 (Ericsson, Nokia, Siemens) "L2 Enhancements," TSG-RAN WG2#56bis, Sorrento, Italy, January 15-19, 2007.

[R2-072766] 3rd Generation Partnership Project, contribution R2-072766 (Ericsson) "Draft CR for lossless reconfiguration between fixed and flexible RLC PDU size," TSG-RAN WG2#58bis, Orlando, FL, USA, June 25 – 29, 2007.

[RFC793] J. Postel, "Transmission Control Protocol," Internet Engineering Task Force, RFC 793, September 1981.

[RFC813] David D. Clark, "Window and Acknowledgement Strategy in TCP," Internet Engineering Task Force, RFC 813, July 1982.

[RFC1122] R. Braden "Requirements for Internet Hosts - Communication Layers," Internet Engineering Task Force, RFC 1122, October 1989.

[RFC2018] M. Mathis, J. Mahdavi, S. Floyd, A. Romanow, “TCP Selective Acknowledgement Options,” Internet Engineering Task Force, RFC 2018, October 1996.

[RFC2205] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, “Resource ReSerVation Protocol (RSVP),” Internet Engineering Task Force, RFC 2205, September 1997.

[RFC2414] M. Allman, S. Floyd, C. Partridge, “Increasing TCP's Initial Window,” Internet Engineering Task Force, RFC 2414, September 1998.

[RFC2461] T. Narten, E. Nordmark, W. Simpson, “Neighbor Discovery for IP Version 6 (IPv6),” Internet Engineering Task Force, RFC 2461, December 1998.

[RFC2474] K. Nichols, S. Blake, F. Baker, D. Black, “Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers,” Internet Engineering Task Force, RFC 2474, December 1998.

[RFC2581] M. Allmann, V. Paxson, W. Stevens, “TCP Congestion Control,” Internet Engineering Task Force, RFC 2581, April 1999

[RFC2988] V. Paxson, M. Allman, “Computing TCP's Retransmission Timer,” Internet Engineering Task Force, RFC 2988, November 2000.

[RFC3032] E. Rosen, A. Viswanathan, R. Callon, “Multiprotocol Label Switching Architecture,” Internet Engineering Task Force, RFC 3032, January 2001.

[RFC3042] M. Allman, H. Balakrishnan, S. Floyd, “Enhanced TCP’s Loss Recovery Using Limited Transmit,” Internet Engineering Task Force, RFC 3042, January 2001.

[RFC3234] B. Carpenter, S. Brim, “Middleboxes: Taxonomy and Issues,” Internet Engineering Task Force, RFC 3234, February 2002.

[RFC3261] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, “SIP: Session Initiation Protocol,” Internet Engineering Task Force, RFC 4067, June 2002.

[RFC3303] P. Srisuresh, J. Kuthan, J. Rosenberg, A. Molitor, A. Rayhan, “Middlebox communication architecture and framework,” Internet Engineering Task Force, RFC 3303, August 2002.

[RFC3344] C. Perkins (ed.), “IP Mobility Support for IPv4,” Internet Engineering Task Force, IETF RFC 3344, August 2002.

[RFC3374] J. Kempf, “Problem Description: Reasons For Performing Context Transfers Between Nodes in an IP Access Network,” Internet Engineering Task Force, IETF RFC 3344, September 2002.

[RFC3453] M. Luby, L. Vicisano, J. Gemmel, L. Rizzo, M. Handley, J. Crowcroft, “The Use of Forward Error Correction (FEC) in Reliable Multicast,” Internet Engineering Task Force, RFC 3453, December 2002.

[RFC3517] E. Blanton, M. Allman, K. Fall, L. Wang, “A Conservative Selective Acknowledgement (SACK)-based Loss Recovery Mechanism for TCP,” Internet Engineering Task Force, RFC 3517, April 2003.

[RFC3775] D. Johnson, C. Perkins, J. Arkko, “Mobility Support in IPv6,” Internet Engineering Task Force, RFC 3775, June 2004.

[RFC3810] R. Vida, L. Costa, “Multicast Listener Discovery Version 2 (MLDv2) for IPv6,” Internet Engineering Task Force, RFC 3810, June 2004.

[RFC3819] P. Karn, C. Bormann, G. Fairhurst, D. Grossman, R. Ludwig, J. Mahdavi, G. Montenegro, J. Touch, L. Wood, “Advice for Internet Subnetwork Designers,” Internet Engineering Task Force, RFC3819, July 2004

[RFC3828] L-A. Larzon, M. Degermark, S. Pink, L-E. Jonsson, G. Fairhurst, “The Lightweight User Datagram Protocol (UDP-Lite),” Internet Engineering Task Force, RFC 3828, July 2004.

[RFC3926] T. Paila, M. Luby, R. Lehtonen, V. Roca, R. Walsh, “FLUTE - File Delivery over Unidirectional Transport,” Internet Engineering Task Force, RFC 3926, October 2004.

[RFC3963] V. Devarapalli, R. Wakikawa, A. Petrescu, P. Thubert, “Network Mobility (NEMO) Basic Support Protocol,” Internet Engineering Task Force, RFC 3963, January 2005.

[RFC3971] J. Arkko, J. Kempf, B. Zill, P. Nikander, “SEcure Neighbor Discovery (SEND),” Internet Engineering Task Force, RFC 3971, March 2004

[RFC3990] B. O'Hara, P. Calhoun, J. Kempf, “Configuration and Provisioning for Wireless Access Points (CAPWAP) Problem Statement,” Internet Engineering Task Force, RFC 3990, February 2005.

[RFC4066] M. Liebsch, A. Singh, H. Chaskar, D. Funato, E. Shim, “Candidate Access Router Discovery (CARD),” Internet Engineering Task Force, RFC 4066, July 2005.

[RFC4066] J. Loughney, M. Nakhjiri, C. Perkins, R. Koodli, “Context Transfer Protocol (CXTP),” Internet Engineering Task Force, RFC 4067, July 2005.

[RFC4068] R. Koodli, “Fast Handovers for Mobile IPv6,” Internet Engineering Task Force, RFC 4068, July 2005.

[RFC4080] R. Hancock, G. Karagiannis, J. Loughney, S. Van den Bosch, “Next Steps in Signaling (NSIS): Framework,” Internet Engineering Task Force, RFC 4080, June 2005.

[RFC4118] L. Yang, P. Zerfos, E. Sadot, “Architecture Taxonomy for Control and Provisioning of Wireless Access Points (CAPWAP),” Internet Engineering Task Force, RFC 4118, June 2005.

[RFC4140] H. Soliman, C. Castelluccia, K. El Malki, L. Bellier, “Hierarchical Mobile IPv6 Mobility Management (HMIPv6),” Internet Engineering Task Force, RFC 4140, August 2005.

[RFC4301] S. Kent, K. Seo, “Security Architecture for the Internet Protocol,” Internet Engineering Task Force, RFC 4301, December 2005.

[RFC4429] N. Moore, “Optimistic Duplicate Address Detection (DAD) for IPv6,” Internet Engineering Task Force, RFC 4429, April 2006.

[RFC4555] P. Eronen, “IKEv2 Mobility and Multihoming Protocol (MOBIKE),” Internet Engineering Task Force, RFC 4555, June 2006.

[RFC4566] M. Handley, V. Jacobson , C. Perkins, “SDP: Session Description Protocol,” Internet Engineering Task Force, RFC 4566, July 2006.

[RFC4907] B. Aboba, “Architectural Implications of Link Indications,” Internet Engineering Task Force, RFC 4907, June 2007.

[RFC4988] R. Koodli, C. Perkins, “Mobile IPv4 Fast Handovers,” Internet Engineering Task Force, RFC 4988, October 2007.

[RFC5053] M. Luby, A. Shokrollahi, M. Watson, T. Stockhammer, “Raptor Forward Error Correction Scheme for Object Delivery,” Internet Engineering Task Force, RFC 4988, September 2007.

[RFC5113] J. Arkko, B. Aboba, J. Korhonen, F. Bari, “Network Discovery and Selection Problem,” Internet Engineering Task Force, RFC 5113, January 2008.

[RFC5213] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, B. Patil, “Proxy Mobile IPv6,” Internet Engineering Task Force, RFC 5213, August 2008.

Reports, White Papers, etc.

[AN] Ambient Networks project, http://www.ambient-networks.org/ (accessed July 2007).

[AN D1-5] Ambient Networks, “AN Framework Architecture,” Technical Report (Deliverable D1-5), December 2005.

[AN D2-1] [67] Ambient Networks, "Multi Radio Access – Architecture Concepts,” Technical Report (Project Deliverable D2-1), June 2004.

[AN D2-2] [67] Ambient Networks, "Draft Multi-Radio Access Architecture,” Technical Report (Project Deliverable D2-2), December 2004.

[AN D2-4] [67] Ambient Networks, "Multi-Radio Access Architecture,” Technical Report (Project Deliverable D2-4), December 2005.

[AN D2C1] [67] Ambient Networks, "Multi-Access and ARI, Design and Initial Specification,” Technical Report (Project Deliverable D2-C1), December 2006.

[AN D3G1A] Ambient Networks, “SNAP: A symmetric version of QNAP,” Technical Annex to Technical Report (Project Deliverable D3-G1) “Design of Composition Framework,” November 2006.

[AN D4-2] Ambient Networks, “D-4-2: Ambient Networks Mobility Architecture & Concepts,” Technical Report, March 2004

[AN D6H2] Ambient Networks, “First System Evaluation Results,” Technical Report (Project Deliverable D6-H2), January 2007.

[AN D7-1] Ambient Networks, "Ambient Networks Intermediate Security Architecture,” Technical Report (Project Deliverable D7-1), February 2005.

[AN D7-2] Ambient Networks, “Quick NAP - Secure and Efficient Network Access Protocol,” Technical Annex 4 to Technical Report (Project Deliverable D7-2) “Ambient Network Security Architecture,” December 2005.

[AN D7A2a] Ambient Networks, “Draft System Description,” Technical Report (Project Deliverable D7-A2), December 2006.

[AN D7A2b] Ambient Networks, “Quick NAP - Secure and Efficient Network Access Protocol,” Technical Annex to Technical Report (Project Deliverable D7-A2) “Draft System Description,” December 2006.

[AN D15C2] [67] Ambient Networks, “Multi-Access System Design and Specification,” Technical Report (Project Deliverable D15-C2), December 2007.

[AN D18A4] Ambient Networks, “AN System Description,” Technical Report (Project Deliverable D18-A4), December 2007.

[AN D20B2] Ambient Networks, “Mobility support: System specification, implementation and evaluation,” Technical Report (Project Deliverable D20-B2), December 2007.

[67] Co-authored reports.

[AN D21C3] [67] Ambient Networks, “Multi-Access Evaluation and Assessment,” Technical Report (Project Deliverable D21-C3), December 2007.

[AN MC2] [67] Ambient Networks, “Multi Access Simulation Definition,” Technical Report (Milestone M-C.2), June 2006.

[AN MC4] [67] Ambient Networks, “Multi-Access System Design Overview and Interfaces Specification Step 1,” Technical Report (Milestones M-C.3, M-C.4), July 2006.

[AN MC7] [67] Ambient Networks, “Initial Multi Access Simulation Implementation,” Technical Report (Milestone M-C7), October 2006.

[AN MC10] [67] Ambient Networks, “Final Multi Access Simulation Implementation,” Technical Report (Milestone M-C.10), February 2007.

[AN MC12] [67] Ambient Networks, “Multi-Access Intermediate Performance Evaluation,” Technical Report (Milestone M-C.12), May 2007.

[AN MC15] [67] Ambient Networks, “Multi-Access Final Performance Evaluation,” Technical Report (Milestone M-C.15), November 2007.

[AN R1-8] [67] Ambient Networks, “Common Ambient Networks Use Cases,” Technical Report (R1-8), December 2005.

[AN R2-4] [67] Ambient Networks, “Generic Link Layer Functionality,” Technical Report (R2-4), November 2004.

[AN R2-6] Ambient Networks, “Economic Feasibility,” Technical Report (R2-6), June 2005.

[AN R2-7] Ambient Networks, “MRRM Architecture Feasibility,” Technical Report (R2-7), June 2005.

[AN R2-8] [67] Ambient Networks, “Feasibility of Generic Link Layer,” Technical Report (R2-8), August 2005.

[AN R2-9] Ambient Networks, “Non-Conventional Low-Cost Concept Feasibility,” Technical Report (R2-9), July 2005.

[AN R2-10] [67] Ambient Networks, “Proof of Multi-Radio Access Concepts,” Technical Report (R2-10), September 2005.

[AN-ANA] [67] Ambient Networks, "Ambient Network Attachment,” Technical Report, January 2007.

[AN-CC] [67] Ambient Networks, "Ambient Networks Task Force Access Selection and Mobility & Common Use Case Continuous Connectivity,” Technical Report, December 2005.

[AN-MP+06] [67] P. Magnusson, M. Prytz, J. Sachs, T. Rinta-Aho, "Link Cost Abstraction,” Ambient Networks Technical Report, March 2006.

[AN-MRA] [67] Ambient Networks, "White Paper: Multi-Radio Access in an Ambient Networks World,” Technical Report, September 2004.

[AN-SM+06] [67] J. Sachs, P. Magnusson, M. Prytz, T. Rinta-Aho, "Link Performance Abstraction,” Ambient Networks Technical Report, March 2006.

[AN-SO07] [67] J. Sachs, M.F. Orué , "Investigation of Composite Access Selection Strategies Based on Different Interests in Collaborative Multi-Access Networks," Ambient Networks Technical Report, May 2007.

