DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

COMPUTER COMMUNICATION TECHNIQUES IN VOICE COMMUNICATIONS

by
Eren DENİZ

August, 2005
İZMİR

COMPUTER COMMUNICATION TECHNIQUES IN VOICE COMMUNICATIONS

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfilment of the Requirements for the Degree of Master of Science in Electrical and Electronics Engineering

by

Eren DENİZ

August, 2005

İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"COMPUTER COMMUNICATION TECHNIQUES IN VOICE COMMUNICATIONS"** completed by **Eren DENİZ** under supervision of **ASSOC. PROF. DR. ZAFER DİCLE** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc.Prof Dr. Zafer DICLE
Supervisor

(Committee Member)

(Committee Member)

Prof.Dr. Cahit HELVACI
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

Firstly, I would like to give my thanks to my supervisor, Assoc. Prof. Dr. Zafer Dicle for his guidance along the fulfilment of this project.

I wish to express my sincere appreciation to my friends Utku Ergül, Tolga Narbay and Uğur Gül for their assistance during the course of this thesis.

Eren DENİZ

COMPUTER COMMUNICATION TECHNIQUES IN VOICE COMMUNICATIONS

ABSTRACT

In the near future, if you make a telephone call, it is more than likely that it would be over the Internet or some other packet network. The aim of this master thesis is to induce the technologies the future mobile network will consist of. In order to do this the future prospects of the techniques for mobile data communications as well as network architectural issues are analysed.

The thesis aims to answer the core issues: To compare the techniques for mobile data communication and to answer how to organize the network.

In particular signalling protocols like H.323 or the Session Initiation Protocol (SIP), and transport protocols such as the Real-Time Transport Protocol (RTP) are explained in detail. An overview of 802.11 and IEEE 802.16 standards and technologies are presented.

Keywords : VoIPoW, WiMAX, IP Telephony, Wireless Networks

BİLGİSAYAR AĞLARI ÜZERİNDEN SES İLETİŞİMİ UYGULAMALARI

ÖZ

Yakın bir gelecekte tüm “telefon” görüşmelerinin, İnternet veya paket ağları üzerinden yapılacağını rahatlıkla söyleyebiliriz. Bu çalışma da nasıl bir teknoloji ve altyapıyla bu varsayımda bulunabildiğimizi açıklamayı amaçlamaktadır.

Bu tez ses iletimini bilgisayar ağları üzerinden sağlamaya yönelik tüm çalışmaların yüzleşmek zorunda olduğu sorunları ve mevcut çözümlerin karşılaştırmalı analizlerini vermektedir. Kablosuz ağların ve bilgisayar ağları üzerinden ses iletiminin prensipleri incelenmekte ve ikisini biraraya getirmeye yönelik çalışmalar irdelenmektedir.

Başlıca sinyalleşme protokolleri olan H.323 ve SIP ile 802.11 ve 802.16 standartları tanıtılmış, birbirlerine olan üstünlükleri ve zayıf yönleri belirlenmiş ve kablosuz ağlar üzerinden ses iletimine uygunlukları araştırılmıştır.

Keywords : Internet protokolü üzerinden ses iletişimi, WiMAX, Kablosuz Ağlar

CONTENTS

Page

CHAPTER ONE
INTRODUCTION

1. Introduction

The vast majority of information exchanged over the public telecommunication network has been voice. The present voice communication networks, public telephone and Integrated Services Digital Network (ISDN) networks utilize digital technology via circuit switching. Circuit switching establishes a dedicated path (circuit) between the source and destination. This environment provides fixed bandwidth and short and controlled latency (delay). It provides satisfactory quality service and does not require a complicated encoding algorithm. The capacity of the circuit, however, is not shared by other users, thereby hindering the system's overall efficiency.

Packet switched networks such as Internet had been developing very fast in the past decades. The advantages of packet switched networks, such as efficiency and flexibility, make them eventually become the terminator of traditional circuit switch networks, i.e. Public Switch Telephone Network (PSTN). Generally, Voice Over Packet Services are the real time delivery of packetized voice traffic across packet switched networks such as Internet. It provides economical communication expense and suitable speech quality compared with traditional telephone networks.

Recently, wireless/mobile communication has been growing rapidly and providing more and more convenient services. It's not a surprise that there's a great demand to add voice services to wireless IP networks and wireless handsets. In this thesis, we will define what Voice Over Packet Network is and how do we use it over Wireless Networks.

1.1. Voice Over Packet Networks

IP, ATM and frame relay networks are considered candidates for a backbone supporting integrated voice and data applications. However, the first two are seen to be the hot favorites for wide area networks and internetworks. Frame relay is a better candidate for smaller networks. First we describe the technologies in brief. Then we compare voice over ATM and IP and state why we feel the latter is the preferred technology.

1.1.1. Frame Relay

Frame relay is a protocol standard for LAN internetworking which provides a fast and efficient means of transmitting information from a user device to the bridges and routers. It is a service for people who want an absolute bare-bones connection-oriented way to move bits at a reasonable speed and low cost. Its existence is due to changes in technology over the years. It is a direct evolution of the traditional packet-switching technology, which used complex protocols and a great deal of overhead for error detection, correction and flow control, as the user terminals were not able to do so themselves (Davidson & Peters, 2000). The situation has changed radically. Leased lines and now fast, digital and considerably reliable. The terminals now boast of superior processing power, capabilities at much lower costs. This suggests the use of simple protocols for data link, with most of the work done by the terminals rather than the network.

Frame relay could be understood as a basic virtual leased line. The customer leases a PVC (permanent virtual circuit) between two points and can then send frames of up to 1600 bytes each between them. It is also possible to lease PVCs between a given site and multiple other sites. Each frame carries a 10-bit address called the DLCI (Data Link Connection Identifier), which identifies the virtual circuit to be used. The difference between frame relay service and a permanent leased line is that the latter permits the customer to send data all day long at maximum data rate. For the frame relay virtual line, data bursts may be sent at the

maximum rate with the long-term average data rate being within a permissible value. That is exactly the reason why frame relay is cheaper than leased lines. The standard frame relay speed is around 1.5 Mbps.

Frame relay provides a minimum service. It is primarily a way to determine the start and end of a frame; source and destination address and detect errors. It does not provide flow control or acknowledgement service. If a frame is corrupted it is discarded. How the data is recovered is up to the higher service layers at the end terminals. This is exactly the kind of service real-time traffic, such as voice, needs. However, the success of frame relay is incumbent on an inherent reliability of the transport network.

As can be deduced for an Internetwork, frame relay service proves inadequate for voice transport. However, it is ideal for LANs (Local Area Networks) and medium sized MANs (Metropolitan Area Networks).

1.1.2. Asynchronous Transfer Mode (ATM)

The phrase ‘Voice over ATM’ has two aspects to it. We can either transport voice packets (cells) directly over the ATM architecture or tunneling IP packets over base ATM transport layers. The latter option is out of the question since it goes leads to inefficiency due to excess protocol overhead. This goes against the very reason why ATM was developed: An efficient, integrating, multiservice architecture. Unlike frame relay or IP, ATM is not just a protocol. It is not confined to a single layer in some architecture. It is an architecture itself. It is part of the envisioned B-ISDN (Broadband ISDN).

ATM is so called because it is not synchronous i.e. tied to a master clock. The basic idea behind ATM is to transmit all information in small, fixed size packets called cells which are 53 bytes long: 5 bytes of header and 48 bytes of payload, in our case voice (Handley et al., 1999). ATM service is also called cell relay.

Cell switching is highly flexible and can handle both constant rate traffic (audio, video) and variable rate traffic (data) easily. Second, at the very high speeds envisioned digital switching of cells is easier than using traditional multiplexing techniques, especially using fiber optics. ATM networks are connection-oriented. Making a call requires first, sending a message to setup the connection. After that, subsequent cells follow the same path to the destination. Similar to frame relay, cell delivery is not guaranteed but their order is. The intended speeds for ATM networks are 155.52 MB/s (for compatibility with SONET) and 622 MB/s (for four 155 MB/s channels) (Woodard, 2002).

The ATM architecture consists of two main layers. Just above the physical layer is the ATM layer, which deals with cells and cell transport. It defines the layout of the cell and what the header means. It also deals with establishment and release of virtual circuits. Congestion control is also located here. Because most applications do not work differently with cells, a layer above the ATM layer has been defined that allows users to send packets larger than a cell. This is called the ATM Adaptation Layer (AAL). There are five types of AALs defined for different types of services. AAL 1 and AAL 2 are used for transporting voice directly over ATM whereas, if we want to use voice over IP over ATM, then we would need to use AAL 5.

1.1.3. IP Networks

IP networks are networks that use the ubiquitous internet protocol i.e. IP. They are generally based on the TCP/IP stack. IP is the network level protocol that encapsulates the higher layer PDU (protocol data unit) into IP datagram. The most significant feature of IP is its 32-bit IP address: a virtual address given to each host and router in the network. The actual physical address of the device is obtained using some address resolution protocol (ARP). The IP address was developed as a reference format for an address field understood by all devices as they use different formats for physical addresses based on the standards used. A characteristic of all IP networks is that they are best-effort networks i.e. the network does not implicitly provide a guaranteed or differentiated quality of service (QoS) or class of service

(CoS) (Schulzrinne, 2000). Higher layers such as TCP need to compensate for this. This is a major drawback for real-time traffic over IP networks, which require a certain QoS for the service to be acceptable. Thus separate protocols need to be developed and implemented in order to transfer voice over IP networks with an acceptable quality. These are discussed in detail in the next section.

1.1.4 IP vs ATM

IP networks have been mainly designed to support a best-effort service, which is suitable for data. Thus as mentioned before, they still face many challenges to supporting high-quality LD voice in a multiservice environment. The new architectures and protocols developed will go a long way to help meet these challenges.

ATM protocols and standards for multiservice networks are largely defined. The main challenges remaining for LD voice services are related to the selection of speech processing algorithms, a transport protocol above the ATM layer i.e. AAL1 or AAL2 and the right network architecture. ATM interoperates with PSTN networks rather well as they can use similar signaling standards. ATM also beautifully integrates different types of traffic encompassing all type of CBR (Constant Bit Rate) and VBR (Variable Bit Rate) traffic. It supports QoS guarantee and inbuilt QoS differentiation. ATM also has tremendous support from physical layer standards such as SONET. Thus, it appears to the perfect technology for voice. However, being multifaceted has its drawbacks. Congestion control, flow control and management issues are yet to be solved in their entirety.

While an IP network incurs significantly higher delay and jitter at lower link speeds, the difference diminishes as the link speed increases. The continuing growth of data traffic will justify the higher-speed links needed to support IP data traffic. Carrying voice at high priority on these IP networks will then be very attractive. While IP packets are allowed to be as large as 64 KB, the maximum size for packets carried on today’s Internet is 1,536 bytes, with an average of about 350 bytes. If this

maximum packet size does not change, then excellent delay jitter performance is possible with link speeds of OC-3 and higher. For ATM, we do not need to limit the size of IP packets to allow control of delay jitter for voice.

With the current protocol stack, voice over IP is less bandwidth efficient than voice over ATM using either AAL-1 or AAL-2. Innovations discussed in this report will help narrow or eliminate the gap. As mentioned before, for data traffic originating from IP endpoints, IP is more bandwidth efficient than IP over ATM, owing to the ATM and AAL-5 overheads, as well as the partial fill of ATM cells using AAL-5. As the proportion of IP data traffic grows, the overall bandwidth efficiency of an integrated IP network becomes more favorable than that of an integrated ATM network.

Both ATM and IP can offer multi-application VPN (Virtual Private Network) services. The ability to offer such services is built into ATM standards and products. The currently emerging Layer 3/4 IP switches have flexible, hierarchical bandwidth management that will help make IP networks increasingly more suitable to offer such services. The emerging DiffServ standardization using the DS field will further add to the flexibility of IP networks that provide LD voice and multi-application VPN services. (KUNDAJE et al., 2001)

The biggest hindrance in the widespread deployment of ATM for integrated voice-data services is the cost. ATM equipment is very expensive currently. Also, IP networks are the most prevalent in the world today. To convert them to ATM requires a complete revamping which is too expensive for most. The success of voice over packet networks is based on the lower cost factor. If we take that away there is no reason for going ahead with the idea. Thus, we feel it is better to stick to IP networks and enhance their capabilities. Although a daunting task it would be worth it. Time is the essence; since if the current PSTN carriers do provide better services at lower costs sometime soon, the market for VoIP services may die out.

CHAPTER TWO
IP TELEPHONY

The text-based communication services do not need guaranteed Quality of Service (QoS). That is the reason why they are perfect for use in the packet-switched Internet. In contrast to these services the telephony system was invented more than 100 years ago and is based on a circuit switched network. This Public Switched Telephone Network (PSTN) is very reliable and telecommunication companies have invested much money to build it up and maintain it. Therefore they need good reasons to migrate the speech-data to a packet-switched network.

This chapter first describes the basics of the Public Switched Telephone Network and then explains how IP telephony works. Also the advantages and problems of IP telephony are discussed. After a brief introduction to audio codecs also fax over IP, video over IP and data conferencing are described.

2.1. The Public Switched Telephone Network

Although this chapter deals with IP telephony, it is especially for software engineers very important to understand the basics of PSTN. This section briefly explains the architecture and the most important components.

2.1.1. Architecture

Figure 2.1 shows the structure of PSTN. A call from a telephone is transferred over an access line (local loop) to a Central Office. There, a Class 5 voice switch digitizes the call to a 64 kbps Pulse Code Modulation (PCM) voice stream (according to ITU G.711)

Using Time Division Multiplexing (TDM), this stream is multiplexed onto trunk lines, which connect to Class 4 voice switches or to Private Branch Exchanges (PBXs). A PBX is used for example by companies or universities to bundle their telephone lines. Like Class 5 switches, they handle the routing of the calls to the telephones, but often they have additional features.

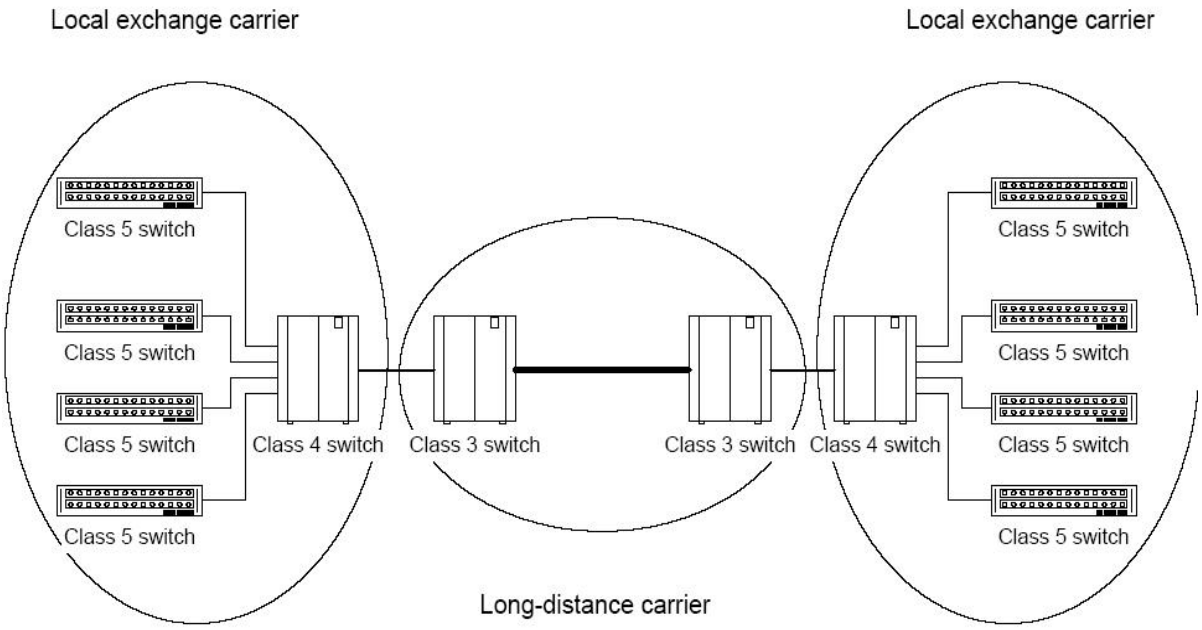

Figure 2.1 Public Switched Telephone Network

2.1.2. Signaling

Before a person can talk to another, a connection must be established. This functionality is realized by signaling protocols. A distinction is drawn between user-to-network signaling and network-to-network signaling.

2.1.2.1. User-to-Network Signaling

The most frequented method for analog user-to-network signaling is Dual Tone Multi Frequency (DTMF). Every numeral is assigned to a unique frequency. This tone is sent over the voice channel to the switch (In-Band signaling).

Integrated Services Digital Network (ISDN) uses Out-of-Band signaling. In this case, a separate Data channel (D-channel) is used for the signaling information. This channel has a transfer rate of 16 kbps. The voice is transmitted over a so-called Bearer channel (B-channel) with 64 kbps. An ISDN Basic Rate Interface (BRI) has two B-channels and one D-channel. The more powerful Primary Rate Interface (PRI) uses 30 B-channels and one D-channel (in Europe).

2.1.2.2. Network-to-Network Signaling

Typical In-Band signaling methods for network-to-network signaling on trunk lines are Single Frequency (SF), Multi Frequency (MF) or Robbed-Bit signaling. When SF is used, no tone is sent while the circuit is up. When one party hangs up, a disconnection is signaled by sending a 2600 Hz tone over the circuit. MF uses different frequencies to signalize either events like seizure, release, answer or acknowledgement or to send information like the phone number. While SF and MF is used for analog telephone systems, Robbed-Bit signaling was developed for digital ones. Here, the least-significant bit from the frames of the voice bit stream is dedicated to signaling.

Now let us take a look at Signaling System 7 (SS7), the most widespread type of Out-of-Band signaling. It was defined by the International Telecommunications Union (ITU) in 1980 and includes three classes of devices: Service Switching Points (SSPs) are switches that originate or terminate calls, Service Control Points (SCPs) offer access to databases with additional routing information and Signal Transfer Points (STPs) route SS7 messages. These devices are called SS7 nodes.

The SS7 protocol stack consists of four layers. Message Transfer Part (MTP) 1, MTP 2 and MTP 3 are equivalent to the three layers in the Open Systems Interconnection (OSI) reference model (physical, data link and network layer). Following protocol sets form layer 4: Telephone User Part (TUP), ISDN User Part (ISUP), Transaction Capabilities Application Part (TCAP) and Signaling Connection Control Part (SCCP). TUP was created to perform basic phone calls. Because it does not support ISDN and intelligent network functions like call forwarding or selective call blocking, ISUP was developed. It originates, manages and terminates ISDN and non-ISDN connections. TCAP enables connections to external databases. The obtained information is transported in form of a TCAP message. Finally, SCCP provides end-to-end routing and is required to route TCAP messages to their proper database.

2.2. Overview of TCP/IP

TCP/IP is defined as an industry standard suite of protocols that computes use to find, access and communicate with each other over a transmission medium. In this context, a protocol is the set of standards and rules that a machine’s hardware and software must follow in order to be recognized and understood by other computers. The protocol suite is implemented via a software package most commonly known as the TCP/IP stack, which breaks the job into a number of tasks. Each layer corresponds to a different facet of communication. The TCP/IP architecture consists of four “layers” performing certain functions: 1) Application layer, 2) Transport layer, 3) Internet layer and 4) Physical (network interface) layer. Each layer contains

protocols, which will be briefly summarized here. Figure 2.2 describes the TCP/IP reference model (OSI/RM), the standard that all other protocols follow.

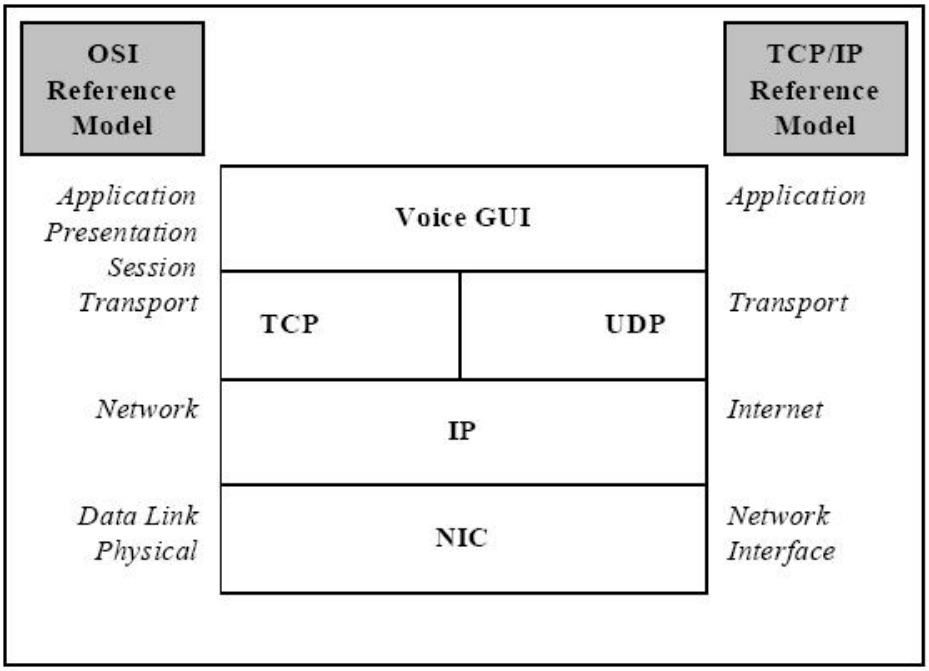

Figure 2.2 OSI and TCP/IP Reference Models

When transmitting voice over the Internet, the data being sent starts from the application layer (typically via the GUI), traverses down the “stack” to the network interface card (NIC) with each layer adding header and trailer frames. It is then sent to the receiver, where the data goes up the “stack” in reverse order, and this time stripping the appropriate header and trailer frames.

2.3. Packetization

Given the nature of addition and removal of header/trailer data in each packet, there is an innate packetization and processing delay. For a latency-sensitive application such as voice, it is imperative that this delay be minimized. Conversely, it is desired to efficiently transmit packets over the Internet to fully utilize the bandwidth. There is clearly a trade-off in the desire to maintain small packets to

minimize delay and the desire to send a large payload to minimize the overhead due to header content, thus maximizing payload efficiency. Packets are efficient for data transfer, but are not so attractive for real-time services such as voice. That is where the selection of an optimum voice coder is necessary, which is discussed later. Since this paper is focusing on voice over IP as opposed to voice over ATM, techniques on minimizing latency in the IP environment will be discussed.

The smallest packetization delay obviously occurs if only one sample of voice was sent at a time. However, that would cause a great more number of packets to be sent which would strain packetization processing as the single sample traversed the TCP/IP stack. "…If voice were digitized at 8000 samples/s, where each sample is 1 byte, then a 500 byte packet would take 62,5 ms to fill. For a desired delay of no more than 100 ms, it would mean that 62,5 percent of the delay budget is spent in packetization!..." (MEHTA & UDANI, 2001)

Each voice packet incurs an uncompressed 40 byte header that comprises 20 bytes for the IP header, 8 bytes for the UDP header and 12 bytes for the Real Time Transport Protocol (RTP) header. The IP header consists of several fields, including version, its length, type of service, flags, time to live, protocol, header checksum and source and destination IP addresses. The UDP header contains 8 bytes of Protocol Control Information (source and destination ports, UDP length and checksum). Finally, the RTP packet is used on top of UDP to allow the transport of isochronous data across a packet network, which introduces jitter and may send packets out of order.

2.4. An Overview of IP Telephony

Figure 2.3 shows several possibilities to use IP telephony. If people use their computers to communicate they must be online to receive calls. It is also possible to use gateways that connect a packet-switched network with PSTN. Then, connections from and to a conventional telephone are possible, too. A third alternative would be

that both dialog partners use telephones but the call is routed through a packet-switched network.

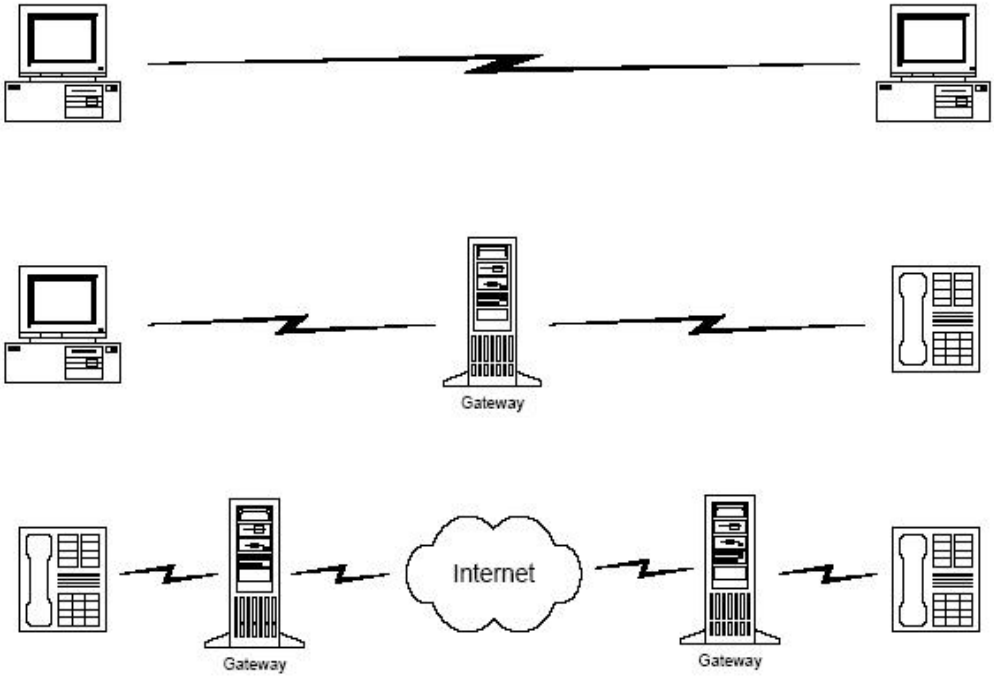

Figure 2.3 Possible ways to use IP telephony

2.4.1. Advantages

IP telephony was developed for the transmission of speech-data over the Internet Protocol. One of its most important advantages is, that a voice transmission over a packet-switched network is much cheaper because of the more efficient use of bandwidth. While a phone call over the PSTN needs a full-duplex 64 kbps channel for the duration of the call a VoIP-transmission requires about 14 kbps with compression. This bandwidth is used only when something has to be transmitted. Another advantage is, that just one instead of two different networks has to be maintained. Either computers with multimedia capabilities or IP phones that can be plugged into a conventional RJ-45 socket can be used for phone calls.

People often do not use just one kind of communication service. They send and receive emails or faxes, make phone calls or participate at video conferences. It is very hard to integrate all these services when different kinds of networks are used. If all services run over a packet-switched network, it is possible to use all these communication services within one application. A Unified Messaging System (UMS), which is described in chapter 7 realizes such an integration.

Another advantage of VoIP is the much easier development of additional services. This leads to a higher functionality of VoIP systems.

2.4.2. Problems

The quality of IP telephony depends on the network that transports the speech packets. While a Local Area Network (LAN) can send these packets in an appropriate time and with a negligible packet-loss, a Wide Area Network (WAN) often can not. This is a big problem for VoIP solutions. Normal data-packets are almost not very time-critical, but VoIP-data needs a guaranteed QoS. The goal of QoS is to provide the bandwidth and latency an application needs. This can be realized by prioritizing the data, whereas time-critical data are transferred faster than non-critical.

2.4.2.1. Latency

Latency or delay is the duration of the voice from the speaker to the listener. The ITU-T G.114 recommendation defines 150 ms as limit for a good quality. This limit is often unreachable, but up to 300 ms delay is in the majority of the cases also accepted. Three kinds of latency are defined (Davidson & Peters, 2000):

Propagation Delay: Because the light travels in a vacuum with a speed of 300000 kilometers per second and the electrons in copper cables or the light in light-wave cables travel with 200000 kilometers per second it takes some time to pass the route from the sender to the receiver. For example a signal over a light-wave cable

around the half globe takes about 100 ms. Propagation delay is not avoidable, but often negligible.

Handling Delay: This kind of latency depends on the active components. The more routers and switches a packet has to pass, the more delay arises. The coding and decoding of the audio signal takes about 20 ms each. When proxies or firewalls must be passed, a delay up to 500 ms can occur. Of course, such latency would be unacceptable.

Queuing Delay: Queuing delay occurs when a component like a switch or a router must handle more packets than it can process. It is dependent on the degree of utilization of the network. A queue at the endpoint is needed to reduce the influence of jitter on the overall delay.

2.4.2.2. Jitter

When sending two voice-packets with a defined offset, but these two packets are received with a larger offset, this effect is called jitter. Jitter also affects the latency of the voice-transmission, because the more jitter appears the larger the queue at the receiver must be. Under normal conditions, this queue should store a packet for at least 30 ms.

2.4.2.3. Echo

In analog telephone systems, not tuned electric impedances at transitions between 4-wire trunks and 2-wire local-loops can cause signal reflections. Modern telephones can suppress these echoes. In packet-based networks so-called echo cancellers are used. When User A talks to User B the echo canceller stores the voice signal. Then it is adding the negation of the stored signal at the appropriate place to the signal from User B.

If a computer as endpoint is used, also feedback must be kept in mind. When User A talks to User B, its voice is emitted by the loudspeaker and is then, through the microphone of User B, retransmitted to the originator. A good way to reduce or even eliminate feedback is to use headsets.

2.4.2.4. Packet Loss

In an overloaded network it frequently can happen that routers discard data-packets. If TCP is used the lost packet is retransmitted. This is an excellent strategy when sending conventional data, but for streamed-data it is not a good idea. That is why VoIP uses UDP as underlying protocol.

When a packet-loss occurs, the receiver has several possibilities. It can either send nothing (silence), send a noise or replay the last received packet (concealment-strategy). If just one packet was lost, silence is the worst, while replaying a packet is the best solution. Another possibility is to add to every transmitted packet the last transmitted one. This alternative wastes more bandwidth than the concealment-strategy does.

2.5. Related Communication Services

In addition to IP telephony there are other techniques from the PSTN domain which have alternatives based on IP. This section explains Fax over IP (FoIP), video over IP and data conferencing. While the first one becomes less important and is more and more replaced by emails, the popularity of video transmissions over the Internet rises due to the increasing bandwidth of backbones and last-mile technologies. Also data conferencing is becoming more and more important.

2.5.1. Fax over IP

FoIP provides the possibility to send faxes over an IP-based network. An endpoint of a connection can be a software application on the IP-network side or a conventional fax device on the PSTN side. The translation between Internet and

PSTN is realized by a FoIP-gateway (see Figure 2.4). When both endpoints are conventional fax devices and the message is transferred over a packet-based network, it passes two gateways.

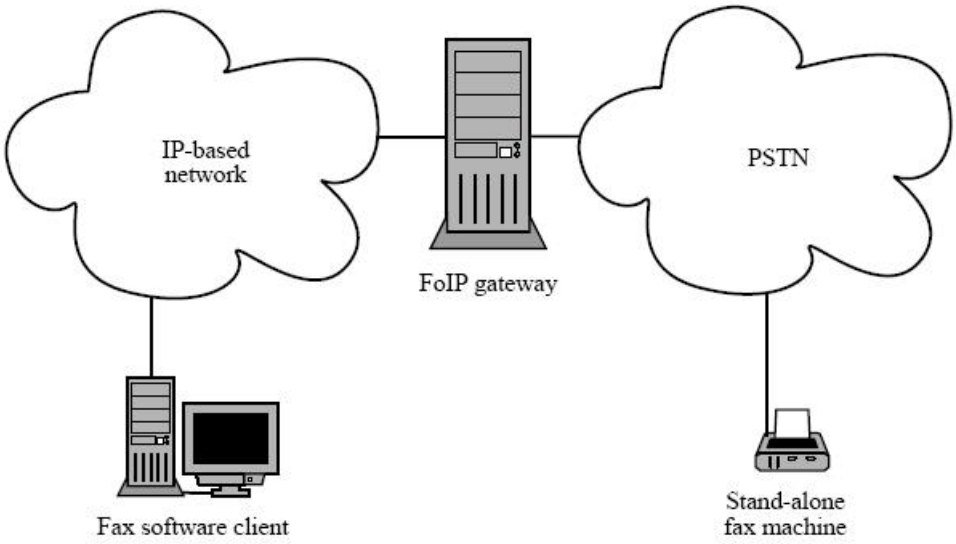

Figure 2.4 Fax over IP Architecture

A gateway can be implemented as a store-and-forward or as a real-time gateway. When using a store-and-forward gateway the message is stored at the gateway until the endpoint confirms the reception of the fax-message. The gateway then forwards the confirmation of the correct transmission to the originator. The protocol for store-and-forward FoIP was specified by the ITU-T in T.37. A better solution is real-time FoIP, because it complies with conventional fax sending. While store-and-forward FoIP simulates an existing connection to the receiver, this FoIP method directly sends the message to the recipient. T.38 is the standard for real-time FoIP. The protocol used on the PSTN-side is called T.30.

In contrast to voice over IP, FoIP does not allow lost data. Therefore gateways can use T.30 to inform the fax machine to delay or to interrupt the transmission or to resend parts of the message.

Because real-time FoIP requires guaranteed delay, which is possible in local area networks but not on the Internet, near real-time FoIP is a compromise between store-and-forward and real-time FoIP.

The main aspect why companies migrate to FoIP is the reduced costs. Overseas connections are very expensive and therefore it is a good idea to send the data via a FoIP-gateway. If a company does not have an own gateway, Internet-fax service provider offer the possibility to send messages via their gateway. If both communication partners have an IP-connection and a fax software, a gateway is not necessary. Another advantage of FoIP is the much easier integration into a unified messaging system where faxes can be converted into emails and vice versa. An important point which is often ignored is the security aspect. Unfortunately, a fax machine is usually used by the complete department or even the entire company. Insofar it can happen, that employees can read the faxes of colleagues. If FoIP in combination with a unified messaging system is used, the faxes are directly sent to the inbox of the recipient of the message.

2.5.2. Video over IP

Although H.323 is known as a standard for VoIP, it was also specified for video transmissions over IP and data conferencing. This section introduces the basics of video telephony and video conferencing, other kinds of video over IP, like video on demand or real-time video are not explained.

While audio codecs are mandatory elements on H.323 terminals, video codecs are optional elements. However, if they exist, video telephony with two users and video conferencing with two or more users are possible. If connections to the PSTN are required, the gateways must be extended in a way that they can communicate with conventional ISDN or PSTN video phones and video conferencing systems. Video phones normally use H.320 (for ISDN) or H.324 (for PSTN) as communication protocol.

When a multipoint video conference is performed, the Multipoint Control Unit (MCU), which manages the conference, must decide which of the incoming data streams are sent to the other participants. The audio channels of these streams are simply mixed together and sent to all endpoints. For the video channels there are

several possibilities. The incoming video streams can be merged to one video image (i.e. the resulting video image includes the downsized video images of each participant). Alternatively the resulting video stream can be one of the incoming streams. For example if a teacher performs a video conference with his pupils (distance learning) it is a good idea to lock the focus on the teacher. Finally, the MCU can transmit the video stream of the most prominent (loudest) participant to the others.

2.5.2.1. Video Codecs

There are several standards defined to encode and decode a video stream. If an H.323 terminal supports video conferencing, it should at least have H.261 QCIF implemented, which is explained later. To encode and decode a video stream, following steps are necessary:

- Recording and digitization of data
- Compression and encryption (on sender-side)
- Decryption and decompression (on receiver-side)

At this point, just the compression/decompression part is discussed. An often used technique for compressing a single image or individual pictures in a video is called intraframe compression. As compression method JPEG (Joint Photographic Experts Group) or JPEG-2000, which is based on the wavelet-technology are used. Because intraframe compression leads to high bit rates the interframe compression technique is mostly preferred. Because there are relatively little changes from one video frame to the next one, this technique exploits the similarities between two frames. This reduces the volume of data enormous.

The Moving Pictures Experts Group (MPEG) defined some standards that are based on interframe compression. All these standards have three different types of frames:

I-Frame (intra-coded frame): The I-frame is the only type of frame that uses intraframe compression. It is stored in the JPEG-format and can be decoded without any information of other frames in the video stream.

P-Frame (predictive-coded frame): In a P-frame just the differences to the last I- or P-frame are stored. Therefore these frames are also needed for the decompression of this frame.

B-Frame (bidirectionally predictive-coded frame): This type is similar to the P-frame, except that not only preceding but also succeeding frames are used for encoding this frame.

The more P- and B-frames are used in a video stream the smaller is the needed bandwidth for the transmission of an MPEG video.

