MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY

Autonomous Institution – UGC, Govt. of India

Department of COMPUTATIONAL INTELLIGENCE (CSE-AIML, AIML)

B.TECH(R-20 Regulation)

(III YEAR – II SEM)

2023-24

MOBILE COMPUTING

(R20A0523)

LECTURE NOTES

MALLA REDDY COLLEGE OF ENGINEERING & TECHNOLOGY

(Autonomous Institution – UGC, Govt. of India)

Recognized under 2(f) and 12(B) of UGC ACT 1956

(Affiliated to JNTUH, Hyderabad, Approved by AICTE-Accredited by NBA & NAAC – 'A' Grade - ISO 9001:2015 Certified)

Maisammaguda, Dhulapally (Post Via. Hakimpet), Secunderabad–500100, Telangana State, India

Department of COMPUTATIONAL INTELLIGENCE

(CSE-AIML, AIML)

MOBILE COMPUTING

(R20A0523)

LECTURE NOTES

Prepared by

MRS. PRIYA ANIKET GHUGE,

ASSISTANT PROFESSOR

Department of Computational Intelligence

Artificial Intelligence and Machine Learning

Vision

To be a premier center for academic excellence and research through innovative interdisciplinary collaborations and making significant contributions to the community, organizations, and society as a whole.

Mission

- To impart cutting-edge Artificial Intelligence technology in accordance with industry norms.
- To instill in students a desire to conduct research in order to tackle challenging technical problems for industry.
- To develop effective graduates who are responsible for their professional growth, leadership qualities and are committed to lifelong learning.

QUALITY POLICY

- To provide sophisticated technical infrastructure and to inspire students to reach their full potential.
- To provide students with a solid academic and research environment for a comprehensive learning experience.
- To provide research development, consulting, testing, and customized training to satisfy specific industrial demands, thereby encouraging self-employment and entrepreneurship among students.

For more information: www.mrcet.ac.in

Syllabus

M R C E T CAMPUS | AUTONOMOUS INSTITUTION - UGC, GOVT. OF INDIA

B.Tech III Year II Sem-CSE(AI&ML) **L/T/P/C**

3 -/-/-3

(R20A0523) MOBILE COMPUTING

PROFESSIONAL ELECTIVE -III

Course Objectives:

1. To make the student understand the concept of mobile computing paradigm, its novel applications and limitations.
2. To understand the typical mobile networking infrastructure through a popular GSM protocol.
3. To understand the issues and solutions of various layers of mobile networks, namely MAC layer, Network Layer & Transport Layer
4. To understand the database issues in mobile environments & data delivery models.
5. To understand the ad hoc networks and related concepts.
6. To understand the platforms and protocols used in mobile environment.

UNIT – I

Introduction: Mobile Communications, Mobile Computing – Paradigm, Promises/Novel Applications and Impediments and Architecture; Mobile and Handheld Devices, Limitations of Mobile and HandheldDevices. GSM – Services, System Architecture, Radio Interfaces, Protocols, Localization, Calling, Handover, Security, New Data Services, GPRS, CSHSD, DECT.

UNIT – II (Wireless) Medium Access Control (MAC): Motivation for a specialized MAC (Hidden and exposed terminals, Near and far terminals), SDMA, FDMA, TDMA, CDMA, Wireless LAN/(IEEE 802.11) Mobile Network Layer: IP and Mobile IP Network Layers, Packet Delivery and Handover Management,Location Management, Registration, Tunneling and Encapsulation, Route Optimization, DHCP.

UNIT – III Mobile Transport Layer: Conventional TCP/IP Protocols, Indirect TCP, Snooping TCP, MobileTCP, Other Transport Layer Protocols for Mobile Networks.
Database Issues: Database Hoarding & Caching Techniques, Client-Server Computing & Adaptation,Transactional Models, Query processing, Data Recovery Process & QoS Issues.

UNIT – IV Data Dissemination and Synchronization: Communications Asymmetry, Classification of DataDelivery Mechanisms, Data Dissemination, Broadcast Models, Selective Tuning and Indexing Methods, Data Synchronization – Introduction, Software, and Protocols

UNIT – V Mobile Adhoc Networks (MANETs): Introduction, Applications & Challenges of a MANET, Routing, Classification of Routing Algorithms, Algorithms such as DSR, AODV, DSDV, etc. , Mobile Agents, Service Discovery. Protocols and Platforms for Mobile Computing: WAP, Bluetooth, XML, J2ME,Java Card, Palm OS, Windows CE, Symbian OS, Linux for Mobile Devices, Android.

TEXT BOOKS:

1. Jochen Schiller, ―Mobile Communications‖, Addison-Wesley, Second Edition, 2009.

2. Raj Kamal, ―Mobile Computing‖, Oxford University Press, 2007, ISBN: 0195686772.

REFERENCE BOOKS:

1. Jochen Schiller, ―Mobile Communications‖, Addison-Wesley, Second Edition, 2004.
2. Stojmenovic and Cacute, ―Handbook of Wireless Networks and Mobile Computing‖, Wiley, 2002, ISBN 0471419028.
3. Reza Behravanfar, ―Mobile Computing Principles: Designing and Developing Mobile Applicationswith UML and XML‖, ISBN: 0521817331, Cambridge University Press, Oct 2004.

Course Outcomes:

1. Able to think and develop new mobile application.
2. Able to take any new technical issue related to this new paradigm and come up with a solution(s).
3. Able to develop new ad hoc network applications and/or algorithms/protocols.
4. Able to understand & develop any existing or new protocol related to mobile environment
5. To create software systems using scripting languages such as Perl, PHP, and Ruby.

Unit-1

1.1. Introduction to Mobile Communication:

The rapidly expanding technology of cellular communication, wireless LANs, and satellite services will make information accessible anywhere and at any time. Regardless of size, most mobile computers will be equipped with a wireless connection to the fixed part of the network, and, perhaps, to other mobile computers.

Mobility and portability will create an entire new class of applications and, possibly, new massive markets combining personal computing and consumer electronics.

Mobile communication entails transmission of data to and from handheld devices. The location of the device can vary either locally or Globally.

Mobile Communication takes place through a wireless, distributed or diversified network and it is a two-way transmission or reception of data streams. Signals from a system can be transmitted through a fiber, wire, or wireless medium

1.1.1. GUIDED TRANSMISSION:

- Metal wires and optical fibres guided or wired transmission of data.
- Guided transmission of electrical signals takes place using four types of cables
  - Optical fiber, Coaxial cable, Twisted-pair cable, Power line
- Fibre- and wire- based transmission and their ranges

Advantages:

- Transmission along a directed path from one point to another
- Practically no interference in transmission from any external source or path
- Using multiplexing and coding, a large number of signal-sources simultaneously transmitted along an optical fibre, a coaxial cable, or a twisted-pair cable

Disadvantages:

- Signal transmitter and receiver fixed.
- Number of transmitter and receiver systems limits the total number of interconnections possible

1.1.2. UNGUIDED (WIRELESS) TRANSMISSION:

- Wireless or unguided transmission is carried out through radiated electromagnetic energy.
- Electromagnetic energy flows in free space (air or vacuum).
- The radiated energy is of frequency in MHz or GHz spectrum range.
- Spectrum means a set of frequencies in a range.

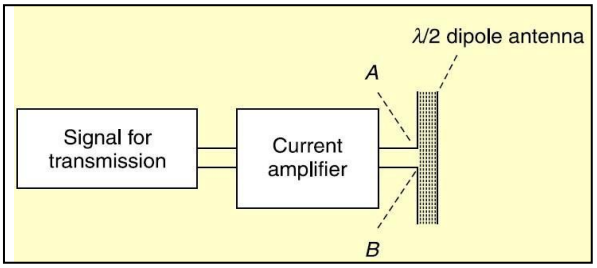
a) Signal Propagation Frequencies:

- Electrical signals transmitted by converting them into electromagnetic radiation.
- These radiations are transmitted via antennae that radiate electromagnetic signals.
- There are various frequency bands within the electromagnetic spectrum.
- The various types of frequencies are,
  - Long Wavelength (LW) radio, very low frequency.
  - Medium Wavelength (MW) radio, medium frequency.
  - Short Wavelength (SW) radio, high frequency.
  - FM radio band frequency.
  - Very High Frequency (VHF).

- Ultra High Frequency (UHF).
- Super High microwave Frequency (SHF).
- Extreme High Frequency (EHF).
- Far infrared
- Infrared
- Visible Light and Ultra-violet

b) Antennae:

- Devices that transmit and receive electromagnetic radiations.
- Most antennae function efficiently for relatively narrow frequency ranges.
- If an antenna not properly tuned to the frequency band in which the transmitting system connected to it operates, the transmitted or received signals may be impaired.



- The forms of antennae used are chiefly determined by the frequency ranges they operate in and can vary from a **single piece of wire** to a **parabolic dish.**

c) Modulation:

- Modulation means modification to original action so that the modification in the action is clearly presented.
- For example, a professor's voice is modulated and reflects his command over the subject.
- Similarly, electrical signals are modulated with information or electrical signals which is then communicated over long distances.
- Types of modulation are,
  - Analog signal modulation, Digital signal modulation, Amplitude modulation
  - Amplitude shift keying, Frequency modulation, Frequency shift keying
  - Phase modulation, Phase shift keying
  - Binary phase shift keying, Gaussian minimum shift keying,
  - Quadrature phase shift keying, Eight phase shift keying
  - Quadrature amplitude modulation (QAM).
  - 64-QAM

d) **Modulation methods and standards for voice-oriented data communication standards:**

1G:- Devices have only voice-oriented communication.

2G:- 2G devices communicate voice as well as data signals.

-Came onto the market in 1988.

-Support data rates up to 14.4 kbps.

2.5G and 2.5G+: -Support data rates up to 100 kbps.

3G:

-Higher data rates than 2G and 2.5G.

-Uses 2 Mbps or Higher for short distance transmission.

-Uses 384 Kbps for long distance transmission.

-Supports voice, data and multimedia streams.

-enable transfer of video clips and faster multimedia communication.

4G:

-Higher data rates than 3G.

-Support streaming data for video.

-Enables multimedia news paper, high resolution mobile TV.

-Support data rates up to 100 Mbps.

GSM and CDMA based standards and mobile communication network for long distance communication.

GSM: Global System for Mobile communication. It was developed by Groupe Speciale Mobile

(GSM) and Founded in Europe in 1982. It Support data rates up to 14.4 Kbps and Supports Cellular networks.

GSM900: It is using GMSK for transmitting 1's and 0's.

-Uses FDMA for channels and TDMA for user access in each deployed channel.

ii) EDGE and GPRS 2.5G and 3G:

-GSM has been enhanced to tri-band series and packet oriented data communication.
-GPRS is a packet-oriented service for data communication of mobile devices.
-Utilizes the unused channels in the TDMA mode in a GSM network.
-EDGE is an enhancement of the GSM phase 2.
-it has the data rates up to 48Kbps per 200KHz channel.

EGSM – Extended GSM.
GPRS – General Packet Radio Service.
EDGE – Enhanced Data rates for GSM Evolution.
EGPRS – Enhanced GPRS.

e) Modulation methods and standards for data and voice communication:

- CDMA – Code Division Multiple Access,
- FDMA – Frequency Division Multiple Access,
- TDMA – Time Division Multiple Access,
- WCDMA – Wireless CDMA
- UMTS – Upgraded WCDMA methods for downlink and uplink:
- High Speed Packet Data Access is provided by HSPDA and HSUPA.

1.2. **Mobile Computing** is a technology that allows transmission of data, voice and video via a computer or any other wireless enabled device without having to be connected to a fixed physical link.

Mobile Computing is also an umbrella term used to describe technologies that enable people to access network services anyplace, anytime, and anywhere.

A communication device can exhibit any one of the following characteristics:

- **Fixed and wired**: This configuration describes the typical desktop computer in an office. Neither weight nor power consumption of the devices allow for mobile usage. The devices use fixed networks for performance reasons.
- **Mobile and wired**: Many of today's laptops fall into this category; users carry the laptop from one hotel to the next, reconnecting to the company's network via the telephone network and a modem.
- **Fixed and wireless**: This mode is used for installing networks, e.g., in historical buildings to avoid damage by installing wires, or at trade shows to ensure fast network setup.
- **Mobile and wireless**: This is the most interesting case. No cable restricts the user, who can roam betweendifferent wireless networks. Most technologies discussed in this book dealwith this type of device and the networks supporting them. Today's most successful example for this category is GSM with more than 800 million users.

1.3. MOBILE COMPUTING ARCHITECTURE:

- It represents the architectural requirements for programming a mobile device.
- The requirements are Programming Languages, Functions of OS, and Functions of middleware for mobile systems.