[AN-ST07] [67] J. Sachs, A. T. Tran, "Evaluation of Access Advertisement, Discovery and Attachment Schemes for Multi-Access Networks," Ambient Networks Technical Report, May 2007.

[BMBF04] [67] German Federal Ministry of Research and Education (BMBF) "An Overview of Energy Efficiency Techniques for Mobile Communication Systems," White Paper Research Focus Mobile Communications, Report of the Working Group 7 "Low-Power Broadband Wireless Communications," June 2004.

[CO99] COST. Action 231, Final report, Digital Mobile Radio Towards Future Generation Systems. Technical Report, European Commission, Brussels, 1999. Citation: ch. 3.6

[ERI07] Ericsson Press Release, "Ericsson reports continued solid performance," July 20, 2007. http://www.ericsson.com/ericsson/press/releases/20070720-1140997.shtml (accessed July 2007)

[IM04] IP Monitoring Project, measurements from February 6th, 2004, available at http://ipmon.sprint.com/

[IPONAIR] IPonAir Project, http://www.iponair.de/ (accessed July 2007).

[ISTAT07] Internet World Stats, "World Internet Usage and Population Statistics," http://www.internetworldstats.com/stats.htm (accessed July 2007).

[Khu05] B. S. Khurana, "Performance Evaluation of a Generic Link Layer for Multi-Radio Wireless Networks," Diploma Thesis, Institute for Wireless Networks, Aachen University (RWTH), Germany, December 2005.

[LLB05] S. Landström, L-Å. Larzon, U. Bodin, "Buffer management for TCP over HS-DSCH," Technical Report LTU TR 05⁄09 SE, Luleå University of Technology, Department of Computer Science and Networking, Sweden, September 2005.

[Tra07] A. T. Tran, "Evaluation of Access Advertisement, Discovery and Attachment Schemes for Multi-Access Networks," Master Thesis, Institute for Wireless Networks, Aachen University (RWTH), Germany, March 2007.

[UMTSF03] UMTS Forum, "Mobile Evolution – Shaping the Future," White Paper, August 2003. http://www.umts-forum.org/ (accesses July 2007)

[WWRFa02] [67] Wireless World Research Forum, "Reconfigurable Software Defined Radio Equipment and Supporting Networks - Research Thematics," Special Interest Group Reconfigurability, White Paper, Version 2.0, September 2002.

[WWRFb02] [67] Wireless World Research Forum, "Reconfigurable Software Defined Radio Equipment and Supporting Networks - Reference Models and Architecture," Special Interest Group Reconfigurability, White Paper, Version 2.0, Wireless World Research Forum, September 2002.

[Yil04] O. Yilmaz, “Access Selection in Multi-Access Cellular and WLAN Networks,” Master Thesis, Radio Communication Systems Laboratory, Royal Institute of Technology (KTH), Sweden, February 2005.

[YW06] S. Yankov, S. Wiethölter, "Handover Blackout Duration of Layer 3 Mobility Management Schemes,” Technical Report TKN-06-002, Telecommunication Networks Group, Technische Universität Berlin, May 2006.

List of Acronyms

2G	Second Generation RAT
3G	Third Generation RAT
3GPP	Third Generation Partnership Project
AAA	Authentication, Authorisation and Accounting
AAS	Active Access Set
AC	Access Controller
ACK	Acknowledgement
AGW	Access Gateway
AHO	Access Handover
ANAP	Ambient Networks Attachment Protocol
AP	Access Point
API	Application Programming Interface
AR	Access Resource
ARA	Access Resource Area
ARQ	Automatic Repeat Request
AS	Access Selection
ASP	Access Selection Period
AvS	Available Set of Radio Accesses
AWGN	Additive White Gaussian Noise
BBM	Break-Before-Make
BEP	Bearer Endpoint
BER	(Residual) Bit Error Rate
BDP	Bandwidth-Delay Product
BLER	Block Error Rate
BM	Bearer Management
BSC	Base Station Controller
BSS	Base Station Subsystem (GERAN)
BSS	Basic Service Set (WLAN)
BSSID	Basic Service Set Identifier
BTS	Base Transceiver Station
CARD	Candidate Access Router Discovery
CAPWAP	Control and Provisioning of Wireless Access Points
CAS	Candidate Access Set
CCA	Clear Channel Assessment
CDMA	Code Division Multiple Access
CMP	Context Management Protocol
CQI	Channel Quality Indicator
CRC	Cyclic Redundancy Check
CS	Circuit-Switched
CSAA	Connectivity Setup, Advertisement and Attachment
CSCF	Call Session Control Function
CSMA/CA	Carrier-Sense Multiple Access with Collision Avoidance
CQI	Channel Quality Indicator
CW	Contention Window
CWND	Congestion Window
CXTP	Context Transfer Protocol
DAD	Duplicate Address Detection
DAS	Detected Access Set
DCF	Distributed Coordination Function
DCH	Dedicated Channel
DIFS	DCF Interframe Space
DNS	Domain Name System
DSCH	Downlink Shared Channel
DSCP	DiffServ Codepoint
DSL	Digital Subscriber Line
DSLAM	Digital Subscriber Line Access Multiplexer
DUPACK	Duplicate Acknowledgement
EDGE	Enhanced Data Rates for GSM Evolution

EIRP	Equivalent Isotropically Radiated Power
EGPRS	Enhanced General Packet Radio Service
EIRP	Equivalent Isotropic Radiated Power
EPC	Evolved Packet Core
ESS	Extended Service Set (WLAN)
ETSI	European Telecommunications Standards Institute
E-UTRAN	Evolved UMTS Terrestrial Radio Access Network
FACH	Forward Access Channel
FC	Forwarding Control
FCS	Frame Check Sequence
FDMA	Frequency Division Multiple Access
FE	Functional Entity
FEC	Forward Error Correction
FEP	Flow Endpoint
FMIP	Fast Mobile IP
FP	Forwarding Point
FRAT	Future Radio Access Technology
GAN	Generic Access Network
GANC	Generic Access Network Controller
GGSN	Gateway GPRS Support Node
GLL	Generic Link Layer
GLL_{CA}	Generic Link Layer – Context Anchor
$GLL_{I\text{-}CT}$	Generic Link Layer – Interface and Context Anchor
GLL-IW	Generic Link Layer Interworking
GLL-MAC	Generic Link Layer – Medium Access Control
GLL-PDCP	Generic Link Layer – Packet Data Convergence Protocol
GLL-RLC	Generic Link Layer – Radio Link Control
GMSC	Gateway Mobile Switching Centre
GPRS	General Packet Radio Service
GSC	Generic Session Continuity
GSM	Global System for Mobile Communication
GTP	GPRS Tunnelling Protocol
GTP-U	GPRS Tunnelling Protocol – User Plane
GW	Gateway
HARQ	Hybrid Automatic Repeat Request
HIP	Host Identity Protocol
HMIP	Hierarchical Mobile IP
HOLM	Handover and Locator Management
HS	High-Speed Packet Access
HSDPA	High-Speed Downlink Packet Access
HSPA	High-Speed Packet Access
HSUPA	High-Speed Uplink Packet Access
HSS	Home Subscriber Server
IAPP	Inter-Access Point Protocol
I1, I2	Initiator messages of ANAP (based on HIP Base Exchange)
IE	Information Element
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IKE	Interne Key Exchange
IMS	IP Multimedia Subsystem
IP	Internet Protocol
IPsec	Internet Protocol Security
IPv4	Internet Protocol Version 4
IPv6	Internet Protocol Version 6

ISDN	Integrated Services Digital Network
IWLAN	Interworking WLAN
L2	Layer 2
L3	Layer 3
LLC	Logical Link Control
LTE	Long Term Evolution
MAA	Multi-Access Anchor
MAC	Medium Access Control
MAC-d	Dedicated MAC
MAC-r	RAT-specific MAC
MA-GW	Multi-Access Gateway
MBB	Make-Before-Break
MIP	Mobile IP
MLD	Multicast Listener Discovery
MOBIKE	Mobility and Multihoming for Internet Key Exchange
MPLS	Multi-Protocol Label Switching
MR-GLL	Multi-Radio Generic Link Layer
MRRM	Multi-Radio Resource Management
$MRRM_{ANF}$	Multi-Radio Resource Management – Access Network Function
$MRRM_{ASF}$	Multi-Radio Resource Management – Access Selection Function
$MRRM_{CMF}$	Multi-Radio Resource Management – Connection Management Function
MSC	Mobile Switching Centre
NAD	Network Advertisement and Discovery
NSIS	Next Steps in Signalling
OBR	Object Bit Rate
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OSI	Open System Interconnection
PAN	Personal Area Network
PCF	Point Coordination Function
PCRF	Policy Control and Charging Resource Function
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
PDU	Packet Data Unit
PER	(Residual) Packet Error Rate
PF	Policy Function
PHY	Physical Layer
PLCP	Physical Layer Convergence Protocol
PMIP	Proxy Mobile IP
PN	Peer Network
PS	Packet-Switched
PSTN	Public Switched Telephone Network
QNAP	Quick Network Attachment Protocol
QoS	Quality of Service
R1, R2	Responder messages of ANAP (based on HIP Base Exchange)
RA	Radio Access
RACH	Random Access Channel
RAT	Radio Access Technology
RBG	Radio Bearer Gateway
RLC	Radio Link Control
RNC	Radio Network Controller
RNS	Radio Network Subsystem
RSSI	Received Signal Strength Indicator
RSVP	Resource Reservation Protocol
RTT	Round-Trip Time
SACK	Selective Acknowledgements
SAE	System Architecture Evolution

SC-FDMA	Single-Carrier Frequency Division Multiple Access
SDF	Service Data Flow
SDMA	Space Division Multiple Access
SDP	Session Description Protocol
SDU	Service Data Unit
SEND	Secure Neighbor Discovery
SGSN	Serving GPRS Support Node
SIFS	Short Interframe Space
SIP	Subscriber Identity Module
SINR	Signal to Interference and Noise Ratio
SIP	Session Initiation Protocol
SNAP	Symmetric Network Attachment Protocol
SNDCP	Subnetwork Dependent Convergence Protocol
SRNS	Serving RNS
STA	Station (WLAN)
SSID	Service Set Identifier
TCP	Transmission Control Protocol
TDMA	Time Division Multiple Access
TTI	Transmission Time Interval
TV	Television
UDP	User Datagram Protocol
UE	User Equipment
UMA	Universal Mobile Access
UMB	Universal Mobile Broadband
UMTS	Universal Mobile Telecommunications System
UN	User Network
UT	User Terminal
UTRAN	UMTS Terrestrial Radio Access Network
VAS	Validated Access Set
VCC	Voice Call Continuity
VLR	Visitor Location Register
VPN	Virtual Private Network
WCDMA	Wideband Code Division Multiple Access
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area Network
WTP	Wireless Termination Point
WWW	World Wide Web

List of Symbols

Symbol	Meaning	Context
α	Constant path loss component	Channel capacity (Chapter 4, Annex B)
α_A, α_B	Link layer transmission expansion factor due to ARQ for links A, B	Access handover (Chapter 6)
α_I	Service availability	Access Resource Abstraction (Chapter 5)
β	Distance-dependent path loss exponent	Channel capacity (Chapter 4, Annex B)
δ_{avail}	Relative amount of currently available resources	Access Resource Abstraction (Chapter 5)
Δd	Width of a ring around the centre of a radio cell	Channel capacity (Chapter 4, Annex B)
δ_{free}	Relative amount of currently free resources	Access Resource Abstraction (Chapter 5)
Δk	Width of a ring around the centre of a radio cell	Stochastic knapsack model (Section 4.6)
$\Delta SINR$	degradation term compared to Shannon capacity	Channel capacity, stochastic knapsack model (Chapter 4, Annex B)
ε	Spectral efficiency	Channel capacity (Annex B)
ε	Granularity of radio link characteristics used to determine traffic classes	Stochastic knapsack model (Section 4.6)
ε_{max}	maximum spectral efficiency	Channel capacity (Chapter 4, Annex B)
η_k	User density distribution of service class k	Stochastic knapsack model (Section 4.6)
λ	Wavelength	Channel capacity (Chapter 4, Annex B)
λ_k	Arrival rate of service request of class k	Stochastic knapsack model (Section 4.6)
λ_n	Service request	Channel capacity (Chapter 4, Annex B)
μ_k	Inverse of the mean holding time of service request of class k	Stochastic knapsack model (Section 4.6)
σ	User distribution density	Channel capacity (Annex B)