The ITU-T also defined standards for video compression. H.261 describes a coding method for compressing a standard television color-signal into a video stream with a bandwidth from 64 kbps to 2 Mbps. It includes two formats: The Common Intermediate Format (CIF) has a resolution of 352x288 pixels with 30 pictures per second. The Quarter Common Intermediate Format (QCIF) defines a resolution of 176x144 pixels which is a quarter of the pixels of CIF. QCIF provides between 7.5 and 15 pictures per second. Sub-QCIF (SQCIF) again has a smaller resolution than QCIF (128x96 pixels). H.263 is an ITU-T standard that was defined to offer video compression with higher resolution and lower bit rates (between 15 kbps and 20 kbps) than H.261.

2.5.3. Data Conferencing

T.120 is a data conferencing standard that is part of H.323. It simplifies collaboration by allowing users to share data and applications, or use whiteboards from different computers. In fact, T.120 is not a single protocol but defines a suite of protocols on the networking and applications level to enable real time multimedia

transmissions, multipoint data connections and conferencing. It is beyond the scope of this thesis to explain all these protocols in detail. Just an overview of some features is given.

Protocols on the networking level are T.122, T.123, T.124 and T.125. T.122 is a standard for multipoint services. It allows the participants of a conference to send data to the other ones. T.123 is responsible for transporting and sequencing data, and for controlling the flow of data across networks. It includes also an error-correction for a reliable data transport. T.124 provides the generic conference control for the initiation and administration of multipoint data conferences and T.125 specifies defines private and broadcast channels to transport the data. Together, T.122 and T.125 make up the T.120 multipoint communication services.

The T.126 and T.127 standards define the applications level of T.120. T.126 enables whiteboard functionality. It specifies how applications must send and receive data. Data transfer can either be compressed or uncompressed. T.127 is responsible for the correct file exchange among conference participants.

2.6. Components of VoIP

The PSTN is the collection of all the switching and networking equipment that belongs to the carriers that are involved in providing telephone service. In this context, the PSTN is primarily the wire line telephone network and its access points to wireless networks, such as cellular. VoIP is being promoted to augment, if not eventually replace, the current PSTN infrastructure. As previously mentioned, the overall technology requirements of an IP telephony solution can be split into four categories: signaling, encoding, transport and gateway control. These are succinctly described below but are discussed in the next four chapters, respectively.

The purpose of the signaling protocol is to create and manage connections between endpoints, as well as to create and manage calls. Next, when the conversation commences, the analog signal produced by the microphone from the

human voice needs to be encoded in a digital format suitable for transmission across an IP network. The IP network itself must then ensure that the real-time conversation is transported across the available media in a manner that produces acceptable voice quality. Finally, it may be necessary for the IP telephony system to be converted by a gateway to another format – either for interoperation with a different IP based multimedia scheme or because the call is being placed onto the PSTN.

2.7. Summary

To establish a connection in the PSTN, two kinds of signaling protocols are used: user-to-network and network-to-network signaling. An example for analog user to-network signaling is DTMF. ISDN has a separate data channel for the signaling information. The most widespread type of network-to-network signaling is SS7, which uses Out-of-Band signaling.

IP telephony has some advantages compared to common telephony systems. It uses the bandwidth more efficient. This is due to speech-data is being transmitted over a packet-based network. Also the integration of different communication services is much easier if the same type of network is used. Latency, jitter, echo and packet loss must be considered when implementing an IP telephony system.

The main speech coding techniques can be divided into waveform, source, and hybrid codecs. While source codecs create low bit rates, waveform codecs generate a good speech quality. Hybrid codecs use the methods of source and waveform codecs, trying to minimize the disadvantages of both.

FoIP, video over IP and data conferencing are communication services related to IP telephony. FoIP offers the possibility to transmit faxes over the Internet. Video over IP is based on the same protocols as IP telephony, but needs more bandwidth. Also other kinds of codecs, like MPEG-4 are used. Data conferencing allows users to share data and applications or use white boards.

As previously mentioned, the overall technology requirements of an IP telephony solution can be split into four categories: signaling, encoding, transport and gateway control. These are succinctly described below but are further discussed in the next four sections, respectively.

CHAPTER THREE
SIGNALLING

Once a user dials a telephone number (or clicks a name hyperlinked to a telephone number), signaling is required to determine the status of the called party – available or busy – and to establish the call. These are multiple and complex levels of signaling that must take place in order to initiate and complete a call; these complexities escalate when VoIP users in packet networks communicate with PSTN subscribers. Neither H.323 nor the Session Initiation Protocol (SIP) alone makes up a complete set of IP telephony protocols; these protocols are merely competing standards for signaling. Both of these schemes will be explicated herein. Moreover, a means to achieve PSTN services from VoIP, namely interacting with Signaling System 7 (SS7), will be briefly examined.

3.1. The H.323 Protocol Suite

H.323 is a standard from the ITU-T for multimedia collaboration on packet-based networks like IP or Asynchronous Transfer Mode (ATM). It consists of several ITU-T recommendations which include protocols for the interaction of H.323 components and the communication with Switched Circuit Networks (SCNs). Figure 3.1 shows some of the H.323 protocols on the TCP/IP protocol stack.

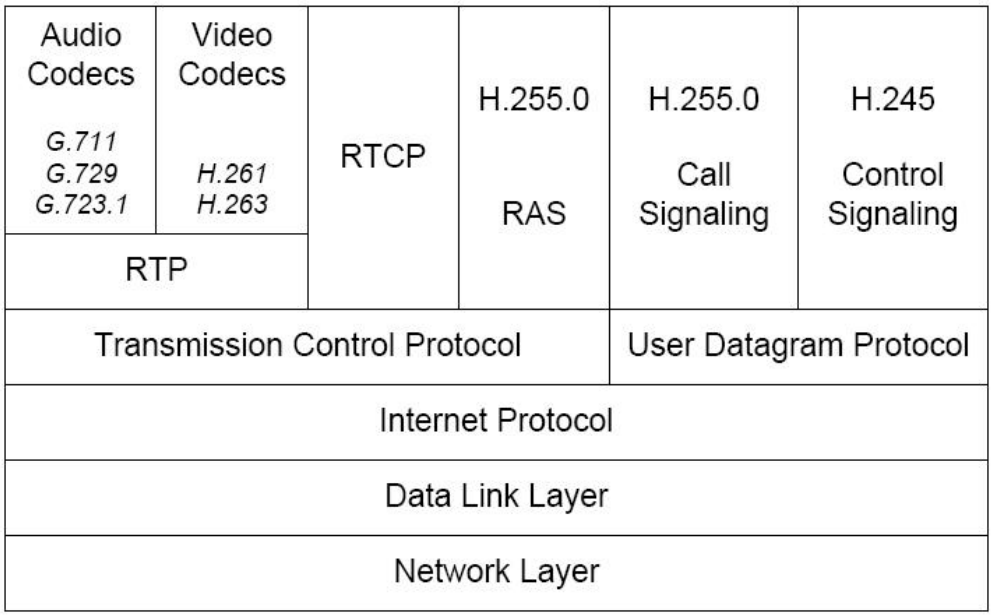

Figure 3.1 H.323 protocols on TCP/IP stack

This chapter describes the architecture of H.323 and defines the components that are collaborating. Finally, the most important protocols of the H.323 protocol suite are explained.

3.1.1. Architectural Overview

Figure 3.2 shows an example of two H.323 IP telephony networks which are connected to the Internet and to PSTN. Both networks include elements like terminals, gateways and gatekeepers. H.323 zone B also includes a Multipoint Control Unit (MCU). At the PSTN-side a conventional telephone and an H.324 video phone are connected.

The term H.323 zone was defined by ITU-T as follows: A zone is the collection of all terminals, gateways and MCUs managed by a single gatekeeper. A zone includes at least one terminal, and may or may not include gateways or MCUs. A zone has one, and only one gatekeeper. A zone may be independent of network topology and

may be comprised of multiple network segments which are connected using routes or other devices.

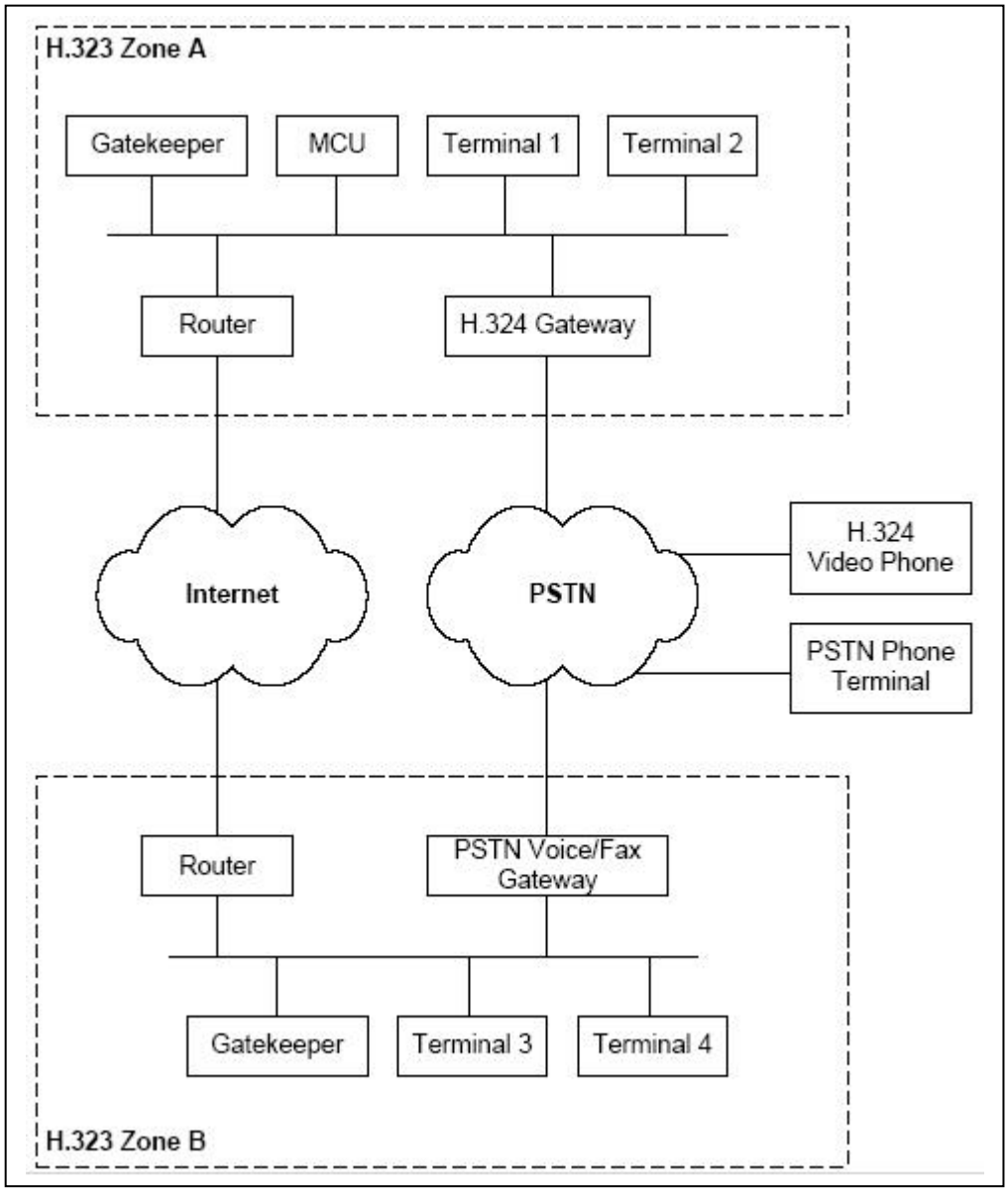

Figure 3.2 H.323 IP telephony networks

3.1.1.1. Terminal

An H.323 terminal can be a personal computer (PC) with multimedia capabilities or a stand-alone device. It must include an H.323 protocol stack. The basic service of

a terminal is audio communication. Optional services are video or data communication. Audio and video streams run over the Real-Time Transport Protocol (RTP). The Real-Time Transport Control Protocol (RTCP) provides feedback on the quality of the transmission. Another function a terminal must implement is Registration, Admission and Status (RAS). It is used to register endpoints (terminals and gateways) at gatekeepers. RAS also supports admission control and bandwidth changes. Call signaling provides functions to establish a connection between endpoints and control signaling is used to send control messages.

3.1.1.2. Gateway

A gateway is an element that links different networks (e.g. a packet-based network with the PSTN). For this purpose they convert media formats and translate the protocols for call setup and release. Gateways act on the H.323-side like an H.323 terminal and on the SCN-side like an SCN terminal. Therefore, gateways have to implement services like RAS, call signaling, control signaling, RTP and RTCP. If no connection to another type of network or protocol is needed, gateways are not necessary.

Modern gateways are split into two functional units: a Signaling and a Media Gateway. A Signaling Gateway connects to SS7 and provides signaling control information to the Media Gateway Controller (MGC) or Call Agent. The MGC takes this information and provides call routing control and billing information. Finally, the Media Gateway transfers and converts the different media streams from one network to another. Figure 3.3 shows distributed switch architecture.

The protocol the Media Gateway uses to communicate with the Media Gateways is called Media Gateway Controller Protocol (MGCP). A newer standard jointly specified by ITU and IETF is called H.248 (by ITU) or Megaco (by IETF).

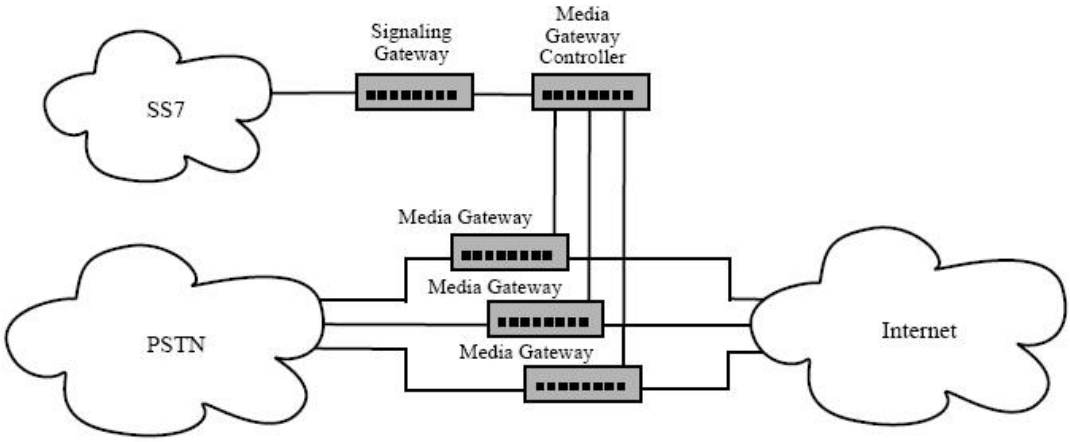

Figure 3.3 Distributed Switch Architecture

3.1.1.3. Multipoint Control Unit

An MCU is an endpoint which provides support for multipoint conferences. It shall consist of a Multipoint Controller (MC) and zero or more Multipoint Processors (MPs).

The MC checks the capabilities of three or more endpoints and then sends the possible operating modes to all endpoints attending at the conference. The MP reads one or more input streams, switches or mixes them and then sends the resulting stream to all endpoints in the conference.

3.1.1.4. Gatekeeper

A gatekeeper is an optional element in an H.323 environment. However, if it exists in a network, it must be used by terminals and gateways. Gatekeepers provide several call control services for H.323 endpoints. Some of these services are mandatory while others are optional.

Mandatory services a gatekeeper must provide include address translation (IP address to standard E.164 telephone number) or admissions control. Also bandwidth control and zone management must be provided.

Optional services include call-control signaling (based on the Recommendation H.225 of the ITU-T), call authorization (a gatekeeper can accept or reject a call depending on access-based or time-based restrictions to and from terminals or gateways) or call management (e.g. rerouting a call to achieve load balancing).

3.1.2. H.225.0 Call Signaling Protocols

The H.225.0 recommendation [ITU, 2000a] from the ITU includes H.225.0 RAS for the registration of an endpoint at the gatekeeper and H.225.0-Q.931 for setting up and terminating a call between two endpoints. All kinds of H.225.0 messages (as well as the H.245 messages described in section 3.3) are ASN.1 (Abstract Syntax Notation One) encoded. This is a binary format based on so called Packet Encoding Rules (PER).

3.1.2.1. H.225.0 RAS - Registration, Admission and Status

H.225.0 RAS defines a protocol between an endpoint (terminal or gateway) and the gatekeeper. RAS messages are transferred over an unreliable protocol like the User Datagram Protocol (UDP). Therefore, timeout checks and retries must be considered when implementing the protocol.

H.225.0 RAS provides a separate channel - the RAS channel - to transfer RAS messages. They are used to find the gatekeeper, to register endpoints at the gatekeeper or to check the admission of an endpoint to initiate or to receive a call.

Gatekeeper Discovery

Before an endpoint can register and authenticate at the gatekeeper it has to identify it in the network. If the gatekeeper discovery is done statically, then the endpoint is preconfigured with the transport address of the gatekeeper. If the discovery is done dynamically the endpoint multicasts a Gateway Request (GRQ) message. This GRQ message includes the transport address of the endpoint, the identifier of the gatekeeper that normally provides the service to the endpoint (or a null string) and a series of alias addresses that identify the endpoint. When a gatekeeper receives a GRQ message it determines whether or not to allow the endpoint to use the services it provides. If the endpoint is accepted, the gatekeeper will reply with a Gateway Confirm (GCF) message which includes the address and the identifier of the gatekeeper. If the endpoint is not accepted, the gatekeeper sends a Gateway Reject (GRJ) message.

Endpoint Registration

The next step after discovering the gatekeeper is to register the user and the associated endpoint with the gatekeeper. Therefore, it sends a Registration Request (RRQ) message. This message includes the transport address of the endpoint at which the gatekeeper should send H.225.0-Q.931 messages. This address is necessary when the signaling is routed via the gatekeeper (gatekeeper routed call). When the gatekeeper receives an RRQ message it performs checks to determine whether or not to allow the endpoint to register. If the endpoint is allowed to register, the gatekeeper sends a Registration Confirm (RCF), otherwise a Registration Reject (RRJ) message. The RCF message includes the transport address of the gatekeeper where the endpoint should send H.225.0-Q.931 messages to. By registering at a gatekeeper, the endpoint joins an H.323 zone.

Admission Control

Admission control enables the gatekeeper to authorize each outgoing or incoming call to and from an endpoint. By sending an Admission Request (ARQ) message the endpoint requests an alias address resolution and a call authorization. The gatekeeper determines the address and checks the authorization for the call. For the authorization it may contact a user policy server. A user policy server handles a profile for each user. Such a profile defines the geographic locations where calls can be made, or restrictions on the duration of a call. If the gatekeeper allows the user to make the call it sends an Admission Confirm (ACF), otherwise an Admission Reject (ARJ) message.

Accounting

For billing it is important to know the participants as well as the duration of a call. Because H.225.0-Q.931 signaling is between endpoints, the gatekeeper must be notified of the start and the release of a call. This is done by sending an Information Request (IRR) message to the gatekeeper. When a call was released either the gatekeeper itself or an accounting server generates a Call-Detail Record (CDR). This CDR includes information like starting time, duration, identification of the two parties and so on.

Call Termination

Not just due to billing but also because of bandwidth reasons the gatekeeper must be notified when a call was released. If a gatekeeper has allocated bandwidth for a call it then can reallocate them. Therefore, the endpoint sends a Disengage Request (DRQ) message which includes the call identification. But also the gatekeeper can use the DRQ message to force the endpoint to terminate a call. This is necessary for example if a user of a prepaid card has run out of credit.

User Deregistration

To cancel a registration of a user, either the gatekeeper or the endpoint sends an Unregistration Request (URQ) message to the other side, which replies with an Unregistration Confirm (UCF) message. If an endpoint sends a URQ message for a registration that does not exist, the gatekeeper replies with an Unregistration Reject (URJ) message.

If an endpoint fails unexpectedly it cannot unregister anymore. This can cause problems when a gatekeeper is at full capacity and has to reject new registrations. To avoid such situations either the gatekeeper can ping all registered endpoints periodically and delete the not responding ones or the endpoint can provided a heartbeat in defined intervals. The latter method is called lightweight registration.

3.1.2.2 H.225.0-Q.931 - Call Signaling

H.225.0-Q.931 defines the communication between two endpoints (point-to-point calls) for setting up a call and terminating a connection. It uses a subset from the Q.931 ISDN signaling messages, which enables an easy interconnection with ISDN networks. Each H.225.0-Q.931 message has the same information elements as the corresponding Q.931 recommendation message.

H.225.0-Q.931 is running over a reliable protocol like TCP. Before call signaling can start, the transport address of the other endpoint must be known. The transport address consists of an IP address and a TCP port number. The default port number for H.225.0-Q.931 is 1720. To obtain a transport address from a URL the H.225.0 Annex G protocol is used.

The following scenario describes the call signaling messages between Terminal A and Terminal B, where Terminal A initiates the call.

Call Initiation

At the beginning of a call Terminal A connects to the H.225.0-Q.931 transport address of Terminal B. Then a setup message is sent to Terminal B. The setup message includes two transport addresses of terminal A. One is the H.225.0-Q.931 and the other is the H.245 transport address (see section 3.3). Terminal A will wait at least four seconds for a response from Terminal B before disconnecting.

Call Proceeding

If between Terminal A and Terminal B a gateway exists, this gateway can use a call proceeding message to inform Terminal A that the setup message was received but it can not be relayed to Terminal B within four seconds. The call proceeding message is optional.

Call Alerting

When Terminal B has received the setup message it informs the user of the incoming call. This can be done by popping a dialog box on the user-interface display or in form of a ringing tone. Then Terminal B sends an alerting message to Terminal A, which waits at least 180 seconds for another H.255.0-Q.931 message before disconnecting.

Call Connection

If the user at Terminal B takes the call a connect message is sent to Terminal A. This message includes the transport address of Terminal B for the H.245 call control. Terminal A will then open a TCP channel for call control to the correct address. Now the call has been established and the transmission of multimedia data can begin.

Call Termination

When the call is finished or the user at Terminal B has not taken it, either endpoint sends a release complete message. Finally, both TCP connections are closed.

3.1.3. H.245 - Call Control

After the call between endpoints is established and before the multimedia transmission can begin the capabilities of the endpoints must be determined. This is necessary to use audio or video codecs that are available on both sides. The H.245 protocol provides messages to gather this information. A Terminal Capability Set message includes the receiver's capabilities for audio/video transmissions, data applications and user input. The sender specifies the kind of data with the Open Logical Channel message, which includes the type of the data stream (audio, video or application data). All messages defined in H.245 are ASN.1-encoded.

3.1.4. Supplementary Services

In addition to the basic telephony service, the ITU-T has defined a set of supplementary services in the H.450 protocols. Most of these services already exist in PSTN (a modern PBX provides up to 300 supplemental services). Examples for standardized supplementary services are call transfer, call forwarding, call hold, call park and pickup, call waiting, and message waiting.

3.1.5. Summary

H.323 is a protocol suite for multimedia collaboration on packet-based networks. There are four components defined in the H.323 standard. A terminal is an endpoint that initiates and terminates calls. It can either be a PC with multimedia capabilities or a stand-alone device.

Gateways link different networks. They convert media formats and translate the protocols for call setup and release. An MCU provides support for multipoint conferences. It reads one or more input streams, switches or mixes them and sends the resulting stream to all connected endpoints.

Gatekeepers provide call control services like address translations, admissions and bandwidth control as well as zone management. The H.323 standard consists of a number of protocols. Examples are the H.225.0 Call Signaling protocols and H.245 for Call Control. The H.225.0 recommendation includes H.225.0 RAS for the registration of an endpoint at the gatekeeper and H.225-Q.931 for setting up and terminating a call between two endpoints. H.245 provides functionality to determine the capabilities of the endpoints.

3.2. Session Initiation Protocol

The Session Initiation Protocol (SIP) is a signaling protocol standardized by the IETF. Although SIP can be used to establish IP telephony calls, it is not limited to that kind of sessions. Also sessions for instant messaging, network games or video conferencing can be managed. SIP operates over any packet-based network, reliable or unreliable, but usually SIP uses UDP as underlying protocol.

In contrast to H.323 which is ASN.1-coded, SIP is a text-based protocol using the ISO 10646 character set in UTF-8 encoding. It is more similar to the Hypertext Transfer Protocol (HTTP) rather than to legacy signaling.

This chapter first explains how SIP addresses are defined. Then the components of SIP are specified. Also the structure of a message and the different types of messages are described. Next, some scenarios like the registration of a user agent or the set-up of a call are shown. Finally, SIP-specific event notifications and the Session Description Protocol (SDP) are explained.

3.2.1. SIP Addressing

A SIP address is necessary to identify a user. A user who has a SIP address is globally reachable. Such an address has the form of a Uniform Resource Locator (URL) and is similar to a mailto URL, i.e. sip:userinfo@host. The user info can either be a username or a telephone number. The host is a domain name or an IP address. A SIP URL can designate an individual, a group or a defined person in a group.

When the host indicates an Internet telephony gateway the user info part of the SIP address can specify a telephone number. This can either be a local or a global number (a global number begins with a '+'). Because this number could also be a valid username, a parameter denotes that it is a telephone number (e.g. +1234567890@gateway.xy user=phone). SIP addresses can be inserted in a webpage. By clicking the address-link a connection to the specified user is initiated. Because a SIP URL can often be guessed from the email-address of the user, SIP offers authentication and access control mechanisms.

3.2.2. SIP Components

The SIP specification defines several components that are necessary to implement a SIP environment.

User Agent: A user agent is an end-device which can be implemented in hard or software. It consists of a user agent client and a user agent server. The user agent client is a client application that initiates the SIP request while the user agent server listens for incoming calls and notifies the user when a call is requested.

Proxy Server: A proxy server is an intermediary program that acts as both, a server and a client for the purpose of making requests on behalf of either clients. Requests are serviced internally or by passing them on, possibly after translation, to

other servers. A proxy server interprets, and, if necessary, rewrites a request message before forwarding it.

Redirect Server: A redirect server is a server that accepts a SIP request, maps the address into zero or more new addresses and returns these addresses to the client. Unlike a proxy server, it does not initiate its own SIP request. Unlike a user agent server, it does not accept calls.

Registrar: A registrar is a server that accepts Register requests. A registrar is typically co-located with a proxy or redirect server and may offer location services.

3.2.3. SIP Messages

Messages in SIP are text-based and use the ISO 10646 character set in UTF-8 encoding. Lines must be terminated with a CR/LF. A SIP message is either a request from a client to a server or a response from a server to a client. A message sent by a SIP component consists of a start-line, a header, a blank line and an optional body. Figure 3.4 shows an example of an Invite message and its response from the server. This example is used to explain the most important elements of a SIP message.

The request message has following structure. The start-line states the kind of message, the caller's SIP URL and the SIP version. The Via header field describes the path of the request so far. The From field states the sender and the To field the recipient of the message. The Call-ID field uniquely identifies a particular invitation or all registrations of a particular client. This identifier should be generated by a good random number generator to avoid session hijacking. Consecutive request with the same Call-ID must have an increasing CSeq header field. The Subject field can contain a string that describes the session while the Contact field provides information how the user can be reached for further communication. Finally, Content-Type and Content-Length describe the kind and the length of the content in the message body, respectively. In the example the content in the message body is in

the SDP data format to specify the media type (audio), the codec (G.711) as well as the host name (100.101.102.103) and the port number (49172).

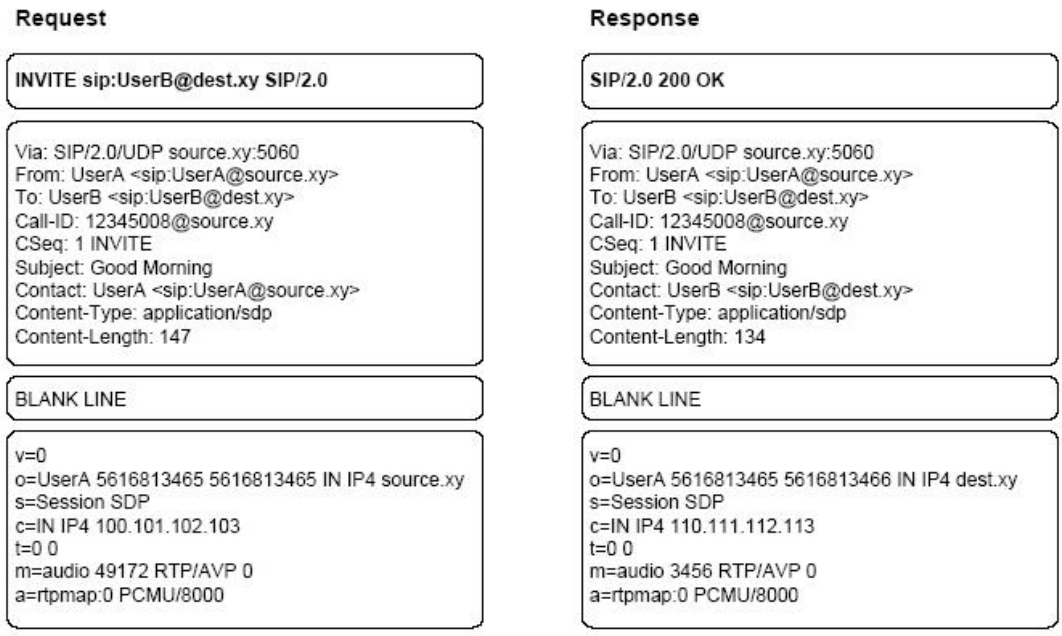

Figure 3.4 SIP Messages

In the response message the start-line includes the SIP version, the response code and a keyword that specifies the correct execution of the request. The Contact header field states the current location of the caller. This information can be used to directly contact the user and bypass any proxy servers. All other fields in the response message are similar to the request message.

The following list describes the standardized SIP messages. SIP extensions (e.g. SIP Extensions for Instant Messaging, see section 6.2) may define additional message types.

Register: The Register message binds a permanent SIP address to a current location. It includes a To header field which states the SIP URL and a Contact header field with the current contact address. If the Contact header field is empty the

registrar returns all current contacts of the user. The Expires header field states how long the current location is valid (in minutes).

Invite: With this kind of message a client asks the caller to join a particular conference or to establish a two-party conversation. The message body contains a description of the session. Re-invites can be used to change the session state.

Options: This message is used to inquire the capabilities of a SIP component. Proxy and redirect servers simply forward an options message.

Ack: Ack is used to tell the callee that his confirmation of the Invite message was received.

Bye: The Bye message sent by the user agent client indicates that the user wishes to quit the session.

Cancel: If Invite requests are pending, the Cancel message is used to abort them. The response codes in a reply message have the form "xyz explanatory text".

Receivers of such a response code just have to understand "x".

1yz Informational	100 - Trying
	180 - Ringing
	181 - Call is being forwarded
2yz Success	200 - OK
3yz Redirection	300 - Multiple choices
	301 - Moved permanently
	302 - Moved temporarily
	305 - Use proxy

4yz Client Error	400 - Bad request
	401 - Unauthorized
	402 - Payment required
	486 - Busy here
5yz Server Error	500 - Server internal error
	501 - Not implemented
	503 - Service unavailable
6yz Global Failure	600 - Busy everywhere
	604 - Does not exist anywhere

3.2.4. SIP Transactions

This section describes transactions to register an endpoint at a current location, to initiate and to terminate a session. A user agent client can initiate a session either directly or via an intermediate server. The intermediate server can act as a proxy and forward the request or as a redirect server.

3.2.4.1. Registration

A user agent registers at a SIP registrar to bind the user's permanent SIP URL to a current location. This is realized by sending a Register message to the registrar. The Register message does not have to be sent to a dedicated registrar, also a multicast address ("sip.mcast.net") where a registrar listens to, can be used.

Figure 3.5 shows an example of a registration. The user agent client sends a Register message to the registrar. Then the registrar forwards the SIP URL together with the current location (from the Contact header field) to a location server. A location server is used by a SIP redirect or proxy server to obtain information about a caller's possible location (Handley et al., 1999). Finally, the registrar sends a reply to the user agent.

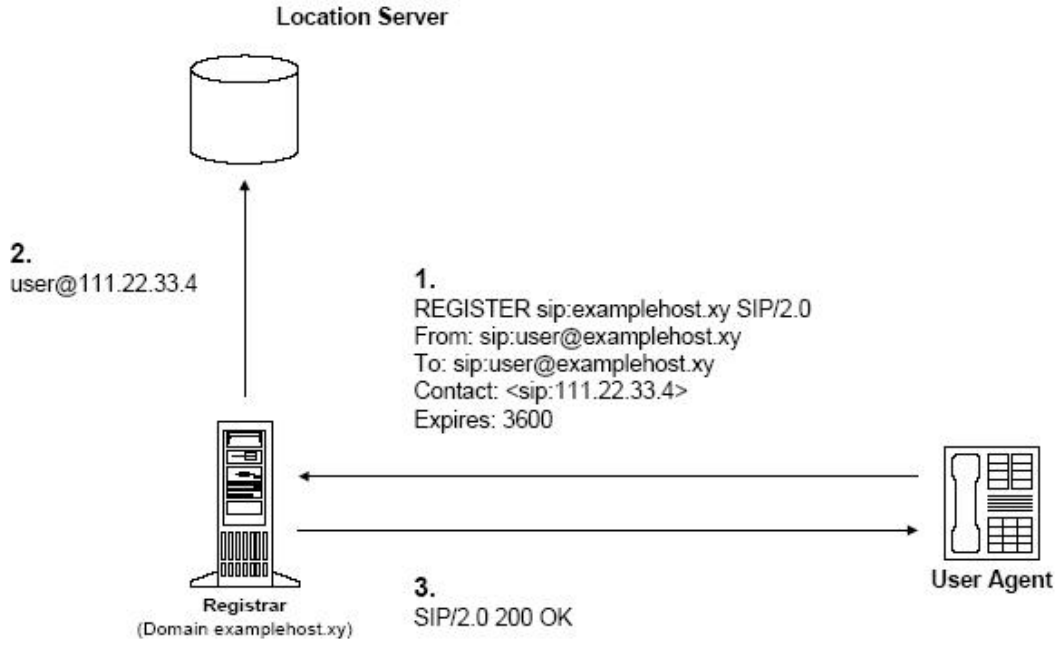

Figure 3.5 SIP Registration

If the user changes his location again he first sends a register message with the current location and sets the Expires field to zero. This removes the current location from the database. Then the user can register with the new location.

3.2.4.2 Session Initiation and Termination

When a person wants to establish a call, the user agent client first sends an Invite message to ask the callee to join the conversation. Then the callee can accept the call by sending a response (response code 200 OK). Before the callee accepts a call the callee's user agent client can send responses with informational response codes (e.g. 100 TRYING or 180 RINGING). If the callee has accepted the call the caller confirms that it has received the response by sending an Ack request. If the caller no longer wants to communicate, it can send a Bye message. To release an existing call, one of the participants sends a Bye request. A party receiving a Bye request must stop transmitting media streams.

The following two sections describe the invitation to a session when between the two parties one or more a proxy or a redirect servers are located.

3.2.4.3 SIP Invitation in Proxy Mode

When an Invite message passes a proxy server, the server requests the current location from a location service. Then it adds itself on top of the Via header field list, so that the packet can be routed back the same path and forwards the Invite message to the current location. If the location server returns more than one location, the proxy server either can try the addresses one after another or parallel.

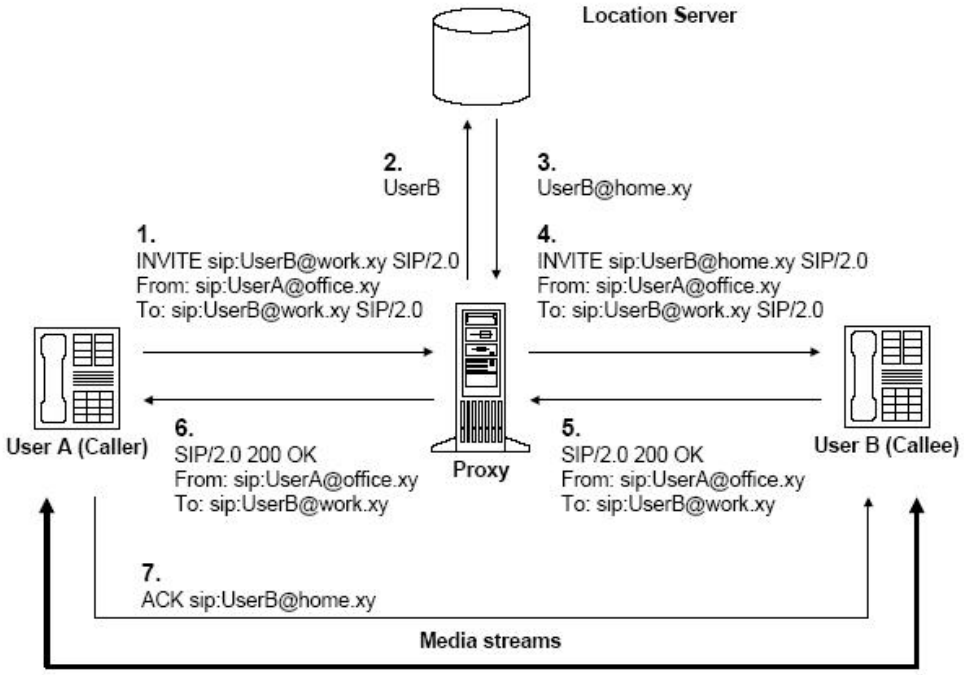

Figure 3.6 SIP Invitation in Proxy Mode

When the callee's user agent responses, each host on the path must remove its address from the Via field of the response. This is necessary to hide the internal routing information from the caller.

