LAPPEENRANTA UNIVERSITY OF TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY

MASTER'S THESIS

THE ARCHITECTURE OF A FOURTH GENERATION MOBILE NETWORK

The council of the Department of Information Technology approved the subject of the thesis on March 21, 2001.

Supervisor and instructor: Professor Olli Martikainen

Lappeenranta, June 3, 2001

Marko Myllynen
Latvapolku 5
FIN-48400 Kotka
Finland

ABSTRACT

Author: Myllynen, Marko

Subject: **The Architecture of a Fourth Generation Mobile Network**

Department: Information technology

Year: 2001

Place: Lappeenranta

Master’s thesis. Lappeenranta University of Technology. 59 pages and 11 figures.

Supervisor: Professor Olli Martikainen

Keywords: 4G, mobile, network, Internet

Fourth generation mobile networks seamlessly combine telecommunication networks, the Internet and their services. Initially, the Internet has been accessed only from stationary computers while traditional telecommunication networks provided telephone and data services. The users of fourth generation mobile networks are able to use both Internet based services and those of traditional telecommunication networks even while roaming.

This thesis presents an overall architecture of a fourth generation mobile network. The basic components of the architecture are described and the architecture is compared to second and third generation mobile networks. Relevant Internet standards are introduced and their applicability to mobile networks is discussed. Wireless short-range, high-speed network access technologies are presented. Terminal and personal mobility management methods needed in fourth generation mobile network are introduced.

The presented architecture is based on wireless short-range, high-speed network access technologies and Internet standards. The architecture enables connections to other users without knowledge about their current terminal or location. The services of the Internet can be used anywhere in the fourth generation mobile network area. A general purpose mobility management method for application within a single network area is proposed. This method can be used alongside with the presented architecture.

TIIVISTELMÄ

Tekijä: Myllynen, Marko

Nimi: **Neljännen sukupolven mobiiliverkon arkkitehtuuri**

Osasto: Tietotekniikan osasto

Vuosi: 2001

Paikka: Lappeenranta

Diplomityö. Lappeenrannan teknillinen korkeakoulu. 59 sivua ja 11 kuvaa.

Tarkastaja: Professori Olli Martikainen

Hakusanat: 4G, mobiili, verkko, Internet

Neljännen sukupolven mobiiliverkot yhdistävät saumattomasti televerkot, Internetin ja niiden palvelut. Alkuperin Internetiä käytettiin vain paikallaan pysyviltä tietokoneilta perinteisten televerkkojen tarjotessa puhelin- ja datapalveluita. Neljännen sukupolven mobiiliverkkojen käyttäjät voivat käyttää sekä Internetiin perustuvia että perinteisten televerkkojen palveluita liikkuessaankin.

Tämä diplomityö esittelee neljännen sukupolven mobiiliverkon yleisen arkkitehtuurin. Arkkitehtuurin perusosat kuvaillaan ja arkkitehtuuria verrataan toisen ja kolmannen sukupolven mobiiliverkkoihin. Aiheeseen liittyvät Internet-standardit esitellään ja niiden soveltuvuutta mobiiliverkkoihin pohditaan. Langattomia, lyhyen kantaman nopeita liitäntäverkkotekniikoita esitellään. Neljännen sukupolven mobiiliverkoissa tarvittavia päätelaitteiden ja käyttäjien liikkuvuuden hallintamenetelmiä esitellään.

Esitelty arkkitehtuuri perustuu langattomiin, lyhyen kantaman nopeisiin liitäntäverkkotekniikoihin ja Internet-standardeihin. Arkkitehtuuri mahdollistaa yhteydet toisiin käyttäjiin ilman tietoa heidän senhetkisestä päätelaitteesta tai sijainnista. Internetin palveluita voidaan käyttää missä tahansa neljännen sukupolven mobiiliverkon alueella. Yleiskäyttöistä liikkuvuuden hallintamenetelmää yhden verkon alueelle ehdotetaan. Menetelmää voidaan käyttää yhdessä esitellyn arkkitehtuurin kanssa.

PREFACE

This thesis has been written for the Laboratory of Telecommunications in the Department of Information Technology of Lappeenranta University of Technology. The thesis is part of the “4G” project realized in the Laboratory of Telecommunications, beginning in late 2000 and continuing until 2002.

I would like to thank my supervisor and instructor, Professor Olli Martikainen, for his valuable advices and suggestions, which helped me to write this thesis. His visions and commitment made this thesis more mature than it would have ever been without his help.

My colleagues Kalle Ikkelä, Ossi Kauranen, Jari Kellokoski, Sami Lindström, and Mika Yrjölä have created a pleasant and inspirational working atmosphere. I am grateful for them also for their comments and suggestions.

Special thanks go to all my friends and my family for all their support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

2G	Second Generation
3G	Third Generation
4G	Fourth Generation
4GLA	4G Location Area
4GSA	4G Service Area
ACL	Asynchronous Connectionless
AH	Authentication Header
AMPS	Advanced Mobile Phone Service
AN	Access Network
AP	Access Point
ARP	Address Resolution Protocol
ARPANET	Advanced Research Projects Agency Network
ARQ	Automatic Repeat Request
AuC	Authentication Center
BS	Base Station
BSC	Base Station Controller
BSS	Base Station System
BTS	Base Transceiver Station
BU	Binding Update
CIP	Cellular IP
CN	Core Network

CN	Correspondent Node
CRC	Cyclic Redundancy Check
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
DARPA	Defense Advanced Research Projects Agency
DHCP	Dynamic Host Configuration Protocol
DHCPv6	Dynamic Host Configuration Protocol for IPv6
DNS	Domain Name System
DRCP	Dynamic Registration and Configuration Protocol
DSSS	Direct Sequence Spread Spectrum
EDGE	Enhanced Data rates for GSM Evolution
EIR	Equipment Identification Register
ESP	Encapsulating Security Payload
ETSI	European Telecommunications Standards Institute
FA	Foreign Agent
FEC	Forward Error Correction
FHSS	Frequency Hopping Spread Spectrum
FTP	File Transfer Protocol
GFA	Gateway Foreign Agent
GGSN	Gateway GPRS Support Node
GHz	gigahertz
GMSC	Gateway MSC
GPRS	General Packet Radio Service

GSM	Global System for Mobile communications
HA	Home Agent
HAWAII	Handoff-Aware Wireless Access Internet Infrastructure
HIPERLAN	High Performance Radio Local Area Network
HLR	Home Location Register
HMIP	Hierarchical Mobile IP
HomeRF	Home Radio Frequency
HSCSD	High-Speed Circuit Switched Data
HTTP	HyperText Transfer Protocol
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IHL	IP Header Length
IP	Internet Protocol
IPng	Internet Protocol next generation
IPsec	IP security
IPv4	Internet Protocol version 4
IPv6	Internet Protocol version 6
ISDN	Integrated Services Digital Network
ISM	Industrial, Scientific, and Medical
ISO	International Organization for Standardization
IWU	Inter Working Unit
Kbps	kilobits per second

LA	Location Area
LAN	Local Area Network
LLC	Logical Link Control
MAC	Medium Access Control
Mbps	megabits per second
MHz	megahertz
MIP	Mobile IP
MIPv6	Mobile IPv6
MN	Mobile Node
MS	Mobile Station
MSC	Mobile services Switching Center
MTU	Maximum Transmission Unit
NCP	Network Control Protocol
NMT	Nordic Mobile Telephone
NSS	Network Subsystem
OFDM	Orthogonal Frequency Division Multiplexing
OSI	Open Systems Interconnection
PDA	Personal Digital Assistant
PGP	Pretty Good Privacy
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RFA	Regional Foreign Agent

RFC	Request For Comments
RM	Reference Model
RTP	Real-Time Transport Protocol
SA	Service Area
SCO	Synchronous Connection Oriented
SDP	Session Description Protocol
SGSN	Service GPRS Support Node
SIG	Special Interest Group
SIM	Subscriber Identity Module
SIP	Session Initiation Protocol
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
URL	Uniform Resource Locator
UTRAN	UMTS Terrestrial Radio Access Network
VLR	Visitor Location Register
VoIP	Voice over IP
WECA	Wireless Ethernet Compatibility Alliance
WLAN	Wireless Local Area Network
WWW	World Wide Web

1 INTRODUCTION

Speech, the basic form of human communication, originates from the dawn of mankind. Through biological and cultural evolution speech has become the most opulent way to express oneself. Advancing technology has now offered new opportunities for the usage of speech. Telephone provides the means to speak with one another on the other side of the world. Mobile telephones enable users to roam around while speaking on a phone. Technical devices have evolved but speech itself has not changed or lost its significance. There is no doubt that speech will maintain its status as an essential part of human communication, regardless of technological advances or devices to come.

There are other important communication means as well. It has been long since the first symbols were adopted to represent words and speech in the ancient Middle East. An alphabet consists of letters that singly have no meaning but grouped together may represent every word a man may breathe. A writing may reach thousands of readers even after time has passed the writer. Even with the power of the written word, one image still may say more than words ever could. Technology has allowed us to joint images into sequences, forming lively reflections of the past or worlds of imagination. The fashionable term multimedia means text, still images, audio, video and data digitally combined together.

Computers were initially not used for communication. However, during the last decade of the 20th century, a certain computer network turned into a worldwide platform for communication, a wide range of digital services and multimedia content. The network now consists of thousands of smaller, independent networks. There are millions of computers and devices attached to this worldwide network through these smaller networks around the world. The users of this worldwide network may send electronic mail to their colleagues or friends and can access multimedia content. Companies may advertise their presence to thousands of people almost for free. Individuals may take care of their bills and do their shopping via the network. This worldwide network is known as the Internet.

Traditional telecommunications and content services are separated into different, dedicated networks. Telephone users are dependent upon telephones and telephone networks. Telephones can only be used for receiving or making voice calls. Internet users are able to access the Internet with their computers, but still need to use a telephone to make or receive voice calls. Second and third generation mobile networks provide telecommunications services with limited communications speeds for other content and services. However, Internet based multimedia content and services require high-speed communication channels for a pleasant user experience. By combining dedicated networks and providing truly high-speed access also for mobile users, the number of available services and content would explode. There would be a vast number of new opportunities for people to communicate and to share and access information, services, and entertainment. The combination would be called a fourth generation mobile network.

This thesis presents an overall architecture of a fourth generation mobile network.

2 THE INTERNET

The Internet has rapidly evolved from a small research network into a worldwide communications and information sharing entity. Millions of people use the Internet on a daily basis to communicate with each other and to access information online. Today, the Internet is accessed mostly from stationary computers but in the future mobile terminals will be used widely for the same purpose. This will increase the usage of the Internet but will also raise challenges for engineers working with new Internet and mobility management related technologies.

The background and current technologies of the Internet are briefly introduced in this chapter. Some limitations of the technologies are explained and solutions for problems are described in this and the following chapters. Expectations for the future are introduced later in this chapter.

2.1 Background

Packet switching is a connectionless networking method in which the data to be sent is divided into packets. Packets are labeled and sent independently through the network to be reassembled at the destination. The communication channel is not reserved exclusively, it is used only when sending or receiving packets. This allows the same communication channel to be shared among many users in the network. An alternative method to packet switching is circuit switching. Circuit switching is based on an obtained and dedicated connection between two ends. The communications channel is reserved even when data is not being sent.

The first ideas of packet switching theory and computer communications though networking were recorded in the beginning of the 1960s [Lic62, Kle61]. First wide area network, consisting of two long distance computers, was tested in 1965. It proved that time-shared computers could work together, sending and receiving data as necessary. It also proved that the circuit switched telephone system was not adequate for computer communications. The need for packet switching was thus confirmed [Rob66].

These fundamental ideas of the Internet were further developed during the 1960s and 1970s under the Defense Advanced Research Projects Agency (DARPA). DARPA researchers initiated ARPANET (Advanced Research Projects Agency Network) as a four-node research network in 1969. During the following years, hosts were added quickly to the initial ARPANET. To enable communications between different kinds of computer systems and applications, a network protocol (i.e., a set of rules) was needed for ARPANET. In December 1970, Network Control Protocol (NCP), the initial ARPANET host-to-host protocol, specification was finished. As soon as NCP became available, network users could begin developing network programs.

The client/server model has become one of the central ideas of network computing. It describes the relationship between two computer programs in which one program, the client, makes service requests to another program, the server, which fulfills the requests. The model provides a convenient way to interconnect programs that are distributed efficiently across different locations.

