

Chapter 1

MOBILITY IN WIRELESS COMMUNICATION NETWORKS

RONAN SKEHILL*, PADRAIG SCULLY, EDUARDO CANO,
JOSEPH JOHNSON, SEAN MCGRATH and JOHN NELSON

*Wireless Access Research Centre,
Department of Electronic and Computer Engineering,
University of Limerick, Ireland*

**ronan.skehill@ul.ie*

Underlying all wireless technologies is the fundamental principle that mobile users communicate with a base station via a wireless channel. In this chapter, a basic wireless channel is explored and the effects of distance, and thereby mobility, on the wireless channel is shown. The chapter then presents a generic wireless communication network and identifies the role and function of each component i.e., mobile device, access network and core network. The merger of new technologies and data services with mobile networks has created new issues in network management. The chapter provides a tutorial on the various mobility management methods used to mitigate and solve mobility issues. Concepts such as personal mobility, session mobility and service mobility can be viewed as high level mobility management solutions while concepts such as ad hoc mobility, mode mobility and terminal mobility are viewed as low level solutions. The chapter explores these concepts and presents current protocols/methods. A guide is provided for practitioners who may implement some of the mobility management methods. The chapter highlights research directions by presenting a range of projects and standard bodies working in the mobility management field.

1. Introduction

In every wireless network a transmitter and a receiver node is needed for communication. Radio waves start from a certain point, typically the transmitter antenna, and with a certain level of power or energy. Assuming the antenna is omnidirectional, i.e., it radiates power equally in all directions, and assuming perfect conditions then radio waves continue in an ever-expanding circle. However, since the energy is finite, at any distance from the transmitter the original energy is now spread over a much larger space. The signal is said to be attenuated. If the signal is too weak, the receiver node cannot correctly receive the transmitted information. If this happens, the receiver lies outside the coverage area of the transmitter. The fundamental issue behind the coverage problem is the phenomenon of fading. Fading can be considered to be the time variation of the channel strengths due to multipath

fading, as well as path loss via distance attenuation. Because receiver nodes can be mobile, the possibility of this happening increases greatly and methods to prevent out-of-coverage issues must be implemented.

A significant disadvantage of wireless networks is interference. Tse *et al.* [1] illustrate that a transmitter–receiver pair can often be thought of as an isolated point-to-point link, and that wireless users communicate over the air and there is significant interference between them. The interference can be between transmitters communicating with a common receiver (e.g., uplink of a cellular system), between signals from a single transmitter to multiple receivers (e.g., downlink of a cellular system), or between different transmitter-receiver pairs (e.g., interference between users in different cells). If the effects of interference between the isolated point-to-point link is not mitigated the signal could be completely lost.

As highlighted with fading and interference, wireless networks have some disadvantages compared to a wired network. However, the advantages of wireless networks surpass the disadvantages. The key advantage is mobility, i.e., the ability of the transmitter–receiver pair to move around and still avail of wired network functionality and services while in the coverage of the wireless network. Planning and network management techniques for fading and interference can overcome problems like coverage. Cell coverage varies as it is dependent on the wireless access technology in use. Technologies like GSM and UMTS can have large cell coverage, up to a radius of 2 km, while wireless technologies like IEEE 802.11 typically have a cell radius of less than a 100 m. Cellular systems like GSM and UMTS function on the assumption that a base station controls a limited coverage area, termed a cell. Cells can be clustered and controlled by a collection of base stations. This cluster represents a service area in which mobile users have the freedom to move/roam and gain network access simultaneously.

Movement in this service area has brought new challenges in wireless communication at all levels of the network. If relating a wireless network to the Open System Interconnected (OSI) stack, mobility primarily affects the physical

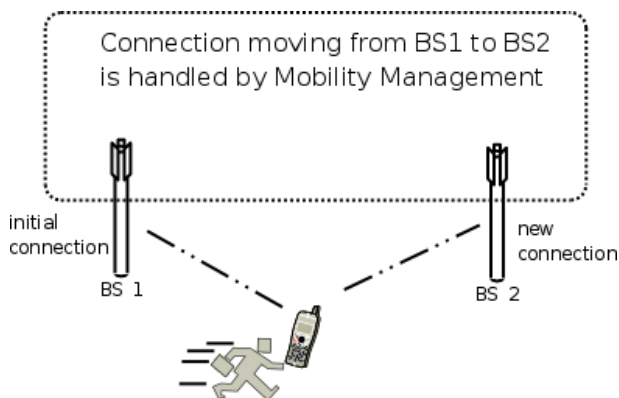


Fig. 1. Mobility management handles the connection from one base station to another.

and media access layers. Handling mobility at this level requires the network to handle mobile devices at a radio level. However, current wireless networks allow applications, including realtime and non-realtime, to run over Internet Protocol (IP) which brings issues in maintaining TCP/IP connections while the user moves from base station to base station. The Transmission Control Protocol (TCP) relies on the source and destination IP address to create a checksum; if either IP address change (which can happen if changing base station, network, etc.) then the checksum fails. If the checksum fails then the TCP connection is broken. Resolving this issue in a wireless network can be achieved by providing mobility management at the IP layer and above.

Underlying all these management techniques are the mobility characteristics of the mobile device in the wireless network. It is the movement of the mobile device that causes the management technique to be executed. Therefore, it is imperative that the movement is correctly understood. In a real-world scenario movement occurs at a multitude of levels: global movement, intercontinental, intercity, and on a smaller scale — campus, street or office. Furthermore, mobility can occur in groups or individually. Naturally, there is a speed associated with each of the movement levels; for example, in an office the speed of a mobile device is limited to walking speeds — whereas with a vehicle on a road the speed of the mobile device is now limited by road speed conditions. Associated with mobile device speed is mobile device density (measured in devices/m²). In low speed mobility areas the mobile device density can be quite high, and with high speed mobility (cars traveling at 100 km/h) the mobile device density is low. Different wireless technologies are best suited for different mobility characteristics. For example cellular networks like GSM or UMTS are flexible and can cover high speed movements of a car on a road and pedestrians in a street, whereas IEEE 802.11 wireless LAN (WLAN) is best suited for low mobility scenarios only. Both technologies provide mobility support but ultimately the trade-off between GSM/UMTS and WLAN is data rates. WLAN can achieve data rates of over 100 Mb/s whereas GSM/UMTS networks can provide data rates up to 2 Mb/s.

Mobility management techniques must therefore be able to accommodate all these wireless technologies and be able to handle mobility speeds in all geographical areas. A single solution is not feasible so a multi-task solution is currently used. Before describing mobility management solutions, the wireless channel is explained to understand how conventional terminal mobility management works. A reference model of a generic future wireless communication network is presented that supports a multi-task mobility management technique.

2. Wireless Channel

The wireless channel is the physical medium that supports the propagation of all signals in a wireless network. In contrast with wired communications systems, the transmitted signals often do not reach their destination directly due to the presence

of obstructions in the line-of-sight path. Obstacles in a communications link lead to the occurrence at the receiver of different replicas of the transmitted signal. This phenomenon, known as multipath propagation, weakens, delays and distorts the transmitted signal in a time-varying fashion. In addition to multipath propagation, the motion between the transmitter and the receiver also has a negative effect on the received signal strength. Thus, the characterization and understanding of the mobile radio channel is essential to combat these distortion effects and to acquire a desirable system performance, which in turn can satisfy the required quality of service (QoS).

An accurate estimation of multipath channel behavior allows for improving the bit error rate using equalization techniques at the receiver, optimization of the maximum capacity, and effectiveness of the shared medium access mechanisms. Furthermore, the design of a mobile radio system that satisfies the stipulated QoS metrics under realistic wireless channel conditions constitutes a major challenge, especially when service requirements demand very high data rates coupled with high-speed mobility.

2.1. Propagation Mechanisms

The communications path between transmitter and receiver is time, space and frequency variant. This variability of the wireless channel causes loss of signal strength at the receiver device referred to as *fading*. A good understanding of the propagation media and the circumstances that cause the fading of received signals is required for a complete characterization of the wireless channel. The variations of the channel can be roughly classified into (i) large-scale fading and (ii) small-scale fading, depending on how fast the average received power fluctuates in the distance between a receiver and a transmitter.

2.1.1. Large-Scale Fading Effects

Large-scale fading is predominantly caused by the path loss phenomenon. Large scale fading manifests itself as fluctuations in the voltage envelope of the received signal over relatively large distances between the transmitter and the receiver. It is common to model large scale fading effects using a path loss model that obeys some sort of power law.

Large scale fading represents the average signal power attenuation or the path loss caused by the motion of the mobile device over large areas (usually of the order of the wireless communication network cell size). This phenomenon occurs when the receiver is *shadowed* by the presence of prominent terrain contours (e.g., buildings, trees, hills, etc.).

Different approaches for modeling the path loss may be considered depending on the propagation environment. The simplest model corresponds to the direct line-of-sight transmission between transmitter and receiver, referred to as *free space loss* [2]. The principle of energy conservation establishes that the integral of the

power density over any closed surface surrounding the transmit antenna must be equal to the transmit power. As the integration area is a circle of radius d centered at the transmitter position, the received power collected by the receiver antenna is obtained according to the Friis transmission equation [2] as:

$$P_{Rx}(d) = P_{Tx} G_{Tx} \frac{1}{4\pi d^2} A_{Rx} = \frac{P_{Tx} G_{Tx} G_{Rx}}{L_{FS}}, \quad (1)$$

where P_{Tx} is the transmitter power, L_{FS} is the free-space path loss and G_{Tx} is the transmit antenna gain in the direction of the receive antenna. The parameter A_{Rx} is the effective area of the receiver, which follows the relationship $A_{Rx} = \lambda^2/4\pi G_{Rx}$, with G_{Rx} being the gain of the receive antenna and $\lambda = c/f_c$ being the wavelength with c the speed of light and f_c the carrier frequency.

More realistic models take into account the effects caused by *reflections*, *scattering* and *diffraction*. A simple model that considers the reflective effect of the earth's surface is named *plane earth loss* [3]. The received power is computed for the situation where only line-of-sight ray and ground-reflected wave exists, obtaining

$$P_{Rx}(d) = \frac{P_{Tx} G_{Tx} G_{Rx}}{L_{PE}} \approx P_{Tx} G_{Tx} G_{Rx} \left(\frac{h_{Tx} h_{Rx}}{d^2} \right)^2, \quad (2)$$

where h_{Tx} and h_{Rx} are the height of the transmitter (mobile device) antenna and the receiver (base station) antenna, respectively. This path loss model is valid for distances larger than $d \geq 4h_{Tx}h_{Rx}/\lambda$. In this model, the received power decays with a factor d^{-4} , resulting in a very inaccurate measure for different types of propagation media (rural, office, etc.). The frequency dependence of the path loss is also not taken into account in this model. Thus, a more accurate model, *Egli path loss* [4], is typically employed if replacing L_{PE} by $L_{EG} = L_{PE} \cdot (40/f[\text{MHz}])^2$.

More approximate models have been developed after carrying out empirical measurements. Empirical analyses confirm that the path loss can suffer significant variations in certain conditions as a function of the range, frequency and antenna heights. The empirical path loss is modeled as

$$L_E[\text{dB}] = A - B \log_{10}(d[\text{km}]) - C, \quad (3)$$

where $A = f(h_{Tx}, h_{Rx}, f_c)$, $B = f(h_{Rx})$ and $C = f(f_c)$. The most commonly employed empirical model for urban areas is the Okumura-Hata [5].