- Mobile computing Architectural Layers, protocols and Layers

i) Programming Languages:

- A variety of programming languages are used in mobile computing architecture.
- Popular language used is Java
- J2ME and Javacard (Java for smartcard)
- J2EE is used for web and enterprise server-based applications of mobile services.
- DOTNET and Python 2.7 are also used.

ii) Functions of OS:

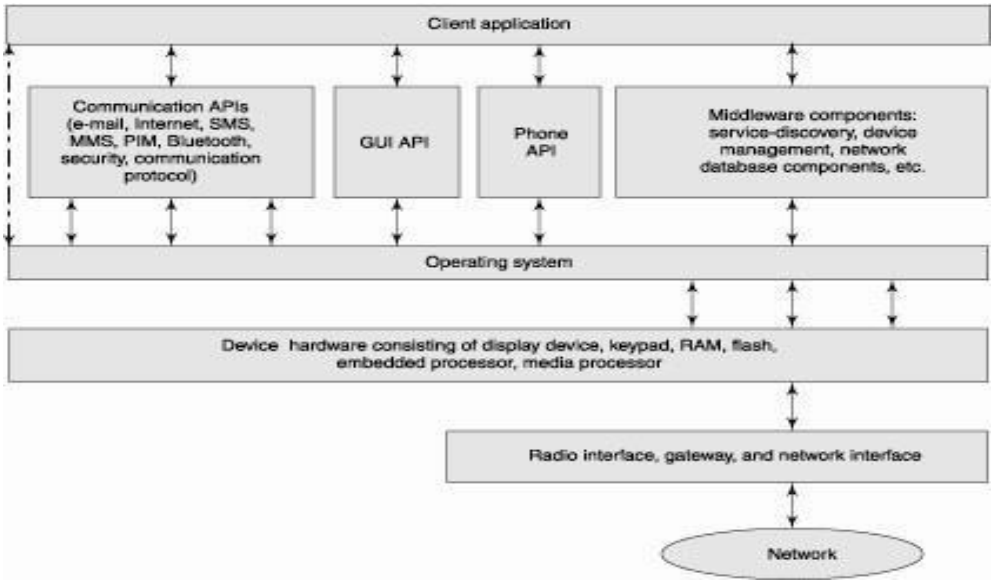
- An OS enables the user to run an application without considering the hardware specifications and functionalities.
- Scheduling multiple tasks.
- Management functions for tasks and memory.
- Interfaces for communication.
- Configurable libraries for the GUI in the device.
- Middleware.

iii) Functions of middleware for mobile systems:

- Middleware are the software components that link the application components with the network-distributed components.
- Mobile OS also provides middleware components.
- Examples are

  -to discover the nearby Bluetooth device, discover the nearby hotspot.
  -to retrieve data from a network database.
  -for service discovery

Mobile computing Architectural Layers:



- It refers to defining various layers between the user applications, interfaces, devices and network hardware.
- A well-defined architecture is necessary for systematic computations and access to data and software objects in the layers.

Protocols:

- GSM900, GSM900/1800/1900, CDMA, WCDMA, HSPA, UMTS, i-Mode, LTE, and WiMax.
- WPAN protocols such as Bluetooth, IrDA, and Zigbee.
- WLAN protocols such as 802.11a, and 802.11b and WAP

Layers: The OSI (open standard for interchange) seven-layer format is

-Physical for sending and receiving signals (TDMA or CDMA coding)
-Data link (Multiplexing), Networking (for linking to the destination)
-Wireless transport layer security (for establishing end-to-end connectivity)
-Wireless transaction protocol, Wireless session protocol, and Wireless application environment (for running a mob.appln., e.g., mobile e-business)

1.4. Novel Applications:

- A large number of applications are available.
- Recently made mobile TV realizable and developed ultra-mobile PC in march-2006
- Some applications include,
    - Smart-phones, Enterprise solutions, Music, Video and E-books, Mobile Cheque, Mobile Commerce, and Mobile based supply chain management.

SmartPhones:

- A Smartphone is a mobile phone with additional computing functions so as to enable multiple applications.
- For example, **Blackberry 8530 curve** has additional computational capabilities

-SMS, MMS, phone, email, address book, web browsing
-PIM software
-Integrated attachment viewing
-QWERTY style layout
-Send/End keys
-Bluetooth capability for hand-free talking via headset, ear buds, and car kits
-Speakerphone
-polyphonic ringtones for personalizing your device
-bright high resolution display, supporting over 65,000 colours
-WiFi 802.11b and WiFi 802.11g
-GPS Tracker
-Media voice or video or camera picture recording and communication
-Live TV
-MicroSD card 256MB

Enterprise Solutions:

- Enterprises or large business networks have huge database and documentation requirements.
- The term 'enterprise solutions' therefore refers to business solutions for corporations or enterprises.
- It includes specialized hardware or software programming for,

-Storage management
-Security
-Revision
-Distribution and so on.

Music, Video and E-books:

- The Apple iPods or iPhones or iPads have made it possible to listen one's favourite tunes anytime anywhere.
- iPads have made it possible to read one's favourite book anytime anywhere.
- Besides storing music these players can also be used to view photo albums, slide shows and video clips.

Mobile Cheque and Mobile Commerce:

- Mobile Cheque is a mobile based payment system employed during a purchase.
- The service is activated through text message exchanges between the customer, a designated retail outlet, and the mobile service provider.

- M-commerce is also a new trend, such as buying or selling of items through mobile internet between customers and organizations.
- Mobile devices also used for e-ticketing, i.e, for booking cinema, train, flight and bus tickets.

Mobile-based Supply Chain Management:

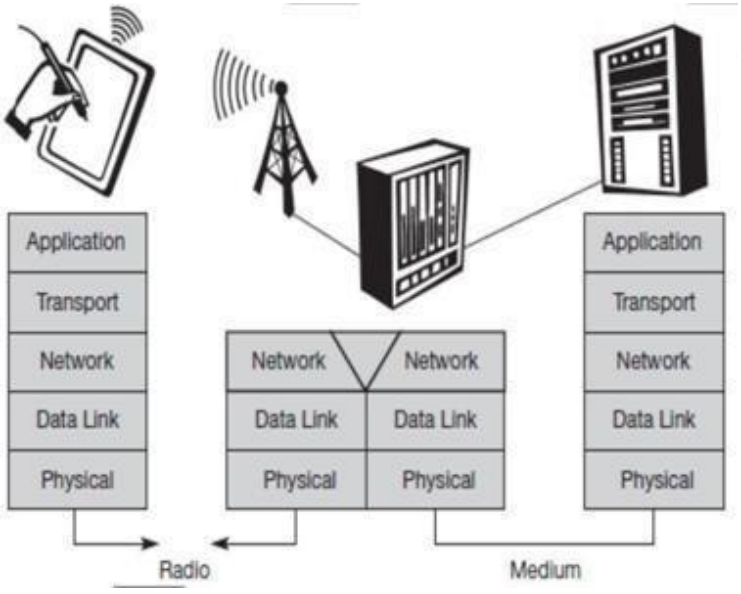
- The producer-consumer problem is called as SCM problem.
- Leading IT companies have developed mobile device software for SCM systems.
- The sales force and the manufacturing units use such mobile devices to maintain SCM.

1.5. Limitations of Mobile Computing:

- Resource Constraints : Battery
- Interference: *Radio transmission cannot be protected against interference using shielding and result in higher loss ratesfortransmitted data or higher bit error rates respectively.*
- Bandwidth: *Although they are continuously increasing, transmission rates are still very lowfor wireless devices compared to desktop systems. Researchers look for more efficient communication protocols with lowoverhead.*
- Dynamic changes in communication environment: *variations in signal power within a region, thus link delays and connection losses*
- Network issues : *discoveryof theconnection-service todestinationandconnectionstability*
- Interoperability issues*:*
- Security constraints: *Not only can portable devices be stolen more easily, but the radio interface is also prone to the dangers of eavesdropping. Wireless access must always include encryption, authentication, and other security mechanisms that must be efficient and simple to use.*

1.6. A simplified reference model

The figure shows the **protocol stack** implemented in the system according to the reference model. **End- systems** (such as the PDA and computer) need a full protocol to handle the application layer, transport layer, network layer, data link layer, and physical layer? Applications on the end-systems communicate with each other using the lower layer services. **Intermediate systems**, such as the interworking unit, do not necessarily need all of the layers.



A Simplified Reference Model

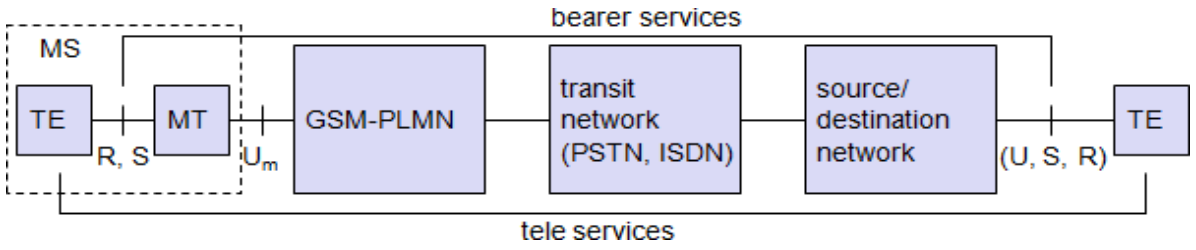
- **Physicallayer**: This is the lowest layer in a communication system and is responsible to convert the stream of bits into signals that can be transmitted on the sender side. The physical layer of the receiver then transforms the signals back into a bit stream. For wireless communication, the physical layer is responsible for frequency selection, generation of the carrier frequency, signal detection, modulation of data onto a carrier frequency and encryption.

- **Data link layer**: The main task of this layer is to access the medium, multiplexing of different data streams, correction of transmission errors, and synchronization (i.e., detection of a data frame). Therefore, the data link layer is responsible for a reliable point-to-point connection between two devices or a point-to-multipoint connection between one sender and several receivers.
- **Network layer**: This third layer is responsible for routing packets through a network or establishing a connection between two entities over many other intermediate systems. Important functions are addressing, routing, device location, and handover between different networks.
- **Transport layer**: This layer is used in the reference model to establish an end-to-end connection
- **Application layer**: Finally, the applications are situated on top of all transmission oriented layers. Functions are service location, support for multimedia applications, adaptive applications that can handle the large variations in transmission characteristics, and wireless access to the world-wide web using a portable device.

1.7. ***GSM : Mobile services, System architecture, Radio interface, Protocols, Localization and calling,Handover,Security,andNewdataservices***.

1.7.1. GSM Services: GSM is the digital mobile telecommunication system in the world today. It is used by over 800 million people in more than 190 countries. GSM permits the integration of different voice and data services and the interworking with existing networks. Services make a network interesting for customers. GSM has defined three different categories of services: bearer, tele and supplementary services.

Bearer services: GSM specifies different mechanisms for data transmission, the original GSM



allowing for data rates of up to 9600 bit/s for non-voice services. Bearer services permit transparent and non-transparent, synchronous or asynchronous data transmission.

Transparent bearer services: Transfer of data using physical layer is said to be transparent when the interface for service uses only physical layer protocols. Physical layer is the layer which transmits or receives data after formatting or multiplexing or insertion of **forward error correction (FEC)** using a wired (fiber) or wireless (radio or microwave) medium.

Forward error correction (FEC) bits: The physical layer protocol in a GSM bearer service also provides for FEC. Bluetooth also provides FEC. The FEC bring out redundant bits along with the data to be transmitted. This redundant data allows the receiver to detect and correct errors.

- **Non-transparent bearer services** use protocols of layers two and three to implement error correction and flow control. These services use the transparent bearer services, adding a **radio link protocol (RLP).** This protocol comprises mechanisms of **high-level data link**

control (HDLC), and special selective-reject mechanisms to trigger retransmission of erroneous data.

- Synchronous and asynchronous data transmission:
    - Synchronous means data is transmitted from a transceiver at a fixed rate with constant phase differences are maintained b/w two devices. It means establish a constant clock rate b/w receiver and sender. (i.e not using handshaking technique).
    - Asynchronous means data is transmitted by the transceiver at variable rate b/w two devices. It means, first set the bandwidth and provide the clock rate b/w two devices.

GSM specifies several bearer services for interworking with PSTN, ISDN, and packet switched public data networks (PSPDN) like X.25, which is available worldwide. Data transmission can be full-duplex, synchronous with data rates of 1.2, 2.4, 4.8, and 9.6 kbit/s or full-duplex, asynchronous from 300 to 9,600 bit/s.

Tele services: GSM mainly focuses on voice-oriented tele services. These services encrypted (such as voice transmission, message services, and basic data communication) with terminals and send to / received from the PSTN or ISDN (e.g., fax).

The primary goal of GSM was the provision of high-quality digital voice transmission. Special codes (coder/decoder) are used for voice transmission, while other codes are used for the transmission of analog data for communication with traditional computer modems used in, e.g., fax machines.

Another service offered by GSM is the **emergency number** (eg 911, 999). This service is mandatory for all providers and free of charge. This connection also has the highest priority, possibly pre-empting other connections, and will automatically be set up with the closest emergency center.

It also offers the **Short Message Service (SMS)** for message transmission up to 160 characters. The successor of SMS, the **Enhanced Message Service (EMS),** offers a larger message size, formatted text, and the transmission of animated pictures, small images and ring tones in a standardized way. But with MMS, EMS was hardly used. MMS offers the transmission of larger pictures (GIF, JPG, WBMP), short video clips etc. and comes with mobile phones that integrate small cameras.

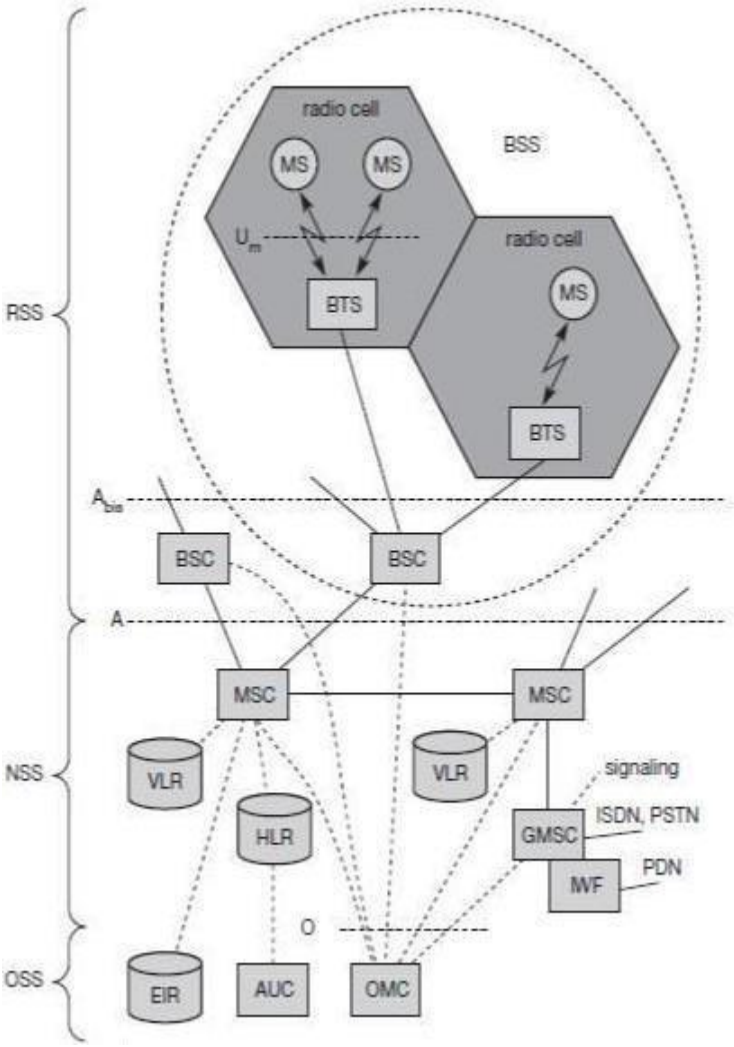
Supplementaryservices: GSMproviderscan offer **supplementary services**. Theseservicesoffer various enhancements for the standard telephony service, and may vary from provider to provider. Typical services are user **identification**, call **redirection**, or **forwarding** ofongoing calls, barring of incoming/outgoing calls, Advice of Charge (AoC) etc. Standard ISDN features such as **closed user groups** and **multiparty** communication may beavailable.