Figure 3.6 shows an invitation passing a proxy server. Note that Ack may be sent either directly to the callee using the callee's address from the Contact header field of the response message or via the proxy server again.

3.2.4.4. SIP Invitation in Redirect Mode

Using a redirect server is useful when a person has moved or changed the provider. Then the caller does not need to try the same server next time. The redirect server responses to an Invite message with a 301 (moved permanently) or a 302 (moved temporarily) response code. Figure 4.4 shows an example for the redirection of an Invite message.

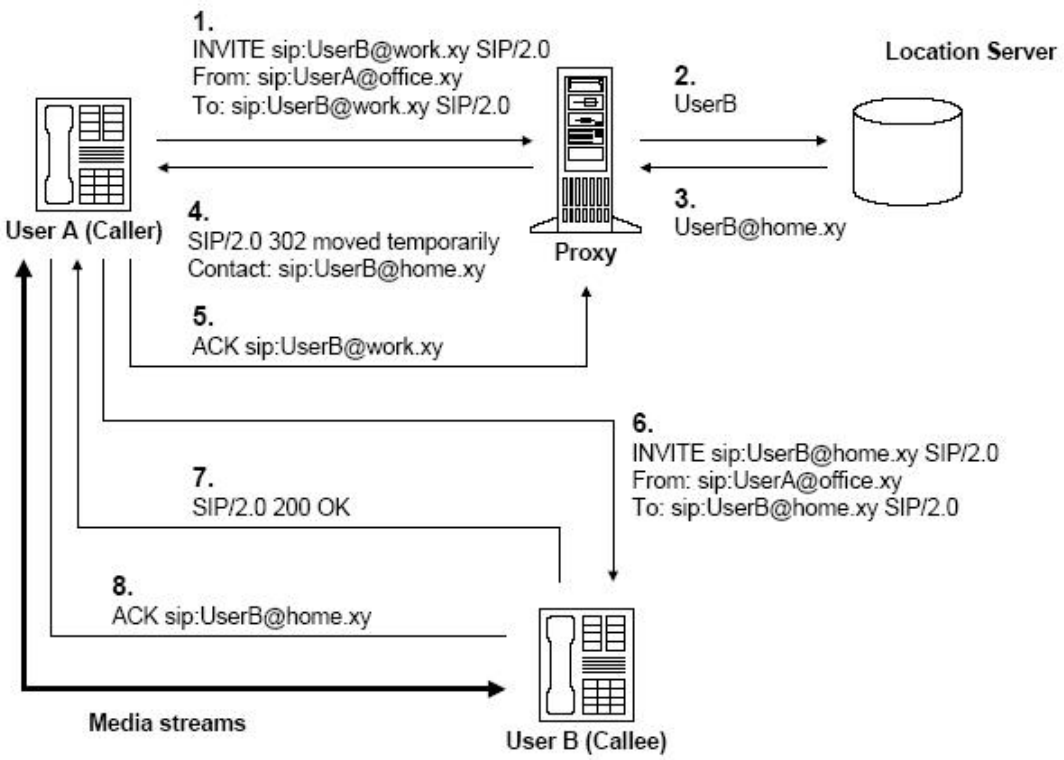

Figure 3.7 SIP Invitation in Redirect Mode

3.2.5. SIP-specific Event Notification

To receive the current state of a remote node (and to receive updates of the state) an entity (the subscriber) has to send a Subscribe message to the remote node. This node (the notifier) decides whether or not the subscriber is allowed to receive its notifications. If the subscription request was accepted, the notifier sends immediately after the acknowledgement message a Notify message with the state of the node. When the state changes, the notifier sends again a Notify message to all accepted subscribers. Figure 3.8 shows a typical flow of messages.

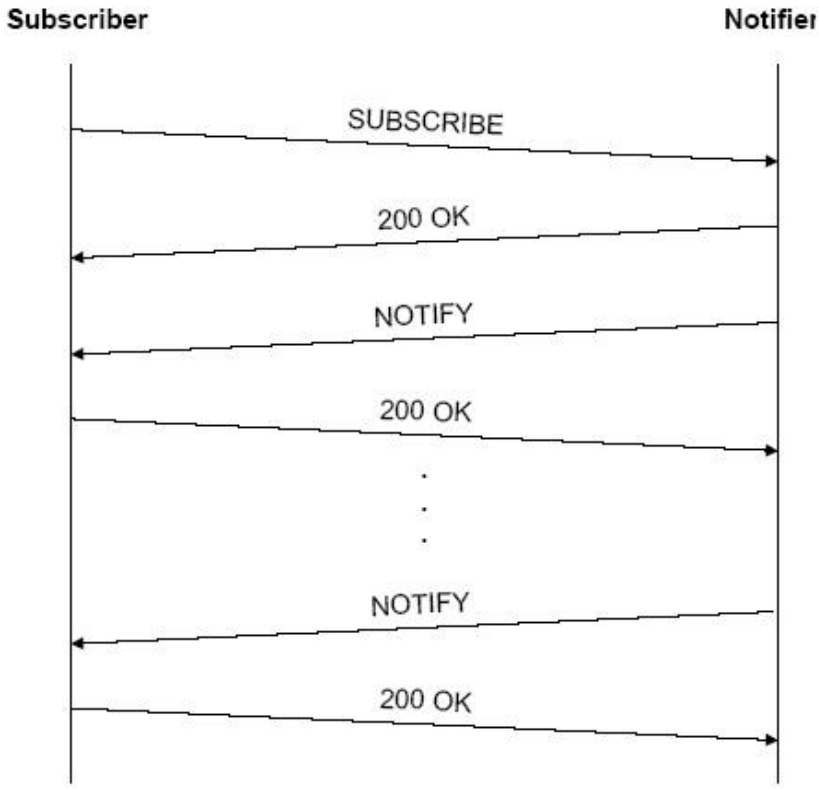

Figure 3.8 Typical flows of Subscribe and Notify messages

Subscribers must include exactly one Event header in Subscribe requests. This header field indicates which event or class of events the subscription concerns. The Expires header field indicates the duration of a subscription. To unsubscribe, a Subscribe message with the Expires header set to ’0’ must be sent.

3.2.6. Session Description Protocol

Data in the body of an Invite message like shown in Figure 3.4 is often encoded using SDP. The purpose of this data is to provide information of the multimedia capabilities of the caller. The body of the response message includes the capabilities of the callee. SDP includes following information:

- Session name and purpose
- Time(s) the session is active
- The media comprising the session
- Information to receive those media (addresses, ports, formats and so on)
- Information about the bandwidth to be used
- Contact information

To convey enough information to be able to join a multimedia session, following additional data about the kind of media used must be given:

- The type of media (e.g. video or audio)
- The transport protocol (e.g. RTP/UDP/IP)
- The format of the video (e.g. MPEG video)

For an IP multicast session also the multicast address of the media and the transport port for the contact address must be provided.

3.2.7. Summary

SIP is a signaling protocol that can be used to establish many kinds of sessions. It is a text-based protocol similar to HTTP. SIP addresses have the form of a URL (e.g. sip:user@host).

The following SIP components are defined in the IETF specification: A user agent is an end-device which initiates calls and waits for incoming calls. A proxy server is

an intermediary program that forwards SIP messages. A redirect server maps the address of a SIP request into zero or more new addresses and returns them. A registrar accepts Register messages and is typically co-located with a proxy or redirect server.

A registration is a SIP transaction where a user agent registers at a SIP registrar to bind the user's permanent SIP URL to a current location. To establish a call, the caller sends an Invite message to the callee. An invitation can be executed either in proxy or redirect mode.

Data in the body of an Invite message is often encoded using SDP. This data provides information of the multimedia capabilities of the caller. To enable SIP nodes to request event notifications from remote nodes, the IETF has specified an abstract framework. This framework defines the Subscribe and Notify messages.

3.3. H.323 vs. SIP

H.323 and SIP are both competing for the dominance of IP telephony signaling. There is much debate in the industry as to which protocol is superior, H.323, SIP, or perhaps another protocol that may be in the early stages of development. Currently, there is no clear-cut winner; however, the standards appear to be evolving such that the best features of each are being implemented in the other protocol. For example, the evolution of H.323 from versions 1 through 4 has focused on decreasing call setup delay from six or seven round trips to be on par with SIP's 1,5 round trips. Obviously, this convergence is highly desirable for interoperability issues between the two protocols and thereby reduces signaling overhead.

Both H.323 and SIP support the majority of required end-user functions comparatively equally, such as call setup and teardown, call holding, call transfer, call forwarding, call waiting and conferencing. Yet, functional differences remain, such as H.323's support for message waiting indication and SIP's support for third-party control. In addition, the third version of H.323 provides a more robust

mechanism for capabilities exchange – the process by which it is determined whether a particular feature is supported by both participating entities – than does SIP.

Furthermore, H.323 and SIP differ in terms of advantage in Quality of Service (QoS) and management, scalability and flexibility, and interoperability, as described in Table 3.1. It appears that H.323 has exceptional QoS, management and interoperability, due to H.323's support for the emerging Differentiated Services/Policy Management to QoS and the protocol's extensive history, respectively. On the other hand, because SIP is a significantly less complex protocol than its bloated counterpart, it scales much better.

Table 3.1 Advantages of H.323 and SIP in VoIP Features

Feature	Similar	Strengths of H.323 v. 3	Strengths of SIP
QoS and Management	Call setup delay, packet loss recovery, lack of resource reservation capability	Fault tolerance, admission control, policy control	Loop detection
Scalability and Flexibility	Stateless processing, UDP support, inter-server communications for endpoint location	Location of endpoints in other administrative domains	Less complexity, greater extensibility, ease of customization
Interoperability		PSTN signaling interoperability, inter-vendor interoperability	

In terms of impacting VoIP applications, vendors are implementing an assortment of protocols, ranging from the varieties of H.323 to SIP to a proprietary signaling protocol. Presumably, major vendors will support the two major protocols until it becomes clear that either one protocol will fade away or the two approaches will merge. The latter scenario is more likely unless either protocol makes significant advances that the other does not incorporate. Finally, Table 3.2 shows a detailed version of comparison of H.323 and SIP.

Table 3.2 Comparison of H.323 version 2 and SIP

	H.323 version 2	SIP
FUNCTIONALITY		
CALL CONTROL SERVICES:		
Call Holding	Yes	Yes
Call Transfer	Yes	Yes
Call Forwarding	Yes	Yes
Call Waiting	Yes	Yes
ADVANCED FEATURES:		
Third party control	No	Yes
Conference	Yes	Yes
Click-for-dial	Yes	Yes
Capability exchange	Yes & Better	Yes
QUALITY OF SERVICE		
Call setup delay	3~4 RT	2~3 RT
RELIABILITY:		
Packet loss delivery	Through TCP	Better
Fault detection	Yes	Yes
Fault tolerance	N/A	Good
MANAGEABILITY		
Admission Control	Yes	No
Policy Control	Yes	No
Resource reservation	No	No
SCALABILITY		
Complexity	More	Less
Server processing	Stateful	Stateful or Stateless
Inter-server communication	No	Yes
FLEXIBILITY		
Transport Protocol Neutrality	TCP	TCP/UDP
Extensibility of Functionality	Vendor Specified	Yes, IANA
Ease of Customization	Harder	Easier
INTEROPERABILITY		
Version Compatibility	Yes	Unknown
SCN Signaling Interoperability	Better	Worse
EASE OF IMPLEMENTATION		
Protocol Encoding	Binary	Text

3.4. Signaling System 7

SS7 is the set of protocols used for call setup, teardown and maintenance in the PSTN. It is the current suite of protocols used in the North American public network to establish and terminate telephone calls. SS7 is implemented as a packet switched network, which typically uses dedicated links, nodes and facilities. In general, SS7 is a non-associated, common channel, out-of-band signaling network – allowing switches to communicate during a call. However, SS7 signaling may traverse real or virtual circuits on links that also carry voice traffic. The goal of this section is to provide a concise overview of the signaling functions and interfaces of SS7, because they impact the implementation of internetworking between IP based telephony and the PSTN.

3.4.1. SS7 Network Topology

SS7 network topologies are constructed using three types of components that are arranged throughout the network in a manner that offers maximum reliability, flexibility and speed for accomplishing several instrumental tasks in providing telephone service. These elements are Service Switching Points (SSPs), Signaling Transfer Points (STPs) and Service Control Points (SCPs). An SSP is the local telephone exchange, which employs subscriber circuits and trunks connecting to other exchanges. An STP offers transfer and routing services of SS7 messages originating at the SSP. An SCP offers access to the telephone companies' databases via the STP network.

3.4.2. Integrated SS7 and IP

An SS7-IP interface coordinates the SS7 view of IP elements and IP view of SS7 elements. There are three methods to integrate an IP based network with SS7, each with its advantages and shortcomings. The first approach is to give the access concentrator the ability to interface directly to SS7. The advantage of this approach is that it keeps all the functionality of the SS7/IP integration contained within a single

device, making it the most manageable solution. The limitation, however, is scalability, because each access concentrator would require its own connection to the SS7 network.

A simple way to gain an SS7 connection for access concentrators is to use an external converter to handle the translation of SS7 to PRI signaling. On the other hand, the converter is also limited in scalability. The final technique bridges the existing PSTN and IP networks, translating the signaling information between the two incompatible network types. Unlike simple converters, however, gateways provide added intelligence for security and control and can be equipped for greater redundancy, resiliency, and scalability. This disadvantage of an SS7 gateway is that it uses a special (and currently nonstandard) interface protocol to talk to the access concentrator.

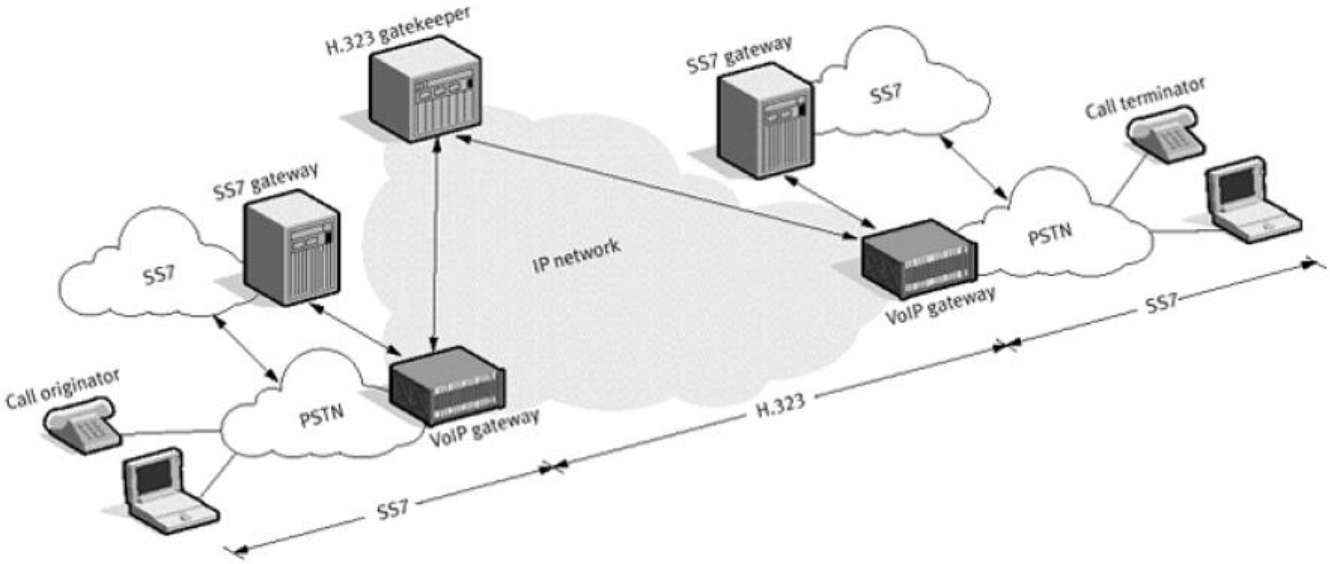

Figure 3.9 SS7 based VoIP Network

This industry is moving toward converged network infrastructure to provide a more efficient and effective way of handling increased call volumes as well as delivering new, enhanced services. The integration of SS7 and IP is an important evolutionary step that will also provide significant short-term benefits. Figure 3.9 illustrates a type of VoIP network employing an SS7 to IP gateway. SS7 provides the call control on either side of the traditional PSTN, while H.323 provides call control in the IP network. The media gateway provides the circuit to voice conversion.

CHAPTER FOUR
VOICE CODERS

This section deals with the conversion of the analog voice-signal into a (compressed) digital signal and vice versa. The requirements for the digital signal are a high quality and a low bit rate. Coding methods which produce a very low bit rate often need a fast hardware to compute the bit stream in an appropriate time. In contrast, coding methods that enable a high voice-quality with the drawback of a high bit rate need less computational power but require a faster network. Therefore it is very important to adjust the coding method to the given equipment.

The main speech coding techniques can be divided into waveform codecs, source codecs and hybrid codecs (Woodard, 2002). Figure 4.1 shows the varying of bit rate and speech quality for the three classes.

4.1. Waveform Codecs

Waveform codecs quantize the original signal without any knowledge of the type of the signal. Therefore they are signal independent and work also for non-speech signals. Low computational costs and a good speech-quality are the advantages of waveform codecs.

The most popular waveform codec is Pulse Code Modulation (PCM). At the beginning the amplitude and the bandwidth of the analog signal is limited. Then the amplitude of the signal is measured with a fixed rate (sampling rate). The duration

between two samples is defined by the Shannon-theorem, which states that a signal can be sampled without a loss of information when the sampling-frequency is at least twice than the highest frequency that occurs in the analog signal. A higher sampling rate does not lead to a higher quality of the signal and is therefore not necessary. So the sampling rate for a signal that is limited to 4 kHz should be at least 8 kHz. The precision of the measured value is defined by the number of bits used for one value. The error that occurs by converting the analog value to a digital one is called quantization error. Since this error is more significant for lower signal levels, often nonlinear converters are used. This means that digital values at lower levels lie closer together than values at higher levels. Two nonlinear quantization schemes are standardized: μ-law is used in North America, Japan and South Korea while A-law is used in the rest of the world.

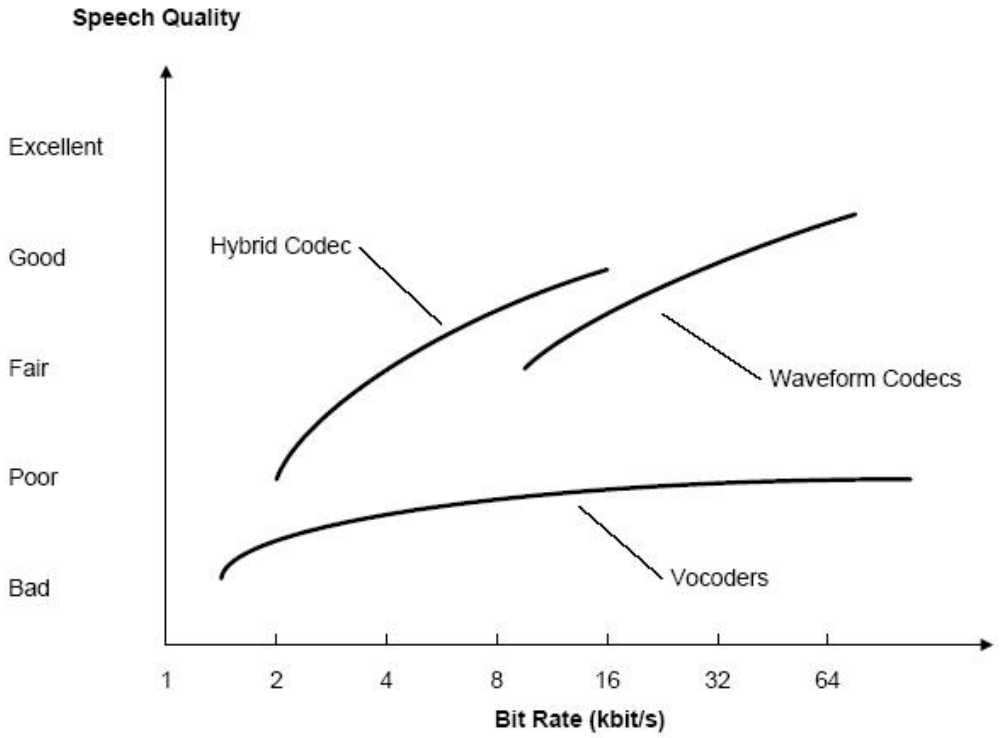

Figure 4.1 Speech Quality versus Bit Rate for Common Classes of Codecs

To reobtain the analog signal, the quantized values are converted to analog ones with the same sampling rate as in the encoding phase. To smooth the resulting signal, it is sent through a low pass filter. The ITU-T defines G.711 as a PCM coder for narrowband signals (3,4 kHz). G.711 is part of H.323 and produces a bit rate of 64 kbps with a delay of about 5 ms. The speech-quality of G.711 is often used as a reference for other coding methods.

A variation of PCM is the Differential Pulse Code Modulation (DPCM) where the PCM values are encoded as differences between the previous and the predicted current value. Because this difference has a lower variance than the original signal, the number of bits required for an audio signal can be reduced by about 25 percent.

An improvement of DPCM is the Adaptive Differential Pulse Code Modulation (ADPCM). This method tries to adjust the parameters for prediction and quantization to the input-signal. This leads to lower bit rates with the drawback of higher computational costs. An example for an ADPCM-standard is G.721 with 32 kbps, that was later integrated into the ITU-T G.726 standard.

All the codecs described above analyze the signal in the time domain. But there exist waveform codecs that also use the frequency domain. An example of such a codec is Sub-Band Coding (SBC). In this method the signal is split into several frequency bands. Each of these bands is then coded using PCM, DPCM or ADPCM. The splitting into frequency bands is done because some frequencies are perceptually more important than others. These frequency bands can be coded with a higher bit rate and therefore lower noise which leads to an improved quality of the speech-signal. The ITU-T standard G.722 specifies an SBC codec for 7 kHz signals. It produces 48, 56 or 64 kbps with a delay of about 3 ms.

4.2. Source Codecs

Source codecs for speech are called vocoders. In contrast to waveform coders, they do not try to reproduce the original signal but to detect relevant parameters in the signal. In doing so, the content is preserved while the sound of the speaker's voice is lost. Because just the detected parameters are transferred, the bit rate of vocoders is significant lower than for waveform coders. Despite of this advantage, vocoders are not used in IP telephony because the speakers can not be identified by the sound of their voice. Source codecs are mainly used in military applications where low bit rates are more important than natural sounding speech.

An example for source coders is so-called Linear Predictive Coding (LPC) vocoders. They process an LPC analysis, where the signal is split into small segments. It is assumed that the characteristic of the signal does not change within a segment. The predicted values in a segment are generated by using a linear combination of the previous values. The coefficients of the linear combination (LPC coefficients) are specified by the minimization of the difference between sampled and predicted value. The quality of an LPC-vocoder depends on the number of LPC coefficients and the duration of a segment.

4.3. Hybrid Codecs

Hybrid Coders use methods of waveform and source codecs, whereas they try to minimize their disadvantages. Hybrid codecs usually have bit rates between 2 and 16 kbps.

The Residual Excited Linear Prediction Coding (RELP) uses a method where the signal is represented by the coefficients of an LPC analysis and a wave form coded residual signal. The LPC coefficients are recalculated every 10 to 20 milliseconds. RELP coders usually have bit rates between 4.8 and 16 kbps. A variant of RELP is used by the Global System for Mobile Communications (GSM). This codec has a bit rate of 13 kbps.

To reduce the bit rate, Code Excited Linear Predictive Coding (CELP) does not quantize each sampled value for its own (scalar quantization) but uses vector quantization. This method combines several sampled values to a vector. The coder compares this vector with reference vectors listed in a codebook and transfers the code of most similar reference vector.

There are several standardized hybrid codecs. Low Delay CELP (LD-CELP) is a CELP-coder that is defined as G.728. Its advantage is a very low processing delay of about 2 ms. Its bit rate is 16 kbps. Another standard is G.732.1 with bit rates of 6.3 kbps and 5.3 kbps, respectively. This standard either uses Multipulse Maximum Likelihood Quantization (MP-MLQ) or Algebraic CELP (ACELP) as coding schemes. Finally, G.729 uses Conjugate Structure - Algebraic Code Excited Linear Prediction (CS-ACELP) as coding method. Its bit rate is 8 kbps. All of the standards mentioned above are part of H.323.

4.4. Mean Opinion Score

The quality of a reconstructed signal is an important characteristic of a voice codec. Therefore it is necessary to measure the quality. While humans perform subjective tests, a computer measures the quality of a signal in relation to the original signal in an objective way. A subjective measuring for the evaluation of a codec is the Mean Opinion Score (MOS). A MOS-test of a codec is performed as follows. A group of people listens to different samples of voice recordings and rates each recording with one (bad quality) up to five points (excellent quality). The mean of the ratings of all people states the MOS score of a codec. Table 4.1 shows the MOS scores for some ITU-T codecs.

Table 4.1 ITU-T codec MOS values

Compression Method	Bit Rate	Framing Size	MOS Score
G.711 PCM	64 kbps	0,125 ms	4,1
G.711 PCM	64 kbps	0,125 ms	4,1
G.726 ADPCM	32 kbps	0,125 ms	3,85
G.728 LD-CELP	15 kbps	0,625 ms	3,61
G.729 CS-ACELP	8 kbps	10 ms	3,92
G.729a CS-ACELP	8 kbps	10 ms	3,7
G.723.1 MP-MLQ	6,3 kbps	30 ms	3,9
G.723.1 ACELP	5,3 kbps	30 ms	3,65

4.5. Capacity of IEEE 802.11b Wireless LAN supporting VoIP (G.711 vs G.729)

In this section we evaluate the capacity of an IEEE 802.11b network carrying voice calls in a wide range of scenarios, including varying delay constraints, channel conditions and voice call quality requirements. We consider both G.711 and G.729 voice encoding schemes and a range of voice packet sizes.

We first present an analytical upper bound and, using simulation, show it to be tight in scenarios where channel quality is good and delay constraints are weak or absent. We then use simulation to show that capacity is highly sensitive to the delay budget allocated to packetization and wireless network delays. We also show how channel conditions and voice quality requirements affect the capacity. Selecting the optimum amount of voice data per packet is shown to be a trade-off between throughput and delay constraints: by selecting the packet size appropriately given the delay budget and channel conditions, the capacity can be maximized.

Unless a very high voice quality requirement precludes its use, G.729 is shown to allow a capacity greater than or equal to that when G.711 is used, for a given quality requirement.

4.5.1. The Scenario

Transmitting voice over wireless communication links has been in widespread use for many years - this is clearly shown by the huge take up of mobile telephony around the world. Cellular networks provide coverage by locating antennae every few kilometers, either on purpose-built masts or on top of existing buildings. While these provide coverage over a large area, reception inside buildings is often very poor compared with the signal quality available outside.

In this section we investigate whether IEEE 802.11 devices (the price of which has fallen significantly in recent years) could be used to create a low-cost wireless voice network that could be integrated with wired Voice over IP networks, or connected directly to cellular networks.

The scenario we are considering is shown in Figure 4.2. The network comprises a single IEEE 802.11 basic service set (BSS) with one access point (AP), and a number of wireless users. The AP is connected to a wired network, to which other users are directly connected. Voice calls take place between a user in the BSS and a user connected to the wireless network (e.g. between users A and A').

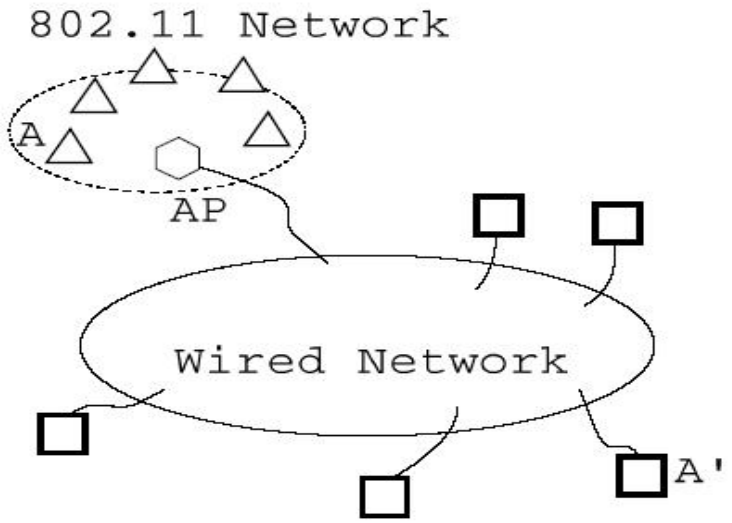

Figure 4.2 Network Scenario

We focus on the capacity of the wireless network as the principal metric of interest; this is important not only for deployment of these networks, but also as a means of comparing protocols and techniques in future works. We define capacity in this context to be the maximum number of simultaneous, bidirectional calls that can be supported, subject to a minimum voice quality requirement.

Voice traffic is generated by packetizing the output of a voice encoder (we consider both G.711 and G.729 schemes, without the use of silence suppression), creating packets each containing a fixed amount of voice data; we consider 10, 20, 30 or 50ms of voice data per packet. These packets are transmitted over the network using RTP over UDP/IP. As an example of parameters used in current VoIP implementations, Cisco 7960 VoIP phones can use either the G.729 or G.711 codecs, and the default amount of voice data per packet is 20ms.

The various IEEE 802.11 protocols provide a ‘best effort’ service: packets are carried without any delay guarantee, and may be dropped within the network. Here we consider that the wireless network operates using the Distributed Coordination Function (DCF) MAC protocol (without the RTS/CTS mechanism enabled); although the Point Coordination Function (PCF) protocol was designed to better handle stream-type traffic, this has not been widely implemented. The DCF protocol is based on CSMA/CA, whereby stations must detect that the medium is idle before transmitting. Because of the possibility of collisions and/or decoding errors (e.g. due to poor channel conditions), frames may require multiple transmissions before being successfully received. After an unsuccessful transmission (indicated by the lack of acknowledgment), the random back off duration is selected from an exponentially increasing range. After a certain number of retransmissions, the sending station will drop the frame. Hence, the MAC protocol introduces significant randomness in the packet delay as well as the possibility of packet loss.

Although the play out buffer at the receiver reduces the effect of variation in delay, and packet loss concealment (PLC) can mitigate the effects of packet loss, the combined effects of delay (i.e. the end-to-end delay as set at the play out buffer) and

packet loss (either within the network, or at the play out buffer due to excessive delay) on the quality of the voice call must be taken into consideration. For this, we use the "Mean Opinion Score" (MOS) metric. Call quality is rated on a scale of 1 (worst) to 5 (best) - typically, a score of 3.6 or higher is considered satisfactory.

In the following section, we present an analysis of the MAC protocol which, by making simplifying assumptions, leads to an upper bound on the capacity of the network. The tightness of this bound for error-free scenarios are then evaluated by simulation. In section 4.5.3., we use simulation to show how delay constraints and channel conditions affect the capacity of a network.

4.5.2. Mathematical Analysis of Network Capacity

In this section we present an upper bound on the network capacity, by making certain assumptions about the performance of the network. These assumptions are that:

i) no collisions occur
ii) frames are always received without errors
iii) all frames arrive at the play out buffer before their respective play out deadline

We refer to such a scenario as being throughput constrained.

The analysis is based on the following argument. At any point in time, one of the following is taking place on the wireless network:

- The frame sequence for the transmission of data from the AP to the stations is ongoing. We define a frame sequence as the transmission of the voice data frame (with transmission time T_{VOICE}), the Short Interframe Space (SIFS), the transmission of the acknowledgement (T_{ACK}) and the DCF Interframe Space (DIFS) following the acknowledgement.

- The frame sequence for transmission of data from a station to the AP is ongoing.

- The medium is idle, and the AP is counting slots as part of a back off procedure. Recall that all stations must wait ('back off') for a random number of idle slots following each transmission.

- The medium is idle, and the AP is not counting down idle slots.

We consider an arbitrarily long period of time T seconds during which N calls are in progress. Let R be the number of packets generated by each encoder per second.

Of the T seconds, the time required for the frame sequences for transmissions to and from the AP is given by $2NRT(T_{VOICE} + SIFS + T_{ACK} + DIFS)$. Similarly, the minimum amount of idle time required for the AP to complete its back off procedures is $\left\lfloor \sum_{i=1}^{i=NRT} CWi \right\rfloor$ x T_{SLOT}, where T_{SLOT} is the slot duration, and CWi is the number of slots picked from a uniform distribution over $(0,CW_{MIN})$ for the i th transmission. For large T this expression converges to $NRT(T_{SLOT}xCW_{MIN}/2)$. (Hole & Tobagi, 2004).

Considering only the first three possible uses of the time then, we require that, in order to support the offered load,

$$T \geq [2NRT(T_{VOICE} + SIFS + T_{ACK} + DIFS)] + [NRT(T_{SLOT} \text{ x } CW_{MIN}/2)]$$

This expression becomes an equality when the amount of idle time which is not counted towards the AP's back off, T_{IDLE}, is zero. It is very hard to find an expression for T_{IDLE} since it depends both on the load on the network and the way in which stations carry out their back off procedures. However, we argue that as the load approaches capacity, this value will become very small. Since devices may count down back off slots concurrently with each other, the most efficient use is made of the network when idle slots are counted by many devices simultaneously. In

this network, one single device (the AP) is transmitting half of all the traffic, and so has the greatest single requirement for non-overlapping idle time. We therefore argue that as the load (i.e. the number of calls N) increases, stations will take advantage of the idle time required by the AP to fulfill all of their back off requirements, thereby minimizing the amount of time during which the medium is idle, and not being counted towards the AP's back off requirements. By making the assumption that, at capacity T_{IDLE}=0, we obtain the upper bound on the value of N given in

$$N = \left\lfloor \frac{1}{R(2(T_{VOICE} + SIFS + T_{ACK} + DIFS) + (T_{SLOT} \times CW_{MIN}/2))} \right\rfloor$$

For 802.11b, CW_{MIN}, SIFS, T_{SLOT} and DIFS are respectively 31, 10us, 20us and 50us. Assuming a data rate of 11Mbps, T_{VOICE} and T_{ACK} are comprised of the component times shown in Table 4.2.

Table 4.2 Component Times

COMPONENT TIMES OF T_{VOICE} & T_{ACK} AT 11MBPS

T_{VOICE}	PLCP Preamble & Header	192.0us
	MAC Header + FCS	20.4us
	IP/UDP/RTP header	29.1us
	Voice Data	(Voice octets ×8/11) us
T_{ACK}	PLCP Preamble & Header	192.0us
	ACK Frame	10.2us

In order to assess the tightness of this upper bound, we compared the value of the upper bound with the capacity obtained by simulation. These results are shown in Table 4.3 (the calculated upper bound is in parentheses). Note that the capacity values shown are independent of any quality requirement: no packet loss occurs at or below capacity, while packet loss due to queue overflow at the Access Point for higher loads is extremely high (typically 10% and higher).

These results support our assertion that T_{IDLE} tends to 0 as the load increases, and also show that the effect of collisions on capacity is small. Indeed, we observed that the percentage of transmissions involved in a collision ranged from around 1.5% to 4% for the AP's transmissions and from 2% to 9% for the other stations' when the network is operating at maximum capacity. Because of the low rate of collisions, no frames were dropped by the MAC due to an excessive number of retransmissions (using the default retry limit of 7).

Table 4.3 Capacity Results

CAPACITY RESULTS FROM SIMULATION (ANALYSIS)

	Voice Data per frame			
	10ms	20 ms	30ms	50ms
G.711	6 (6)	12 (12)	17 (18)	25 (26)
G.729	7 (7)	14 (14)	21 (22)	34 (35)

These results also highlight the effect of the large overhead at the MAC and physical layers. The majority of this overhead comprises the transmission time of the PLCP header and preamble (accounting for over 50% of the total time) and idle time, when no station is transmitting (typically 20-30% of the time). As a result, the capacity is affected more by the rate of packet generation R (and hence the amount of voice data per packet) than by the bit rate of the encoder. For example, although the output bit rate of a G.711 encoder is eight times that of a G.729 encoder, the reduction in capacity when G.711 is used is less than 50%.

Note that the use of the short PLCP preamble can significantly reduce this overhead, leading to an increase in capacity: our simulations showed an increase of around 25-50%. However, since this capability is optional, it cannot be relied upon, and for the remainder of the paper we assume the use of the long preamble only.