$\sigma_{i,min}$	Relative resource efficiency with minimum performance	Access Resource Abstraction (Chapter 5)
$\sigma_{i,extra}$	Relative resource efficiency with extended performance	Access Resource Abstraction (Chapter 5)
A	Access allocation	Access selection (Chapter 4)
a_k	Area element of a radio cell with similar radio link properties	Stochastic knapsack model (Section 4.6)
B, B_i	Channel carrier bandwidth	Channel capacity (Chapter 4, Annex B)
BDP	Bandwidth-delay-product	Access handover (Chapter 6)
b_k	Resource requirements of service class k in the stochastic knapsack	Stochastic knapsack model (Section 4.6)
B_k	Blocking probability of service class k in the stochastic knapsack	Stochastic knapsack model (Section 4.6)
$BLER$	Block error rate	Access handover (Chapter 6)
c	Speed of light	Channel capacity (Chapter 4, Annex B)
C	Capacity of the stochastic knapsack	Stochastic knapsack model (Section 4.6)
C_i, C_{link}, C_n, $C_{Shannon}$	Capacity of a link	Channel capacity (Chapter 4, Annex B)
C_{AWGN}, $C_{Rayleigh}$	Capacity of a link for additive white Gaussian noise or Rayleigh fading	Channel capacity (Chapter 4, Annex B)
C_s	Cost-dependent component of end user business utility	Access selection (Chapter 4)
CW_{min}	Minimum contention window size of WLAN	Channel capacity, WLAN, Access discovery and attachment (Chapter 4, Chapter 5)
d	Distance between transmitter and receiver	Channel capacity (Chapter 4, Annex B)
D	Distance between radio cell centre to interferer	Channel capacity (Chapter 4, Annex B)
D_D	Amount of packet duplication at access handover	Access handover (Chapter 6)
D_L	Amount of packet loss at access handover	Access handover (Chapter 6)

d_k	Distance from cell centre	Stochastic knapsack model (Section 4.6)
D_R	Amount of packet re-ordering at access handover	Access handover (Chapter 6)
D_{Ro}	Offset of packet re-ordering at access handover	Access handover (Chapter 6)
e	Number of unused (empty) WLAN channels	WLAN, Access discovery and attachment (Chapter 5)
$E\{\}$	Expectation value	Channel capacity (Annex B)
$E_i\{\}$	Integral exponential function	Channel capacity (Annex B)
E_i	Component of access resource utility depending on resource efficiency	Access selection (Chapter 4)
f	Signal carrier frequency	Channel capacity (Chapter 4, Annex B)
f_{AS}	access selection function	Access selection (Chapter 4)
G_l	Combined antenna gain of transmitter and receiver	Channel capacity (Chapter 4, Annex B)
I	Interference power	Radio propagation, channel capacity, stochastic knapsack model (Chapter 4, Annex B)
k	Service class of the stochastic knapsack	Stochastic knapsack model (Section 4.6)
K	Number of service classes of the stochastic knapsack	Stochastic knapsack model (Section 4.6)
L_D	Distance-dependent path loss component	Radio propagation (Annex B)
L_i	Size of a message i	Access discovery and attachment (Chapter 5)
L_i	Load-dependent component of access resource utility	Access selection (Chapter 4)
L_L	Sum of antenna gain and feeder losses in path loss	Radio propagation (Annex B)
L_M	Multi-path component of path loss	Radio propagation (Annex B)

L_P	Path loss	Radio propagation, channel capacity, stochastic knapsack model (Chapter 4, Annex B)
L_S	Shadowing component of path loss	Radio propagation (Annex B)
M	Number of channels	Access Discovery and Attachment (Chapter 5)
MaxChannelTime	Maximum time for WLAN channel scanning	Access Discovery and Attachment (Chapter 5)
MinChannelTime	Minimum time for WLAN channel scanning	Access Discovery and Attachment (Chapter 5)
n_k	Number of elements of service class k in the stochastic knapsack	Stochastic knapsack model (Section 4.6)
N	Noise power	Radio propagation, channel capacity, stochastic knapsack model (Chapter 4, Annex B)
N	Number of active users in a radio cell	Channel capacity, WLAN, Access discovery and attachment (Chapter 4, Chapter 5)
OBR	Object bit rate	Access handover (Chapter 6)
p_{Ba}	Access provider preference (for business utility)	Access selection (Chapter 4)
p_{Bs}	Service priority (for business utility)	Access selection (Chapter 4)
$P_C(N)$	Proportion of collisions experienced in WLAN for each successfully acknowledged MAC frame	WLAN, Access discovery and attachment (Chapter 4, Chapter 5)
$p_{i,j}$	Allocation probability of a service request to an access system	Stochastic knapsack model (Section 4.6)
P_{Int}	Transmit power of the interferer	Channel capacity (Chapter 4, Annex B)
p_R	Resource preference (for access resource utility)	Access selection (Chapter 4)
P_{Rx}	Received signal power	Channel capacity (Chapter 4, Annex B)
p_s	Service priority (for service utility)	Access selection (Chapter 4)

P_{Tx}	Transmitted signal power	Channel capacity (Chapter 4, Annex B)
Q_A	Queue size at link layer A at time of access handover	Access handover (Chapter 6)
$Q_{i,offered}$	Offered quality	Access Resource Abstraction (Chapter 5)
$q_{i,min}$	Relative resource usage with minimum performance	Access Resource Abstraction (Chapter 5)
$q_{i,extra}$	Relative resource usage with extended performance	Access Resource Abstraction (Chapter 5)
r, r_A, r_B	Data rate (on links A, B)	Access handover (Chapter 6)
R_i	Effective data rate of a user i	Channel capacity (Chapter 4)
R_i	Resource-dependent component of access resource utility	Access selection (Chapter 4)
R_{max}	Maximum data rate ($R_{max}=B\cdot\varepsilon_{max}$)	Channel capacity (Chapter 4, Annex B)
r_{max}	Maximum relative amount of useable resources	Access Resource Abstraction (Chapter 5)
r_{min}	Minimum required relative amount of resources	Access Resource Abstraction (Chapter 5)
r_{occ}	Occupied relative amount of resources	Access Resource Abstraction (Chapter 5)
R_s	Revenue component for network business utility	Access selection (Chapter 4)
RTT	Round-trip time	Access handover (Chapter 6)
RTT_{E2E}	End-to-end round-trip time	Access handover (Chapter 6)
S	Service data flow	Access selection (Chapter 4)
S	Signal power	Channel capacity (Chapter 4, Annex B)
s_B	Packet size of first packet on link B	Access handover (Chapter 6)
s_k	Size of an area element	Stochastic knapsack model (Section 4.6)
$SINR$	Signal-to-interference-and-noise ratio	Channel capacity, stochastic knapsack model (Chapter 4, Annex B)

SNR	Signal-to-noise ratio	Channel capacity (Chapter 4, Annex B)
T	Average throughput of elements entering the stochastic knapsack.	Stochastic knapsack model (Section 4.6)
T_A, T_B	Transmission time for a packet on link A, B	Access handover (Chapter 6)
T_{A0}, T_{B0}	Transmission latency on link A, B	Access handover (Chapter 6)
t_{ACK}	Duration of WLAN MAC acknowledgement	Channel capacity, WLAN, Access discovery and attachment (Chapter 4, Chapter 5)
$T_{AL2sync}$	Time for the L2 transmission status update procedure	Access handover (Chapter 6)
T_{Amax}	Maximum time that link A remains active after the access handover	Access handover (Chapter 6)
T_{Arem}	Time to transmit the remaining queue over link A	Access handover (Chapter 6)
T_{Aup}, T_{Bup}	Time to setup link layer connectivity for link A, B	Access handover (Chapter 6)
t_{busy}	Time that the WLAN channel is occupied by other users	WLAN, Access discovery and attachment (Chapter 5)
t_{cont}	Transmission overhead due to contention in WLAN	Channel capacity, WLAN, Access discovery and attachment (Chapter 4, Chapter 5)
T_c	Probe request transmission time in WLAN	WLAN, Access discovery and attachment (Chapter 5)
$T_{collissions}$	Total time of collisions in WLAN	WLAN, Access discovery and attachment (Chapter 5)
T_{CT}	Time to transmit context from link A to B	Access handover (Chapter 6)
t_{DIFS}	Duration of WLAN DCF interframe space	WLAN, Access discovery and attachment (Chapter 5)
$T_{discovery}$	Discovery delay for scanning WLAN channels	Access Discovery and Attachment (Chapter 5)

T_e	Time to scan an empty WLAN channel	WLAN, Access discovery and attachment (Chapter 5)
$T_{FP\text{-}A}$, $T_{FP\text{-}B}$	Transmission delay from forwarding point to GLL on link A, B	Access handover (Chapter 6)
T_i	Effective transmission time of a packet *i* in WLAN	Channel capacity, WLAN (Chapter 4)
T_I	Time of interruption at access handover	Access handover (Chapter 6)
t_{jam}	Channel occupancy of frames subject to collisions in WLAN	WLAN, Access discovery and attachment (Chapter 6)
$T_{N\text{-}user}$	WLAN channel occupancy of *N* users	WLAN, Access discovery and attachment (Chapter 6)
T_{other}	Channel usage of other users in WLAN	WLAN, Access discovery and attachment (Chapter 6Chapter 1)
t_{ov}	Transmission overhead for a packet in WLAN	Channel capacity, WLAN, Access discovery and attachment (Chapter 4, Chapter 5)
T_{own}	Channel usage of the considered user in WLAN	WLAN, Access discovery and attachment (Chapter 5)
$T_{PhyPreambe}$	Duration of WLAN physical layer preamble	WLAN, Access discovery and attachment (Chapter 5)
$T_{procedure}$	Transmission time of a signalling procedure	Channel capacity, WLAN (Chapter 4)
t_{SIFS}	Duration of WLAN short interframe space	WLAN, Access discovery and attachment (Chapter 5)
t_{slot}	Slot period for the contention window of WLAN	Channel capacity, WLAN, Access discovery and attachment (Chapter 4, Chapter 5)
t_{tr}	Physical layer transmission time of a packet in WLAN	Channel capacity, WLAN, Access discovery and attachment (Chapter 4, Chapter 5)

T_u	Time to scan an empty WLAN channel	WLAN, Access discovery and attachment (Chapter 5)
$T_{user\text{-}i}$	WLAN channel occupancy of user *i*	WLAN, Access discovery and attachment (Chapter 5)
u	Number of used WLAN channels	WLAN, Access discovery and attachment (Chapter 5)
U	Utilisation of the stochastic knapsack	Stochastic knapsack model (Section 4.6)
u_{Bu}	Business utility for end user	Access selection (Chapter 4)
u_{Bi}	Business utility for a network *i*	Access selection (Chapter 4)
u_d	Delay-dependent component of service utility	Access selection (Chapter 4)
u_G	Global utility for access selection	Access selection (Chapter 4)
u_q	Reliability-dependent component of service utility	Access selection (Chapter 4)
u_r	Rate-dependent component of service utility	Access selection (Chapter 4)
u_R	Access resource utility	Access selection (Chapter 4)
$U(R)$	Total number of users in cell of radio R	Annex B
u_S	Service utility	Access selection (Chapter 4)
u_{se}	Component of service utility depending on security level	Access selection (Chapter 4)
$VR(H)$	Upper edge of RLC ARQ receiver window	Access handover (Chapter 6)
$VR(R)$	Lower edge of RLC ARQ receiver window	Access handover (Chapter 6)
$VT(S)$	Upper edge of RLC ARQ transmitter window	Access handover (Chapter 6)
$VT(A)$	Lower edge of RLC ARQ transmitter window	Access handover (Chapter 6)
Z	Expected available resources in an access system	Stochastic knapsack model (Section 4.6)

Deutsche Zusammenfassung

Drahtlose Kommunikationsnetze erlauben Endnutzern jeder Zeit und an jedem Ort von Kommunikationsdiensten Gebrauch zu machen. Dabei existiert eine Vielzahl von Funkzugangstechniken, weitere werden entwickelt. Funkzugangstechniken unterscheiden sich beispielsweise in Reichweite und Abdeckung, ihrer spektralen Effizienz, Kapazität und maximalen Datenrate, in ihrer Komplexität und ihren Kosten, zudem unterstützen sie verschiedene Dienste und gestatten dem Nutzer Mobilität in unterschiedlichem Umfang. Je nach Funkumgebung, erwartetem Verkehrsaufkommen und den in Anspruch genommenen Diensten sind für den Netzbetreiber lokal unterschiedliche Funkzugangstechniken jeweils die geeignetsten. Eine dynamische Auswahl des Funkzugangssystems anhand der Kommunikationsanforderungen und des Netzzustands ermöglicht gleichzeitig, dass Endnutzer allzeit bestmöglich angebunden sind. Deshalb ist es erstrebenswert, mehrere Funkzugangstechniken in ein einheitliches Kommunikationsnetz zu integrieren (sogenannte *Netzkonvergenz*) und ein Zusammenspiel verschiedener Kommunikationsnetze mit unterschiedlichen Funkzugangstechniken zu ermöglichen. Diese Arbeit untersucht technische Lösungen der Netzkonvergenz einschliesslich der dazu erforderlichen Verwaltungsaufgaben und geeigneter Netzarchitekturen. Sie definiert dabei drei grundlegende Funktionen zur Verwaltung mehrerer Funkzugangstechniken: die *Funkzugangswahl*, die *Funkzugangsüberwachung* und den *Funkzugangswechsel*.

Die *Funkzugangswahl* beurteilt und vergleicht verschiedene Funkzugangssysteme und bestimmt, welches als geeignetstes für die jeweilige Kommunikationsverbindung zu nutzen ist. Die Entscheidung kann auf einer Vielfalt von Parametern beruhen, wie z.B. Eigenschaften der Funkverbindung, Verfügbarkeit von Ressourcen in den einzelnen Funkzugangssystemen oder Dienstanforderungen. Durchgeführt wird die Funkzugangswahl von einer oder mehreren *Multifunkverwaltungsfunktionen* (*multi-radio resource management)*. Anhand von Systemsimulationen wurde gezeigt, dass durch eine dynamische Funkzugangswahl in den meisten Szenarien ein deutlicher Mehrwert erzielt werden kann. Die entscheidenden Systemparameter sind dabei die vorhandenen Funkzugangstechniken, der Funknetzplan, die geographischen Nutzerverteilung und verschiedene Algorithmen der Funknetzwahl.

