Solution Manual for Mobile Communications 2nd ed.

Jochen H. Schiller, Freie Universität Berlin, Germany

schiller@computer.org, www.jochenschiller.de

1. Introduction

1.1 Good sources for subscriber numbers and other statistics are, e.g., www.gsmworld.com, www.3gpp.org, www.3gpp2.org, www.emc-database.com, www.3g.co.uk, www.regtp.de …

1.2 Today's GSM operators add the new 3G air interfaces of UMTS to their existing GSM/GPRS infrastructure networks. Current GSM/GPRS networks already offer packet and circuit switched data transmission following the Release 99 of UMTS. The operators have to install new radio access networks, i.e., antennas, radio network controller etc. as described in chapter 4. The situation is similar for operators using cdmaOne (IS-95) technology. However, these operators go for cdma2000 as this system allows them to reuse their already existing infrastructure. Thus, based on the separation of the mobile phone systems into (very roughly) CDMA and GSM operators will lead to two different major 3G systems, cdma2000 and UMTS (and their future releases). Right now, it does not seem that there is a place for a third 3G system. Current TDMA operators might move to EDGE enhanced systems or join the UMTS system. However, it is still open what will happen in China – the Chinese system TD-SCDMA was pushed by the government, but networks and devices are still missing. Currently, the majority of Chinese subscribers use GSM, some operators offer CDMA.

2. Wireless Transmission

2.1 Check also the WRCs that try to harmonize global frequency plans.

2.2 Below 2 MHz radio waves follow the ground (ground wave propagation). One factor for this is diffraction (waves are bound towards obstacles that have sizes in the order of the wavelength), another factor is the current induced in the Earth's surface, which slows the wavefront near the earth, causing the wavefront to tilt downward. Several reasons make low frequencies unusable in computer networks:

- Lower frequencies also mean lower data rates according to Nyquist/Shannon as the available bandwidth is less.
- Lower frequencies also require large antennas for efficient transmission and reception. This might work for submarines, not for mobile phones.
- Lower frequencies penetrate material more easily. Thus SDM is more difficult – cell size would increase dramatically and frequency reuse would be almost impossible.

2.3 Frequencies in the THz range, e.g., infrared, visible light, are easily blocked by obstacles and, thus, do not interfere with other transmissions. In this case, only the standard safety regulations apply (e.g., laser emission). Most radio systems stay well beyond 100 GHz as it is not that simple to generate higher frequencies (in the lower THz range).

2.4 The classical European approach was based on standardisation and regulation before any products were available. The EU governments founded ETSI to harmonize all national regulations. ETSI created the standards, all countries had to follow. In the US companies develop systems and try to standardize them or the market forces decide upon success. The FCC, e.g., only regulates the fairness among different systems but does not stipulate a certain system. The effects of the two different approaches are different. Many "governmental" standards in Europe failed completely, e.g., HIPERLAN 1, some succeeded only in Europe, e.g., ISDN, and however, some soon became a worldwide success story, e.g., GSM. For most systems the US approach worked better, first some initial products, then standards. One good example is the wireless LAN family 802.11, a good counter example is the

mobile phone market: several different, incompatible systems try to succeed, many features, well established in Europe since many years, are not even known in the US (free roaming, MMS, GPRS roaming, no charges for being called etc.).

2.5 Computers, in contrast to, e.g., TV sets, travel around the world as laptops, PDAs etc. Customers want to use them everywhere. Thus it is very important to be able to use built-in WLAN adapters around the globe without reconfiguration and without licensing. Furthermore, it is much cheaper for WLAN manufacturers to design a single common system compared to many different systems for probably small markets.

2.6 No. Loss-free transmission of analogue signals is not possible. Attenuation, dispersion etc. will always distort the signal. Additionally, each digital signal is transmitted as “bundle” of analogue sine waves (think of Fourier!). A perfect digital signal with rectangular shape requires an infinite number of sine waves to be precisely represented (the digital signal can be represented as infinite sum of sine waves according to Fourier). However, no medium can transmit infinite high frequencies. Thus, the digital signal can never be transmitted without any loss.

2.7 Without any additional “intelligence” directional antennas are not useful in standard mobile phones as users may not want to direct the phone to a certain antenna. Users move, rotate, flip the phones etc. Phones are in bags, pockets, … while operated hands-free. There is no chance of directed transmission. However, new developments comprising fast signal processors and multiple antennas may exploit directed characteristics of antennas (beam forming). There are several ways of improving the gain of an antenna: right dimensioning (e.g., half the wavelength), multiple antennas plus a signal processor combining the signals, active and passive components attached to the antenna (compare with traditional TV antennas, satellite dishes etc.).

2.8 Problems: attenuation, scattering, diffraction, reflection, refraction. Except for attenuation all other effects can divert the waves from a straight line. Only in vacuum and without gravitational effects radio waves follow a straight line. Without reflection

radio reception in towns would be almost impossible. A line.-of-sight almost never exists. However, reflection is the main reason for multipath propagation causing ISI.

2.9 ISI mitigation: large enough guard spaces between symbols/low symbol rate (used in OFDM: distribute the symbol stream on many different carriers), channel estimation/calculate the n strongest paths and adapt the receiver accordingly. Using higher frequencies reduces the effects of multipath propagation and thus ISI (waves more and more behave like light). The higher the symbol rate the stronger the ISI. If senders and/or receivers move fast the chances for ISI are higher because the location of obstacles changes, hence the number, magnitude, and timing of the secondary pulses – it is difficult to follow the signals and adjust the delays for recombination. ISI lowers the bandwidth of a TDM scheme as the guard spaces require some time.

2.10 Several mechanisms exist to mitigate narrowband interference (which might be caused by other senders, too):

- Dynamic Frequency Selection: Senders can sense the medium for interference and choose a frequency range with lower/no interference. HiperLAN2 and 802.11h use this scheme. Network operators can also this scheme to dynamically assign frequencies to cells in mobile phone systems. DFS has a relatively low complexity.
- Frequency hopping: Slow frequency hopping (several symbols per frequency) may avoid frequencies with interference most of the time with a certain probability. This scheme may be used in GSM. Furthermore, wireless systems can use this principle for multiplexing as it is done in Bluetooth systems (still slow hopping as Bluetooth sends many symbols, indeed a whole packet, on the same frequency). Fast hopping schemes transmit a symbol over several frequencies, thus creating a spread spectrum. FH systems have medium complexity. Main topic is synchronisation of the devices.

- Direct sequence spread spectrum: Data is XORed with a chipping sequence resulting in a spread signal. This is done in all CDMA systems, but also in WLANs using, e.g., Barker sequences for spreading (e.g., 802.11b). The signal is spread over a large spectrum and, thus, narrowband interference only destroys a small fraction of the signal. This scheme is very powerful, but requires more powerful receivers to extract the original signal from the mixture of spread signals.

2.11 Worldwide regulation always uses FDM for separating different systems (TV, WLAN, radio, satellite, …). Thus, all radio systems must modulate the digital signal onto a carrier frequency using analogue modulation. The most prominent system is the traditional radio: all music and voice use frequencies between, e.g., 10 Hz and 22 kHz. However, many different radio stations want to transmit at the same time. Therefore, all the original signals (which use the same frequency range) must be modulated onto different carrier frequencies. Other motivations for digital modulation are antenna and medium characteristics. Important characteristics for digital modulation are spectral efficiency, power efficiency and robustness. Typical schemes are ASK, PSK, FSK.

2.12 The receiver may “check” the distance between the received point and the neighbouring points. The receiver then chooses the closest neighbour and assumes that the sender originally transmitted data represented by the chosen point. The more points a PSK scheme uses the higher are chances that interference (noise) shifts a transmitted “point” onto another. If the gaps between the points are too small, in particular smaller than noise added during transmission, chances are very high that the receiver will map received data onto the wrong point in the constellation diagram (please note: data is coded using PSK, the points in the constellation diagram represent codes, these codes are then transmitted – it is just simpler to think in “points”…).

2.13 Main benefits: very robust against interference, inherent security (if the spreading code is unknown it is very difficult to tap the transmission), basis for CDMA

technologies, can be used in the “background” of existing systems if the signal level is low enough. Spreading can be achieved by XORing a bit with a chipping sequence or frequency hopping. Guard spaces are now the orthogonality of the chipping sequences or hopping patterns. The higher the orthogonality (well, that is not very mathematical, but intuitive), the lower the correlation of spread signals or the lower the collision probability of frequency hopping systems. DSSS system typically use rake receivers that recombine signals travelling along different paths. Recombination results in a stronger signal compared to the strongest signal only.