4.5.3. The Effects of Delay Constraints, Non-Ideal Channel Conditions and Quality Requirements

In the previous section, it was assumed that the play out deadline was met by all packets and that no packets were received with error(s). The results showed that, to maximize capacity under such assumptions, G.729 is to be preferred over G.711 and those packets should contain as much voice data as possible. However, when considering the impact of poor channel conditions, call quality requirements and/or delay constraints, many further issues must be considered.

Because of the coding and modulation used in 802.11b, packet error rates are highly dependent on packet length: long packets are more susceptible to error than short packets. As the channel quality deteriorates, more retransmissions are required, resulting in a lower throughput-constrained capacity (more time is required per successful transmission), and higher per-packet delays.

Furthermore, although G.729 has been shown to permit greater capacity, the quality of such calls is limited by the encoding scheme's intrinsic MOS of 3.65, compared to 4.15 for G.711. The G.729 algorithm also requires a 5ms look ahead, delaying packets by an additional 5ms compared to G.711.

Finally, for a given encoding scheme, the packetization delay is higher for larger packets. The loss of such packets is also harder to conceal (using PLC) than smaller packets.

In this section we first evaluate the effect of a delay constraint in the context of ideal channel conditions, and then present results showing the combined effects of delay constraints and non-ideal channel conditions.

4.5.3.1. The Delay Constraint

The delay constraint is set by the play out buffer at the receiver, which drops packets that have incurred an excessive end-to-end delay and arrive after their play out deadline. Throughout this section, we assume a fixed play out deadline for packets, allowing a maximum of 150ms end-to-end delay. This allows for the encoder's algorithmic and packetization delay, the propagation delay through the wired network, queuing delay at the wireless network interface, channel access delay and propagation time over the wireless medium.

Since an end-to-end delay of 150ms causes a negligible decrease in quality, the MOS score is reduced from the intrinsic value only by the effect of loss. The effects of loss on MOS are given for G.729 and G.711, for 10ms and 20ms packets. The loss rates corresponding to MOS values of 3.6 and 4.0 are shown in Table 4.4. We apply the values for 20ms packets to 30ms and 50ms packet sizes considered here.

Table 4.4 Maximum Loss Rates (%) For Minimum MOS Requirements

	Minimum MOS	
	4.0	3.6
G.711 (10ms)	1	4.9
G.729 (10ms)	N/A	0.33
G.711 (20ms)	1	3
G.729 (20ms)	N/A	0.19

In Figures 4.2 and 4.3, we plot the complementary cumulative distribution function (CCDF) for the sum of the delay incurred within the wireless network and the packetization delay for several scenarios with ideal channels, using G.711 and G.729 respectively. On these plots, the y-axis gives the probability of a packet incurring a delay greater than the value on the x-axis. (We plot here the delay CCDF for packets transmitted from the wired nodes to the wireless stations, since the delay

in this direction is almost always higher than in the opposite direction, due to queuing at the AP).

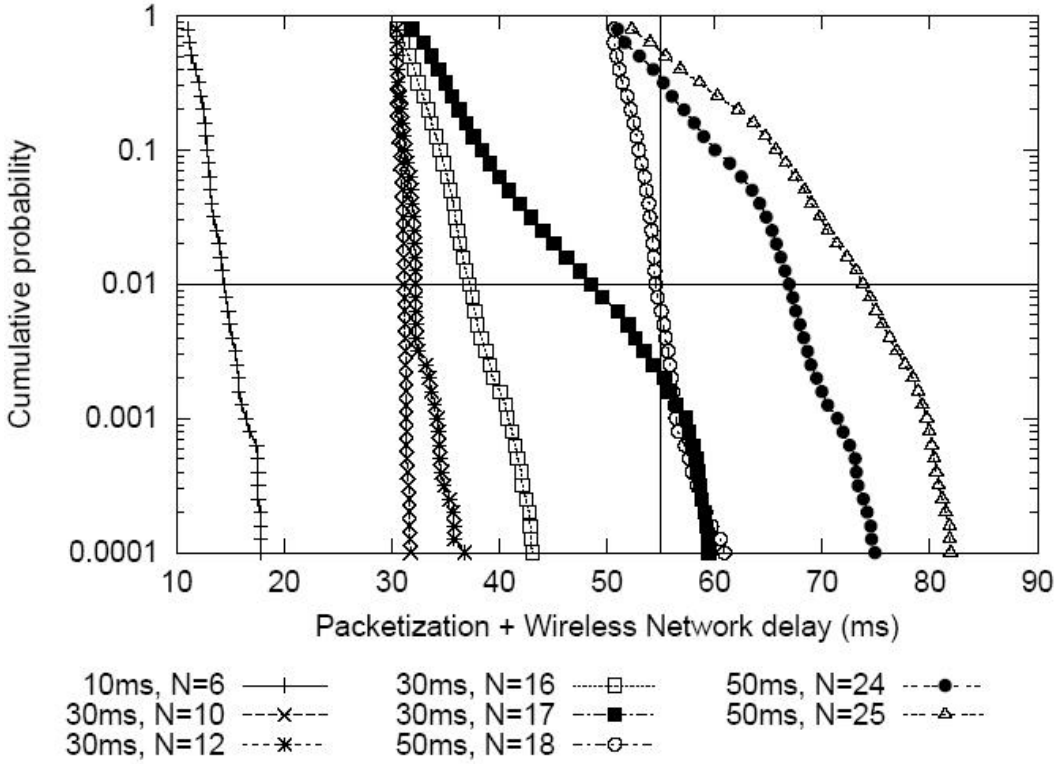

Figure 4.3. CCDF for delay for G.711 with various packet sizes and number of calls

These figures can be used to determine the fraction of packets that will be dropped at the play out buffer due to late arrival as a function of the delay budget allocated for packetization (including the 5ms look-ahead required by G.729) and queuing and transmission in the wireless network (assuming the delay in the wired network to be constant). For example, using G.711 with 50ms packets, and with a delay budget of 55ms, 18 concurrent calls can be supported with a loss rate of less than 1% (as indicated on Fig. 4.3).

Using the loss requirements specified in Table 4.4 together with the delay statistics, we can evaluate the capacity for various packet sizes and delay budgets. Clearly there is a tradeoff between the throughput constraint (favoring larger packets) and the delay constraint (favoring smaller packets). In Table 4.5 we show the

capacity that may be attained by selecting the optimum packet size appropriate to the wireless network and packetization delay budget. The optimum packet size (in ms) is given in parentheses following the capacity value. (Note that G.729 cannot be used for a minimum call quality MOS requirement of 4.0)

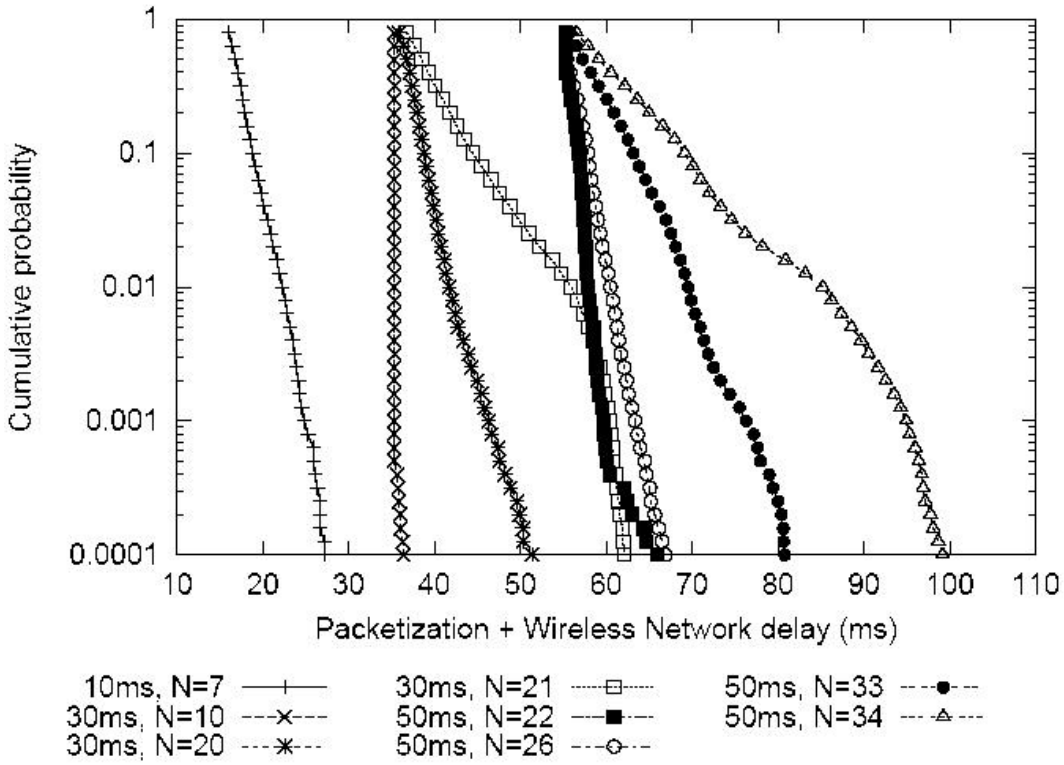

Figure 4.4. CCDF for delay G.729 with various packet sizes and number of calls

Considering a MOS requirement of 3.6, it can be seen from the table that G.711 and G.729 provide similar capacity where the delay budget is less than 70ms: in general, G.711 is constrained by throughput, G.729 by delay. In particular, the 5ms look-ahead time required by a G.729 encoder makes the delay constraint particularly stringent in such scenarios. Where the delay budget is higher, throughput constraints dominate in both cases, and G.729 provides a much greater capacity.

Comparing the capacity for MOS requirements of 3.6 and 4.0, we are clearly constrained to the lower capacity of G.711 for the higher MOS value. However, comparing the capacity using G.711 with different quality requirements, we observe that the difference is minimal even though the maximum loss rate is reduced from 4.9% to 1%.

Table 4.5 Capacity of an 802.11b Network (Assuming Ideal Channel)

Delay budget	MOS = 3.6		MOS = 4.0
(ms)	G.711	G.729	G.711
≤ 10	0	0	0
20	6 (10)	6 (10)	6 (10)
30	11 (20)	11 (20)	11 (20)
40	16 (30)	16 (30)	16 (30)
50	17 (30)	20 (30)	17 (30)
60	23 (50)	23 (50)	22 (50)
70	25 (50)	31 (50)	24 (50)
80	25 (50)	33 (50)	25 (50)
≥ 90	25 (50)	34 (50)	25 (50)

4.5.3.2. Non-ideal Channel Conditions

To assess the effect of non-ideal channel conditions, we use a simplified channel model, represented by a constant Bit Error Rate (BER). We assume that the channels between all pairs of nodes are subject to this BER value, and that all bit errors occur independently. We maintain the abstraction of the channel at the BER level, rather than consider the Signal-to-Noise Ratio (SNR) at the receiver, since the SNR to BER mapping is implementation-dependent. Finally, we assume that, unless a collision occurs, the PLCP preamble and header are received and decoded correctly.

As has been described, poorer channel conditions, leading to higher BER values, cause an increase in per-packet delay. To illustrate this, we plot in Figure 4.5 the delay CCDF for a network carrying 10 calls using 30ms G.711 packets for BER values ranging from 0 to 10^{-4}. The effect on the throughput constraint is shown in

Table 4.6, which lists the throughput constrained capacity (i.e. without any delay constraint) for a minimum MOS requirement of 3.6. As in the ideal channel case, packet loss only occurs at the AP's interface queue.

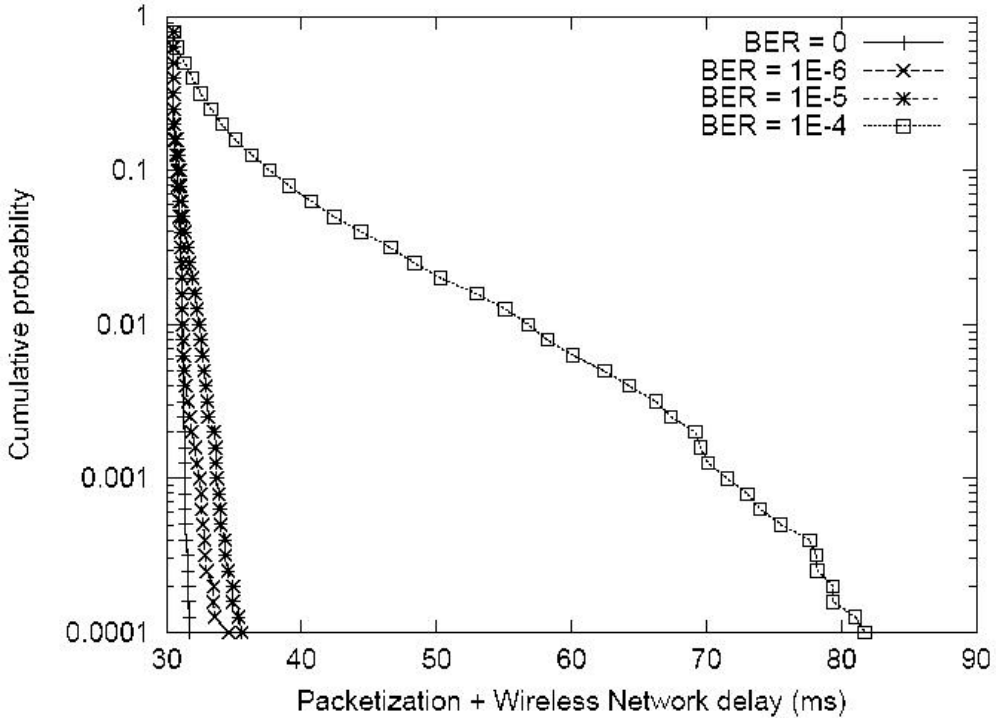

Figure 4.5. CCDF for delay for 30ms G711 packets, 10 calls

Table 4.6 Throughput Bounded Capacity For Various Ber Values

	Voice Data per frame (ms)							
	G.711				G.729			
BER	10	20	30	50	10	20	30	50
0	6	12	17	25	7	14	21	34
10^{-6}	6	12	17	25	7	14	21	34
10^{-5}	6	12	16	24	7	14	20	33
10^{-4}	5	9	12	15	6	12	18	29
2×10^{-4}	4	7	8	7	5	11	16	25

We observe that for BER $\leq 10^{-6}$, the packet error rate for both G.711 and G.729 is so low that the difference in capacity between such a channel and an error-free channel is negligible. For $10^{-6} < \text{BER} \leq 2 \times 10^{-4}$, the capacity decreases by varying degrees, depending on the amount of voice data contained in a packet (longer packets are more susceptible to errors, hence are more likely to require retransmission). For BER $\geq 10^{-3}$, not even one voice call can be supported by the network due to the very high packet error rate, which causes the MAC to retransmit frames so often that the probability of a frame being dropped due to excessive retransmissions becomes significant.

As we have already seen for the ideal channel case, adaptation to the delay constraint can be used to maximize capacity for a given quality (MOS) requirement. If stations can adapt jointly to the channel conditions and delay budget, the network capacity can be similarly maximized for non-ideal channel scenarios. In Table 4.7, we show the maximum capacity and corresponding packet size for the case of BER = 10^{-4} for delay budgets up to 110ms. Figures 4.6, 4.7 and 4.8 show the maximum capacities that can be achieved using G.711 and G.729 with optimum adaptation, for MOS requirements of 3.6 and 4.0 for the range of BER values considered.

Table 4.7 Capacity of an 802.11b Network (BER = 10-4)

Delay budget	MOS = 3.6		MOS = 4.0
(ms)	G.711	G.729	G.711
≤ 10	0	0	0
20	4 (10)	5 (10)	4 (10)
30	7 (20)	6 (10)	5 (10)
40	9 (30)	11 (20)	7 (20)
50	10 (30)	16 (30)	9 (30)
60	11 (30)	17 (30)	10 (30)
70	11 (30)	21 (50)	10 (30)
80	11 (30)	27 (50)	11 (30)
90	11 (50)	28 (50)	11 (30)
100	13 (50)	29 (50)	11 (30)
110	13 (50)	29 (50)	11 (50)

From these figures we observe that, for a 3.6 MOS requirement, G.729 provides equal or greater capacity than G.711 in all cases. In particular, for a given BER, G.729 requires fewer retransmissions due to channel errors than G.711, since it generates smaller packets and has a correspondingly lower packet error rate. Although fragmentation can (in principle) be used to create smaller packets, and hence lower the effective packet error rate, its use in this application is limited by the 802.11 specification which requires a fragmentation threshold of no less than 256 bytes. In fact, we have observed through simulation that fragmentation cannot improve the capacity beyond that achievable with optimum packet size selection.

For most channel conditions (BER < 10^{-4}) the delay is dominated by the packetization delay alone; then, the optimum packet size can be determined without knowledge of the state of the channel by selecting the highest packet size smaller than the delay budget. However, as can be seen from Table 4.7, this is not the case for higher BER values.

4.5.3.3. Discussion

In presenting our results so far, we have assumed a fixed play out scheme, resulting in an end-to-end delay of 150ms. However, this value may not always be the most suitable choice. The additional degradation in MOS that would result if this delay were increased by less than 50ms is shown to be very small, e.g. increasing the allowed delay by 20ms to 170ms results in MOS degradation of less than 0.1.

Since we have shown that the capacity is highly sensitive to the delay constraint, relaxing the delay constraint by increasing the play out deadline would allow more calls to be supported with minimal degradation in MOS. In order to maximize call capacity and quality, an adaptive play out scheme would then be required, the details of which are outside the scope of this thesis.

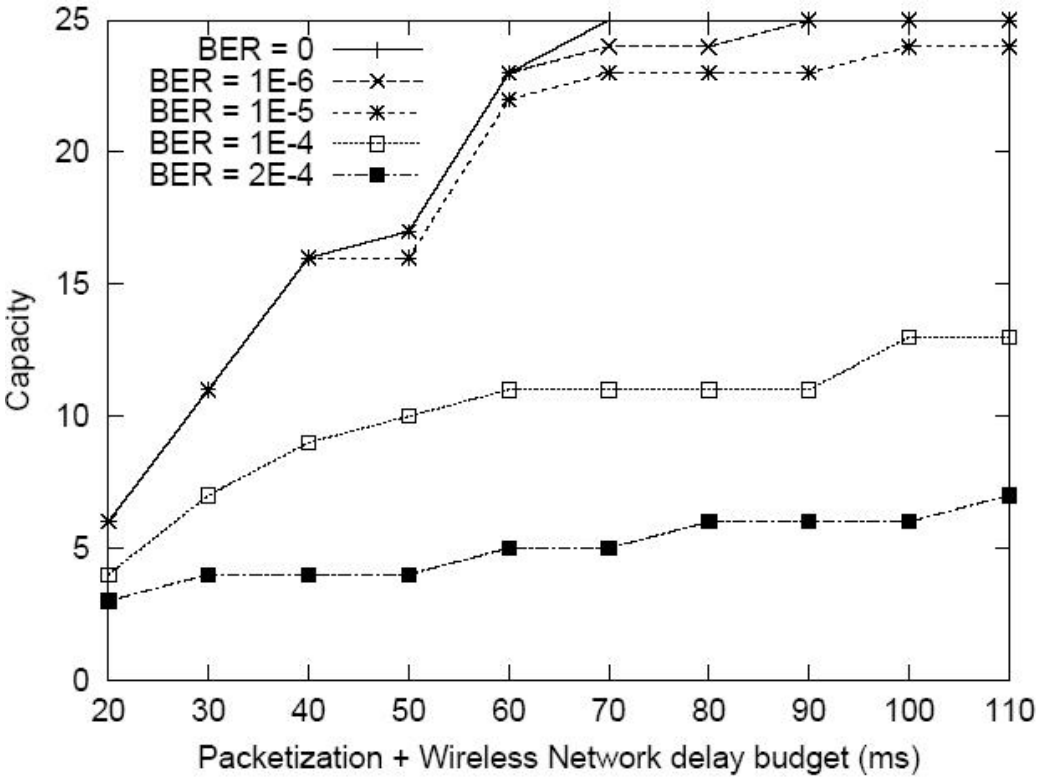

Figure 4.6. Number of G.711 calls, MOS ≥ 3.6, that can be supported

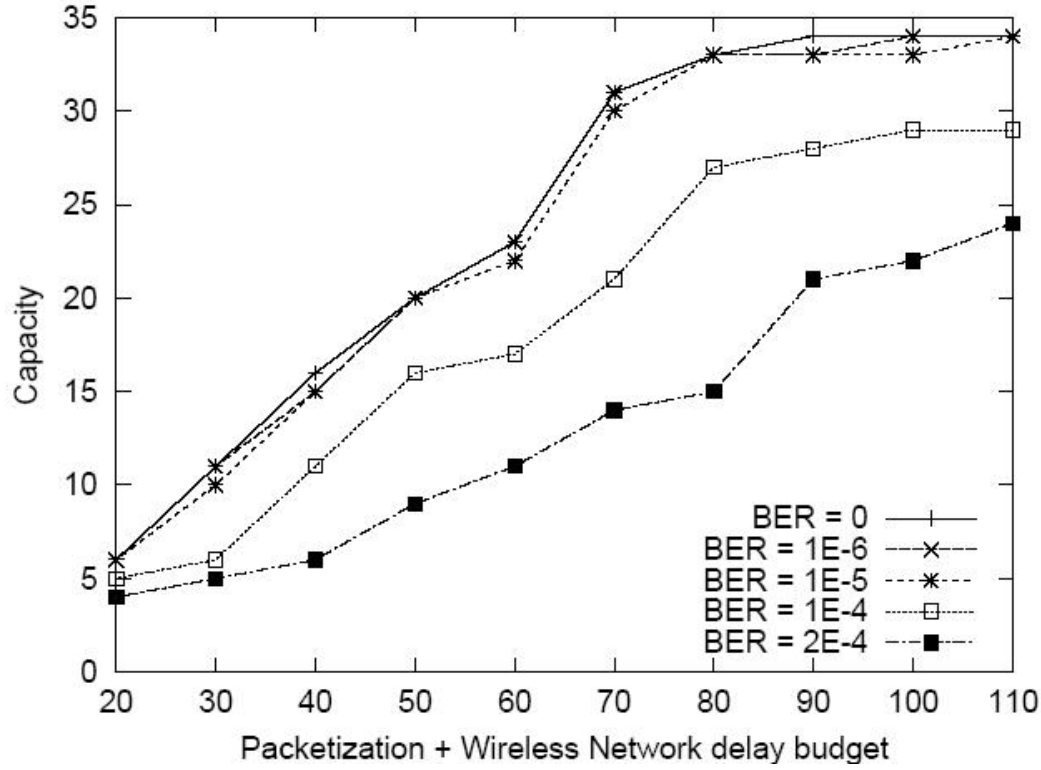

Figure 4.7. Number of G.729 calls, MOS ≥ 3.6, that can be supported

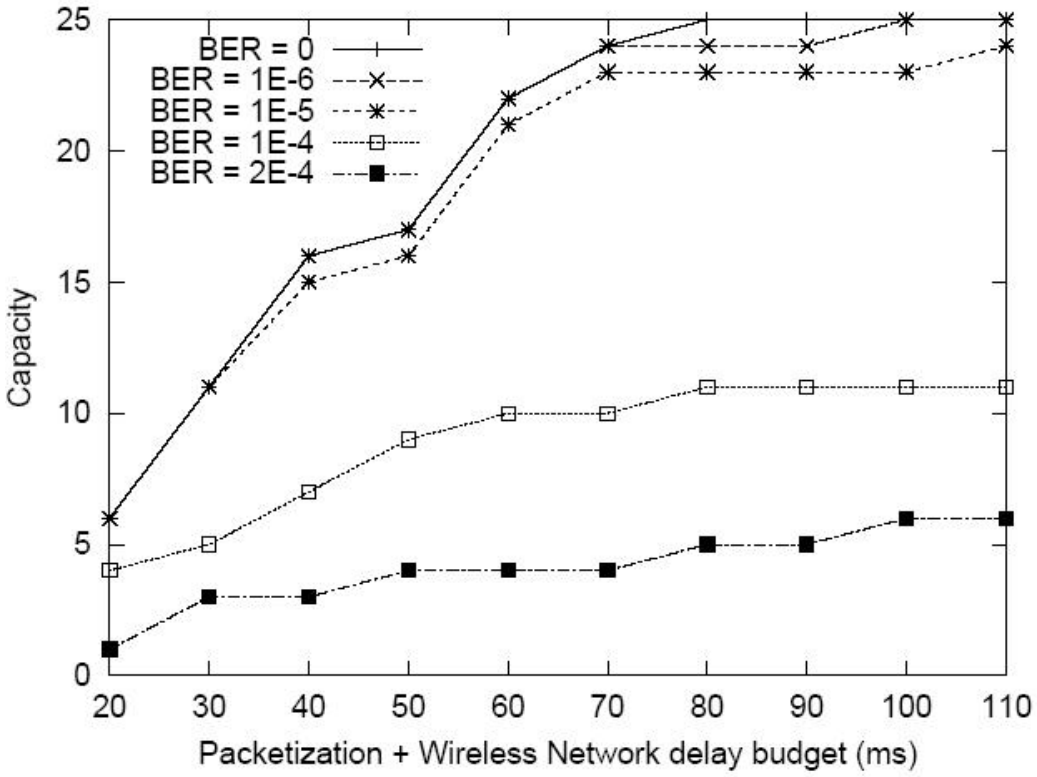

Figure 4.8. Number of G.711 calls, MOS ≥ 4.0, that can be supported

4.6. Conclusion

In this chapter we have evaluated an upper bound on the capacity of an IEEE 802.11b network carrying voice calls, and found it to be tight in scenarios where channel quality is good and delay constraints are weak or absent.

We have then shown that capacity is highly sensitive to the delay budget allocated to packetization and wireless network delays. Concerning non-ideal channel conditions, we have shown that capacity is very close to that in an error-free channel for BER values of less than 10^{-5}; in more adverse conditions capacity is reduced considerably, and is zero for channels with BER $\geq 10^{-3}$. By selecting the packet size appropriately given the delay budget and channel conditions, the capacity can be maximized; furthermore, in the majority of cases optimum packet size selection can be made without knowledge of the channel conditions.

Throughout, the use of G.729 has been shown to allow greater capacity than the use of G.711, unless a voice quality corresponding to a MOS of greater than 3.65 is required, in which case G.729 cannot be used.

CHAPTER FIVE
MULTIMEDIA TRANSPORT OVER IP

H.323 and SIP focused on the initiation, termination and the control (i.e. the signaling) of multimedia sessions. Both, H.323 and SIP use the Real-Time Transport Protocol (RTP) and the Real-Time Transport Control Protocol (RTCP) to transmit and control the media stream, respectively.

Real-time data has requirements different to non-time-critical transmissions. For example they do not need a highly reliable packet transport. Requirements of multimedia transport over IP networks are:

Sequencing: If packets are received out of order they must be reordered in real-time. Lost packets have to be detected and are not requested again.

Intra-media synchronization: It is important that the time between two packets at the receiver is the same as at the sender. Therefore, the receiver has to be informed about the amount of time that should elapse between two frames. Usually, the packets are played out at a fixed interval. An exception are silence periods, were no packets are sent.

Inter-media synchronization: Often different media types like audio and video are used. It is necessary to synchronize them so that the audio that is played out matches the video (lip-sync).

Payload identification: It is often necessary to change the encoding for the media on the fly. For example this is the case when the available bandwidth has changed). Then a mechanism is needed to identify the encoding of each packet.

Frame indication: Multimedia data are sent in logical units called frames. It must be guaranteed that it can be determined where the frame begins and ends.

RTP and RTCP fulfill these requirements. Together they also provide functionality beyond sequencing and loss detection (Schulzrinne, 2000):

Multicast-friendly: RTP and RTCP have been designed to operate in small multicast groups like a three-person phone call as well as in huge ones like multimedia broadcast events.

Media Independence: RTP provides services needed for real-time media like voice or video. For every codec a separate specification with the additional header fields and the semantics has to be defined.

Mixers and Translators: A mixer is a system that receives RTP packets from one or more sources, combines these packets to one RTP stream and sends it out. A translator takes a single media stream and converts it to another format. Then the new stream is sent away.

Quality of Service (QoS) Feedback: RTCP allows the receiver of a media stream to provide feedback on the quality of the reception. RTP sources can use this information to adjust the data rate.

Loose Session Control: The Real-Time Transport Control Protocol enables users to distribute identification information periodically. This enables other users to see who is participating in a session. Identification information can be for example the name, a phone number or the email address.

Encryption: RTP media streams can be encrypted using keys that are exchanged by (non-RTP) methods.

5.1 Real-Time Transport Protocol

RTP provides end-to-end network transport functions suitable for applications that transmit real-time data, such as audio, video or simulation data, over multicast or unicast network services. Although RTP can run over any suitable transport or network protocol, it usually runs on top of UTP and uses its multiplexing and checksum services.

5.1.1 RTP Header Format

Figure 5.1 shows the RTP header format with an optional payload. The RTP header is 12 bytes long. The length of the payload is just limited by the underlying protocol, not by RTP itself.

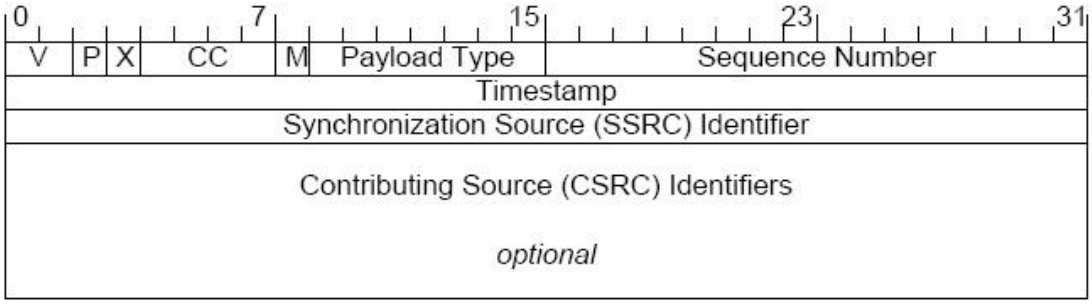

Figure 5.1 RTP header format

Version (V): Indicates the version of the protocol. The current version defined in is two.

Padding (P): If this bit is set, the packet contains one or more additional padding octets at the end which are not part of the payload. The last octet of the padding contains a count of how many padding octets should be ignored. Padding may be

needed by some encryption algorithms with fixed block sizes or for carrying several RTP packets in a lower-layer protocol data unit.

Extension (X): If the extension bit is set, the fixed header is followed by exactly one header extension. The format of this extension is defined in section 5.1.2.

CSRC Count (CC): The Contributing Source (CSRC) count contains the number of CSRC identifiers that follow the fixed header.

Marker (M): The interpretation of the marker bit is defined by a profile. It is intended to allow significant events such as frame boundaries to be marked in the packet stream. A profile may define additional marker bits or specify that there is no marker bit by changing the number of bits in the payload type field.

Payload Type: This field identifies the format of the RTP payload and determines its interpretation by the application. A profile specifies a default static mapping of payload type codes to payload formats. Additional payload type codes may be defined dynamically through non-RTP means. An RTP sender emits a single RTP payload type at any given time; this field is not intended for multiplexing separate media streams.

Sequence Number: The sequence number increments by one for each RTP data packet sent, and may be used by the receiver to detect packet loss and to restore packet sequence. The initial value of the sequence number is random (unpredictable) to make known-plaintext attacks on encryption more difficult, even if the source itself does not encrypt, because the packets may flow through a translator that does.

Timestamp: The timestamp reflects the sampling instant of the first octet in the RTP data packet. The sampling instant must be derived from a clock that increments monotonically and linearly in time to allow synchronization and jitter calculations. The resolution of the clock must be sufficient for the desired synchronization accuracy and for measuring packet arrival jitter (one tick per video frame is typically

not sufficient). The clock frequency is dependent on the format of data carried as payload and is specified statically in the profile or payload format specification that defines the format, or may be specified dynamically for payload formats defined through non-RTP means. If RTP packets are generated periodically, the nominal sampling instant as determined from the sampling clock is to be used, not a reading of the system clock. As an example, for fixed-rate audio the timestamp clock would likely increment by one for each sampling period. If an audio application reads blocks covering 160 sampling periods from the input device, the timestamp would be increased by 160 for each such block, regardless of whether the block is transmitted in a packet or dropped as silent.

The initial value of the timestamp is random, as for the sequence number. Several consecutive RTP packets may have equal timestamps if they are (logically) generated at once, e.g., belong to the same video frame. Consecutive RTP packets may contain timestamps that are not monotonic if the data is not transmitted in the order it was sampled, as in the case of MPEG interpolated video frames. (The sequence numbers of the packets as transmitted will still be monotonic.)

Synchronization Source (SSRC) Identifier: The SSRC field identifies the synchronization source. This identifier is chosen randomly, with the intent that no two synchronization sources within the same RTP session will have the same SSRC identifier. Although the probability of multiple sources choosing the same identifier is low, all RTP implementations must be prepared to detect and resolve collisions. If a source changes its source transport address, it must also choose a new SSRC identifier to avoid being interpreted as a looped source.

Contributing Source (CSRC) Identifier List: The CSRC list identifies the contributing sources for the payload contained in this packet. The number of identifiers is given by the CC field. If there are more than 15 contributing sources, only 15 may be identified. CSRC identifiers are inserted by mixers, using the SSRC identifiers of contributing sources. For example, for audio packets the SSRC

identifiers of all sources that were mixed together to create a packet, are listed, allowing correct talker indication at the receiver.

5.1.2. RTP Header Extension

If the extension bit in the RTP header is set, a variable-length header extension follows the basic RTP header. The header extension enables individual implementations to create payload-format-independent functions that require additional information. Figure 5.2 shows the RTP header extension format.

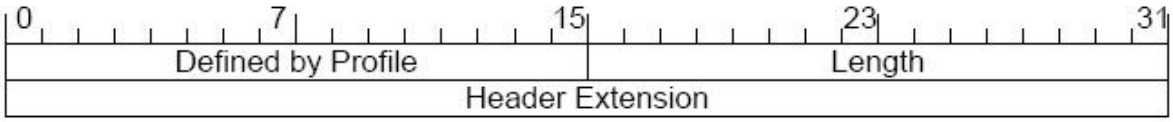

Figure 5.2: RTP header extension format

The first two octets of the header are left open for distinguishing identifiers or parameters. The format of this field has to be defined by the profile specification under which the implementations are operating. The next two octets contain the number of 32-bit words in the extension (excluding the four-octet extension header).

5.2 Real-Time Transport Control Protocol

RTCP is based on the periodic transmission of control packets to all participants of a session. These packets are using the same distribution mechanism as the data packets. Following four functions are performed by RTCP:

Provide Quality Feedback: The primary function is to provide feedback on the quality of the data distribution. The feedback may be directly useful for control of adaptive encodings, but experiments with IP multicasting have shown that it is also critical to get feedback from the receivers to diagnose faults in the distribution. Sending reception feedback reports to all participants allows one who is observing

problems to evaluate whether those problems are local or global. With a distribution mechanism like IP multicast, it is also possible for an entity such as a network service provider who is not otherwise involved in the session to receive the feedback information and act as a third-party monitor to diagnose network problems. This feedback function is performed by the RTCP sender and receiver reports, described later in this chapter.

Carrying Transport-level Identifier: RTCP carries a persistent transport-level identifier for an RTP source called the canonical name or CNAME. Since the SSRC identifier (see section 5.1.1) may change if a conflict is discovered or a program is restarted, receivers require the CNAME to keep track of each participant. Receivers also require the CNAME to associate multiple data streams from a given participant in a set of related RTP sessions, for example to synchronize audio and video.

Control Rate of Packets: The first two functions require that all participants send RTCP packets, therefore the rate must be controlled in order for RTP to scale up to a large number of participants. By having each participant send its control packets to all the others, each can independently observe the number of participants. This number is used to calculate the rate at which the packets are sent.

Control Session: A fourth, optional function is to convey minimal session control information, for example participant identification to be displayed in the user interface. This is most likely to be useful in loosely controlled sessions where participants enter and leave without membership control or parameter negotiation. RTCP serves as a convenient channel to reach all the participants, but it is not necessarily expected to support all the control communication requirements of an application.

5.2.1 RTCP Packets

Each RTCP packet begins with a fixed part (similar to that of RTP), followed by structured elements that may be of variable length. The length depends on the packet type, but it always ends on a 32-bit boundary. RTCP packets are stackable. This means, that multiple packets can be concatenated without any separators. This compound packet can be sent within one packet of the underlying protocol. The specification of RTCP defines five RTCP packet types:

Sender Reports: Users who are sending media generate sender reports. A sender report consists of a header, the sender information (timestamps, data sent so far) and report blocks. A report block contains statistics on the reception of RTP packets from a single synchronization source.

Receiver Reports: The receiver report is created by users who receive media. Each report contains one block for each RTP source.

Source Description Items (SDES): This packet type is used for session control. It contains the CNAME (see section 5.2), which is used to associate different media streams generated by the same source and for resolving conflicts in the SSRC value. SDES packets also identify the participant through its name, email and phone number.

