1 Introduction

1.1 Historical perspective on radio resource management

As J. C. Maxwell had predicted in the 1850s, wireless transmission of electrical energy was feasible. Several decades later Heinrich Hertz managed to experimentally verify Maxwell's daring ideas with his award-winning experiment in 1888. He was able to demonstrate that his 600 MHz transmitter was capable of producing a spark in his simple receiver a few meters away in his laboratory. Although several scientists and inventors would like to claim the fame of inventing radio as we know it today, it took an engineer to bring this groundbreaking research into practical use. The Italian pioneer Guglielmo Marconi was the first to make practical and commercial use of the so-called Hertzian waves. After some initial experiments on his father's estate in 1895, his wireless apparatus gradually became a commercial success. It eventually made Marconi the first, but certainly not the last, millionaire in the wireless business. From humble beginnings, transmitting messages a few hundred meters in his first experiments, in 1901 he was finally able to demonstrate wireless communication across the Atlantic Ocean from Poldhu in Cornwall, England, to Newfoundland, Canada. In the decades to follow, wireless communications became an essential technology onboard ships. The early 1920s saw the advent of radio broadcasting, bringing wireless receivers into every home. We know what happened later—wireless has created a deep impact in our daily lives through success stories such as TV broadcasting, worldwide shortwave communication, satellite communications, and in recent decades mobile telephony and wireless and mobile Internet access.

The latest chapter in this story started to be written in the early 1980s with the commercial success of automated mobile telephony and mobile data. Examples of so-called first generation mobile telephone systems are the NMT system in Scandinavia (1981), AMPS in the USA (1984), TACS in the UK (1984) and other systems. These systems were targeting limited markets, terminals were expensive and they never reached very high user penetrations. The first-generation systems were analog designs—only the switching logic relied on digital technology.

It took another decade and the introduction of global standards for digital mobile systems to put a cellular phone in almost every person's hands. These systems basically provide the same service as previous analog systems, but employ advanced digital signal processing to lower the cost of production and to improve the range and the tolerance to interference, allowing more users in the system without compromising the speech

quality. In particular, systems based on the global GSM (Global System for Mobile communications) standard had become immensely popular around the globe. Recent statistics in 2011 showed that there are more than 6 billion mobile subscriptions world wide, corresponding to an 87% penetration globally.

As new wireless systems evolved to complement the second-generation wireless access systems, a distinct shift in design criteria can be noted. From being primarily systems for voice communication, 3G wireless systems appearing on the market in the early 2000s were designed to also deal with data and various multimedia and web-based applications. The globally most popular 3G system, UMTS (Universal Mobile Telecommunications System), is not *per se* very much more efficient than its 2G predecessors but is geared to provide a combination of packet- and circuit-switched low-level bearer services, initially with on-air data rates between 384 kbit/s wide area coverage to 2 Mbps indoor/microcell coverage. Already a few years later high-speed data-only evolutions 3.5G or Turbo 3G were introduced in UMTS providing raw on-air data rates up to 14 Mbps with significantly lower latency, improving in particular the performance of web-originated downlink traffic.

Early 3G systems were based on the concept of generic bearer services, each mapping to some specific class of end-user applications envisaged by the system designers. In recent years it has become more and more obvious that it is impossible to predict what applications will be so popular, so-called killer applications, that we should design the wireless access infrastructure around them. Most 3G systems nowadays provide only two types of bearer services—voice and best-effort packet services—the latter to provide generic IP (Internet Protocol) access. One may say that IP access has become the completely dominant communication service—the killer service—both for fixed and for mobile applications. As a consequence, 4G systems, e.g. Long-Term Evolution (LTE) being deployed currently, are data-only systems providing wireless IP access with high data rates up to 100 Mbps and low latency. In parallel, we have seen the evolution of wireless local area networks, which are aiming to provide similar services but at much higher data rates at short ranges in office environments.