As more research networks were built the idea of the Internet was born with multiple independent networks interoperating. Beginning with the original ARPANET, networks were added to the Internet and by 1980, there were almost 20 operational networks on the Internet. As more and more networks and computers were attached to the Internet, it became clear that a new protocol architecture was needed to overcome the limitations of NCP. For example, NCP had no end-to-end host error control mechanism since it was developed for ARPANET which was supposed to be so reliable that no error control would be needed. This was contrary to the Internet where networks may or may not be reliable.

2.2 Layered architecture

The Internet protocol architecture consists of layers. In a layered architecture, each layer has specified functions it is responsible for. Upper layers use the services of lower layers to carry out their own functions. Usually one protocol is used to implement the functionality of one layer. A protocol stack consists of several protocols implementing a layered architecture.

When programs on different computers communicate with each other using some upper layer protocol, the data flow from the upper layer protocol goes through the whole protocol stack to the lowest protocol of the stack on the sender side and vice versa on the receiving side. The protocols involved add their own management data needed to transport the original datagrams, which is removed by the corresponding receiving side protocols. Thus, the receiving protocol receives the datagrams in the form they were sent by the corresponding protocol on the sender side. The receiving protocol does not know how many lower layer protocols were involved in the transport nor how the lower layer protocol received the datagram. The receiving protocol just accepts the datagrams from the protocol below and process them as required. Figure 1 presents the protocols used in the Internet and their relationships to the well-known Open System Interconnection (OSI) Reference Model (RM) specified by ISO (International Organization for Standardization).

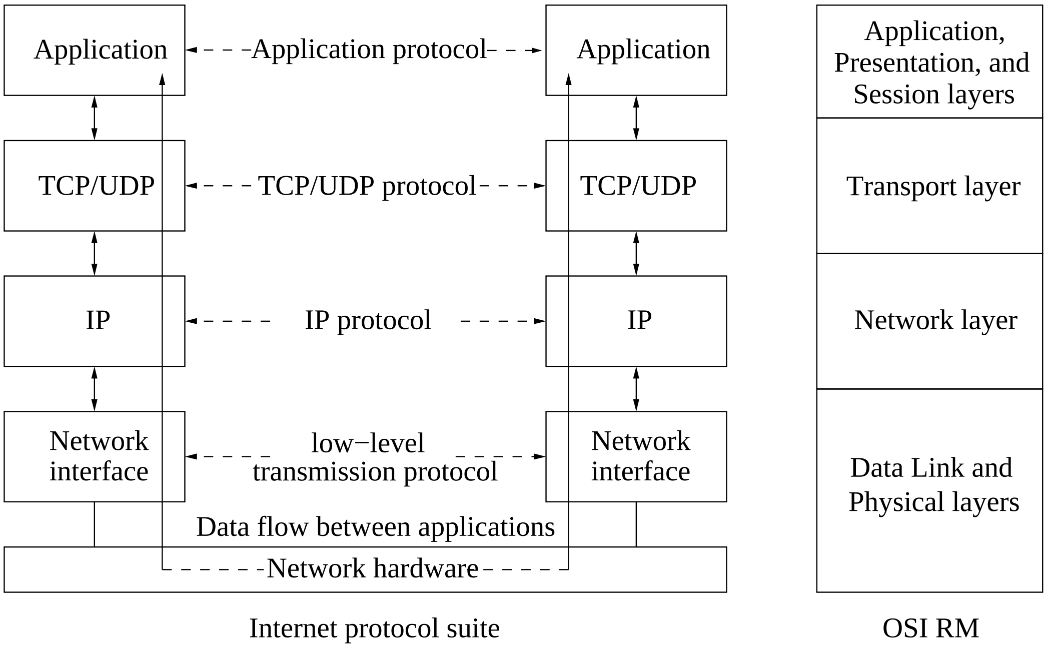


Figure 1: The protocols of the Internet and OSI RM

2.3 Physical and Data Link Layers

The physical and data link layers of OSI Reference Model are responsible for transmitting data over a communications channel. The physical layer defines the physical and electrical characteristics of a network. It is concerned with transmitting raw bits, ones and zeroes, over a communications channel using network hardware. The communications channel may be wired or wireless, reliable or unreliable.

The data link layer controls transmissions over the communications channel. It transforms a connection channel into a link which appears free of transmission errors to the network layer. Usually this is accomplished by having the sender break the input data up into data frames, which are then transmitted sequentially. The receiver sends acknowledgment frames when receiving frames so the sender can retransmit frames if not acknowledged by the receiver.

Traditional network technologies used to implement layer 1 and 2 functionality of the OSI Reference Model are, for example, Ethernet, ISDN (Integrated Services Digital Network), and Token Ring. Today, wireless technologies, such as Wireless Local Area Network (WLAN) and Bluetooth, which are presented later in this paper, are also used.

2.4 Internet Protocol (IP)

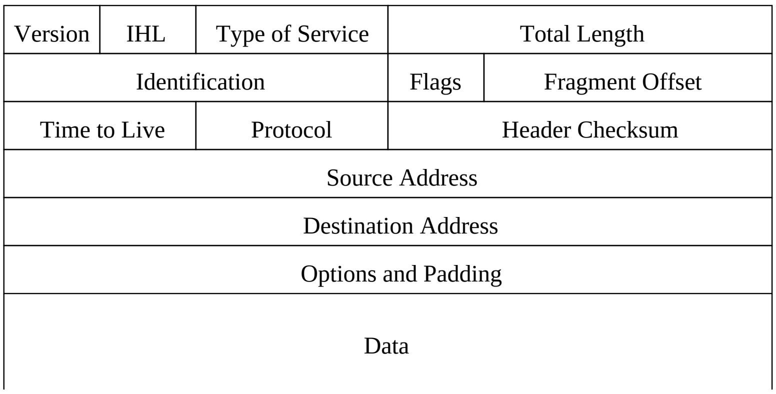
The Internet Protocol [RFC791] corresponds to layer 3, the network layer, of the OSI Reference Model. The Internet Protocol is responsible for routing packets from a source to a destination in the Internet. IP does not specify the physical details of a transport medium; this is done by a lower layer protocol. Because IP is able to utilize different kinds of lower layer protocols, packets may travel through different kinds of networks during transportation. Different networks of the Internet are connected to other networks by gateways, which are also known as routers. Gateways are, in short, computers which pass packets between networks.

A unique 32-bit identifier, the IP address, identifies all nodes attached to the Internet. The 32-bit IP address can be presented in symbolic or numeric form. The numeric form (for example, *216.239.37.100*) is used in routing and other network operations, the corresponding symbolic form (www.google.com) is for user convenience only. The forms are mapped between each other as needed by the Domain Name System (DNS).

ARP (Address Resolution Protocol) is a protocol for mapping network layer addresses to physical link layer (hardware) addresses. Before sending an IP packet to another host in the network, a host must find the hardware address of the recipient. This is done by broadcasting an ARP request to all hosts in the network. The ARP request contains the sender's IP and hardware addresses and the recipient's IP address. The recipient recognizes its own IP address in the ARP packet and replies to the sender with an ARP reply containing its hardware address. The sender of the ARP request has now both the IP and hardware address of the recipient and is able to send IP packets. ARP has different specifications for different link types. For example, ARP for Ethernet is specified in [RFC826].

An IP address specifies a network and a node's interface in the network. IP packets in the Internet are routed based on the network identifier part of the IP address (netid). As an IP packet arrives to a destination network, a local router sends the packet to the receiver, determined by the host identifier part of the IP address (hostid). Netid identifies just a network in the Internet; hostid identifies a single interface in that particular network. All IP packets are routed independently, so IP packets may travel through different routes when exchanged with hosts. Since all networks of the Internet are treated equally, they all appear to be part of the same global network.

An IP packet contains the actual data to be transmitted from upper layer protocols and also source and destination addresses as well as other information needed to route the packet from a source to a destination. This routing and data processing related information is stored in the IP header. The format of an IP packet is presented in Figure 2.



<table>
<tr><td>Version</td><td>IHL</td><td colspan="2">Type of Service</td><td colspan="4">Total Length</td></tr>
<tr><td colspan="4">Identification</td><td>Flags</td><td colspan="3">Fragment Offset</td></tr>
<tr><td colspan="2">Time to Live</td><td colspan="2">Protocol</td><td colspan="4">Header Checksum</td></tr>
<tr><td colspan="8">Source Address</td></tr>
<tr><td colspan="8">Destination Address</td></tr>
<tr><td colspan="8">Options and Padding</td></tr>
<tr><td colspan="8">Data</td></tr>
</table>

Figure 2: The format of an IP packet

The Internet Protocol also provides fragmentation and reassembly of the packets and error reporting. Since IP uses a lower-layer transmission protocol for transmitting the data, it is tied to the limitations of the lower-layer protocol. Usually, those protocols have a limitation on the maximum transmission unit (MTU), which is the maximum size of data that can be sent over a network link at a time. Thus, routers must fragment IP packets that are larger than the MTU before sending them over a link. Fragmented packets must be reassembled in the destination before passing them to an upper layer protocol.

IP does not guarantee that packets reach the destination. Packets may be lost, delayed, delivered non-consecutively (in an order other than that in which they were sent), or damaged in transmission. It is up to upper layer protocols to cope with these problems.

2.5 Upper Layer Protocols

The two most used transport layer protocols in the Internet are the Transmission Control Protocol [RFC793] and the User Datagram Protocol [RFC768]. Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) receive data for delivery from an application layer protocol and use IP to actually transmit the data from source to destination.

TCP divides data from an application layer protocol into datagrams, labels the datagrams and passes them individually to IP, which will route them to the destination. On the receiving side, the TCP module reassembles the data to the original form and passes it to the application layer protocol. Common application layer protocols that use TCP are the HyperText Transfer Protocol (HTTP) and the File Transfer Protocol (FTP). TCP is the most widely used transport layer protocol, thus the acronym TCP/IP is often used.

TCP guarantees that data will be delivered intact. This is achieved by acknowledging received datagrams. The sender side TCP module will resend datagrams that are not acknowledged in time. The receiving side is able to reconstruct the data to the original form by using labels in the datagrams. TCP is a connection-oriented protocol, which means that a connection is first established by a three-way handshake mechanism, then the data is transmitted, and in the end the connection is terminated by closing mechanisms.

UDP offers a limited amount of services compared to TCP. UDP does not divide data into datagrams or guarantee sequencing of the datagrams on the receiving side. Application layer protocols using UDP must be able to deal with the limitations of UDP. Typically, protocols using UDP send only data in which its disappearance would not be catastrophic for an application.

2.6 Internet Protocol version 6 (IPv6)

The Internet Protocol, also known as Internet Protocol version 4 (IPv4), was designed around 1980. At that time, computers were rare and the Internet was seen mainly as a research network. Under those circumstances, IPv4 worked well. However, during the 1980s and 1990s, as more and more computers and networks were connected to the Internet, problems arose. The TCP/IP architecture in general had proven to be effective; the problems were mostly related to addressing and routing. Thus, it became clear that a new network layer protocol was needed. For this need, the Internet Protocol version 6 was developed.

IPv6 [RFC2460], also known as Internet Protocol next generation (IPng), offers several enhancements compared to IPv4. Most notably, the size of IPv6 addresses is 128 bits. The address size expansion was needed due to fact that IPv4 32-bit addresses are expected to run out around 2005-2010. In contrast to the limited IPv4 address space, the IPv6 address space is vast; 2^{128} is nearly 10^{39}. Even if IPv6 addresses were assigned very inefficiently, still hundreds of addresses would be available for every square meter of the planet Earth's surface [RFC1715]. To simplify the transition from IPv4 to IPv6, several transition phase technologies have been specified, including a method to present IPv4 addresses as IPv6 addresses [RFC2373]. The transition from IPv4 to IPv6 will not happen in one day, it will probably take years. During the transition phase, both IPv4 and IPv6 networks co-exist.

IPv6 supports the automatic configuration of network addresses of devices newly attached to a network [RFC2462]. This will make it easier to attach mobile terminals and new equipment to IPv6 based networks. The usage of autoconfiguration also means that numeric forms of IPv6 addresses (e.g., *FEDC:BA98:7654:3210:FEDC:BA98:7654:3210*) are likely to change more often than those of IPv4 addresses. Therefore, it is recommended to use symbolic forms of IPv6 addresses, which are similar to the symbolic forms of IPv4 addresses.