Eine Funkzugangswahl erfordert eine *Funkzugangsüberwachung*, um Informationen über die vorhandenen Funkzugangssysteme zu erhalten. Voraussetzung hierzu ist zunächst, verschiedene Funkzugangssysteme in einer generischen und vergleichbaren Form bewerten zu können. Dies leistet eine *generische Verbindungsschicht* (*generic link layer*), die aufgrund von funkzugangsspezifischen Metriken eine allgemeine Beschreibungsform liefert. Wie welche der erforderlichen Informationen nun erlangt werden, hängt von dem Verbindungszustand des Endgerätes ab. Ohne Netzverbindung kann ein Endgerät bereits lokale Messungen durchführen oder diese vom Zugangsnetz in Pilotkanälen erhalten. Andere Arten von Information erfordern, dass das Endgerät eine Verbindung zu dem Funkzugangssystem aufgebaut hat und dort angemeldet ist. In der Arbeit werden verschiedene Lösungen aufgezeigt, wie Informationen über ein Funkzugangssystem in verschiedenen Verbindungszuständen von einem Endgerät ermittelt werden können. Ein neues Anschlussverfahren erlaubt, unterschiedliche Informationen über das Zugangssystem in die Prozeduren des Verbindungsaufbaus und -anschlusses zu integrieren. Dieses Anschlussverfahren wird in einem Szenario untersucht, in dem die Endgeräte der Nutzer umliegende WLAN-Funkzugangssysteme bewerten. Demnach kann sowohl der Aufwand an

Signalisierung als auch die Zeit, die benötigt ist, um eine WLAN-Funkzugangssystem zu bewerten, deutlich reduziert werden.

Die Zugangswahl kann bestimmen, eine Datenverbindung auf ein neues Zugangssystem zu übertragen. Ein solcher *Funkzugangswechsel* ist auf verschiedene Arten zu realisieren. Ein mögliches Problem dabei kann die Verzerrung der übertragenen Daten darstellen. Für Dienste, die auf dem Übertragungssteuerungsprotokoll (TCP) basieren, wurde dies gezeigt und die Abhängigkeit von den Eigenschaften der Verbindungsschicht des alten und neuen Zugangs untersucht. Durch verschiedene Methoden der Kontextübertragung kann ein größerer Verlust an Dienstgüte verhindert werden.

Annex A Business Scenarios for Multi-Access Networks

In Section 4.4.1 we have introduced different business roles in the service provisioning chain for multi-access networks. There are a number of different scenarios of how these different business roles can be adopted by market players leading to different market constellations. In the following we present some scenarios and discuss how different accesses are provided by business players and how they are perceived by the user. The presented scenarios are merely examples; more business scenarios can exist.

Multi-access mobile network operator

A multi-access mobile network operator embeds all different business roles as depicted in Figure A.1. The operator has a business relationship with the end user; it acts as connectivity provider and access provider and owns the different access networks and the core network to offer connectivity to the user. At the same time the multi-access mobile network operator acts as service provider for certain types of services, like e.g. multimedia telephony. It furthermore provides content services to the end user, e.g. via a web portal. The operator may have a business relationship to other third party content providers which provide some content to the user; these content services may be enriched with specific user context information provided by the multi-access mobile network operator. For example, the operator may provide payment services for certain content providers; or it may provide location services for location-specific services like traffic telematics and navigation. The end user may furthermore have direct relationships to other service providers that are independent from the multi-access mobile network operator, like e.g. corporate services, or Internet services. The multi-access mobile network operator merely acts as communication provider to connect to such services. The functionality for multi-access management is provided by the multi-access mobile network operator. Today's cellular operators act as multi-access mobile network operators; they provide access for the user via their one or more mobile radio access network and possibly also local or regional hotspot networks. The user can make use of any access network provided by the operator with which it has a business agreement (typically in form of a SIM card with subscription).

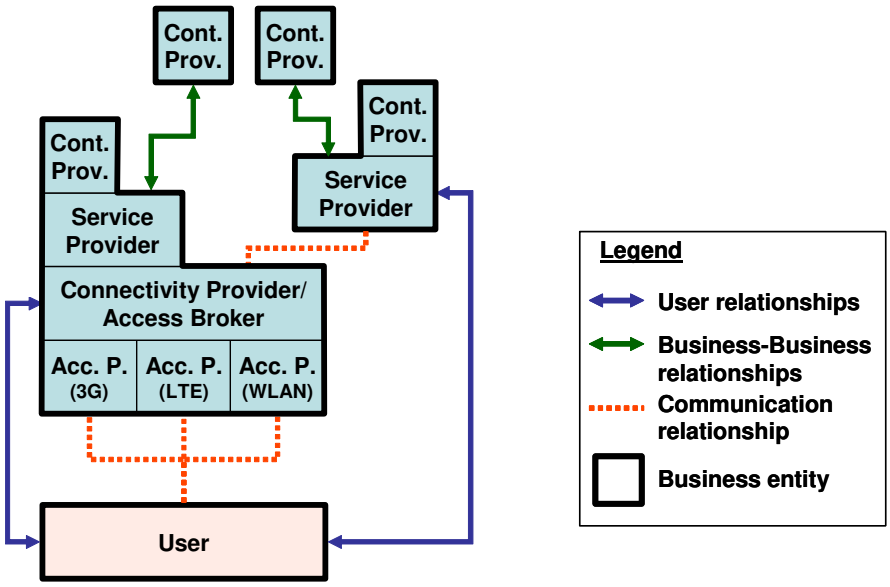


Figure A.1: Multi-access mobile network operator.

Roaming between mobile network operators

For (multi-access) mobile network operators it is possible to provide additional types of access networks to the users via roaming networks. This makes use of the roaming model, which was originally intended to provide communication services to users that are outside the coverage of the home operator. By cooperation between mobile network operators users can use the access network of a visited network to inter-connect to the home network and gain access to home network services, as shown in Figure A.2. Roaming agreements can also be established between mobile network operators that provide access services in the same geographic area allowing for multiple access networks being provided to end users. For mobile network operators this provides an efficient solution to offer a certain access capacity or access technology to its customers without an own infrastructure investment. At the same time revenues need to be shared between roaming partners. Roaming networks may be provided by other mobile network operators; it may also be a regional network operator providing e.g. WiMAX access. A roaming network may also be operated by a group of cooperating operators. For example, multiple operators with own 2G/3G mobile networks could operate a common converged mobile network; this would enable them to share costs for the converged network investment and operation. For the end user the access connectivity provided by the roaming network provides the same services as from the home network without a need for a separate business relationship to the visited network. The usage of access connectivity from the visited network may result in increased usage costs depending on the roaming agreements. The main functionality for multi-access management is provided by the home network, and partly also by the visited network.

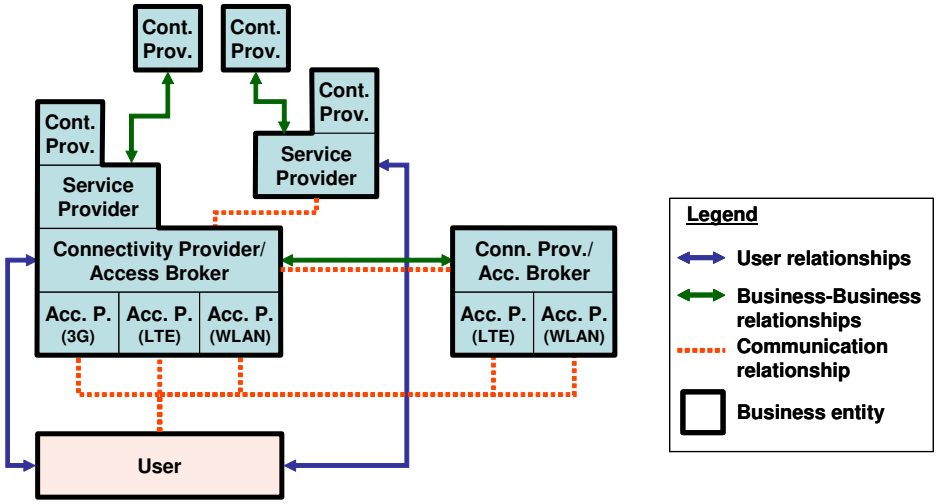


Figure A.2: Roaming between mobile network operators.

Mobile network operators with access brokers

Wide-area mobile communication services are provided by mobile network operators typically with a national coverage; conversely, local area communication services are confined to local regions, typically they are provided by private organisations and intended for closed user groups. Increasingly, also public local area wireless networks emerge; operators can be local operators or even municipalities. Local access providers can collaborate with wide-area mobile network operators according to the roaming model described above. However, with a large number of local access providers this leads to a very large number of roaming agreements. To simplify the business relationships, an access broker can bundle the access provided by multiple local access providers and establish roaming agreements with one or more mobile network operators, as shown in Figure A.3. The access broker maintains business relationships to the individual access providers. For the end user this results in the situation as in the roaming case; no individual agreements with access providers ort he access broker are required. The multi-access management functionality is mainly provided by the mobile network operator; and to some extent possibly also by the access broker.

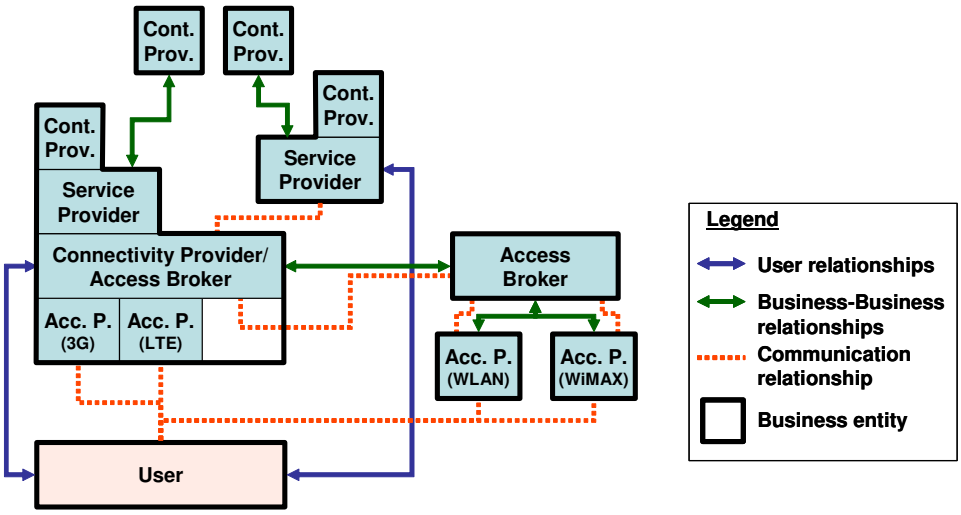


Figure A.3: Mobile network operators with access brokers.

Combined access brokers and connectivity providers

An operator can combine the role of access broker and connectivity provider. In this case all access connectivity is provided by independent access providers which have a business relationship with the combined access broker and connectivity provider (see Figure A.4). The access broker and connectivity provider maintains the business relationship to the end user. If a substantial number of access providers provide access connectivity, the access broker can make a valuable service offering to end users. The multi-access management functionality is provided by the access broker and connectivity provider.

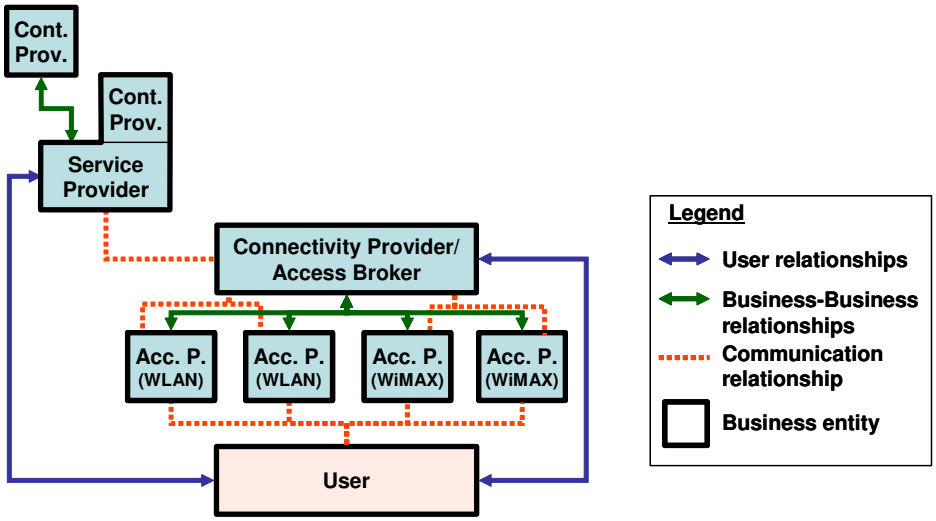


Figure A.4: Access brokers and connectivity providers.

Competing connectivity providers

A user may have business relationships with a multitude of connectivity providers and associated access providers without any cooperation between the connectivity providers, as shown in Figure A.5. The multi-access management functionality is mainly located in the user network.

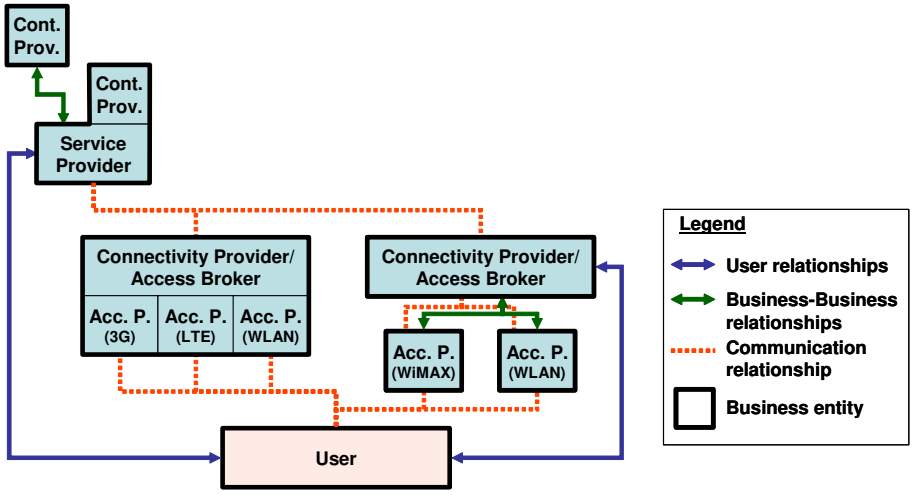


Figure A.5: Competing connectivity providers.