1.7.2. GSM Architecture:

A GSM system consists of three subsystems, the radio sub system (RSS), the network and switching subsystem (NSS), and the operation subsystem (OSS).

Network Switching Subsystem: The NSS is responsible for performing call processing and subscriber related functions. The switching system includes the following functional units:

- Some location register (HLR): The HLR has a database that used for storage and management of subscriptions. HLR stores all the relevant subscriber data including a subscribers service profile such as call forwarding, roaming, location information and activity status.



- Visitor location register (VLR): It is a dynamic real-time database that stores both permanent and temporary subscribers data which is required for communication b/w the coverage area of MSC and VLR.
- Authentication center (AUC): A unit called the AUC provides authentication and encryption parameters that verify the users identity and ensure the confidentiality of each call.
- Equipment identity register (EIR): It is a database that contains information about the identity of mobile equipment that prevents calls from stolen, unauthorized or defective mobile stations.
- Mobile switching center (MSC): The MSC performs the telephony switching functions of the system. It has various other functions such as 1. Processing of signal. 2. Control calls to and from other telephone and data systems. 3. Call changing, multi-way calling, call forwarding, and other supplementary services. 4. Establishing and terminating the connection b/w MS and a fixed line phone via GMSC.

Radio Subsystem (RSS): The **radio subsystem (RSS)** comprises all radio specific entities, i.e., the **mobile stations (MS)** and the **base station subsystem (BSS)**. The figure shows the connection between the RSS and the NSS via the A **interface** (solid lines) and the connection to the OSS via the **O interface** (dashed lines).

- Base station subsystem (BSS): A GSM network comprises many BSSs, each controlled by a base station controller (BSC). The BSS performs all functions necessary to maintain radio connections to an MS, coding/decoding of voice, and rate adaptation to/from the wireless network part. Besides a BSC, the BSS contains several BTSs.
- Base station controllers (BSC): The BSC provides all the control functions and physical links between the MSC and BTS. It is a high capacity switch that provides functions such as handover, cell configuration data, and control of radio frequency (RF) power levels in BT". A number of BSC's are served by and MSC.

- Basetransceiver station(BTS):The BTS handles the radio interface to the mobile station. A BTS can form a radio cell or, using sectorized antennas, several and is connected to MS via the Um interface, and to the BSC via the A BTS interface. The Um interface contains all the mechanisms necessary for wireless transmission (TDMA, FDMA etc.). The BTS is the radio equipment (transceivers and antennas) needed to service each cell in the network. A group of BTS's are controlled by an BSC.

Operation Service Subsystem (OSS): The OSS facilitates the operations of MSCs. The OSS also handles the Operation and maintenance (OMC) of the entire network.

Operation and Maintenance Centre (OMC): An OMC monitors and controls all other network entities through the 0 interface. The OMC also includes management of status reports, traffic monitoring, subscriber security management, and accounting and billing. The purpose of OSS is to offer the customer cost-effective support for centralized, regional and local operational and maintenance activities that are required for a GSM network. OSS provides a network overview and allows engineers to monitor, diagnose and troubleshoot every aspect of the GSM network.

The **mobile station (MS)** consists of the mobile equipment (the terminal) and a smart card called the Subscriber Identity Module (SIM). The SIM provides personal mobility, so that the user can have access to subscribed services irrespective of a specific terminal. By inserting the SIM card into another GSM terminal, the user is able to receive calls at that terminal, make calls from that terminal, and receive other subscribed services.

The mobile equipment is uniquely identified by the International Mobile Equipment Identity (IMEI). The SIM card contains the International Mobile Subscriber Identity (IMSI) used to identify the subscriber to the system, a secret key for authentication, and other information. The IMEI and the IMSI are independent, thereby allowing personal mobility. The SIM card may be protected against unauthorized use by a password or personal identity number.

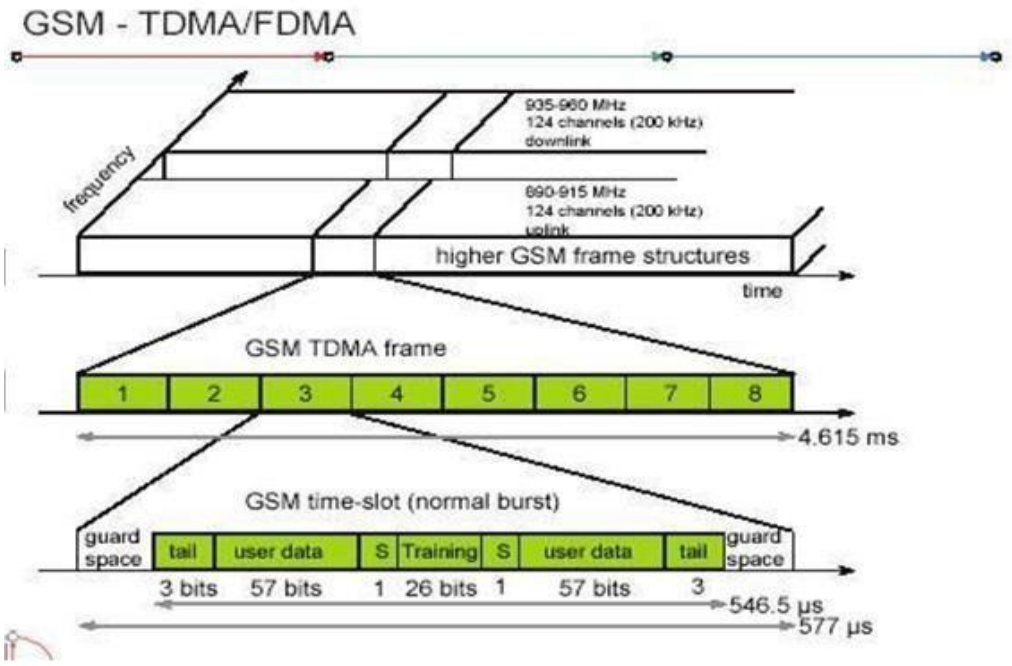
1.7.3. Radio Interface

The most interesting interface in a GSM system is the radio interface, as it contains various multiplexing and media access mechanisms.

Electric signals are given to antenna. The antenna radiates the electromagnetic waves. Electromagnetic waves propagate b/w the transmitter and receiver. Two electrical signals of two sources are not have same frequency at the same time. GSM TDMA Frame, Slots and Bursts

In the below figure, the GSM implements SDMA using cells with BTS and assigns an MS to a BTS. The diagram shows GSM TDMA frame. A frame is again subdivided into 8 GSM time slots, where each slot represents a physical TDM channel and lasts for 577 μs. Each TDM channel occupies the 200 kHz carrier for 577 μs every 4.615 ms. Data is transmitted in small portions, called bursts. As shown, the burst is only 546.5 μs long and contains 148 bits. The remaining 30.5 μs are used as guard space to avoid overlapping with other bursts due to different path delays and to give the transmitter time to turn on and off.

The first and last three bits of a normal burst (**tail**) are all set to 0 and can be used to enhance the receiver performance. The **training** sequence in the middle of a slot is used to adapt the parameters



ofthe receiver to thecurrent pathpropagation characteristics and to select thestrongest signal in case ofmulti-pathpropagation. A flag **S** indicateswhetherthe **data** field contains user or network control data.

Apart from the normal burst, ETSI (1993a) defines four more bursts for data transmission: a **frequency correction** burst allows the MS to correct the local oscillator to avoid interference with neighbouring channels, a **synchronization burst** with an extended training sequencesynchronizes the MS withthe BTS in time, an **access burst** is used for the initialconnectionsetupbetween MS and BTS, andfinallya **dummyburst** is used if no datais available for aslot.

Physical, logical channels and frame hierarchy: Twotypes ofchannels,namelyphysicalchannels andlogicalchannelsarepresent.

Physical channel: channel defined by specifying both, a carrier frequency and a TDMA timeslot number.

Logic channel: logical channels are multiplexed into the physical channels. Each logic channel performs a specific task. Consequently the data of a logical channel is transmitted in the corresponding timeslots of the physical channel. During this process, logical channels can occupy a part of the physical channel or even the entire channel.

Frame hierarchy:TDMA frames are grouped into two types of multiframes:

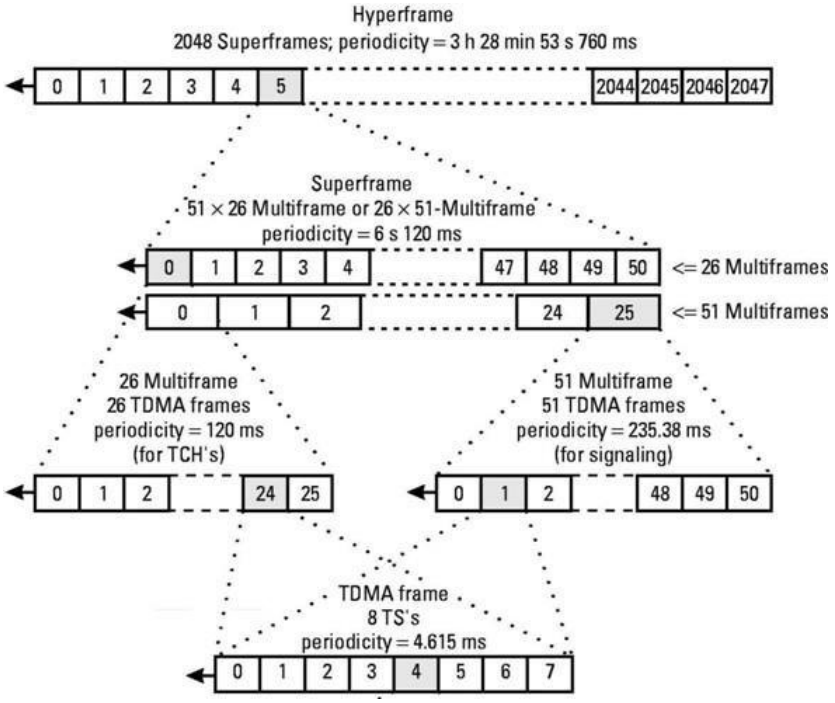
- 26-frame multiframe (4.615ms x 26 = 120 ms) comprising of 26 TDMA frames. This multiframe is used to carry traffic channels and their associated control channels.
- 51-frame multiframe (4.615ms x 51 □ 235.4 ms) comprising 51 TDMA frames. This

multiframe is exclusively used for control channels.

The multiframe structure is further multiplexed into a single superframe of duration of 6.12sec. This means a superframe consists of

- 51 multiframes of 26 frames.
- 26 multiframes of 51 frames.

The last multiplexing level of the frame hierarchy, consisting of 2048 superframes (2715648 TDMA frames), is a hyperframe. This long time period is needed to support the GSM data encryption mechanisms. The frame hierarchy is shown below:



GSM Frame Hierarchy

There are two different types of logical channel within the GSM system: Traffic channels (TCHs), Control channels (CCHs).

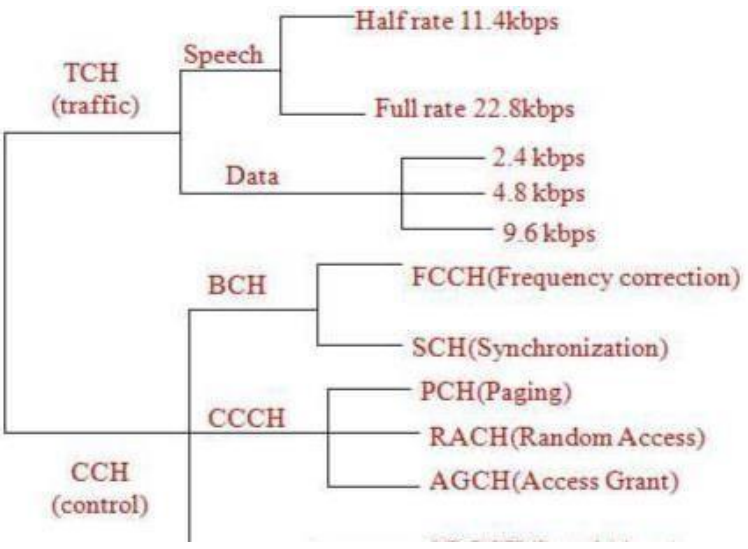
Traffic Channels: Traffic channels carry user information such as encoded speech or user data. Traffic channels are defined by using a 26-frame multiframe.

Two general forms are defined:

i. Fullratetrafficchannels(TCH/F), at a grossbit rateof22.8 kbps(456bits / 20ms)

ii. Half rate traffic channels (TCH/H), at a gross bit rate of 11.4 kbps.

Uplink and downlink are separated by three slots (bursts) in the 26-multiframe structure. This simplifies the duplexing function in mobile terminals design, as mobiles will not need to transmit and receive at the same time. The 26-frame multiframe structure, shown below multiplexes two types of logical channels, a TCH and a Slow Associated



Control CHannel (SACCH).