2.1.2. Small-Scale Propagation Effects

Small-scale fading describes the rapid fluctuation in signal amplitude and phase that occurs in the received signal over a small period of time or a short traveled distance (on the order of a wavelength). These types of fluctuations are manifested in two different manners, time-spreading of the signal due to *multipath propagation* and time-variant behavior of the channel caused by the *Doppler effect*.

Multipath fading is caused by the presence of multiple replicas of the transmitted signal that occur at the receiver due to reflections against obstacles. Multiple copies of the transmitted signal sum together in either a constructive or destructive manner depending on the phase of each partial wave. These additions create fading dips in the received signal power and distort the frequency response characteristics of the transmitted signal. These distortions are linear and need to be combated at the receiver by applying equalization and diversity techniques.

The multipath wireless channels can be initially classified according to their coherence bandwidth. The coherence bandwidth represents the range of frequencies over which the channel response is flat, meaning that all the frequency components are equally attenuated. *Flat fading* occurs when the signal bandwidth is less than the channel coherence bandwidth. In this situation, the presence of deep fades due to multipath can result in deep fades in the signal over all its bandwidth. On the other hand, when the signal bandwidth is greater than the channel coherence bandwidth, the multipath fading is considered to be *frequency selective*. That is, multipath fades are unlikely to affect the signal over its entire bandwidth without experiencing fades in received energy.

The *Doppler effect* represents the variations of the observed frequency (or wavelength) of a received wave caused by the motion of the mobile device during the communications process. The Doppler effect leads to a frequency spreading of the wave. The Doppler frequency of the n th incident wave is determined by the angle of arrival α_n , which is defined by the direction of arrival of the n th incident wave and the direction of the motion of the mobile device. This Doppler frequency follows the relationship:

$$f_n = \frac{v_{Tx}}{c} f_c \cos(\alpha_n), \quad (4)$$

where v_{Tx} is the velocity of the transmitter mobile device relative to the receiver. The spectrum of the transmitted signal undergoes a frequency expansion during transmission due to Doppler effects. If the information bandwidth is smaller than the spectral broadening, given by the *Doppler spread*, the channel is *slow fading*. On the other hand, if the Doppler spread is larger than the information bandwidth, the channel is *fast fading*.

2.2. Wireless Channel Models

The characterization of the wireless channel requires the extraction of the main characteristics of channel sounding campaigns in order to express them in a mathematical manner. The *channel impulse response* models analytically the behavior of the channel. The relationship between the received signal $r_s(t)$, the transmitted signal $s(t)$ and the channel impulse response $h(t)$ is expressed as

$$r_s(\tau) = s(\tau) \otimes h(\tau) + n(\tau) \stackrel{FT\{r_s(t)\}}{\iff} R(f) = S(f)H(f) + N(f), \quad (5)$$

where \otimes represents the convolution operation, τ is the time variable, and $FT\{x\}$ is the Fourier transform of x . The function $n(t)$ is a thermal noise signal modeled as a zero-mean additive white Gaussian noise (AWGN) with constant power spectral density value $N_0/2$. The following channel models are classified and described in terms of their channel impulse responses.

2.2.1. Additive Gaussian White Noise Channels

The additive Gaussian white noise (AWGN) channel is the simplest model and mathematically describes the free-space channel without considering fading, interference, nonlinearity or dispersion phenomena. The channel impulse response is given by

$$h(t) = \sqrt{L_{FS}}\delta_D(\tau - \tau_d), \quad (6)$$

where $\delta_D(\tau)$ is the delta function and τ_d is the time taken by the transmitted wave to arrive at the input of the receiver.

2.2.2. Narrowband Channels

Narrowband refers to a wireless communications where the bandwidth of the transmitted signal does not significantly exceed the *coherence bandwidth* of the channel. The coherence bandwidth is a statistical measurement of the channel that accounts for the frequency interval over which two frequencies of a signal are likely to experience comparable or correlated amplitude fading. Thus, narrowband transmission uses radio signals that see flat fading. The channel impulse response of the narrowband channel can be expressed as

$$h(\tau, t) = \sqrt{L}A(t)\delta_D(\tau - \tau_d), \quad (7)$$

where $A(t)$ is the channel amplitude that varies with respect to the time t over a small area. This attenuation is typically modeled as a complex Gaussian random variable and, therefore, with an absolute value that follows a Rayleigh distribution [6].

2.2.3. Wideband Channels

In wideband communications systems the channel is frequency-selective since the channel's coherence bandwidth is larger than the information bandwidth. Therefore, multipath propagation effects must be taken into account in the channel impulse response, which is described as

$$h(\tau, t) = \sqrt{L} \sum_{i=0}^{N-1} A_i(t)\delta_D(\tau - \tau_i(t) - \tau_d), \quad (8)$$

where N is the total number of multipath components. The multipath amplitudes $A_i(t)$ and multipath delays $\tau_i(t)$ constitute the main parameters of the linear time-varying system.

The influence of the multipath fading on the wireless channels can be characterized by analyzing the main parameters of the *power delay profile*. The power delay profile illustrates the amount of energy collected at the receiver and the delays associated with this reception. The main channel parameters are the first moment of the power delay profile (*mean excess delay*), the variance of the power delay profile (*delay spread*) and the *number of significant paths* that accumulate most of the energy of the multipath channel [7]. The number of significant paths (N_P) is a valuable parameter to be estimated for reducing the complexity of multipath-based receiver implementations, such as RAKE receivers [8].

Several types of wideband channel models can be found in the literature. The most relevant models are the two-path fading channel [9], the Rummmler channel [10] and the Saleh–Valenzuela channel [11].

Example: Bit error rate.

The bit error rate (BER) function is evaluated for two different channel scenarios in order to illustrate the effects of the wireless channel on the received information. The first environment considers the transmission of BPSK modulated signals over an AWGN channel. In the other scenario, the same modulated signal is now transmitted through a Rayleigh narrowband channel. The theoretical BER functions are computed in Ref. 12 and expressed as

$$\begin{aligned} BER_{BPSK-AWGN} &= 0.5 \operatorname{erfc}(E_b/N_0), \\ BER_{BPSK-Ray} &= 0.5 \left(1 - \frac{1}{\sqrt{1 + (E_b/N_0)^{-1}}} \right), \end{aligned} \quad (9)$$

where $\operatorname{erfc}(\cdot)$ is the complementary error function and E_b is the energy per bit. Figure 2 shows the BER performance degradation of the Rayleigh channel in comparison with that of the AWGN.

3. Background: Wireless Communication Networks

As wireless technologies evolve, the demand for better and faster communication systems also increases. Since its introduction in the 1980s, the growth of the business of commercial cellular systems has been rapid. The number of wireless communication users as well as the spectrum of the available services has increased at an unexpected rate. The main reason for this growth was the newly introduced notion of terminal, service and user mobility.

As technological enhancements in modern communication networks strive to facilitate ubiquitous connectivity, the future will be increasingly wireless. To date deployment of wireless networks such as GSM, UMTS and IEEE 802.11 has risen in

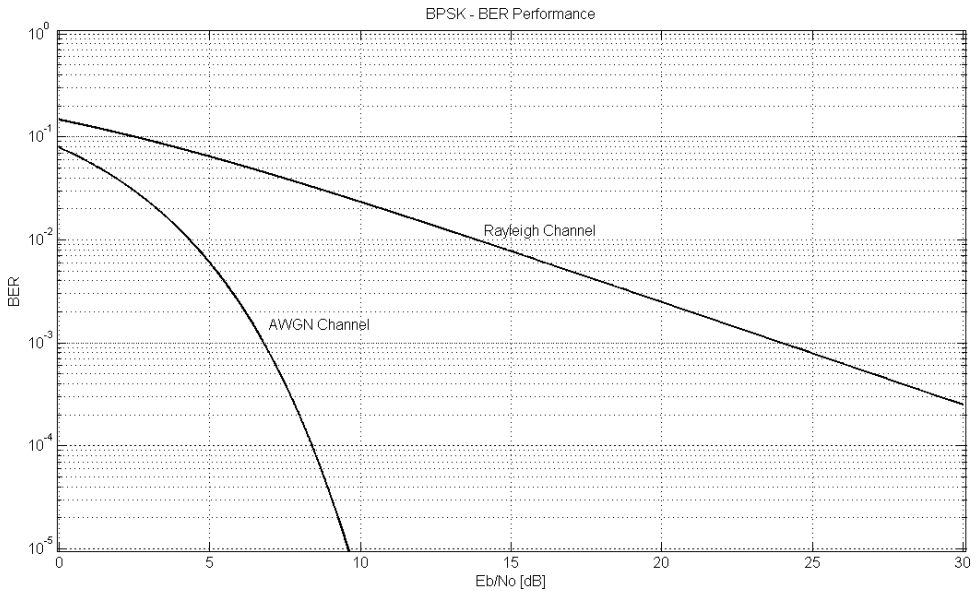


Fig. 2. BER performance for BPSK signals over AWGN and Rayleigh channels.

an effort to provide the desired pervasive access. Wireless technologies must continue to evolve in order to meet increasing demand for faster communication systems and better service availability anytime, anyplace, anywhere. It is expected that wireless communications will become the dominant mode of data access technology in the near future. With that new research challenges are arising — what services the user requires, how to provide the services effectively in the network, and how to manage the network.

Fundamental to all wireless communication systems is the mobility of the mobile device. The movement is the determinant factor in the systems' signaling, traffic and mobility management scheme. Before investigating mobility management techniques an example wireless network is illustrated in terms of architecture, cell network planning, quality of service and supported services.

3.1. Architecture

Existing terrestrial wireless communication networks are based on the cellular concept as outlined by MacDonald [13]. The network structure is composed of a fixed core network with wireless last hops between radio station, and mobile devices. Different cell size environments such as macrocell, microcell and picocell are grouped together as they are dependent on the wireless access technologies. The wireless communication network cell architecture is comprised of these environments tiled and layered to create a tiered cell structure. In a wireless communication service network, as illustrated by Fig. 3, three fundamental hardware

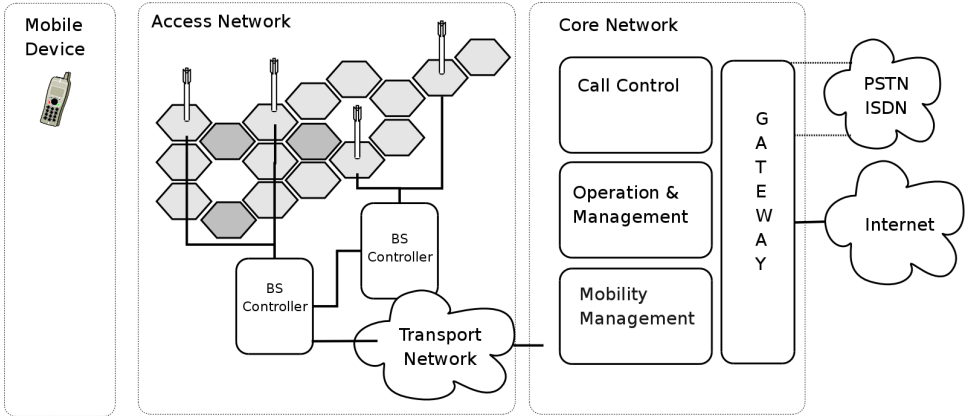


Fig. 3. Wireless communication network.

entities exist: (i) core network system, (ii) access network system and (iii) the mobile device.