2.14 The main reason is the support of more users. Cellular systems reuse spectrum according to certain patterns. Each cell can support a maximum number of users. Using more cells thus results in a higher number of users per km². Additionally, using cells may support user localisation and location based services. Smaller cells also allow for less transmission power (thus less radiation!), longer runtime for mobile systems, less delay between sender and receiver. Well, the downside is the tremendous amount of money needed to set-up an infrastructure with many cells. Typically, each cell holds a certain number of frequency bands. Neighbouring cells are not allowed to use the same frequencies. According to certain patterns (7 cluster etc.) cellular systems reuse frequencies. If the system dynamically allocates frequencies depending on the current load, it can react upon sudden increase in traffic by borrowing capacity from other cells. However, the “borrowed” frequency must then be blocked in neighbouring cells.

2.15 TDM/FDM-systems have a hard upper limit of simultaneous users. The system assigns a certain time-slot at a certain frequency to a user. If all time-slots at all frequencies are occupied no more users can be accepted. Compared to this “hard capacity” a CDM system has a so-called “soft-capacity” (compare filling a box with bricks or tissues). For CDM systems the signal-to-noise-ratio typically limits the number of simultaneous users. The system can always accept an additional user. However, the noise level may then increase above a certain threshold where

transmission is impossible. In TDM/FDM systems additional users, if accepted, do not influence other users as users are separated in time and frequency (well, there is some interference; however, this can be neglected in this context). In CDM systems each additional user decreases transmission quality of all other users (the space for the tissues in the box gets tighter).

3. Medium Access Control

3.1 Stations in a wired network “hear” each other. I.e., the length of wires is limited in a way that attenuation is not strong enough to cancel the signal. Thus, if one station transmits a signal all other stations connected to the wire receive the signal. The best example for this is the classical Ethernet, 10Base2, which has a bus topology and uses CSMA/CD as access scheme. Today’s wired networks are star shaped in the local area and many direct connections forming a mesh in wide area networks. In wireless networks, it is quite often the case that stations are able to communicate with a central station but not with each other. This lead in the early seventies to the Aloha access scheme (University of Hawaii). So what is CS (Carrier Sense) good for in wireless networks? The problem is that collisions of data packets cause problems at the receiver – but carrier sensing takes place at the sender. In wired networks this doesn’t really matter as signal strength is almost the same (ok, within certain limits) all along the wire. In wireless networks CS and CD at the sender doesn’t make sense, senders will quite often not hear other stations’ signals or the collisions at the receiver.

3.2 In case of Aloha stations do not care about other stations but simply access the medium if they have to send data. There are no stations exposed as stations do not perform carrier sensing. Hidden stations may cause collisions. The same is true for slotted Aloha the only difference being the slotted character of medium access. Reservation schemes typically work with a central reservation station which can be heard by all others. Without this condition or equivalent means of distributing

reservations the whole scheme will not work. Thus, there are no hidden or exposed terminals. MACA is designed to handle hidden and exposed terminals in a distributed WLAN without central reservation station. However, MACA may fail in case of asymmetric communication conditions or highly dynamic topologies (stations may move fast into collision range).

3.3 As long as a station can receive a signal and the signal arrives at the right time to hit the right time-slot it does not matter in TDMA systems if terminals are far or near. In TDMA systems terminals measure the signal strength and the distance between sender and receiver. The terminals then adapt transmission power and send signals in advance depending on the distance to the receiver. Terminals in CDMA systems have to adapt their transmission power very often (e.g., 1500 times per second in UMTS) so that all signals received, e.g., at a base station, have almost the same strength. Without this one signal could drown others as the signals are not separated in time.

3.4 Typically, SDMA is performed or supported by a network provider. The provider plans the network, i.e., places the base stations according to certain topologies, geographic situations, capacity planning etc. If the system is running, base stations support the infrastructure in the decision of assigning a certain base station to a terminal. This is often based on received signal strength or the current capacity. The mobile terminal supports the infrastructure by transmitting information about the received signal strengths. The terminal can furthermore initiate the change of the access point.

3.5 Modulation – Transmitters must shift all baseband signals to a carrier frequency. This is typically an analogue process and requires analogue components. Classical receivers also need filters for receiving signals at certain frequencies. Depending on the carrier frequency different antennas may be needed. Pure TDMA systems stay on one frequency, all receivers can wait on the same frequency for data. In FDMA systems receivers have to scan different carrier frequencies before they can receive signals. MAC is performed on many different layers. The WRCs (World Radio

Conferences) are used for worldwide frequency assignments such as the 2 GHz range for IMT-2000. ITU controls worldwide frequency usage. National authorities regulate frequencies in different nations. On the next lower layers network operators perform MAC: frequencies usage is controlled by network planning and current load. Finally, base stations in mobile phone systems assign frequencies to terminals depending on the current availability. In WLANs network administration assigns frequencies thus forming cells.

3.6 Wireless networks can use different frequencies, different time slots or even different codes to implement duplex channels. Typical wired networks simply use different wires (however, more elaborated schemes such as echo cancellation are feasible, too).

3.7 If communication systems use fixed TDM patterns terminals can be very simple. The only requirement is to stay synchronised to be able to receive the right data. This is the standard system in classical telecommunication networks (e.g., ISDN, PCM-30 systems, SDH etc.). Ethernets, the Internet, wireless LANs etc. work demand driven. Here the advantage is the low overhead when starting communication: terminals don't have to setup connections reserving time slots prior to communication. However, users transmit more and more data compared to voice. Most networks of today are data dominated (if the amount of data is considered, not the revenue). Thus, data transmission should be optimised. While WLANs are optimised for data from the beginning (and isochronous audio transmission causes some problems), wide area mobile phone systems started as almost voice only systems. The standard scheme is circuit switched, not packet switched. As more an more data is transmitted these networks have to integrate more and more data oriented technologies: GPRS in GSM, IP in the core network of UMTS etc.

3.8 Interference and countermeasures in:

- SDMA: Interference happens if senders are too close to each other. Terminals or base stations have to keep a minimum distance.

- TDMA: Interference happens if senders transmit data at the same time. Countermeasures are tight synchronisation and guard spaces (time gap between transmissions).
- FDMA: Interference happens if senders transmit data at the same frequency. Thus, different frequencies have to be assigned to senders by organisations, algorithms in base stations, common frequency hopping schemes etc. Furthermore, guard bands between used frequency bands try to avoid interference.
- CDMA: Interference happens if senders transmit data using non-orthogonal codes, i.e., the correlation is not zero. Thus, senders should use orthogonal or quasi-orthogonal codes.

3.9 Even in vacuum radio waves have limited velocity: the speed of light. As soon as matter is in the way waves travel even slower. Thus, it can happen that a sender senses the medium idle, starts the transmission and just in a moment before the waves reach another sender this second sender senses the medium idle and starts another transmission. This is the reason for CD (listen while talk) in classical CSMA/CD Ethernets.

3.10 After reservation of the medium succeeded no more collisions can occur (if the system is error free). Reservation schemes can also guarantee bandwidth, delay, and maximum jitter. Thus, during the transmission nothing can happen. Compared to classical Aloha the collision probability is lower because the contention period is kept short compared to the contention-free period where transmission takes place. A disadvantage of reservation schemes is the latency for data transmission. Before terminals can start transmission they have to reserve the medium. This wastes time in case of a very lightly loaded medium.

3.11 Think of asymmetric transmission conditions and, for example, the hidden terminal scenario. What if station C in figure 3.10 transmits with a lot of power while it cannot

receive anything from B? Then MACA fails because CTS is not received but C causes a collision at B.

3.12 Fixed TDMA schemes can give hard guarantees – that's why they are used in classical phone systems (ISDN, SDH, GSM/CSD, …). Also implicit reservations can give guarantees after the reservation succeeded. Furthermore, all centralistic systems, i.e., systems with a base station or access point controlling data transfer, can give guarantees. All non-deterministic schemes, such as CSMA/CA, MACA, cannot give any hard guarantees.

3.13 The guard space between users in CDMA systems is the orthogonality between the spreading codes. The lower the correlation is, the better is the user separation.

3.14 The transmitted signal in this simplified example is (-2,0,0,-2,+2,0) + (1,-1,0,1,0,-1) = (-1,-1,0,-1,+2,-1). The receiver will calculate for A: (-1,-1,0,-1,+2,-1) * (-1,+1,-1,-1,+1,+1) = 1-1+0+1+2-1 = 2. For B the result is: (-1,-1,0,-1,+2,-1) * (+1,+1,-1,+1,-1,+1) = -1-1+0-1-2-1 = -6. The receiver can decide "more easily" for the binary 0 in case of B compared to the binary 1 in case of A. Noise can obviously affect the signal. But still the receiver can distinguish between the two signals – our simple example uses perfectly synchronised signals (the spread symbols are in phase). Adding the near/far problem to our simplified example does not change much: still the receiver can detect the signal – unless noise becomes too strong compared to the signal. Simply multiply the noise and B's signal by, let us say, 20. The transmitted signal is then: A_s+20*B_s+20*noise = (-1,+1,-1,-1,+1,+1) + (-20,-20,+20,-20,+20,-20) + (+20,-20,0,+20,0,-20) = (-1,-39,19,-1,+21,-39). The receiver then receives for A: (-1,-39,19,-1,+21,-39) * (-1,+1,-1,-1,+1,+1) = 1-39-19+1+21-39 = -74, and for B: (-1,-39,19,-1,+21,-39) * (+1,+1,-1,+1,-1,+1) = -1-39-19-1-21-39 = -120. Both results are negative, the receiver can not reconstruct the original data of A, but that of B. This example should just give a rough feeling what the problems are. For our simple problem here we don't see all the effects: the spreading codes are much too short, everything is synchronised.