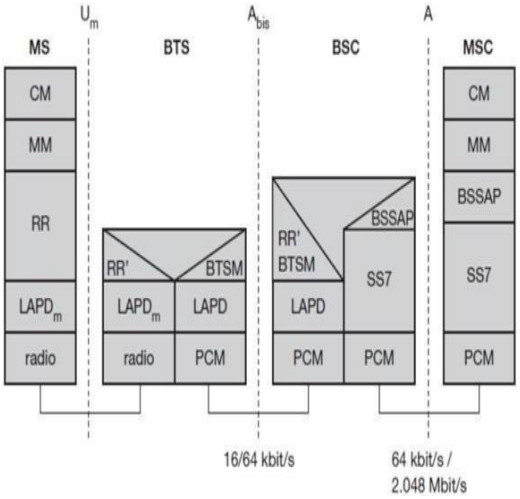
However, if required, a Fast Associated Control CHannel (FACCH) can steal TCH in order to transmit control information at a higher bit rate. This is usually the case during the handover process. In total 24 TCH/F are transmitted and one SACCH.

Control Channels: Control channels carry system signaling and synchronization data for control procedures such as location registration, mobile station synchronization, paging, random access etc. between base station and mobile station. Three categories of control channel are defined: Broadcast, Common and Dedicated. Control channels are multiplexed into the 51-frame multiframe.

- **Broadcast control channel (BCCH)**: A BTS uses this channel to signal information to all MSs within a cell. Information transmitted in this channel is, e.g., the cell identifier, options available within this cell (frequency hopping), and frequencies available inside the cell and in neighboring cells. The BTS sends information for frequency correction via the **frequency correction channel (FCCH)** and information about time synchronization via the **synchronization channel (SCH)**, where both channels are sub channels of the BCCH.
- **Common control channel (CCCH)**: All information regarding connection setup between MS and BS is exchanged via the CCCH. For calls toward an MS, the BTS uses the **paging channel (PCH)** for paging the appropriate MS. If an MS wants to set up a call, it uses the **random access channel (RACH)** to send data to the BTS. The RACH implements multiple access (all MSs within a cell may access this channel) using slotted Aloha. This is where a collision mayoccur with other MSs in a GSM system. TheBTSusesthe**accessgrant channel (AGCH)** to signalanMS that it canusea TCH or SDCCH for further connection setup.
- **Dedicated control channel (DCCH)**: While the previous channels have all been unidirectional, the following channels are bidirectional. As long as an MS has not established a TCH with the BTS, it uses the **stand-alone dedicated control channel (SDCCH)** with a low data rate (782 bit/s) for signaling. This can comprise authentication, registration or other data needed for setting up a TCH. Each TCH and SDCCH has a **slow associated dedicated control channel (SACCH)** associated with it, which is used to exchange system information, such as the channel quality and signalpowerlevel. Finally, if moresignaling informationneedsto be transmittedand a TCH already exists, GSM uses a **fast associated dedicated control channel (FACCH)**. The FACCHusesthe timeslotswhich are otherwise used by the TCH. This is necessary in the case of handovers where BTS and MS haveto exchange larger amounts of data in less time.

1.7.4. **GSM Protocols:** The signaling protocol in GSM is structured into three general layers depending on the interface, as shown below.

Layer 1 is the physical layer that handles all radio-specific functions. This includes the creation of bursts according to the five different formats, multiplexing of bursts into a TDMA frame, synchronization with the BTS, detection of idle channels, and measurement of the channel quality on the downlink.



The main tasks of the physical layer contain channel coding and error detection/ correction, which are directly combined with the coding mechanisms. Channel coding using different forward error correction (FEC) schemes.

Signaling between entities in a GSM network requires higher layers. For this purpose, the **$LAPD_m$** protocol has been defined at the U_m interface for **layer two**. $LAPD_m$ has been derived from link access procedure for the D-channel (**LAPD**) in ISDN systems, which is a version of HDLC.

The **network lay**er in GSM contains several sub-layers. The lowest sub-layer is the radio resource management (RR). The functions of RR' are supported by the BSC via the BTS management (BTSM). The maintasks of RR are setup, maintenance, and release of radio channels. Mobility management (MM) contains functions for registration, authentication, identification, location updating, and the provision of a temporary mobile subscriber identity (TMSI).

Finally, the call management (CM) layer contains three entities: call **control (CC), short message service (SMS**), and **supplementary service (SS).**

SMS allows for message transfer using the control channels SDCCH and SACCH, while SS offers the services like user identification, call redirection, or forwarding of ongoing calls.

CC provides a point-to-point connection between two terminals and is used by higher layers for call establishment, call clearing and change of call parameters.

Data transmission at the physical layer typically uses pulse code modulation (PCM) systems. LAPD is used for layer two at Abis, BTSM for BTS management.

Signaling system No. 7 (SS7) is used for signaling between an MSC and a BSC. This protocol also transfers all management information between MSCs, HLR, VLRs, AuC, EIR, and OMC. An MSC can also control a BSS via a BSS application part (**BSSAP**).

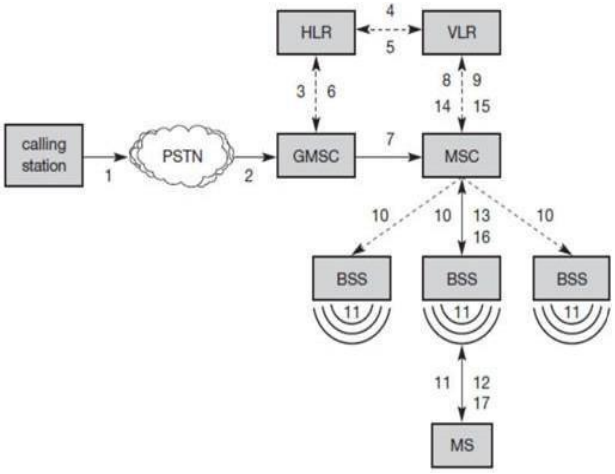
1.7.5. **Localization and Calling:** The fundamental feature of the GSM system is the automatic, worldwide localization of users for which the system performs periodic location updates. The HLR always contains information about the current location and the VLR currently responsible for the MS informs the HLR about the location changes. Changing VLRs with uninterrupted availability is called roaming.

Roaming can take place within a network of one provider, between two providers in a country and also between different providers in different countries.

To locate and address an MS, several numbers are needed:

- **Mobile station international ISDN number (MSISDN)**:- The only important number for a user of GSM is the phone number. This number consists of the country code (CC), the national destination code (NDC) and the subscriber number (SN).
- **International mobile subscriber identity (IMSI):** GSM uses the IMSI for internal unique identification of a subscriber. IMSI consists of a mobile country code (MCC), the mobile network code (MNC), and finally the mobile subscriber identification number (MSIN).
- Temporary mobile subscriber identity (TMSI): To hide the IMSI, which would give away the exact identity of the user signalling over the air interface, GSM uses the 4 byte TMSI for local subscriber identification.
- **Mobile station roamingnumber(MSRN)**: Another temporary address that hides the identity and location of a subscriber is MSRN. The VLR generates this address on request from the MSC, and the address is also stored in the HLR. MSRN contains the current visitor country code (VCC), the visitor national destination code (VNDC), the identification of the current MSC together with the subscriber number. The MSRN helps the HLR to find a subscriber for an incoming call.

For a Mobile Terminated Call*(MTC)*, thefollowingfigureshowsthedifferentstepsthattake place:



Step 1: User dials the phone number of a GSM subscriber.

Step 2: The fixed network (PSTN) identifies the number belongs to a user in GSM network and forwards the call setup to the Gateway MSC (GMSC).

Step 3: The GMSC identifies the HLR for the subscriberandsignalsthecallsetupto HLR

Step 4: The HLR checks for number existence and its subscribed services and requests an MSRN from the current VLR.

Step 5: VLR sends the MSRN to HLR

step 6: Upon receiving MSRN, the HLR determines the MSC responsible for MS and forwards the information to the GMSC

Step 7: The GMSC can now forward the call setup request to the MSC indicated

Step 8: The MSC requests the VLR for the current status of the MS

Step 9: VLR sends the requested information

Step 10: If MS is available, the MSC initiates paging in all cells it is responsible for.

Step 11: The BTSs of all BSSs transmit the paging signal to the MS

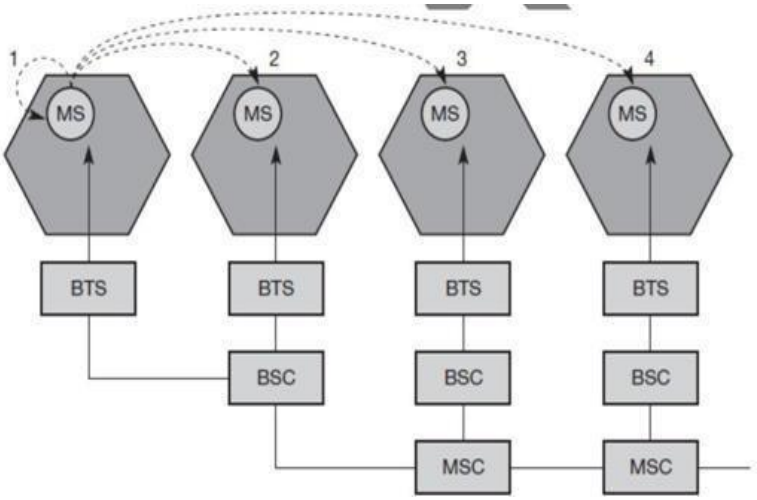
Step 12: Step 13: If MS answers, VLR performs security checks

Step 15: Till step 17: Then the VLR signals to the MSC to setup a connection to the MS

1.7.6. **Handover:** Cellular systems require handover procedures, as single cells do not cover the whole service area. However, a handover should not cause a cut-off, also called call drop.

GSM aims at maximum handover duration of 60 ms. There are two basic reasons for a handover:

1. The mobile station moves out of the range of a BTS, decreasing the received signal level increasing the error rate thereby diminishing the quality of the radio link.
2. Handover may be due to load balancing, when an MSC/BSC decides the traffic is too high in one cell and shifts some MS to other cells with a lower load.

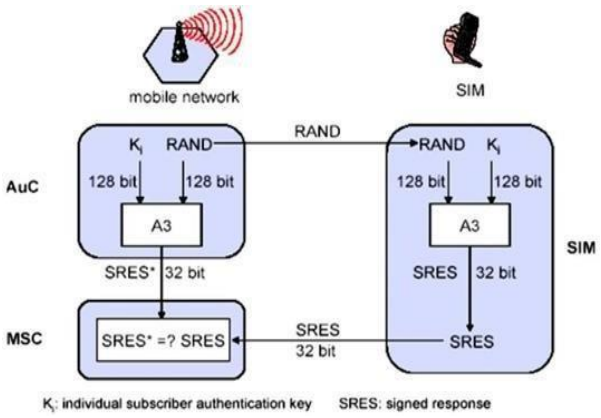


The four possible handover scenarios of GSM are shown below:

- **Intra-cell handover:** Within acell, narrow-band interference could maketransmission at a certain frequency impossible. The BSC could then decide to change the carrier frequency (scenario
- **Inter-cell, intra-BSC handover:** This is a typical handover scenario. The mobile station moves from one cell to another, but stays within the control of the same BSC. The BSC then performs a handover, assignsanewradiochannelinthenewcellandreleasesthe old one.
- **Inter-BSC, intra-MSC handover:** As a BSC only controls a limited number of cells; GSM also has to perform handovers between cells controlled by different BSCs. This handover then has to be controlled by the MSC.
- **Inter MSC handover:** A handover could be required between two cells belonging to different MSCs. Now both MSCs perform the handover together

1.8. **Security**: GSM offers several security services using confidential information stored in the AuC and in the individual SIM. The SIM stores personal, secret data and is protected with a PIN against unauthorized use. Three algorithms have been specified to provide security services in GSM.

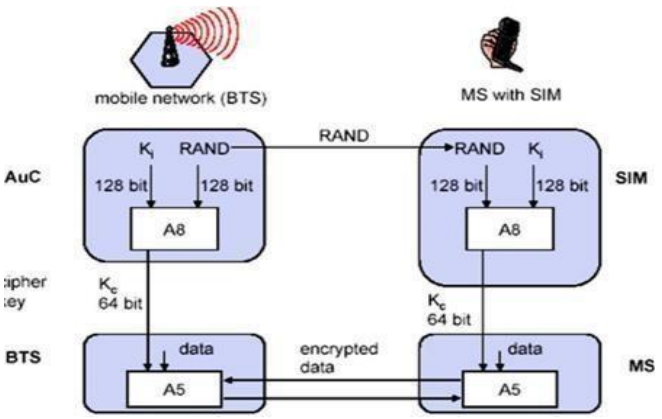
Algorithm A3 is used for **authentication**, **A5** for **encryption**, and **A8** for the **generation of a cipher key**. The various security services offered by GSM are:



Access control and authentication: The first step includes the authentication of a valid user for the
SIM. The user needs a secret PIN to access the SIM. The next step is the subscriber authentication. This step is based on a challenge-response scheme as shown in figure.

Confidentiality: All user-related data is encrypted. After authentication, BTS and MS apply encryption to voice, data, and signaling as show below fig.

Anonymity: To Provide user anonymity, all data is encrypted before transmission, and user identifiers are not used over the air-instead, GSM transmits a temporary identifier (TMSI), which is newly assigned by the VLR after each location update. Additionally, the VLR can change the TMSI at any time.



1.9. **New Data Services:** To enhance the data transmission capabilities of GSM, two basic approaches are possible. As the basic GSM is based on connection-oriented traffic channels, e.g., with 9.6 kbit/s each, several channels could be combined to increase bandwidth. This system is called HSCSD {high speed circuit switched data}. A more progressive step is the introduction of packet- oriented traffic in GSM, i.e., shifting the paradigm from connections/telephone thinking to packets/internet thinking. The system is called GPRS {general packet radio service}.