Fueled by the introduction of flat rate tariffing, the commercial success of mobile and wireless access to the Internet has been monumental in recent years. Initially thought of as a way of selling excess capacity in 3G networks, or providing some simple value-added services, it has, together with the proliferation of smartphones, created an explosion of traffic volumes. This trend is already threatening to overrun many networks. It is obvious that new technologies and careful management of resources in terms of cost, energy and radio frequency spectrum is needed to keep up with the pace of these developments. These problems are indeed the main focus of this book.

1.2 Key problems in wireless systems

The designers of wireless systems have struggled with a series of fundamental design bottlenecks, or key problems, each typical of their respective phase of development.

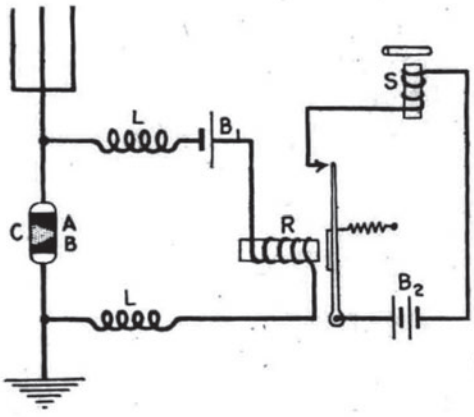

Figure 1.1 Early radio: Marconi-type passive coherer radio receivers. From *Elements of Radiotelegraphy*, E. W. Stone, 1919.

The removal of each bottleneck made the development take a significant leap forward, just to face another bottleneck. Let's briefly review these key problems.

1.2.1 Path loss—the early days

The dawn of wireless communication was dominated by wireless telegraphy, which became widespread in the early years of the 20th century. As electronic amplifiers were yet to be invented, the receivers were passive devices, mainly consisting of a simple tuned circuit, i.e. a bandpass filter tuned to the dominant frequency of the transmissions. As the signals could not be amplified in the receiver, all the energy at the receiver output (e.g. a sound in an earphone or the energy for pulling on a magnet that would operate a pen) had to be generated at the transmitter (see Figure 1.1). The loss of energy over a wireless connection, the path loss, is gigantic, in particular over large distances. This meant that the transmitters were large and bulky, capable of radiating enormous amounts of power. Needless to say, this was a severe limitation to mobile wireless communication, maybe with the exception of use on larger ships.

As the electronic tube amplifier [L. de Forest, 1907] became available, the path loss problem could be solved. Now, as receivers could amplify the weak received signals almost arbitrarily, the path loss could be completely compensated for. Moderately sized transmitters and sound transmission took us into the era of radio broadcasting in the early 1920s. The word "radio" was now synonymous with radio broadcasting to the man on the street. In the western world, a radio receiver appeared in virtually every home. Sound broadcasting was soon followed by TV broadcasting. In the USA this occurred in the 1930s, but elsewhere the commercial success of TV had to wait for the 1950s. Wireless communication also played a pivotal role in the Second World War. Soon after the war two groundbreaking inventions revolutionized wireless communication. The first was the invention of the transistor. Originally intended as a

tube replacement in order to enable low-power, portable radios, the term "transistor" became synonymous with the small pocket broadcast receiver in the 1950s and 1960s.

At this stage, when there was now no limit to the amplification of signals in the receivers, communication engineers became aware of the next bottleneck, the thermal noise.

1.2.2 Thermal noise

Thermal noise is caused by the Brownian dance of electrons and is everywhere. It appears in all materials and electronic components. No matter how much the received signals are amplified, the noise will also be amplified alongside them. The second key discovery at the end of the Second World War was the recognition of the fundamental limits to the amount of information that could be transmitted reliably in the presence of noise.

As Claude E. Shannon published his "A mathematical theory of communication" in 1948, he laid the foundation of digital communications. At the time, the findings were not very practicable as the advent of integrated circuits and digital signal processing (DSP) devices was still decades away. We can now push the performance of today's wireless communication systems very close to the Shannon limits. The most remarkable achievements made possible by this new way of thinking are probably satellite communication and communicating with deep space probes. From a commercial angle, probably the most revolutionary item is the digital cellular telephone. The latter managed to provide acceptable voice quality even in the most adverse environments, in moving cars or indoors. Over the years, digital communication engineers have been quite successful in pushing the performance close to the constraints manifested in the laws of physics and in Shannon's theory. However, as we got more and more cellular phones, another fundamental problem became evident: the limited radio spectrum.