4. Telecommunication systems

4.1 Key features: GSM (wide area coverage, bandwidth 9.6-50 kbit/s, voice, SMS, MMS), DECT (local coverage, voice, data, high density), TETRA (regional coverage, ad-hoc mode, very fast connection set-up, group call, voice, data, very robust), UMTS (medium coverage, higher data rates 384 kbit/s, flexible bandwidth assignment). Common features are traditional voice support (circuit switched), integration into classical fixed telecommunication network, ISDN core network. The systems have different, unique properties: GSM has wide area coverage, TETRA ad-hoc mode and fast connection setup, DECT can support high user densities. If allowed from licensing GSM could replace DECT, if modified GSM can replace TETRA (e.g., GSM-Rail) – under certain conditions (GSM does not offer an ad-hoc mode). But also UMTS has specific advantages – higher data rates compared to classical GSM (but lower coverage) and higher coverage than WLANs (but lower data rates).

4.2 Systems optimised for voice transmission support certain fixed data rates and operate circuit switched. Data transmission happens quite often spontaneous with varying data rates. Thus either too much bandwidth is reserved to accommodate the maximum expected data rate or data transmission experiences long delays due to connection setup. One possible step towards the support of data transmission is the introduction of packet switched services as known from the Internet. An example is GPRS in GSM. Instead of time-based billing providers can now bill based on volume (however, application based billing would make even more sense as customers are not interested in bytes but useful applications).

4.3 The three big categories are bearer, tele, and supplemental services. Separation of services supports phased introduction of services and separation of concerns: network providers, service providers, device manufacturers etc. can focus on certain sets of services (e.g., tele services between terminals) and rely on certain interfaces to other services (e.g., to the underlying bearer services).

4.4 Main reason is the forward error correction to mitigate transmission errors. Furthermore, bandwidth is needed for signalling, guard spaces.

4.5 See figure 4.4. Specifying all (or at least many) internal interfaces allows for a larger variety of vendors. As long as vendors stay with the standardised interfaces equipment of different vendors can be combined and network operators are not completely dependent from one manufacturer. However, reality often looks different and network operators often use only equipment from one or two vendor(s).

4.6 The MS contains all device related functions: device ID, coders/decoders, radio etc. The SIM contains subscriber related functions and data: authentication, PIN, user id etc. This separation helps changing phones while keeping personal data: users simply insert their SIM in a new mobile phone and can use, e.g., their personal phone book, PIN etc. Exceptions are so-called SIM locked phones – in this case a mobile phone accepts only a certain SIM. However, this is rather a marketing than technical reason. Besides the SIM also the mobile phone itself can store user-related data. Additional user-related data is stored in the VLR responsible for the location area a user is currently in and the HLR of the network operator the user has a contract with. User data is protected in several ways: authentication centres are protected parts of the HLR residing at the network operator. Inside the core network only temporary identifiers are used, data is encrypted over the air interface (weak, but still encrypted), and the content of the SIM is protected via a PIN (some cards destroy themselves after being attacked too many times). Localisation could be terminal assisted: the terminal could gather the current signal strength from all surrounding base stations. Furthermore, using the time of arrival helps calculating the distance. Reflection and attenuation makes the calculation more difficult.

4.7 GSM uses only two levels of hierarchy: Network operators store all user related information in the HLR and all information related to visitors within a certain location area in a VLR. Capacities of HLRs is up to some million customers, that of VLRs up to a million. I.e., within the location area a maximum of, e.g., one million users can be

active (registered). If many users move between location areas updates have to take place, i.e., the HLR always gets the information about the new VLR. These updates happen independently on the users' activity (data transmission, calls etc.). For standard scenarios – most users stay most of the time within their location area – the 2-level hierarchy works well. However, if, e.g., many tourists move frequently the updating process puts some load on the network as the HLR in the home network of the tourists always requires update information – probably around the globe. More levels of hierarchy could improve scalability but also raises complexity.

4.8 HSCSD still operates circuit switched as CSD does. It "simply" combines several connections. GPRS introduces a new paradigm in GSM, packet switching. Basically, the core network needs routers handling the packet stream. These routers (SGSN, GGSN) operate on IP and rely on the traditional GSM network for user localisation. Another new component located at the HLR is a registry for subscribed GPRS services. Furthermore, the system has to set up a context for each active user, account transmitted data, assign IP addresses etc.

4.9 Traditional GSM has cell diameters of up to 70 km, i.e., a user may have a maximum distance of 35 km to the base station. This limitation is not because of too strong attenuation, but because of the delay the signals experience. All signals must arrive synchronised at the base station, timing advance adjust the sending point (the further away a terminal is the earlier it has to send its data). With some tricks the diameter can be doubled. The capacity is limited by the number of channels * number of time slots – signalling overhead. The number of channels is operator and regulation dependent. The capacity is independent of the usage of GSM/CSD, HSCSD or GPRS – all three systems use the same basic frame structure and modulation. New modulation schemes can offer higher capacity, EDGE is an example. Furthermore, systems like GPRS offer different levels of error protection – this may increase user data rates under good propagation conditions, but does not increase the system capacity.

4.10 GSM uses SDM, FDM and TDM:

- SDM: Operators design the cell layout, place base stations and reuse frequencies according to certain cluster patterns.
- FDM: Regulation authorities assign channels to operators, operators assign channels to base stations, and base stations assign a certain channel to a terminal during data transmission.
- TDM: Base stations assign a time-slot or several time-slots to a terminal for transmission.

4.11 The BSS has to create a frame structure. Terminals listen into the medium, receive signals over broadcast channels and synchronise to the frame structure. Within each time-slot during transmission a midample further improves synchronisation. The terminal itself is responsible for precise synchronisation within the cell. This is very important in TDM systems as otherwise neighbouring data may be destroyed.

4.12 Examples for delays in packet transmission:

- CS: connection setup (some seconds), FEC coding/decoding and interleaving (about 100 ms), propagation delay (some ms)
- PS: channel access (depending on the current load), FEC coding/decoding and interleaving (about 100 ms), propagation delay and routing (some ms). Experiments show that packets in GPRS may experience heavy delays due to channel access delays: 200-500 ms for 128 byte packets, several seconds for 1-4 kbyte packets.

4.13 Besides problems due to interference, collisions in GSM systems can only occur during connection setup. Terminals have to access the base station suing a slotted Aloha scheme for the layer 2 signalling connection. During this connection attempt several terminals may collide and have to repeat the connection attempt. During data transmission or voice call no collision can occur. Data transmission in standard GSM (CSD) behaves just as voice calls. HSCSD has the additional problem of requesting

several channels. These may be occupied. However, this does not cause a collision but a simple denial of the connection request for several channels. Channel assignment and release is handled dynamically in GSM systems. For GPRS, too, data transmission can not cause a collision as the terminal wanting to transmit has to request time-slots firs. After the assignment of time-slots the terminal may access these slots without further collisions. Depending on the current load, not too many slots may be available; however, network operators try to offer at least one slot per cell for GPRS traffic to offer a minimum data rate.

4.14 GSM comprises many different channels for signalling control data. If no traffic ch channel (TCH) exists, an MS uses an SDCCH for signalling, e.g., authentication and registration data required prior to TCH establishment. TCH and SDCCH use an SACCH for signalling channel quality/signal strength. If a TCH exists and more signalling is required (e.g., during handover), an MS uses a FACCH, which is located in the time-slots otherwise used by the TCH.

4.15 The GSM system only stores the current location area for a user in the VLR. Each time a user changes the location area this change is reflected in the VLR. Additionally, periodic updates are possible. Roaming includes changing the network operator. This can happen within the same country (national roaming) or when going to another country (international roaming). The latter is the most common scenario as national roaming typically involves direct competitors. Prerequisite are roaming agreements between the different operators. The HLR always stores the current VLR for the user, no matter if inside the own or inside a foreign network. Precise localisation of users is performed during call setup only (paging within the location area).