The main function of IP (both IPv4 and IPv6), routing, requires that routers maintain routing tables. A router investigates its routing table for every IP packet the router receives to find out where to send the IP packet next. IPv6 provides more efficient routing compared to IPv4 in different ways. With the IPv6 addressing architecture [RFC2373], it is possible to reduce the size of current IPv4 routing tables. Smaller routing tables combined with other changes in IPv6, like simplified headers (presented in Figure 3) and MTU discovery, IPv6 routers will be able to operate more efficiently than present IPv4 routers. MTU discovery is a method in which intermediate routers will not fragment packets which are too large. Instead, these packets are discarded and an error message is returned to the sender. Thus, the sender is able to determine the total path MTU and does all fragmentation initially. All packets are therefore processed through the network without delay at routers, without the overhead of intermediate fragmentation. As well as MTU discovery, simplified headers also reduce overhead at intermediate routers. [Hui98]

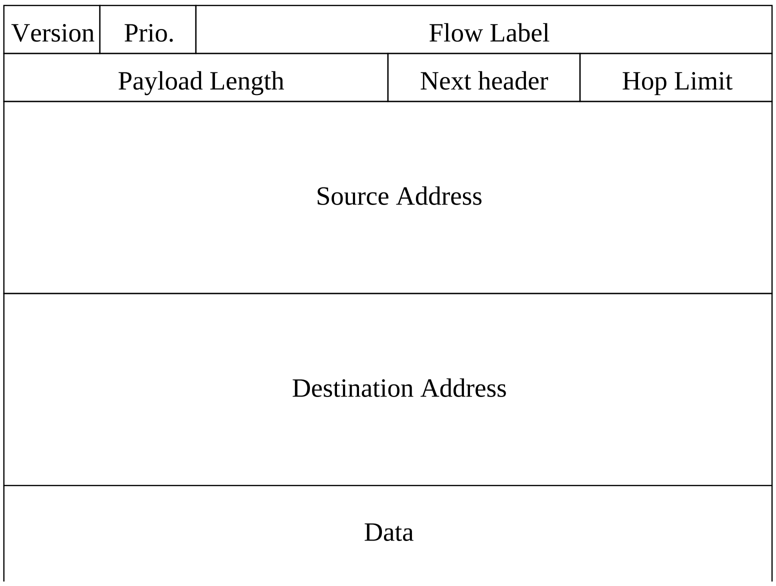


Figure 3: The format of an IPv6 packet

Other new features of IPv6 include support for IPsec [RFC2401, RFC2402, RFC2406] and Quality of Service (QoS). With an IP Authentication Header [RFC2402] it is possible to verify that a packet really came from the origin indicated by the source address field. The IP Encapsulating Security Payload [RFC2406] is a mechanism to provide for encrypted information transmission. While an Authentication Header (AH) prevents third parties from introducing spurious information into the Internet, the Encapsulating Security Payload (ESP) header prevents third parties from extracting information from the Internet that they are not entitled to have. IPsec (IP security) features are especially important in public wireless environments, where networks could easily be used by malicious users.

Quality of Service is a concept that guarantees a certain level of transmission rates, error rates, and other characteristics in advance. The current IPv4 based Internet does not provide any guarantees; all routing is based on best-effort principle. QoS features are needed by, for example, real-time interactive multimedia and high-rate telemetry applications.

2.7 The Future of the Internet

As discussed in the previous sections, the initial Internet was built as a research network at a time when computers were rare. Nowadays, most companies and many individuals are connected to the Internet in developed countries. The shift from a small research network to a global one has happened in just a few decades. Nevertheless, this is just the beginning.

Computers used until the mid 1990s were simply too heavy and large to be moved easily from one place to another. However, as technology has evolved, mobile computers have become convenient to use and mobile computing is gaining more and more popularity. The benefits of mobile computing are many. One mobile computer can be used at home, at office, and elsewhere. Users may roam around using their computers at the same time and, in the future, mobile terminals can be connected to the Internet wirelessly while roaming. Advancing technology enables connecting new kinds of devices other than computers to the Internet. Toasters, fridges and traffic lights have already been connected to the Internet for testing purposes.

Traditional telecommunications and content services are vertically integrated. Each service depends on a dedicated network and corresponding terminals. Examples of such vertical services are fixed telephone services, traditional data services and mobile phone services. The Internet changes this vertical structure to a horizontal one: all terminals and services will be Internet compatible. Instead of vertical service "pipes" there will be a horizontal structure of services, network, and access, illustrated in Figure 4. [Mar00]

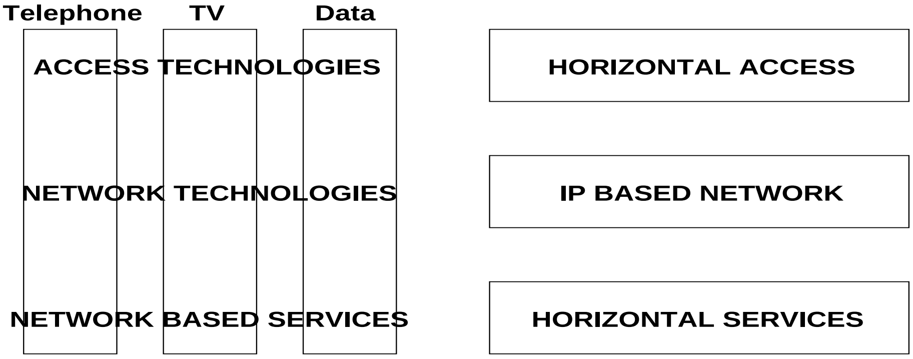


Figure 4: Vertical and horizontal service structure

The horizontal structure will change terminals, services and the way services are managed. The horizontal structure will allow different combinations of service functionalities in terminal equipment. Horizontal networks will not only make existing services easier and more widely applicable, but also create a platform for the integration of various new services and applications into the same terminals [Mar00]. The common factor for new services and terminals is that they all use IP as the basis of their communications.

The development of new services, terminals, applications and content combined with increased mobility will lead to the need for new techniques to provide means for high-speed wireless connections and mobility management in the Internet. For technically advanced mobile terminals, new kinds of networks will be built with the help of high-speed wireless technologies. Several promising high-speed wireless technologies are already in use or under development. Emerging technologies include HIPERLAN (High Performance Radio Local Area Network), HomeRF (Home Radio Frequency) and OFDM (Orthogonal Frequency Division Multiplexing) based systems. In the next chapter, two high-speed wireless technologies already available for consumers, the Wireless Local Area Network and Bluetooth, are introduced. Mobility management protocols needed in future IP based mobile networks are presented in chapter 4.

3 WIRELESS COMMUNICATIONS

The first generation of mobile phone networks were adopted during the 1980s. In Europe NMT (Nordic Mobile Telephone) and in the United States AMPS (Advanced Mobile Phone Service) systems were mainly used. Both systems were analog and offered low data transfer rates.

Due to the popularity and limitations of first generation mobile networks, second generation (2G) mobile networks were developed. During the 1990s, GSM (Global System for Mobile communications) became the de facto mobile phone standard in Europe. GSM [Mou92] is a digital, circuit switched technology that offers 9.6 Kbps and 14.4 Kbps data transfer rates. Clearly, these are not enough for high-speed communications, but the GSM network architecture has been proven effective and capable to serve users en masse.

3.1 The Architecture of the GSM Network

An overview of the GSM network architecture is presented in Figure 5. The GSM network is composed of several functional entities, whose functions and interfaces are defined in the GSM standards specified by the European Telecommunications Standards Institute (ETSI). The GSM network can be divided into three main functional parts, which are Mobile Station, Base Station System and Network Subsystem.

Mobile Station (MS) is the user terminal. The terminal consists of a radio transceiver, signal processors, display, and a Subscriber Identity Module (SIM) card. One SIM card can be used with different terminals. The SIM card is used to identify the user and enabling the routing of incoming calls to the user's current terminal.

Base Station System (BSS) consists of one or more Base Transceiver Stations (BTS) and one Base Station Controller (BSC). Radio transceivers are in BTSs and a BTS manages the radio link protocols with an MS. The BSC manages radio resources such as frequency hopping and handovers for all BTSs in the BSS area. The BSC is the connection between an MS and the Mobile service Switching Center.

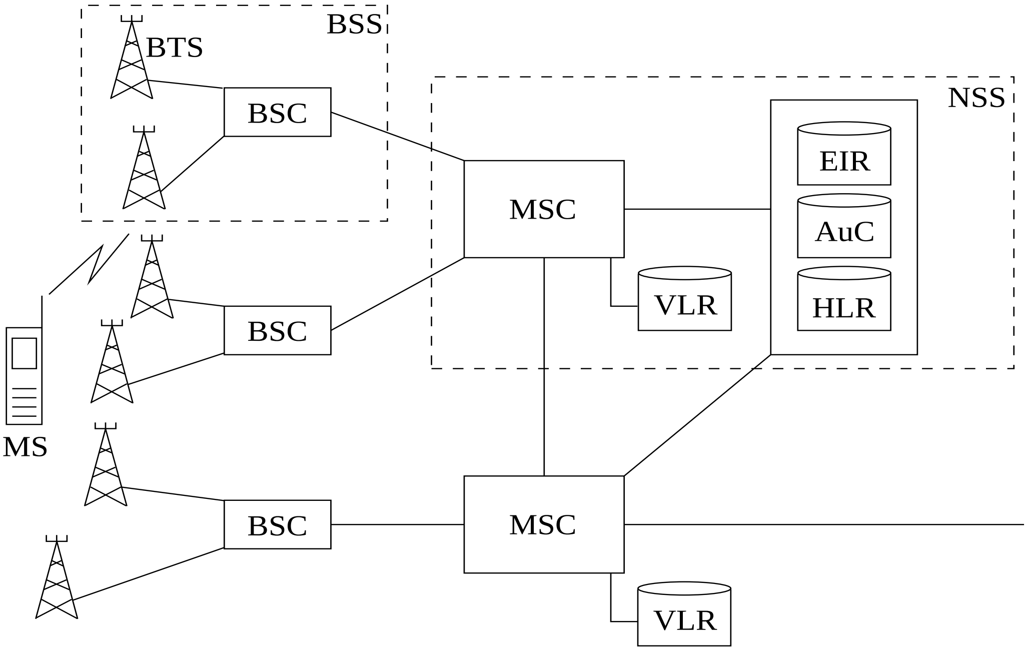


Figure 5: The GSM network architecture

Network Subsystem (NSS) is the third part of the GSM network. It includes databases for authentication, registration and call management. It also provides connections to external networks, like the traditional Public Switched Telephone Network (PSTN) and the Internet. The relevant components of the NSS are:

Mobile services Switching Center (MSC) is the main component the NSS. It coordinates the setting up of calls to and from the Mobile Stations. The MSC has no information about particular Mobile Stations; this information is stored in the location registers.

Home Location Register (HLR) is the database in charge of the management of mobile subscribers. The HLR contains subscriber information and subscriber location information. When a call is set up, a query is sent to the target MS's HLR which responds with the MS's current location information enabling routing of the call toward the MSC area where the target MS currently is.

Visitor Location Register (VLR) controls all Mobile Stations located in the MSC area it is in charge of. When a new MS arrives to the MSC area, the VLR and the MS's HLR exchange information to allow the proper handling of calls involving the MS.

The GSM network can also be divided into several operational areas. For reference, relevant areas are described.

Location Area (LA) is an area in which a Mobile Station may move freely without updating location registers. A Location Area may consist of several cells. A cell is an area covered by a BTS's radio transceiver (or one of the transceivers, if many). A Location Area is the area associated with one VLR.

Service Area (SA) is an area in which a Mobile Station is obtainable by a GSM or PSTN subscriber without the subscriber's knowledge of the actual location of the Mobile Station within the Service Area. A Service Area consists of several Location Areas and could be the size of about one country.

System area consists of one or more Service Areas with fully compatible MS-BSS interfaces. One system area is typically owned by one network operator. The Mobile Station's HLR knows the location of the MS regardless of the operator it is using.

Roaming and Handover

Roaming and handover are essential terms when discussing mobility management. The distinction of the terms varies from one source to another, so, for clearness, the meaning of the terms are explained.

Roaming is the movement of a mobile terminal from one part of the network area to another part while retaining the capability of making or receiving calls. In the GSM network, roaming is movement from one location area to another.

Handover is the action of switching a call in progress from one cell to another (or between radio channels in the same cell). Handover is used to allow established calls to continue when mobile terminals move from one cell to another.

In the GSM network, handover can be carried out in several ways:

Intracell handover: A Mobile Station is switched from one radio channel to another within the cell area.

BTS-BTS handover: A Mobile Station is switched from one Base Transceiver Station to another under the control of the same Base Station Controller.

BSC-BSC handover: A Mobile Station is switched from one BSC to another and from one BTS to another, at the same time. The target BSC controls the handover.

MSC-MSC handover: A Mobile Station is switched from one Mobile services Switching Center to another. The target MSC controls the handover.