1.2.3 Interference—the limited spectrum

Although Shannon had already noted some aspects of this problem in the study of bandlimited channels, it has become evident that this is not entirely a technical problem. Since there is only one ether, it is clear that extensive and concurrent use of the same natural resource will inevitably lead to conflicts, in this case to unwanted interference between different users. This was obvious to radio listeners in the old days as they tried to receive a radio program on the medium wave AM band at night. Hundreds of radio stations competed for the attention of the listeners creating devastating mutual interference. The signal strength is in most cases sufficiently high that it would be possible to properly receive most of these stations if they were alone. This means that the problem is something different from the struggle against nature, i.e. the thermal noise discussed above. Rather, this problem, as with all resource-sharing problems, also has a social dimension. This dimension was already recognized in the very early years of radio, when the sharing of the frequency spectrum was given an administrative solution.

The International Telecommunication Union (ITU) was formed just after the Second World War specifically to deal with these problems. A technique that has been popular for spectrum resource sharing since the advent of radio communication is frequency multiplexing. The available spectrum is split into narrow frequency bands, mainly because early modulation schemes produced narrowband signals. This was an excellent way to separate different users of the spectrum and to avoid unintended interference. Within the ITU, the countries of the world have collaborated to closely regulate the use of the frequency spectrum. This spectrum hierarchy starts at the ITU where the frequency spectrum from 10 kHz to 200 GHz is meticulously split down into almost 100 bands or allocations. These allocations are in turn assigned to services in a document called the Radio Regulations (RR). Among these services you find fixed, mobile, broadcasting, radar, amateur and other similar uses of the radio spectrum. The frequency allocations are not at this level assigned to owners, nor to any country. Assigning spectrum to individual users is normally done by the National Regulatory Agencies (NRAs) in each of the ITU member countries. Frequencies are let to different users and user groups using various licensing arrangements. Licenses are typically issued for considerable periods of time, to match the technical/economical lifespan of the radio equipment used. The NRAs guarantee (police) that the provisions of the RR are maintained. When it comes to the lower frequency bands with long-distance propagation properties, decisions cannot be taken by individual NRAs, but instead all permissions for new transmitter sites have to be internationally coordinated. In principle, this requires that for every new transmitter, the NRA of that country has to collect the consent of all other countries that could be affected within some reasonable range. Needless to say, as the usage of the spectrum has been increasing, this has become a complicated matter. As new technical developments have meant new requirements on the spectrum, the rather slow administrative process has had difficulties in keeping up.

At the ITU level, making any significant changes in the RR, e.g. to allocate frequency bands to new systems and technologies, has been a demanding task, since a consensus decision between over 170 member states of the ITU has to be reached. Such changes are discussed at the World (Administrative) Radio Conferences (WARC or WRC) which are held every fourth or fifth year. One reason for this is the large differences in wealth and technological development in different parts of the world. Whereas some countries, e.g. western highly developed countries, demand new systems with high capacity and performance, other countries may have the view that the old technology is still viable and that already-made investments in equipment, receivers and so on should be protected. Major changes, if even possible to decide on, may require decades of careful planning and lobbying.

As the 1980s saw the transition from land mobile radio to (automated) mass-market mobile telephony, it became clear that a radically different solution was needed. It is obvious that individual users due to their sheer number cannot be given individual frequency assignments by the NRAs, and even more that the NRAs would be able to protect their reception quality. Instead, in a cellular telephone system the frequency administration is largely handed over to the owner of the system, the operator. The operator is given a license and frequency assignments by the NRA. When designing

a system the operator has to organize the use of the spectrum in such a way that interference between the users of the system is kept to an acceptable level. Cellular telephone systems use a combination of careful planning and automatic schemes that adapt the spectrum utilization to the current user requirements. In the planning stage, the base stations are placed at carefully chosen locations and each base station is assigned a certain set of frequencies. The choice of which base station is to be used for connecting a mobile telephone and which actual frequency channel is to be used is done automatically while the system is in operation.