The **core network system** or simply the core network (CN) is responsible for maintaining information (such as identity, location, authentication, and billing) about the users in the network. This information is stored in registers or a database. The core network provides access to other backbone networks — Internet, circuit switched, etc. When a call arrives for the mobile device, the CN interrogates its registers and processes the command request.

The **access network system** comprises base stations (BS) and base station controllers (BSC). The BS is the last fixed wired link in the wireless communication network, that is, the BS provides a radio link to the mobile device. The geographical area the BS is in provides radio coverage of an area called a cell. The size of the cell is dependent on the power of the BS and the height of the antenna. A BSC has control over several BSs and thus clusters a group of cells together; this cluster of cells forms an area known as the location area. Combining all the location areas together provides the network's service area.

The **mobile device** is carried by the user. In wireless communication networks these are small, lightweight battery powered devices. Since the mobile device is a battery powered device, the need to prolong battery life is essential. To conserve power the mobile device has two states, *active* and *idle*. When in the active state, the mobile device is actively sending and receiving information or partaking in a call. While in the active mode, its exact cell location is known. While the mobile device is in an idle state, it is passively connected to the network. Depending on the mobility management scheme used in the network, the exact location may not be known, instead the general location (that is, the location area) is known.

In essence the mobile device is any device used directly by an end user to communicate. It can be a hand-held telephone, a card in a laptop computer, or other

device. It connects to the base station via a wireless link and can communicate with the core network system once the link is in place.

3.2. Cell Planning

When planning a location/service area for a wireless network, three main approaches exist: equilateral triangles, mesh (square or rectangle), and hexagonal coverage. Although propagation considerations recommend the circle as a cell shape, the circle is impractical for design purposes. An array of circular cells produces ambiguous areas, which are contained either in no cell or in multiple cells. On the other hand, any regular polygon approximates the shape of a circle and three types, the equilateral triangle, the square and the regular hexagon, can cover a plane with no gaps or overlaps. A service area could be designed with square or equilateral triangles, but for economical reasons the regular hexagonal shape has been adopted. The economical motivation for choosing the hexagon is as follows: Assume a base station located at the center of each cell, the center being the unique point equidistant from the vertices. The vertices are in fact the worst case points since they lie at the greatest distance from the nearest base station. Restricting the distance between the cell center and any vertex to a certain maximum value helps to assure satisfactory transmission quality at the worst case points. If an equilateral triangle, a square, and a regular hexagon all have the same center-to-vertex distance, the hexagon has a substantially larger area. Consequently, to serve a given total coverage area, a hexagonal layout requires fewer cells, therefore fewer transmitter sites [13, 14]. It is worth noting that in real wireless networks, the shape and size of the cell does vary, but for analytical modeling only homogeneous cells sizes are considered. Furthermore, in real wireless networks rectangular and sectorized cell layout is common in urban environments. However, in the scope of this chapter only hexagonal shaped homogeneous cells are considered.

Example: Available channels and cell planning.

Correct cell planning allows for a radio frequency to be reused in a different area for a completely different transmission. Second generation networks like GSM require two communication links — one link for the mobile device to communicate with the base station (uplink) and another link for the base station to communicate with the mobile device.

Consider we have a total of S duplex channels available for use with:

- k channels available per cell, and
- N total cells.

The total number of radio channels may be expressed as: $S = k \times N$.

If these N cells collectively use all the available frequencies then these group of cells are called a cluster.

If a cluster is repeated M times throughout the system then the system capacity may be roughly calculated as:

$$C = M \times k \times N = M \times S.$$

Let us assume we have 30 MHz of bandwidth.

Each voice call requires two 30 kHz channels for uplink and downlink communication, i.e., a 60 kHz duplex channel is needed. Therefore, 500 channels are available in the system. This is calculated from: 30 MHz/60 kHz. If we say that the cluster size is 7 ($N = 7$), then the total number of channels per cell = 500/7 = 71 approximately.

If we consider that a hexagonal cell layout ensures that each cell has six equidistant neighbors, how can we reuse a duplex channel (f_1) with causing interference to neighboring cells in the cluster?

The hexagon is an ideal choice for cellular coverage areas because it closely approximates a circle and offers a wide range of tessellating reuse cluster sizes. A cluster size of N can be constructed using:

$$N = i^2 + i \times j + j^2,$$

where i and j a positive integers.

One can find the nearest co-channel by

- moving i cells along any chain of hexagons,
- turning 60 degrees counter clockwise,
- moving j cells.

For example consider: $N = 7, i = 1, j = 2$.

The reuse distance is an important factor with this cluster size. For example consider D to be the reuse distance and r the cell radius. Then, the relationship

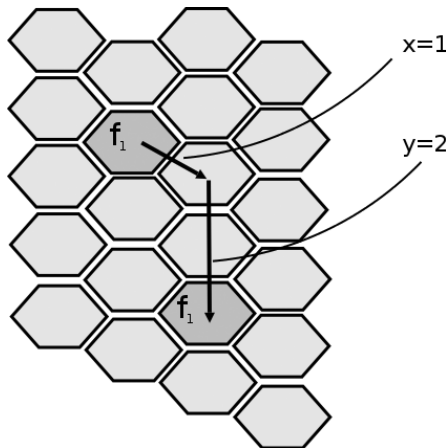


Fig. 4. Tessllated cluster.

between D and r is

$$\frac{D}{r} = \sqrt{3N}.$$

This is important when the cell size is small and when planning network topology, base station layout, etc.

3.3. Quality of Service

The fundamental purpose of all communication networks is to provide a connection between end users. As networks have evolved, a multitude of services have been offered to users in the form of voice, SMS, video conferencing, email and web browsing. The success or failure of that network is dependent on how the user perceives performance of the service on offer. For a user to be satisfied with a network, a number of criteria must be fulfilled. The user should be able to connect easily, the call or service should last for the intended duration without prematurely ending, and with respect to interactive data services (web browsing) download rate is a significant factor. QoS then is a measure of the perceived quality of the service being used and an indication of user satisfaction. Rather than being a single measure of a user's satisfaction, QoS is the collective term for the quantifying of network performance from an end user point of view, and for the measures taken to ensure reliable service.

To ensure that networks are performing at the desired level, one or more measurements of desired performance and priorities of a communications system are considered as part of QoS. QoS measures may include service availability, maximum bit error rate (BER), frame error rate (FER), minimum committed bit rate (CBR) and other measurements that are used to ensure quality communications service.

In heterogeneous environments, i.e., future wireless communication networks, the ability to offer end-to-end QoS while making the different radio access networks interoperate with each other is an important issue. QoS management mechanisms deal with such aspects to ensure services meet QoS requirements.

3.4. Supported Services

A wireless network should be designed with a diverse service environment in mind and similar to those offered on fixed networks. To realize a true mobile multimedia platform, emphasis is placed on the delivery of these services with the specification of service classes. A wireless network should have four service classes:

Conversational: The conversational class specifies realtime traffic, such as voice or voice/video conferencing. These services are bi-directional and traffic is symmetric on this link. These services have very strict timing constraints in order to preserve the integrity of the traffic, both in terms of delay and packet delivery. Thus, this class has the highest priority value associated with it.

Streaming: The streaming class represents video/audio streaming services. Streaming information is highly asymmetric traffic, as data is continuous in one direction. Streaming services are high priority services in terms of packet loss and preserving data integrity, but data is buffered so realtime delivery is not critical.

Interactive: The interactive class type refers to events such as web browsing, database queries and FTP, i.e., information being requested by the user from a remote host. Interactive traffic is asymmetric in nature because data is transferred mainly in one direction. Interactive traffic is not subject to strict timing requirements, but traffic is affected by round-trip delay and response time constraints. As with other classes, transmission errors are not tolerated. However, due to the more relaxed delay restrictions, the focus can be placed on packet delivery.

Background: Background traffic is generated by email and other applications without time constraints. The information is downloaded and assembled in the background without high priority. However, maintaining the integrity of the data is important.

Examples of traffic class applications and associated QoS characteristics (data rates, delay and FER) are presented in Table 1.

Table 1. Traffic class and associated QoS.

Traffic class	Fundamental characteristics	Examples of service	Expected data rates	Delay	Loss (FER)
Conversational	Preserve time variation between information entities of the stream.	Voice	4–25 Kbps	<150 ms	<3%
		Video telephony	32–384 Kbps	<150 ms	<1%
	Conversational pattern (stringent and low delay).	Video games	n/a	<250 ms	zero
Streaming	Preserve time variation between information entities of the stream.	Streaming audio	32–128 Kbps	<10 s	<1%
		Streaming video	32–128 Kbps	<10 s	<1%
Interactive	Request/ response pattern. Preserve payload content.	Web browsing	4–16 Kbps	<4 s	<3%
Background	Destination is not expecting the data within a certain time. Preserve payload.	Download of emails	4–16 Kbps	Can tolerate very high delay time	zero

4. Mobility Management

Traditionally mobility management contains two distinct but related components: location management and handover management. These tasks allow the network to track a mobile device in the wireless communication network. However, as wireless communication networks evolve towards incorporating new IP services, a multi-tier mobility management scheme is needed. The mobility of users is generally tracked through paging/registration procedures often referred to as terminal mobility. However, with increasing user requirements and QoS becoming a prerequisite in wireless communication networks, mobility management has evolved over time from terminal mobility to incorporate personal mobility, service mobility, session mobility, ad hoc and mode mobility.

These mobility management tasks can be divided into distinct levels of network tasks; *high level* and *low level* tasks. Generally low level tasks are associated with layer 1 and 2 of the OSI model, that is, the physical and media access layers. High layer tasks are executed in layer 3 and above, that is, the network, transmission, session, presentation and application layers.

4.1. Low Level Tasks

4.1.1. Terminal Mobility

Terminal mobility refers to (i) the ability of a mobile device to use telecommunication services from any location, while in motion, and (ii) the ability of the network to locate and identify the mobile device as it moves. Depending on the device design, part of the mobile device functions may reside on a removable and portable device or a smart card. According to Ref. 15, terminal mobility is associated

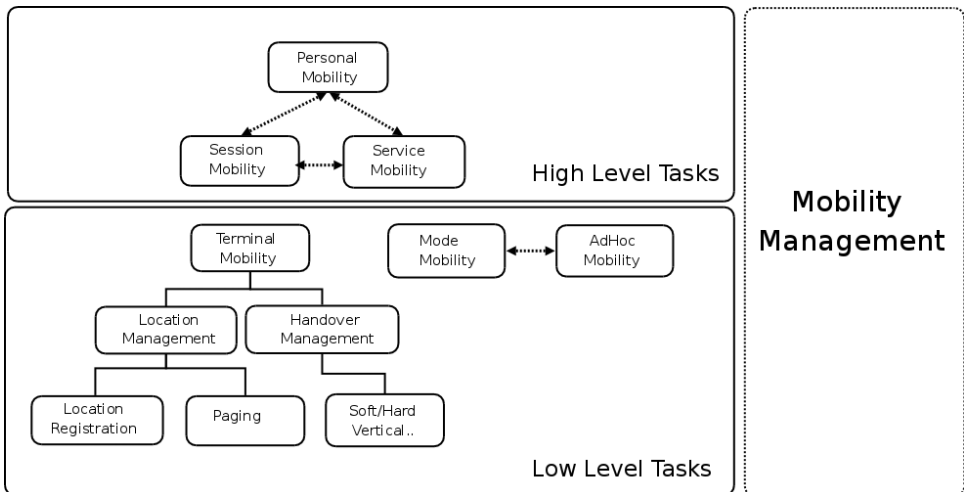


Fig. 5. Mobility management in wireless communication networks can be divided up into high level and low level tasks.

with wireless access and requires the user to carry a device and be within the radio coverage area. The coverage area is directly related to the wireless channel and thereby terminal mobility suffers from the effects of propagation and noise.