3.2 Evolution toward Third Generation Mobile Networks

The original GSM is a circuit switched network technology. General Packet Radio Service (GPRS) is a GSM based packet switched technology. GPRS [0260, 0360], sometimes referred as a phase 2+ technology, adds new functional entities to the GSM network architecture and enhances existing components. GPRS offers data transfer rates up to 172 Kbps. Other GSM based phase 2+ technologies include EDGE (Enhanced Data rates for GSM Evolution) and HSCSD (High-Speed Circuit Switched Data), which promise data transfer rates up to 384 Kbps and 57 Kbps, respectively.

UMTS (Universal Mobile Telecommunications System) is a third generation (3G) broadband, packet switching based radio technology based on GSM and GPRS. UMTS [21101] offers data transfer rates up to and possibly higher than 2 Mbps. In order to access GPRS or UMTS networks a GPRS or UMTS capable user terminal is required. Multi-mode user terminals may support different technologies enabling roaming between different kinds of networks. Users can be charged in packet switched networks by the amount of data transferred, not by the length of a connection.

The evolution from GSM toward UMTS will happen gradually. First, only urban areas are upgraded to GPRS and UMTS networks, and only later other areas follow. The UMTS network architecture following an evolutionary path from GSM toward UMTS is illustrated in Figure 6.

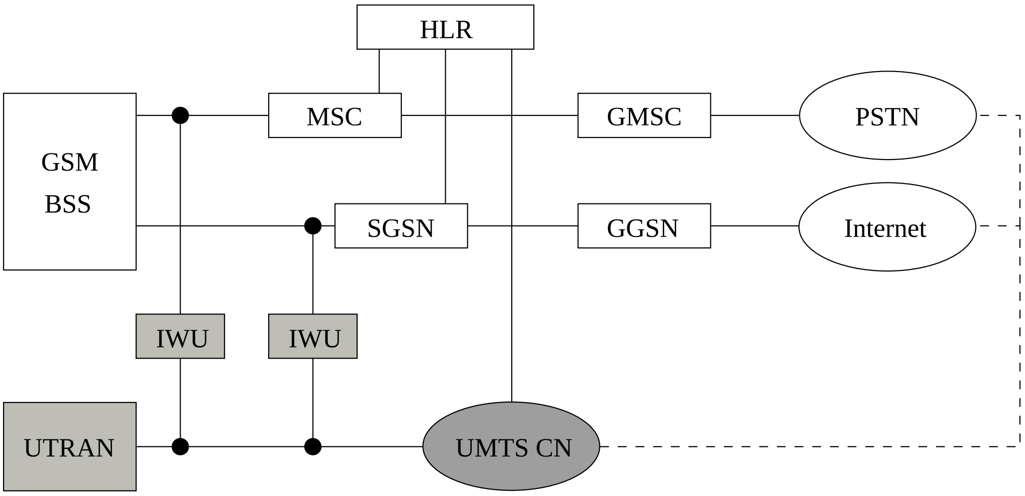


Figure 6: The evolution from GSM toward UMTS

The UMTS network architecture is divided into two separate parts: the Access Network (AN) part and the Core Network (CN) part connected to each other via an IWU (Inter Working Unit). In the beginning of the UMTS era, it is likely that the UMTS Access Network, i.e., UMTS Terrestrial Radio Access Network (UTRAN), will be interconnected with the GSM phase 2+ NSS functioning as the Core Network. The GSM NSS in phase 2+ will be capable of handling both the conventional circuit switched transmission already introduced in the original GSM and the packet switched transmission provided by GPRS. The circuit switched transmission path between the GSM BSS and external networks is routed through the GSM network via the GSM MSC and the UMTS component GMSC (Gateway MSC). The packet switched transmission is routed via the GPRS components SGSN (Serving GPRS Support Node) and GGSN (Gateway GPRS Support Node). UTRAN will be interconnected to this core network via two IWUs. This architecture makes it possible for GSM, GPRS, and UMTS customers to be connected both to circuit switched networks (e.g., PSTN) and packet switched networks (e.g., the Internet).

The architecture of a fourth generation mobile network is analogous to second and third generation networks in personal mobility management. Network technologies and components differ.

3.3 Wireless Local Area Network (WLAN)

A local area network (LAN) is a group of computers and associated devices that share a common communications link within a small geographic area. A LAN typically consists of one or a few office buildings. A wireless local area network is a LAN to which devices can be attached through a wireless connection. In a WLAN, either all devices or only part of them are wireless. If all devices are wireless, the network is called an ad hoc network. IEEE (Institute of Electrical and Electronics Engineers) standard 802.11 [80211] specifies wireless data transmission methods for wireless local area networks. WLAN is often used as a synonym for the 802.11 standard.

The original 802.11 was approved in 1997 and in 2001, WLAN products are widely available for consumers. Outside of the standards bodies, wireless industry leaders have united to form the Wireless Ethernet Compatibility Alliance (WECA). WECA's mission is to certify cross-vendor interoperability and compatibility of the 802.11 standard based wireless networking products. The members of WECA include such companies as 3Com, Apple, Cisco, IBM, Lucent, and Nokia.

IEEE standard 802.11 defines two pieces of equipment: a wireless station and an access point (AP). Typical wireless stations are mobile terminals and laptop computers. Access points include both wireless and wired network interfaces acting as bridges between wireless and wired networks.

Connection speeds of 1 Mbps and 2 Mbps are defined in the original version of the 802.11 standard. Amendment 802.11b (also known as 801.11 High Rate) to the standard provides for connection speeds of 5.5 Mbps and 11 Mbps. The recent 802.11g and 802.11a amendments increase connection speeds up to 22 Mbps and 54 Mbps, respectively. Speeds of several megabits per second are suitable for fourth generation mobile networks.

The 802.11 standard does not specify technology or implementation details but simply specifications for the physical layer, Medium Access Control (MAC) layer, and security functions. This is enough to allow manufacturers of WLAN equipment to build interoperable network hardware. Most WLANs provide interconnection with wired networks such as Ethernet.

Specifications for infrared connections and two types of radio connections are defined in the standard. An infrared connection requires that devices communicating share a line of sight. This kind of requirement limits mobility, thus 802.11 based infrared devices are not widespread.

The radio connections specified by the 802.11, 802.11b and 802.11g standards use the unlicensed, globally available 2.4 GHz ISM (Industrial, Scientific, and Medical) radio frequency band. The newer 802.11a standard uses the 5.8 GHz ISM band. The range of the radio connection can vary from a few meters up to 100 meters, depending on the power of the radio antenna, receiver design and the propagation path. Like with any other radio system, typical building objects like walls and metal disturb the connection and reduce the range.

The 802.11 standards focus on the two bottom layers of the OSI RM, the physical and data link layers. Figure 7 illustrates the relationship between 802.11 and the OSI RM. The 802.2 Logical Link Control (LLC) sublayer uses the services of the Medium Access Control sublayer and provides medium independent data link functions for network layer protocols, like IP. The 802.2 LLC is also used by other IEEE network standards, allowing for very simple bridging from WLAN to wired networks following IEEE standards.

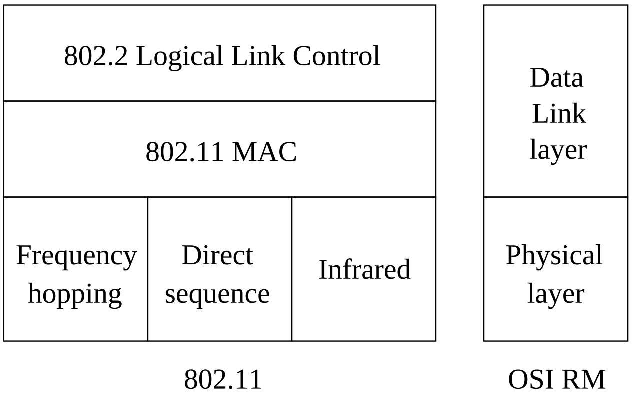


Figure 7: IEEE 802.11 and OSI RM

The 802.11 MAC sublayer provides functions for media access and error checking. The 802.11 standard specifies that Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) should be used as the method for transmitting information in a WLAN. The method provides the means to sense whether the transport medium is currently busy with another transmission. If the medium is not busy, data can be sent. If two or more stations sense a quiet network and start send data simultaneously, collisions will occur and the data will not reach its destination. For this reason, the standard requires that the receiving station dispatches an acknowledgment to inform the sending station that a collision did not occur. If the sending station did not receive an acknowledgment, it will assume that the original packet did not arrive and will resend until an acknowledgment is received.

Direct Sequence Spread Spectrum (DSSS) and Frequency Hopping Spread Spectrum (FHSS) are the radio connection methods defined by the 802.11, 802.11b, and 802.11g standards. The methods are fundamentally different ratio mechanisms and will not interoperate with one another. Spread spectrum techniques are used to enable many unrelated products to share the spectrum without explicit cooperation.

In the case of DSSS, the transmission signal is simultaneously spread over a wide range of the radio spectrum. Only a small fragment of the data is sent in any one frequency. DSSS is the method adopted by the majority of WLAN vendors. FHSS based devices transmit a small fragment of data in one frequency and then hop to another to send the next fragment.

OFDM (Orthogonal Frequency Division Multiplexing) is used by the recent 802.11a standard. OFDM is a multicarrier modulation scheme that sends a high-speed data signal interleaved into parallel bit streams. Bandwidth is used efficiently enabling very high data transfer rates. OFDM is the technology which is considered to be the cornerstone of the next generation of high-speed wireless data products.

3.4 Bluetooth

Bluetooth is intended to be a robust, low-cost, low-power, short-range radio link. It was originally developed as a cable replacement and supports both voice and data, providing standardized wireless communications between any electrical devices. Bluetooth can be used instead of 802.11 in WLANs. The Bluetooth specification consists of two documents; the foundation Core [Blu01a], which provides design and security specifications and the foundation Profile [Blu01b], which provides interoperability guidelines.

The first Bluetooth devices became available for consumers during spring 2001 and the final Bluetooth breakthrough is expected to happen in 2002 or in 2003. Bluetooth is developed by the Bluetooth Special Interest Group (SIG) whose core members are 3Com, Ericsson, IBM, Intel, Lucent, Microsoft, Motorola, Nokia and Toshiba. By April 2001 the Bluetooth SIG included more than 2000 member companies.

The Bluetooth radio operates in the 2.4 GHz ISM band. Bluetooth radio transmission uses FHSS combined with ARQ (Automatic Repeat Request), CRC (Cyclic Redundancy Check) and FEC (Forward Error Correction) to achieve appropriate reliability on the wireless link. Radio range is from around 10 meters up to 100 meters, depending on the power of the transmitter and the propagation path. The usage of lower ranges reduces power consumption.

The Bluetooth system provides a point-to-point connection when only two Bluetooth units are involved, otherwise a point-to-multipoint connection is used. Two or more units sharing the same channel form a piconet. In a piconet, one Bluetooth unit acts a master, others act as slaves. Up to seven slaves can be active and many more can remain locked to the master in a parked state. Slaves can participate in different piconets on a time division multiplex basis. In addition, a master in one piconet can be a slave in another piconet. Multiple piconets with overlapping coverage form a scatternet. For example, computers can act as masters and peripherals as slaves.

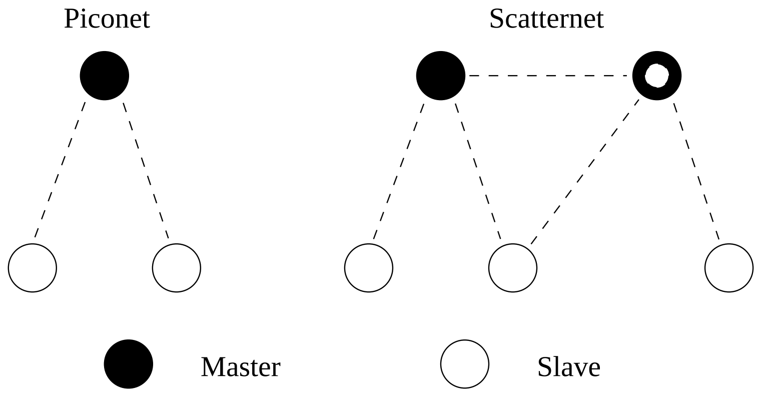


Figure 8: A piconet and a scatternet

Bluetooth uses a combination of circuit and packet switching. Bluetooth can support an Asynchronous Connectionless (ACL) link for data and up to three simultaneous Synchronous Connection Oriented (SCO) links for voice, or a channel that simultaneously supports asynchronous data and synchronous voice. Each voice channel supports a 64 Kbps synchronous channel in each direction. The asynchronous channel can support a maximum data transfer rate of 723 Kbps to one direction and 57 Kbps to the other. 433 Kbps symmetric links are also possible. The next version of Bluetooth is expected to support connection speeds of up to 10 Mbps.