BYE: A BYE packet is included when a user leaves a session.

APP: This type of packet can be used to add application-specific information to RTCP packets.

5.3. TCP

In the 1970s, the Defense Advanced Research Projects Agency (DARPA) of the U.S. Department of Defense (DOD) developed the Transmission Control Protocol (TCP) to provide communication among hosts manufactured by different vendors. DARPA designed TCP to work within a layered hierarchy of networking protocols, using the Internet Protocol (IP) to transfer data.

Built upon the IP layer suite, TCP is a connection-oriented, end-to-end protocol that provides the packet sequencing, error control, and other services required to provide reliable end-to-end communications. IP takes the packet from TCP and passes it along whatever gateways are needed, for delivery to the remote TCP layer through the remote IP layer.

TCP services are required to support upper-layer protocols, such as Telnet and FTP, which are part of the TCP/IP suite. TCP does not require reliability of the communication protocols below itself. Therefore, TCP functions with lower-level protocols that are simple, potentially unreliable datagram services. TCP uses IP for a lower-level protocol.

5.4. UDP

UDP provides users access to IP-like services. UDP packets are delivered just like IP packets - connection-less datagram that may be discarded before reaching their targets. UDP is useful when TCP would be too complex, too slow, or just unnecessary.

UDP provides a few functions beyond that of IP:

Port Numbers: UDP provides 16-bit port numbers to let multiple processes use UDP services on the same host. A UDP address is the combination of a 32-bit IP address and the 16-bit port number.

Checksum services: Unlike IP, UDP does checksum its data, ensuring data integrity. A packet failing checksum is simply discarded, with no further action taken.

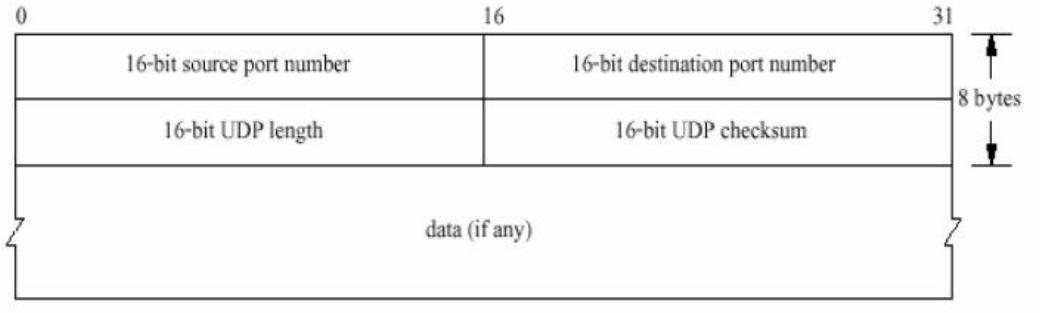

Figure 5.3 UDP Header

Source Port (16 bits): This is an optional field. If it is not used, it is set to zero. Otherwise, it specifies the port of the sender.

Destination Port (16 bits): The port, this packet is addressed to.

Length (16 bits): The length in bytes of the UDP header and the encapsulated data. The minimum value for this field is 8.

Checksum (16 bits): This is computed as the 16-bit one's complement of the one's complement sum of a pseudo header of information from the IP header, the UDP header, and the data, padded as needed with zero bytes at the end to make a multiple of two bytes. If the checksum is set to zero, then checksum is disabled. If the computed checksum is zero, then this field must be set to 0xFFFF.

5.5. IP

The IP resides within layer 3 (network layer) of the OSI Model. It provides end-to-end transport of data units through internets using connectionless services. Being connectionless, IP does not provide reliable data transfer, but this is not an issue if the upper layers provide reliability and error control.

An IP host must encapsulate data into IP headers, which are then passed to the data link (such as Ethernet). The protocol at the data link layer then encapsulates the IP header with the data into its own data unit (the datagram). The datagram is then passed down to the physical layer, where it is passed over the network as a serial bit stream (with possible encapsulation again, depending on the technology used).

For data to leave the local network, it must be sent to a router. Routers are network layer devices and are capable of processing the Ethernet and the IP headers. If the data is to be passed to another network, the Ethernet (or data link header) is stripped from the data, and the IP header is then processed.

Before transmitting the data over a port to the next network, the router must create a new IP header and place the data (consisting of the TCP header, possibly application header, and user data) into the IP header. The datagram is then given to the data link layer (which may now be X.25 ISDN, Frame Relay, or even Switched 56), and the whole process is repeated.

IP does have its limitations, the biggest being the number of addresses available. As you will see when we discuss IP addresses, there is a severe limitation in the number of addresses that IP can support. This issue has brought about the need for a replacement to IP. Internet Protocol next generation (IPng) provides a 16-byte address rather than a 4-byte one.

The primary function of IP is to provide routing information for data being transported through internets. Any error control is provided by the Internet Control

Message Protocol (ICMP), which resides at layer 3 as well. This protocol does not provide error control but merely reports errors to the originating hosts.

IP is not a requirement for TCP. The TCP protocol can use almost any network layer protocol for delivery as long as the protocol is capable of providing routing services and supports the interfaces between the two layers. Remember that the concept of layering was to allow various layers of a protocol to be changed without affecting the layers above or below it.

5.6. Summary

Protocols transmitting data over packet-based networks in real-time have requirements different to non-real-time protocols.

In the majority of cases, a highly reliable packet transport is not needed. Lost packets are not requested again. Out-of-Order packets must be re-ordered in real time. It is also important that at the receiver and at the sender the time between two packets is the same. If different types of media are transferred, they must be synchronized correctly. Other requirements for real-time protocols are frame indication and the possibility to change the encoding of the media on the fly.

RTP and RTCP fulfill these requirements. RTP is used to transport multimedia data over an IP network, while RTCP provides feedback of the quality of the transmission and conveys minimal session control information.

CHAPTER SIX
GATEWAY CONTROL

Gateways are responsible for converting packet-based audio formats into protocols understandable by PSTN systems. The aforementioned signaling protocols such as H.323 and SIP provide more services than are necessary, such as service creation and user authentication, which are irrelevant for gateways. Vendors have gravitated towards simplified Device Control Protocols (DCPs), rather than all-encompassing signaling protocols.

Figure 6.1 displays processing that must occur in a gateway to convert PSTN to IP and vice versa. The network interface in a gateway includes any hardware or software that connects the gateway to the telephone system or network. Digital signal processing is typically achieved with dedicated hardware and associated software algorithms that perform voice coding in a previous section. Specifically, the DSP subsystem compresses and decompresses speech, detects tones and silence, generates tones and comfort noise, and cancels echo. To efficiently perform vocoding, DSP implementations depend on processing entire frames of data at once. Finally, between the DSP processing and passing the data to the WAN, there are a number of packet-handling processes that must be encountered. A nontrivial amount of gateway-incurred latency is present, which affects perceived voice quality.

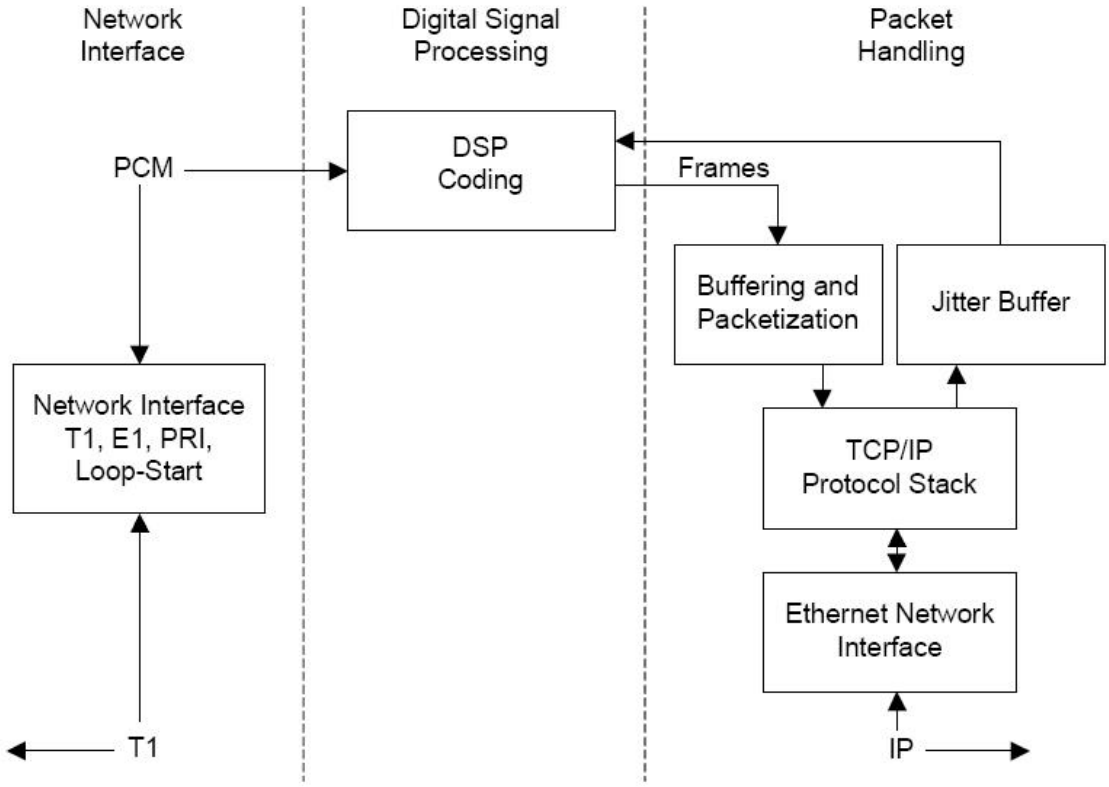

Figure 6.1. Gateway Processing

The Internet Protocol Device Control (IPDC) was a first generation DCP whose goal was to create “dumb” endpoints or gateways. It separated services from the gateway and placed network intelligence on a server. The Simple Gateway Control Protocol (SGCP) was created at approximately the same timeframe as IPDC, and it similarly resolves around intelligent servers and simple endpoints. The two standards merged to combine the best features of each and formed the Media Gateway Control Protocol (MGCP).

6.1. Media Gateway Control Protocol (MGCP)

A media gateway is a network element that provides conversion between the information carried on telephone circuits and packets carried over the Internet or over other IP networks. MGCP is an IETF standard that defines gateway control. It is the light weight telephony protocol tha aims to reduce complexity and increase reliability and interoperability for Internet telephones. MGCP also enhances security since all critical information is stored on trusted servers, thereby, MGCP devices are treated as distrusted network elements. Unfortunately, MGCP partially overlaps with signaling protocols, which obscures the boundary between signaling and gateway control.

6.2. Megaco (H.248)

While MGCP was evolving, a parallel effort was underway at ITU, which was developing H.GCP (a protocol that contains the minimal features necessary to create gateway. The ITU and IETF pooled their efforts and created the Megaco protocol (H.248). Although Megaco is still being refined, it contains all of the MGCP's functionality, plus superior controls over analog telephone lines and the ability to transport multiple commands in a single packet.

The Megaco framework could potentially enable service providers to offer a wide variety of converged telephone and data services. Media gateways will be the junctions that provide a path between switched and packet networks for voice. Megaco implementations can also be enhanced using extension methods: packages. These packages are sets of commands, related events, and statistics that can be added to a basic Megaco device. When the media gateways are initially set up for communication, a vocoder approach will normally be used. Megaco-related standards will enable support of existing and new applications of telephone service over hybrid telephone networks that will contain a mix of switched, IP and ATM technology.

Table 6.1 compares the popular DCPs. Megaco appears to incorporate the desired features of gateway control.

Table 6.1 Comparison of DCPs

Feature	IPDC	SGCP	MGCP	Megaco
ASCII-based	No	Yes	Yes	Yes
Binary	Yes	No	No	No
Trunking controls	No	No	No	Yes
Event controls	Yes	No	Yes	Yes
Packages & Extensibility	Yes	No	Yes	Yes

CHAPTER SEVEN
WIRELESS NETWORKS

Wireless networks serve as the transport mechanism between devices and among devices and the traditional wired networks (enterprise networks and the Internet). Wireless networks are many and diverse but are frequently categorized into three groups based on their coverage range:

WWAN, WLAN, and WPAN. WWAN, representing wireless wide area networks, includes wide coverage area technologies such as 2G cellular, Cellular Digital Packet Data (CDPD), Global System for Mobile Communications (GSM), and Mobitex. WLAN, representing wireless local area networks, includes 802.11, Hyperlan, and several others. WPAN, represents wireless personal area network technologies such as Bluetooth and Infrared. All of these technologies are “tetherless” –they receive and transmit information using electromagnetic (EM) waves. Wireless technologies use wavelengths ranging from the radio frequency (RF) band up to and above the IR band. The frequencies in the RF band cover a significant portion of the EM radiation spectrum, extending from 9 kilohertz (kHz), the lowest allocated wireless communications frequency, to thousands of gigahertz (GHz). As the frequency is increased beyond the RF spectrum, EM energy moves into the IR and then the visible spectrum. Because wireless network and technology are so diverse, we primarily focus on the WLAN and WPAN technologies.

7.1. Wireless LANs

WLANs allow greater flexibility and portability than do traditional wired local area networks (LAN). Unlike a traditional LAN, which requires a wire to connect a user's computer to the network, a WLAN connects computers and other components to the network using an access point device.

An access point communicates with devices equipped with wireless network adaptors; it connects to a wired Ethernet LAN via an RJ-45 port. Access point devices typically have coverage areas of up to 300 feet (100 meters). This coverage area is called a cell or range. Users move freely within the cell with their laptop or other network device. Access point cells can be linked together to allow users even to "roam" within a building or between buildings.

7.2. Ad Hoc Networks

Ad hoc networks such as Bluetooth are networks designed to dynamically connect remote devices such as cell phones, laptops, and PDAs. These networks are termed ad hoc because of their shifting network topologies. Whereas WLANs use a fixed network infrastructure, ad hoc networks maintain random network configurations, relying on a system of mobile routers connected by wireless links to enable devices to communicate. In a Bluetooth network, mobile routers control the changing network topologies of these networks. The routers also control the flow of data between devices that are capable of supporting direct links to each other. As devices move about in an unpredictable fashion, these networks must be reconfigured on the fly to handle the dynamic topology. The routing protocol Bluetooth employs allows the routers to establish and maintain these shifting Networks

The mobile router is commonly integrated in a device such as a PDA (Figure 7.1). This mobile router, when configured, ensures that a remote, mobile device, such as a mobile phone, stays connected to the network. The router maintains the connection and controls the flow of communication. (Figure 7.1 also illustrates how with

emerging technologies the mobile phone will be capable of connecting to the network, synchronizing the PDA address book, and downloading e-mail on an IEEE 802.11 WLAN all at the same time.)

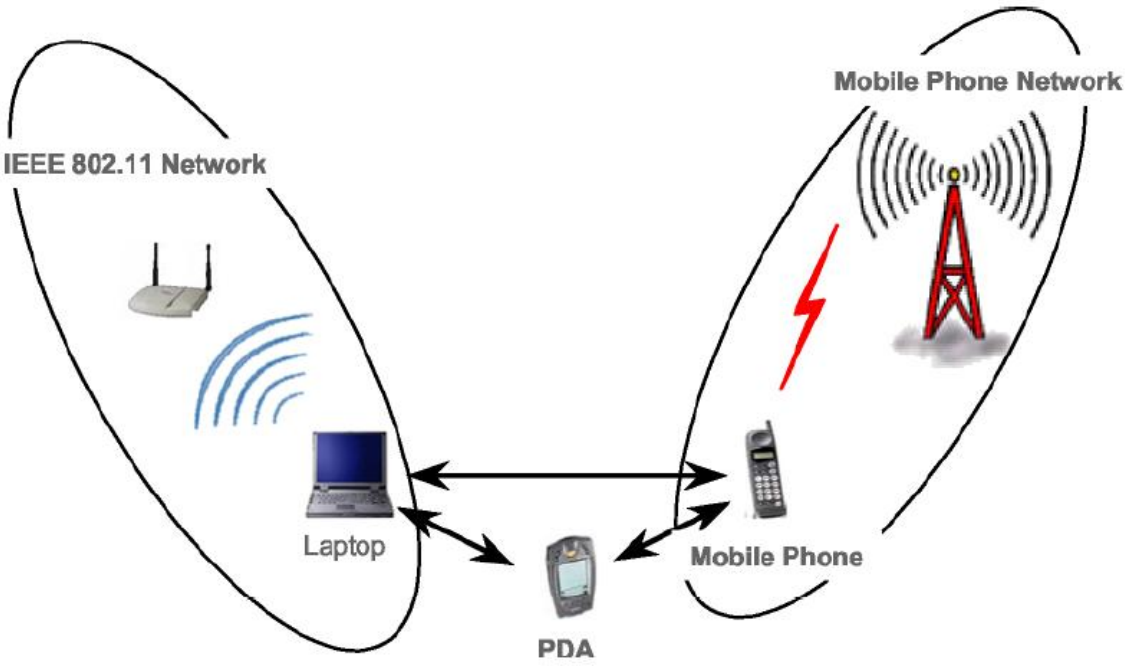

Figure 7.1 Ad Hoc Network

7.3. IEEE 802.11 Standards Overview

IEEE 802.11 is a standard for WLANs designed to provide high-speed data communication between portable devices. It is intended to allow flexible wireless networks to be created within local area without the need for the wired infrastructure and it can be used as an extension of a wired LANs. As any IEEE 802.x standard, for instance 802.3 (Ethernet) and 802.5 (Token Ring) standards for wired LANs, the 802.11 standard defines both the physical layer and the Medium Access Control (MAC) layer. As shown in Figure 7.2, the 802.11 standard together with the IEEE 802.2 standard defines two lowest layers of the well-known seven-layer ISO Open System Interconnection (OSI) networking model – the physical layer and the data link layer. The IEEE 802.2 standard defines the Logical Link Control (LLC) layer,

which is common for the 802.x family of standards. The IEEE 802.11 standard was adopted in 1997. Since then, several extensions to the standard have been developed, and more are emerging. The complete family of the current and emerging 802.11 standards is listed in Table 7.1. This section provides an overview of the original 802.11 standard and its extensions.

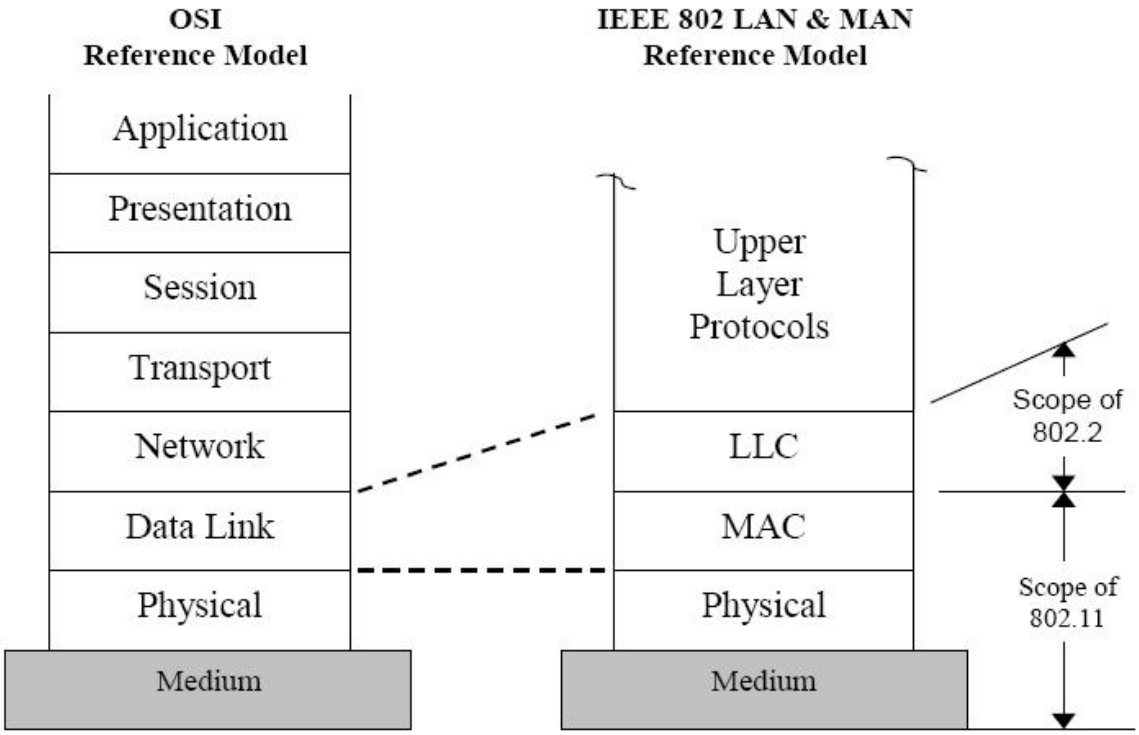

Figure 7.2 IEEE 802.11 standard and OSI reference model

7.3.1. Network architecture

The station is the most basic element of the 802.11 WLANs. A station is any device that contains the functionality of the 802.11 protocol. The basic service set (BSS) is the basic building block of 802.11 WLAN and consists of two or more stations. Figure 7.3 illustrates the concept of the BSS when applied to two types of networks defined in the IEEE 802.11 standard: independent and infrastructure. The ovals used to depict a BSS illustrate the coverage area within which the member stations of the BSS may remain in communication. The Independent BSS, often

referred as an ad-hoc, is stand-alone self-configuring network, providing direct communication between stations. The Infrastructure BSS uses fixed location access points (AP) to provide connectivity to stations.

Table 7.1 Summary of IEEE 802.11 standards

Standard	Description	Status
802.11	Original standard	Completed
802.11a	Physical layer, 54 Mbps, 5 GHz	Completed
802.11b	Physical layer, 11 Mbps, 2.4 GHz	Completed
802.11c	Access Point bridging	Completed
802.11d	Regulatory extensions	Completed
802.11e	Quality of Service	Estimated completion in 2004
802.11f	Inter Access Point roaming	Completed
802.11g	Physical layer, 54 Mbps, 2.4 GHz	Completed
802.11h	Transmit power control, Dynamic frequency selection	Completed
802.11i	Enhanced security	Completed
802.11j	Japanese regulatory extensions	Estimated completion in 2004
802.11k	Radio resource measurement	Ongoing
802.11m	Maintenance	Ongoing
802.11n	Physical layer, high throughput study group 100+ Mbps	Estimated completion in 2006-2007

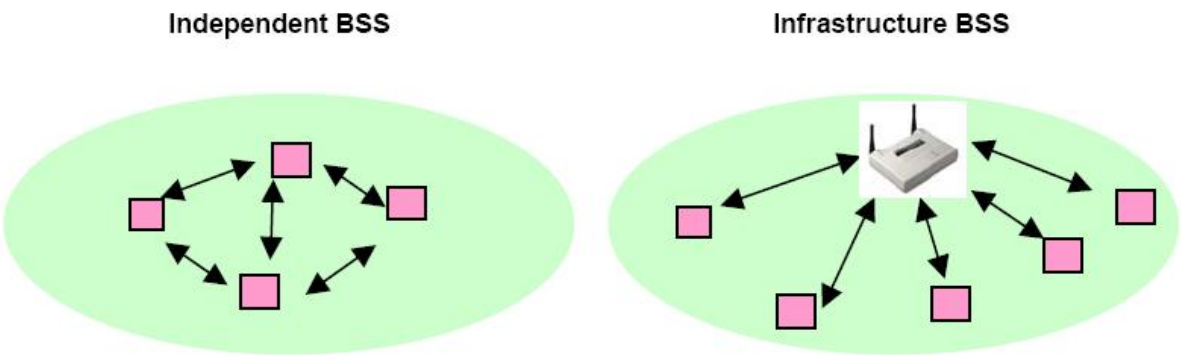

Figure 7.3 Independent BSS and Infrastructure BSS

In order to extend the operational range of a BSS, the 802.11 standard defines an Extended Service Set (ESS), as illustrated in Figure 7.4. An ESS consists of multiple BSS interconnected by distribution system, wired or wireless backbone network. The 802.11 standard does not define the distribution system itself but the distribution services only. ESS can be interconnected with other wired or wireless networks, allowing stations within this ESS access to other networks' resources. Each BSS and ESS has its unique identification called BSSID and ESSID respectively, which are required to implement addressing.

To join an Infrastructure BSS, a station must select an AP and associate with it. The association service creates a mapping between the station and the AP that can be provided to the distribution system. The station can then sends and receives messages via the associated AP. The dissociation service terminates an existing connection. The re-association service allows a station with an established association with one certain AP to move its association to another AP. A station uses the distribution service every time it sends MAC frames across the distribution system. The integration service connects the 802.11 WLAN to other LANs, including one or more wired LANs or 802.11 WLANs. A portal performs the integration service. The portal is an abstract architectural concept that translates 802.11 frames to frames that may traverse another network, and vice versa. The authentication service can be used by station to establish the identity of the other station. Also the privacy service is available, preventing the contents of messages from being read by anyone other than the intended recipient stations.

7.3.2. Physical layer

The 802.11 physical layer defines three basic transmission techniques: Direct Sequence Spread Spectrum (DSSS), Frequency Hopping Spread Spectrum (FHSS), and Diffuse Infrared. FHSS and Diffuse Infrared have received little attention and were not used in extensions of 802.11 standard, hence will be neglected in this thesis.

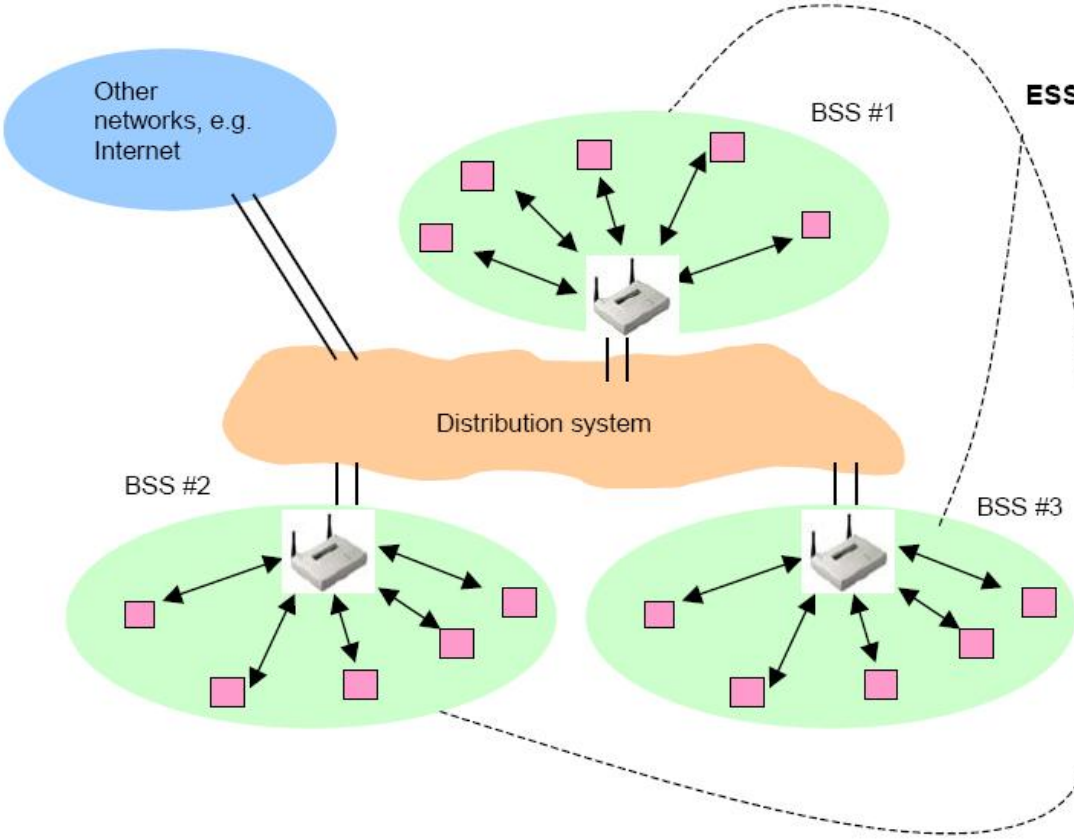

Figure 7.4 Extended service set (ESS)

The DSSS 802.11 system is aimed for globally available unlicensed 2.4 GHz band, known also as the band designated for the ISM (Industrial, Scientist and Medical) applications. By supporting different power levels allowed by different countries' regulations, it became possible to develop a wireless LAN standard that could be used on a global basis, which was the most important reason for choosing this band.

Two physical layer data rates are defined: 2 Mbps and 1 Mbps. The data is modulated using DQPSK and DBPSK for the 2 Mbps and 1 Mbps data rates respectively. The 802.11 system changes data rates to match the radio channel conditions. As a station moves further away from another station or if interference source is present, the highest data rate may not provide reliable transmission of data. To coupe with that, the 802.11 system decreases the data rate, since lower rates are more tolerant to the noise and thus more reliable than higher data rates. The 802.11

standard does not define the criteria to use to decide which data rate to use. The standard only requires that all compliant products must support all specified data rates for compatibility purpose.

To create a DSSS signal, the data symbol is multiplied with the spreading sequence. The following 11-chip Barker code has been chosen as the spreading sequence due to good autocorrelation properties and relatively short length: +1, –1, +1, +1, –1, +1, +1, +1, –1, –1, –1. After spreading operation, the bandwidth of the transmitted signal is increased by a factor of 11. This provides a spreading gain of 10*log10(11)=10.4 dB against narrowband interference signals and makes a DSSS signal appear as background noise to a narrowband receiver. On the receiver side, the received data is correlated with the spreading sequence to obtain the originally sent data. As every user in the network uses the same spreading sequence, no multiple access (as opposite to more complexes CDMA technique) or security is provided in the 802.11 DSSS system by means of the data spreading.

The bandwidth of the transmitted signal is always about 22 MHz regardless of the data rate. Therefore, the 2.4 GHz ISM band with bandwidth of 83.5 MHz can accommodate up to three non-overlapping channels as shown in Figure 7.5. The 802.11 standard defines totally fourteen partly overlapping channels in the 2.4 GHz ISM band.

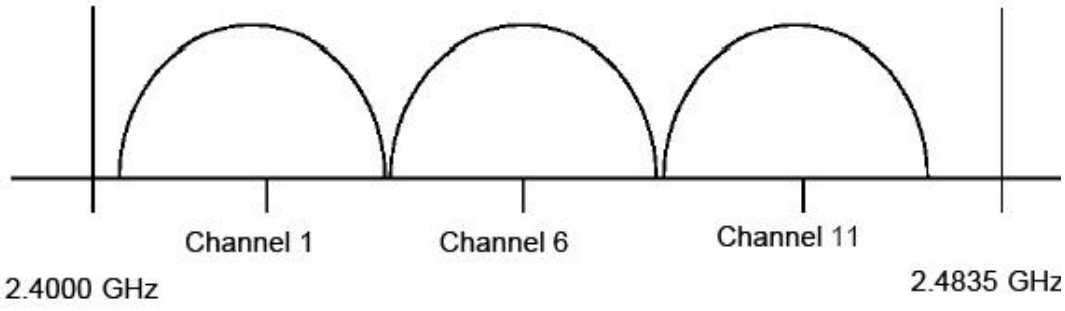

Figure 7.5 Three 802.11 non-overlapping channels in the 2.4 GHz ISM band

Figure 7.6 shows the format of the DSSS physical layer frame. It starts with 128 synchronization bits that the receiver uses to detect the presence of the signal. The 16-bit start delimiter is used for bit synchronization. The signal field indicates the modulation that is to be used for transmission and reception of payload data, 1 Mbps DBPSK or 2 Mbps DQPSK. The 8-bit service field is reserved for future use. The 16-bit length field indicates the number of bytes in the payload data. The CRC field, short for Cyclic Redundancy Check, is used for error detection.

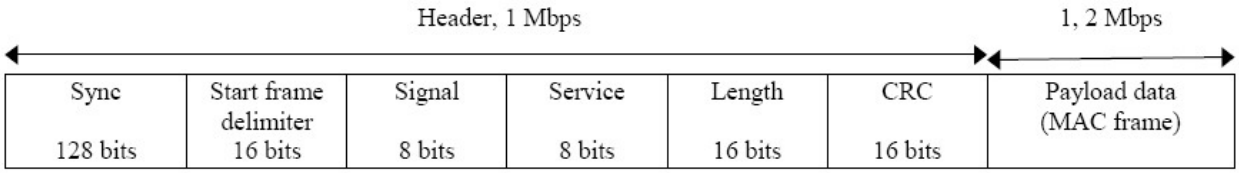

Figure 7.6 802.11 DSSS physical layer frame format

7.3.3. Physical layer extensions

Ratified in 1999, the 802.11b standard adds 5.5 Mbps and 11 Mbps data rates to the original 1 Mbps and 2 Mbps 802.11 modes. Currently, 802.11b is the most popular 802.11 technology.

7.3.3.1. 802.11b

The higher data rates are achieved by using complementary code keying (CCK) DSSS technology. The CCK technology codes more data bits per 11 spread bits, 4 bits and 8 bits for 5.5 Mbps and 11 Mbps respectively, than 1 or 2 bits in the plain 802.11 standard while keeping the same bandwidth of the transmitted signal. It does this by first using 8 bit spreading sequence instead of the original 11-bit sequence. However, this 8-bit sequence still runs at a rate of 11 Mbps, which result in the same spreading factor of 11. Thus, the clock rate for data is increased from 1 Mbps to

1.375 Mbps (8*1.375=11). The CCK encoding does not use a static spreading sequence; six of the 8 bits are used to choose 1 of 64 complementary spreading codes. Different spreading codes are chosen based on the incoming data. The same DQPSK is used to modulate spreaded data. Figure 7.7 shows the format of the physical layer frame. The frame header still runs at 1 Mbps while payload data can run at four different rates depending on the channel conditions.

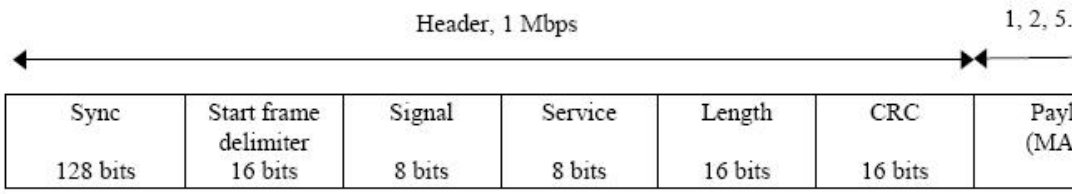

Figure 7.7 802.11b physical layer frame format

7.3.3.2. 802.11a

The 802.11a standard, introduced at the same time as 802.11b, is intended for the 5 GHz license-free UNII band and provides data rates up to 54 Mbps. The 5 GHz band has an advantage of large bandwidth allocated for the unlicensed operations. There are 455 MHz available (5.15 – 5.35 MHz and 5.470 – 5.725 MHz) for use by WLAN systems in Europe. This allows 19 non-overlapping channels in the 5 GHz band versus 3 non-overlapping channels in the 2.4 GHz band.

The 802.11a is based on Orthogonal Frequency Division Multiplexing (OFDM) modulation, which allows achieving higher data rates within about the same channel bandwidth as 802.11b. OFDM is a multi carrier transmission technique. The OFDM signal consists of multiple sub carriers, each one being modulated by a low rate data stream. Low rate data streams are formed by demultiplexing one high data rate stream. Sub carriers are kept orthogonal, so data symbols modulated on these sub carriers can be recovered without mutual interference. Since the symbol rate on each sub carrier is slower than the original data rate, the OFDM technique is particularly efficient in time dispersive environments.

The 802.11a OFDM signal consists of 52 carriers. Data is sent on 48 carriers simultaneously, with 4 carriers used as pilots to aid in channel estimation at the receiver. Forward error correction coding (convolution coding) is used to provide error detection and correction. Table 2 shows supported data rates. Various data rates are provided by changing the redundancy in the error correction coding and by changing modulation scheme.

7.3.3.3. 802.11g

Adopted in 2003, the 802.11g extension enables 54 Mbps data rates, the same data rate as provided by the 802.11a standard, but now in the 2.4 GHz band. This is achieved by using the same data rates and modulation formats as used in the 802.11a standard. Additionally, the 802.11g standard is backward compatible with the 802.11b standard, i.e. the 802.11b modulation formats and data rates are supported.

Table 7.2 802.11a and 802.11g data rates and rate-dependent parameters

Data rate, Mbps	Modulation	Coding rate	Coded bits per subcarrier	Coded bits per OFDM symbol	Data bits per OFDM symbol
6	BPSK	1/2	1	48	24
9	BPSK	3/4	1	48	36
12	QPSK	1/2	2	96	48
18	QPSK	3/4	2	96	72
24	16-QAM	1/2	4	192	96
36	16-QAM	3/4	4	192	144
48	64-QAM	2/3	6	288	192
54	64-QAM	3/4	6	288	216

7.3.3.4. 802.11n

The 802.11n is the next generation extension of the physical layer. It is expected that 802.11n will support throughput (useful data rates) of over 100 Mbps. The standard is still in the earlier development phase. Among the proposed approaches to provide such high data rates are smart antenna technology, enhanced modulation, and increased channel bandwidth (using both 2.4 and 5GHz bands).