Annex B Wireless Transmission Characteristics

In wireless communication systems a radio signal is transmitted from the transmitter to the receiver and thereby information is transferred. The wireless connectivity is referred to as the radio link or radio channel. The link performance, or link capacity, of the radio link depends on the radio link quality and the channel bandwidth. The radio link quality is expressed by the amount of signal energy that is received at the receiver in relation to the amount of distortion caused by thermal noise and interference stemming from other undesired signals. The radio wave is attenuated on its way between transmitter and receiver according to a channel attenuation. Next we describe the radio link performance and the propagation loss. We discuss how the usage of radio resources is distributed within a radio cell.

B.1 Link Performance

The capacity of a communication channel with *additive white Gaussian noise* (AWGN) has been determined in the seminal work by Shannon as a function of the link quality measured at the receiver expressed as *signal-to-noise-ratio* (SNR) [Sha48]

$$\begin{aligned} C_{Shannon} &= B \cdot \log_2\left(1+\frac{S}{N}\right) \\ &= B \cdot \log_2\left(1+SNR\right) \end{aligned} \quad . \qquad \text{(B.1)}$$

where

SNR : signal-to -noise ratio at receiver,

S : signal power at the receiver,

N : (thermal) noise power at the receiver,

B : channel bandwidth.

Eq. (B.1) is also known as the Shannon-Hartley law of information theory[68] and it is typically referred to as the Shannon capacity of a channel or link. The Shannon capacity gives the theoretical upper bound of the transmission rate for error-free transmission of information in the presence of white Gaussian noise at a given SNR. It assumes ideal channel coding and modulation without time constraints. Later the law has been extended to include interference (*signal-to-noise-plus-interference-ratio*, SINR):

[68] R.V.L. Hartley first introduced a technical definition and a "measure" for information and described the information with the logarithm of the symbol alphabet size [Har28]. A brief overview of the history of information theory can be found, e.g., in [Lun02] [Los97] [WP06a].

$$C_{Shannon} = B \cdot \log_2\left(1 + \frac{S}{I+N}\right) = B \cdot \log_2(1+SINR) \quad \text{(B.2)}$$

From the link capacity the spectral efficiency can be derived, which is defined as the capacity normalised to the signal bandwidth:

$$\varepsilon = \frac{C_{link}}{B} = \log_2(1+SINR) \quad \text{(B.3)}$$

For a radio channel the SINR is time-varying; in an environment with multi-path propagation the SINR follows a Rayleigh distribution (in case of non-line-of-sight) or a Ricean distribution (in case of line-of-sight). In this case only a statistical mean for the channel capacity can be formulated, which is lower than the Shannon capacity for Gaussian noise, as has been shown by Lee [Lee90]. The channel capacity in a Rayleigh fading environment approaches again the Shannon capacity if the carrier bandwidth becomes very large or if maximum ratio combining of a large number of independent signal components is performed. Mohr has shown [Moh02] that the channel capacity for a Rayleigh fading channel is slightly smaller than for additive white Gaussian noise. He derived the following correction term:

$$\frac{C_{Rayleigh}}{C_{AWGN}} = \frac{e^{\frac{N \cdot E\{L_P\}}{P_{Tx}}} \cdot Ei\left\{-\frac{N \cdot E\{L_P\}}{P_{Tx}}\right\}}{\ln\left(1 + \frac{P_{Tx}}{N \cdot E\{L_P\}}\right)}, \quad \text{(B.4)}$$

where

$E\{\}$: expectation value,

$Ei\{\}$: “Integral Exponential Function”.

The difference compared to the Shannon capacity for AWGN channels depends on the SINR; in the typical range of -20 dB - +40 dB it reaches its maximum, with a difference in the order of 10%-15%.

Realistic transmission systems differ from the idealistic assumptions contained in the Shannon capacity formula; they deploy imperfect coding with finite memory and also only support a limited number of modulation and coding schemes. As a result the achievable transmission rate remains below the Shannon capacity. In different regions of SINR, different modulation and coding schemes achieve different transmission rates, at low SINR a robust modulation and coding scheme is best suited, whereas at high SINR a coding scheme with higher-order modulation and low code rate achieves the highest transmission rate. In most radio access systems link adaptation is deployed, which automatically selects the modulation and coding depending on the link quality. An example of link adaptation in the EDGE system according to [FNO99] is depicted Figure B.6.

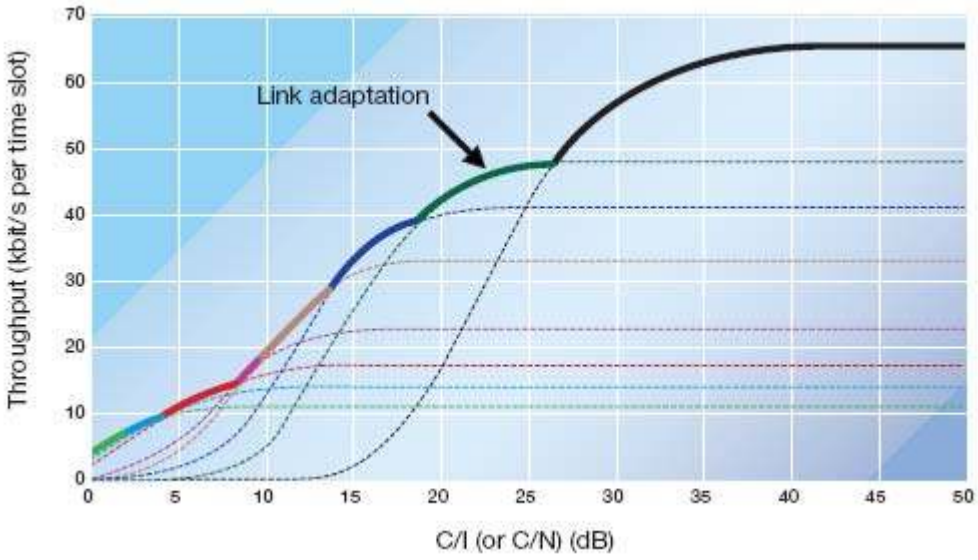


Figure B.6: Link adaptation for EDGE with eight different modulation and coding schemes (source [FNO99]).

Mohr [Moh03] has developed a model for the description of a general radio access technology with ideal link adaptation. This model approximates link capacity of a realistic radio link by an adapted Shannon capacity which is modified by two parameters: the capacity is shifted by a degradation term $\Delta SINR$, and it is upper bounded by a maximum spectral efficiency, ε_{max}, and a corresponding maximum data rate $R_{max}=B\cdot\varepsilon_{max}$:

$$C_{link} = \min\left\{B\cdot\log_2\left(1+10^{\frac{1}{10}\cdot\left(SINR_{[dB]}-\Delta SINR_{[dB]}\right)}\right); R_{\max}\right\}$$
$$= \min\left\{B\cdot\log_2\left(1+\frac{SINR}{\Delta SINR}\right); R_{\max}\right\} \qquad \text{(B.5)}$$

Typical values of ΔSINR are according to [Moh03] in the order of 10dB. The maximum data rate is determined by the highest modulation and coding scheme of the RAT. An approximate capacity of radio access technologies based on this generic model is shown in Figure B.7. The figure indicates the link capacity of HSPA (denoted as HS), WLAN 802.11 a and b and a future RAT (FRAT) that uses 20 MHz channel bandwidth. The maximum spectral efficiency is set according to the highest coding and modulation format. For FRAT the same maximum spectral efficiency is used as for HS. In Figure B.7 all RATs use the same ΔSINR, therefore the differences in link capacity result only from the differences in channel bandwidth and the differences in maximum spectral efficiency.

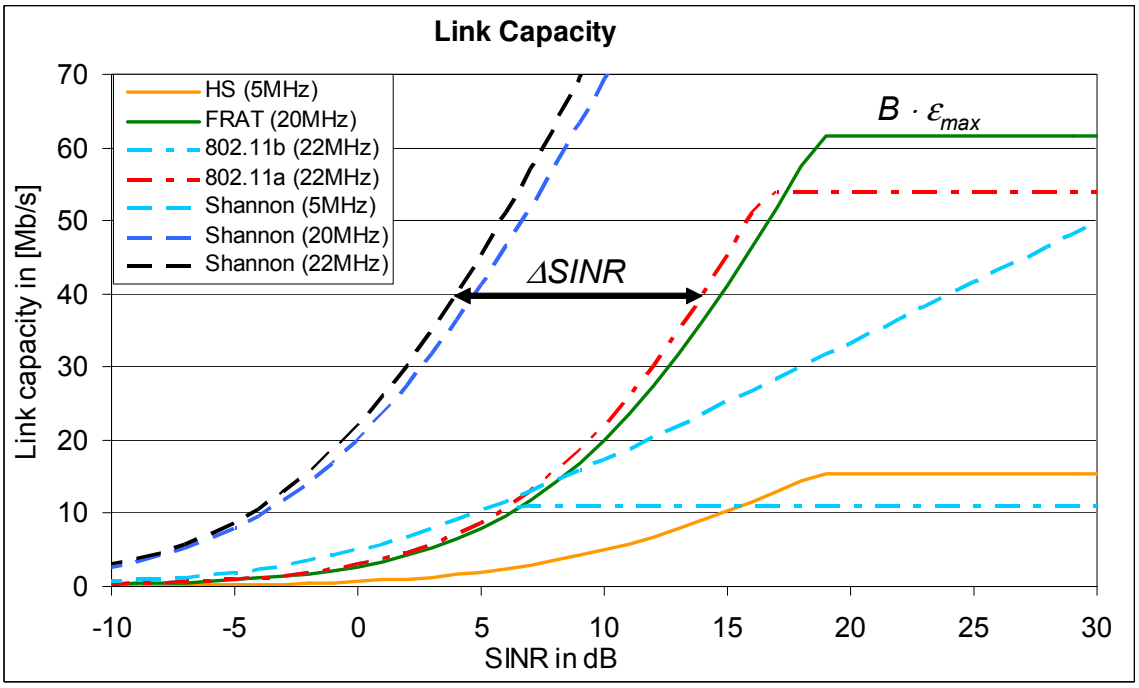


Figure B.7: Generic description of the performance of an adaptive radio interface according to [Moh03] (*ΔSINR*=10dB) with idealised examples for HS, FRAT, 802.11a and 802.11b and corresponding Shannon capacities.

B.2 Path Loss Models for Radio Propagation

In wireless communication systems radio signals are transmitted as electromagnetic waves from the sender to the receiver. The propagation of the radio waves follows the Maxwell's equations. When a radio wave travels away from the transmitter in free space the power flux density decreases due to the increase in the surface area of the wave front. The power P_{Rx} that is received at a receiver located at distance d from a transmitter that transmits with power P_{Tx} is

$$P_{Rx} = P_{Tx} \cdot G_l \cdot \left(\frac{\lambda}{4 \cdot \pi \cdot d} \right)^2 = P_{Tx} \cdot G_l \cdot \left(\frac{c}{4 \cdot \pi \cdot d \cdot f} \right)^2, \qquad \text{(B.6)}$$

where G_l is the combined antenna gain of the transmitter and receiver and λ is the wavelength of the signal. The free space path loss thus depends on the distance and the signal frequency f (with the speed of light $c = \lambda \cdot f$). In terrestrial wireless networks the radio channel is affected by a variety of objects along the transmission path; the radio wave is reflected, refracted or scattered at objects and it is attenuated when penetrating objects. Realistic values for the distance-dependent exponent of the path loss are in the range of 3.5 – 6, instead of 2 as in the case of free-space path loss (see eq. (B.6)). A description of propagation characteristics derived from Maxwell's equations is unfeasible in most practical situations due to the complex geometry of real world environments and missing knowledge about the correct material constants. Instead a stochastic description of the radio propagation is typically applied that describes the path loss accommodating for stochastic variations of terrain factors,

as can be seen in Figure B.8. The path loss, L_P, is typically expressed in logarithmic form and is considered to consist of four components:

$$L_P = L_D + L_L + L_S + L_M \,. \tag{B.7}$$

L_D describes the distance-dependent attenuation in signal power due to electro-magnetic propagation; it determines the global mean path loss. The signal power decays exponentially with the distance between sender and receiver. Around the global mean path loss an additional shadowing component L_S exists, which depends on local terrain-specific obstacles like tall buildings. It is typically described as a zero-mean log-normal distributed random component with a variance in the order of 6 dB – 10 dB. Finally, the superposition of multiple reflected signals leads to signal fluctuations when a mobile terminal moves distances in the order of the signal wavelength. This multi-path component, L_M, can cause signal fluctuations up to 30dB and is typically modelled by a Rayleigh distributed random component. The term L_L describes the sum of losses and gain in the antenna and feeder system of the sender and receiver. For example, an antenna typically has a directional antenna gain compared to an isotropic radiator. The cabling that connects the antenna to the transceiver provides a loss of the signal strength.