4 MOBILITY AND THE INTERNET

The Internet Protocol assumes that a node's IP address uniquely identifies its point of attachment to the Internet. Datagrams destined to a node are routed toward the network indicated by the network identifier part of the node's IP address. In order to receive datagrams destined to it, a node must be located in the network indicated by its IP address. If the node would move to another network, datagrams destined to it would be undeliverable. Therefore, for a node to change its point of attachment without losing its ability to communicate, either the node's IP address should be changed or routers on the Internet should be informed about the node's new location. Both of these alternatives are often unacceptable.

Changing a node's IP address makes it impossible for a node to maintain its transport and upper layer connections when the node moves to a new network. Transport layer connections depend upon a constant IP address. Changing the IP address would cause all upper layer connections to terminate. The alternative, informing the Internet routing fabric about the node's new location, has obvious and severe scaling problems. For example, the node's current location information should be available for most routers almost immediately after the node has changed its location, which clearly is impossible in the case of the global Internet. In addition, the size of the routing tables, already considered as a problem in the Internet, would grow dramatically as numerous individual mobile nodes would have to be added into them.

4.1 Mobile IP

Mobile IP [RFC2002] provides the means for a node to maintain its IP address while moving from one network to another. Mobile IP, also known as IP Mobility Support and Mobile IPv4, is a network layer protocol developed by the Internet Engineering Task Force (IETF). The Mobile IP architecture consists of routers and nodes supporting Mobile IP enhancements. Both nodes not supporting Mobile IP and nodes using Mobile IP enhancements are able to communicate with each other.

The three components of the Mobile IP (MIP) architecture are [RFC2002]:

Mobile Node (MN) is a host or router that changes its point of attachment from one network or subnetwork to another. A mobile node may change its location without changing its IP address; it may continue to communicate with other Internet nodes at any location using its constant IP address, assuming link layer connectivity to a point of attachment is available. From a Mobile IP point of view, link layer connections can be either wired or wireless since Mobile IP only specifies network layer methods.

Home Agent (HA) is a router on a mobile node's home network that tunnels datagrams to the mobile node's visited network for delivery to the mobile node when it is away from its home network and maintains current location information for the mobile node.

Foreign Agent (FA) is a router on a mobile node's visited network that provides routing services to the mobile node while registered on the network. The foreign agent detunnels and delivers datagrams to the mobile node that were tunneled by the mobile node's home agent. For datagrams sent by a mobile node, the foreign agent may serve as a default router for registered mobile nodes.

A mobile node's home network is the network that the mobile node's home address (i.e., the node's long term, constant IP address) indicates. Other networks are considered as foreign networks. Without mobility functions, a mobile node would have to change its IP address when moving to a foreign network. A correspondent node (CN) is a node, either mobile or stationary, with which the mobile node is currently communicating.

Home agents and foreign agents may advertise their presence by sending Agent Advertisement messages on each link (i.e., a possible point of attachment) they are providing mobility services. With a lack of Agent Advertisements, a mobile node may send a solicitation for an Agent Advertisement on the link to learn if any prospective agents are present. As a mobile node receives Agent Advertisements, it is able to determine whether it is on its home network or on a foreign network. If a mobile node is located on its home network, it operates without mobility functions. All routing concerning the MN is done as if it was never away from its home network.

When a mobile node detects that it has moved to a foreign network, it obtains an additional IP address, a care-of address, on the foreign network. The care-of address can either be determined from an Agent Advertisement sent by a foreign agent, or by some external assignment mechanism, such as Dynamic Host Configuration Protocol [RFC2131]. The mobile node on a foreign network then registers its new care-of address with its home agent. If the care-of address was obtained from an Agent Advertisement, registration is done via the foreign agent that sent the Advertisement message. If a different foreign agent was operating on each link, then the MN must obtain and register a new care-of address every time it changes to a new link. If one foreign agent was providing its services for several points of attachments, then the same care-of address will be valid for all those points.

While a mobile node is registered on a foreign network, only the MN's home agent and the foreign agent the MN registered via, if such, are aware of the mobile node's care-of address. Other nodes and routers still use the mobile node's home address when sending datagrams to it. When datagrams destined to the mobile node arrive to its home network, the MN's home agent intercepts the datagrams and tunnels them to the mobile node's care-of address. Tunneling [RFC1853, RFC2003] is a method to encapsulate datagrams and forward them to another destination. There the datagrams are decapsulated and delivered to the final destination in their original form. Thus, the mobile node receives the datagrams just as it would receive them as if it was on its home network. Datagrams sent by the mobile node to correspondent nodes are generally delivered using standard IP routing mechanisms, not necessarily passing through the MN's home agent. Figure 9 illustrates the Mobile IP routing scheme.

Usually, foreign agents are used to provide care-of addresses for mobile nodes. A care-of address provided by a foreign agent is the IP address of the foreign agent itself. In this way, a single foreign agent can act as a tunnel end-point for several mobile nodes. No unnecessary demands are placed on the already burdened IPv4 address space. If other care-of address assignment mechanisms would be used, each mobile node would need an individual care-of address.

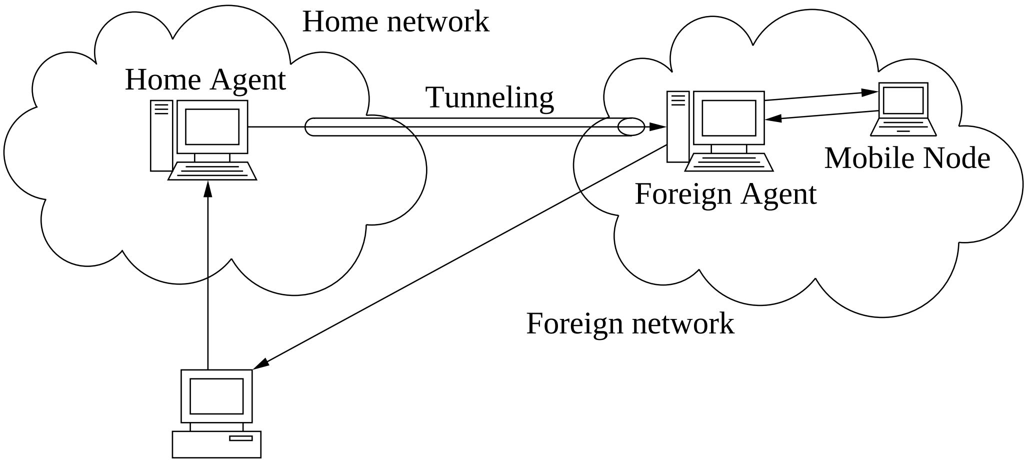


Figure 9: Mobile IP routing

Although Mobile IP has been proven to be a working concept for mobility management on the Internet, certain problems exist. As described, all datagrams destined to a mobile node are routed via the MN's home network. This triangle routing is inefficient and causes additional latencies that can disturb, for example, real-time applications. Also, if a mobile node frequently changes its care-of address, part of the datagrams are lost when they are sent to the old care-of address before the MN's home agent receives information about the new care-of address. Extensions to Mobile IP have been proposed and specified to overcome Mobile IP related problems. For example, Route Optimization in Mobile IP [Per00] is used to avoid triangle routing. Although the extensions solve many problems, it cannot be guaranteed that all parties implement all extensions.

4.2 Mobile IPv6

Mobile IPv6 [Joh00] uses the features of IPv6 to overcome the problems of Mobile IPv4 and to provide an efficient mobility management method for IPv6 networks. The Mobile IPv6, also referred to as Mobility Support in IPv6, specification is not yet released as an official IETF RFC (Request For Comments) standard, but several drafts have been made available. The final standard specification is expected to be released in the end of 2001 or beginning of 2002.

The basic concepts of Mobile IPv6 (MIPv6) are similar to Mobile IPv4. Each mobile node has a home address and a home network. If a mobile node is located on its home network, all routing is done using standard IP routing mechanisms without mobility functions. Mobile node movement detection is based on Router Advertisements [RFC2461, Joh00], sent by standard IPv6 routers.

On a foreign network, a mobile node obtains a care-of address and informs the home agent about the new care-of address. Due to the vast address space of IPv6, each mobile node can be provided an individual care-of address. Care-of addresses are formed by the means of address autoconfiguration, provided by standard IPv6. No foreign agents or other external address assignment mechanisms, such as the Dynamic Host Configuration Protocol (DHCP), are needed.

Route Optimization, an additional extension to Mobile IPv4, is an integral part of Mobile IPv6 used to avoid inefficient triangle routing. A mobile node may send Binding Update (BU) messages to inform correspondent nodes about the MN's care-of address. After receiving a Binding Update, the correspondent node is able to send datagrams directly to the mobile node's care-of address, enabling end-to-end communications. If a correspondent node has not received a Binding Update, or is unable to process one, the CN will send datagrams to the MN's home network. There, the mobile node's home agent will tunnel the datagrams to the MN's care-of address.

Mobile IPv6 also enables a mobile node to inform any Mobile IPv6 home agent on its previously visited network about the MN's new care-of address. The home agent on the previously visited network can then forward datagrams arriving to the MN's old care-of address to the MN's new care-of address, minimizing the amount of lost datagrams.

In addition to more efficient routing, Mobile IPv6 utilizes IPsec to meet security requirements. Binding Updates are authenticated to prevent malicious BUs and datagrams may be encrypted to prevent third parties from extracting the information transferred.

Unfortunately, not everything is yet ready for a wide scale Mobile IPv6 deployment and usage. First of all, there is no way to use Mobile IPv6 on an IPv4 network. Therefore, the transition from IPv4 to IPv6 must happen before Mobile IPv6 can be widely used.

Another problem is that in order to authenticate Binding Updates sent to correspondent nodes, a mobile node should share a security association with each correspondent node. The creation of such numerous security associations is a heavy process for both mobile nodes and correspondent nodes, often impossible to accomplish. Solutions for the problem have been proposed (such as [Nik01]), but are still under development. It is also important to notice that if a correspondent node does not support Mobile IPv6, the correspondent node is unable to process a Binding Update and the inefficient triangle routing will be used.

Macro mobility refers to the movement of a node between different networks or subnetworks. A node's movement inside a network or a subnetwork is called micro mobility. Mobile IP in general is intended for solving the macro mobility problem. However, with short-range radio access technologies emerging, it may not be extraordinary that several foreign agents should be running within one network area (for example, one foreign agent running in each WLAN access point). However, for scalability reasons, it is inefficient to inform the home agent, possibly on the other side of the world, every time the mobile node changes its point of attachment within a single network.

4.3 Hierarchical Mobile IP

Hierarchical Mobile IP is an extension to Mobile IP in order to provide more scalable micro mobility management. In each foreign network there are one or more Gateway Foreign Agents (GFAs) that are connected to the Internet. Other foreign agents on the network are connected to the GFAs. Multiple hierarchy levels of foreign agents beneath the GFA level can exist, if Regional Foreign Agents (RFAs) are used. The terms GFA and RFA and their functions are specified in Mobile IP Regional Registration [Gus01]. Some implementations differ in details, but the basic idea is very similar in all implementations. The architecture of Hierarchical Mobile IP (HMIP) is illustrated in Figure 10. Hierarchical MIPv6 mobility management [Sol00] proposes a similar architecture for Mobile IPv6.

When registering for the first time on a foreign network implementing HMIP, a mobile

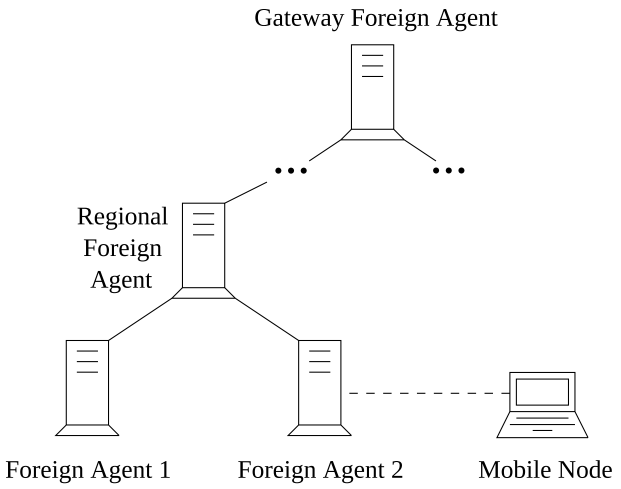


Figure 10: Hierarchical Mobile IP

node registers its new care-of address with its home agent. The care-of address is the IP address of a GFA and the registration is done via the GFA. After the initial registration, as the mobile node registers a new care-of address obtained from a newly found foreign agent under the same GFA, the registration is done with regional registration messages. The care-of address registered at the home agent (i.e., the GFA address) will not change when the mobile node changes its foreign agent under the same GFA. The lowest FA that the MN is already registered to, replies to regional registration messages. The mobile node's home agent will tunnel all packets to the GFA that will in turn deliver them to the MN, possible through other foreign agents on the network. A mobile node and the foreign network must support HMIP and other extensions to ensure that the hierarchical architecture will be used. If the extensions are not supported, only the basic Mobile IP model is employed.