4.16 Users of the GSM systems work with telephone numbers. That is all users should see. These phone numbers are completely independent of the current location of the user. The system itself needs some additional information; however, it must not reveal the identity of users. The international identification of users is done with the

IMSI (=country code + network code + subscriber ID). During operation within a location area, only a temporary identifier, the TMSI is needed. This hides the identity of a user. The TMSI is not forwarded to the HLR. These are already some examples for identifiers; however, GSM provides some more:

- *IMEI:* MS identification (like a serial number); consists of type approval code (centrally assigned), final assembly code, serial number, and spare (all three manufacturer assigned).
- *IMSI:* Subscriber identification, stored in the SIM. Consists of the mobile country code (3 digits, e.g., 262 for Germany), the mobile network code (2 digits, e.g., 01 for the German T-Mobile), and the mobile subscriber identification number (10 digits). The mobile network code together with the mobile subscriber identification number forms the national mobile subscriber identity.
- *MSISDN:* Mobile subscriber ISDN Number, i.e., the phone number, assigned to a subscriber, not a telephone! The MSISDN is public, not the IMSI nor the mapping MSISDN-IMSI. An MSISDN consists of the country code (up to 3 digits, e.g., 49 for Germany), the national destination code (typically 2 or 3 digits), and the subscriber number (up to 10 digits).
- *MSRN:* The mobile station roaming number is a temporarily assigned, location based ISDN number. The VLR assigns MSRNs and forwards them to the HLR/GMSC for call forwarding. Assignment happens either upon entering a new LA or upon request from the HLR (call-setup).
- *LAI:* The location area identity describes the LA of a network operator. It consists of a country code (3 digits), a mobile network code (2 digits), and a location area code (16 bit). The LAI is broadcasted on the BCCH for LA identification.
- *TMSI:* The VLR currently responsible for an MS can assign a 32 bit temporary mobile subscriber identity to an MS with a SIM. The tuple (TMSI, LAI) uniquely identifies a subscriber. Thus, for ongoing communication IMSI is replaced by (TMSI, LAI).

- *LMSI:* An additional local mobile subscriber identity (32 bit) can be used by the VLR/HLR for fast subscriber look-up.
- *CI:* Within a LA each cell has a unique cell identifier (16 bit). Thus, the tuple (LAI, CI) uniquely identifies a cell worldwide (global cell identity).
- *BSIC:* The base transceiver station identity code identifies base stations (6 bit) and consists of a 3 bit network colour code and a 3 bit base transceiver station colour code.
- All MSCs, VLRs and HLRs have unique ISDN numbers for identification.

4.17 The typical reason for a handover is a weaker signal from the current base station compared with a neighbouring base station. Another reason could be the current load situation: the network could decide to offload some users from a crowded cell. For the typical steps and types of handover see figures 4.11-4.13. For HSCSD to succeed the same resources are needed in the new cell as were available in the old one. I.e., there must be enough time-slots available to handle the same number of simultaneous connections. Otherwise the available bandwidth will decrease. Sure the probability of having several channels available is much lower than having a single channel. For GPRS data rates fluctuate anyway depending on the current load. The same happens during and after handover. Without pre-reservation neither HSCSD nor GPRS can give any QoS guarantees. There is not even a QoS guarantee for a voice call – if the next cell is already completely booked the connection will break upon entering this cell.

4.18 The first step is the authentication of the user against the SIM. This is done using a simple PIN. Then, the SIM authenticates itself against the GSM system. This second authentication is much stronger compared to the PIN. This is because the operator is not really interested in who is using the system as long as it is a valid and paying customer. Authentication with the system uses a challenge response scheme with a shared secret on the SIM and in the AuC. Neither the SIM nor the AuC will transmit

this secret over the air or reveal it to customers. Encryption only takes places between the MS and the BSS. GSM does not provide strong encryption end-to-end or MS to the gateway into the fixed network. System designers decided for over-the-air encryption only as they thought that the system itself is trustworthy. Thus, authentication of base stations against MSs was neglected, too. This opened ways to fake base stations. UMTS introduces full authentication of all components.

4.19 The classical data rate of GSM is 9.6 kbit/s. Using less FEC 14.4 kbit/s are available, too. These data rates are achievable using a single time-slot per frame in a certain channel. HSCSD combines several time-slots but leaves coding untouched. GPRS can dynamically use several time-slots per frame plus offers 4 different coding schemes that allow for higher data rates per slot. EDGE finally introduces another modulation scheme (PSK) in addition to GMSK, which offers even higher data rates under good propagation conditions. Only EDGE can really increase the capacity of a GSM cell. Independent of the coding and modulation schemes the complexity of handover signalling, handover delay and high delay due to coding/interleaving remain.

4.20 Real devices can (currently) not offer all data rates specified in the standards. While the standards in principle specify devices that use all 8 time-slots in both directions, real devices can often not send and receive at the same time. Furthermore, older devices even need some time to switch from sending into receiving mode, thus wasting another slot or even several slots. Additionally, current GPRS phones often do not offer all coding schemes.

4.21 The delay is specified between the MS and the exit point of the GPRS core network. The best average delay is 0.5 s. Assuming a data rate of 115.2 kbit/s (a common rate using serial adapters connected to the mobile phone) and a delay of 2*0.5 sec, 115.2 kbit = 14.4 kbyte are in transit. TCP was made for streaming larger amounts of data, i.e., file transfers etc. TCP allows for fair sharing of bandwidth as soon as it is in stable state. This requires the reception of acknowledgements, the adaptation of

sending windows and thresholds. However, if the whole transfer is 10 kbyte only, TCP either never gets an acknowledgement back during transmission to adapt sender characteristics (only if the initial sending window is large enough), or TCP wastes bandwidth by using a too small starting sending window (standard case). Real measurements with GPRS exhibit high latencies (examples are round trip times for different packet sizes, class 8 mobile phone): 0.8 s/64 byte, 1.4 s/128 byte, 2.2 s/1024 byte, 2.9 s/2048 byte, and 4.8 s/4096 byte. Additionally, measurements show high jitter. Under these conditions, TCP performs poorly. Chapter 9 lists several proposed changes to TCP (e.g., large initial sending window).

4.22 GPRS still needs the classical CS core for localisation, authentication etc. However, for data transfer the MSCs are not needed any more. The routers in the PS core (SGSN and GGSN) perform data forwarding (see figure 4.16).

4.23 DECT offers 120 full duplex channels, each with a standard rate of 32 kbit/s (unprotected). DECT applies TDM for structuring the frames and multiplexing users (24 slots per frame, typ. 12 up/12 down link). Furthermore, FDM is applied to increase capacity (several DECT cells at the same location, 10 channels). Users can also apply SDM by placing access points further apart. All the multiplexing schemes together result in very high capacities of the system, which is needed, e.g., in office buildings. Compared to GSM the system is simpler. Although data bases have been defined, too, typical DECT systems consist of a simple base station and several mobile devices. Most scenarios do not require complicated handover (although possible in DECT). Most systems furthermore do not need accounting and billing mechanisms as they are simply connected to the fixed phone network or a PBX.

4.24 Police, fire brigades, ambulances, disaster relief teams, public transportation authorities, taxi drivers, etc. are typical users of trunked radio systems. Trunked radio systems are attractive because of special features like very fast connection setup (sub second), group calls, paging, high robustness, cheap operation, reliable and fast messaging, and ad-hoc capabilities. Existing systems for these special purposes are

still often analogue systems operating on special frequencies without strong encryption. This makes it very difficult to cooperate for, e.g., fire brigades, the police and ambulances during disaster relief operations – the teams have to exchange equipment in order to be able to communicate. Trunked radio systems can be cheaper compared to GSM as they can have higher coverage with fewer base stations due to the lower expected load. Furthermore, complex billing and accounting mechanisms are quite often not needed.

4.25 Main features: higher and more flexible data rates, better voice quality due to new codecs, usage of CDMA (in almost all systems), operation at 2 GHz. Higher cell capacities and higher data rates are mainly achieved by more powerful modulation schemes, better codecs with higher compression rates for voice, CDMA as additional multiplexing scheme, and more powerful devices (more precise power adaptation, utilisation of multipath propagation, …). UMTS implements asymmetrical data rates and different data rates in the same direction via different spreading factors. As the chipping rate of UMTS is always constant, data rates depend on the spreading factor. The more the data is spread the lower the data rate is.

4.26 Currently, the situation is not absolutely clear as the different countries are in different stages implementing 3G systems. Right now no one believes in a common worldwide system, not even the same frequencies are available everywhere:

- Europe: After a much discussed licensing process (beauty contests and auctions) many operators are currently deploying 3G systems. Some operators already dropped out, some filed bankruptcy. All operators for 3G will use UMTS, in the beginning the UTRA/FDD mode only (no one knows when and if UTRA/TDD will be deployed). Although licensing did not prescribe the usage of UMTS, there were only a few operators thinking of different systems in the beginning. Start of the system was 2002, 50% of the population should have access to UMTS in 2005 (in Germany).

- Japan: Two different 3G systems are available in Japan. NTT DoCoMo uses a variation of UMTS in their W-CDMA system marketed as FOMA. KDDI deploys a cdma2000 system, which is 3G from the version 1xEV-DO on.
- China: While most 2G users today use GSM creating the biggest national market for this system, it may be speculated that UMTS will be a major 3G system in China, too, as this system can easily reuse the existing core network in its Release 99. The Chinese development TD-SCDMA was incorporated into UMTS (UTRA/TDD, slow chipping option, Release 4). However, it is currently not clear when and if this system will be deployed. There are also some cdma-operators in China which might opt for cdma2000.
- North America: The situation in the US and Canada is quite unclear. Already today many systems exist in parallel without a clear winner (compared to GSM in Europe). Furthermore, licensing of 3G spectrum takes a long time and the availability of spectrum is not clear yet. Thus, it could be the case that EDGE enhanced systems (TDMA and GSM) will be deployed offering higher data rates with EGPRS compared to today's networks. The cdma-operators will go for cdma2000.