HSCD: A straightforward improvement of GSM's data transmission capabilities is high speed circuit switched data (HSCSD) in which higher data rates are achieved by bundling several TCHs. An MS requests one or more TCHs from the GSM network, i.e., it allocates several TDMA slots within a TDMA frame. This allocation can be asymmetrical, i.e. more slots can be allocated on the downlink than on the uplink. A major disadvantage of HSCD is that it still uses the connection-oriented mechanisms of GSM, which is not efficient for computer data traffic.

GPRS: The next step of data transmission is GPRS. It also avoids the problems of HSCSD. The **general packet radio service (GPRS)** providespacket modetransfer for applicationsthat exhibit traffic patterns such as frequent transmission of small volumes (e.g., typical web requests) or infrequent transmissions of small or medium volumes (e.g., typical web responses) according to the requirement specification

1.10. **GPRS Overview:** General Packet Radio Service (GPRS) is a Mobile Data Service accessible to GSM and IS-136 mobile phones users. This service is packet-switched and several numbers of users can divide the same transmission channel for transmitting the data.

General Packet Radio System is also known as **GPRS** is a third-generation step toward internet access. GPRS is also known as GSM-IP that is a Global-System Mobile Communications Internet Protocol as it keeps the users of this system online, allows to make voice calls, and access internet on- the-go. Even Time-Division Multiple Access (TDMA) users benefit from this system as it provides packet radio access. GPRS also permits the network operators to execute Internet Protocol (IP) based core architecture for integrated voice and data applications that will continue to be used and expanded for 3G services.

The packet radio principle is employed by GPRS to transport user data packets in a structure way between GSM mobile stations and external packet data networks. These packets can be directly routed to the packet switched

networks from the GPRS mobile stations.

GPRS also permits the network operators to execute Internet Protocol (IP) based core architecture for integrated voice and data applications that will continue to be used and expanded for 3G services.

In the current versions of GPRS, networks based on the Internet Protocol (IP) like the global internet or private/corporate intranets and X.25 networks are supported.

Who owns GPRS ?

The GPRS specifications are written by the European Telecommunications Standard Institute (ETSI), the European counterpart of the American National Standard Institute (ANSI).

Key Features

Following three key features describe wireless packet data:

- **The always online feature -** Removes the dial-up process, making applications only one click away.
- **An upgrade to existing systems -** Operators do not have to replace their equipment; rather, GPRS is added on top of the existing infrastructure.
- **An integral part of future 3G systems -** GPRS is the packet data core network for 3G systems EDGE and WCDMA.

Goals of GPRS

GPRS is the first step toward an end-to-end wireless infrastructure and has the following goals:

- Open architecture
- Consistent IP services
- Same infrastructure for different air interfaces
- Integrated telephony and Internet infrastructure
- Leverage industry investment in IP
- Service innovation independent of infrastructure

Benefits of GPRS

Higher Data Rate: GPRS benefits the users in many ways, one of which is higher data rates in turn of shorter access times.

Easy Billing**:** GPRS packet transmission offers a more user-friendly billing than that offered by circuit switched services. In circuit switched services, billing is based on the duration of the connection. This is unsuitable for applications with bursty traffic. The user must pay for the entire airtime, even for idle periods when no packets are sent (e.g., when the user reads a Web page).

The advantage for the user is that he or she can be "online" over a long period of time but will be billed based on the transmitted data volume.

GPRS Applications:

GPRS has opened a wide range of unique services to the mobile wireless subscriber. Some of the characteristics that have opened a market full of enhanced value services to the users. Below are some of the characteristics:

- **Mobility -** The ability to maintain constant voice and data communications while on the move.
- **Immediacy -** Allows subscribers to obtain connectivity when needed, regardless of location and without a lengthy login session.
- **Localization -** Allows subscribers to obtain information relevant to their current location.

- Using the above three characteristics varied possible applications are being developed to offer to the mobile subscribers. These applications, in general, can be divided into two high-level categories:
- Corporation
- Consumer

These two levels further include:

- **Communications -** E-mail, fax, unified messaging and intranet/internet access, etc.
- **Value-added services -** Information services and games, etc.
- **E-commerce -** Retail, ticket purchasing, banking and financial trading, etc.
- **Location-based applications -** Navigation, traffic conditions, airline/rail schedules and location finder, etc.
- **Vertical applications -** Freight delivery, fleet management and sales-force automation.
- **Advertising -** Advertising may be location sensitive. For example, a user entering a mall can receive advertisements specific to the stores in that mall.

Along with the above applications, non-voice services like SMS, MMS and voice calls are also possible with GPRS. Closed User Group (CUG) is a common term used after GPRS is in the market, in addition, it is planned to implement supplementary services, such as Call Forwarding Unconditional (CFU), and Call Forwarding on Mobile subscriber Not Reachable (CFNRc), and closed user group (CUG).

GPRS Architecture: GPRS architecture works on the same procedure like GSM network, but, has additional entities that allow packet data transmission. This data network overlaps a second- generation GSM network providing packet data transport at the rates from 9.6 to 171 kbps. Along with the packet data transport the GSM network accommodates multiple users to share the same air interface resources concurrently.

Following is the GPRS Architecture diagram:

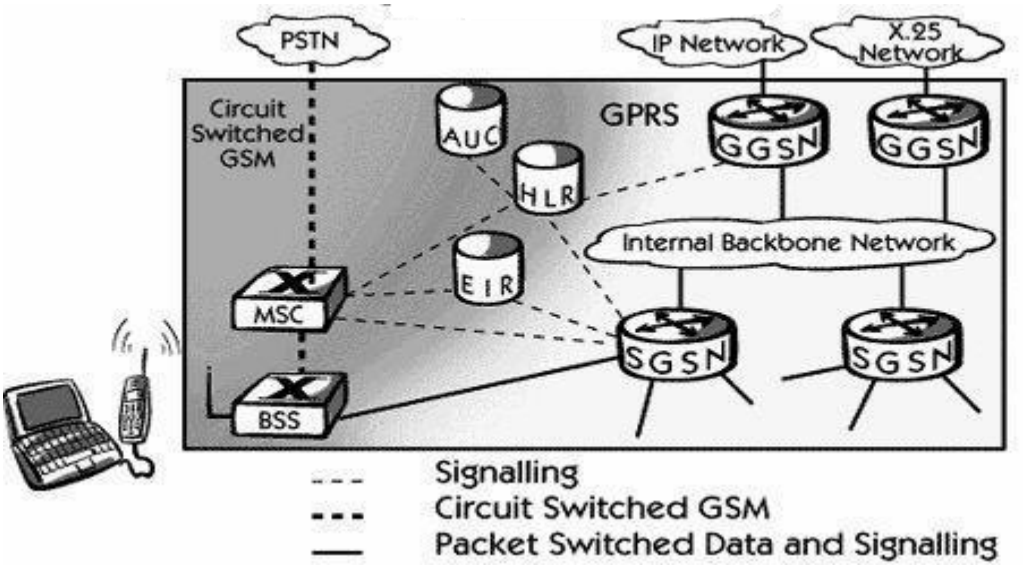


Figure: GPRS Architecture

GPRS Base Station Subsystem

Each BSC requires the installation of one or more Packet Control Units (PCUs) and a software upgrade. The PCU provides a physical and logical data interface to the Base Station Subsystem (BSS) for packet data traffic. The

BTS can also require a software upgrade but typically does not require hardware enhancements.

When either voice or data traffic is originated at the subscriber mobile, it is transported over the air interface to the BTS, and from the BTS to the BSC in the same way as a standard GSM call. However, at the output of the BSC, the traffic is separated; voice is sent to the Mobile Switching Center (MSC) per standard GSM, and data is sent to a new device called the SGSN via the PCU over a Frame Relay interface.

GPRS Support Nodes

Following two new components, called Gateway GPRS Support Nodes (GSNs) and, Serving GPRS Support Node (SGSN) are added:

Gateway GPRS Support Node (GGSN)

The Gateway GPRS Support Node acts as an interface and a router to external networks. It contains routing information for GPRS mobiles, which is used to tunnel packets through the IP based internal backbone to the correct Serving GPRS Support Node. The GGSN also collects charging information connected to the use of the external data networks and can act as a packet filter for incoming traffic.

Serving GPRS Support Node (SGSN)

The Serving GPRS Support Node is responsible for authentication of GPRS mobiles, registration of mobiles in the network, mobility management, and collecting information on charging for the use of the air interface.

Internal Backbone

The internal backbone is an IP based network used to carry packets between different GSNs. Tunnelling is used between SGSNs and GGSNs, so the internal backbone does not need any information about domains outside the GPRS network. Signalling from a GSN to a MSC, HLR or EIR is done using SS7.

Routing Area

GPRS introduces the concept of a Routing Area. This concept is similar to Location Area in GSM, except that it generally contains fewer cells. Because routing areas are smaller than location areas, less radio resources are used while broadcasting a page message.

GPRS Protocol Stack:

The flow of GPRS protocol stack and end-to-end message from MS to the GGSN is displayed in the below diagram. GTP is the protocol used between the SGSN and GGSN using the Gn interface. This is a Layer 3 tunneling protocol.

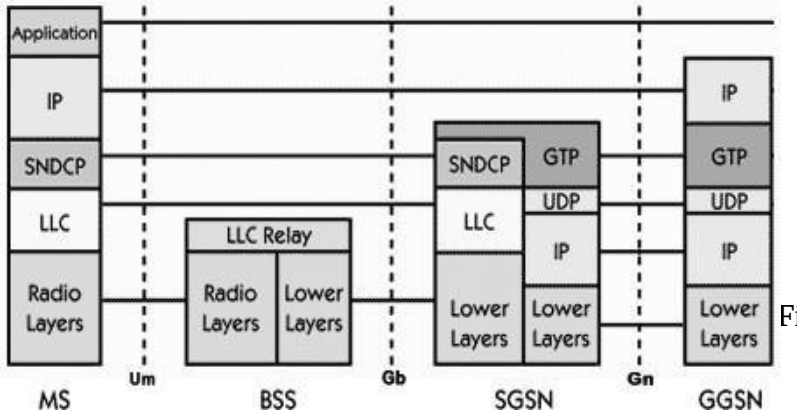


Figure: GPRS Protocol Stack

The process that takes place in the application looks like a normal IP sub-network

for the users both inside and outside the network. The vital thing that needs attention is, the application communicates via standard IP, that is carried through the GPRS network and out through the gateway GPRS. The packets that are mobile between the GGSN and the SGSN use the GPRS tunneling protocol, this way the IP addresses located on the external side of the GPRS network do not have deal with the internal backbone. UDP and IP are run by GTP.

Sub Network Dependent Convergence Protocol (SNDCP) and Logical Link Control (LLC) combination used in between the SGSN and the MS. The SNDCP flattens data to reduce the load on the radio channel. A safe logical link by encrypting packets is provided by LLC and the same LLC link is used as long as a mobile is under a single SGSN.

In case, the mobile moves to a new routing area that lies under a different SGSN; then, the old LLC link is removed and a new link is established with the new Serving GSN X.25. Services are provided by running X.25 on top of TCP/IP in the internal backbone.

APPLICATIONS OF MOBILE COMPUTING In many fields of work, the ability to keep on the move is vital in order to utilise time efficiently. The importance of Mobile Computers has been highlighted in many fields of which a few are described below:

- **Vehicles:** Music, news, road conditions, weather reports, and other broadcast information are received via digital audio broadcasting (DAB) with 1.5 Mbit/s. For personal communication, a universal mobile telecommunications system (UMTS) phone might be available offering voice and data connectivity with 384 kbit/s. The current position of the car is determined via the global positioning system (GPS). Cars driving in the same area build a local ad-hoc network for the fast exchange of information in emergency situations or to help each other keep a safe distance. In case of an accident, not only will the airbag be triggered, but the police and ambulance service will be informed via an emergency call to a service provider. Buses, trucks, and trains are already transmitting maintenance and logistic information to their home base, which helps to improve organization (fleet management), and saves time and money.
- **Emergencies:** An ambulance with a high-quality wireless connection to a hospital can carry vital information about injured persons to the hospital from the scene of the accident. All the necessary steps for this particular type of accident can be prepared and specialists can be consulted for an early diagnosis. Wireless networks are the only means of communication in the case of natural disasters such as hurricanes or earthquakes. In the worst cases, only decentralized, wireless ad-hoc networks survive.
- **Business:** Managers can use mobile computers say, critical presentations to major customers. They can access the latest market share information. They can communicate with the office about possible new offers and call meetings for discussing responds to the new proposals. A travelling salesman today needs instant access to the company's database: to ensure that files on his or her laptop reflect the current situation, to enable the company to keep track of all activities of their travelling employees, to keep databases consistent etc. With wireless access, the laptop can be turned into a true mobile office, but efficient and powerful synchronization mechanisms are needed to ensure data consistency.

- **Replacement of Wired Networks**: wireless networks can also be used to replace wired networks, e.g., remote sensors, for tradeshows, or in historic buildings. Due to economic reasons, it is often impossible to wire remote sensors for weather forecasts, earthquake detection, or to provide environmental information. Wireless connections, e.g., via satellite, can help in this situation. Other examples for wireless networks are computers, sensors, or information displays in historical buildings, where excess cabling may destroy valuable walls or floor.

Infotainment: wireless networks can provide up-to-date information at any appropriate location. The travel guide might tell you something about the history of a building (knowing via GPS, contact to a local base station, or triangulation where you are) downloading information about a concert in the building at the same evening via a local wireless network. Another growing field of wireless network applications lies in entertainment and games to enable, e.g., ad-hoc gaming networks as soon as people meet to play together.

Mobile and Handheld Devices:

Mobile phones are handheld mobile devices with large functionalities. Each mobile device essentially includes a computer. Most Mobile phones are now smart phones which communicate with other phones using a cellular service provider's network. The smartphones has multimedia features, digital music players, iPod's Bluetooth, camera, etc.

Mobile and Handheld devices are using Operating System Software that provides interfaces, perform allocation and management functions, and act as platform for running the increasingly sophisticated software that are created for mobile computing devices. These operating systems are using middleware for enabling the functions of the device.