1.2.4 Infrastructure cost and energy consumption

As we move from mobile telephony to mobile data systems with a several orders-of-magnitude increase in capacity and data rates, we are not able to find enough spectrum to match. In addition, at very high data rates our systems will also again become limited by the thermal noise (basically the second key problem again). To some extent, we can counteract the latter problem with more transmit power at the expense of increased energy consumption and low battery lifetime in mobile devices. Another, more effective, method to increase data rate and capacity is to use more base stations, thereby effectively limiting the range of transmissions. The price we pay for this is more investment in infrastructure and equipment. Balancing the infrastructure cost, the energy consumption and the available spectrum is today the key problem in high-capacity wireless systems. This problem set will be the main focus of this book.

1.3 Wireless access networks—the issues

There are many different types of wireless communication systems, ranging from broadcasting to satellite systems, from shipborne systems to police, fire brigade and military systems. In this book our focus will be on systems for mobile and nomadic (data) network access. The reason for this is that almost all IT services today rely on network access—information is retrieved, stored or processed remotely rather than in the mobile devices themselves. IP connectivity is becoming the dominant design in service provisioning, commercially dwarfing other types of specialized network solutions (e.g. peer-to-peer systems). We communicate with other people in distant locations and most of the apps in our smartphones are becoming cloud-based. Cloud computing is a consequence of efficient and virtually free communication. We compute and store information wherever in the world it may be cheapest and most effective, neglecting the cost of communication. The physical terminal we are using is of no consequence. In mobile access this has not really been the case—poor coverage, high cost, latency and limited data rates have been limitations that have prevented service mobility and convergence. If and when the mobile and wireless networks provide better connectivity and access speeds, cloud computing will inevitably also become the ruling paradigm in wireless. A striking example of this is when two friends meet in the street and would like to exchange digital content, say photos. In principle, short-range

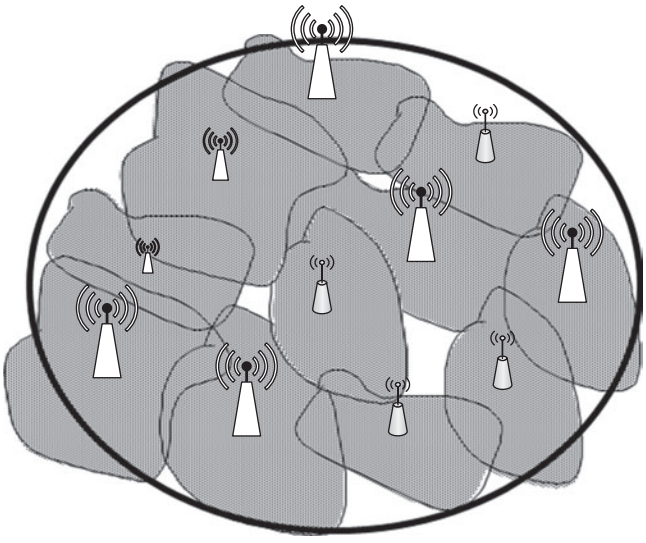

Figure 1.2 Schematic coverage map of a wireless communication system.

peer-to-peer radio connectivity, e.g. Bluetooth, would be the most effective way from an engineering perspective. However, instead of wasting time and effort in peering, you simply email your photo to your friend or put it on Flickr using cellular or WiFi access. The reason for this is that, even though this operation may consume significantly more network resources, the marginal cost for the user is zero.

In a wireless access system, the primary goal is to provide fixed network access to a large number of mobile or stationary users dispersed over a geographical area.