4.27 OVSF offers only certain fixed data rates (certain multiples of 15 kbit/s). If users want to send with a data rate in-between the system either drops data (which can be recovered using FEC) or inserts dummy data. In the FDD mode adjusting the spreading factor is the only way for offering different data rates. TDD offers additionally the possibility of requesting more or less slots for up or downlink.

4.28 The spreading codes can always be the same in UTRA FDD to lower the system complexity. However, each UE has an individual scrambling code that is quasi-orthogonal to other scrambling codes. In UTRA TDD, scrambling is cell specific.

4.29 The important characteristic of combining/splitting is that it is never performed inside the (traditional) core network. I.e., the MSCs do not notice anything from the new possibilities offered by CDMA (reception of data via more than one base station).

Depending on the location of the handover (between two antennas at the same Node B, between two Node Bs, between two RNCs) the Node Bs or the serving RNC have to perform splitting/combination. The interface I_{ur} is needed to transfer data between RNCs for combination/splitting without any interaction with the CN. For CDMA receiving signals from different base stations looks like multipath propagation. The rake receivers can thus handle both. The handover is then as soft as a change in the strongest signals in a multipath scenario. TDMA/FDMA systems like GSM cannot do this because the currently used time-slot and/or frequency may not be available in the next cell.

4.30 The terminals have to measure and adapt their transmission power 1500 times per second in UTRA/FDD to achieve equal signal strength at the base station. In GSM this is no problem as it never happens that two stations send at the same time on the same frequency.

5. Satellite systems

5.1 The traditional application for satellites is the “big cable in the sky.” I.e., satellites interconnect distant locations. Today, this traditional usage for satellites is not dominant anymore. Thousands of fibres through all oceans connect all continents offering more capacity than currently needed. However, satellites are still required for TV/radio distribution and access to telecommunication networks at remote places, places with destroyed infrastructure, hostile environments etc.

5.2 The delay earth-GEO satellite-back to earth is always about 250 ms. This is very high compared to delays in fibre optics. Nothing can change this fact as (currently) the speed of light is the upper limit for the signal propagation speed and the distance of the GEOs is almost the circumference of the earth.

5.3 The inclination determines the coverage of the satellite. At an inclination of 0° the equator is covered. With a 90° inclination a satellite orbits over the poles. Geostationary satellites are only possible over the Equator, but then reception is poor

at higher latitudes. The elevation determines the signal quality. At an elevation of 0° reception is almost impossible. Typically, a signal has a usable quality starting from an elevation of 10°. Optimum signal quality can be achieved at 90°. High elevations are also required in urban or mountainous areas where buildings or mountains block signals from satellites with low elevation.

5.4 Characteristics, pros/cons of different orbits (see chapter 5 for further figures):

- GEO: Satellites seem to be pinned to the sky; pros: fixed antennas possible, wide area coverage, simpler system design; cons: long delays, high transmission power, low system capacity (difficult SDM), weak signals at high latitudes, and crowded positions over the equator.
- LEO: low orbiting satellites; pros: low delay, lower transmission power, inter-satellite routing; cons: high complexity, high system cost
- MEO: somewhere in-between GEO and MEO
- HEO: non-circular orbits; pros: higher capacity over certain points; cons: complex systems

5.5 Attenuation caused by the atmosphere, dust, rain, fog, snow, … Blocking of signals due to obstacles (buildings, mountains). The lower the elevation the longer is the way for the signals through the atmosphere. Without beam forming high output power is needed.

5.6 Classical satellites were simple amplifiers that amplified the incoming analogue signal and transmitted it again on a different frequency. The next step came with digital signals. Satellite could then work as repeater. This includes regeneration of the digital data and transmission of signals representation the received data without noise (compared to analogue amplifiers that also amplify noise). Many of today's satellites are repeaters. The next steps are switches/routers. Satellites can perform data forwarding functions depending on receiver addresses and can even route data through space from satellite to satellite.

5.7 Without any additional repeaters on earth, satellite phones only work outdoor (or close to windows). Satellite signals are typically too weak to penetrate roofs. Furthermore, satellite phones often require a line-of-sight even outdoor. Thus, skyscrapers blocking the LOS may block communication, too.

5.8 In order to stay synchronous with the earth's rotation, GEOs have to use the common orbit at 35786 km. Furthermore, the inclination must be 0°. This leads to satellites stringed on this orbit like stones on a thread. Additionally, the satellites should orbit above populated regions. Thus, areas above the equator looking towards Europe, America, Asia etc. are crowded. This is also the reason why all satellites must spare some propellant to catapult them out of the orbit after their lifetime. They must not block their position.

6. Broadcast systems

6.1 DAB and DVB both offer much higher data rates compared to 2G/3G networks. But they operate only unidirectional and bandwidth is shared (well, the capacity of a 2G/3G cell is shared, too). Thus, broadcast systems are good for distributing mass data relevant to many (in the best case all) users. Good examples are radio and TV, but also system updates, popular web content, news etc. Typically, it is to expensive to broadcast individual data. However, if broadcast bandwidth is available this is feasible, too. DAB/DVB can be complementary to 2G/3G systems. In particular if downloads are needed at higher relative speeds. Mobile phone systems have to lower their bandwidth dramatically at high speeds, while broadcast systems may still work at full bandwidth.

6.2 Examples are news, search engines, weather reports, big portals, i.e., web pages that are relevant to many users. But also within individual web pages common parts (commercials, video streams) could use broadcast systems, while the individual parts use mobile telecommunication systems.

6.3 If the location of a user is known to the system, LBSs may offer individual, location dependent services (next pizzeria, next ATM, cheapest bookstore in close proximity, gaming partners within a certain radius etc.). Depending on the current location, the LBS may program broadcast disks of broadcast providers for individual users or groups of users. If an LBS discovers a group of people standing in front of a museum, it could trigger a video stream on a DVB device showing pictures from the current exhibition.

7. Wireless LAN

7.1 Without further mechanisms mobility in WLANs is restricted to the coverage of a single access point. In order to support roaming additional inter access point protocols are needed. The access points have to inform each other about the current active stations within their coverage. This approach is only feasible for local areas, otherwise location registers etc. similar to GSM are required. The access points simply operate as transparent, self-learning bridges that need additional information to “forget” stations faster compared to the aging mechanisms in fixed network bridges. Station identification is based on MAC addresses. Roaming typically requires a switched layer-2-network.

7.2 Differences: coverage (GSM 70km cells, WLAN 100m), data rates (GSM 50 kbit/s, WLAN 50 Mbit/s), quality of service (WWAN voice/data rate, WLAN none/some with HiperLAN2), transmission power (powerful base stations for WWANs, some hundred mW for WLANs), operation (WWAN licensed, WLAN license exempt), administration (WWAN public operators, WLAN private), frequencies (WWAN many different national frequencies, WLAN almost common international ISM bands). Common characteristics: similar propagation characteristics, similar problems.

7.3 WLANs introduce the air interface which is very simple to eavesdrop. Thus, many WLAN standards introduce more or less strong encryption mechanisms. The most famous one, WEP, has been cracked soon after introduction. Furthermore, the most

prominent WLAN family, 802.11, does not provide powerful authentication mechanisms. New standards introduce more security (802.11i), however, users should always use an additional VPN on top of the WLAN to protect privacy and data integrity. WLANs following Bluetooth or HiperLAN2 offer more advanced security functions compared to 802.11.

7.4 All three standards offer ad-hoc functionality, although only Bluetooth was designed with the focus on ad-hoc networking. 802.11 heavily relies on an access point for many functions (e.g., power control, frequency selection, QoS in polling mode, access control etc.). Bluetooth on the other hand implements all functions in all nodes enabling all devices to set up a network. Main focus of HiperLAN2 is the infrastructure mode, too. Roughly, it can be said that 802.11 covers all standard office applications, Bluetooth focuses on inter-device connectivity, while HiperLAN2 was designed for QoS support (no products yet).

7.5 One reason for infrared is still cost – IR devices are very cheap and very simple to integrate. Another advantage is the simple protection from eavesdropping. Attackers can much more easily tap Bluetooth communication, incautious users even let their Bluetooth devices open for public access (simply scan for Bluetooth devices at public devices - many are detectable). IR communication is much more secure as the devices have to face each other (directed IR).

7.6 802.11 covers a whole family of WLAN standards. Depending on transmission technology, bandwidth etc. different PHY layers exist. They all share a common MAC. In order to adapt the different lower parts of the PHY layer a sublayer offers common functions required by MAC, e.g., carrier sensing. The HiperLAN family specifies several PHY layers. However, currently it seems that only HiperLAN2 has a chance to survive – this standard specifies only one PHY layer. All Bluetooth systems use the same layers.