The OS software for Mobile devices are

1. Windows CE Based Devices
2. Mac OS 4 Based Devices. (eg. Apple iPhone4)
3. Symbian OS Based Devices.

Desktop

Personal Computers

Laptop/Tablet PC

PDA

Personal Digital Assistants (PDAs) and Cell Phones

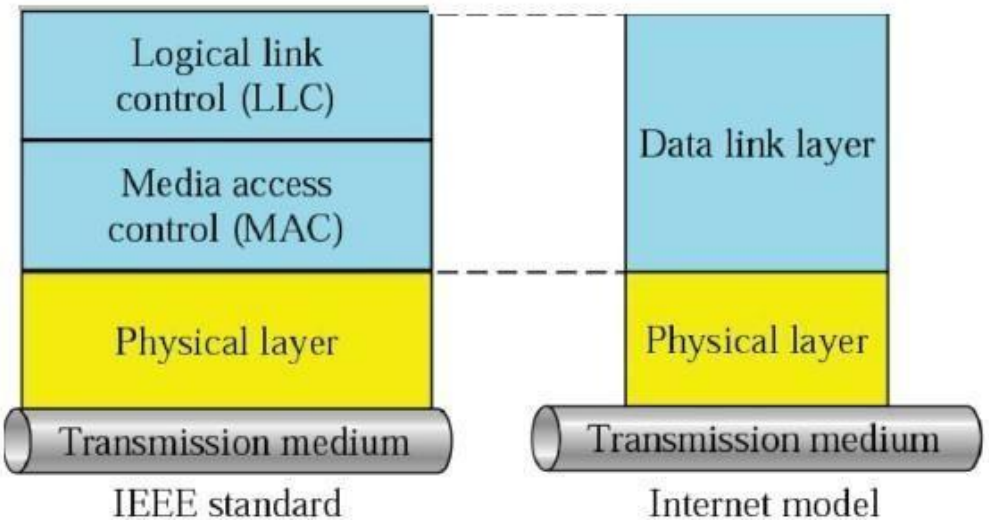
Smartphone

Unit-2

(Wireless) Medium Access Control (MAC): Motivation for a specialized MAC (Hidden and exposed terminals, Near and far terminals), SDMA, FDMA, TDMA, CDMA, Wireless LAN/(IEEE 802.11).

2.0. MAC: The **Media Access Control** (**MAC**) data communication protocol sub-layer, also known as the Medium Access Control, is a sub-layer of the Data Link Layer specified in the seven-layer OSI model (layer 2). The hardware that implements the MAC is referred to as a **Medium Access Controller**. The MAC sub-layer acts as an interface between the Logical Link Control (LLC) sub layer and the network's physical layer. The MAC layer emulates a full-duplex logical communication channel in a multi-point network. This channel may provide unicast, multicast or broadcast communication service.

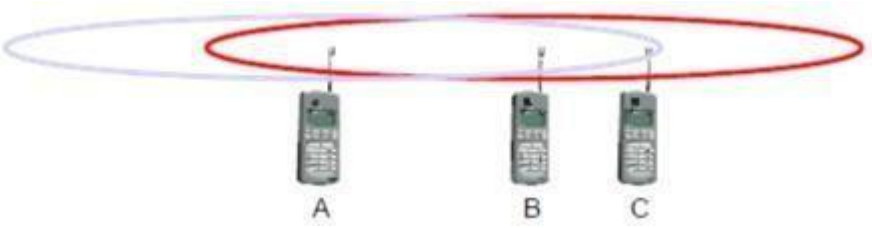
2.1. Motivation for a specialized MAC



One of the most commonly used MAC schemes for wired networks is carrier sense multiple access with collision detection (CSMA/CD). In this scheme, a sender senses the medium (a wire or coaxial cable) to see if it is free. If the medium is busy, the sender waits until it is free. If the medium is free, the sender starts transmitting data and continues to listen into the medium. If the sender detects a collision while sending, it stops at once and sends a jamming signal. But this scheme doest work well with wireless networks. The problems are:

- Signal strength decreases proportional to the square of the distance
- The sender would apply CS and CD, but the collisions happen at the receiver
- It might be a case that a sender cannot “hear” the collision, i.e., CD does not work Furthermore, C“ might not work, if for e.g., a terminal is “hidden”

2.1.1. Hidden and Exposed Terminals: Consider the scenario with three mobile phones as shown below. The transmission range of A reaches B, but not C (the detection range does not reach C either). The transmission range of C reaches B, but not A. Finally, the transmission range of B reaches A and C, i.e., A cannot detect C and vice versa.



Hidden terminals

- A sends to B, C cannot receive A
- C wants to send to B, C senses a "free" medium (C" fails) and starts transmitting
- Collision at B occurs, A cannot detect this collision (CD fails) and continues with its transmission to B
- A is "hidden" from C and vice versa

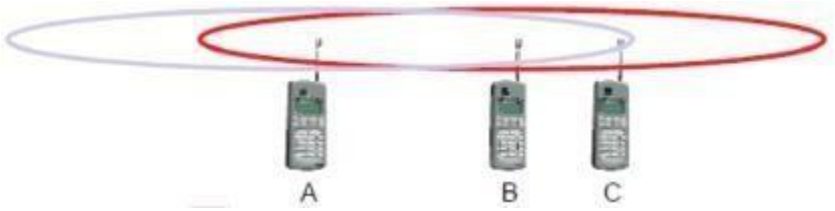
Exposed terminals

- B sends to A, C wants to send to another terminal (not A or B) outside the range
- C senses the carrier and detects that the carrier is busy.
- C postpones its transmission until it detects the medium as being idle again
- but A is outside radio range of C, waiting is **not** necessary
- C is "exposed" to B

Hidden terminals cause collisions, where as Exposed terminals causes unnecessary delay.

2.1.2. Near and far terminals

Consider the situation shown below. A and B are both sending with the same transmission power.



- Signal strength decreases proportional to the square of the distance
- "o, B's signal drowns out A's signal making C unable to receive A's transmission
- If C is an arbiter for sending rights, B drown out A's signal on the physical layer making C unable to hear out A.

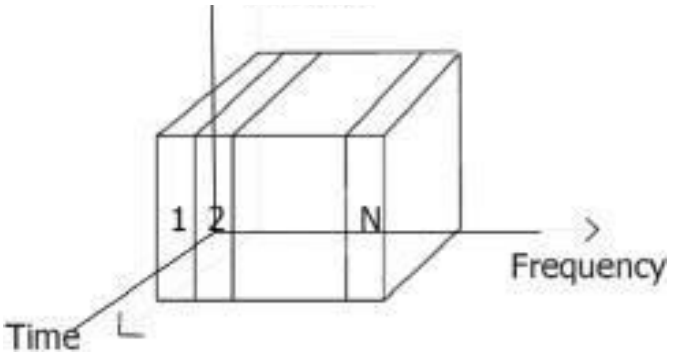
The **near/far effect** is a severe problem of wireless networks using CDM. All signals should arrive at the receiver with more or less the same strength for which Precise power control is to be implemented.

2.2. SDMA: **Space Division Multiple Access** is used for allocating a separated space to users in wireless networks. No of application are assigning on **base station** to a mobile phone user. The mobile phone may receive several base stations with different quality.

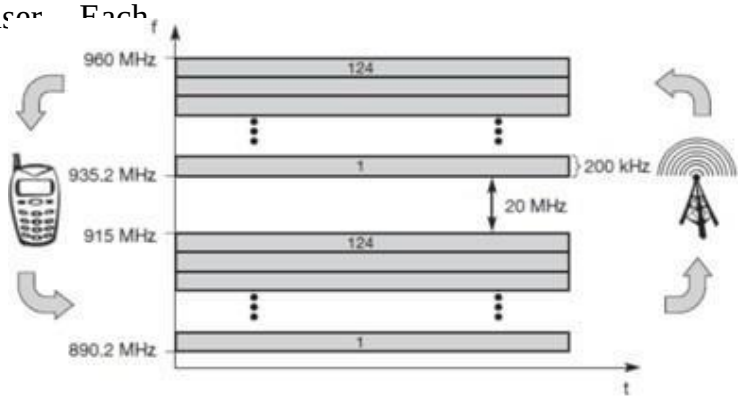
A MAC algorithm can decide which base station is best, taking into account which frequencies (FDM), time slots (TDM) or code (CDM) are still available. The SDMA algorithm is formed by cells and sectorized antennas which constitute the infrastructure implementing **space division multiplexing (SDM).** SDM has the unique advantage of not requiring any multiplexing equipment. It is usually combined with other multiplexing techniques to better utilize the

individual physical channels.

2.3. FDMA: Frequency division multiplexing (FDM) describes schemes to subdivide the frequency dimension into several non-overlapping frequency bands.



requency Division Multiple Access is a method employed to permit several users to transmit simultaneously on one satellite transponder by assigning a specific frequency within the channel to each user. Each conversation gets its own, unique, radio channel. The channels are relatively narrow, usually 30 KHz or less and are defined as either transmit or receive channels. A full duplex conversation requires a transmit & receive channel pair. FDM is often used for simultaneous access to the medium by base station and mobile station in cellular networks establishing a duplex channel. A scheme called frequency division duplexing (FDD) in which the two directions, mobile station to base station and vice versa are now separated using different frequencies**.**



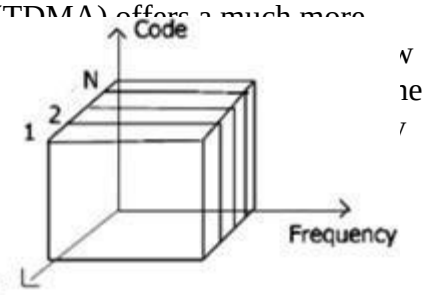
FDM for mUltiple access and DUplex

The two frequencies are also known as **uplink**, i.e., from mobile station to base station or from ground control to satellite, and as **downlink**, i.e., from base station to mobile station or from satellite to ground control. The basic frequency allocation scheme for GSM is fixed and regulated by national authorities. All uplinks use the band between 890.2 and 915 MHz, all downlinks use 935.2 to 960 MHz. According to FDMA, the base station, shown on the right side, allocates a certain frequency for up- and downlink to establish a duplex channel with a mobile phone. Up- and downlink have a fixed relation. If the uplink frequency is fu = 890 MHz + n·0.2 MHz, the downlink frequency is fd = fu + 45 MHz,

i.e**., fd = 935 MHz + n·0.2 MHz** for a certain channel n. The base station selects the channel. Each channel (uplink and downlink) has a bandwidth of 200 kHz.

This scheme also has disadvantages. While radio stations broadcast 24 hours a day, mobile communication typically takes place for only a few minutes at a time. Assigning a separate frequency for each possible communication scenario would be a tremendous waste of (scarce) frequency resources. Additionally, the fixed assignment of a frequency to a sender makes the scheme very inflexible and limits the number of senders.

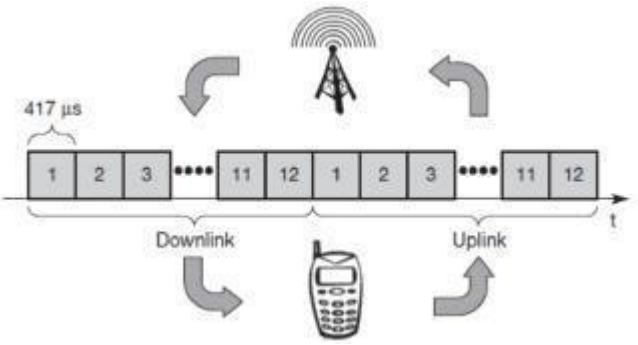
TDMA: A more flexible multiplexing scheme for typical mobile communications is time division multiplexing (TDM). Compared to FDMA, time division multiple acc



flexible scheme, which comprises all technologies that allocate certain
synchronization between sender and receiver has to be achieved in the
by using a fixed pattern similar to FDMA techniques, i.e., allocating a
using a dynamic allocation scheme.

Listening to different frequencies at the same time is quite
channels separated in time at the same frequency is simpl
identification, but are not as flexible considering varying bandwidth requirements.

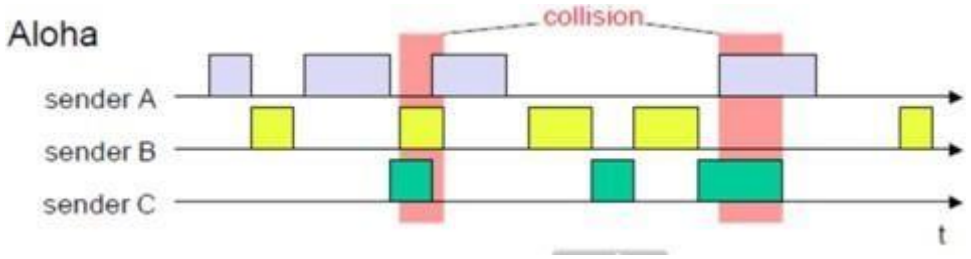
2.4. Fixed TDM: The simplest algorithm for using TDM is allocating time slots for channels in a fixed pattern. This results in a fixed bandwidth and is the typical solution for wireless phone systems. MAC is quite simple, as the only crucial factor is accessing the reserved time slot at the right moment. If this synchronization is assured, each mobile station knows its turn and no interference will happen. The fixed pattern can be assigned by the base station, where competition between different mobile stations that want to access the medium is solved.



The figure shows how these fixed TDM patterns are
used to implement multiple access and a duplex channel between a base station and mobile station. Assigning different slots for uplink and downlink using the same frequency is called **time division duplex (TDD)**. As shown in the figure, the base station uses one out of 12 slots for the downlink, whereas the mobile station uses one out of 12 different slots for the uplink. Uplink and downlink are separated in time. Up to 12 different mobile stations can use the same frequency without interference using this scheme. Each connection is allotted its own up- and downlink pair. This general scheme still wastes a lot of bandwidth. It is too static, too inflexible for data communication. In this case, connectionless, demand- oriented TDMA schemes can be used.