From the network point of view, terminal mobility management consists of (i) location management and (ii) handover management. These techniques correspond to the state of the mobile device. For example, with handover management, the mobile device is considered to be in an active state, the user may be partaking in a call or data session and is actively using a wireless channel. With location management, the mobile device is usually in an idle (power saving) mode, waking up periodically to determine its location and whether to update the network of its location. Location management is a twofold process. Firstly, gathering information of a mobile device's location is referred to as the location registration process. A mobile device assists in the process by updating the network of its current location by *location updating* or *registering* at certain times or events. The second process is interrogating the information gathered by the first process (which is stored in the core network) and locating the mobile device on the network. This occurs when a call arrives at the network gateway for the mobile device, a process called paging.

Registration and *paging* processes, collectively known as location management, are therefore concerned with the procedures required by a network to maintain location information about the user, or specifically, for each active user, and setting up incoming calls to non-active users. Non-active mobile devices in wireless networks conserve resources (such as radio or battery) and therefore there is no constant connection with the network. To route a call to the user, the network must find the user's location, that is, the cell in which the user resides. Terminal mobility brings fundamental problems of information management: if the network does not know exactly where the mobile device is, the amount of information stored and processed is lower than if the network knows the exact location of the user. A three-way trade-off consists of: (i) *currency*, i.e., how up-to-date the information is, (ii) *availability*, that is, the information available at all sites/locations or centrally stored, and (iii) the *precision* of the user's location. The extremes of the problem are: no up-to-date knowledge of the user location at any of the sites and up-to-date information of the user's exact cell location available to all the sites.

4.1.1.1. Handover management

Handover management enables the network to maintain a user's connection as the mobile device continues to move and change its access point to the network. According to Akyildiz *et al.* [16] the three-stage process for handover first involves initiation, where either the user, a network agent, or changing network conditions identify the need for handover — this is determined largely by the movement of the user. The second stage is new connection generation, where the network must find new resources for the handover connection and perform any additional routing operations. Under network-controlled handover (NCHO), or mobile-assisted

handover (MAHO), the network generates a new connection, finding new resources for the handover and performing any additional routing operations. For mobile-controlled handover (MCHO), the mobile device finds the new resources and the network approves. The final stage is data-flow control, where the delivery of the data from the old connection path to the new connection path is maintained according to agreed-upon service guarantees.

The handover schemes can be classified according to the way the new channel is set up and the method with which the call is handed off from the old base station to the new one. At call level, there are two classes of handover schemes, namely hard handover and soft handover [17].

In **hard handover**, the old radio link is broken before the new radio link is established and a mobile terminal communicates at most with one base station at a time. The mobile device changes the communication channel to the new base station with the possibility of a short interruption of the call in progress. If the old radio link is disconnected before the network completes the transfer, the call is forced to terminate. Thus, even if idle channels are available in the new cell, a handover call may fail if the network response time for link transfer is too long. Second generation mobile communication systems based on GSM fall into this category.

In **soft handover**, a mobile device may communicate with the network using multiple radio links through different base stations at the same time. The handover process is initiated in the overlapping area between cells some short time before the actual handover takes place. When the new channel is successfully assigned to the mobile device, the old channel is released. Thus, the handover procedure is not sensitive to link transfer time. The second and third generation Code-division multiple access (CDMA) based mobile communication systems lie in this category.

Soft handover decreases call dropping at the expense of additional overhead (two busy channels for a single call) and complexity (transmitting through two channels simultaneously) [18, 19]. Two key issues in designing soft handover schemes are the handover initiation time and the size of the active set of base stations the mobile device is communicating with simultaneously.

Example: Soft handover.

Handover from one cell to the other is an important mechanism in cellular systems. Traditionally, handovers are hard: users are either assigned to one cell or the other but not both. However, CDMA enables overlapping of the repeater coverage zones, so that every mobile device is always well within range of at least one of the base stations. This enables the concept of soft handover. The soft handover process is mobile-initiated and is illustrated in Figs. 6 and 7. While a mobile device is tracking the downlink signal strength of the cell, it is currently in Cell 1; it can be searching for signal of adjacent cells (Cell 2). It is important to remember that the measurements used for the handover event are made on the downlink only.

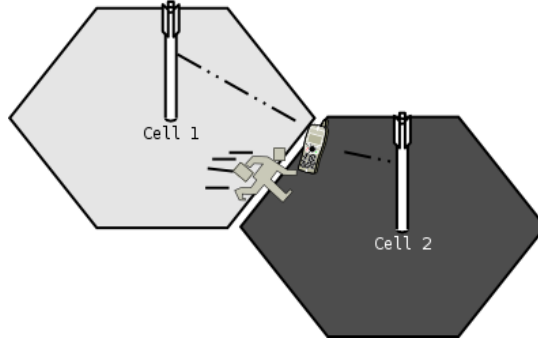


Fig. 6. Moving from Cell 1 to Cell 2 introduces handover management.

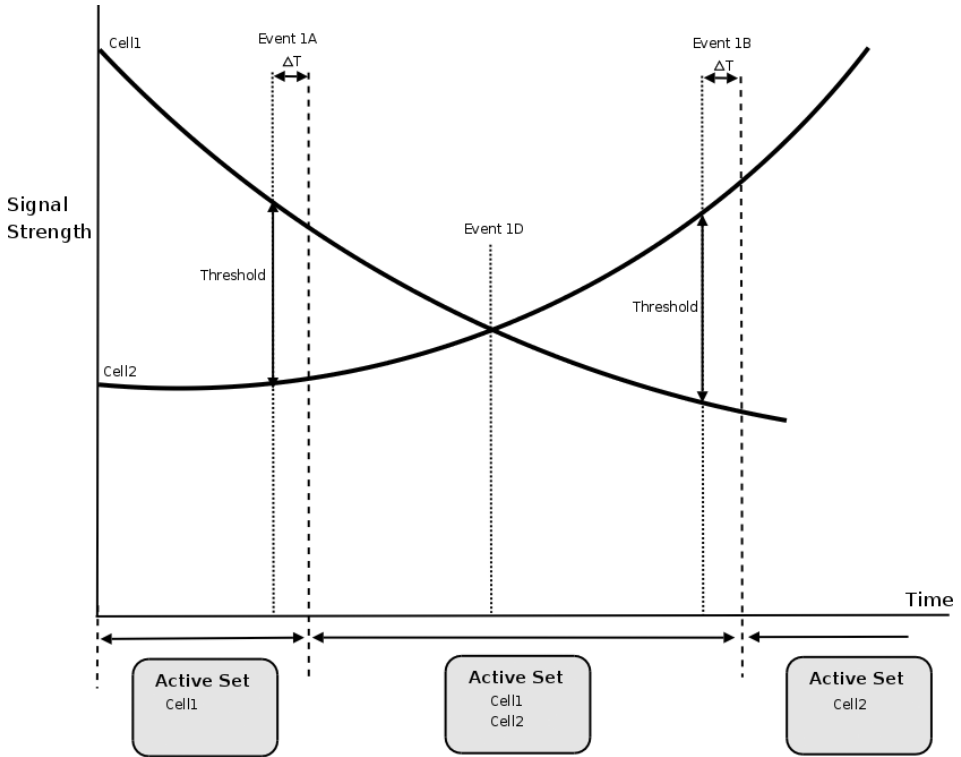


Fig. 7. Soft handover.

Idling in Cell 1 the user has Cell 1 in its active set. The active set is a set of radio links simultaneously involved in a specific communication service between the user equipment and the network. As the user moves in Cell 1 and towards Cell 2, the signal from Cell 2 increases in strength. The mobile device monitors

the received the signal from Cell 2. If the measured signal of Cell 2 is greater than the best measured cell in the active set (Cell 1) minus a threshold for a period of time Δt then Cell 2 is added to the active set. This is referred to as Event 1A in WCDMA networks. In a similar manner, if the measured signal is below the best measured cell in the active set minus the threshold for a period of time Δt then Cell 1 should be removed from the active set. This is referred to as Event 1B in WCDMA networks.

The “threshold” parameter is defined as a “guard band” in order to avoid event triggering due to insignificant measurement fluctuations in the received signal.

4.1.2. Mode and Ad hoc Mobility

These mobility management techniques are more for the mobile device than core network functions and only a brief overview of the functions is given. Mode mobility allows a wireless user to switch between different technologies and infrastructures. For example, a device may support several different wireless technologies and thus mode mobility allows the user to switch between them [20]. Ad hoc mobility allows users to communicate with one another without any permanent infrastructure being in place. Any user can act as a router to relay a session for others. The caller and the callee can also directly establish a session if near enough.

4.2. High Level Tasks

4.2.1. Introduction

High level mobility management tasks are not concerned with underlying wireless/wired technologies and are designed to solve mobility of personal applications, services and sessions. Session Initiation Protocol (SIP) is gaining acceptance as the signaling protocol of multimedia and Internet telephony, and it is envisaged it will be used in providing mobility management solutions in next generation wireless networks. The following section defines each of the high level tasks and presents how SIP can provide mobility management.

SIP supports the concepts of personal, session and even service mobility. It is a text-based protocol similar to HTTP and defined in IETF RFC 3261 [21]. It operates between the application and transport layers and it can use reliable or unreliable transport protocols. Any transport protocol can be used but UDP and TCP support must be implemented. TCP or some other congestion-controlled transport protocol must be used when sending large SIP messages.

A SIP message consists of a header and a body like HTTP. The SIP header is used for signaling to manage the user’s identity, location, authentication and authorization, etc. The body of the message is used to carry information about the session to be established or just some plain information like text or an XML document. Here is an example of a SIP message carrying a message in the

Session Description Protocol (SDP) format:

```

INVITE sip:bob@biloxi.com SIP/2.0
Via: SIP/2.0/UDP
pc33.atlanta.com;branch=z9hG4bKnashds8
To: Bob <bob@biloxi.com>
From: Alice <alice@atlanta.com>;tag=1928301774
Call-ID: a84b4c76e66710
CSeq: 314159 INVITE
Max-Forwards: 70
Contact: <sip:alice@pc33.atlanta.com>
Content-Type: application/sdp
Content-Length: 155

v=0
o=UserA 2890844526 2890844526 IN IP4 pc33.atlanta.com
s=Session SDP
c=IN IP4 pc33.atlanta.com
t=0 0
m=audio 49172 RTP/AVP 0
a=rtpmap:0 PCMU/8000

```

The header ends with a single carriage return and the SDP parameters begin in this example with “ $v = 0$ ”. The media in the session should use the parameters described here such as the IP addresses, codecs and ports listed. The media in the session can be sent and received to any location and can be independent of the route and destination that the signaling messages follow.

SIP is an application-layer signaling protocol that solves many high level mobility management issues.

4.2.2. Personal Mobility

Personal mobility is the ability of a user to access services at any device based on a unique personal identifier, for example, the Universal Personal Telecommunication (UPT) number or Personal User Identity (PUI) [22]. Furthermore, personal mobility is the capability of the network to provide those services delineated in the user’s service profile. Personal mobility involves the network capability to find the device associated with the user for addressing, routing and charging of the UPT user’s calls. According to Chung *et al.*, personal mobility [23] is one of the key issues in realizing wireless communication networks in emerging third generation mobile communication networks such as International Mobile Telecommunication 2000 (IMT-2000). It can be realized through UPT service.