The number of users, their service demand and locations are not *a priori* known. An example from the early days of history is the (national) radio/TV broadcasting systems. In these systems, one-way wireless connections to individual mobile or stationary listeners are provided by a collection of broadcast transmitters connected to a program distribution network. Another, more recent example, which we will cover in somewhat more depth, is a mobile (cellular) telephone or mobile data system. In this example, the fixed infrastructure that the mobile users are attempting to acquire services from is the Public Switched Telephone Network (PSTN) or the Internet. To provide the services of the network, i.e. to connect mobile users to fixed (or other mobile) users in the network or servers, the network is extended by a set of radio base stations. The base stations provide the physical two-way radio connections to the mobile terminals. Figure 1.2 illustrates the principles of wireless network design. The network consists of a fixed network part and a wireless network part. The fixed network provides connections between base stations or Access Points (APs) which in turn provide the wireless connections to the mobiles. The APs are distributed over the geographical area where mobile users are provided with communication services.

This area is called the service area. The mobile terminals that are to be provided with the required service may be anywhere within the service area and will be assigned a connection to some AP. The assignment is done by the system and without any user

intervention. The area around an AP where the transmission conditions are favorable enough to maintain a connection or provide a service of the required quality is termed the coverage area of the AP. The transmission quality (e.g. the voice quality, the data rate, etc.) and thus the shape of these regions will depend significantly on the propagation conditions and the interference from other users in the system.

The coverage areas of the individual APs are, in practice, of highly irregular shape. The areas may contain coverage holes, i.e. there are locations close to an AP that are not covered, e.g. due to shadowing from buildings. It is more common that the system is designed to create overlap areas, i.e. there are areas where a terminal may communicate with several APs. Also the opposite situation may occur, i.e. a situation where the terminal is in a white spot, a region where communication with sufficient quality is not possible. The fraction of the service area that is not affected by white spots is called the coverage or the area availability of the system. These quantities are both defined as the probability that communication is maintained at some given randomly chosen location (chosen from a uniform distribution) in the service area. Another measure of great interest is the population availability, i.e. the probability that a randomly selected user can be provided with adequate communication service. This measure can be calculated by weighting the covered areas with the (user) population density.

In two-way communication systems (such as mobile telephone or mobile IP access systems) links have to be established both from the AP to the mobile (called the downlink or forward link) and between the mobile terminal and the AP (the uplink or reverse link).

The first casual look may suggest that these links have very similar properties. There are, however, distinct differences from a radio propagation perspective. For example, in wide area cellular systems, the AP (base station) usually has its antennas at highly elevated locations, free from obstacles. The terminals, on the other hand, are usually located at street level, where buildings and other obstacles create shadowing and multipath reflections. Also the interference situation in the up- and downlinks will be different since there are many terminals with varying locations and relatively few APs at fixed locations.

For obvious economic reasons, a network owner wants to provide the required service at minimum cost. His/her objective will therefore be to provide sufficient coverage with as few APs as possible. This would not only minimize the cost of installation, towers, radio equipment and other AP hardware, but also minimize the fixed, wired part of the infrastructure. Various propagation effects limit the coverage and will thus put a lower limit on the number of APs that need to be installed. If the distance between two APs becomes too large, eventually there will be points between the APs where the signal level will drop too low, which will in turn result in poor voice quality or low data rate. Shadowing and multipath phenomena will add to these problems. Stated simply, the transmission range of the APs is too small compared to the inter-AP distance. Such a system where this type of problem is dominant is termed a range-limited system. Typically, mobile cellular systems are range-limited systems in their initial stages of development when the key objective is to quickly and at a low cost cover the service

area for a low number of subscribers. Other examples are early broadcasting systems, where radio stations were few compared to the bandwidth available.

As systems evolve and become popular, cellular and broadcasting systems alike, the number of transmitters in the system eventually becomes large compared to the available bandwidth. These system are not primarily troubled by weak received signals, but interference from other APs and mobile terminals. Such systems are said to be bandwidth or interference limited. The main problem in these systems is the proper management of the scarce resources, e.g. bandwidth. The objective of this management task is to satisfy both the provider of services (the operator) and the user of these communication services. The former wants a high and efficient utilization of the system since he/she derives more revenues by providing higher data rates or services to more users, or he/she is capable of providing a given service with less resource consumption (less power, spectrum or APs). The user expects good Quality of Service (QoS). In cellular systems, such user requirements can be expressed in terms of probabilistic measures such as the average data rate, the probability of being denied making a voice call with acceptable quality (blocking) or when dropping an ongoing connection. In the following chapters, we will demonstrate that as in most resource management problems, the operators aim to increase the data rate or the number of users served. Such quantities we will loosely refer to as the capacity of the system. These objectives are in conflict with the users' desire to achieve a higher service quality. Squeezing more users into the system will inevitably cause more interference resulting in poorer transmission quality, lower data rates and/or longer waiting times. Striking the proper balance between these aims is a delicate problem for the operator when he/she makes offers to the users, in particular in a competitive situation. Efficient frequency resource management, i.e. employing schemes that either avoid some of the interference or that better resist the interference between users, can both increase the capacity and improve the service quality in the system.