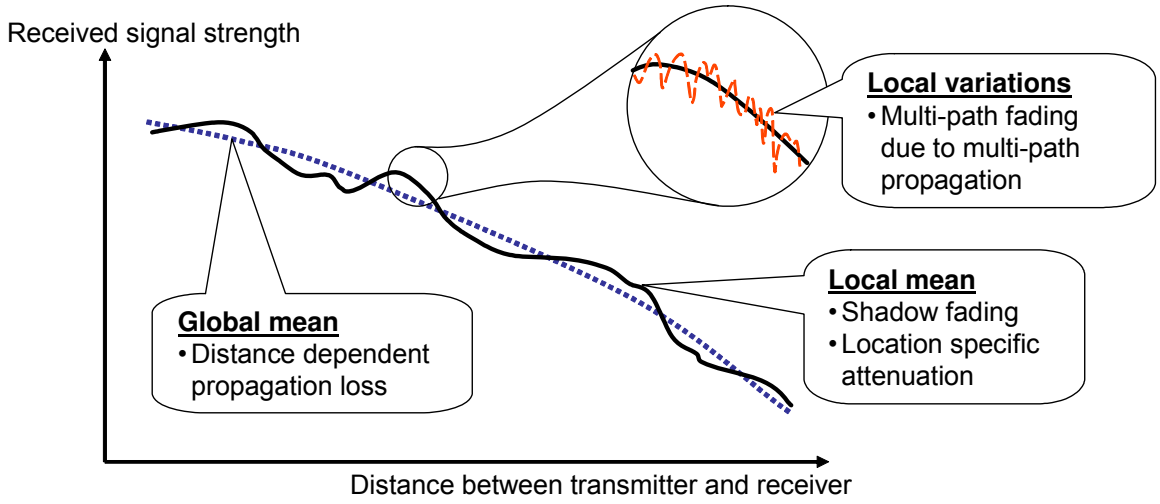


Figure B.8: Path loss characteristics of a radio link due to attenuation loss, shadow fading and multi-path fading.

Two types of propagation models are often used to describe the global mean path loss L_D: a single-slope model or a dual-slope model. The former is typically used when no direct line-of-sight connection exists (e.g. antennas are located above roof-top); the latter is typically used when propagation changes at a specific break point (e.g. when a line-of-sight connection exists up to a certain distance). For mobile cellular networks the Okumura-Hata model and its extension [CO99] are most frequently used (see Figure B.9). It is a single-slope propagation model. The path loss $L_D[dB]$ in decibel scale can be described as

$$L_{D[dB]} = 10 \cdot \alpha + 10 \cdot \beta \cdot \log(d)\,, \tag{B.8}$$

where d is the distance between sender and receiver, α is a constant that depends, for example, on the carrier frequency, antenna heights, and the type of radio environment, and β is the distance-dependent path loss exponent, which depends on the network environment. For WLAN networks often a dual-slope model is used, which uses the free-space path loss exponent up to a breakpoint distance (assuming line-of-sight) and then changes to a larger path loss exponent.

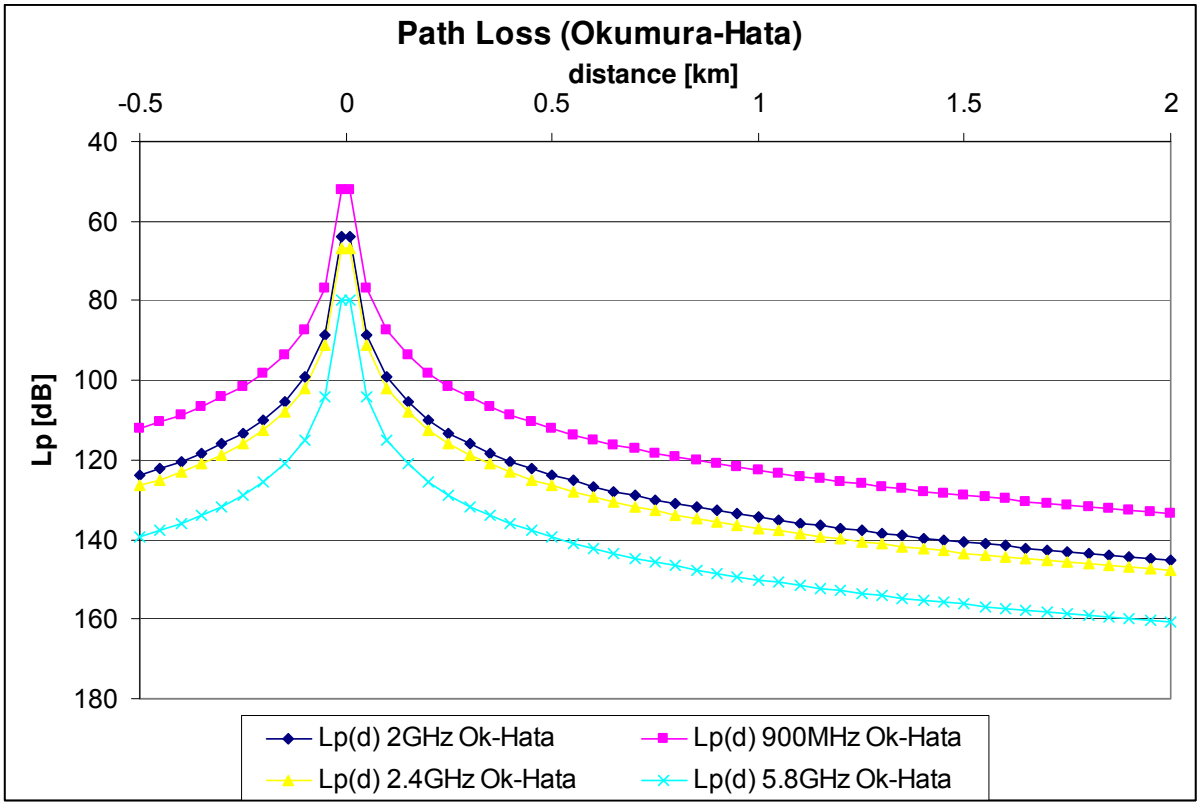


Figure B.9: Path loss (L_D) according to Okumura-Hata model at different carrier frequencies.

The path loss determines how the signal is received at the receiver. The received radio signal quality is described as the *signal-to-noise-and-interference ratio* (SINR); it is the transmitter power P_{Tx} divided by the path loss normalised to the thermal noise N plus interference I from other UNs:

$$SINR(d) = \frac{P_{Tx}}{L_p(d)} \cdot \frac{1}{I(d) + N} \qquad \text{(B.9)}$$

The interference *I(d)* also depends on the distance *d*. Interference can be divided into inter-cell interference and intra-cell interference. Inter-cell interference is caused by interfering radio cells using the same radio spectrum. These interfering cells are located at a certain interfering distance. The interference is thus larger at the cell edge, i.e. closer to neighbouring cells, compared to the cell centre. In cellular networks this distance is planned according to a cell-reuse plan to limit inter-cell interference. Intra-cell interference is caused by transmission of users in the same radio cell. It depends on the location and activity of those users. Some orthogonal radio access technologies can prohibit intra-cell interference owing to orthogonality of the different signals.

B.3 Resource Usage Distribution Within a Radio Cell

The radio link performance over a wireless channel depends on the radio link quality. The radio link quality is described as the signal-to-noise-and-interference-ratio, SINR, which is measured at the receiver:

$$SINR_i = \frac{S_i}{N + \sum_{\substack{j=1 \\ j \neq i}}^{K} I_j}. \qquad \text{(B.10)}$$

Apart from the thermal noise N the SINR depends on the interference caused by all other users in the system. Interferers can be differentiated into intra-cell interferers, where the interfering signals are from users within the same cell (see Figure B.10), and inter-cell interferers, where the interfering signals are from users in neighbouring cells (see Figure B.11). The amount of intra-cell interference depends on the multiple access scheme that is used. Typically orthogonal multiple access schemes are used, for example, code division multiple access (CDMA as used in UMTS/HSPA) or orthogonal frequency division multiple access (OFDMA as used in WiMAX and LTE). The orthogonal signal constellations prohibit intra-cell interference. However, due to the multipath characteristic of the wireless channel the orthogonality can be partly lost. This can result in intra-cell interference, which can be removed with additional mechanisms (e.g. multi-user detection in CDMA or a cyclic prefix in OFDMA).

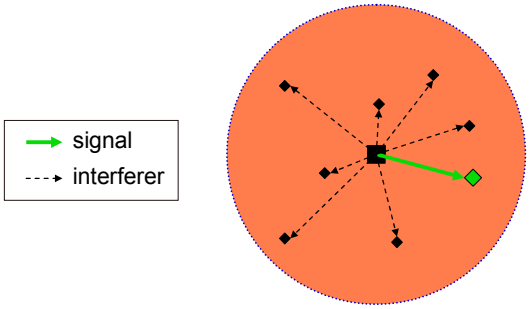


Figure B.10: Downlink intra-cell interference from other users in the same cell.

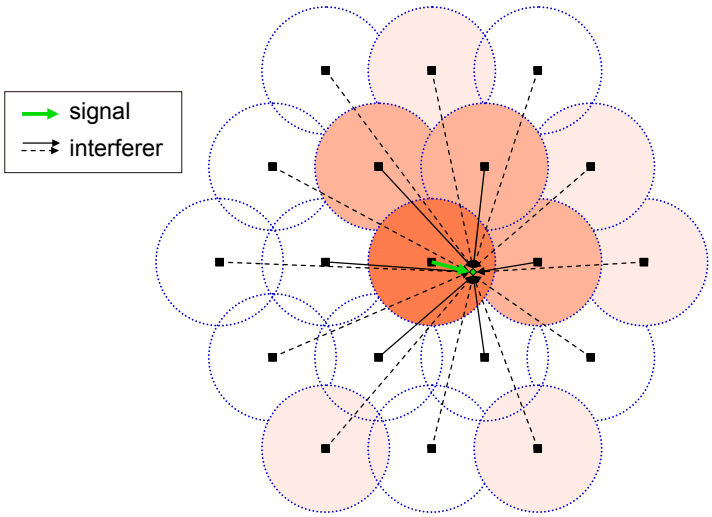


Figure B.11: Downlink inter-cell interference from neighbouring cells.

In order to estimate the amount of resources that are needed for the transmission to a particular user, it is required to know the amount of interference at the user location. For this we use a simplified interference model, as depicted in Figure B.12. We focus here on downlink transmission. All accumulated interference for a particular user can be modelled as being caused by a single interferer, which is located at distance D from the cell centre. The signal is transmitted over the distance r, whereas the interference is transmitted over distance D-r.

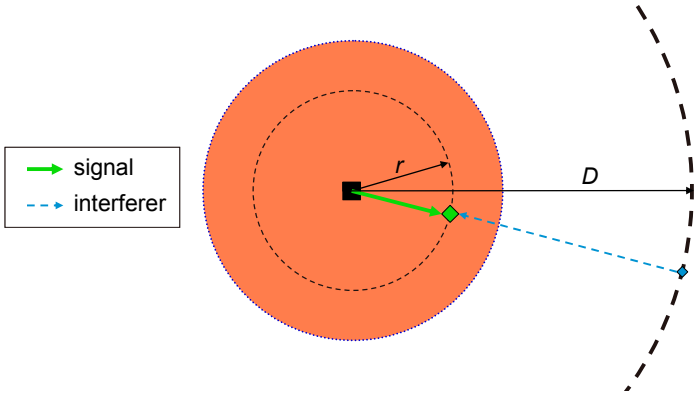


Figure B.12: Simplified interferer model, where all interferers are modelled as a single interferer.

The radio link quality required in order to achieve a particular channel capacity C can be determined from the modified Shannon formula of eq. (B.5) as

$$SINR = \Delta SINR\left(2^{C/B} - 1\right), \text{ with } C \le B \cdot \varepsilon_{\max}. \qquad \text{(B.11)}$$

We assume a scenario which is interference limited, which means that the amount of thermal noise is negligible compared to the interference. The SINR for the user is then according to our model in Figure B.12

$$SINR = \frac{P_{Tx}}{L_P(r)} \cdot \frac{L_P(D-r)}{P_{Int}}, \tag{B.12}$$

with

P_{Tx} : signal transmit power at the base station,

P_{Int} : transmit power of the interferer,

L_P : propagation path loss.

Solving eq. (B.12) for P_{Tx} and taking the SINR from eq. (B.11) we obtain

$$P_{Tx} = \left(2^{\frac{C}{B}} - 1\right) \cdot \Delta SINR \cdot P_{Int} \frac{L_P(r)}{L_P(D-r)}. \tag{B.13}$$

Assuming a typical path loss of the form

$$L_P(r)_{\mathrm{dB}} = 10\alpha + 10\beta \cdot \log r \tag{B.14}$$

$$L_P(r) = 10^{\alpha} \cdot r^{\beta} \tag{B.15}$$

eq. (B.13) becomes

$$P_{Tx} = \left(2^{\frac{C}{B}} - 1\right) \cdot \Delta SINR \cdot P_{Int} \frac{r^{\beta}}{(D-r)^{\beta}}. \tag{B.16}$$

The transmission power required to provide a certain link capacity to a user is depicted in Figure B.13 depending on distance *r*. A radio technology with 5 MHz bandwidth and a transmission power of 20 W is assumed; an interferer transmits with 20 W power is located at 2km distance, with an Okumura-Hata propagation model without shadowing and *ΔSINR* of 5 dB.