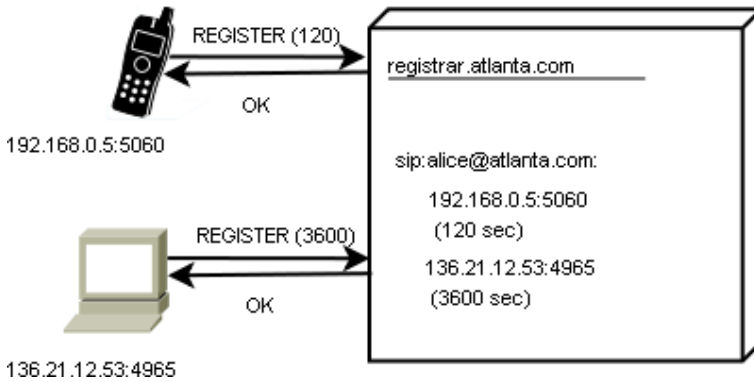


Fig. 8. SIP registration.

Example: SIP user agent registration.

SIP supports the concept of personal mobility by having a personal identifier known as the address-of-record (AOR), which takes the form of a Uniform Resource Identifier (URI). Unlike the Uniform Resource Locator (URL), the SIP URI only indicates the user’s identity and is not sufficient by itself to locate the user’s device. User location is achieved using a location service provided by an entity known as the “registrar.” The registrar can have a list of the devices that the user is currently registered on which is kept up-to-date using the value in the “expires” parameter for each URI in the “contact” header field and if not present, the value in the “expires” header field. User agents on a device register initial values and can update them by sending additional REGISTER messages. A REGISTER message with an “expires” value of “0” is a de-registration.

The diagram in Fig. 8 shows a user called Alice being registered in the atlanta.com domain using two devices with different expiry intervals. The shorter expiry time is suited to mobile devices, which move around often and may change IP address, whereas clients such as desktops can refresh their presence less often. When someone attempts to contact Alice, both devices will “ring” and Alice can accept or reject the offer on both devices. Each device will only accept the type of media it can receive. Alice can register new devices and de-register a device at any time.

```
REGISTER sip:registrar.atlanta.com SIP/2.0
Via: SIP/2.0/UDP
192.168.0.5:5060;branch=z9hG4bKnashds7
Max-Forwards: 70
To: Alice <sip:alice@atlanta.com>
From: Alice <sip:alice@atlanta.com>;tag=456248
Call-ID: 843817637684230@998sdasdh09
CSeq: 1826 REGISTER
```

```

Contact: <sip:alice@192.168.0.5>;expires=120
Expires: 7200
Content-Length: 0

REGISTER sip:registrar.atlanta.com SIP/2.0
Via: SIP/2.0/UDP
136.21.12.53:4965;branch=z9hG4bKfdsds-ewef
Max-Forwards: 70
To: Alice <sip:alice@atlanta.com>
From: Alice <sip:alice@atlanta.com>;tag=456248
Call-ID: dav23kv3@136.21.12.53
CSeq: 1 REGISTER
Contact: <sip:alice@136.21.12.53>
Expires: 3600
Content-Length: 0

```

4.2.3. Session Mobility

Session mobility is ensures that active sessions are not disrupted while devices, persons or applications are moving or being relocated. For example, a caller may want to continue a session that originally began on a mobile device to a desktop PC when entering the office. A user may also want to move parts of a session, for example, if the user has specialized devices for audio and video, such as a video projector, video wall or speakerphone [24].

Example: How a SIP session is initiated and modified.

To initiate or modify a session, the invitation request must be routed to the target user agent (UA). This is done using proxies which use DNS queries and the registrars in the target domain to locate the target user.

Initially Alice is unaware of Bob's IP address and so uses her default proxy to send the request. Alice's proxy looks at the domain of Bob's URI and locates the SIP proxy for Bob's domain. This proxy then queries the registrar in that domain for Bob's location and sends the request to all the contact addresses returned. Bob's UA may start "ringing" and this is indicated to Alice via a 180 response message. If Bob accepts the session invitation, then his UA sends a 200 OK response via the proxies to Alice. In Bob's response, the location of that UA is included in the "contact" header field and Alice's location was included in the "contact" header field of the INVITE. Now both UAs know the location of each other and can send messages in a peer-to-peer fashion. The ACK confirms that Alice received the 200 OK from Bob. The session is terminated by either UA sending a BYE which is accepted with a 200 OK response.

The session can be modified to change the parameters of the session or even the address of one of the parties involved. The procedure is simply to send an INVITE again (re-INVITE) with the new parameters in the SIP header/body.

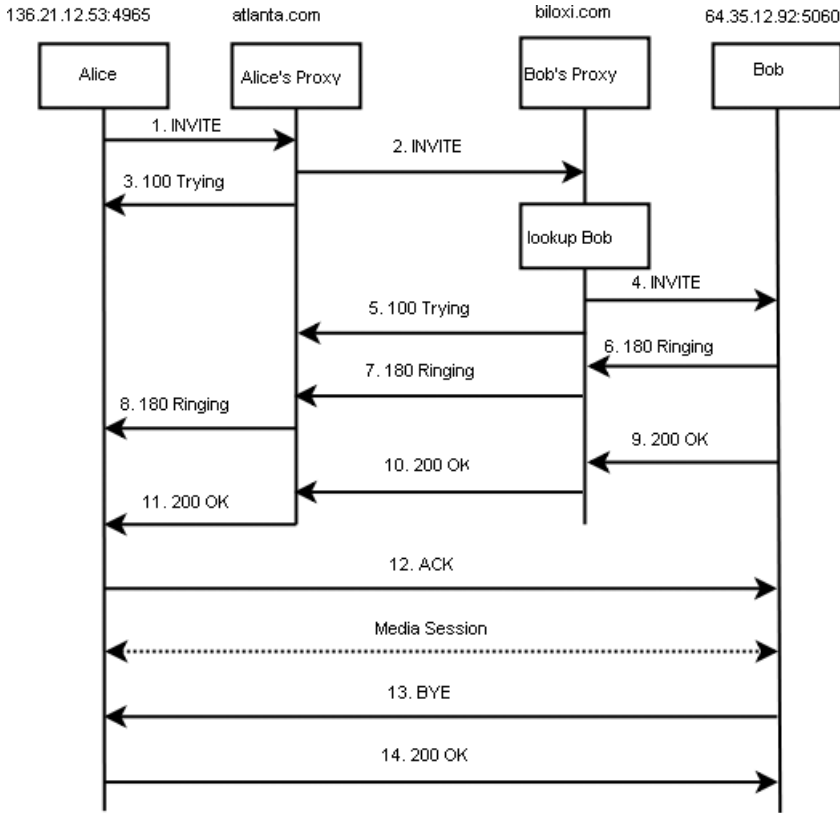


Fig. 9. SIP message sequence chart.

Parameters such as the Call-ID must remain the same so that it can be recognized as a modification to an existing session and not an entirely new session. The receiving UA can ask the user for confirmation of the change and if it is rejected, the session continues without modification. If the modification is accepted, both UAs change the parameters accordingly.

4.2.4. Service Mobility

Service mobility refers to the network capability to provide subscribed services at the device or location selected by the user. The services the user can invoke at the device, of course, depend on the capability of the device and the network serving the device. Service mobility is accomplished through keeping an up-to-date service profile of the user and interrogating it as necessary [15].

Example: How SIP provides service mobility.

While all the examples so far have focused on SIP providing a mechanism for setting up sessions using a SIP URI, it is also possible to register other types of

contact addresses using any suitable URI. Other URIs could identify phones, fax, or email addresses. The “q” parameter also identifies a user’s preference for each contact address where the highest value is the most preferred address. For example, for an AOR sip:carol@chicago.com, the “contact” header field in a REGISTER could be

```
Contact: <sip:carol@work.com>;q=0.7
Contact: <sip:carol@home.com >;q=0.3
Contact: <mailto:carol@email.com>;q=0.1
```

From the example above, we can see that Carol prefers to be contacted at her work.com SIP address. In a later registration, she can change her addresses and modify her preferences, e.g.,

```
Contact: <sip:carol@home.com>;q=0.8
Contact: <mailto:carol@webmail.com>;q=0.1
```

While each request from other users will be addressed to sip:carol@chicago.com, the destination of the request will depend on the current contact addresses and preferences registered by Carol.

5. Thoughts for Practitioners

5.1. *Implementing Quality of Service*

Supporting QoS in multi-system networks is an important issue to achieve user mobility across heterogeneous RANs with a unified service experience. Traffic management must take into account the different characteristics of each RAN to provide seamless service continuity as well as end-to-end QoS. For example, significant variation in transmission capacity can be found between RANs, etc. In addition, the traffic can cross multiple domains where each of them has its own administration, rendering end-to-end QoS provision even more complex. A mapping of the QoS services between different RANs is required to provide unified QoS traffic classes. It is also necessary to manage the QoS in each domain and to permit interaction between the different QoS management entities thanks to policy-based QoS management.

In the case of services, the traffic classes and the QoS attributes are defined as described above. As the end-to-end path in the multi-system network may include the Internet environment, the Internet QoS attributes and classes will be mapped to the classes of each RAN. There are two main QoS approaches in this area, namely, Integrated Services (IntServ) and Differentiated Services (DiffServ).

IntServ uses the per-flow approach to provide guarantees to individual streams. This approach adopts the resource reservation along the flow path using the resource reservation protocol (RSVP). The latter sets up some flow state (e.g., bandwidth reservations, accounting) in the routers that a flow goes through. An admission

control is used to guarantee the QoS. It grants or rejects the flow requests, based on availability of resources and the guarantees provided to other flows. In addition to the well-known best effort service, two service classes are proposed: the guaranteed service class for the traffic requiring a firm bound on delay and the load controlled class where flow receives a quality of service closely approximating QoS that flow would receive from an unloaded network element. This is generally not seen as a scalable QoS solution for individual flows on all IP networks due to the signaling requirements.

DiffServ follows the philosophy of the aggregation of multiple flows in the edge of the DiffServ domain into a few classes of service or Per Hop Behaviors (PHBs). There is no reservation, but Differentiated Services Code Point (DSCP) etiquette has been included in the type of service (TOS) field of the IP packet header to differentiate between different PHBs. The DSCP is set at the edge of the domain where a conditioning and shaping mechanism is executed in accordance with the requirements or rules of each service. Using the DSCP, the node inside the network determines how packets are forwarded. There are two PHBs being defined with the best effort: expedited forwarding (EF) PHB and assured forwarding (AF) PHB. EF is intended to support services with tightly bounded loss, delay and jitter. AF offers different levels of forwarding assurances for packets belonging to an aggregated flow. Packets are marked with one of three drop precedences, such that those with the highest drop precedence are dropped with lower probability than those marked with the lowest drop precedence. AF gives the customer the assurance of a minimum throughput, even during periods of congestion. The per-hop basis, lack of firm reservations and statistical nature of the guarantees given by DiffServ are seen as more scalable for Internet applications.

As DiffServ and IntServ models shall be supported for PDP contexts (3GPP 23.107), the mapping between Internet QoS and UMTS QoS is under investigation. For example, when DiffServ is used in the backbone of the UMTS, the Conversational class is transported in the EF PHB; streaming is mapped to EF or AF. Interactive and background classes are mapped into AF PHBs.