2.5. Classical Aloha

In this scheme, TDM is applied without controlling medium access. Here each station can access the medium at any time as shown below:

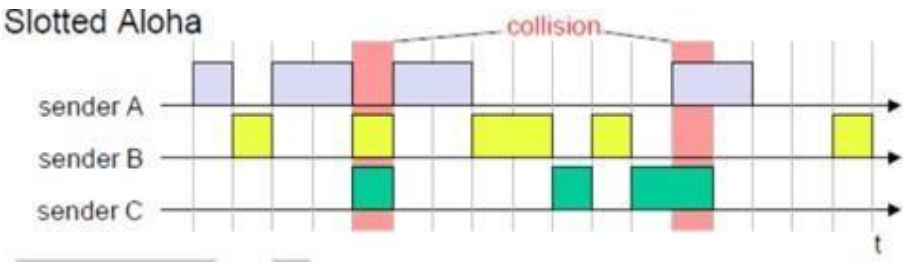


This is a random access scheme, without a central arbiter controlling access and without coordination among the stations. If two or more stations access the medium at the same time, a **collision** occurs and

the transmitted data is destroyed. Resolving this problem is left to higher layers (e.g., retransmission of data). The simple Aloha works fine for a light load and does not require any complicated access mechanisms.

2.6. Slotted Aloha

The first refinement of the classical Aloha scheme is provided by the introduction of time slots (**slotted Aloha**). In this case, all senders have to be **synchronized**, transmission can only start at the beginning of a **time slot** as shown below.



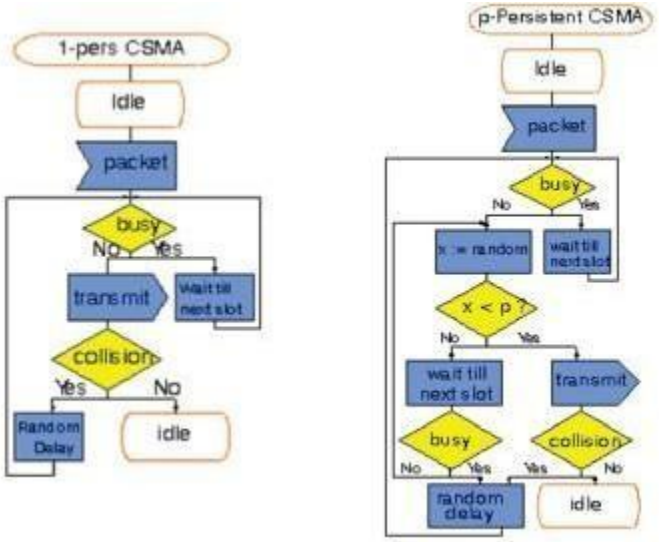
The introduction of slots raises the throughput from 18 per cent to 36 per cent, i.e., slotting doubles the throughput. Both basic Aloha principles occur in many systems that implement distributed access to a medium. Aloha systems work perfectly well under a light load, but they cannot give any hard transmission guarantees, such as maximum delay before accessing the medium or minimum throughput.

2.7. Carrier sense multiple access: One improvement to the basic Aloha is sensing the carrier before accessing the medium. Sensing the carrier and accessing the medium only if the carrier is idle decreases the probability of collision. But, as already mentioned in the introduction, hidden terminals cannot be detected, so, if a hidden terminal transmits at the same time as another sender, a collision might occur at the receiver. This basic scheme is still used in most wireless LANs. The different versions of CSMA are:

- **1-persistent CSMA**: Stations sense the channel and listens if its busy and transmit immediately, when the channel becomes idle. It's called 1-persistent CSMA because the host transmits with a probability of 1 whenever it finds the channel idle.
- **Non-persistent CSMA**: stations sense the carrier and start sending immediately if the medium is idle. If the medium is busy, the station pauses a random amount of time before sensing the medium again and repeating this pattern.
- **p-persistent CSMA**: systems nodes also sense the medium, but only transmit with a probability

of p, with the station deferring to the next slot with the probability 1-p, i.e., access is slotted in addition

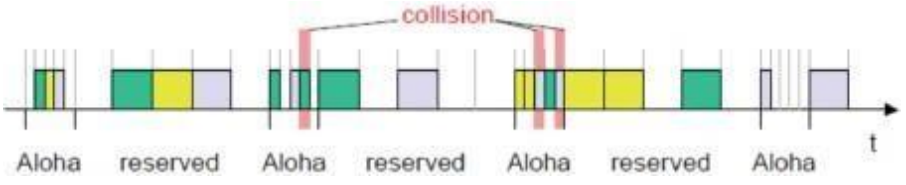
CSMA with collision avoidance (**CSMA/CA**) is one of the access schemes used in wireless LANs following the standard IEEE 802.11. Here sensing the carrier is combined with a back-off scheme in case of a busy medium to achieve some fairness among competing stations.



2.8. Demand assigned multiple access: Channel efficiency for Aloha is 18% and for slotted Aloha is 36%. It can be increased to 80% by implementing reservation mechanisms and combinations with some (fixed) TDM patterns. These schemes typically have a reservation period followed by a transmission period. During the reservation period, stations can reserve future slots in the transmission period. While, depending on the scheme, collisions may occur during the reservation period, the transmission period can then be reservation period, stations can reserve future slots in the transmission period. While, depending on the scheme, collisions may occur during the reservation period, the transmission period can then be accessed without collision.

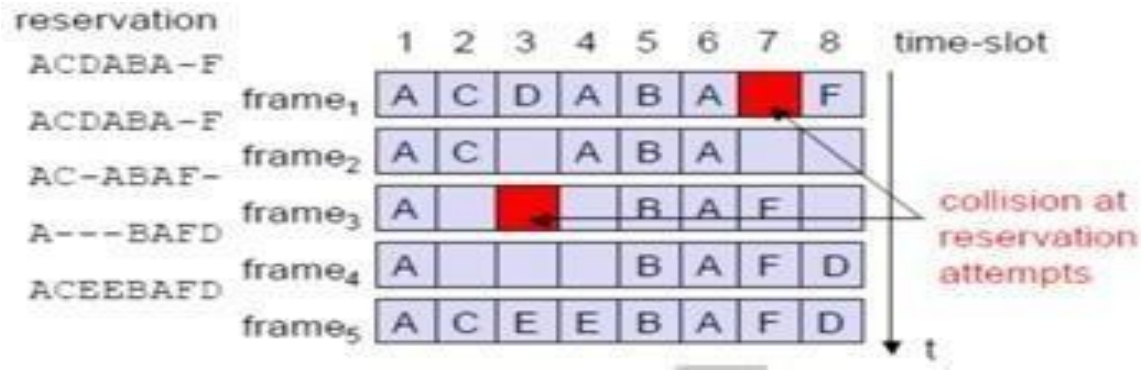
One basic scheme is **demand assigned multiple access (DAMA)** also called **reservation Aloha**, a scheme typical for satellite systems. It increases the amount of users in a pool of satellite channels that are available for use by any station in a network. It is assumed that not all users will need simultaneous access to the same communication channels. So that a call can be established, DAMA assigns a pair of available channels based on requests issued from a user. Once the call is completed, the channels are returned to the pool for an assignment to another call. Since the resources of the satellite are being used only in proportion to the occupied channels for the time in which they are being held, it is a perfect environment for voice traffic and data traffic in batch mode.

It has two modes as shown below.



During a contention phase following the slotted Aloha scheme; all stations can try to reserve future slots. Collisions during the reservation phase do not destroy data transmission, but only the short requests for data transmission. If successful, a time slot in the future is reserved, and no other station is allowed to transmit during this slot. Therefore, the satellite collects all successful requests (the others are destroyed) and sends back a reservation list indicating access rights for future slots. All ground stations have to obey this list. To maintain the fixed TDM pattern of reservation and transmission, the stations have to be synchronized from time to time. DAMA is an **explicit reservation** scheme. Each transmission slot has to be reserved explicitly.

PRMA packet reservation multiple access: It is a kind of implicit reservation scheme where, slots can be reserved implicitly. A certain number of slots form a frame. The frame is repeated in time i.e., a fixed TDM pattern is applied. A base station, which could be a satellite, now broadcasts the status of each slot to all mobile stations. All stations receiving this vector will then know which slot is occupied and which slot is currently free.



The

base station broadcasts the reservation status ‘ACDABA-F’ to all stations, here A to F. This means that slots one to six and eight are occupied, but slot seven is free in the following transmission. All stations wishing to transmit can now compete for this free slot in Aloha fashion. The already occupied slots are not touched. In the example shown, more than one station wants to access this slot, so a

collision occurs. The base station returns the reservation status ‘ACDABA-F’, indicating that the reservation of slot seven failed (still indicated as free) and that nothing has changed for the other slots. Again, stations can compete for this slot. Additionally, station D has stopped sending in slot three and station F in slot eight. This is noticed by the base station after the second frame. Before the third frame starts, the base station indicates that slots three and eight are now idle. Station F has succeeded in reserving slot seven as also indicated by the base station.

As soon as a station has succeeded with a reservation, all future slots are implicitly reserved for this station. This ensures transmission with a guaranteed data rate. The slotted aloha scheme is used for idle slots only; data transmission is not destroyed by collision.

2.9. Reservation TDMA: In a fixed TDM scheme N mini-slots followed by N·k data-slots

form a frame that is repeated. Each station is allotted its own mini-slot and can use it to reserve up to k data-slots.

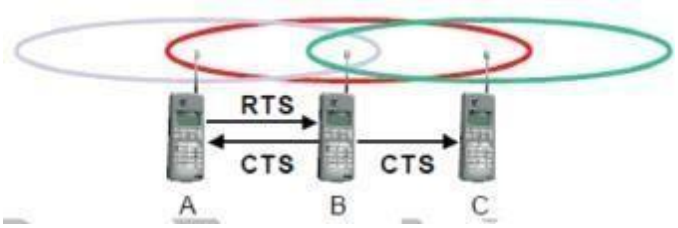
This guarantees each station a certain bandwidth and a fixed delay. Other stations can now send data in unused data-slots as shown. Using these free slots can be based on a simple round-robin scheme or can be uncoordinated using an Aloha scheme. This scheme allows for the combination of, e.g., isochronous traffic with fixed bitrates and best-effort traffic without any guarantees.

Multiple access with collision avoidance

Multiple access with collision avoidance (MACA) presents a simple scheme that solves the hidden terminal problem, does not need a base station, and is still a random access Aloha scheme – but with dynamic reservation. Consider the hidden terminal problem scenario.

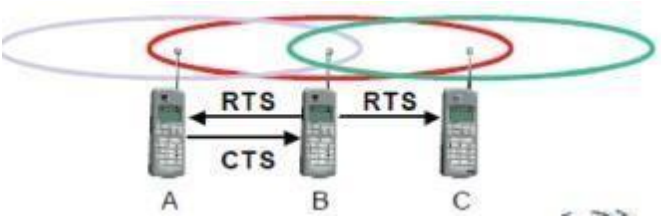
A starts sending to B, C does not receive this transmission. C also wants to send something to B and senses the medium. The medium appears to be free, the carrier sense fails. C also starts sending causing a collision at B. But A cannot detect this collision at B and continues with its transmission. A is **hidden** for C and vice versa.

With MACA, A does not start its transmission at once, but sends a **request to send (RTS)** first. B receives the RTS that contains the name of sender and receiver, as well as the length of the future transmission. This RTS is not heard by C, but triggers an acknowledgement from B, called **clear to send (CTS)**. The CTS again contains the names of sender (A) and receiver (B) of the user data, and the length of the future transmission.



This CTS is now heard by C and the medium for future use by A is now reserved for the duration of the transmission. After receiving a CTS, C is not allowed to send anything for the duration indicated in the CTS toward B. A collision cannot occur at B during data transmission, and the hidden terminal problem is solved. Still collisions might occur when A and C transmits a RTS at the same time. B resolves this contention and acknowledges only one station in the CTS. No transmission is allowed without appropriate CTS.

Now MACA tries to avoid the **exposed terminals** in the following way:



With MACA, B has to transmit an RTS first containing the name of the receiver (A) and the sender (B). C does not react to this message as it is not the receiver, but A acknowledges using a CTS which identifies B as the sender and A as the receiver of the following data transmission. C does not receive this CTS and concludes that A is outside the detection range. C can start its transmission assuming it will not cause a collision at A. The problem with exposed terminals is solved without fixed access patterns or a base station.

2.10. Polling: Polling schemes are used when one station wants to be heard by others. Polling is a strictly centralized scheme with one master station and several slave stations. The master can poll the slaves according to many schemes: round robin (only efficient if traffic patterns are similar over all stations), randomly, according to reservations (the classroom example with polite students) etc. The master could also establish a list of stations wishing to transmit during a contention phase. After this phase, the station polls each station on the list.

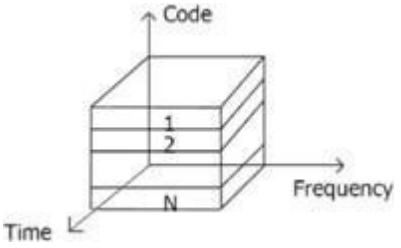
Base station signals readiness to all mobile terminals

- terminals ready to send transmit random number without collision using CDMA or FDMA
- the base station chooses one address for polling from list of all random numbers (collision if two terminals choose the same address)
- the base station acknowledges correct packets and continues polling the next terminal
- this cycle starts again after polling all terminals of the list

Inhibit sense multiple access: This scheme, which is used for the packet data transmission service Cellular Digital Packet Data (CDPD) in the AMPS mobile phone system, is also known as **digital sense multiple access (DSMA)**. Here, the base station only signals a busy medium via a busy tone (called BUSY/IDLE indicator) on the downlink.

After the busy tone stops, accessing the uplink is not coordinated any further. The base station acknowledges successful transmissions; a mobile station detects a collision only via the missing positive acknowledgement. In case of collisions, additional back-off and retransmission mechanisms are implemented.

2.11. CDMA: Code division multiple access systems apply codes with certain characteristics to the transmission to separate different users in code space and to enable access to a shared medium without interference.