4.4 Micro Mobility Management Methods

HAWAII (Handoff-Aware Wireless Access Internet Infrastructure) is another micro mobility management method, also an extension to Mobile IP. HAWAII [Ram00] is based on hierarchical domains and is transparent to mobile nodes supporting certain other extensions. Transparency means it is enough that only the foreign network supports HAWAII, while mobile nodes are unaware of it. Like HMIP, HAWAII reduces the changes of care-of addresses with the MN's home agent, resulting in less disruptions for packet transmissions.

Cellular IP (CIP) provides mobility support for frequently moving nodes on a single network. Both Cellular IPv4 [Cam99] and Cellular IPv6 [She00] have been specified. For global macro mobility support, Mobile IP can be used in conjunction with Cellular IP, but Mobile IP is not a prerequisite for the usage of Cellular IP. The usage of CIP requires that both the network and the mobile nodes support Cellular IP.

A Cellular IP Network is connected to the Internet through a Cellular IP Gateway. All packets to and from a Cellular IP Network are routed through the Gateway. Cellular IP mobile nodes are connected to a CIP Network via base stations (BSs), which can be, for example, WLAN access points. All base stations are connected to the Gateway, possibly via other base stations.

Each mobile node uses a single IP address while on a CIP Network. If the node and the Gateway both support Mobile IP, the mobile node is able to use the Gateway's address as its care-of address. The possibility to use several Gateways on one CIP Network is currently under further study.

For idle nodes on a Cellular IP Network, only the area a node is located in (paging area in CIP terms) is known, not the specific base station it is connected to. Incoming packets for the idle nodes are routed to all base stations in that area. As a node sends or receives packets, it moves to active state. For active nodes, the exact location is known and packets are routed directly to them. If a node does not send or receive any packets for a time, it moves to idle state. Both idle and active nodes receive beacon signals from base stations and send location update messages when changing location. Idle nodes send location update messages only when moving to a new paging area while active mobile nodes send location updates every time they move to the area of a new base station.

For a general micro mobility management solution, Hierarchical Mobile IP and HAWAII are not ideal, since they both rely on Mobile IP. Although Cellular IP can be used without Mobile IP, it requires additional components for mobile nodes and visited networks even when used with Mobile IP. A general solution should be independent of other mobility management protocols and not require anything other than standard components on mobile nodes. A general solution for the micro mobility problem, transparent to mobile nodes, is proposed later in this paper.

4.5 Session Initiation Protocol (SIP)

Mobile IP provides for terminals a mean to be reachable by a constant IP address, the home address. In order to initiate a connection with a mobile node (and with the person using it), the connection must be initiated using the MN's home address. However, it is expected that in the future a user may have several different kinds of mobile terminals and devices. All these devices will have an individual IP address. A user may use them at random, depending on their current activity. If one user wants to contact another, it would be awkward to try contacting the party's devices one by one until the currently used one is found. In addition, if a user is on, for example, a public terminal or newly acquired equipment, other users would probably be unaware of its IP address.

The Session Initiation Protocol [RFC2543] enables personal, terminal independent mobility by providing the capability to reach a called party at a single, location-independent address. The Session Initiation Protocol is an application layer control protocol for creating, modifying, and terminating end-to-end sessions with one or more participants. SIP is designed to be independent of the lower layer transport protocol and can be extended with additional capabilities.

SIP can be used to create sessions between two or more participants. Callers and callees are identified by unique SIP addresses, SIP URLs (Uniform Resource Locators). The SIP address takes a form similar to an e-mail address, i.e., user@host. The user part is a user name or a telephone number. The host part is either a domain name or a numeric network address. A SIP address can identify a communications device, a service, an individual, the first available person or service from a group or a whole group. A device can be configured with a SIP URL by using a SIM card, manually or by some other means depending on the device the user is using.

SIP invitations used to create sessions carry session descriptions, which allow participants to agree on a set of compatible media types. The Session Description Protocol [RFC2327] can be used for this purpose. The Session Description Protocol (SDP) is used to describe, for example, contact information for the person responsible for the session, the media comprising the session and the transport layer protocol used to transfer data during the session.

For SIP session data transportations, the Real-Time Transport Protocol (RTP) is one of the most commonly used transport layer protocols. RTP [RFC1189] provides end-to-end network transport functions suitable for applications transmitting real-time data, such as audio, video, or simulation data over the Internet. VoIP (Voice over IP) can be defined as a technology that enables all circuit switched PSTN based services on IP based networks. VoIP sessions are generally initiated with SIP and the data is transferred with RTP or UDP. TCP or UDP are usually used for transporting session control messages (i.e., signaling), but they can also be used for data transportation. The ability of SIP to use different protocols makes it suitable for a general session management protocol.

The usage of SIP in a fourth generation mobile network for mobility management is described in more detail in the next chapter.

5 A FOURTH GENERATION MOBILE NETWORK ARCHITECTURE

The need for fourth generation mobile networks originates with ever increasing mobility and the convergence of traditional telecommunication networks and content services. A fourth generation (4G) mobile network combines several traditional technologies to provide personal communication and content services for its mobile users.

5.1 Background

Previous chapters have introduced standards and techniques that are used on the Internet and in 2G and 3G mobile networks. The protocol architecture of the Internet has been proven to be both reliable and scalable, providing a solid foundation for a wide range of services. New wireless high-speed network access technologies and Internet mobility management methods enable extending and combining traditional technologies into a single entity, the 4G mobile network. The experiences from second and third generation mobile networks help us to understand how to manage a large number of mobile users.

The 4G mobile network architecture is based on Internet standards and wireless short-range, high-speed network access technologies. The architecture enables connections to other users without knowledge about their current terminal or location. Internet services can be used anywhere in the 4G mobile network area. The 4G network can be accessed with different kinds of terminals and devices. Client devices for 4G mobile network applications can be roughly categorized as follows:

- Laptop or PDA (Personal Digital Assistant) with 4G mobile network access
- Dual-mode wireless phone with GSM or UMTS and 4G mobile network access
- Wireless phone with 4G mobile network access
- Other specialized devices (e.g., wearables, small appliances, or public fixed terminals) with 4G mobile network access.

Fourth generation terminals are used to access various services, including traditional Internet services as well as services targeted to 4G terminal users only. A factor common to these services is that they all use IP as the basis of their communications. Thus, clearly 4G terminals must support IP. Hybrid mobile terminals, which support both IP and some other communications technologies such as 3G UMTS, may also be used. Since these hybrid terminals are also IP compatible, 4G operators are only required to offer IP based access to their networks and services.

5.2 Requirements for the Architecture

The fourth generation mobile network architecture should provide a scalable platform for service and content providers, enable reliable personal communication and provide personal and terminal mobility management. The following list summarizes the most essential requirements for the architecture.

Minimal Software Required in Terminals

Terminals should avoid unnecessary power consumption and processing activity. All electrical activity on a terminal increases its power consumption. Especially mobile terminals should minimize their electrical activity since they are running on batteries. Increased battery capacity or processing capability is not a good solution, since it also increases the cost of terminals. Therefore, all possible mobility management and administration related functions should be executed by the network, leaving as few functions as possible for terminals to execute. Another benefit in centralizing software to the network is that when software needs to be upgraded, no changes are required for terminals.

As Simple Base Stations as Possible

WLAN and Bluetooth access points have a maximum coverage range of roughly about 100 meters, depending on the radio transmitter and surroundings. With such a short coverage range, a relatively large number of access points (base stations) are needed. It is essential that access point cost is minimal since expensive access points could dramatically increase the overall cost of a network. The cost of an access point can be reduced by using limited processing capability. This also backs the idea of centralizing software to the network. Possible software upgrade effects on APs are little, facilitating network and service administration.

Open Standards Based Protocols

To avoid interoperability problems and service dispersion all communications should be based on open, standardized protocol definitions. Closed, proprietary protocols could easily lead to a situation where a user cannot access all the services needed because of protocol incompatibilities. This would cause additional costs for service providers who would be required to support several communication protocols instead of just one global standard. Users would be frustrated to find a fourth generation mobile network incompatible with their 4G terminal. Higher costs or any kind of incompatibilities could reduce general interest in the whole 4G concept.

Scalability

The architecture should be scalable both upwards and downwards. At the beginning of the 4G era, most of the population of the world will not be immediately able to use 4G technology. Later, globally even up to 10^{10} users, terminals and devices may be using 4G mobile networks. Locally user rates are highly dependent upon local population density. The architecture should be ready for high user data rates to avoid scalability and other problems which could lead to costly architecture changes.

A base station communicating with a mobile terminal indicates the current location of the terminal. The less coverage one BS has the more accurately the locations of terminals are known. Low coverage base stations result in more handovers than high coverage base stations, but provide more precise terminal location information. Therefore, some operators may prefer low coverage base stations even if high coverage BSs are available. The architecture should not set limits on operators choosing base stations for their networks.

Internet and VoIP Compatibility

The Internet and World Wide Web (WWW) provide a vast amount of services. Not all services could reasonably be modified for 4G terminals, hence the terminals and the architecture must be able to use standard Internet protocols to access Internet based services. Internet standards compliance benefits all. Users can access all available services with a single terminal. Service providers need to provide only one interface for their services in order to reach both Internet users and the users of 4G mobile networks.

Speech is the basic form of human communication. There is no doubt that people want to speak to each other no matter how much Internet based services and multimedia content are available. VoIP provides support for transferring voice over the Internet. Connections to other users should be possible without information about their whereabouts or current device.

5.3 Basic Components of a 4G Mobile Network

The physical components of fourth generation mobile networks are mobile terminals, base stations (access points) and gateways connecting local 4G networks to other networks, in particular the Internet. For administration, local content and service providing purposes, additional routers and servers may be used alongside the gateway. Mobile terminals communicate with the network via base stations using wireless radio connections. Base stations are usually connected to the gateway with wired connections.

The base stations of second and third generation mobile networks have rather high coverage. Base stations are expensive and due to their high coverage, the locations of terminals and users are not known accurately. For a fourth generation mobile network, typical base stations are low-cost and have lower coverage. Low coverage base stations are used to locate users more accurately, enabling location dependent services. With low-cost base stations which use unlicensed ISM radio bands for communications, a fourth generation mobile network may be set up without heavy financial investments.

Both the mobile terminal and network must support the same communication technology. Data transfer rates in fourth generations mobile networks are several or tens of megabits per second, depending on the used technology. Different technologies do not necessarily mean incompatibilities though. For example, the standard 802.11g based devices provide data transfer rates of up to 22 Mbps but are still compatible with 802.11 based devices which provide only 2 Mbps data rates. Compared to 3G UMTS, which serves data rates up to and possibly higher than 2 Mbps, fourth generation mobile networks have much more potential to provide real-time and multimedia content services.

One of the central ideas of fourth generation mobile networks are “hot spots”. A hot spot is an area covered by 4G base stations and usually owned by a single entity. The diameter of a hot spot can vary from a hundred meters to a few kilometers. Corporate offices, shopping centers, hotels and airports could form hot spots, each offering different kinds of services. Hot spots can be connected to private corporate LANs, public administration LANs or mobile WLANs installed in trains, ships, airplanes or buses. Areas not belonging to hot spots are covered by either high range 4G, 3G or 2G base stations. Dual-mode terminals may operate at higher speeds in hot spot areas and at lower speeds elsewhere.

The Internet acts as the core network of fourth generation mobile networks. Local 4G mobile networks are connected to each other via the Internet using standard Internet protocols. The Internet is an open network for all and requires no additional financial investments, making it suitable for use as a core network.

By connecting 4G mobile networks to the Internet, access can be easily provided to mobile users. If wanted, services and content offered by a 4G mobile network can be made available also for Internet users, not only for mobile users of that network. However, only mobile users are able to take full advantage of the location dependent services offered by fourth generation mobile networks.

Figure 11 presents an overall architecture of a fourth generation mobile network. The functions of the illustrated components are described in the next sections.