When viewing multi-system networks, the minimum set of service and classes can be those as defined by 3GPP, i.e., conversational, streaming, interactive, and background. An example set of service and classes are outlined in [Ref. 25]. The latter outlines the QoS requirements that shall be provided to the end user/applications and describes them as requirements between communicating entities, i.e., end-to-end. Figure 10 summarizes the major groups of applications in terms of QoS requirements. Applications and new applications may be applicable to one or more groups. However, there is no strict one-to-one mapping between the groups of application/service defined in [Ref. 25] and the traffic classes as defined in [Ref. 26]. For instance, an Interactive application/service can very well use a bearer of the Conversational traffic class if the application/service or the user has tight requirements on delay.

Error Tolerant	Conversational voice and Video	Voice Messaging	Streaming audio and video	Fax
Error Intolerant	Telnet, interactive games	E-commerce, Web browsing	FTP, still image paging	Email arrival notification
	Conversational (Delay <<1 sec)	Interactive (Delay approx 1 sec)	Streaming (Delay << 10 sec)	Background (Delay > 10 sec)

Fig. 10. Multi-system applications in terms of QoS requirements.

The reference user performance expectations are presented in Table 2 for conversational services, in Table 3 for interactive services and in Table 4 for streaming services. The QoS values in the tables represent end-to-end performance, including mobile-to-mobile calls and satellite components. Delay values represent one-way delay (i.e., from originating entity to terminating entity). The values included in the following tables are commonly accepted values from an end user viewpoint. The delay contribution within the mobile network should be kept to a minimum since there may be additional delay contributions from external networks. Note that the overall one-way delay in the mobile network (from UE to PLMN border) is approximately 100 ms.

5.2. QoS and SIP

SIP is not a resource reservation protocol and therefore by itself cannot guarantee a level of QoS. However, it does support QoS in conjunction with other protocols and extensions.

Example: How SIP can support QoS.

The most basic form of QoS, within SIP is the ability of a user agent to reject session requests that it cannot accept. If the invitation for a session includes a list of options, the terminating user agent has the ability to select the session description parameters that it believes are most suitable. Within a session, it is possible to renegotiate the parameters of the session to improve QoS, but it is the responsibility of the user agent application to know that change is required and what changes to make.

There is an optional extension to SIP which provides for resource management using SIP and SDP defined in [Ref. 27] and updated in [Ref. 28]. It provides for

Table 2. End user performance expectations — Conversational/realtime services.

Medium	Application	Degree of symmetry	Data rate	Key performance parameters and target values		
				End-to-end one-way delay	Delay variation within a call	Information loss
Audio	Conversational voice	Two-way	4-25 kb/s	<150 msec preferred <400 ms limit	<1 ms	<3% FER
Video	Videophone	Two-way	32-384 kb/s	<150 ms preferred <400 ms limit Lip-synch: <100 ms		<1% FER
Data	Telemetry — two-way control	Two-way	<28.8 kb/s	<250 ms	n/a	zero
Data	Interactive games	Two-way	<1 KB	<250 ms	n/a	zero
Data	Telnet	Two-way (asymmetric)	<1 KB	<250 ms	n/a	zero

Table 3. End user performance expectations — Interactive services.

Medium	Application	Degree of symmetry	Data rate	Key performance parameters and target values		
				One-way delay	Delay variation	Information loss
Audio	Voice messaging	Primarily one-way	4–13 kb/s	<1 s for playback <2 s for record	<1 ms	<3% FER
Data	Web-browsing — HTML	Primarily one-way		<4 s/page	n/a	zero
Data	Transaction services — high priority, e.g., e-commerce, ATM	Two-way		<4 s	n/a	zero
Data	E-mail (server access)	Primarily one-way		<4 s	n/a	zero

Table 4. End user performance expectations — Streaming services.

Medium	Application	Degree of symmetry	Data rate	Key performance parameters and target values		
				Start-up delay	Transport delay variation	Packet Loss at Session Layer
Audio	Speech, mixed speech and music, medium and high quality music	Primarily one-way	5–128 kb/s	<10 s	<2 s	<1% packet loss ratio
Video	Movie clips, surveillance, realtime video	Primarily one-way	20–384 kb/s	<10 s	<2 s	<2% packet loss ratio
Data	Bulk data transfer/retrieval, layout and synchronization information	Primarily one-way	<384 kb/s	< 10 s	n/a	zero
Data	Still image	Primarily one-way		<10 s	n/a	zero

a SIP “preconditions” header field and QoS parameters for SDP. Using RFC 3261 procedures alone, it would be possible for a user agent to receive an invitation to a session that it could not take part in because of a lack of network resources. The user agent would ring, and when the user answered, the session would fail. The offer from the caller can contain a list of possible session codecs and requirements, and it is necessary for the callee to answer this offer before network resources can be reserved. This extension solves the problem by requiring an exchange of a session description offer and answer before the user-agents begin to ring. If the user-agents cannot reserve the resources required, the session fails without interrupting the callee.

5.3. Implementing SIP in a Wireless Communication Network

SIP aims to be transport-independent and can operate on almost all IP-based networks, which makes it suitable for use as a high-level signaling protocol. It can be used to bridge ATM, UMTS and Ethernet-based networks as long as they support IP. It can also be used on both IPv4 and IPv6 networks as long as the user agents are capable of supporting the relevant IP version.

For these reasons, SIP has been chosen as a core signaling protocol by a number of standardization bodies for their next-generation network architectures. The 3GPP has defined the IP Multimedia Subsystem (IMS) [29], which defines how SIP fits into GPRS/UMTS networks. The 3GPP IMS standard has been adopted as a central part of the NGN architectures being developed by the 3GPP2, ITU-T, ETSI Telecoms & Internet converged Services & Protocols for Advanced Networks (TISPAN) and CableLabs.

They have defined the procedures and security requirements for authentication, authorization and accounting, including how to interface with existing user-information repositories within the network such as the Home Subscriber Server (HSS). The benefit of following this standard is the ability for international networks to interoperate and provide services across networks. It also allows for network operators to provide a standard interface to third party application providers. They can set trigger points based on elements in the SIP headers and provide services such as voicemail and call forwarding. They can also be contacted directly via a SIP URI to provide content services or directory lookup services. By using SIP, the aim is to simplify integration and reduce development time.

The framework also provides the interfaces and procedures required to implement QoS using SIP and SDP in conjunction with a policy decision point and a policy enforcement point using UMTS control mechanisms in 3GPP TS 23.107 [26]. Specifications for QoS specifically involving GPRS-based networks are contained in 3GPP TS 23.207 [30]. These elements can intercept session initiations to ensure that the requested QoS can be met and they can modify a session to ensure continuing QoS. ETSI and the ITU, on the other hand, define a more generic system known as the Resource and Admission Control System (RACS).

It is also possible for the SIP proxies within an IMS network to redirect the media within a session through a media gateway. This media gateway can perform functions such as transcoding to a format optimal to the end user device or the available bandwidth.

5.4. Base station planning

The demand for new mobile communications services has grown rapidly in recent years. Wireless networks play an important role in application areas, such as industry, medicine, personal communications, etc. To satisfy user requirements, the mobile network operators must provide these applications with a satisfactory QoS, even when the traffic increases considerably in a particular zone. Under these circumstances, the operators are obliged to increase their capacity. Since the spectral resources are finite, an option for increasing the capacity is to miniaturize and sectorize the radio zones (i.e., cells). This causes a steady increase in the number of base stations and, therefore, an increase in the network complexity. A more effective solution consists of distributing the base station within the cell and estimating its optimal antenna height taking into account the propagation conditions. Therefore, a good understanding of the wireless propagation channel is essential for cell planning strategies.

The effects on the received power of the mobile device caused by selecting different heights of the base station antennas are analyzed in the following example. The propagation model considered in this analysis is the Okumura–Hata defined for urban and suburban areas. The average received power of the mobile device is obtained from (1) in logarithmic expression as

$$P_{Rx}[dB_m] = P_{Tx}[dB_m] - G_B[dB] - G_m[dB] - L_E[dB], \quad (10)$$

where the transmitter is the base station ($G_B = G_{Tx}$) and the receiver unit is the mobile device ($G_m = G_{Rx}$). The unit dB_m is the logarithmic value of measurement related to the power in milliwatts ($dB_m = 10 \log(\text{power [mW]}/1 \text{ mW})$).

The Okumura–Hata path loss model is expressed in (3) with the following values:

$$A = 69.55 - 26.16 \log_{10}(f_c) - 13.82 \log_{10}(h_B) - a(h_m, f_c), \quad (11)$$

$$B = 44.9 - 6.5 \log_{10}(h_B), \quad (12)$$

$$C = \begin{cases} 0, & \text{(urban)} \\ 2[\log_{10}(f_c/28)]^2 + 5.4, & \text{(suburban)} \end{cases} \quad (13)$$

where h_B and h_m are the base station and mobile antenna's heights in meters and f_c is the carrier frequency in MHz. Finally, the parameter $a(h_m)$ is defined according

to the topology of the area as:

$$a(h_m, f_c) = \begin{cases} 3.2[\log_{10}(11.75h_m)]^2 - 4.97, & \text{(urban)} \\ [1.1 \log_{10}(f_c) - 0.7]h_m - [1.56 \log_{10}(f_c) - 0.8], & \text{(suburban)} \end{cases} \quad (14)$$

In this example, a GSM network with downlink transmissions in the 900 MHz band is considered for analyzing the effects of the path loss on the received signal strength. The following parameters are selected for this simulation analysis: $P_{Tx} = 33dB_m$, $G_B = 13dB$, $G_m = 0dB$, $f_c = 935$ MHz and $h_m = 1.75$ m. The received power values as a function of the distance between the base station and the mobile device are represented in Fig. 11 for an urban area, considering four different values of base station antenna heights ($h_B = 10$ m, $h_B = 30$ m, $h_B = 50$ m and $h_B = 75$ m). The chosen scenario considers base station coverage of microcells having a radius of about 500 m. It is important to remark that small-scale fading effects have not been considered.

Figure 11 shows the influence of the distance and the base station antenna height on the received power. In terms of distance, the received power decays approximately 25 dB when the mobile device moves from 25 m to 125 m of the base station for all antenna height values. The same power loss value is experienced when the mobile device moves from 125 m to 500 m. Therefore, the loss in power due to large-scale fading is more significant when the mobile device changes its position near to the base station.