7.7 All systems save power by periodic sleep functions. In particular Bluetooth systems offer several low power modes as they are typically battery operated. Negative effects

of power saving are the increased latency for spontaneous transmissions – the devices have to wake-up first. Thus, the shorter access delay should be the less power a device can save. Furthermore, high data rates require high power. If the periodic sleep function is not synchronised with, e.g., periodic data transfer heavy jitter will result.

7.8 802.11 does not offer QoS in the ad-hoc mode as it needs the access point for polling. HiperLAN2, on the contrary, establishes a central controller for the ad-hoc mode (called direct mode), which controls QoS. Bluetooth always works ad-hoc, well, a master controls up to seven slaves and, thus, forms an ad-hoc access point. Bluetooth therefore can offer QoS in its ad-hoc mode. QoS in Bluetooth is provided by periodic polling through the master. This guarantees certain data rates and access latencies. HiperLAN2 can give hard QoS guarantees as it controls access latency, bandwidth etc. After a master has been found, Bluetooth can give hard guarantees for SCO connections. 802.11 can give guarantees if no contention phase is allowed (polling only). As soon as there is a contention phase, the system cannot guarantee access latencies.

7.9 802.11 uses the MACA mechanism sending RTS/CTS to solve the hidden terminal problem. For HiperLAN2 this problem does not exist as the access point controls all medium access. If a terminal is hidden it cannot communicate at all and, thus, does not interfere. In Bluetooth, too, are no hidden terminals as the master controls all visible slaves. If a terminal does not see the master it cannot participate in communication. If this terminal sends anyway it will not interfere as this terminal then acts as master with a different hopping sequence.

7.10 802.11 implements a back-off mechanism that tries to offer fair access to the medium in the standard case (no polling from the access point). If all systems behave well this mechanism gives a fair share of the overall bandwidth to all stations. In HiperLAN2 and Bluetooth medium access is controlled by an access point or master, respectively. Fairness then depends on these special nodes, which also decide upon

the waiting time of a packet when it will be transmitted. In 802.11 the waiting time directly influences the chances for transmission in the next contention cycle.

7.11 802.11 offers immediate acknowledgement, Bluetooth implements different ARQ and FEC schemes, as well as while HiperLAN2 does.

7.12 During polling, there are no collisions on the MAC layers of HiperLAN2 and Bluetooth as the access point/master controls the medium. However, in order to access the access point, nodes may transmit during a random access phase in HiperLAN2 (random channel with feedback from the access point). At this point collisions may occur on the MAC layer. For 802.11 collisions on the MAC layer are nothing unusual. The MAC algorithm with back-off solves this problem. Collisions on the PHY layer may occur in Bluetooth only if another piconet randomly jumps to the same frequency at the same time. This will destroy data for this time-slot. In HiperLAN2 different networks are separated in frequency, thus there should be not collisions besides the above mentioned during the random access phase. In 802.11 networks MAC collisions are also collisions at the PHY layer. Important packets in 802.11 have higher priorities implemented via shorter waiting times (SIFS, PIFS).

7.13 802.11 has the lowest overhead as each node can simply access the medium if it is free. Thus, 802.11 offers the shortest access latency at zero load and still low latency at light load. The system breaks down at high load as then only collisions will occur and no station is able to send anything. Therefore, 802.11 has a rather soft capacity. HiperLAN2 and Bluetooth require some kind of connection setup. This increases access latency – even if the load is light or zero. As soon as a connection exists, the quality and access latency is almost independent of the load. Both systems can be loaded to the maximum without a system breakdown. For Bluetooth this is true in a piconet, not within scatternets. Scalability is low in general (8 nodes within a piconet). For HiperLAN the number of maximum nodes depends on the QoS requirements. In 802.11 networks the number of supported nodes depends on the traffic patterns.

7.14 HiperLAN 2 and 802.11 need an IAPP, Bluetooth does not support roaming at all. Nodes changing piconets have to resynchronise to the new piconet, there is no signalling between masters for roaming nodes. Typically, inter access point protocols are available in infrastructure networks only (there could be something like a master-to-master protocol in Bluetooth…). For ad—hoc networks the overhead would be too much. Roaming support is typically via self-learning bridges exchanging their filtering databases (which MAC address is visible at which bridge). HiperLAN 2 additionally provides support for key exchange during roaming, sector/radio/network handover etc.

7.15 Forwarding data in Bluetooth between piconets require a node jumping back and forth between these piconets. This also requires authentication in both networks, nodes that are (almost) always active and synchronous clocks if the master jumps into another piconet. If the master jumps away all network traffic in the piconet stops, all slaves have to wait until the master returns. All hopping sequences must stay synchronous during that time. Up to now not many devices are capable of forming scatternets with nodes jumping back and forth.

7.16 When the development of WATM started much hype came with ATM in the fixed networks. ATM was seen as the big unifying technology handling all different types of traffic with QoS. Well, in principle this is still true, however, it turned out that this technology is much too complicated for many applications (but it is still dominant in WANs). ATM offers hard QoS, end-to-end. The Internet of today do not offer QoS – most QoS architectures failed or did not succeed until now (Integrated Services, Differentiated Services). However, there were almost no applications that could use the QoS offered by ATM at the desktop. Most applications of today can adapt to the varying quality of the Internet. WATM never made it, but many of the ideas survived as people involved in WATM also developed, e.g., UMTS, HiperLAN2 etc. (and again it is questionable if HiperLAN2 will make it into any product).

8. Mobile network layer

8.1 See the introduction of ch. 8. Main problems are the high dynamicity – Internet routing protocols (like the standard fixed network routing protocols in classical phone networks) have never been designed for roaming nodes, not to mention mobile routers. Without additional functions addressing fails, nodes would use topological incorrect addresses etc. Standard routing protocols from the Internet (e.g., OSPF within autonomous systems, BGP between these systems) can handle link and router failures, overload situations etc. if they do not happen too frequently.

8.2 Quick solutions could be the permanent adaptation of the current IP address of a mobile node depending on the current location. But then no correspondent node can find the mobile node (or a lot of signalling this current IP address would be necessary). Alternatively, all routers could change routing table to reflect the current location of the mobile node. This obviously does neither scale nor is it secure – changing routing entries destabilises the whole network.

8.3 See 8.1.1.2. Although mobile IP tries to provide transparency of mobility it cannot hide, e.g., additional delay due to larger distances or lowered QoS due to inferior connections to the mobile node. However, mobility is transparent if only best-effort transmission is considered. Scalability, too, is a problem as soon as many nodes move between subnets. Mobile IP causes a big overhead due to registration messages. This is one of the reasons for micro mobility supporting approaches. Security is also problematic, as topological incorrect addresses do not work together with firewalls and route optimisation reveals location.

8.4 See figure 8.2. Encapsulation is required between the HA and the COA, which could be located at an FA or at the MN. This is needed to make mobility transparent – the inner data packet should not notice data transfer through the tunnel, thus TTL remains untouched.

8.5 See figure 8.4. Layer 2 registration is handled by, e.g., the WLAN or fixed LAN.

8.6 Assume that the MN detects a stronger signal from an access point compared to the current signal. If available on layer 2 the MN could detach from the old access point after attaching to the new one. It would first set-up a layer two association and listen for agent advertisements. Alternatively, it could send agent solicitations. After receiving the advertisement and attaching to a new FA authentication could start. Concurrently, the FA could inform the old FA about the node. See figure 8.13 plus layer 2, e.g., 8.3.5.3 for 802.11.

8.7 If MN_a and MN_b are both in foreign networks attached to FA_a and FA_b the packet flow is as follows. MN_a sends packets to MN_b via the Internet to HA_b (actually, MN_a sends to MN_b's address, the packets are only intercepted by HA_b). HA_b encapsulates the packets to FA_b, which then forwards the packets to MN_b. If reverse tunnelling is required, the packet flow is as follows: MN_a sends its packets via FA_a through the reverse tunnel via HA_a and the Internet to HA_b. HA_b then forwards the packets through the tunnel to FA_b, which in turn forwards the packets to MN_b.

8.8 Tunnelling simply means that a packet is encapsulated at tunnel entry and decapsulated at tunnel exit. The packet is thus payload of the outer packet inside the tunnel. IP-in-IP encapsulation is the simple case of using IP for encapsulating other IP packets. This is simple because all devices already know how to insert payload into an IP packet. Bandwidth is wasted by transferring the same field several times. Minimal encapsulation tries to avoid this waste of bandwidth, however, it cannot be used in case of fragmentation. GRE is a more general scheme, not only for IP traffic but also, e.g., encapsulation of Ethernet packets into IP packets. Additionally, it may control the level of encapsulation. Several versions exist.