7.3.4. MAC layer and MAC layer extensions

The main functions of the 802.11 MAC are following:

- Beaconing and frame exchange at the MAC layer to deliver data
- Frame formatting
- Multiple accesses to the shared wireless medium
- Power management
- Quality of service (QoS)
- Security

7.3.4.1. Beaconing

A beacon frame is sent periodically to synchronize the stations in the BSS and to inform the stations of impending data. In an independent BSS, the synchronization mechanism is distributed among the stations in the BSS. In an infrastructure BSS, the AP is responsible for transmitting the bacon frames regularly.

7.3.4.2. Frame exchange

About thirty types of frames are defined for the MAC to provide management and information data exchange between the stations. All stations are required to decode and react to the information in the MAC header of every frame they receive. Since wireless medium is not as reliable as wired, the basic frame exchange consists of two

frames: the frame sent and the frame acknowledgement. If the source does not receive the acknowledgement, the source attempts to retransmit the frame.

7.3.4.3. Frame format

The basic MAC frame is shown in Figure 7.8. Some of the MAC packets do not include all of the fields. Up to four addresses can be used depending on the frame type. For example, if two stations communicating with each other are associated with different APs, then the MAC addresses of both APs and both stations will be present in the four address fields. Addresses are 48-bit IEEE 802 MAC address (common address space is shared between 802.11 WLAN, 802.3 Ethernet and other 802.x LAN standards). Each station has its own unique MAC address.

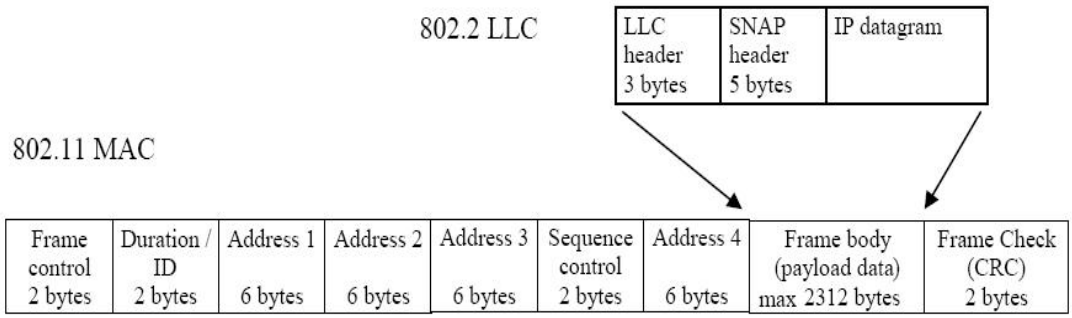

Figure 7.8 802.11 MAC frame format

7.3.4.4. Multiple Accesses

Several multiple access mechanisms are defined in the standard to determine when a station in a BSS is allowed to transmit and when it may be able to receive data packets over the shared wireless medium. The basic access method of 802.11 is Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA), which is defined as a part of the Distributed Coordination Function (DCF) in the standard. DCF provides support for best-effort asynchronous data transfer. CSMA/CA is a "listen before talk" access mechanism. It relies on the physical carrier sense form the

Physical Layers and the virtual carrier sense implemented in a special field of every frame to determine the state of the medium.

The CSMA/CA protocol avoids collisions among stations sharing the same medium by utilizing a random back off time. The period of time immediately following a busy medium is the highest probability of collisions occurring, especially under high utilization. The CSMA/CA scheme implements a minimum time gap between frames from a given user. Once a frame has been sent from a given transmitting station, that station must wait until the time gap is up to try to transmit again. Once the time has passed, the station selects a random amount of time (the back off interval) to wait before "listening" again to verify a clear channel on which to transmit. If the channel is still busy, another back off interval is selected that is less than the first. This process is repeated until the waiting time approaches zero and the station is allowed to transmit.

An optional RTS/CTS handshake procedure is used to operate with CSMA/CA to handle a problem referred to as a “hidden terminal” problem. This problem occurs when a receiving station is in range of two transmitting stations, which are not in range of one another. In this case attempting to detect if the medium is free does not necessarily work because two transmitting stations, which are not in range of one another, can not detect one another’s transmissions. Thus, the packets from two transmitting stations will collide at the receiving station. In RTS/CTS technique instead of transmitting a data packet after waiting for a free medium, a station transmits a short ready-to-send (RTS) packet to request the use of the medium. If this succeeds, the receiver will quickly reply with a short clear-to-send (CTS) packet. After the successful exchange of an RTS/CTS pair the actual transmission takes place. This method allows hidden terminals to hear either CTS or RTS packets. It also means that if packets do collide only a short RTS or CTS packet is lost, which is preferable than to have collisions of long data packets. For example, RTS is 20 bytes and CTS is 14 bytes, whereas data packets can be up to 2300 bytes long. If this optional function is available at a station it is enabled in one of three modes: always on, always off, or on for packet sizes above a certain threshold.

The 802.11 standard defines one more optional media access protocol, called as Point Coordination Function (PCF). The PCF uses a polling procedure to provide connection-oriented contention-free service. This function is performed by an AP, which polls stations within the BSS and allows them to transmit. In this way, delay-sensitive packets such as voice or video can be given priority over other data.

7.3.4.5. Power management

Power management is at great importance in battery-powered devices such as mobile phones. The 802.11 standard specifies an optional power management function. This function allows stations in a BSS to enter a low power mode of operation while remain associated within the BSS. Because of difference in operations between an Independent BSS and Infrastructure BSS, two mechanisms for power management have been developed.

In an Individual BSS, power management is controlled by the mobile stations. It is implemented through a beacon frame. The station in a low power mode is required to wake up to receive every beacon frame and stay awake for a certain period of time after each beacon for data reception. Sending stations buffer the frames to be sent to the destination station in the low power mode until the destination station awakens. No stations that send a bacon frame are allowed to enter the power save mode until they receive a bacon frame from another station within the BSS. This restriction ensures that there is at least one station in the BSS that is in active mode and is able to process requested frames.

In an Infrastructure BSS, the power management mechanism is centralized in the AP. It is implemented through frame exchange and specific information transferred in the beacon frame. Transmitted by an AP, the beacon frame indicates whether a station has frames to receive. Stations are not required to wake up for each beacon frame. A station is required to inform an AP when the station enters a low power mode and the number of beacon frames the station will remain in a low power mode.

7.3.4.6. Transmit power control and dynamic frequency selection, 802.11h

Recently adopted, the 802.11h extension allows WLANs to meet regulations initially adopted by European countries for operation in the 5GHz band (later it was adopted as a global requirement). The regulations call for WLANs to detect the presence of radars, satellites and space research systems and then protect them from interference by selecting another operating channel or reducing transmit power. The 802.11h provides a standard method to avoid interference by introducing two techniques: transmit power control and dynamic frequency selection.

The transmit power control is a coordination mechanism in which stations located in close proximity to access points would decrease their transmit power, and stations located further away from access points would increase transmit power. The mechanism allows to reduce the average transmit power across a WLAN system, resulting in decreased interference level to other WLANs.

The dynamic frequency selection mechanism is intended to avoid interference between WLANs or between a WLAN system and other radio systems, such as radars. It does this by detecting the presence of other systems and switching to the channel with the lowest interference level.

7.3.4.7. QoS, 802.11e

The 802.11e task group is currently working on two main QoS issues: to improve the efficiency of the MAC protocol and differentiation between different types of data traffic. Efficiency is an important issue in 802.11. For instance, a current 11 Mbps 802.11b device in the best case (two communicating devices in close proximity to each other, no interference) can provide to the network layer the throughput, the actual data rates of about 5 Mbps. The reason for this is overheads in the MAC protocol. Differentiation will enable enhanced multimedia and voice capabilities by giving higher priority for time-sensitive data packets (video and audio

streaming, VoIP) over general data packets whose delivery time is less critical (e-mail, http, ftp).

7.3.4.8. Security, 802.11i

The recently ratified 802.11i security extension was probably the most awaited, especially by enterprises, where security issues are at premium. Wireless technologies release users from having to be physically attached to the network, but using radio waves as a transmission medium makes wireless networks susceptible to interceptions and attacks. The traffic can be captured at any location as long as the signal reaches the receiver. The typical range of 802.11b/g APs is about 100 meters (indoor environment), which hackers can extend well beyond 500 meters by using directional antennas. This enables so call “war driving” or “parking lot attack”, where hackers can perform traffic analysis and attacks in a car, by driving around or just by parking nearby.

802.11i is intended to fix the problems that are well known in the original 802.11 security protocol called Wired Equivalent Privacy (WEP) and address all known attack. WEP is a single static symmetric shared key system in which 40-bit (also called 64-bit) or 108-bit (also called 128-bit) encryption is applied to packet transmissions. WEP was intended to protect wireless communication from eavesdropping and prevent unauthorized access to a WLAN, so to make WLAN communication as secure as wired LAN data transmission would be. However, several security flaws have been found in the technique. For example, there are freely available tools to crack WEP keys, including AirSnort and Crack. These applications perform statistical analyses on encrypted packets to eventually determine the secret shared key. The current implementations require about 500 Mbytes of data before the secret key can be successfully derived.

The 802.11i standard can be viewed as consisting of three main parts. Two of the parts are enhanced encryption algorithms in form of Temporal Key Integrity Protocol (TKIP) and Advanced Encryption Standard (AES). Both of those standards were

specifically designed to fix the known flaws in WEP, with TKIP being targeted at legacy equipment and AES as long-term replacement of TKIP. Unlike TKIP, the encryption method based on AES was not designed for backward compatibility. AES is considered state of the art in encryption technology. It is stronger than TKIP and scales better to higher data speeds, however requires significantly more computational power.

The third part is 802.1x based authentication replacing the non-working 802.11 native WEP authentication. IEEE 802.1x is a standard for network access control at Data Link layer for both wired and wireless networks. It provides a framework for user authentication and encryption key distribution, so it can be used to restrict network access until the user has been authenticated by the network. It is used in conjunction with one of the upper layer authentication protocols, such as Extensible Authentication Protocol with Transport Layer Security (EAP-TLS), supported natively in Windows XP to perform verification of users and generation of encryption keys. Unlike WEP, 802.1x encryption keys are unique for each session between an individual client and AP.

7.3.5. Other miscellaneous 802.11 standards

802.11c (Access Point Bridging). The 802.11c provides required information to ensure proper bridge operation, for example to bridge Ethernet and 802.11 WLAN.

802.11d (Regulatory Extensions). The usage of spectrum differs from country to country. The 802.11d extension allows configuring the 802.11 products to be conformant with worldwide regulations. Also the standard specifies a means for the access point, which is configured to operate in a particular country, to broadcast information about what the local regulatory environment is, specifically which channels are legal in this regulatory domain and what transmit power level is permitted.

802.11f (Inter Access Point Roaming). This technology handles the registration of APs within a network and the exchange of information when a user is roaming among coverage areas supported by different vendors' APs. It helps with fast hand-off from AP to AP.

802.11j (Japanese Regulations). The 802.11j extension will allow operations in new bands in Japan.

802.11k (Radio resource measurement). The 802.11k is a manageability extension. It will allow APs to request the client stations about the state of the medium around them and report the state to some central point. Thus, the information about the overall state of the network will be provided at the central point.

802.11m (Maintenance). The 802.11m is not really a standard as such. It is a task group whose job is to provide interpretations of the standard.

7.3.6. 802.11 in a Mobile Phone

The purpose of this section is to discuss the need for 802.11 in a mobile phone. This section provides comparison of the 802.11 technology with other wireless technologies, such as widely adopted in mobile phones but also new emerging, and describes possible usage scenarios for 802.11 in a mobile phone.

7.3.7. Wireless technologies in a mobile phone

Many wireless technologies destined for mobile phones have been developed, and more standards are emerging. Table 7.3 provides comparison of the most popular technologies (cellular systems are limited to implement in Europe). The technologies listed in Table 7.3 are divided into groups according to distances they can cover:

- Wireless personal area network (WPAN) – covers personal operating area
- Wireless local area network (WLAN) – covers buildings or campuses
- Wireless metropolitan area network (WMAN) - covers city or county
- Wireless wide area network (WWAN) - provides national or global coverage
- Satellite network - provides true global coverage

Figure 7.9 also shows a comparison of those technologies in terms of bit rate versus range. If benefits of each technology are examined, it can be seen that each technology has its strengths and weaknesses, making it more suitable for specific type of application scenarios.

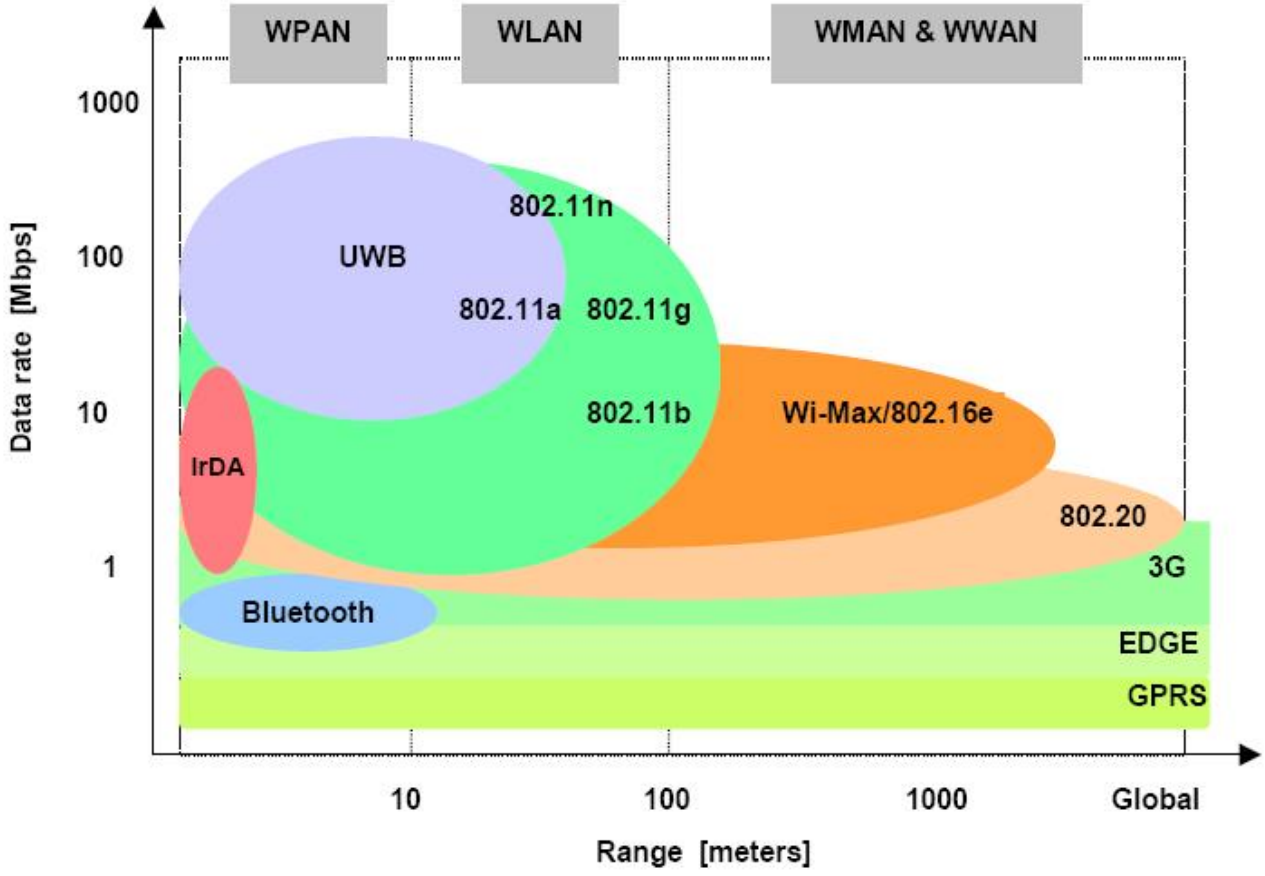

Figure 7.9 Comparison of wireless technologies in terms of bit rate versus range

7.3.8. 802.11 and the Current Mobile Phone Technologies

Compared to the cellular technologies, 802.11 has an advantage of much higher data rates at lower costs. Against this, 802.11 is a short range technology and provides low mobility (the specification stops at the MAC layer). Hence, the coverage of the 802.11 WLANs is mainly limited to certain areas such as homes, offices, universities, and public places (airports, hotels, coffee houses, conventional centers, city centers, etc.) and will never provide the ubiquitous coverage of cellular systems due to practical reasons. The low service cost of 802.11 WLANs is mainly because of unlicensed radio spectrum being used. However, there is a significant drawback in using unlicensed spectrum – much lower signal quality compared to licensed spectrum. Thus unlike cellular systems, one WLAN system can undergo interference from other WLANs, wireless systems or devices, and regulatory bodies will not help to solve such kind of situations in unlicensed radio spectrum unless interfering systems and devices exceed regulatory requirements.

Compared to Bluetooth and IrDA, 802.11 has the advantage of longer range and in most cases higher data rates, but cannot compete with those technologies in terms of power consumption and price. Thus, Bluetooth and IrDA domination as a cable replacement technology will not be affected by 802.11 on the market of accessories and low power peripherals, such as wireless headsets.

Table 7.3 Comparison of wireless technologies

Type of network	Technology	Peak bit rate (Mbps)	Coverage/ range (meters)	Freq. (GHz)	Associated relative cost	Status
Satellite	GPS	Location service	Global	~1,5	High	Widely deployed
WWAN (Cellular)	GSM/GPRS	0,115	Global	~0,9; ~1,8; ~1,9	High	Widely deployed
	EDGE	0,384		~0,9; ~1,8; ~1,9	High	Under deployment. Expected to be widely deployed in 2004-2005
	WCDMA, HSDPA	2 uplink/ downlink; 10 downlink		~2;	High	Under deployment. Expected to be widely deployed in 2005-2006
WMAN/ WWAN	802.20	>1	15000	-	-	Standard is under development.
WMAN	WiMAX / 802.16e	15	5000	2 - 6	-	Estimated to be completed in 2005
WLAN	802.11b	11	100	~2,4	Medium	Widely deployed
	802.11a	54	50	~ 5	Medium	Under deployment.
	802.11g	54	100	~2,4	Medium	Under deployment.
	802.11n	>100	-	-	-	Standard is under development. Estimated completion in 2006-2007
WLAN/ WPAN	UWB / 802.15.3a	>100	10 - 20	3,1 – 10,6	Low	Standard is under development. Estimated completion in 2004
WPAN	Bluetooth 1.2	0,720	10	~2,4	Low	Widely deployed
	IrDA 1.4	16	2	Optical, 850 nm	Very low	Widely deployed

7.3.9. 802.11 Applications and Usage Scenarios

The following list illustrates the range of applications that can be covered by using 802.11 in a mobile phone:

- Voice over IP (VoIP)
- Web browsing
- E-mail
- Corporate intranet access
- Messaging (instant messaging, SMS, MMS, etc)
- Push-to-talk
- File up/downloading
- Audio and video streaming
- PC synchronization and backup (diary, address book, files, etc)
- Multiplayer gaming
- Positioning

The possible usage scenarios in a mobile phone come from the advantages of 802.11. Three most perspective scenarios from the author point of view are described below. In the first scenario, the 802.11 mode can be used for fast and low cost Internet connectivity and low cost VoIP calls wherever the mobile phone is within coverage of 802.11 WLANs, for example at homes, enterprises, and public places. Thus, the 802.11 phone will be as a low cost replacement for traditional wired, cordless and DECT phones at homes and enterprises, and also will provide data services and will allow having low-cost connection at public places. The cellular mode, providing lower data rates at higher cost but having advantage of the ubiquitous coverage of cellular systems, will be employed as soon as the mobile phone moves out of the 802.11 WLANs coverage. In the simplest case, 802.11 WLANs and cellular networks will be completely separated, thus users will have to select manually the network they prefer. In more complicated cases, there will be internetworking between 802.11 WLAN and cellular networks, providing an increased level of service for users. 3GPP, an organization developing technical

specifications for a 3rd Generation Mobile System based on evolved GSM, has defined six 3GPP-WLAN internetworking scenarios with increased technical complexity :

• Common billing and customer care. This is the simplest internetworking scheme where both networks remain completely separate. However, the customer will receive one bill for use of both networks and will have a single customer care relationship.

• 3GPP system based access control and charging. This is the scenario where authentication, authorization and accounting are provided by the 3GPP system. The user data traffic will remains completely separate in both networks.

• Access to 3GPP system packet-switched based services. The goal of this scenario is to provide the WLAN user with at least some of the packet-switched services provided by 3GPP.

• Service continuity. This scenario is about providing handover between the systems for packet-switched services, but the handover does not have to be seamless (changes in quality, delays and gaps are allowed)

• Seamless services. The goal of this scenario is to provide seamless handover between the systems for packet-switched services. No noticeable interruption in the service is allowed.

• Access to 3GPP circuit-switched services. This scenario will provide WLAN users with seamless access to 3GPP circuit-switching services.

Additionally, many vendors are developing system solutions, which will allow seamless handover between VoIP over 802.11 and GSM circuit switched voice. One of the driving forces is that mobile operators have become involved into this issue. This move of mobile operators would seem strange as one can think that deployment

of 802.11 WLANs will cut operators' revenues. However, one rationale for this is that operators will have possibility to use WLAN where it is expensive for them to use traditional cellular networks to provide services because of the expense of building additional infrastructure or buying additional licensed spectrum.

The second possible scenario is a wireless access through 802.11 to the following local services at public WLANs in airports, railway stations, trains, ferryboats, etc.:

• Information services. For example, in the case of railway stations and airports this can be schedules or information about restaurants located nearby (menu, prices, open hours)

• Positioning services. For example, an interactive map can be provided with the current location of the person, locations of the restaurants and the best paths to get to those restaurants.

• Tickets sale, check-in

• Entertainment services (music, video, games)

The third possible scenario is cable replacement. As mobile phones' memory size and size of external memory cards is constantly increasing, 802.11 can be used for faster wireless data transfer between a mobile phone and a laptop or any other 802.11 devices. For example, to fully upload with data the 1 Gbytes SDIO card, recently introduced by SanDisk, Bluetooth 1.2 will require about 3.5 hours, IrDA 1.4 about 12 minutes, and 802.11a or 802.11g about 4 minutes. However, it should be noted that introduction of new high-speed cable replacement technologies, such as UWB and based on this technology Wireless USB standard (recently announced by Intel), will most likely limit 802.11 usage in this scenario. With target data rate of 480 Mbps, Wireless USB will require about 20 seconds to upload with data a 1 Gbytes memory card.

Generally, 802.11 has application to a broad range of types of mobile phones, from a fairly basic phones that offer basic voice and data services through middle-class phones that have such features as cameras, MP3 players, MMS to smart phones. In basic phones 802.11 can be used for VoIP calls and simple data applications while in higher-class phones it can be used for all applications described earlier.

7.4. Bluetooth

Bluetooth is the specification used as a blueprint for IEEE's 802.15 wireless personal area network (WPAN) initiative. It is typically used for providing device yo device connectivity on an adhoc basis, whereas WLAN systems target as a wireless replacement or extension of the LAN infrastructure. Bluetooth is meant to be more than just a radio channel. It is proposed to be an intelligent and robust method for allowing devices to seek and provide one another services in ways that streamline mobile computing and enable more responsive behaviour from wire-line networks. Bluetooth operates in a band of radio frequencies that is just above 2.4 GHz, like IEEE 802.11b, and can thus cause interference. This will be explained shortly.

Bluetooth was specifically designed to accommodate both synchronous (such as voice) communications and asynchronous (data) communications. This technology is meant more as a wire replacement than a LAN topology; thereby, it is likely that Bluetooth will coexist with other standards that are more LAN oriented. The Bluetooth approach aims to dramatically reduce the complexity of the protocol and reduce the transmitter power, and consequently, the coverage range to lower cost and simplify operation.

To accomplish these goals, Bluetooth uses an arrangement of very small "piconets" that can support only eight nodes at a time. All network connectivity is adhoc, which means there is no network established until a device chooses to communicate. When communications are established, devices within the piconet determine a master node, which synchronizes timing and controls communications.

Communication rates between Bluetooth devices can reach 1Mbps at a radius of 10 meters, but this is highly dependent on how Bluetooth is implemented on those devices. For example, some vendors are pushing to expand the range of Bluetooth connections to 100 meters.

Bluetooth uses spread spectrum, in which multiple users share a single spectrum slice but use sophisticated information processing to identify their own signals while ignoring others, like 802.11. Specifically, Bluetooth uses frequency hopping, wherein senders and receivers follow pre-planned sequences of moves between narrow channels within an agreed upon range. This rapid movement (1600 hops per second) is essentially to avoid collisions with other packets.

7.4.1. Bluetooth Security

In any wireless implementation, security is paramount. Like 802.11x, Bluetooth addresses the area of security. Devices connecting via Bluetooth enjoy automatically negotiated link-level security, with key sizes up to 128 bits. However, Bluetooth's protocols only establish the identity of a device, not its user. Security negotiations take place only when a connection first established, not on subsequent connection exchanges. This means that Bluetooth alone cannot enforce one way transfers of data. Therefore, any applications that run on top of Bluetooth connectivity must implement user authentication and database or service access control to enhance overall security. This can be considered as a trade off between security and convenience from the user's point of view.

7.4.2. Interference with 802.11b

Since applications for both IEEE 802.11b and Bluetooth are targeted for similar users and environments, it is likely that both radios will come in close proximity to each other. Moreover, both technologies operate at the 2.4 GHz, along with the nascent Home RF standard and some microwave ovens and cordless telephones, making it possible for the wireless radios to be adversely affected from these devices.

Studies have been conducted to determine the degree of harmful, mutual interference caused by the radios. The degree in which an 802.11 device is susceptible to interference from nearby Bluetooth transmitter is clearly dependent upon the strength of the desired DSSS signal from the access point, which in turn is dependent on the range. According to the cited study, 802.11b WLANs show graceful degradation and acceptable reliability in presence of significant levels of Bluetooth interference.

CHAPTER EIGHT
WiMAX (802.16)

This chapter will explore the IEEE 802.16 standards, the capabilities they enable and their advantages over current wireless networking technologies. We will begin with a general discussion of the standard, followed by a brief comparison of the IEEE 802.16 and IEEE 802.11 standards. Subsequent sections will discuss network architectures, and features of the standard.

8.1. What is IEEE 802.16

The IEEE 802.16 is a standard, designed by the IEEE, for local and metropolitan area network (MAN) fixed broadband wireless access. The IEEE 802.16 standard itself is titled "Air Interface for Fixed Broadband Wireless Access Systems" and was approved by the IEEE on 6 December 2001. The standard applies to frequencies between 10 and 66 GHz, while the IEEE 802.16a standard covers frequencies between 2-11 GHz. However, the MAC portion of the standard is entirely frequency independent, and thus leaves open the possibility of future adaptations of the standard.

Systems designed using the IEEE 802.16 standard will be capable of performance comparable to cable, DSL or T1 systems, with shared data rates up to 120 Mbps for LOS transmission in the 10-66 GHz frequency range and 70 Mbps NLOS in the 2-11

GHz frequency range. These systems will be able to provide simultaneous support to "more than 60 businesses at T1 level and hundreds of homes with DSL rate connectivity at 20 MHz bandwidth". In addition to these capabilities, IEEE 802.16 systems will be capable of providing:

- Long range operation: radius up to 30 miles
- Non Line of Sight (NLOS) performance
- Ability to operate in high multi-path environment
- Guaranteed service levels
- Superior scalability
- QoS capable of supporting voice and video applications
- High Spectral efficiency
- Routable networks within an IEEE 802 framework
- Ability to support multicast traffic

The primary advantages of IEEE 802.16 systems over wired systems include: cost savings, quick setup and more complete coverage. While IEEE 802.16 systems are not inexpensive, the costs are still much less than those associated with wired systems. Cost savings are achieved by eliminating the need for wired infrastructure investment and monthly leasing expenses. Installing an IEEE 802.16 system and establishing service requires relatively little time when compared to the three months it might take to establish T1 service in some areas. While DSL services may not be available in areas that are too far from the local telephone company switch, and similar services are often not available in areas of low subscriber density, IEEE 802.16 service can easily and cost effectively reach these areas.

Typical applications for IEEE 802.16 in the commercial sector may include cellular backhaul, broadband on demand and best connected wireless service. IEEE 802.16 is particularly well suited for providing these services. In a cellular backhaul role, IEEE 802.16's robust bandwidth management makes it a reliable alternative to leased wire. This technology is particularly well suited for businesses that relocate frequently within a metropolitan area, such as construction companies, and trade

shows. These companies are able to provision wireless broadband service quickly as they move from one location to another without the need to re-wire. Similarly, the development of hand off procedures between IEEE 802.16 networks will allow a user to roam from network to network, connecting to the best available service in each area.

8.1.1. Comparison of IEEE 802.11 and IEEE 802.16

In recent years IEEE 802.11 has experienced a widespread adoption in residential, corporate and even military settings. The IEEE 802.11 has been used primarily in a data access role, through the creation of "hotspots", a small area where network users can roam unencumbered by wires. Additionally, IEEE 802.11 has been used to provide extension of existing networks into areas where cabling might be impractical or cost prohibitive, building to building connectivity, last mile data delivery, and connectivity for small office /home office (SOHO) networks and mobile offices.

For reasons that will be outlined below, IEEE 802.11 is not well suited for "backbone" or core data distribution roles within a network, or as a public access medium. Among IEEE 802.11's primary limitations are its relatively short range, poor scalability, and security vulnerabilities. Table 8.1 shows a comparison of IEEE 802.11 and IEEE 802.16 standards.

As shown in Table 8.1, the signal from the typical IEEE 802.11 access point (AP) propagates only about 200 yards. This limits the mobility of users and requires the use of many access points for large coverage areas. In addition to IEEE 802.11's inherent range limitations, this standard is very vulnerable to the effects of multi-path, and zone blocking. These vulnerabilities limit IEEE 802.11x's ability to operate in environments with many vertical obstructions and to support NLOS communications.

IEEE 802.11's use of the Carrier Sense Multiple Access with Collision Avoidance (CSMA-CA) access control protocol lies at the heart of its poor scalability. In this

protocol, APs "sense" whether there is any traffic on the wire prior to transmitting, and transmit only when the medium is clear. Unfortunately, due to signal propagation delays and hidden node problems, the possibility of collisions always exists, and this probability increases dramatically as more users are added to the network. As users increase, collisions increase, eventually creating a situation where retransmissions and collisions begin to severely limit throughput.

Table 8.1 Comparison of 802.11 vs. 802.16

Feature	802.11	802.11b	802.11a	802.11g	802.16	802.16a
Assigned Spectrum	2.4 GHz	2.4 GHz	5.8 GHz	2.4 GHz	10-66 GHz	2-11 GHz
Access Control	CSMA- CA	CSMA-CA	CSMA-CA	CSMA-CA	TDMA / DAMA	TDMA / DAMA
Maximum Throughput	2 Mbps / user	11 Mbps / user	54 Mbps / user	54 Mbps / user	124 Mbps / channel	70 Mbps / channel
Propagation Distance	200 yards	200 yards	200 yards	200 yards	> 1 mile	Several miles
Network Architectures	PMP	PMP	PMP	PMP	PTP, PTCM	PMP, PTCM, Mesh
Modulation	Frequency hopping-direct sequence	Frequency hopping - direct sequence	OFDM	OFDM	QUAM, PSK	OFDM
Adaptive Modulation?	No	No	No	No	Yes	Yes
Full Mobility?	No	No	No	No	No	Upcoming
QOS?	No	No	No	No	Yes	Yes

The IEEE 802.11 standard has been plagued by security vulnerabilities associated with the Wireless Equivalent Privacy (WEP) encryption protocol. For reasons that are beyond the scope of this thesis, the WEP protocol is particularly vulnerable to encryption cracking. WEP has been considered so vulnerable, that the IEEE has developed a replacement, the WIFI Protected Access (WPA) protocol, which will be available in equipment following the 802.11i protocol.

The IEEE 802.11 standards enjoy two advantages: price and prevalence. Currently IEEE 802.11 network interface cards available can be purchased for about $60, and APs can be had for less than $100. The second advantage IEEE 802.11 networks currently enjoy is that they are more prevalent than ever before. Today it is not uncommon to find hotspots in airports, bookstores, coffee shops, etc. This prevalence results in more users of this protocol, which in turn produces a more widespread acceptance of the technology by vendors and providers.

In contrast to the IEEE 802.11 standards, equipment based on the IEEE 802.16 standards boasts longer ranges, more robust signals capable of NLOS communication, the ability to handle many users while supporting high QoS and guaranteed service levels, and superior security. In addition to these advantages, future versions of the standard will support full mobility and mesh networking capabilities. It is also important to note that the price of IEEE 802.16 equipment is expected to drop once it becomes more commonly available. IEEE 802.16's capabilities and standards will be covered in more detail in the next section.

The IEEE 802.16 standard is the ideal standard for a public access medium. Its ability to support thousands of users simultaneously is primarily due to its use of time division multiple accesses (TDMA) with demand assigned multiple accesses (DAMA) scheduling for MAC procedures. The specifics of IEEE 802.16's MAC protocol will be examined in greater detail in later sections.

8.1.2. WiMax and Interoperability

The Worldwide Interoperability for Microwave Access (WiMax) forum is an organization of equipment and component suppliers dedicated to promoting the adoption of IEEE 802.16 compliant equipment. This organization tests and certifies products for interoperability and standards compliance. Additionally, the WiMax forum creates what it calls system profiles, which are specific implementations, selections of options within the standard, to suit particular ensembles of service offerings and subscriber populations. The goal of these system profiles is to increase

the adoption rate of IEEE 802.16 equipment by simplifying the setup of this equipment. Prominent members of WiMax include Intel Corporation, Fujitsu, Motorola, AT&T, and many others.

8.2. The IEEE 802.16 Standards

The creation of the Wireless MAN standard is important because it results in a research and development costs savings to equipment manufacturers, which in turn insures interoperability of the equipment they produce and ultimately leads to a reduced risk on the part of equipment operators. The fact that the standard has been developed within the IEEE 802.x framework means that it is possible to both bridge and route traffic to other IEEE 802.x networks (e.g., .11, .3, etc.). In addition to these benefits, a standard provides minimum performance criteria for equipment manufacturers to meet.

The following is a list of the IEEE 802.16 family of standards along with a brief summary and the current status of each. It is important to note that in July 2004, the IEEE approved the draft standard known as IEEE 802.16 - 2004 which combines the IEEE 802.16, IEEE 802.16a, and the IEEE 802.16.c standards into one document.

• IEEE 802.16- The "Air Interface for Fixed Broadband Wireless Access Systems" was approved on December 2001. Designed for Wireless MANs operating in the 10-66 GHz frequency range.

• IEEE 802.16.2- Addresses recommended practices for the operation of multiple fixed broadband wireless systems. Published in 2001, this standard applies to the 10-66 GHz frequency range.

• IEEE 802.16a- This extension to the IEEE 802.16 standard addresses the operation of systems in the 2-11 GHz frequency range, for both licensed and unlicensed operation. This standard was approved in Jan 2003. A substandard that addresses Mesh network architectures is included as part of this standard.

• IEEE 802.16c- Specifies system profiles designed to improve interoperability in the 10-66 GHz frequency range. This standard was approved in December 2002.

• IEEE 802.16e- Addresses both fixed and mobile operations in licensed bands in the 2-6 GHz frequency range. Mobile operation is designed for vehicles moving up to 150 km/hour.

• IEEE 802.16f - Addresses mesh networking architectures.

8.3. Deployment Architectures

A typical IEEE 802.16 network is made up of one central base station (BS) that communicates with one or more Subscriber Stations (SS). This communication can take place in several different network architectures to include:

• Point-to-point (PTP): Connections between two nodes, in this case a BS and a SS. PTP links have the advantage of extended range over PMP links.

• Point-to-multipoint (PMP): A connection between one BS and multiple SS nodes. Generally involves the use of sector or omni-directional antennas to create a coverage area with more than one SS. This architecture supports multicast communication.

• Point-to-consecutive point (PTCM): Involves the creation of a closed loop through multiple PTP connections.

• Mesh: IEEE 802.16a substandard, where each node is able to route data adaptively to its destination. Mesh architectures are self organizing and self healing.

8.4. The Physical Layer (PHY)

The IEEE 802.16 standard and the IEEE 802.16a standard each specify a separate air interface due to differences in frequency range, but they both use the same MAC protocol. This ability to apply one MAC to multiple PHY interfaces has much potential application in both commercial and military applications. The two separate air interface standards make it possible for operators to take advantage of the strengths of either frequency range dependent on the deployment situation. For military purposes, it may be possible to adapt the IEEE 802.16 standard to employ a PHY that is better suited to military operations.