The frequency spectrum is not the only resource wireless operators and their customers have to be concerned with—there are other scarce resources as well. One obvious such resource is the infrastructure of APs, including networks of switches/routers. It will become clear from our analysis that a denser system with more access ports (i.e. more expensive infrastructure), has the potential of providing more capacity and higher QoS to the users. Another important resource to be managed is the energy consumption in the system. Since most modern wireless networks are designed for lightweight and portable use, the battery energy is severely limited. Moreover, the biomedical restrictions on emitting electromagnetic fields from handheld devices also impose limits on the transmitter power. Limitations in available energy at the portable terminal may also lead to restrictions in the complexity of the signal processing algorithms employed at the terminal. In all these cases, lower transmitter power leads to either lower transmission quality or lower radio range. Each of these effects has to be countered by adding more access ports (i.e. a more expensive infrastructure). On the other hand, in mobile data systems with high data rate the cost of energy in the APs has also become a significant concern. In a similar way to trading off power requirements and infrastructure density, it will be seen that frequency spectrum bandwidth and power

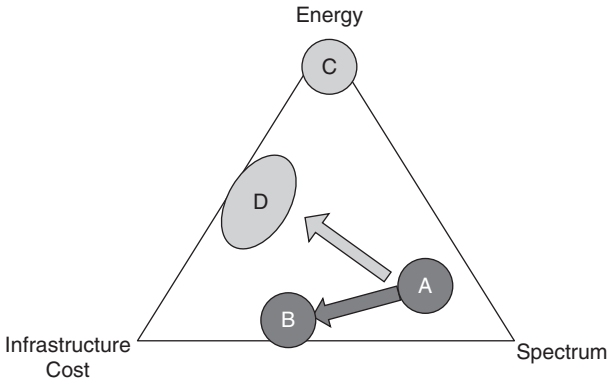

Figure 1.3 Trade-off of resources in wireless networks.

can be traded off. Figure 1.3 illustrates this interdependency. It illustrates how the traditional design of mobile communication systems has been spectrum limited (A). This situation is studied in Chapter 4. As systems require more capacity, wireless transmitters are packed closer and closer and the infrastructure cost (Chapter 11) starts to limit the design (B). As the cost for energy goes up, we have to be aware of this element as well. Completely energy-minimizing systems (C) we will see are not really realistic but a reasonable compromise has to be struck (D; Chapter 9).

1.4 Outline of the book

This book is intended as a textbook for an advanced course in wireless networks. We will assume that the students know the fundamentals of radio propagation and digital communications over wireless channels. Chapter 2 provides a more stringent definition of those models and performance metrics that were introduced in a more hand-waving fashion in Section 1.3 above. We also outline the basic methodology used to analyze wireless access networks. After this modeling introduction, the book is then basically divided into two parts.

Part I, Radio Resource Management (RRM) in wireless systems, discusses various techniques to manage the interference and to maximize the capacity in an existing (already deployed) wireless access network, first from an orthogonal access perspective and then a non-orthogonal access perspective. Chapter 3 discusses various medium access schemes and Chapter 4 various scheduling approaches. In both chapters, we assume orthogonal access and that no simultaneous transmissions are allowed on the same radio resources because of heavy interference between different users. Chapters 6 to 8 then discuss more advanced RRM techniques such as power control, interference management, handover, and other inter-cell interference management techniques. Chapter 11 illustrates how various RRM techniques are applied in 4G Long Term Evolution (LTE) systems.