8.9 Triangular routing via CN-HA-FA-MN is inefficient. One optimisation is the binding update at the CN. A CN can enter the COA of a MN in its routing table. This lets the CN directly send its data to the MN. This solution reveals the current location of the MN and is not transparent anymore (the CN now knows that the MN is mobile, furthermore, it knows the location via the COA).

8.10 Many mobility supporting function are already integrated in IPv6. An explicit FA is not needed any more, all routers are capable of agent advertisements, tunnelling, forwarding of data, setting up security associations. Authentication is built-in as well as optimisation functions.

8.11 Mobile IP does not increase security compared to IP, on the contrary. The only additional security related function is the authentication of MN and HA. However, if MN and HA, together, want to attack an FA, nothing can prevent them. Firewalls and mobile IP do not really go together. Either reverse tunnelling or tunnelling in general drills a hole in the firewall or MNs can not operate in foreign networks. The firewall has to be integrated into the security solution. IP does not support QoS. If QoS supporting approaches like DiffServ or IntServ are used, new functions are needed for mobile IP to support QoS during and after handover. Furthermore, packets requiring certain QoS must be treated according to these requirements also inside the tunnel.

8.12 DHCP is a mechanism for configuring nodes. Parameters acquired via DHCP are, e.g., IP address, default gateway, DNS server, subnet mask etc. Without DHCP all parameters must be configured manually. A DHCP server provides DHCP information, a relay can forward data into different LANs.

8.13 If users only want to access other server, e.g., for WWW browsing, mobile IP is not needed. After obtaining a new IP address via DHCP a node can act as client. However, as soon as a node wants to offer a service, it should keep its IP address. Otherwise it is difficult to find it or other additional mechanisms (DDNS) are required to map, e.g., a node name to the node's address. DHCP can act as source of COAs in mobile IP.

8.14 Ad-hoc networks in general do not require an infrastructure to operate (they can be connected to an infrastructure). Multi-hop ad-hoc networks additionally do not require that all nodes can receive each other. Nodes may forward transmissions for other nodes. Advantages are the lower required transmission power (it's just like whispering

into the neighbour's ear instead of shouting out loud) and the increased robustness (failure of single nodes can be tolerated).

8.15 Routing is complicated because of frequent topology changes, different capabilities of the nodes, varying propagation characteristics. Furthermore, no central instance can support routing.

8.16 Both algorithms assume a more or less stable network – at least changes are very infrequent compared to routing data exchange. Furthermore, both algorithms establish routing tables independent of the necessity for communication. This not only causes a lot of unnecessary bandwidth, but may render useless if the topology changed right before communication should take place.

8.17 AODV is a reactive protocol. Route calculation is only performed if necessary. This improves scalability under light load, but causes a higher initial latency.

8.18 DSR separates finding a route and keeping the route working. If no communication is required DSR does not try to establish any route. As soon as a route is needed, DSR tries to find one. As long as the communication keeps on going DSR tries to maintain the route. In fixed networks routes are always calculated in advance.

8.19 Most algorithms fail if the links are asymmetric (up to the extreme case of unidirectional links). Think of DSR – the algorithm states that the receiver simple sends the packet collecting routers on the way between source and destination back to the source by choosing the routers in the reverse order. But what is some reverse links do not exist? Then DSR has to find a way the other way round, too. Now source and destination both got a way – but in the wrong direction! Somehow this information must reach the other side – without a route quite difficult (broadcast is always a solution…).

8.20 Mobile IP causes too much overhead during registration if used for very mobile nodes (nodes, changing networks quite frequently). Furthermore, all registration messages cross the Internet from the foreign to the home network (plus registrations reveal the

current location). Micro-mobility supporting approaches basically insert another layer of hierarchy to offload some of the complexity from the HA (compare with HLR, VLR).

8.21 Location information may help routing (geo routing) by optimising the route. If one already knows the location it is simpler to choose the right router towards the destination. However, again privacy problems may arise. Not too many people want to reveal their location to everyone.

8.22 For fast moving cars in cities efficient routing is very difficult as the topology changes very fast. Flooding with some optimisations may be the only way to go. However, if the cars are on a highway, it is simpler: cars typically form clusters per direction. On car of the cluster could be the cluster head, all other cars route via this car. Routing can go along the lanes of the highway.

9. Mobile transport layer

9.1 Packet loss due to transmission errors: Relatively low in fixed networks (10^{-10}-10^{-12}), quite high in wireless networks (10^{-2}-10^{-4})/large variation/typically compensated by FEC/ARQ; packet loss due to congestion: no difference between fixed and wireless networks; packet loss due to mobility: happens only in mobile networks…

9.2 TCP typically assumes congestion in case of packet loss. This is the correct assumption in fixed networks, not in wireless networks (transmission errors due to interference and mobility are more frequent). In wired networks TCP helps stabilising the Internet, in wireless and mobile networks standard TCP performs very poorly.

9.3 If only some users replaced TCP by UDP they might experience higher throughput. However, the missing congestion avoidance mechanisms would soon lead to huge packet loss in the Internet. Additionally, reliability has to be added as UDP does not guarantee packet transmission. A lot of research exist for TCP friendly protocols, reliable UDP etc.

9.4 I-TCP splits the connection into two parts – a wired/fixed and a wireless/mobile part. This isolates problems on the wireless link from the fixed network. However, this also

requires that intermediate systems are able to look into IP packets to split the connection. This prevents the usage of IPsec – end-to-end security and I-TCP (or proxy solutions in general) do not go together.

9.5 See figure 8.2 for the packet flow. TCP does not directly interact with IP as mobile IP keeps mobility transparent. TCP may only experience higher loss rates during handover. Mobile IP handles the handover; old FAs may or may not forward packets. If acknowledgements arrive too late, TCP assumes congestion, goes into congestion avoidance and enters slow-start. However, slow-start is absolutely counterproductive. Sending with the same data rate as before would make sense.

9.6 Compare with figure 9.2. FA, CN, HA, MH should work as Mobile IP specifies. Without any PEP TCP would experience packet loss due to the change of the subnet if the old FA does not forward packets. If PEPs are used the old PEP must transfer the whole state (buffers for retransmissions, sockets, …) to the new PEP. The CN and the MH should not notice the existence of PEPs. One place to put a PEP is the FA. However, the PEP could also be located at the edge of the fixed network. PEPs work on layer 4 (in this example), while the Mobile IP components work on layer 3 – they might interact, but they do not have to.

9.7 Using end-to-end encryption prohibits the use of any proxy schemes – unless the proxy is included in the security association. This is quite often not possible as the foreign network together with the proxy belongs to another organisation. As soon as IPsec with encryption is used, no proxy can look inside the packet and examine the TCP header for further processing.

9.8 Selective retransmission is always a good idea. Most of the other optimisations exhibit drawbacks: compare with 9.3. There is no single solution and even the standards/drafts are inconsistent with each other.

9.9 First of all the tricky part. Error rates on links, as given in the question, are always bit error rates. Under the assumption that these errors are independent (and only under this assumption!), the packet loss probability p used in the formulae can be calculated

as: p = 1 - ((1-bit error rate)$^{\text{packet size}}$). Using this formula, you can calculate the packet loss rates (this ignores all FEC and ARQ efforts!).

- Fixed network: BER = 10^{-10}, MSS = 1000 byte = 8000 bit, thus the packet loss rate p = 1 - ((1-10^{-10})^{8000}) ≈ 8*10^{-7}. RTT = 20 ms: Using the simple formula, this yields a max. bandwidth of 0.93 * 8000 / (0.02 * √(8*10^{-7})) bit/s ≈ 416 Mbit/s.
- WLAN: The same calculation with the WLAN error rate 10^{-3} and additional 2 ms delay results in a packet loss rate of 0.99966 and a bandwidth of 0.93 * 8000 / (0.022 * √0.99966) bit/s ≈ 338 kbit/s. This is a good example showing why big packets cause problems in WLANs – and why FEC/ARQ is definitively needed… Real life throughput in WLANs is about 6 Mbit/s for 802.11b WLANs (if there are no other users).
- GPRS: Using GPRS with an additional 2 s RTT and a BER of 10^{-7} (i.e., a packet loss rate of 8*10^{-4}) results in only 0.93 * 8000 / (2.02 * √(8*10^{-4})) bit/s ≈ 130 kbit/s. Well, currently GPRS offers only 50 kbit/s, but that is a limitation the simple formula does not take into account.
- In practice, the performance depends very much on the error correction capabilities of the underlying layers. If FEC and ARQ on layer 2 do a good job, TCP will not notice much from the higher error rate. However, the delay introduced by ARQ and interleaving will decrease bandwidth. Additionally, the slow start mechanisms must be considered for short living connections. Nevertheless, it is easy to see from these simple calculations that offering higher data rates, e.g., for GPRS, does not necessarily result in higher data rates for a customer using TCP.

9.10 No matter what the estimation is, it is not possible to get a TCP bandwidth higher than the link speed. The link speed itself does not appear in the formula as attempting to send faster than the slowest link in the path causes the queue to grow at the transmitter driving the bottleneck. This increases the RTT, which in turn reduces the achievable throughput.