The effects of the antenna height of the base station on the received power can be analyzed by fixing the distance value and comparing the power variations

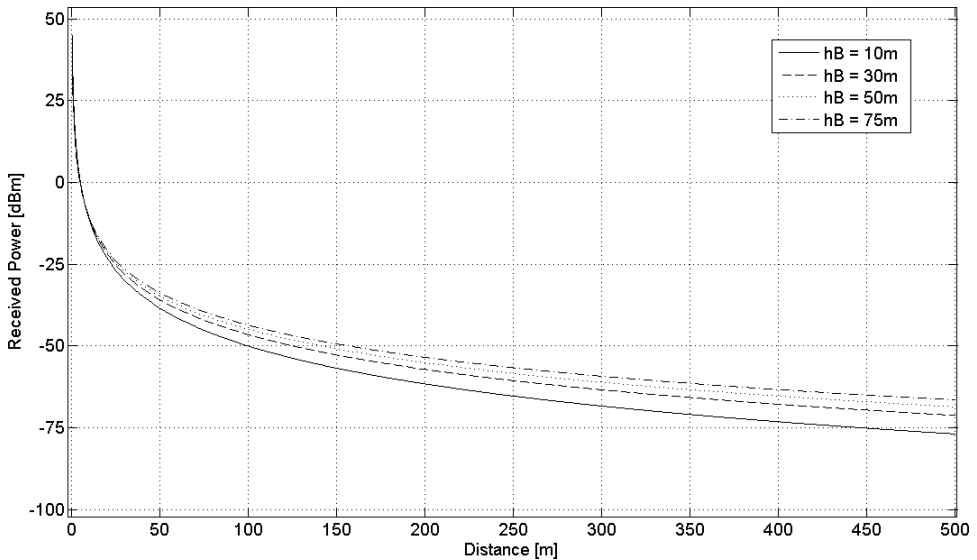


Fig. 11. Received power versus distance based on Okumura–Hata path loss model for urban areas.

when changing antenna height values. It is obvious that the received power will increase if the antenna height increases since the direct path will be less affected by obstacles. However, the improvement in the received power obtained by increasing the antenna height is not always significant, while the complexity of the base station infrastructure increases. Therefore, there is an optimum value of the base station antenna height that must be investigated for cell planning strategies. In Fig. 11, a 5 dB augmentation of the received power is observed for distance $d = 300$ when increasing the antenna height from 10 m to 30 m. However, the increase in the received power is only 2.5 dB when changing the antenna height from 30 m to 50 m.

6. Directions for Future Research

In next-generation wireless networks users will encounter both intra- and inter-system handovers. It is essential that applications running on mobile devices remain unaware of the users' movement, both intra- and inter-system, to ensure uninterrupted services with minimum QoS degradation to the end users. To achieve this continuation of upper layer sessions during handover, a new level of mobility support is required by next-generation networks. However, future mobility management brings different issues and challenges as it requires additional mobility related features such as moving networks, seamless roaming and vertical handover. Many projects and much research on mobility management within heterogeneous networks are currently ongoing with the aim of enabling seamless handover and interoperability between heterogeneous network types including both IEEE 802 and non-IEEE 802 networks. The support of multiple media types will improve the available network coverage, which can improve the user experience of mobile devices. Mobility solutions can be found by either improving existing architecture or by revising the current architecture principles to reflect the changing environment and to efficiently and easily cope with new requirements. In terms of relevance to mobility some current mobility management projects are discussed further in this section. The solutions outlined propose different addressing and packet forwarding schemes, the majority of which are IP based, allowing interoperability and easy integration with existing architectures.

The Internet Indirection Infrastructure (i3) project [31] provides mobility in a unique way by employing an overlay network with rendezvous points. These points are known servers where packets are forwarded with an identifier which is used by the receiver to obtain delivery of the packet. This level of indirection decouples the act of sending and receiving and provides natural support for mobility. However, this is not optimal as data passes via a third party during traversal from source to final destination.

Forwarding directive Association and Rendezvous Architecture (FARA) [32] is a project that aims to provide mobility by decoupling end-system names from network addresses. This flexible architecture concept separates the host

identifier from location information without the introduction of a new global namespace. The FARA model allows several different forwarding mechanisms to co-exist in the network. However, it does not take into account packet forwarding issues like performance of network nodes or provide any solution for security.

The Host Identity Payload (HIP) project [33] also attempts to optimize handover performance by introducing a separation between the location and host identity information at the IP layer. It introduces a new identity, cryptographic in nature, for host identities which can be dynamically mapped to IP addresses. The host identity is basically a public key that is used to identify a related private party in the network; the separation provides a means to handle mobility and multi-homing in a secure way. However, issues arise when the host moves between IPv4 and IPv6 networks. If a mobile does not have connectivity for both versions, it may have to use a proxy node that performs the address version conversion on behalf of the mobile device.

The IST MIND project [35] is the follow up to the BRAIN project [36] and builds on the concept of an all-IP core accessible by a variety of technologies. It provides micro-mobility solutions by enabling hosts to co-operate with self-organizing wireless ad hoc networks. Among other aspects, the research considers QoS, multi-homing, multicast and security issues.

The DRiVE project [36] supports the use of different access systems using flow based, host-controlled IP forwarding to enable seamless intersystem handover. The multi-access architecture handles macro-mobility with hierarchical mobile IP. Extensions for moving networks define a mobile IP based solution in OverDRiVE [37].

The ongoing research of the IETF NEMO [38] working group is concerned with managing the mobility of an entire network. The network is able to change its point of attachment to the Internet and thus its reachability in the network topology. The mobile network includes one or more mobile routers which connect the rest of the mobile network to the global Internet.

Ambient networks [39] is a new networking architecture currently under investigation, which aims at enabling the cooperation of heterogeneous networks belonging to different operator and/or technology domains. Ambient networks use the *network composition* principle, where networks form the basic building block of communication networks rather than devices. This approach is a more advanced concept than the simple inter-networking of IP. The current Internet assumes homogeneity in the environment in which to provide control. A unifying view like ambient networks has the potential to solve this issue of fragmented control.

In terms of standardization organisations such as the IEEE 802 LAN/MAN Standards Committee and the Internet Engineering Task Force (IETF) are currently working on developing a common framework to extend existing mobility protocols

to facilitate and optimize seamless inter-system handover. The aim is to provide a platform for successfully addressing present challenges related with QoS, security, and seamless mobility as well as allowing for scalability for future possibilities. One such group is the IEEE 802.21.

The IEEE 802.21 standard [40] assumes mobile users are capable of supporting multiple interfaces, both wired and wireless. The goal of the project is to enable mobile users to handover between IEEE 802 networks whether or not they are of different media types and furthermore enable handover between non-802 networks. To achieve this, an abstraction layer is defined, providing Media Independent Handover (MIH) functions, with the aim of simplifying the handover management to and from IEEE 802.11, 802.16, 802.3, 3GPP2 networks and the 3GPP network family. The media-independent interface should allow higher layer network selector entities to be application specific and afford the ability of selecting the best network at a given moment for the user.

To achieve complete and seamless mobility within a pervasive infrastructure, some essential issues need to be addressed, such as the integration of different technologies, composition of networks and connectivity. The progress towards the desired uninterrupted mobility will aid in successful deployment of next-generation heterogeneous networks.

7. Conclusion

In the course of the accelerated contention of wireless technologies with wired technologies, broadband wireless service has become a reality, and wireless Internet is attainable. However, QoS and cost remain as deficiencies of wireless systems. Despite the freedom of mobility, in the data arena wireless technologies have enjoyed limited popularity to speak of. The true advantage of mobility in the context of broadband services is exemplified by the capability to deliver location specific services to the user.

Without any doubt, next-generation wireless networks will be considerably more complex than today's second/third-generation wireless systems. Global roaming and the ability to access the Internet and data resources everywhere are among the key driving factors that will have profound impact on how these networks will be managed. In this changing environment, operators will introduce new services and more powerful and efficient ways of doing business by integrating new technologies with existing ones. This is already beginning to happen with mergers of WLAN networks and cellular networks. Not only will the network elements and communication devices evolve but so too will the management systems and way of managing. Mobility management has widely been recognized as one of the most important and challenging problems for next-generation networks. New management strategies will be needed to enable seamless access to wireless networks and mobile services.

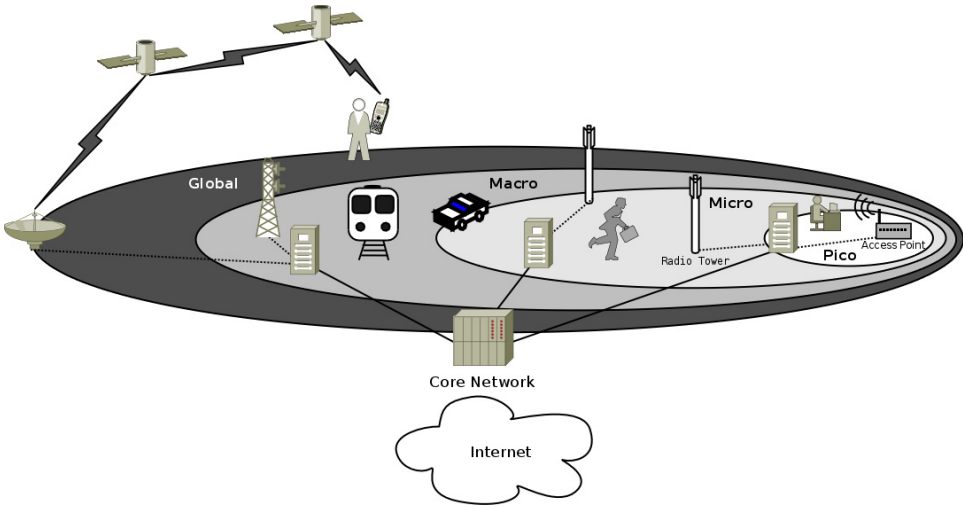


Fig. 12. Wireless communication network landscape.

The landscape of the next-generation network, as illustrated in Fig. 12, will take into consideration wireless technologies providing high-speed data access and voice in cells ranging from pico (10m) to macro (10km). At the same time the network must provide a seamless and transparent means of access to these services at a global level. From a service management and architectural point of view many things will be different in next-generation networks, but one fundamental concept will not change: users will be mobile.

Acknowledgments

This work was supported by the Science Federation of Ireland NCNRC project (03/CE.2/I315) and the Irish Research Council for Science, Engineering and Technology program.

Terminology

Mobility: is the ability to move or change position.

Wireless communication: The transfer of information over a distance without the use of electrical conductors or wires. The distances involved may be short (<100m) or very long (>2km) for radio communications. Information transmitted over distance can be affected by distortions such as fading.

Multipath fading: The propagation phenomenon that results in radio signals reaching the receiving antenna via separate paths. Causes of multipath fading include atmospheric ducting, ionospheric reflection and refraction, and reflection from terrestrial objects. As a result of multipath fading, the received signal is a sum

of many reflections each with different propagation time, phase and attenuation values.

Doppler effect: The shift in frequency and wavelength of waves which results from a source moving with respect to the medium, a receiver moving with respect to the medium, or even a moving medium.

Large-scale fading: The slow variations of the mean (distance-dependent) power of the received signal over time. This phenomenon is caused by many factors including antenna losses and the presence of obstacles in the signal path, which leads to propagation losses.

Wireless Communication Network: Consists of (i) mobile devices, (ii) an access network, and (iii) a core network. A radio link between mobile devices and base stations allows the mobile devices to roam the wireless network and maintain a communications link.

Mobility management: The aim of mobility management is to track where the subscribers are, so that calls, SMS and other mobile phone services can be delivered to them. Furthermore, mobility management ensures ongoing calls are maintained as the subscriber moves.

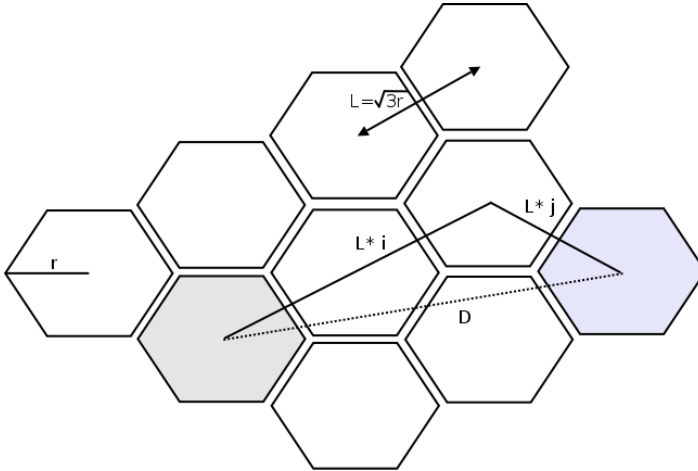
Quality of service (QoS): Comprises all the aspects of a connection, such as time to provide service, voice quality, echo, loss, reliability, etc.