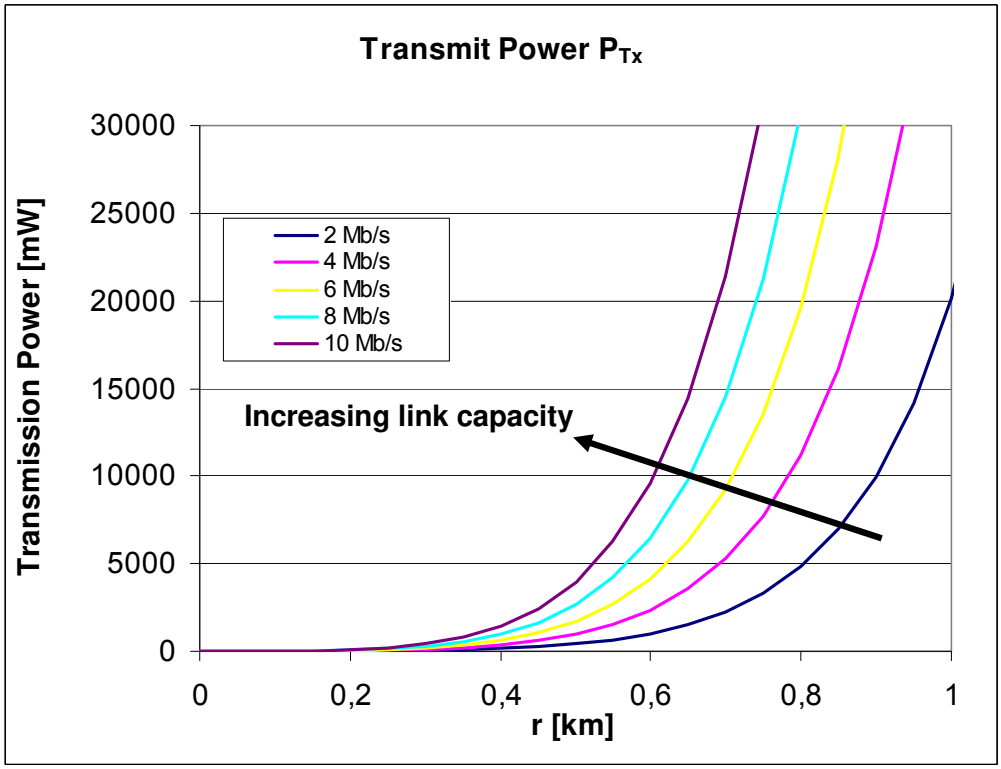


Figure B.13: Required transmission power for different link capacities (2-10 Mb/s).

If we assume a uniform user distribution density σ and a circular radio cell, the total number of users *U(R)* in the cell of radius *R* is

$$U(R) = \int_0^R 2 \cdot \pi \cdot r \cdot \sigma \cdot dr = \pi \cdot \sigma \cdot R^2, \tag{B.17}$$

and the number of users in a ring $[r_0, r_0+\Delta r]$, with inner radius r_0 and width Δr, is

$$\begin{aligned} U(r_0, \Delta r) &= \int_{r_0}^{r_0+\Delta r} 2 \cdot \pi \cdot r \cdot \sigma \cdot dr \\ &= \pi \cdot \sigma \cdot \left[(r_0 + \Delta r)^2 - r_0^{\,2} \right]. \\ &= \pi \cdot \sigma \cdot \left(\Delta r^2 + 2 \cdot r_0 \Delta r \right) \end{aligned} \tag{B.18}$$

The transmit power required by the users in the ring $[r_0, r_0+\Delta r]$ is

$$P_{tot}(r_0, r_0+\Delta r) = \int_{r_0}^{r_0+\Delta r} 2\pi\sigma \cdot r \cdot P_{Tx}(r) \cdot dr$$
$$= \int_{r_0}^{r_0+\Delta r} 2\pi\sigma \cdot r \cdot \left(2^{\frac{C}{B}} - 1\right) \cdot \Delta SINR \cdot P_{Int} \frac{r^{\beta}}{(D-r)^{\beta}}) \cdot dr \,. \qquad \text{(B.19)}$$
$$= 2\pi\sigma \cdot \Delta SINR \cdot P_{Int} \cdot \left(2^{\frac{C}{B}} - 1\right) \cdot \int_{r_0}^{r_0+\Delta r} \frac{r^{\beta+1}}{(D-r)^{\beta}}) \cdot dr$$

The integral in eq. (B.19) can be numerically solved. Figure B.14[69] shows both the transmission power required within a ring of 50m, as well as the accumulated transmission power up to a certain radius from the cell centre. It is clearly visible that the amount of used transmission resources increases towards the cell edges. The distribution of transmission power among the different users is presented in Figure 4.11 in Section 4.5.2.1.1; approximately 90% of the transmission power is consumed by 50% of the users located at the cell edge.

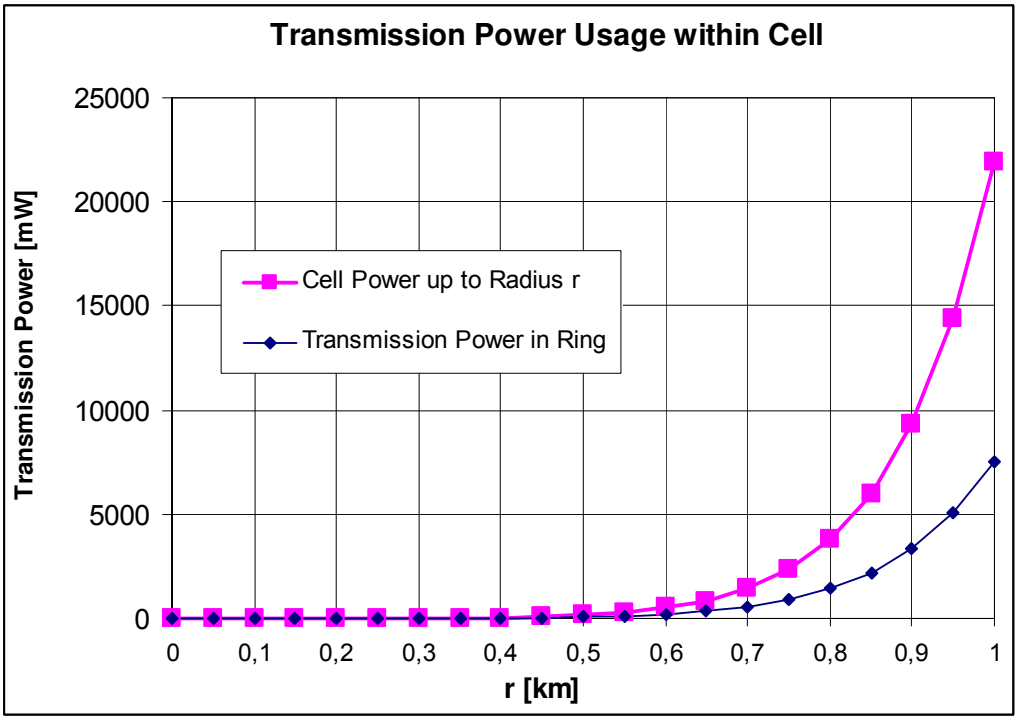


Figure B.14: Transmission power within radio cell with 20 users per cell running services at 500 kb/s.

[69] A user density of 20 users per radio cell with radius 1000m is assumed. All users are transmitting data at 500 kb/s.

Annex C Parameters for the Evaluation of Access Discovery and Attachment

C.1 Information Elements of Network Advertisements and Attachment

Network advertisements are used to provide a user network with sufficient information such that it is able to select an access network (see Section 5.4). Suitable information contained in advertisements has been investigated in [AN D6H2] [AN-ANA] [AN D18A4]. Advertisements contain the identifier of the network as well as services provided by the network. We assume in our analysis, as in [AN D6H2], that four different network service elements are included in the advertisement. Furthermore, during the attachment connectivity parameters are exchanged, like the locator (e.g. IP address) that is assigned to the user network, a gateway address (or *node ID router* according to [AN D18A4]). In addition a security association is established. For this several security parameters are exchanged, including Diffie-Hellman key exchange parameters, signatures and public keys of the networks, puzzle challenges and responses against denial of service attacks. For the performance evaluation of access network discovery and attachment in Section 5.4.4 it is sufficient to consider the lengths of different messages of the procedures. We have derived the values of message sizes from [AN D6H2] as listed in Table C-1 and Table C-2.

Table C-1: Parameters for independent *connectivity setup, advertisement and attachment*.

Messages		Information items	Size [Bytes]	Delay [µs]	Category
Legacy beacon	DIFS	Distributed Interframe spacing		50,00	PHY
	Backoff	Backoff		0,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	MAC Header	24	192,00	MAC
	Frame Body	Timestamp	**8**	**288,00**	MAC
		Beacon interval	**2**		MAC
		Capability information	**2**		MAC
		SSID (*name of network*) (variable: 0-32 bytes)	**18**		MAC
		Supported Rate (variable: 1-8 bytes)	**6**		MAC
	FCS	Frame Check Sequence	4	32,00	MAC
Authentication (Open systems)	DIFS	Distributed Interframe spacing		50,00	PHY
	Backoff	Backoff		0,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	MAC Header	24	192,00	MAC
	Frame Body	Authentication algorithm number (=0: Open)	**2**	**48,00**	MAC
		Authentication transaction sequence number (=1)	**2**		MAC
		Status code (reserved)	**2**		MAC
	FCS	Frame Check Sequence	4	32,00	MAC
ACK (Auth.)	SIFS	Short Interframe spacing		10,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	ACK MAC	14	112,00	MAC
Authentication - Response	DIFS	Distributed Interframe spacing		50,00	PHY
	Backoff	Backoff		192,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	MAC Header	24	192,00	MAC
	Frame Body	Authentication algorithm number (=0: Open)	**2**	**48,00**	MAC
		Authentication transaction sequence number (=2)	**2**		MAC
		Status code (=0)	**2**		MAC
	FCS	Frame Check Sequence	4	32,00	MAC
ACK (Auth.)	SIFS	Short Interframe spacing		10,00	PHY

	PHY	PHY Preamble		192,00	PHY
	MAC	ACK MAC	14	112,00	MAC
Association request	DIFS	Distributed Interframe spacing		50,00	PHY
	Backoff	Backoff		0,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	MAC Header	24	192,00	MAC
	Frame Body	Capability Information	**2**	**224,00**	MAC
		Listen interval	**2**		MAC
		SSID (*name of network*) (variable: 0-32 bytes)	**18**		MAC
		Supported Rate (variable: 1-8 bytes)	**6**		MAC
	FCS	Frame Check Sequence	4	32,00	MAC
ACK (Assoc.)	SIFS	Short Interframe spacing		10,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	ACK MAC	14	112,00	MAC
Association response	DIFS	Distributed Interframe spacing		50,00	PHY
	Backoff	Backoff		224,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	MAC Header	24	192,00	MAC
	Frame Body	Capability Information	**2**	**96,00**	MAC
		Status code	**2**		MAC
		Association ID	**2**		MAC
		Supported Rate (variable: 1-8 bytes)	**6**		MAC
	FCS	Frame Check Sequence	4	32,00	MAC
ACK (Assoc.)	SIFS	Short Interframe spacing		10,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	ACK MAC	14	112,00	MAC
Solicit Advertisement	DIFS	Distributed Interframe spacing		50,00	PHY
	Backoff	Backoff		0,00	PHY
	PHY	PHY Preamble		192,00	PHY
	L2	MAC Header	24	192,00	MAC
		LLC Header	8	64,00	MAC
	Frame Body	Flag to solicit advertisement	**2**	**16,00**	ANAP
	FCS	Frame Check Sequence	4	32,00	MAC
ACK (Solicit Adv)	SIFS	Short Interframe spacing		10,00	PHY
	PHY	PHY Preamble		192,00	PHY

	MAC	ACK MAC	14	112,00	MAC
Advertisement	DIFS	Distributed Interframe spacing		50,00	PHY
	Backoff	Backoff		192,00	PHY
	PHY	PHY Preamble		192,00	PHY
	L2	MAC Header	24	192,00	MAC
		LLC Header	8	64,00	MAC
	Frame Body	ID1	**16**	**448,00**	ANAP
		Each of 4 services available has: *Service type*	**1**		ANAP
		Each of 4 services available has: *Service specification*	**4**		ANAP
		Each of 4 services available has: *Service deployment mode*	**1**		ANAP
		Each of 4 services available has: *Payment method*	**1**		ANAP
		Each of 4 services available has: *Accounting data unit*	**1**		ANAP
		Each of 4 services available has: *Pricing information*	**2**		ANAP
		(other 3 services as above)	**30**		ANAP
	FCS	Frame Check Sequence	4	32,00	MAC
ACK (Adv.)	SIFS	Short Interframe spacing		10,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	ACK MAC	14	112,00	MAC
I1*	DIFS	Distributed Interframe spacing		50,00	PHY
	Backoff	Backoff		0,00	PHY
	PHY	PHY Preamble		192,00	PHY
	L2	MAC Header	24	192,00	MAC
		LLC Header	8	64,00	MAC
	Frame Body	ID1	**16**	**336,00**	ANAP
		ID2	**16**		ANAP
		Session ID	**2**		ANAP
		Proposed Security parameters	**8**		ANAP
	FCS	Frame Check Sequence	4	32,00	MAC
ACK (I1*)	SIFS	Short Interframe spacing		10,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	ACK MAC	14	112,00	MAC
R1*	DIFS	Distributed Interframe spacing		50,00	PHY
	Backoff	Backoff		0,00	PHY
	PHY	PHY Preamble		192,00	PHY
	L2	MAC Header	24	192,00	MAC