5.4 Mobility Management

Personal mobility management in the fourth generation mobile network is based on the Session Initiation Protocol. SIP enables the use of different transport layer protocols in

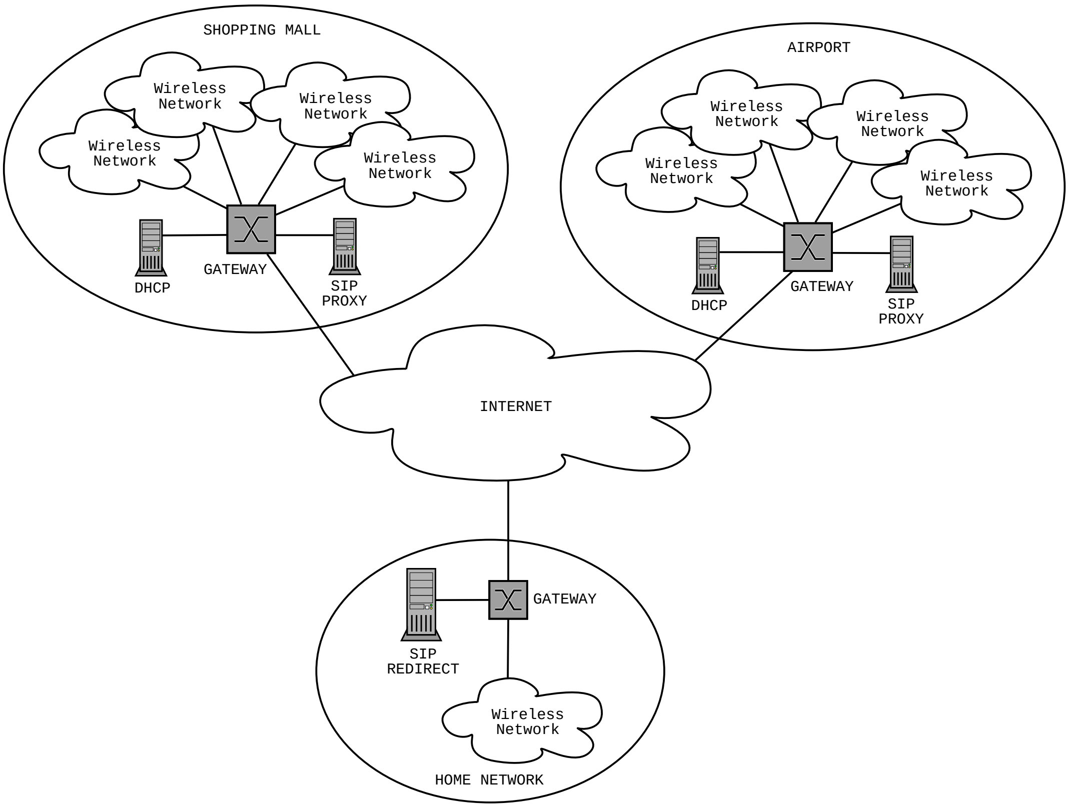


Figure 11: Overall architecture of a 4G mobile network

SIP initiated sessions. This is essential in 4G mobile networks where voice, interactive multimedia and other different kinds of sessions with individual requirements are present. SIP also enables the end-to-end connections required by, for example, VoIP sessions.

5.4.1 SIP Registration

A SIP session is initiated by sending a SIP request to a SIP URL. The request is routed to the home network of the contacted user, regardless of their current location. The information about current locations of users is maintained by SIP redirect servers contained in every network acting as a home network for mobile users. A SIP redirect server accepts a SIP request, maps the address into zero or more new addresses (SIP contact address) and returns these addresses to the client initiating the session. The new address or addresses indicate the current location and terminal of the contacted user. The SIP redirect server on the home network (home SIP redirect server) receives these new addresses from mobile terminals registering on foreign networks or users registering on new devices. After the client initiating the session has received the new addresses from the home SIP redirect server, it is able to initiate an end-to-end session with the contacted user. The home SIP redirect server is not used during the actual session, it is used only for querying the location and terminal of the contacted user while initiating the session. A home SIP redirect server is analogous to the HLR of the GSM network architecture.

A 4G Location Area (4GLA) is either a hot spot or an area between hot spots covered by high coverage 4G base stations. 4GLA is analogous to a GSM Location Area. Every 4G Location Area where visiting mobile terminals are allowed, contains a SIP outbound proxy server [Sch01c] which is analogous to the VLR in a GSM network. Visiting mobile users in a 4GLA register their current location with their home SIP redirect server using the SIP outbound proxy. The registration can be done in several ways, as described in [Sch01b]. In the fourth generation mobile network, either an outbound proxy intercept registration or a user-initiated proxy registration can be used.

Both outbound proxy intercept and user-initiated proxy registrations use the outbound proxy of the 4GLA to forward SIP registration messages to the home SIP redirect server. In the case of outbound proxy intercept, the terminal sends a registration message to its home network containing the terminal's current IP address as the SIP contact address. The outbound proxy in the visited 4GLA intercepts the registration message and changes the contact address to point to itself before forwarding the message. In the case of a user-initiated proxy registration, the terminal recognizes that it is visiting a foreign network and sends a registration message to the outbound proxy using the proxy's address as the contact address. The outbound proxy forwards the message to the terminal's home SIP redirect server without changes. After successful registration, all session initiation messages involving a visiting user travel through the outbound SIP proxy server. This allows network operators to collect statistics and possible billing information.

It is worth noticing that a user may not wish to register their location in certain location areas. If the user configures the terminal in use not to send registrations or uses a public terminal anonymously, the home SIP redirect server is not aware of the user's current location. Thus, the home SIP redirect server returns no contact addresses in response to session initiation messages, disabling incoming calls to the user. The user is still able to make outgoing calls if the terminal has a valid IP address in the visited location area. Network operators may require that a user must register in order to access Internet services, but may provide local services for unregistered users. These and other service related issues in 4G mobile networks are described with more detailed in [Ikk01].

5.4.2 Address Management on Foreign Networks

On a foreign network a mobile terminal needs a local IP address in order to use the foreign network and its services and to register a SIP contact address with its home SIP redirect server. The address could be configured manually, but this requires knowledge about the foreign network's available IP addresses and would cause unnecessary manual work for users. Therefore, the address should be obtained automatically by terminal software.

Mobile IP provides the means to obtain a care-of address on a foreign network. Terminal software automatically registers the care-of address with a foreign agent, if used, and with the home agent. The registered care-of address could also be used as a SIP contact address. The problem is that the terminal's home network must provide Mobile IP services (i.e., a home agent must be present in the home network). If no home agent is operating, the terminal fails to register its care-of address.

The Dynamic Host Configuration Protocol [RFC2131] is the de facto standard protocol for automatically assigning IP addresses for devices without a permanent IP address on a network. DHCP allows a network administrator to supervise and distribute IP addresses from a central point, allowing the collection of statistics and other information. DHCP messages can also be used by 4G mobile terminals to recognize that they are on a foreign network [Sch01a, RFC2132] and to locate the SIP outbound proxy server [Sch01c]. The usage of DHCP does not require any other software than a DHCP client for mobile terminals and a DHCP server for 4G mobile network.

The Dynamic Registration and Configuration Protocol (DRCP) is a proposal for a lightweight dynamic address configuration protocol, suitable especially for mobile hosts. Terminals using DRCP [McA00] are also able to use DHCP servers in the lack of DRCP servers. DRCP is under development at this time of writing, but it provides a notable alternative for address configuration in mobile environments.

IPv6 address autoconfiguration can also be used, but, naturally, this requires that both the terminal and the network support IPv6. The Dynamic Host Configuration Protocol for IPv6 (DHCPv6) offers the automatic allocation of reusable network addresses and additional configuration flexibility. DHCPv6 [Bou01] is a stateful counterpart to "IPv6 Stateless Address Autoconfiguration" [RFC2462], and can be used separately or concurrently with the latter to obtain configuration parameters.

5.4.3 Micro Mobility

This section proposes a general purpose micro mobility management method, experimented with in the "4G" project at the Laboratory of Telecommunications of Lappeenranta University of Technology. The term micro mobility refers to terminal mobility inside a single 4G Location Area. The method enables mobile terminals to use a single IP address while using different base stations and IP subnetworks inside a 4GLA. Micro mobility is controlled by the gateway router responsible for the 4GLA, which is also acting as a default router for mobile terminals visiting the 4GLA. Mobile terminals use only standard DHCP client software to obtain a local IP address, no additional IP mobility management software is required.

Link layer mobility is transparent to the network layer. Base stations, which all are connected to the gateway, only relay link layer frames between terminals and the gateway. The network layer of a mobile terminal is only aware of the gateway, individual base stations are invisible to it. A mobile terminal may use different base stations using a single IP address, if the link layer software (i.e., typically device drivers) in base stations and in the mobile terminal provide link layer mobility management functions. For example, all currently available WLAN equipment provide this capability. The device drivers of the terminal are able to accomplish a handover when a new, more reliable base station is located. The handover and the locating of a new base station is done solely by device drivers, not interfering with any network or upper layer protocols. After the handover, the terminal's device drivers will send all subsequent link layer frames to the newly found base station. Regardless of the base station the terminal is currently using, all IP packets are forwarded to it by the gateway. If all base stations belong to the same IP subnetwork, no additional network layer mobility management is required for any part of the network.

The usage of IP subnetworks effectively reduces collisions during packet transmissions in the physical medium. Large IP networks are usually divided into smaller IP subnetworks for scalability reasons. Without additional mobility management, a terminal should change its IP address when moving to a new IP subnetwork area in order to receive IP packets destined to it. Changing the IP address would terminate all transport layer connections and require a new SIP registration. This is clearly undesirable. By using a single IP address through different IP subnetworks, no additional SIP registrations would be needed nor would transport layer connections would terminate.

The method for preserving IP addresses in different subnetworks, experimented with in the "4G" project, is based on an intelligent gateway, DHCP and Proxy ARP [RFC925]. ARP, described in section 2.4, is meant for a single IP (sub)network. If a terminal sends an ARP request and the recipient is in a different IP subnetwork, the recipient will not receive the request and is unable to respond to it. If the gateway connecting the IP subnetworks replies with its own hardware address on behalf of a terminal in another subnetwork, terminals are able communicate without the need to know that they are in different subnetworks. This is, in short, the principle of Proxy ARP. No changes are required for terminals sending and receiving ARP requests and replies, only the gateway implements Proxy ARP.

Now a mobile terminal is able to communicate with other nodes regardless of the subnetwork or network where they are located. If the terminal moves to another subnetwork, it may continue communicating if the gateway still routes IP packets to and from it despite the fact that the terminal's IP address belongs to another subnetwork. The gateway is required to monitor packet streams involving the mobile terminal. Any packet originating from the terminal is routed normally by the gateway. Based on the packets send by a terminal, the gateway updates its routing table, if necessary, to reflect the terminal's current location.

The gateway can easily monitor packet streams originating from a mobile node. If the terminal is idle, it does not send any packets. This problem can be solved by using a standard DHCP client on the terminal. The IP address the mobile terminal obtained when arriving to the 4GLA must have a short lease time. The terminal starts sending DHCP Renewing Requests when, by default, half of the lease time is passed in order to renew its lease of the address [RFC2131]. The gateway uses these requests to track the movements of the otherwise idle mobile terminal and update its routing table if necessary. Again, the gateway is able to route packets destined to the terminal. No bridging is needed since IP packets are routed based on the movements of terminals. It is important to notice that the terminal's IP address must be renewed by the gateway regardless of the terminal's current IP subnetwork. With this method, a mobile terminal is able to use several IP subnetworks without changing its IP address.

This method requires an intelligent gateway that can perform the needed actions. The method can be used alongside with the presented overall architecture, but its usage is not mandatory. An evaluation of the scalability of the proposed method is currently under research.

5.4.4 Roaming

Roaming in 4G mobile networks is the movement from one 4GLA to another. As a user with a mobile terminal roams to a new 4GLA, the terminal must obtain an IP address provided by the new 4GLA network. Without such an address, the terminal cannot communicate with the network. The address can be assigned as described in section 5.4.2.

Several 4G Location Areas can be combined into one 4G Service Area (4GSA). A 4G Service Area is typically administrated by one network operator. An operator may use hierarchical SIP registrations [Wed99, Sch00] to reduce the number of SIP registrations send to home SIP redirect servers. In the hierarchical scheme, a SIP registration message from a mobile terminal is first received by the SIP outbound proxy server of the new 4GLA. This SIP outbound proxy server forwards the SIP registration message to the SIP outbound proxy managing the 4G Service Area where the mobile terminal is currently located. The home SIP redirect server is updated only when the mobile terminal moves from one 4GSA to another. This effectively reduces the number of SIP registrations sent to the home SIP redirect server, enabling it to serve more users. The mechanism works whether the home SIP redirect server and the visited 4GSA are administrated by the same network operator or not. Moreover, regardless of hierarchy, the SIP sessions are still end-to-end connections, any data sent during a session is exchanged directly between the session participants.