8.4.1. 10-66 GHz Systems

Higher frequency microwave signals in the 10-66 GHz frequency range are addressed in the IEEE 802.16 standard. This standard supports only LOS operation and has shorter ranges of only a few kilometres, when compared to lower frequency systems. This frequency range is capable of supporting data rates up to 120 Mbps. The primary advantage of this frequency range over others is the abundant availability of bandwidth. Unlike the lower frequency ranges where frequency bands are often less than 100MHz wide, most frequency bands above 20GHz can provide several hundred megahertz of bandwidth. Additionally, channels within these bands are typically 25 or 28 MHz wide.

IEEE 802.16 utilizes a single carrier modulation (WirelessMAN-SC) using either (1) quadrature phase shift keying (QPSK), (2) 16-bit quadrature amplitude modulation (QAM) or (3) 64 QAM. Communication on the downlink, which typically involves one BS talking to multiple SSs, is handled using time division multiplexing (TDM). The uplink uses TDMA combined with DAMA techniques. The uplink channel is divided into various time slots and the assignment of those slots is dynamically controlled by the MAC of the BS and based on the moment to moment needs of the system.

IEEE 802.16 allows for both time division duplexing (TDD) and frequency division duplexing (FDD). In TDD, the uplink and downlink take turns transmitting on a shared channel, while FDD allocates separate channels to each. The standard also supports half duplex FDD where the uplink and the downlink share one channel much like in TDD.

Another feature unique to the higher frequency IEEE 802.16 standard is the use of adaptive burst profiling. Adaptive burst profiling makes it possible for the radio to make adjustments to the modulation and coding schemes being used in response to changing environmental conditions and the resulting signal quality. Systems using adaptive burst profiling will constantly monitor signal quality and make adjustments on a frame by frame basis, shifting between the more efficient and less robust QAM to the less efficient but more robust QPSK as needed.

8.4.2. 2-11 GHz Systems

The IEEE 802.16a standard addresses lower frequency microwave signals in the 2-11 GHz frequency range. Signals in this frequency range have many advantages over higher frequency signals to include the ability to penetrate walls, NLOS performance, longer ranges than higher frequency signals (over 30 miles using highly directional antennas), support for more complex modulation, and higher robustness and spectral efficiency. Indeed, many of the IEEE 802.16 PHY's most advantageous capabilities are found in this frequency range.

IEEE 802.16a uses orthogonal frequency-division multiplexing (OFDM) with a 256-point transform. A brief description of OFDM is provided below: Orthogonal FDM's (OFDM) spread spectrum technique distributes the data over a large number of carriers that are spaced apart at precise frequencies. This spacing provides the "orthogonality" in this technique which prevents the demodulators from seeing frequencies other than their own. The benefits of OFDM are high spectral efficiency, resiliency to RF interference, and lower multi-path distortion.

IEEE 802.16a also uses TDM and TDMA to schedule uplink and downlink transmissions. Additionally, it uses TDD and FDD in much the same way that IEEE 802.16 systems do.

8.4.3. Error Control

IEEE 802.16 uses two methods to control errors in the PHY: Forward Error Correction (FEC) and Automatic Retransmission Request (ARQ).

a. Forward Error Correction

FEC is common to both air interfaces. IEEE 802.16 normally uses Reed-Solomon GF (256) FEC, but has the option of using the more robust Block Turbo code to either increase the range of the BS or increase throughput. A brief description of Reed Solomon FEC is provided below:

Reed-Solomon error correction is a coding scheme which works by first constructing a polynomial from the data symbols to be transmitted and then sending an over-sampled plot of the polynomial instead of the original symbols themselves. Because of the redundant information contained in the over-sampled data, it is possible to reconstruct the original polynomial and thus the data symbols even in the face of transmission errors, up to a certain degree of error.

b. Automatic Retransmission Request

ARQ is a PHY characteristic that is used to deal with errors occurring due to propagation anomalies. ARQ involves the retransmission of individual bits of data that may have been lost in the original transmission. The efficiency of retransmitting individual bits makes it possible to correct errors before the data is sent to a higher layer for processing. ARQ is a feature of IEEE 802.16a only and is not specified in the IEEE 802.16 standard.

8.4.4. Framing

The IEEE 802.16 PHY uses frames of 0.5, 1 or 2 milliseconds in duration. Each frame is divided into physical slots that are 4-QAM symbols long. Physical slots are used for bandwidth allocation and PHY transitions. In TDD systems, each frame is divided between the uplink and downlink sub-frame portions. For each frame, the downlink sub-frame is transmitted first, followed by a transmit/receive gap that allows the hardware time to switch between transmitting and receiving, which is then followed by the uplink sub-frame. There is also a brief time gap between frames. In FDD systems, transmitting and receiving occur simultaneously on separate channels.

a. Downlink Sub-frame

As shown in Figure 8.1, each downlink sub-frame begins with a preamble followed by a frame control section that contains a downlink map (DL-MAP) message and an uplink map (UL-MAP) message. The frame start preamble is a 32-symbol sequence generated by repeating a 16-symbol sequence. The frame control section is used to pass control information for the channel to all SSs, and this data is not encrypted.

The DL-MAP portion of the frame control section provides listening SSs with the characteristics of the downlink channel. This information includes: PHY synchronization (i.e., schedule of physical layer transitions to include modulation and FEC changes), a downlink channel descriptor message (DCD), a programmable 48-bit BS identifier, and the number of data elements to follow. The DCD and the BS identifier identify the channel and the BS, respectively, and thus together are useful for situations where a SS is on the border of multiple IEEE 802.16 sectors or cells. The DL-MAP message shall be organized as shown in Table 8.2.

Table 8.2 The DL-MAP Message Format

Syntax	Size	Notes
DL-MAP_Message_Format() {		
Management Message Type = 2	8 bits	
PHY Synchronization Field	Variable	See appropriate PHY specification.
DCD Count	8 bits	
Base Station ID	48 bits	
Number of DL-MAP Elements *n*	16 bits	
Begin PHY Specific Section {		See applicable PHY section.
for (*i* = 1; *i* <= *n*; *i*++) {		For each DL-MAP element 1 to *n*.
DL_MAP_Information_Element()	Variable	See corresponding PHY specification.
if !(byte boundary) {		
Padding Nibble	4 bits	Padding to reach byte boundary.
}		
}		
}		
}		

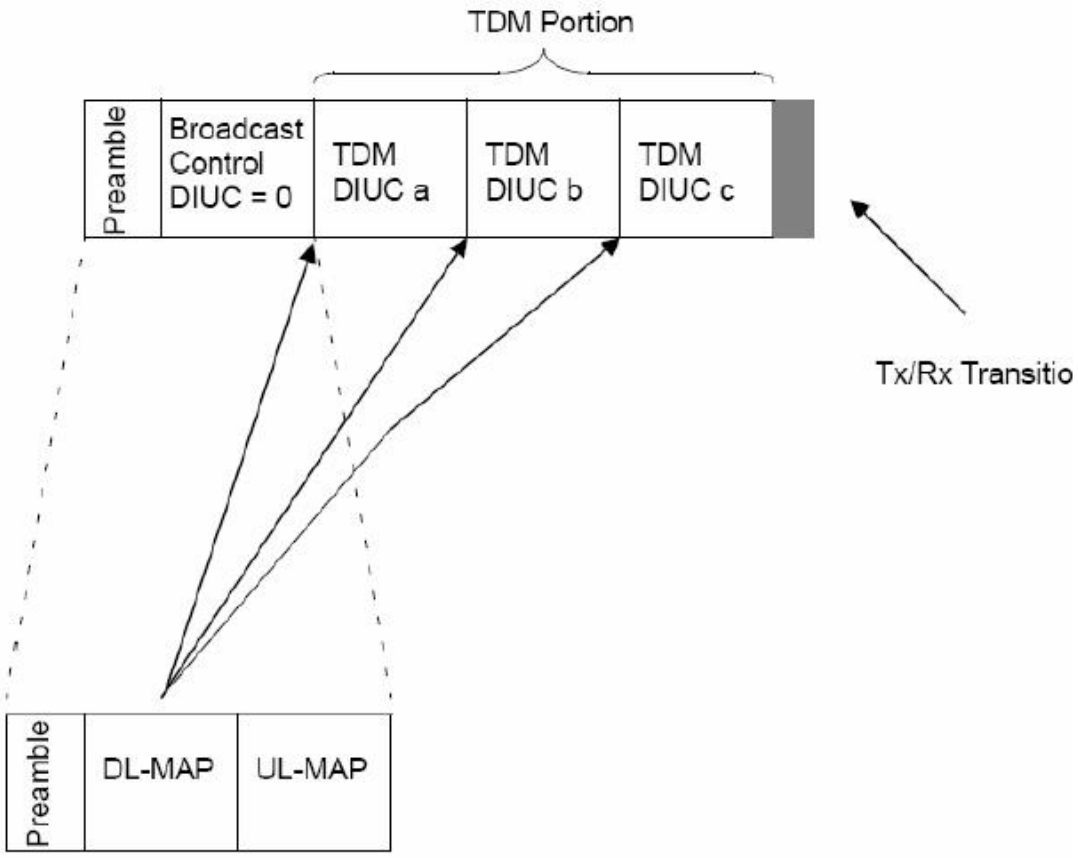

Figure 8.1 TDD Downlink Sub-frame Structure

The UL-MAP is used to communicate uplink channel access allocations to the SSs. Information provided in the UL-MAP include: Uplink channel identifier, uplink channel descriptor (UCD), number if information elements to map, allocation start time and map information elements. The UCD is used to provide SSs with information regarding the required uplink burst profile. The map information elements message identifies the SS this information applies to by using a connection identifier (CID). This message also provides an uplink interval usage code (UIUC) and offsets that are to be used by the SS to transmit on the uplink. The uplink interval usage code is used to specify the burst profile to be used by the SS on the uplink. The UL-MAP message shall be organized as shown in Table 8.3.

The frame control section is typically followed by a TDM portion where downlink data is transmitted to each SS. These TDM sections are used for transmitting data or control messages to specific SSs. Each of these transmissions is carried out according to the burst profile negotiated between the BS and the SS and data is transmitted in order of decreasing robustness.

The recipient SS is specified in the MAC header of the each data transmission, not in the DL-MAP portion of the frame control message. This makes it necessary for full duplex SSs to listen to all downlink sub-frames in order to filter out their data. In FDD systems with half duplex capability, the TDM portion of the downlink sub-frame may be followed by a TDMA portion designed to allow half duplex systems to regain synchronization with the BS. In this case, a separate preamble would precede each TDMA slot as shown in Figure 8.2. Burst profiles parameters and the presence of a TDMA portion will vary on a frame by frame basis as dictated by bandwidth and service demands.

Table 8.3 The UL-MAP Message Format

Syntax	Size	Notes
UL-MAP_Message_Format() {		
Management Message Type = 3	8 bits	
Uplink Channel ID	8 bits	
UCD Count	8 bits	
Number of UL-MAP Elements *n*	16 bits	
Allocation Start Time	32 bits	
Begin PHY Specific Section {		See applicable PHY section.
for (*i* = 1; *i* <= *n*; *i*++) {		For each UL-MAP element 1 to *n*.
UL_MAP_Information_Element()	Variable	See corresponding PHY specification.
}		
}		
}		

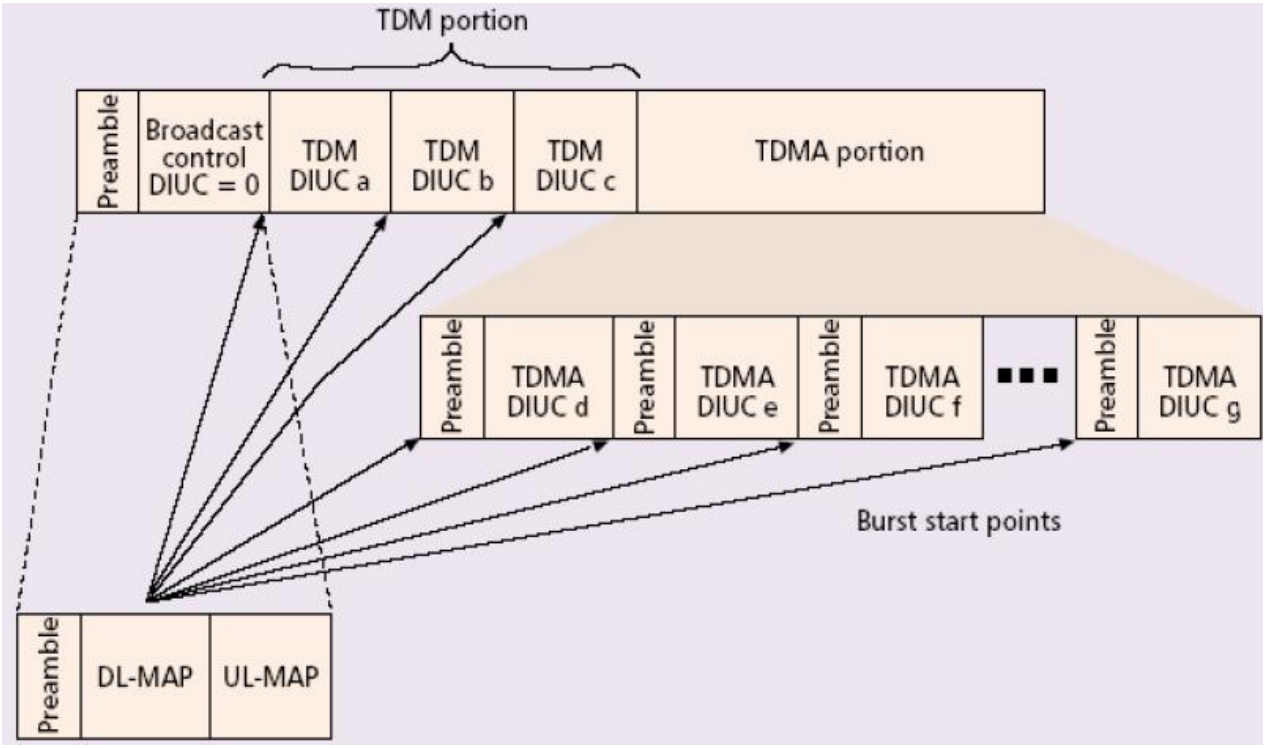

Figure 8.2 The Downlink Sub-Frame Structure

b. Uplink Sub-frame

The uplink sub-frame is used for SSs to transmit information to the BS. A typical uplink sub-frame structure is shown in Figure 8.3.

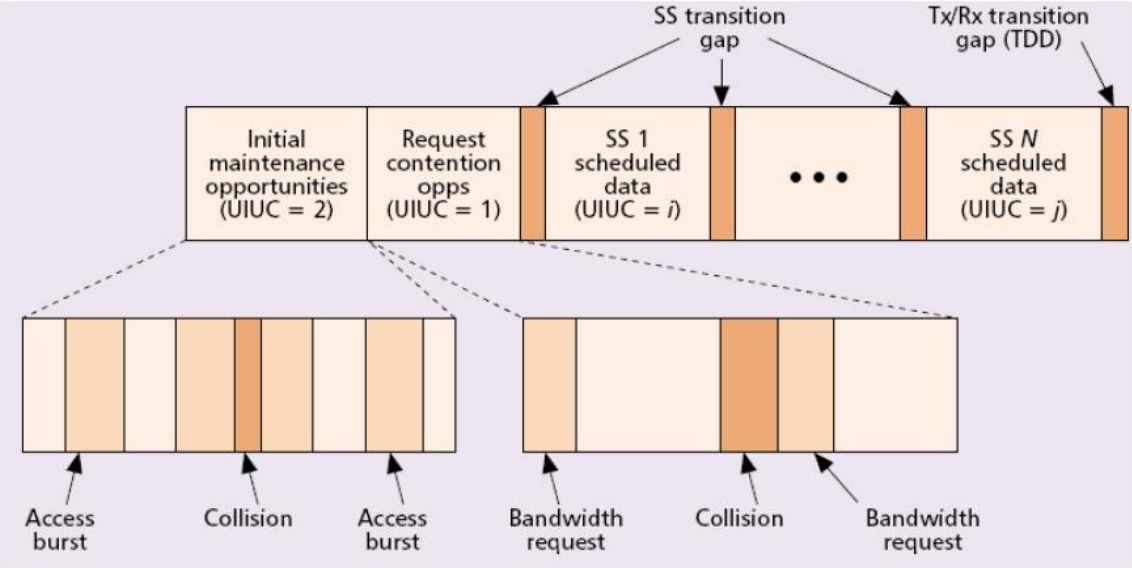

Figure 8.3 The Uplink Sub-Frame Structure

There are three possible burst classes that may be present in any uplink sub-frame:

- Contention based initial maintenance or initial access opportunities
- Contention based opportunities defined by request intervals as a response to multicast or broadcast polling
- Non-contention based and scheduled intervals allocated to specific SSs in UL-MAP bandwidth grants from the BS.

Any of these three burst classes may be present in any frame, in any order and in any quantity per frame as dictated by the BS scheduler in a UL-MAP message.

Initial maintenance/access timeslots include extra guard time to account for SS trying to acquire initial access and who have not yet resolved timing issues related to

their range from the BS. Additionally, collision time gaps, SS transition time gaps and transmit/receive time gaps are used to reduce the possibility excessive collisions.

8.4.5. Transmission Convergence (TC) Sublayer

The TC sub-layer exists between the PHY and the MAC. The TC sub-layer takes variable length MAC protocol data units (PDU) and organizes them within fixed length FEC blocks prior to transmission. A 1-byte pointer is then added to at the beginning of the TC PDU to indicate the first byte of the next MAC PDU within the TC PDU. In the event of lost data transmissions, this pointer allows for resynchronization between the SS and the BS. The TC PDU format is shown in Figure 8.4.

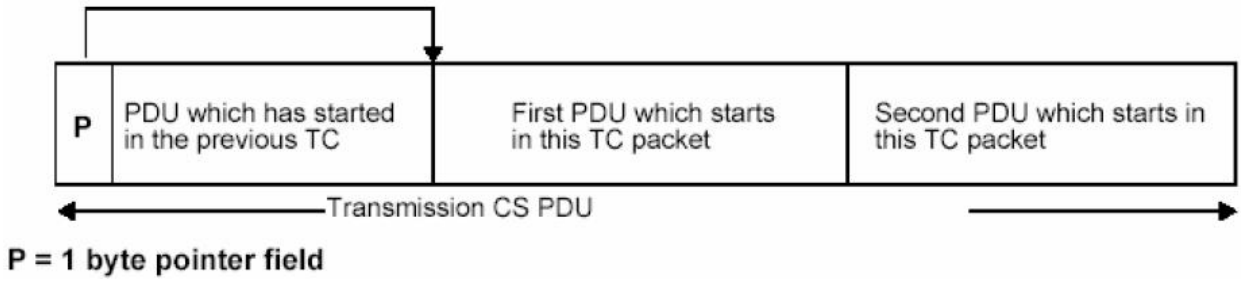

Figure 8.4 The TC PDU format

8.5. Medium Access Controller Layer (MAC)

The IEEE 802.16 MAC is the mechanism responsible for the efficient sharing of the available medium. The IEEE 802.16 MAC is upper layer PHY protocol independent, with the capability of supporting services to include legacy TDM voice and data, IP connectivity, or packetized applications like VOIP. It is also capable of supporting either continuous or bursty traffic and ensuring that QoS is in keeping with the type of traffic being transmitted. Additionally, the IEEE 802.16 MAC is capable of supporting Asynchronous Transfer Mode (ATM) and guaranteed frame rate (GFR) services. Through a variety of methods that we will discuss shortly, the MAC is able to provide differentiated service to users on the same medium. Most

importantly, the MAC is able to guarantee a specified service level and required QoS for each connection. As an example, one sector of a BS is capable of supporting guaranteed T1 service to business customers while simultaneously providing best effort DSL services to other customers within the same service area.

8.5.1. Connection Orientation

A connection is a unidirectional mapping between base station and subscriber station medium access control peers for the purpose of transporting a service flow's traffic. IEEE 802.16 is a connection oriented protocol, where all services are mapped to a connection. This is true even for inherently connectionless services. While each SS has a unique 48-bit MAC address, this number is not used to reference the multiple connections associated with each SS. Instead, connections are referenced using a 16-bit CID. CIDs are used for all interactions with the BS to include bandwidth requests, connection QoS control, and routing data to the appropriate sublayer.

When a SS is first introduced into a network, the BS will assign three management connections in each direction. Each connection is used for transmitting messages of different lengths and urgency. The three management connections and the type of messages they transmit are as follows:

- The basic connection - short, time critical MAC and radio link control messages
- The primary management connection - longer, more delay tolerant messages (ex. authentication or connection setup messages)
- The secondary management connection - standards based messages such as DHCP, TFTP and SNMP messages.

There are various types of connections to support many of the IEEE 802.16 MAC's various functions. A second group of connections, known as transport connections, are established according to the services being supported and the required QoS and traffic parameters. These connections are not to be confused with

layer 4 or Transport layer connections found in the OSI model. Transport connections are typically assigned in pairs. Other connections might be established for contention based initial access, broadcast transmissions, multicast transmissions, etc.

8.5.2. The MAC PDU

a. PDU Description

The definition of a MAC PDU is as follows: The MAC PDU is the data unit exchanged between the MAC layers of the BS and its SSs. A MAC PDU consists of a fixed length header, a variable length payload, and an optional cyclic redundancy check (CRC).

More specifically, PDUs are exchanged among peer entities in the same protocol layer, from higher to lower layers in the downward direction and from lower to higher layers in the upward direction. This exchange of PDUs is shown in Figure 8.5 below. In the downward direction, each layer encapsulates the higher layer PDU into the MAC SDU format before passing it on to the next layer.

Prior to transmission, the MAC can take advantage of several methods of MAC PDU construction to maximize the efficiency of the transmission. The following methods are used in the construction of MAC PDUs:

(1) Concatenation: Involves the concatenation of multiple MAC PDUs into one transmission. May be done for either uplink or downlink transmissions.

(2) Fragmentation: Involves the division of a MAC SDU into several MAC PDUs. May be used to support services where the MAC SDU size may be very large, such as video applications. Fragmentation may also be done in both the uplink or downlink directions.

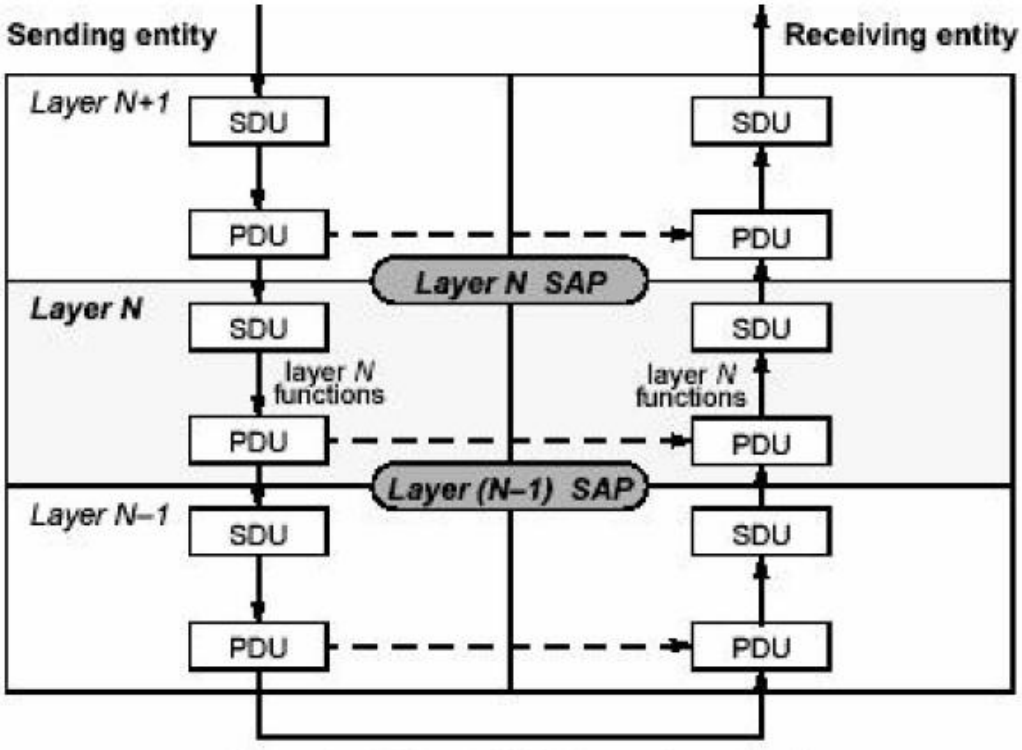

Figure 8.5 PDU and SDU in a Protocol Stack

(3) Packing: Involves the packing of multiple MAC SDUs into one MAC PDU. [Ref 21] The connection must be authorized to carry variable length packets in order to take advantage of packing. Packing may be done in either the uplink or the downlink at the discretion of the transmitting station.

8.5.3. Sublayers

The MAC is made up of three sublayers: the Service Specific Convergence Sublayer (CS), the MAC Common Part Sublayer (MAC CPS), and the Privacy Sublayer. The sublayers are organized as shown in Figure 12, with the CS on top as the interface to higher layers, the MAC CPS below the CS, an the Privacy Sublayer below the MAC CPS. Between each sublayer lies a service access point, which acts as an interface between the two layers it borders. It is important to note that the CS SAP acts as the interface to layer 3 - i.e. to a router or protocol stack in the end system.

8.5.4. Radio Link Control

The IEEE 802.16 Radio Link Controller (RLC) is responsible for the management of adaptive burst profiles, power control and ranging. A different burst profile is used for each channel as determined by the RLC, based on "a number of factors, such as rain region and equipment capabilities". Under favourable link conditions, the RLC will employ the most bandwidth efficient burst profiles available, and will revert to less efficient burst profiles when link conditions become less favourable. Through the use of adaptive burst profiles IEEE 802.16 is able to support a link a planned link availability of 99.999%. The adjustment of burst profiles, power and ranging parameters is controlled by the BS, which monitors signal quality on the uplink and manages requests from associated SSs to make adjustments on the downlink. Power control and initial ranging begin immediately upon initial channel acquisition and will be described below.

8.5.5. Network Entry and Initialization

Figure 8.6 shows the stages of an error free initialization of a SS entering a network. There are many possible branches from this procedure that may be invoked due to errors during initialization. This initialization procedure is designed to eliminate the need for manual configuration of each SS.

Each step in the initialization process will be covered in detail below:

a. Scanning and Synchronization to the Downlink

SSs are designed to scan their frequency lists for active downlink channels immediately upon installation or following any period of signal loss. In the case of signal loss, the SS will store the operational parameters of the last signal and will try to re-establish that connection. After acquiring a channel with a valid downlink signal, the SS will attempt to synchronize the PHY by listening for DL-MAP management messages. The SS will continue to listen for DL-MAP management

messages and in the case of missing DL-MAP messages, the SS will repeat the scanning and synchronization process.

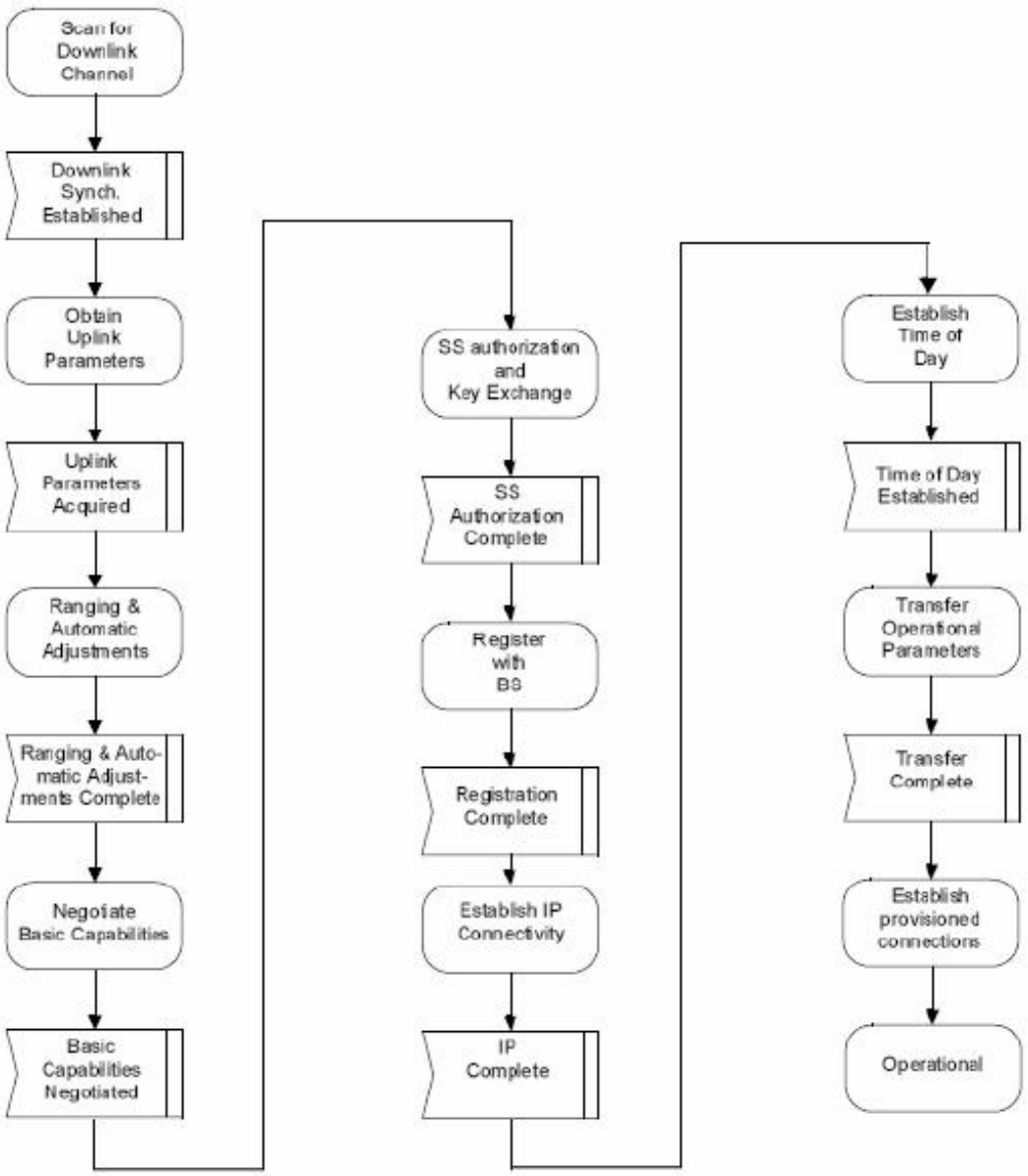

Figure 8.6 SS Initialization Overview

b. Obtaining Transmit Parameters

Once a DL-MAP message has been detected, the MAC sublayer will listen for downlink and uplink transmission parameters. By listening for UCD messages from

the BS, the SS is able to determine a usable uplink channel. UCD messages are broadcast messages, sent out periodically, providing pertinent parameters for all available uplink channels. The SS will collect UCD messages for each available channel, and will attempt to establish communications on a suitable channel. If communications fail on one channel, the SS will move on to the next suitable channel until a connection is established or the list has been exhausted, in which case it will begin the scanning process again.

c. Ranging and Power Adjustment

As described in the IEEE 802.16 standard, Ranging is the process of acquiring the correct timing offset such that the SS's transmissions are aligned to a symbol that marks the beginning of a mini slot boundary. Timing offset is dictated by the distance of the SS to the BS and the corresponding signal propagation delay. The SS begins this process by scanning UL-MAP messages for an available maintenance interval. Once an available maintenance interval has been determined, the SS will send a Ranging Request (RNG-REQ) message, within this contention based initial maintenance period, to the BS at the minimum power level. If this transmission does not receive a response, the SS will increase the power level incrementally as necessary, but not to exceed the maximum specified transmission power. The BS will reply with a Ranging Response (RNG-RSP) message, which specifies the appropriate timing advance and power adjustment for the SS, as well as the basic and primary managements CIDs.

d. Negotiation of Basic Capabilities

The SS will use SS Basic Capability Request (SBC-REQ) messages to report its capabilities to the BS. This message provides the SS's PHY capabilities, supported modulation and coding schemes, and duplexing methods supported. The BS will then respond using the SS Basic Capability Response (SBC-RSP) message to detailing which of the SS's capabilities it will support. This response will be used to adjust the

burst profile to the most efficient usable profile. Up to this point all previous transmissions are carried out using the most robust burst profile available.

e. Authorize SS to Perform Key Exchange

Authorization and key exchange will be covered in more detail in the security section to follow.

f. Registration

According to the IEEE 802.16 standard, Registration is the process by which the SS receives its Secondary management CID and thus becomes manageable. This is accomplished through the Registration Request (REG-REQ) message sent by the SS and the Registration Response (REG-RSP) message sent by the BS.

g. Establish IP Connectivity

The SS may also include the version of IP it uses in the REG-REQ. If not included the BS will authorize the use of the default IPv4 for the Secondary Management Connection. The SS and the BS will then use Dynamic Host Configuration Protocol (DHCP) on the Secondary Management connection to complete IP connectivity.

h. Establish Time of Day

Time of day is used for time stamping of logged events by both the BS and the SS. The SS again uses the Secondary Management connection to retrieve the time from the server. The transmission is sent via user datagram protocol (UDP). The time returned from the server is combined with the SS's timing offset in order to determine the current local time.

i. Transfer Operational Parameters

The SS will use TFTP to transfer the SS configuration file. The configuration file contains the configuration settings for a variety of parameters used in the operation of the SS.

j. Set Up Connections

The SS will next begin to establish connections for pre-provisioned service flows, where a service flow is defined as the unidirectional transport of packets on either the uplink or the downlink. Each service flow is associated with a specific set of QoS parameters for the supported service. These service flows utilize a two phase activation model where a service flow may be admitted (BS has resources reserved, but service is not active), or active (BS has resources reserved and service is active). A third possible state for a service flow is the provisioned state, where the BS has assigned a service flow identifier, but has not reserved any resources for this service flow.

8.5.6. Bandwidth Requests and Grants

IEEE 802.16 manages the allocation of bandwidth by using a request / grant protocol. In this protocol, SSs request bandwidth allocations from the BS through a variety of methods, which will be explored in more detail below. As previously discussed, the BS makes bandwidth assignments by allocating transmission timeslots (via TDMA) only to those SSs that have submitted a request for bandwidth (via DAMA). The BS will use UL-MAP messages to relate the bandwidth allocations to all SSs on the network.

IEEE 802.16 subscriber stations can be divided into two classes based on how they handle bandwidth grants. The first class of SS accepts bandwidth grants for each connection, or on a grant per connection (GPC) basis. The second class of SS is able

to accept grants for all of the SS's bandwidth needs, or on a grant per SS (GPSS) basis. These are covered in more detail below:

a. GPC

The GPC SS receives grants only for specific connections (to include management connections) and as a result must request bandwidth for each individual connection as needed. In addition, the GPC SS must request additional bandwidth to meet any unexpected RLC requirements. For these reasons, GPC systems are less efficient than GPSS systems, but they are also simpler.

b. GPSS

The GPSS SS receives one bandwidth grant, which it uses to meet the needs of all its connections. As a result, the SS itself must manage how much bandwidth is allocated to each connection. In situations where one connection requires more bandwidth than expected, the SS has the option of 'stealing' bandwidth (referred to as bandwidth stealing in the IEEE 802.16 standard) from another connection to cover the temporary bandwidth shortage. The BS is also responsible for priority queuing based on traffic types. The SS can then send a request to the BS requesting that it’s bandwidth grant be increased to meet its new needs. GPSS SSs are the only class of SS available in the 10-66 GHz frequency range.

Bandwidth grants are provided based on a self-correcting protocol as opposed to an acknowledged protocol. In this protocol, if the SS does not receive a bandwidth grant in reply to a bandwidth request, the SS will assume that the request was either lost or could not be fulfilled, and will simply send another request to the BS, without having to wait for some acknowledgement of the original request. This protocol eliminates the overhead associated with acknowledgement messages.

8.5.7. Bandwidth Requests

SSs typically will request bandwidth incrementally as new bandwidth requirements arise, and the BS will add the requested bandwidth to the total perceived requirement for the SS.

a. Request Periods

With incremental requests, the BS has no way of knowing whether it has granted the correct total requirement of bandwidth to the SS, since the total granted bandwidth may be affected by lost grant request packets. Due to this possibility, the, SSs may request bandwidth incrementally or on an aggregate basis. Aggregate requests are used to reset the BS's perception of the total bandwidth requirement of the SS. When a BS receives an aggregate request, it will store the requested bandwidth value as the new total requirement for the requesting SS.