		LLC Header	8	64,00	MAC
	Frame Body	ID1	**16**	**3920,00**	ANAP
		ID2	**16**		ANAP
		Session ID	**2**		ANAP
		Selected Security parameters	**8**		ANAP
		Puzzle challenge (x1)	**64**		ANAP
		Public DH parameters (x2)	**192**		ANAP
		Public key of AN1 (x3)	**128**		ANAP
		Signature of (x1, x2, x3)	**64**		ANAP
	FCS	Frame Check Sequence	4	32,00	MAC
ACK (R1*)	SIFS	Short Interframe spacing		10,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	ACK MAC	14	112,00	MAC
I2*	DIFS	Distributed Interframe spacing		50,00	PHY
	Backoff	Backoff		0,00	PHY
	PHY	PHY Preamble		192,00	PHY
	L2	MAC Header	24	192,00	MAC
		LLC Header	8	64,00	MAC
	Frame Body	ID1	**16**	**5072,00**	ANAP
		ID2	**16**		ANAP
		Session ID	**2**		ANAP
		Puzzle response (x1)	**64**		ANAP
		Proposed Security parameters (x2)	**8**		ANAP
		Public DH parameters (x3)	**192**		ANAP
		Public key of AN2 (x4)	**128**		ANAP
		Additional AN2 info (Node ID) (x5)	**16**		ANAP
		Signature of (x1, x2, x3, x4, x5)	**64**		ANAP
		Encryption of public key of AN2	**128**		ANAP
	FCS	Frame Check Sequence	4	32,00	MAC
ACK (I2*)	SIFS	Short Interframe spacing		10,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	ACK MAC	14	112,00	MAC
R2*	DIFS	Distributed Interframe spacing		50,00	PHY
	Backoff	Backoff		0,00	PHY
	PHY	PHY Preamble		192,00	PHY
	L2	MAC Header	24	192,00	MAC
		LLC Header	8	64,00	MAC
	Frame	ID1	**16**	**2128,00**	ANAP

	Body	ID2	**16**		ANAP
		Session ID	**2**		ANAP
		Selected Security parameters (x1)	**8**		ANAP
		Encrypted public key of AN1 (x2)	**128**		ANAP
		Address of NID router (x3)	**16**		ANAP
		Address assigned to the AN2 (x4)	**16**		ANAP
		Signature of (x1, x2, x3, x4)	**64**		ANAP
	FCS	Frame Check Sequence	4	32,00	MAC
ACK (R1*)	SIFS	Short Interframe spacing		10,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	ACK MAC	14	112,00	MAC

Table C-2: Parameters for integrated *connectivity setup, advertisement and attachment*.

Messages		Information items		Size [Bytes]	Delay [µs]	Category
Extended beacon	DIFS	Distributed Interframe spacing			50,00	PHY
	Backoff	Backoff			0,00	PHY
	PHY	PHY Preamble			192,00	PHY
	MAC	MAC Header		24	192,00	MAC
	Frame Body	Timestamp		**8**	**288,00**	MAC
		Beacon interval		**2**		MAC
		Capability information		**2**		MAC
		SSID (*name of network*) (variable: 0-32 bytes)		**18**		MAC
		Supported Rate (variable: 1-8 bytes)		**6**		MAC
		ID1 (ID of Access AN)		**16**	**448**	ANAP
		4 services available and each of them has	*Service type*	**1**		ANAP
			Service specification	**4**		ANAP
			Service deployment mode	**1**		ANAP
			Payment method	**1**		ANAP
			Accounting data unit	**1**		ANAP
			Pricing information	**2**		ANAP
		(other 3 services as above)		**30**		ANAP
	FCS	Frame Check Sequence		4	32,00	MAC
I1*	DIFS	Distributed Interframe spacing			50,00	PHY
	Backoff	Backoff			0,00	PHY
	PHY	PHY Preamble			192,00	PHY
	L2	MAC Header		24	192,00	MAC
		LLC Header		8	64,00	MAC
	Frame Body	Capability Information		**2**	**224,00**	MAC
		Listen interval		**2**		MAC
		SSID (*name of network*) (variable: 0-32 bytes)		**18**		MAC
		Supported Rate (variable: 1-8 bytes)		**6**		MAC
		ID1		**16**	**336,00**	ANAP
		ID2		**16**		ANAP
		Session ID		**2**		ANAP
		Proposed Security parameters		**8**		ANAP
	FCS	Frame Check Sequence		4	32,00	MAC
ACK (I1*)	SIFS	Short Interframe spacing			10,00	PHY
	PHY	PHY Preamble			192,00	PHY
	MAC	ACK MAC		**14**	112,00	MAC

R1*	DIFS	Distributed Interframe spacing		50,00	PHY
	Backoff	Backoff		0,00	PHY
	PHY	PHY Preamble		192,00	PHY
	L2	MAC Header	24	192,00	MAC
		LLC Header	8	64,00	MAC
	Frame Body	Capability Information	**2**	**96,00**	MAC
		Status code	**2**		MAC
		Association ID	**2**		MAC
		Supported Rate (variable: 1-8 bytes)	**6**		MAC
		ID1	**16**	**3920,00**	ANAP
		ID2	**16**		ANAP
		Session ID	**2**		ANAP
		Selected Security parameters	**8**		ANAP
		Puzzle challenge (x1)	**64**		ANAP
		Public DH parameters (x2)	**192**		ANAP
		Public key of AN1 (x3)	**128**		ANAP
		Signature of (x1, x2, x3)	**64**		ANAP
	FCS	Frame Check Sequence	4	32,00	MAC

ACK (R1*)	SIFS	Short Interframe spacing		10,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	ACK MAC	14	112,00	MAC

I2*	DIFS	Distributed Interframe spacing		50,00	PHY
	Backoff	Backoff		0,00	PHY
	PHY	PHY Preamble		192,00	PHY
	L2	MAC Header	24	192,00	MAC
		LLC Header	8	64,00	MAC
	Frame Body	ID1	**16**	**5072,00**	ANAP
		ID2	**16**		ANAP
		Session ID	**2**		ANAP
		Puzzle response (x1)	**64**		ANAP
		Proposed Security parameters (x2)	**8**		ANAP
		Public DH parameters (x3)	**192**		ANAP
		Public key of AN2 (x4)	**128**		ANAP
		Additional AN2 info (Node ID) (x5)	**16**		ANAP
		Signature of (x1, x2, x3, x4, x5)	**64**		ANAP
		Encryption of public key of AN2	**128**		ANAP
	FCS	Frame Check Sequence	4	32,00	MAC

ACK (I2*)	SIFS	Short Interframe spacing		10,00	PHY

	PHY	PHY Preamble		192,00	PHY
	MAC	ACK MAC	14	112,00	MAC
R2*	DIFS	Distributed Interframe spacing		50,00	PHY
	Backoff	Backoff		0,00	PHY
	PHY	PHY Preamble		192,00	PHY
	L2	MAC Header	24	192,00	MAC
		LLC Header	8	64,00	MAC
	Frame Body	ID1	**16**	**2128,00**	ANAP
		ID2	**16**		ANAP
		Session ID	**2**		ANAP
		Selected Security parameters (x1)	**8**		ANAP
		Encrypted public key of AN1 (x2)	**128**		ANAP
		Address of NID router (x3)	**16**		ANAP
		Address assigned to the AN2 (x4)	**16**		ANAP
		Signature of (x1, x2, x3, x4)	**64**		ANAP
	FCS	Frame Check Sequence	4	32,00	MAC
ACK (R1*)	SIFS	Short Interframe spacing		10,00	PHY
	PHY	PHY Preamble		192,00	PHY
	MAC	ACK MAC	14	112,00	MAC

C.2 WLAN Transmission parameters

For the evaluation of access discovery and network attachment via WLAN the transmission parameters are used as listed in Table C-4 [IEEE802.11] [IEEE802.11a] [IEEE802.11b1] [IEEE802.11g].

Table C-3: WLAN acronyms.

ACK	Acknowledgement
CCA	Clear Channel Assessment
CW	Contention Window
DCF	Distributed Coordination Function
DIFS	DCF Interframe Space
FCS	Frame Check Sequence
LLC	Logical Link Control (IEEE 802.2)
MAC	Medium Access Control
PHY	Physical Layer
PLCP	Physical Layer Convergence Protocol
SIFS	Short Interframe Space

Table C-4: Parameters of WLAN 802.11a/b/g with *distributed coordination function*.

Characteristic	Value				Unit
	802.11b	802.11a	802.11g		
			802.11g - only	compatible with 802.11b	
Slot time	20	9	9	20	µs
SIFS	10	16	10	10	µs
DIFS	50	34	28	50	µs
CCA Time	≤ 15	< 4	<4	< 15	µs
RxTx Turn round Time	≤ 5	< 2	< 5	< 5	µs
RxTx Switch Time	≤ 5	« 1	« 1	≤ 1	µs
MAC Processing Delay	0	< 2	< 2	< 2	µs
PLCP Preamble	144 (72)	20	20	144 (72)	µs
PLCP Header	48 (24)	4	4	48 (24)	µs
CWmin	31	15	15	31	
CWmax	1023	1023	1023	1023	
Max Data rate	11	54	54	54	Mb/s
Supported Data rate	1, 2, 5.5, 11	6, 9, 12, 18,24,36,48, 54	6, 9, 12, 18, 24, 36, 48, 54	1, 2, 5.5, 6, 9, 11, 12, 18, 22, 24, 33, 36, 48, 54	Mb/s
MAC Header	192	192	192	192	bits
LLC Header	64	64	64	64	bits
FCS	32	32	32	32	bits
MAC Layer ACK	112	112	112	112	bits

C.3 Date rate versus Distance for WLAN 802.11

In WLAN systems the physical layer transmission mode is adapted to the radio link quality. The physical layer data rate is adapted according to the sensitivity thresholds for the received signal power as shown in Table C-5.

Table C-5: Receive sensitivity vs. data rate for WLAN 802.11 standards.

Data rate [Mb/s]	*Receive sensitivity [dBm]*		
	WLAN 802.11b	WLAN 802.11a	WLAN 802.11g
1	-90	-	-90
2	-87	-	-87
5.5	-83	-	-83
6	-	-82	-82
9	-	-81	-81
11	-80		-80
12	-	-79	-79
18	-	-77	-77
22	-	-	-76
24	-	-74	-74
33	-	-	-73
36	-	-70	-70
48	-	-66	-66
54	-	-65	-65

The received signal power corresponds to the radiated power minus the path loss. For the sensitivity thresholds also a fading margin is added to compensate fast signal variations during transmission. With the propagation model (see Section B.2 in Annex B) according to Table C-6 the rate can be determined depending on the distance of the user terminal from the WLAN access point as shown in Figure C.15. From this data rate regions can be determined that allow to estimate how frequently the transmission modes change for a moving user within a WLAN cell (see Section 5.4.5).

Table C-6: Propagation model for WLAN radio cell (without shadowing).

Parameters	**802.11a**	**802.11b**	**802.11g**
Frequency [MHz]	5600	2400	2400
Radiated power (EIRP) [dBm]	30 dBm (1W)	20 dBm (0.1W)	20 dBm (0.1W)
Path loss [dB]	Dual slope propagation model (path loss exponent α ; distance d): • α=2 for d≤8m • α=4 for d>8m		
Fading Margin [dB]	10		

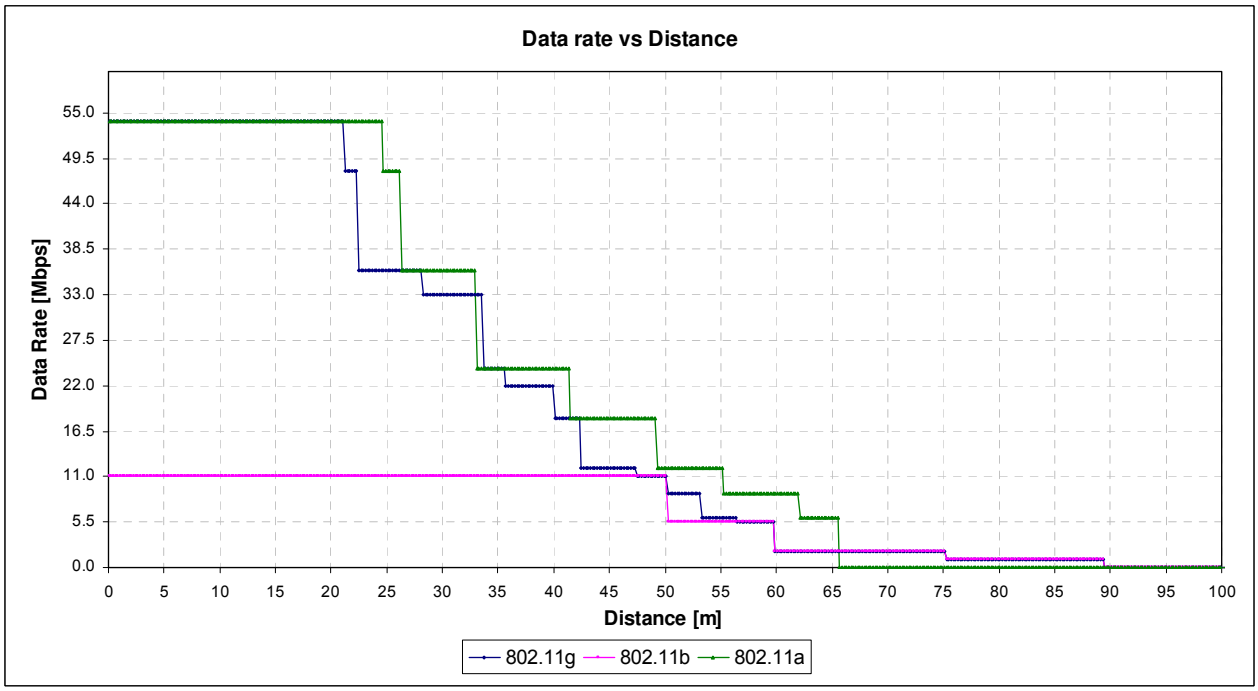


Figure C.15: Data rate vs. distance for IEEE WLAN 802.11a/b/g (without shadowing).