Sessions: A multimedia session is a set of multimedia senders and receivers and the data streams flowing from senders to receivers. A multimedia conference is an example of a multimedia session.

Media Independent Handover (MIH): An IEEE 802.21 function that supports algorithms enabling seamless handover between networks of the same type as well as handover between different network types. MIH may communicate with various IP protocols including SIP for signaling, mobile IP for mobility management, and DiffServ and IntServ for QoS.

Exercises

1. Name the different types of mobility management used in wireless communication networks and classify them into high and low level tasks.
2. In cell planning a frequency is placed in a cluster of N cells and reused in another cluster separated by a distance D . Using the cosine law, $[c^2 = a^2 + b^2 - 2ab \cos(\mu)]$ show that the distance radius ratio is related to the number of cells in the cluster as $D/r = \sqrt{3N}$, where r is the radius of a cell. Let L be distance between two cell centres.



3. Name (i) the four service classes in wireless communication networks, (ii) the characteristics of the traffic, and (iii) outline the QoS requirements for each.
4. User mobility in wireless communication networks varies from low mobility ($<3\text{m/s}$) to very high mobility ($>100\text{km/h}$). Explain the impact mobility has on network connectivity in terms of *terminal mobility* and its subsequent functions if the user is in a cell with radius 1 km.
5. How does the network track mobile users in the network considering that 90% of the time the mobile device is in an idle state (radio off)? For the other 10% the mobile is either on a call (active state/radio on) or periodically waking up.
6. Hard handover is considered, in terms of connection, a “break before make” while soft handover is a “make before break”. Briefly explain why GSM supports only hard handover while UMTS can support hard and software handover.
7. The ITU-T Y-1541 has QoS Classification of 0 to 5 with corresponding characteristics. In the *Applications* column in the table below map common wireless communication applications to the ITU-T QoS Classification.

Common Wireless Applications: Interactive gaming, voice calls, telnet, video conferencing, FAX, network background traffic, FTP, torrent peer to peer, email, web browsing.

QoS	Characteristics	Applications
0	Realtime, jitter sensitive, highly interactive	
1	Realtime, jitter sensitive, interactive	
2	Transaction data, highly interactive	
3	Transaction data, interactive	
4	Low loss only (short data, bulk data)	
5	Traditional applications of default IP networks	

8. A base station is transmitting data to a GSM user located at a distance $d = 2$ km in a suburban area. The base station is using a channel with center frequency $f_c = 935$ MHz. Calculate the minimum height of the base station antenna that satisfies a required received power value of $P_{rx} = -85$ dBm considering the Okumura–Hata path loss model. Multipath fading and Doppler effects can be neglected for this analysis. The main parameters of the base station and the mobile device are:

$G_B = 12$ dB (Gain of the base station antenna),

$G_m = 0$ dB (Gain of the mobile unit antenna),

$P_{tx} = 33$ dBm (Transmitted power), a–d

$h_m = 1.5$ m (Height of the mobile device antenna).

9. How does a user agent update its location using SIP? Fill in the missing elements of the relevant SIP message sent by the user agent from its new IP of 192.168.0.35. The registrar has a domain of “registrar.altanta.com”

____ sip:_____ SIP/2.0

Via:SIP/2.0/UDP _____:5060;branch=z9hG4bKnashes7

Max-Forwards: 70

To: Alice <sip:_____>

From: Alice <sip:_____>;tag=456248

Call-ID: 8438176384230@998sdasdh09

CSeq: 1826 _____

Contact: <sip:_____>;expires=120

Expires: 7200

Content-Length: 0

10. What is the function of the “expires” parameter in a “contact” header field?

References

1. D. Tse and P. Viswanath, *Fundamentals of Wireless Communication* (Cambridge University Press, 2005).
2. T. S Rappaport, *Wireless Communications Principles and Practices*, 2nd edn. (Prentice-Hall, 2002), pp. 107–110.
3. M. J. Feurstein, K. L. Blackard, T. S. Rappaport, Y. Seidel and H. H. Xia, Path loss, delay spread, and outage models as functions of antenna height for microcellular system design, *IEEE Trans. Veh. Tech.* **43**(3), 487–498 (1994).
4. E. N. Singer, *Land Mobile Radio Systems*, (PTR Prentice Hall, 1994), pp. 196.
5. T. Okumura, E. Ohmori and K. Fukuda, Field strength and its variability in VHF and UHF land mobile service, *Rev. Elect. Commun. Lab.* **16**(9/10), 825–873 (1964).
6. A. Papoulis, *Probability and Random Processes, An Introduction for Applied Scientists and Engineers* (McGraw-Hill, New York, 1965).
7. S. Lien and M. Cherniakov, Analytical approach for multipath delay spread power distribution, *IEEE Global Telecommunications Conference*, Vol. 6, Sydney, Australia (IEEE, New York, 1998), 3680–3685.
8. R. Price and P. E Green Jr., A communication technique for multipath channels, *Proc. IRE*, **46**(3), 555–570 (1958).

9. M. Guizani, *Wireless Communications Systems and Networks* (Kluwer Academic Publishers, 2004).
10. W. D. Rummmler, R. P. Coutts and M. Liniger, Multipath fading channel models for microwave digital radio, *IEEE Commun. Mag.* **24**(11), 30–42 (1986).
11. A. A. M. Saleh and R. A. Valenzuela, A statistical model for indoor multipath propagation, *IEEE J. Sel. Area. Commun.* **SAC-5**(2), 128–137 (1987).
12. J. G. Proakis, *Digital Communications*, 4th edn. (McGraw-Hill, New York, 2001), pp. 257–282.
13. V. H. MacDonald, The cellular concept, *Bell Sys. Tech. J.* **58**(1), 15–41 (1979).
14. T. Tugcu, A realistic mobility model and its application to a reservation based call admission scheme for ds-cdma cellular systems, Masters thesis, Institute for Graduate Studies in Science and Engineering, Bogazii University (2001).
15. R. Pandya, Emerging mobile and personal communication systems, *IEEE Commun. Mag.* **33**(6), 44–52 (1995).
16. I. F. Akyildiz, J. Mcnair, J. S. M. Ho, H. Uzunalioglu and W. Wang, Mobility management in next-generation wireless systems, *Proc. IEEE* **87**(8), 1347–1384 (1999).
17. Y.-B. Lin and A.-C. Pang, Comparing soft and hard handovers, *IEEE Trans. Veh. Tech.* **49**(3), 792–798 (2000).
18. D. Wong and T. J. Lim, Soft handovers in CDMA mobile systems, *IEEE Pers. Commun. Mag.* **4**(6), 6–17 (1997).
19. R. Prakash and V. V. Veeravalli, Locally optimal soft handover algorithms, *IEEE Trans. Veh. Tech.* **52**(2), 231–260 (2003).
20. Q. Wang and M. A. Abu-Rgheff, Next generation mobility support, *IET Commun. Eng.* **1**(1), 16–19 (2003).
21. J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley and E. Schooler, *RFC 3261 — SIP: Session Initiation Protocol* (Internet Engineering Task Force 2002).
22. Telecommunication Standardization Sector of ITU, F.852: Universal personal telecommunication UPT service description (service set 2) (2000).
23. Y. Won Chung, M. Young Chung and D. Keun Sung, Effect of personal mobility management in mobile communication networks, *IEEE Trans. Veh. Tech.* **52**(5), (2003).
24. H. Schulzrinne and E. Wedlund, Application-layer mobility using SIP, *ACM SIGMOBILE Mob. Comput. Commun. Rev.* **4**(3), 47–57 (2000).
25. 3rd Generation Partnership Project, 3GPP TS 23.105 — Services and Service Capabilities, Release 7v.7.1.0 (2007).
26. 3rd Generation Partnership Project, 3GPP TS 23.107 — Quality of Service (QoS) Concept and Architecture, Release 7v.7.1.0 (2007).
27. J. Rosenberg, RFC 3312 — *Integration of Resource Management and Session Initiation Protocol (SIP)*, eds. G. Camarillo and W. Marshall (Internet Engineering Task Force, 2002).
28. G. Camarillo and P. Kyzivat, RFC 4032 — *Update to the Session Initiation Protocol (SIP) Preconditions Framework* (Internet Engineering Task Force, 2005).
29. 3rd Generation Partnership Project, TS 23.228 — IP Multimedia Subsystem (IMS) — Stage 2, Release 7v.7.9.0 (2007).
30. 3rd Generation Partnership Project, 3GPP TS 23.207 — End-to-end Quality of Service (QoS) concept and architecture, Release 7v.7.0.0 (2007).
31. I. Stoica, D. Adkins, S. Zhuang, S. Shenker and S. Surana, Internet indirection infrastructure, *IEEE/ACM Trans. Network.* **12**(2), 205–218 (2004).

32. D. Clark, R. Braden, A. Falk and V. Pingali, FARA: Reorganizing the addressing architecture, *Proc. ACM SIGCOMM FDNA Workshop*, Karlsruhe, Germany (ACM, New York, 2003), p. 313.
33. P. Jokela, P. Nikander, J. Melen, J. Ylitalo and J. Wall, Host identity protocol: Achieving IPv4–IPv6 handovers without tunneling, *Proc. Evolute Workshop 2003: “Beyond 3G Evolution of Systems and Services”*, University of Surrey, Guildford, UK (2003).
34. D. Wisely and E. Mitjana, Paving the road to systems beyond 3G — The IST MIND project, *J. Commun. Network. (JCN)*, Special issue (Korea Information and Communications Society, 2002).
35. J. Urban, D. Wisely, E. Bolin, G. Neureiter, M. Liljeberg and T. Robles, BRAIN — an architecture for a broadband radio access network of the next generation, *Wireless Comm. Mob. Comput.* **1**(1), 55–75 (2001).
36. T. Paila, S. Alladin, M. Frank, T. Goransson, W. Hansmann, T. Lohmar, R. Toenjes and L. Xu, Flexible network architecture for future hybrid wireless systems, *IST Mobile Summit*, Barcelona, Spain (2001).
37. M. Ronai, A. Petrescu, R. Tönjes and M. Wolf, Mobility issues in OverDRiVE mobile networks, *IST Mobile Summit*, Aveiro, Portugal (2003).
38. C. Bernardos, A. de la Oliva, M. Calderon, D. von Hugo and H. Kahle, NEMO: Network mobility — Bringing ubiquity to the Internet access, *IEEE INFOCOM 2006 demonstration*, Barcelona, Spain (2006).
39. V. Typpö, J. Eisl, J. Höller, R. A. Calvo and H. Karl, Research challenges in mobility and moving networks: An ambient networks view, *Broadband Satellite Communication Systems and the Challenges of Mobility*, IFIP International Federation for Information Processing Series, Vol. 169/2005 (Springer Boston, 2005).
40. P. Goransson and R. Greenlaw, Roaming between 802.11 and other wireless technologies, *Secure Roaming in 802.11 Networks* (Newnes, 2007), pp. 289–294.