The transport layer connection used in a SIP session will terminate if the mobile terminal's IP address changes. In order to retain a SIP initiated session when roaming to a new 4GLA and obtaining a new IP address, the session data stream must be redirected to the new IP address of the terminal with SIP handoff messages [Big01]. A SIP handoff message enables moving one end of an active session to a new receiver without the need to create a new logical call. The new receiver may be another person or the same person using a different terminal or IP address. Although a new transport layer connection is created, the SIP session itself is uninterrupted. Since IP address can be changed, no Mobile IP is needed with SIP initiated sessions.

A possible approach to retain all transport layer connections is to use DHCP relay agents [RFC2131]. These agents would be operating in each 4G Location Area of a 4G Service Area while only one DHCP server would be operating on each 4GSA. DHCP relay agents pass DHCP messages between DHCP clients and DHCP servers. In this way, the IP address of a mobile terminal could be kept unchanged inside a 4GSA. Smooth location update would be carried out by multicasting IP packets arriving at a terminal to both location areas involved. The scalability of this solution has not been evaluated at this time of writing.

There is, nevertheless, a problem with transport layer connections retained from one network to another. If the data transfer rate provided by the new network is smaller than the previous network, the transport layer connection may clog. This is because the corresponding side of the connection sends datagrams at a faster rate than the new, slower network is able to deliver them. Despite the possible error control mechanisms of a transport layer protocol, the connection may be disturbed. For this reason, all vital data transfers should be initiated with SIP, removing the need for retaining transport layer connections while roaming.

6 DISCUSSION AND FUTURE WORK

This chapter explores some topics not covered by the previous chapters. No straight answers are given for raised issues, the topics are meant for sources of discussion. Some of the areas are currently under research, but the rest are work for the future.

The presented architecture does not cover physical implementation details. Two currently available wireless technologies, WLAN and Bluetooth, were introduced but the architecture is not limited to any particular link layer communication technology. Because of this independence, the architecture can be used with the technology which eventually becomes adopted by major fourth generation mobile terminal manufacturers.

Security was not discussed in the previous chapter, and WLAN and Bluetooth provide security functions only on the link layer. SIP registrations and terminal address obtaining and renewal in a foreign network should be secured and authenticated as well. It is essential that only an authorized user is able to register with her personal SIP URL. This requires, at least, that SIP registration messages are authenticated. Candidates for this purpose are, for example, “HTTP Authentication: Basic and Digest Access Authentication” [RFC2617] and PGP (Pretty Good Privacy, [RFC1991, RFC2440]) but further research is still needed to determine their overall suitability. For secure data transfers, IPsec provides security functions, but should be studied if it is appropriate for low processing capacity mobile terminals.

The Session Initiation Protocol provides a means for personal mobility. SIP handoff messages can be used to move an active session to another IP address, making it possible to change a terminal’s IP address during a SIP session. Non-SIP initiated transport layer connections are not shielded from changing an IP address. Mobile IP could be used for providing a constant IP address (i.e., the home address) for a mobile terminal, but the terminal’s home network should provide Mobile IP services. It should be studied how the Session Initiation Protocol and Mobile IP could most complement each other for providing personal and terminal mobility while shielding all transport layer connections, not only SIP initiated connections.

One interesting scenario is roaming between fourth generation and third or second generation mobile networks. A user with a dual-mode terminal may be having a call when roaming a fourth generation mobile network. If the user leaves the 4G mobile network, the call should be transferred to the control of a suitable mobile MSC, otherwise the call will terminate. On the other hand, if a user roams a 4G mobile network, it could be beneficial to transfer the call to the control of the fourth generation mobile network for billing reasons. It must also be noticed that the current and previous mobile networks can be administrated by different network operators.

7 CONCLUSIONS

This Master's thesis presented an overall architecture of a fourth generation mobile network. Basic components of 4G mobile networks were introduced and essential requirements for the architecture were defined. Basic components of fourth generation mobile networks are mobile terminals, low-cost base stations and gateways connecting local networks, hot spots to the Internet. Hot spots are areas offering high-speed access to 4G mobile network services and to the Internet.

The Internet and its standards form the foundation of the architecture. Using the Internet as the core network enables users to access Internet based services and provides a global framework for mobility management related signaling. VoIP provides the services of traditional telecommunications networks for the users of fourth generation mobile networks.

The architecture enables end-to-end connections to other users without knowledge about their current location or terminal by using the Session Initiation Protocol. SIP provides a means for personal mobility by identifying users with SIP addresses which are location and terminal independent. End-to-end connections between users provide low latencies for data transfers with no overhead to home networks. SIP can be used to initiate different types of sessions, which is essential in 4G mobile networks where sessions with varying requirements are used.

VoIP and other Internet based services require high-speed network access. Two promising wireless high-speed access technologies, WLAN and Bluetooth, were introduced. However, the presented architecture is not limited to any particular access technology, it can be used with the technology eventually adopted by 4G mobile terminal manufacturers.

This thesis also proposed a general purpose micro mobility management method. The proposed method is based on an intelligent gateway which enables mobile terminals to use a single IP address in different IP subnetworks. No additional mobility management software is required for mobile terminals. The proposed method can be utilized with the presented network architecture.

REFERENCES

[RFC2132] Alexander, S. and Droms, R. DHCP Options and BOOTP Vendor Extensions. RFC 2132, March 1997.

[RFC1991] Atkins, D., Stallings, W., and Zimmermann, P. PGP Message Exchange Formats. RFC 1991, August 1996.

[Big01] Biggs, B. and Dean, R. SIP Call Control: Call Handoff. Internet draft, January 2001 (work in progress).

[Blu01a] Bluetooth SIG. Specification of the Bluetooth System, Core. Specification Volume 1 Version 1.1, February 2001.

[Blu01b] Bluetooth SIG. Specification of the Bluetooth System, Profiles. Specification Volume 2 Version 1.1, February 2001.

[Bou01] Bound, J., Carney, M., Perkins, C., and Droms, R. Dynamic Host Configuration Protocol for IPv6 (DHCPv6). Internet draft, April 2001 (work in progress).

[RFC2440] Callas, J., Donnerhacke, L., Finney, H., and Thayer, R. OpenPGP Message Format. RFC 2440, November 1998.

[Cam99] Campbell, A., Gomez, J., Wan C-Y., Turanyi, Z., and Valko, A. Cellular IP. Internet draft, October 1999 (work in progress).

[RFC2460] Deering, S. and Hinden, R. Internet Protocol, Version 6 (IPv6) Specification. RFC 2460, December 1998.

[RFC2131] Droms, R. Dynamic Host Configuration Protocol. RFC 2131, March 1997.

[0260] ETSI. Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS); Service description; Stage 1. GSM 02.60 version 7.5.0 Release 1998, July 2000.

[0360] ETSI. Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS); Service description; Stage 2. GSM 03.60 version 7.4.1 Release 1998, July 2000.

[21101] ETSI/3GPP. Digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); 3rd Generation mobile system Release 1999 Specifications. 3GPP TS 21.101 version 3.3.0 Release 1999, April 2001.

[RFC2617] Franks, J., Hallam-Baker, P., Hostetler, J., Lawrence, S., Leach, P., Luotonen, A., and Stewart, L. HTTP Authentication: Basic and Digest Access Authentication. RFC 2617, June 1999.

[Gus01] Gustafsson, E., Jonsson, A., and Perkins, C. Mobile IP Regional Registration. Internet draft, March 2001 (work in progress).

[RFC2327] Handley, M. and Jacobson, V. SDP: Session Description Protocol. RFC 2327, April 1998.

[RFC2543] Handley, M., Schulzrinne, H., Schooler, E., and Rosenberg, J. SIP: Session Initiation Protocol. RFC 2543, March 1999.

[RFC2373] Hinden, R. and Deering, S. IP Version 6 Addressing Architecture. RFC 2373, July 1998.

[Hui98] Huitema, C. IPv6: The New Internet Protocol, Second Edition. ISBN 0138505055, January 1998.

[RFC1715] Huitema, C. The H Ratio for Address Assignment Efficiency. RFC 1715, November 1994.

[Ikk01] Ikkelä, K. Fourth Generation Mobile Network Service Environment. Master's Thesis, Lappeenranta University of Technology, August 2001 (to appear).

[80211] ISO/IEC 8802-11:1999(E). ANSI/IEEE Std 802.11, 1999 edition. IEEE Standard, 1999.

[Joh00] Johnson, D. and Perkins, C. Mobility Support in IPv6. Internet draft, November 2000 (work in progress).

[RFC2402] Kent, S. and Atkinson, R. IP Authentication Header. RFC 2402, November 1998.

[RFC2406] Kent, S. and Atkinson, R. IP Encapsulating Security Payload (ESP). RFC 2406, November 1998.

[RFC2401] Kent, S. and Atkinson, R. Security Architecture for the Internet Protocol. RFC 2401, November 1998.

[Kle61] Kleinrock, L. Information Flow in Large Communication Nets. RLE Quarterly Progress Report, July 1961.

[Lic62] Licklider, J.C.R. and Clark, W. On-Line Man Computer Communication. August 1962.

[Mar00] Martikainen, O., Naumov, V., and Kolenikov D. Internet Service Management. Smartnet 2000, 2000.

[McA00] McAuley, A., Das, S., Madhani, S., Baba, S., and Shobatake, Y. Dynamic Registration and Configuration Protocol. Internet draft, July 2000 (work in progress).

[Mou92] Mouly, M. and Pautet, M-B. The GSM System for Mobile Communications. ISBN 0945592159, June 1992.

[RFC2461] Narten, T., Nordmark, E., and Simpson, W. Neighbor Discovery for IP Version 6 (IPv6). RFC 2461, December 1998.

[Nik01] Nikander, P. and Perkins, C. Binding Authentication Key Establishment Protocol for Mobile IPv6. Internet draft, April 2001 (work in progress).

[RFC2003] Perkins, C. IP Encapsulation within IP. RFC 2003, October 1996.

[RFC2002] Perkins, C. IP Mobility Support. RFC 2002, October 1996.

[Per00] Perkins, C. Route Optimization in Mobile IP. Internet draft, November 2000 (work in progress).

[RFC826] Plummer, D. An Ethernet Address Resolution Protocol. RFC 826, November 1982.

[RFC791] Postel, J. Internet Protocol. RFC 791, September 1981.

[RFC925] Postel, J. Multi-LAN Address Resolution. RFC 925, October 1984.

[RFC793] Postel, J. Transmission Control Protocol. RFC 793, September 1981.

[RFC768] Postel, J. User Datagram Protocol. RFC 768, August 1980.

[Ram00] Ramjee, R., La Porta, T., Thuel, S., Varadhan, K., and Salgarelli, L. IP micro-mobility support using HAWAII. Internet draft, July 2000 (work in progress).

[Rob66] Roberts, M. and Merrill, T. Toward a Cooperative Network of Time-Shared Computers. Fall AFIPS Conference, October 1966.

[RFC1189] Schulzrinne, H., Casner, S., Frederick, R., and Jacobson, V. RTP: A Transport Protocol for Real-Time Applications. RFC 1889, January 1996.

[Sch00] Schulzrinne, H. and Wedlund, E. Application-Layer Mobility using SIP. Mobile Computing and Communications Review, July 2000.

[Sch01a] Schulzrinne, H. and Nair G. DHCP Option for SIP Servers. Internet draft, March 2001 (work in progress).

[Sch01b] Schulzrinne, H. SIP Registration. Internet draft, April 2001 (work in progress).

[Sch01c] Schulzrinne, H. and Rosenberg, J. SIP: Session Initiation Protocol – Locating SIP Servers. Internet draft, March 2001 (work in progress).

[Sol00] Soliman, H., Casteluccia, C., El-Malki, K., and Bellier L. Hierarchical MIPv6 mobility management. Internet draft, September 2000 (work in progress).

[She00] Shelby, Z., Gatzounas, D., and Campbell, A. Cellular IPv6. Internet draft, November 2000 (work in progress).

[RFC1853] Simpson, W. IP in IP Tunneling. RFC 1853, October 1995.

[RFC2462] Thomson, S. and Narten, T. IPv6 Stateless Address Autoconfiguration. RFC 2462, December 1998.

[Wed99] Wedlund, E. and Schulzrinne, H. Mobility Support using SIP. WoWMoM'99, August 1999.