There are a variety of methods available for a SS to request bandwidth allocations from the BS. Bandwidth requests may be related to the BS during bandwidth request periods specifically dedicated to a SS or during contention periods. The method of polling used by the BS to inform the SSs of upcoming bandwidth request periods is what determines whether the bandwidth request period is a dedicated or contention request period. Polling methods will be covered in the following section.

b. Bandwidth Request Header

In addition to bandwidth request periods allocated via polling, SSs may request bandwidth allocations at any time by sending the BS a bandwidth request MAC PDU with a bandwidth request header and no payload. This method of bandwidth request may be used in any bandwidth grant for GPSS SSs and in either grant request intervals or data grant intervals for a specific connection.

c. Piggyback Request

A similar method for requesting bandwidth is to use a grant management subheader to piggyback a request for additional bandwidth for the same connection within the MAC PDU.

8.5.8. Polling

Polling is the process used by the BS to allocate bandwidth request opportunities to SSs. When the BS wants to notify a SS of an upcoming bandwidth request opportunity, it will use an UL-MAP message information element (IE) to do so. The UL-MAP IE will grant sufficient bandwidth for the SS or SSs to submit their bandwidth requests during the specified request period. Bandwidth request opportunity allocations may be made on a unicast, multicast or broadcast basis as described previously in section 4.b. of this chapter. A brief description of each polling method is provided below:

a. Unicast polling

In unicast polling, a SS is polled individually by the BS. The SS will reply with stuff bytes if the granted bandwidth is not needed.

b. Multicast and Broadcast Polling

The BS will resort to multicast or broadcast polling when insufficient bandwidth is available to individually poll SSs. Multicast and broadcast polling is also done via the UL-MAP message in the same fashion as for unicast polling. The BS reserves some CIDs for multicast or broadcast groups. The primary difference here is that the polling message is directed toward a multicast or broadcast CID instead of an individual CID or SS.

c. Poll-Me Bit

The poll-me bit is used by SSs using the Unsolicited Grant uplink scheduling service (UGS) to notify the BS that they need to be polled. The UGS will be covered in more detail in the following section. The poll-me bit is part of the grant management subheader. Once the poll-me bit has been detected, the BS will issue a unicast poll to the SS requesting it.

8.5.9. Uplink Scheduling Services

IEEE 802.16 uses predefined uplink scheduling services to increase the efficiency of uplink transmissions on each connection based on the service being provided by that connection. The four defined uplink scheduling services are: Unsolicited Grant service, Real Time Polling service, Non-Real Time Polling service, and Best Effort service. The scheduling service that a connection will use is determined at the time of that connection's set up. Each uplink scheduling service is further defined below:

a. Unsolicited Grant Service

This service is used primarily for synchronous, real time services which generate fixed units of data periodically, such as ATM constant bit rate (CBR), T1/E1 over ATM or Voice over IP without silence suppression. In this service, the BS provides periodic fixed size data grants, as negotiated during connection setup, without the need for the SS to send bandwidth requests.

This unsolicited granting of bandwidth eliminates the overhead and latency associated with bandwidth requests and as a result helps to reduce jitter and delay jitter. More stringent jitter requirements may be met through the use of output buffering.

The SS is able to provide feedback to the BS concerning the state of service flows by employing the slip indicator flag in the grant management subheader. The slip

indicator flag is used to indicate a queue backlog, which may be caused by a variety of factors to include lost grants or clock skew with outside networks. Once the BS has been notified of the slippage, it can grant additional bandwidth in order to eliminate the backlog.

b. Real Time Polling Service

This service is designed to meet the needs of real time services needing to transmit periodic, variable sized data packets. This service is well suited for applications such as streaming video or audio, or VoIP. A suitable military application might be in missile guidance systems, where a missile in flight might require periodic tracking information updates. Real time polling works by allocating periodic dedicated (unicast) bandwidth request opportunities to each connection. Because the SS must explicitly request bandwidth, there is more overhead and latency associated with this service than with Unsolicited Grant service, however, some efficiency is gained through the use of variable sized data packets.

c. Non Real Time Polling Service

This service works the same way as the Real Time Polling service, except that connections use contention based access opportunities to transmit bandwidth requests. Unicast polling opportunities are also used to guarantee at least a minimal reserved traffic rate, although these opportunities are less frequent than those found in Real Time polling. Non-Real Time Polling is well suited for supporting services that can tolerate some delay jitter, such as high bandwidth FTP, Internet connections, and ATM GFR. Non-Real Time polling also utilizes the traffic priority parameter, contained in the SS configuration file and established at connection setup, to determine which service flows have priority in relation to others. As stated in the IEEE 802.16 standard, given two service flows identical in all QoS parameters besides priority, the higher priority service flow should be given lower delay and buffering preference.

d. Best Effort Service

There are no throughput or delay guarantees associated with this service. Connections use contention based opportunities to request bandwidth. Additionally the SS may use unicast or unsolicited opportunities to request bandwidth. The availability of unicast opportunities is subject to the load of the network and is not guaranteed. The best effort service is the most bandwidth efficient because it does not reserve bandwidth for a station that may or may not be using it.

8.5.10. Quality of Service

There are various parameters associated with QoS in the IEEE 802.16 standard. These parameters are used at the establishment of a service flow to determine the QoS requirements of a supported service. Below are some of the QoS parameters specified in the IEEE 802.16 standard:

• QoS parameter set type - specifies the proper application of the QoS parameter set to either a provisioned, admitted or active set.

• Traffic priority - used to assign a priority to a service flow's traffic.

• Maximum sustained traffic rate - expressed in bits per second.

• Maximum traffic burst - calculated from the byte following the MAC header to the end of the MAC PDU.

• Minimum reserved traffic rate - specifies the minimum rate reserved for a service flow.

• Vendor specific QoS parameters - can be used by vendors to encode their own QoS parameters.

• Service flow scheduling type - specifies the uplink scheduling service being used for the service flow.

• Request / transmission policy - used to specify various scheduling service rules and restrictive policies on uplink requests and transmissions.

• Tolerated jitter - specifies the maximum delay variation (jitter) for a connection.

• Maximum latency - specifies maximum latency between receipt of packet on the network interface and forwarding to the RF interface.

• Fixed length versus variable length SDU indicator - indicates whether data packets must be fixed length or may be variable length.

8.5.11. Security

The IEEE 802.16 privacy sublayer provides users privacy by encrypting the link between the BS and the SS, and it provides protection against theft of service by encrypting service flows within the network. The privacy sublayer employs an authenticated client/server key management protocol that is capable of supporting the Advanced Encryption Standard (AES). In this protocol the BS, acting as the server, controls key distribution to the SS, which acts as the client.

The privacy sublayer employs to component protocols to carry out all security related tasks. The first is an encapsulation protocol, which is used for the encryption of data packets across the network. This protocol defines the rules associated with using cryptographic suites to encrypt the MAC PDU payload. Cryptographic suites are defined as pairings of data encryption and authentication algorithms.

The second component of the privacy sublayer is the Privacy Key Management Protocol (PKM). PKM is used to provide secure distribution of keys between the BS and SSs. This protocol is further used by the BS and the SS to keep synchronization of keying data between them, and by the BS to control access to network services.

a. Packet Data Encryption

When encryption is enabled on an IEEE 802.16 system, not all packets or even all portions of packets will be encrypted. In order to facilitate ranging and registration, all MAC management messages are sent in the clear. Additionally, encrypted data packets contain an encrypted payload with an unencrypted header. The unencrypted MAC PDU header will contain information specific to the encryption such as an

encryption control field, an encryption key sequence field, and the corresponding CID. This information is used by the receiving BS or SS to decrypt the MAC PDU payload.

b. Key Management Protocol

All IEEE 802.16 SSs shall contain a manufacturer issued X.509 digital certificate, which is used for SS authentication and initial authorization key exchange. The digital certificate will contain the SS's public key as well as its MAC address. Upon authentication, the BS will use the SS's public key to encrypt the authorization key (i.e., a shared secret), and the authorization key will be used to encrypt any subsequent data and key exchange. In addition to digital certificates, all SSs have either factory installed RSA private/public key pairs, or the appropriate algorithms to generate these keys dynamically. The RSA public-key encryption algorithm, and strong symmetric algorithms are used by the PKM protocol to facilitate key exchange.

c. Security Associations

A security association (SA) is defined as the set of security information a BS and one or more of its client SS in order to support secure communications. Upon initialization, each SS will establish at least one SA with the BS. With the exception of the basic and primary connections, all new connections are mapped to a SA.

8.6. Summary

IEEE 802.16 is a well conceived standard from an organization with a good history of producing sound standards. The fact that the WiMax alliance has undertaken the task of ensuring interoperability should accelerate the adoption of the standard and help to produce high quality equipment standards. The IEEE 802.16 standard offers superior performance, support for large numbers of users, robust links, and the future promise of mobility, and mesh networking among other things.

The IEEE 802.16 standard is fertile ground for military experimentation and testing, and with a few adaptations, may produce a communications transformation within DOD.

CHAPTER NINE
THE CHALLENGES OF VoIPoW
(VOICE OVER IP OVER WIRELESS)

9.1. Introduction

Today, the accumulated volume of data traffic is on the verge of surpassing the accumulated volume of voice traffic in all public networks. Given the growth in the areas of wireless voice and data, we see that the combination of mobile and Internet communication constitutes the driving force behind third-generation wireless systems, which promise to support at least 144 kbit/s (384 kbit/s) in all radio environments, and up to 2 Mbit/s in low-mobility and indoor environments.

The standardization of third-generation wireless systems is rapidly progressing in all major regions of the world. These systems which go under the names of IMT-2000 (ITU), UMTS, and EDGE (ETSI/3GPP) will extend the services provided by current second-generation systems (GSM, PDC, IS-136, and IS-95) with high data-rate capabilities. The main application for these services will be wireless packet transfer; for instance, for wireless access to the Internet. However, support will also be provided for high data-rate circuit-switched services, such as real-time video.

9.1.1. UMTS

The universal mobile telecommunications system (UMTS) is being standardized in the Third Generation Partnership Project (3GPP), which is a joint effort between the European Telecommunications Standards Institute (ETSI) and the Association of Radio Industries and Broadcasting (ARIB, Japan). The basic radio-access technology for UMTS/IMT-2000 in all major areas of the world is wideband code-division multiple access (WCDMA). The 1999 release of the UMTS standard was the first to be implemented in commercial products.

The radio-access part the universal terrestrial radio access (UTRA) includes a frequency-division duplex (FDD) mode and a time-division duplex (TDD) mode. The FDD mode is based on pure WCDMA, whereas the TDD mode includes an additional time-division multiple access (TDMA) component.

The WCDMA system, which uses wideband direct-sequence technology (DSCDMA), fully supports the UMTS and IMT-2000 requirements for 384 kbit/s wide-area coverage and 2 Mbit/s local coverage. Particularly noteworthy features of WCDMA are;

- support for interfrequency handover, which is necessary for high-capacity hierarchical cell structures (HCS);
- support for capacity-improving technologies, such as adaptive antennas and multiuser detection.
- built-in service flexibility, which provides spectrum-efficient access for current as well as future applications; and
- efficient handling of bursty applications via an advanced packet-access mode.

WCDMA also provides efficient support for multimedia services; that is, for transferring multiple services on one connection.

9.1.2. EDGE

The GSM and TDMA/136 technologies make up the foundation on which the common radio access for data services will be offered. The enhanced data rates for GSM and TDMA/136 evolution (EDGE) concept, which ETSI and the Universal Wireless Communications Consortium (UWCC) have adopted as the migration path from GSM and TDMA/136, fulfils the requirements for third-generation wireless systems according to IMT-2000. EDGE is capable of offering data services of up to 384 kbit/s and is thus a global complement to the UMTS radio-access network.

The roadmap for EDGE standardization has been divided into two phases. Initial emphasis was placed on enhanced general packet radio service (EGPRS) and enhanced circuit-switched data (ECSD). According to the ETSI time plan, these standards were part of the 1999 release. The second phase of EDGE standardization, which is targeted for release in 2000, will define improvements for multimedia and real-time services. Other objectives will include the alignment of services and interfaces with UMTS, to allow EDGE and UMTS to share a common core network.

9.1.3. Real-time IP Applications Over Wireless

Second-generation radio-access technology brought mobile telephony to the market. Third-generation radio-access technology will extend beyond basic telephony: a common, IP-based transport and service platform will offer mobile users a multitude of real-time and interactive services.

Typical services with real-time requirements are voice and video, as well as delay sensitive applications, such as traffic-signalling systems, remote sensoring, and systems that provide interactive access to WWW servers. The focus of this article, however, is on voice service. The voice service of third-generation wireless systems must, at the very least, offer the same high level of voice quality, and be as spectrum efficient, as present-day second-generation realizations. The challenge is to implement end-to-end service on IP-based transport.

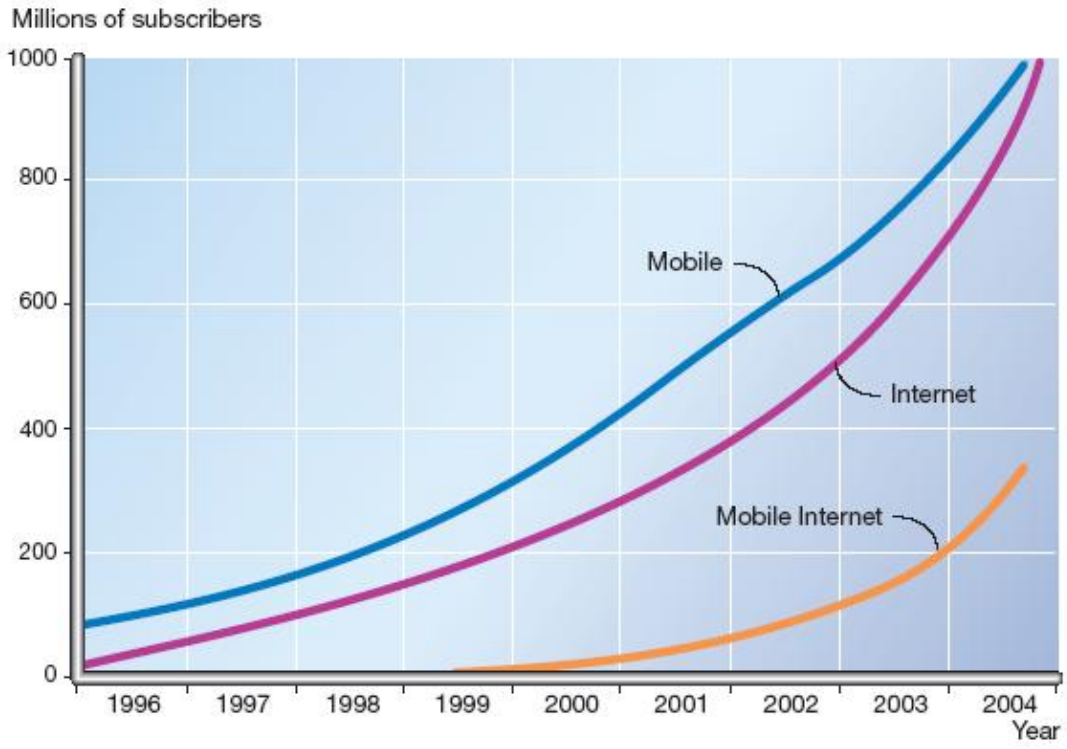

Figure 9.1 The strong growth of mobile communication is expected to continue

Experts predict that by the year 2003/2004 there will be close to one billion subscribers of cellular systems worldwide. Similarly, the Internet will continue to grow. By 2004, the number of subscribers to the Internet is also expected to reach one billion. Of this group, more than 350 million persons will subscribe to the mobile Internet (Eriksson & Olin, “The challenges of voice over IP over wireless”, 2000).

The main advantage of running IP all the way over the air interface is service flexibility. To date, cellular-access networks have been optimized for voice quality and spectrum efficiency. The demand for service flexibility adds a new parameter, as illustrated by Figure 9.2. Since there are no dependencies between an application and the access network, almost anyone can develop new applications. But for services like voice over IP over wireless (VoIPoW), the main challenge is to achieve quality and spectrum efficiency.

To date, all cellular systems that provide voice service have been optimized in a two dimensional space whose X-axis and Y-axis are voice quality and spectrum efficiency, respectively. Now, a third dimension is being added in the form of IP service flexibility. By bridging the radio interface with IP packets, we suffer a lot of protocol overhead, which runs counter to the goal of spectrum efficiency.

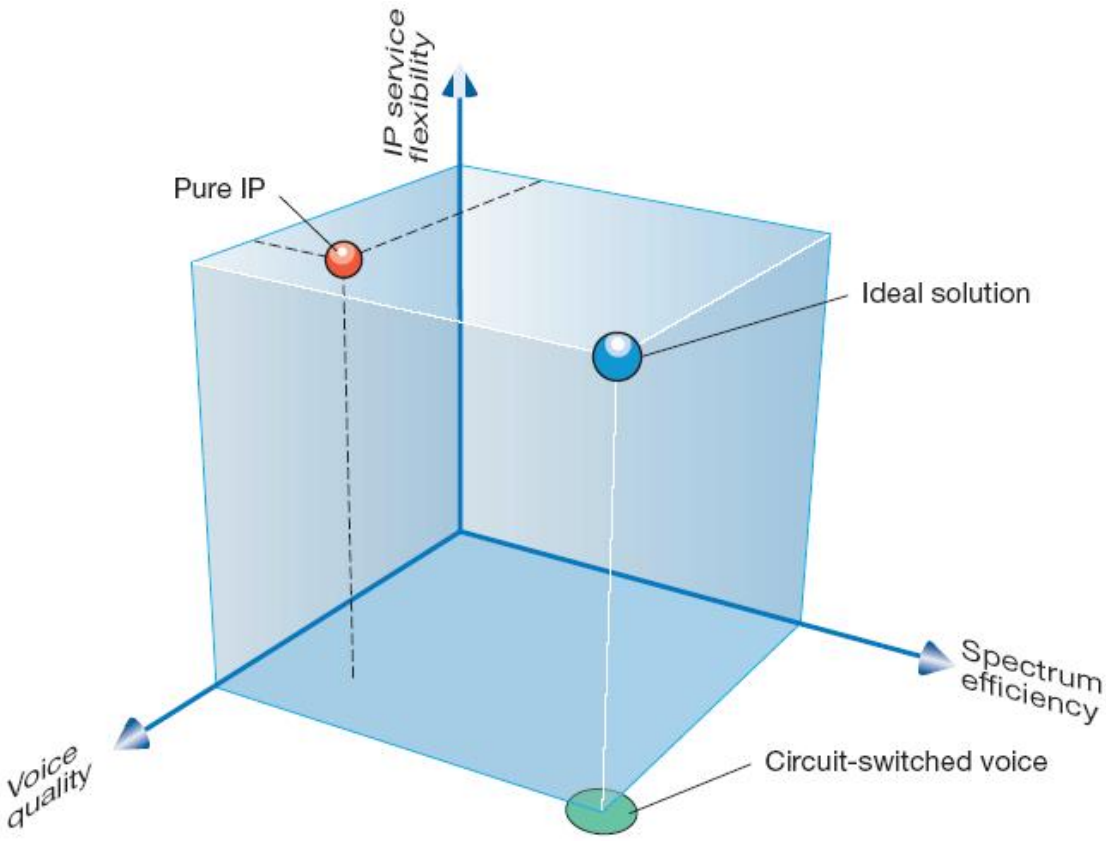

Figure 9.2 The voice-over-IP-over-wireless challenge cube

9.2. Network Architecture Overview

To facilitate our ensuing discussion, let us briefly describe the VoIP service. The basic components of the voice service are two user terminals with IP-based voice applications and a network that provides end-to-end transport between the terminals (Figure 9.3). The terminals exchange voice samples using the real-time transport protocol (RTP), which has been standardized by the IETF.

In some situations, terminals can establish and maintain communication without the involvement of a third-party entity. At other times, however, the two user endpoints cannot establish end-to-end communication without outside intervention; for example, when they do not know one another's IP address or do not use the same voice codec. In these cases, a control plane framework is used to route incoming traffic and to negotiate terminal capabilities (codec support, multiparty conferencing, and so on) in traditional telecommunications this functionality, which is referred to as call control, is provided by, say, a GSM mobile switching center. In the IP world, there are two main methods of providing call-control functionality: ITU-T Recommendation H.323, and the IETF session initiation protocol (SIP).

Originally intended for LAN environments, H.323 is an ITU standard for multimedia applications. Today, however, the standard is being adopted for broader usage. H.323 encompasses a complete architecture and a set of protocols, such as H.225 for call control and H.245 for bearer control. H.323 uses IETF protocols, such as the real-time protocol and the resource reservation protocol (RSVP). Besides end-user terminals, the H.323 architecture encompasses gatekeepers, gateways, and multiparty units. In this context, emphasis is put on the gatekeepers and gateways which constitute the VoIP server. The gatekeeper part is the controlling unit that provides call-control functionality; the gateway part contains the user plane functions. H.323 call control is based on Q.931, which is also used in GSM and ISDN.

The session initiation protocol, which is an IETF standard draft, is only one component in the IETF alternative to the H.323 paradigm for a complete multimedia architecture. Other necessary protocols and components include the session description protocol (SDP), the service access point (SAP), and the real-time control protocol (RTCP).

The session initiation and session description protocols (SIP/SDP) do not make up an architecture; they were designed for session initiation. In contrast to H.323 and GSM/ISDN, SIP/SDP does not provide a complete call-control mechanism - an SIP

proxy primarily provides routing and addressing services; device management is not included. However, the SIP proxy (or VoIP server) can be enhanced to include functionality for offering other services such as transcoding. The session initiation protocol is associated with a paradigm in which call control is distributed over several entities, and in which the user terminal plays a central role in coordinating these entities.

In summary, the two IP-based terminals exchange voice samples that have been encapsulated in RTP over the IP network. The terminals exchange control signalling between themselves or, with assistance from network entities such as a VoIP server, establish and maintain communication sessions through the network according to either the H.323 or SIP paradigm.

Both the session initiation protocol and H.323 support end-to-end solutions in which the network solely functions as a bearer. In this case we assume that an SIP or H.323 network call agent can, if so requested, support the end point (the terminal) with transcoding services.

The mobile terminal supports cellular access (UMTS/WCDMA or EDGE) and a complete VoIP application that is based on either SIP or H.323. We assume that an adaptive multirate (AMR) codec will be supported by future VoIP clients. In addition to basic UMTS packet switched access, the network contains functions for adapting media, routing calls, and for authenticating users and services.

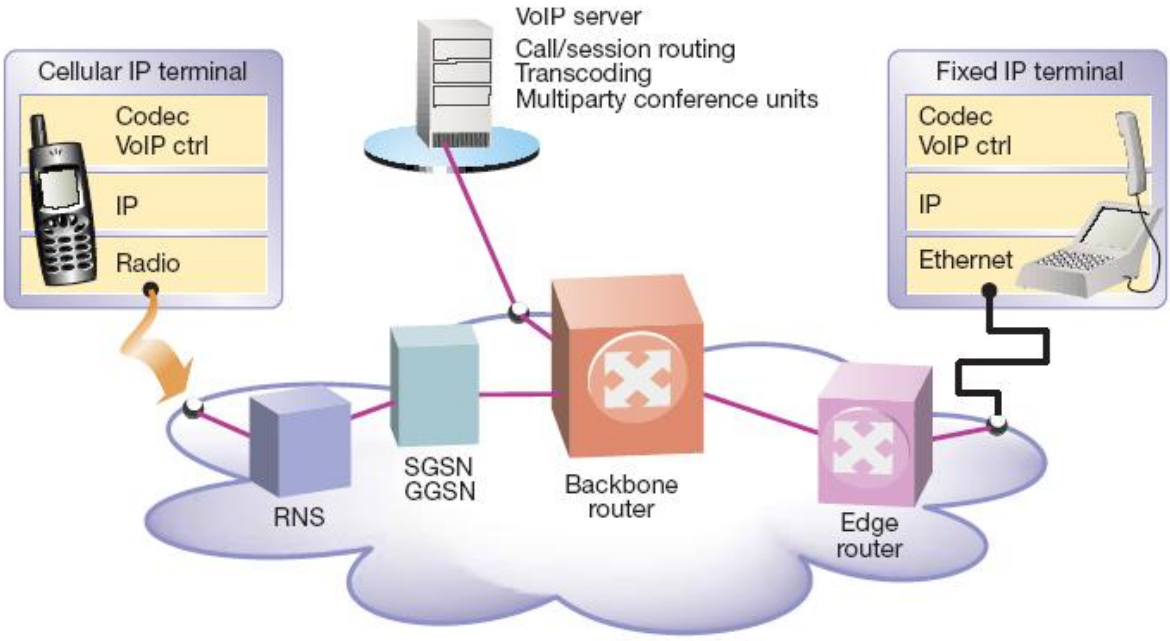

Figure 9.3 Basic VoIP Components

9.3. Antenna Design

The antenna can be provided by a vendor together with the module, purchased separately, or designed in-house. Whichever approach is chosen, it is important to ensure that the antenna provides required performance and will be appropriate for the mobile phone. None of the 802.11 standards regulates the use of antennas and one is free to choose. Antenna must provide required bandwidth, central frequency, and impedance. The following factors are also essential:

• Omnidirectional radiation pattern. This allows transmitting and receiving data to/from all directions.

• Small near field. This decreases the near field energy losses caused by close proximity of the antenna to a human body.

• VSWR equal or less than 2.0:1. VSWR of 1.5:1 is preferred (i.e. a return loss of -14 dB)

• Small dimensions

• Small PCB area required for the ground plane of the antenna
• Operating temperature range -20 0C to +60 0C
• Low cost, easy assembly

9.3.1. Antenna Polarization

There are two design goals to consider when the polarization of the 802.11 antenna is selected:

• Minimize the polarization mismatch loss
• Maximize isolation between the 802.11 and GSM antennas

The polarization state of the antenna can be chosen to be orthogonal to the GSM antenna with purpose to increase the isolation between the antennas (an isolation of 20 dB can be achieved due to orthogonal polarization). Since vertical polarization is employed by the GSM antenna, the 802.11 antenna has to use then the horizontal polarization. However, vertically polarized antennas are typically employed as the transmitting antennas at 802.11 APs. This might seems to cause the polarization mismatch losses, however other issues should be considered as well:

• The orientation of a mobile phone in space is generally random. This results in a varying polarization state of the 802.11 antenna.
• Propagation environment with multiple reflecting and scattering objects, such as typical indoor and urban outdoor environments, results in the polarization state of the received signal to be independent of the transmitted polarization.
• Polarization diversity is often employed on the receive side of the access points which compensates for the polarization mismatches between the transmitted signal by the 802.11 module and the AP receive antenna.

Thus, the decision about the polarization of the 802.11 antenna should be made based on the required isolation between antennas (may be it can be achieved through

other techniques, such as filtering), typical usage scenario, propagation environment, and the polarization properties of the antennas used by network equipment.

9.3.2. Antenna Diversity

Dual receive antenna diversity is supported in a majority of chipsets and can be considered to improve the receiving performance of the 802.11 module in a fading environment. If fading is nearly independent among the antennas then it is much less likely that both signals are weak as compared to only one of them being weak. By selecting the best signal or by combining signals the reduction in the required average received SNR for a given BER can be expected, known as a diversity gain. Thus, better operating range and throughput can be achieved. The diversity gain depends on the correlation of the fading among the antennas. Higher diversity gain can be obtained when the correlation among antenna signals is low. There are three independent methods to achieve low correlation: space diversity, polarization diversity, and pattern diversity. In pattern diversity, antennas with different radiation patterns are used. Since one of the antennas in the 802.11 module is used for both transmission and reception and its radiation pattern is required to be omni directional, pattern diversity is not suitable for this particular application and will not be covered here.

9.3.2.1. Space Diversity

In space diversity, spatial separation of the antennas is used to obtain low correlation of the fading among the antennas. It is derived the following relationship between envelope correlation and antenna separation, assuming multi path with a uniform angle of arrival distribution in azimuth and antennas with omni directional patterns.

9.3.2.2. Polarization Diversity

Two antennas with orthogonal polarization (typically vertical and horizontal or +/-45^{o} slant polarizations) are used to exploit polarization diversity. Polarization diversity is based on the concept that after sufficient random reflections, the polarization state of the signal will be independent of the transmitted polarization (in practice, however, there is some dependence of the received polarization on the transmitted polarization). Thus, in reach multi path environment, multiple versions of the transmitted signal, travelled along different paths, will have the polarization state independent of each other, which results in de-correlation of the signals in orthogonal polarizations.

One vertically polarized transmit antenna and one horizontally polarized diversity receive antenna where used to implement polarization diversity. The antennas separation distance was about 5 centimeters. In fact, polarization diversity can yield even better performance than space diversity if the antenna spacing in the space diversity scheme is less than a half of a wavelength.

Advantage of polarization diversity over space diversity is that it allows two antennas to be collocated. A single physical antenna even can be implemented with different feeds for each polarization. Additionally, polarization diversity can also compensate for polarization mismatches due to random orientation of a mobile phone. As mentioned earlier, polarization mismatch can result in signal losses of up to 20 dB. The polarization diversity scheme can achieve at least half the best-case received signal power for even the worst polarization mismatch. Combined space and polarization diversity scheme can also be considered to further improve the receiver performance.

9.3.2.3. Combining Techniques

Scanning diversity dominates in the current 802.11 chipsets, and is the simplest combining technique. In this method, the antennas are connected to a single receiver through an RF switch. During the preamble (a priori know signal), the receiver scans both antennas and selects the one with the best signal. The best signal can be in terms of signal level or SNR. If during the packet the signal from the selected antenna falls below the threshold then receiver just switches to the second antenna. The diversity gain provided by this technique is the lowest comparing to the other more advanced methods but the advantage with this method is that only one receiver is required.

Antenna diversity is two-folded. It improves the receiver performance but requires extra components - RF switch and the second antenna, i.e. more space and higher price. The RF switch also causes RF losses, thus will partly compensate the achieved diversity gain.

9.4. Conclusion

The widespread growth of the Internet has created a mass market for multimedia and information services. The challenge of providing these services via third-generation wireless systems is twofold: from the market perspective, the challenge is to merge the installed base of users in cellular and Internet environments; and in terms of technology, the challenge is to find common denominators for cellular solutions and efficient Internet access. To succeed in meeting these challenges, third-generation wireless systems must be designed to provide a multitude of services, offering considerable flexibility and cost-effective access with structured quality-of-service handling and ensuring high radio-spectrum efficiency. The UMTS and GSM/EDGE radio-network architecture and quality-of-service concept are designed to support the needs of present-day and future applications. The concept of bearer services at different network levels makes up the basis for providing end-to-end quality-of-service transport through the radio-access network is provided via radio access bearers. The main objective of the VoIPoW concept is to port voice service to

the new packet-data-based platform while maintaining the perceived quality-of-service and spectrum efficiency associated with present-day circuit-switched wireless systems. Given this objective, we see that we cannot choose a single point of implementation. Instead, the main challenge is to find suitable points of implementation that satisfy the voice service requirements for IP service flexibility or spectrum efficiency. By introducing traffic classification and header compression we can offer a spectrum-efficient VoIPoW service with high voice quality and IP service flexibility.

The aim of designing the third generation all-IP wireless network is to separate core and radio-access network components, thereby allowing a common packet switched core network (based on GPRS) to be used for UMTS and GSM/EDGE radio access networks.

A key objective of third-generation all-IP networks is to provide a capable service platform for IP-based applications. The solutions we have described for audio streams and associated control protocols will advance the UMTS network another step toward becoming a full-fledged service platform that can support demanding services, such as IP-based conversational multimedia.

CHAPTER TEN
CONCLUSION

The growth in IP based services the past few years has been explosive. It is projected that this market will continue to grow at an even higher rate for several years to come. IP telephony is expected to benefit from this deluge of IP services. There is a paradigm shift beginning to occur since more communications is in digital form and transported via packet networks, such as IP accelerating traditional voice telephony traffic. While there is more than a century of experience in designing, operating, and managing conventional circuit switched networks, relatively limited data is available about IP based networks. The success of VoIP hinges primarily on a clear understanding of the overall technology and service requirements.

Frost & Sullivan prognosticate that the annual growth rate of global IP telephony service will exhibit triple digits: VoIP products manufactured increase from under four million in 2000 to over one-half billion in 2006 (Guizani et al., 2000). Another survey has ascertained that almost half of industry experts anticipate that 15 to 20 percent of total voice traffic will run over data networks, within a two year timeframe. That number leaps to 91 percent when the time horizon is expanded to three to five years. This suggests that in the immediate future, VoIP usage will be modest. However, within a few years, more business and residential customers will adapt to VoIP as its quality and reliability improves. Eventually (anywhere from the end of this decade to the century's end) circuit switched telephony will be a memory, regulated to museums alongside the telegraph.

VoIP will impact real-time voice traffic in three different ways:

- Voice trunks can replace the analog or digital circuits that are serving as voice trunks or PSTN access trunks.
- PC to PC voice can be provided for multimedia PCs operating over an IP based networks without connecting to the PSTN, including ubiquitous wireless VoIP access.
- Telephony communications appears as a normal telephone to the caller, but may actually consist of various forms of VoIP, all interconnected to the PSTN.

VoIP networks are already incorporating IP based PBXs that emulate the functions of a traditional PBX. These allow both standard telephones and multimedia PCs to connect to either the PSTN or the Internet, providing a seamless migration path to VoIP. Moreover, traditionally telephone service can be enhanced, such as combining real time and non real time communications, high fidelity audio, conference calling and scores of other features.

Forty percent of those who have used VoIP believe it to be “the same” or “superior” to conventional dialing. However, as long as the underlying technologies of VoIP improve to address the issues presented earlier in this thesis, VoIP is poised to rocket. This will be further accelerated as IP version 4 matures to IP version 6 (predominantly due to its built in QoS support), a more reliable network infrastructure (as the Internet evolves to Internet2) and demand for voice communications via wireless devices. In fact, today’s VoIP services are merely a harbinger of the high performance integrated voice, video and data services that will be available in the not too distant future.

The overall technology requirements of an IP telephony solution can be split into four categories: signalling, encoding, transport and gateway control.

Signalling:

In terms of impacting VoIP applications, vendors are implementing an assortment of protocols, ranging from the varieties of H.323 to SIP to a proprietary signalling protocol. Presumably, major vendors will support the two major signalling protocols until it becomes clear that either one protocol will fade away or the two approaches will merge

Voice Coders:

Throughout, the use of G.729 has been shown to allow greater capacity than the use of G.711, unless a voice quality corresponding to a MOS of greater than 3.65 is required, in which case G.729 cannot be used.

Transport:

RTP and RTCP fulfil all the requirements needed by VoIP applications. Where RTP is used to transport multimedia data over an IP network, while RTCP provides feedback of the quality of the transmission and conveys minimal session control information.

Gateway Control:

Gateways are responsible for converting packet-based audio formats into protocols understandable by PSTN systems. Between the DSP processing and passing the data to the WAN, there are a number of packet-handling processes that must be encountered. A nontrivial amount of gateway-incurred latency is present, which affects perceived voice quality.

10.1. Wireless Applications

The IEEE 802.16 standard offers superior performance, support for large numbers of users, robust links, and the future promise of mobility, and mesh networking among other things.

As a second option 3G wireless networks were evaluated. Third-generation wireless systems must be designed to provide a multitude of services, offering considerable flexibility and cost-effective access with structured quality-of-service handling and ensuring high radio-spectrum efficiency. The UMTS and GSM/EDGE radio-network architecture and quality-of-service concept are designed to support the needs of present-day and future applications.

The latest developments in the IEEE 802.16 group are driving a broadband wireless access (r)evolution thanks to a standard with unique technical characteristics. WiMAX will reach its peak by making Portable Internet a reality. When WiMAX chipsets are integrated into laptops and other portable devices, it will provide high-speed data services on the move, extending today's limited coverage of public WLAN to metropolitan areas. The combination of these capabilities makes WiMAX attractive for a wide diversity of people and gives an advantage against all other wireless networks.

REFERENCES

Davidson, J. & Peters, J. (2000). *Voice over IP Fundamentals*. Cisco Press, 1.Edition

Eriksson, G., Olin, B., Svanbro, K. & Turina, D. (2000). *The Challenges of Voice-over-IP-over-wireless*. Ericsson Review

Guizani, A., Moshen, M., Rayes, M. & Ammar, H. (2000). *Internet Telephony*. IEEE Communicaitons

Hole, D. & Tobagi, F. (2004). *Capacity Of An IEEE 802.11b Wireless LAN Supporting VoIP*. Department of Electrical Engineering, Stanford University.

Kundaje, A., Bhatia, G., Dalvi, M., Haridas, M. & Nandi, S. (2001). *Voice Over IP*. University of Bombay

Mehta, C. & Udani, S. (2001). *Overview of VoIP "Technical Report"*. University of Pennsylvania

Woodard J., (2002). *Speech Coding*. University of Southampton, Department of Electronics and Computer Science.

Schulzrinne, H. (2000). *IP Networks*. Retrieved October 26, 2003, from http://www.cs.columbia.edu/~hgs/teaching/ais/slides/videobook.pdf

Speech Coding. (n.d.). Retrieved December 11, 2004, from http://www-mobile.ecs.soton.ac.uk/speech_codecs