10. Support for mobility

10.1 It is simply too expensive to maintain strong consistency. Continuous updates require permanent connectivity, without connectivity all access must be blocked. Alternatives always include weakening the strong consistency. Many schemes include periodic updates, reintegration schemes or, if nothing else works, manual reintegration.

10.2 Either the file systems together with the computer crash or the computer simply indicates that the directory is not accessible. Just try it with different systems (please save everything before).

10.3 No state means no complex state management. Breaking connections does not matter; all necessary state is transferred with the next request. However, state is useful if overhead should be avoided or users should be enabled to resume sessions. Today, all state necessary for HTTP arrives together with the result of the get request. Long term state can be stored in cookies – the state exists as long as the cookie exists. Short term state is stored in the browser. In particular HTTP/1.1 allows for browsers that utilise the history of a session (partial transfer of content, setting of languages, cache handling technologies).

10.4 HTTP is text oriented and human readable. This makes parsing for humans quite simple, however, it wastes bandwidth compared to binary representations. Using HTTP/1.0 additionally wastes bandwidth as each request uses a separate TCP connection. This requires connection setup, data transfer, and connection release for each simple element on a web page. HTTP/1.1 uses persistent connections, i.e., one TCP connection can transfer several requests.

10.5 The closer a cache is located to a browser the better it can serve requests with minimum delay. Furthermore, if caches are located within mobile devices, many request can be handled locally thus avoiding unnecessary transmission over wireless links. If a cache is located at the border of the fixed network towards the wireless

network the cache can isolate the fixed network from problems in the wireless network. Caches in the fixed network may also follow the user.

10.6 Caches can only store content that does not change over time. However, today's web pages contain many individualised components: counters, advertisements, browser adapted content etc. Furthermore, content provides often want users to directly access content from the server to establish usage profiles. Caches without additional mechanisms make access counters and profiling useless. If the client is mobile caches must follow the client.

10.7 HTML today is not only used to describe content but to define layout. The original idea of HTML was to support the description of text structure (headings, bullets, etc.). Now HTML is used for formatting purposes. Most screen designs assume 1024x768 pixels, true colour. Wireless devices still have low resolution displays, 4096 colours, 320x240 pixels etc. Therefore, many pages do not fit onto the small displays. Solutions include downscaling of pictures, content extraction or the description of the same content with special languages (WML, cHTML). Many pages additionally contain content that requires special plug-ins: flash, 3D animations, streaming media etc. Typically they simply will not work on many mobile devices.

10.8 Caching, content transformation, picture downscaling, content extraction, textual descriptions of pictures. Many of the proposed solutions during the nineties were proprietary. WAP is the first standardised common solution supported by many network providers and device manufacturers. It is a different story why WAP was no success from the beginning (wrong marketing, wrong underlying transport system).

10.9 Proxies. Proxies can be located at different places – see the figures – and proxies can even be divided into two halves with special protocols between those halves. Proxies behave like clients towards servers, like servers towards clients. Good locations for proxies are close to the mobile/wireless user, but still in the fixed network. Examples could be foreign agents in mobile IP or routers in a GPRS network.

10.10 See section 10.3. General goals are: very efficient data transfer (binary representation), avoidance of redundant data transfer (stateful protocols), support of heterogeneous, simple devices (WML), access to telephony functions (WTAI), built-in security (WTLS), support of almost all transport platforms.

10.11 Telephony applications are supported via special interfaces in WAP 1.x (WTAI). Special gateways are required to access Internet content. WAP 2.0 combines Internet protocols with WAP 1.x. This allows direct access of Internet content.

10.12 WDP is not a fixed protocol. The interface to WDP is specified, WDP itself depends on the underlying network. The WDP interface provides an unreliable datagram transfer. If the underlying network already provides IP transfer WDP is simply the UDP protocol known from the Internet. Thus, there is no SAP WDP could use.

10.13 Not all mobile phone networks provide the same level of security. GSM, for example, provides only encryption over the air. WAP adds security from the end device to the WAP gateway. But this is also a problem. WAP does not guarantee end-to-end security. The security relation is broken at the gateway. This is the reason for banks to implement their additional security functions. This is also a difference to the usage of SSL/TLS in the Internet.

10.14 Advantage: users can control the acknowledgement process, users may want to know if something went wrong, sometimes it is also possible to slow down a sender by inserting artificial delays in the acknowledgement process, the acknowledgement of a user is “stronger” as it shows the sender that the intended receiver and not the WTP process actually got the message. Disadvantages: users have to interact, this may take some more time. Classical transactional services typically benefit from user acknowledgements, for most push service user acknowledgements are not necessary, still WTP acknowledgements can improve reliability.

10.15 WSP/B together with a class 2 transaction would be a good choice for standard web request/response schemes. The web expects a reliable protocol (that is why typically TCP is used) and works with transactions.

10.16 WSP offers session management, capability negotiation, push and pull, asynchronous request, and efficient content encoding. Some of these features also come with HTTP/1.1. However, in mobile and wireless environments efficient coding directly results in cheaper web usage.

10.17 User experience is typically much better if something happens on the screen. Waiting for all responses to arrive in order might cause too long pauses irritating users. As soon as a response arrives, the browser should display it to signal progress. Additionally, quite often users can continue with browsing although only parts of the page are visible.

10.18 Connectionless services typically have a lower overhead compared to connection setup, data transfer and release. Compared to a pure datagram service the connectionless session service additionally offers transaction identifiers and functions comparable to those of the other WSP services.

10.19 Besides languages and content formats the WAE defines gateways between clients and servers. As mobile devices can often not use the standard formats and protocols of servers (TCP, HTTP, SSL etc.), gateways translate between the classical fixed and the new mobile and wireless world.

10.20 WML offers only a few formatting instructions. It rather defines the intention of the author of a page. If the device should present data to a user this could be done via text of synthesised voice. This approach is more flexible compared to HTML relying on powerful displays. Sure the difference between handheld devices and PCs continuously shrink. Additionally, common commands are binary encoded. Instead of transferring text strings like “http://www.” a single byte can express the same.

10.21 Scripting can help to reduce traffic by checking input on the mobile device. Without scripting support the device must transfer all input for checking to a server. Furthermore, scripting can access many device functions.

10.22 Call indication, call accept, call setup, … WTAI offers special URLs and WTAscript functions for telephony features. A URL can now setup a call, WTAscript functions can change phone book entries etc.

10.23 WTA servers can much better control QoS as they typically belong to the mobile network operator. Standard servers in the Internet might experience many problems providing QoS – there is not even a widespread QoS architecture in the Internet. In principal there are many places for WTA servers: the operator's network, other operators' networks, even the Internet (but then with the above mentioned QoS problems).

10.24 Pushs are useful for indicating unexpected events. Pulling would waste bandwidth and require too much energy from the mobile device. The difference between SI and SL is that in the case of SL the client's user agent decides when to submit the URI (this then is a pull, but not noticed by the user). SIs are indicated and it is up to the user to use the service.

10.25 WAP 1.x was created by a consortium backed by network operators and many device manufacturers while i-mode is a proprietary development by NTT DoCoMo/Japan. I-mode was used from the beginning over packet oriented transport plus it comprises a certain business model: content providers get about 80% of the revenue a customer generates, the network operator handles billing. WAP had problems in its early days as it was marketed as "Internet on the mobile phone", which it is not. Additionally, it started on a connection oriented transport system. However, web usage is highly interactive. These two facts caused the failure of WAP in its early days. Transferring the success of i-mode to other countries is not easy. NTT DoCoMo tried this in Europe but did not have the same success. Reasons are the PC penetration in Europe: many people have already fast Internet connections. Furthermore, not all providers chose i-mode. Other services, such as MMS, attract many people.

10.26 Synchronisation is of major importance for many applications. Several incompatible approaches exist (even applications have their own mechanisms). However, the basic

problems of synchronisation are still unsolved: how to synchronise two updated objects? Without application oriented knowledge synchronisation is not possible.

10.27 WAP 2.0 includes i-mode and Internet components. The development was heavily influenced by the success of Internet applications and the ever increasing power of mobile devices.

10.28 See figure 10.38.

- WAP 1.x stack: This stack supports all “classical” WAP phones and applications. As discussed in this section, there are many reasons for session services and more efficient transactional services. Thus, this stack will remain a useful part of WAP.
- WAP with profiled TCP: This i-mode like scenario offers optimised HTTP and TCP. This might be more efficient then using “pure” Internet solutions, but reqires changes in HTTP and TCP – and the proxy to translate.
- WAP with TLS tunnelling: If end-to-end security is a must, the architecture must not break the connection. Therefore this stack offers end-to-end TLS, but can still benefit from an optimised TCP.
- WAP direct: If the devices are powerful enough and delays are not too high, the standard Internet protocol stack can be used. This would be the simplest solution, which does not require any special WAP protocols any more. While the devices might be powerful enough in the future, the delay problems will remain.