All terminals send on the same frequency probably at the same time and can use the whole bandwidth of the transmission channel. Each sender has a unique random number, the sender XO‘s the signal with this random number. The receiver can “tune” into this signal if it knows the pseudo random number, tuning is done via a correlation function

Disadvantages:

- higher complexity of a receiver (receiver cannot just listen into the medium and start receiving if there is a signal)
- all signals should have the same strength at a receiver

Advantages:

- all terminals can use the same frequency, no planning needed
- huge code space(eg. 2^{32}) compared to frequency space.
- interferences (e.g. white noise) is not coded
- forward error correction and encryption can be easily integrated.

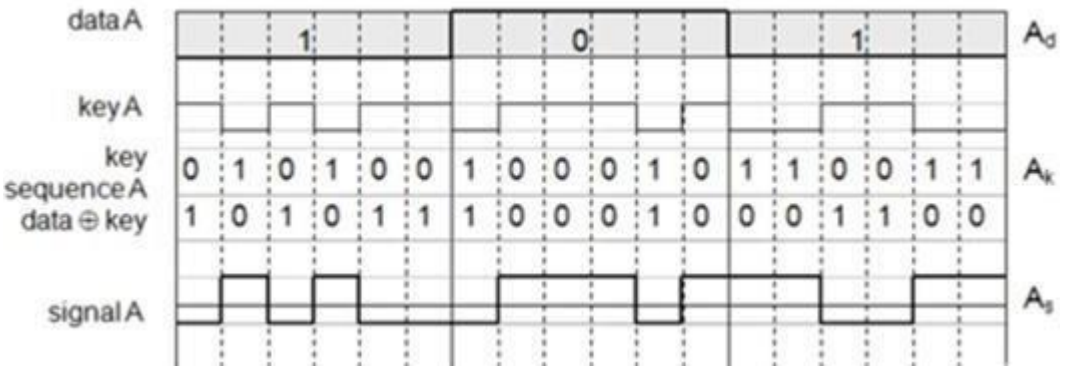
- Sender A wants to transmit the bits 010011.
  - sends $A_d = 1$, key $A_k = 010011$ (assign: "0"= -1, "1"= +1)
  - sending signal $A_s = A_d * A_k = (-1, +1, -1, -1, +1, +1)$
- Sender B wants to transmit the bits 110101
  - sends $B_d = 0$, key $B_k = 110101$ (assign: "0"= -1, "1"= +1)
  - sending signal $B_s = B_d * B_k = (-1, -1, +1, -1, +1, -1)$
- Both signals superimpose in space as
  - $A_s + B_s = (-2, 0, 0, -2, +2, 0)$
- Receiver wants to receive signal from sender A
  - apply key A_k bitwise (inner product)
    - $A_e = (-2, 0, 0, -2, +2, 0)$

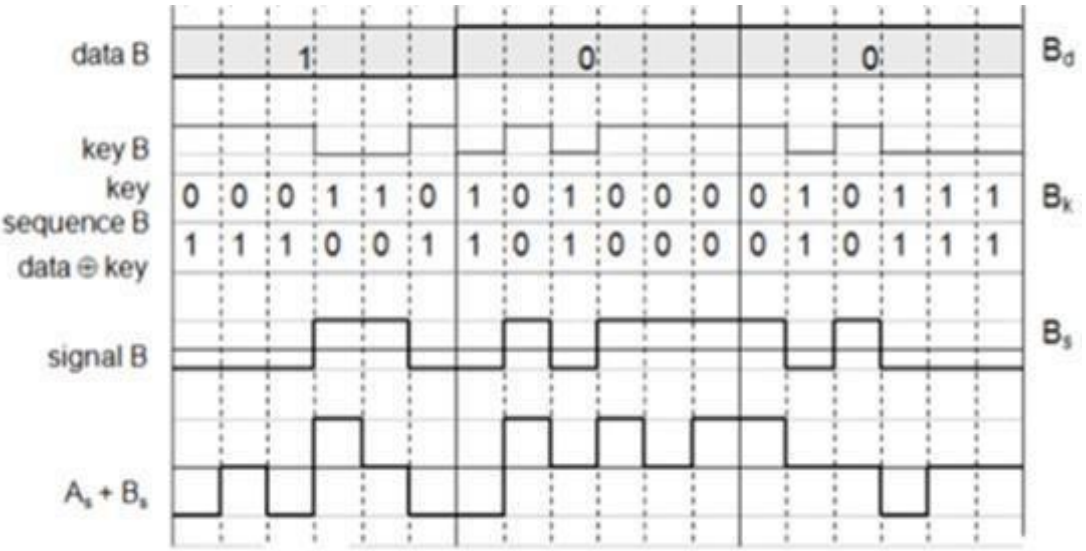
$B_e = (-2, 0, 0, -2, +2, 0)$

- $B_k = -2 + 0 + 0 - 2 - 2 + 0 = -6$, i.e. "0"

The following figure shows a sender A that wants to transmit the bits 101. The key of A is shown as signal and binary sequence A_k. The binary "0" is assigned a positive signal value, the binary "1" a negative signal value. After spreading, i.e., XORing A_d and A_k, the resulting signal is A_s.

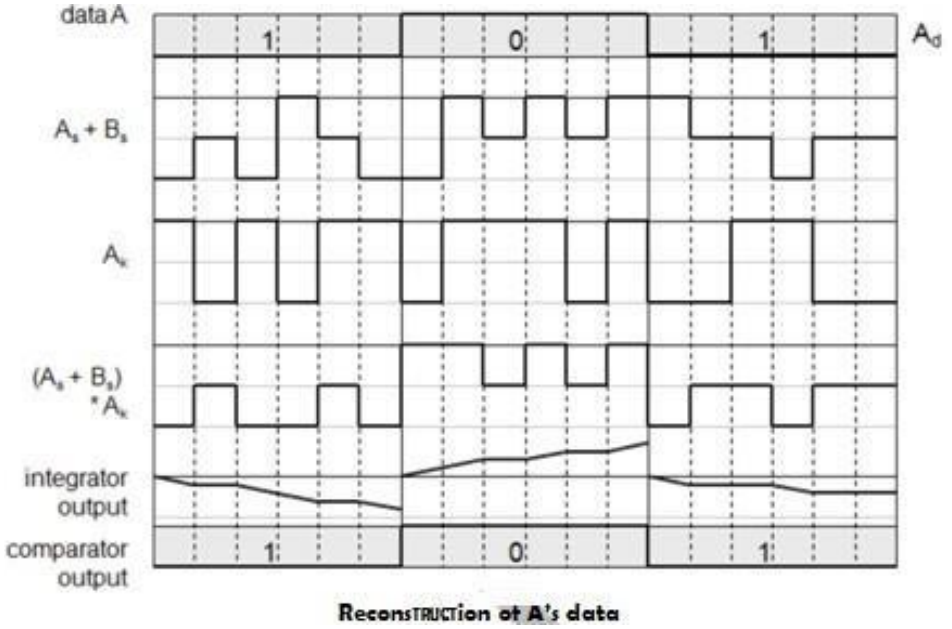
.





Coding and spreading of data from sender A and sender B

The same happens with data from sender B with bits 100. The result is B_S. A_S and B_S now superimpose during transmission. The resulting signal is simply the sum $A_S + B_S$ as shown above. A now tries to reconstruct the original data from Ad. The receiver applies A's key, Ak, to the received signal and feeds the result into an integrator. The integrator adds the products, a comparator then has to decide if the result is a 0 or a 1 as shown below. As clearly seen, although the original signal form is distorted by B's signal, the result is quite clear. The same happens if a receiver wants to receive B's data.



Reconstruction of A's data

Soft handover or **soft handoff** refers to a feature used by the CDMA and WCDMA standards, where a cell phone is simultaneously connected to two or more cells (or cell sectors) during a call. If the sectors are from the same physical cell site (a sectorised site), it is referred to as **softer handoff**. This technique is a form of mobile-assisted handover, for IS-95/CDMA2000 CDMA cell phones continuously make power measurements of a list of neighboring cell sites, and determine

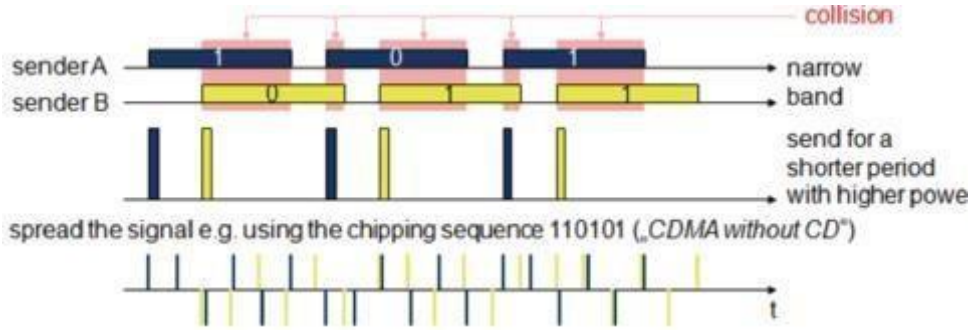
whether or not to request or end soft handover with the cell sectors on the list.

Soft handoff is different from the traditional hard-handoff process. With hard handoff, a definite decision is made on whether to hand off or not. The handoff is initiated and executed without the user attempting to have simultaneous traffic channel communications with the two base stations. With soft handoff, a c*onditional* decision is made on whether to hand off. Depending on the changes in pilot signal strength from the two or more base stations involved, a hard decision will eventually be made to communicate with only one. This normally happens after it is evident that the signal from one base station is considerably stronger than those from the others. In the interim period, the user has simultaneous traffic channel communication with all candidate base stations. It is desirable to implement soft handoff in power- controlled CDMA systems because implementing hard handoff is potentially difficult in such systems..

2.12. Spread Aloha multiple access (SAMA)

CDMA senders and receivers are not really simple devices. Communicating with *n* devices requires programming of the receiver to be able to decode *n* different codes. Aloha was a very simple scheme, but could only provide a relatively low bandwidth due to collisions. SAMA uses spread spectrum with only one single code (chipping sequence) for spreading for all senders accessing according to aloha.

In SAMA, each sender uses the same spreading code, for ex 110101 as shown below. Sender A and B access the medium at the same time in their narrowband spectrum, so that the three bits



The main problem in using this approach is finding good chipping sequences. The maximum throughput is about 18 per cent, which is very similar to Aloha, but the approach benefits from the advantages of spread spectrum techniques: robustness against narrowband interference and simple coexistence with other systems in the same frequency bands.

2.13. Wireless LAN/(IEEE 802.11)

The global goal of WLANs is to replace office cabling, to enable tether less access to the internet and, to introduce a higher flexibility for ad-hoc communication in, e.g., group meetings. **Advantages**

- **Flexibility:** Within radio coverage, nodes can communicate without further restriction. Radio waves can penetrate walls, senders and receivers can be placed anywhere (also non-visible, e.g., within devices, in walls etc.).

- **Planning:** Only wireless ad-hoc networks allow for communication without previous planning, any wired network needs wiring plans. As long as devices follow the same standard, they can communicate
- **Design:** Wireless networks allow for the design of small, independent devices which can for example be put into a pocket. Cables not only restrict users but also designers of small PDAs, notepads etc.
- **Robustness:** Wireless networks can survive disasters, e.g., earthquakes or users pulling a plug. If the wireless devices survive, people can still communicate. Networks requiring a wired infrastructure will usually break down completely.
- **Cost:** After providing wireless access to the infrastructure via an access point for the first user, adding additional users to a wireless network will not increase the cost. This is, important for e.g., lecture halls, hotel lobbies or gate areas in airports where the numbers using the network may vary significantly.

Disadvantages:

- **Quality of service:** WLANs typically offer lower quality than their wired counterparts. The main reasons for this are the lower bandwidth due to limitations in radio transmission (e.g., only 1–10 Mbit/s user data rate instead of 100–1,000 Mbit/s), higher error rates due to interference (e.g., 10–4 instead of 10–12 for fiber optics), and higher delay/delay variation due to extensive error correction and detection mechanisms.
- **Proprietary solutions:** Due to slow standardization procedures, many companies have come up with proprietary solutions offering standardized functionality plus many enhanced features (typically a higher bit rate using a patented coding technology or special inter-access point protocols).
- **Restrictions:** All wireless products have to comply with national regulations. Several government and non-government institutions worldwide regulate the operation and restrict frequencies to minimize interference.
- **Safety and security:** Using radio waves for data transmission might interfere with other high-tech

- **Global operation:** WLAN products should sell in all countries so, national and international frequency regulations have to be considered.
- **Low power:** Devices communicating via a WLAN are typically also wireless devices running on battery power. The LAN design should take this into account and implement special power-saving modes and power management functions.
- **License-free operation:** LAN operators do not want to apply for a special license to be able to use the product. The equipment must operate in a license-free band, such as the 2.4 GHz ISM band.
- **Robust transmission technology:** Compared to their wired counterparts, WLANs operate under difficult conditions. If they use radio transmission, many other electrical devices can interfere with them (vacuum cleaners, hairdryers, train engines etc.).
- **Simplified spontaneous cooperation:** To be useful in practice, WLANs should not require complicated setup routines but should operate spontaneously after power-up. These LANs would not be useful for supporting, e.g., ad-hoc meetings.
- **Easy to use:** In contrast to huge and complex wireless WANs, wireless LANs are made for simple use. They should not require complex management, but rather work on a plug-and-play basis.
- **Protection of investment:** A lot of money has already been invested into ,wired LANs. The new WLANs should protect this investment by being interoperable with the existing networks.
- **Safety and security:** Wireless LANs should be safe to operate, especially regarding low radiation if used, e.g., in hospitals. Users cannot keep safety distances to antennas.
- **Transparency for applications:** Existing applications should continue to run over WLANs, the only

difference being higher delay and lower bandwidth. The fact of wireless access and mobility should be hidden if it is not relevant, but the network should also support location aware applications, e.g., by providing location information.

IEEE 802.11

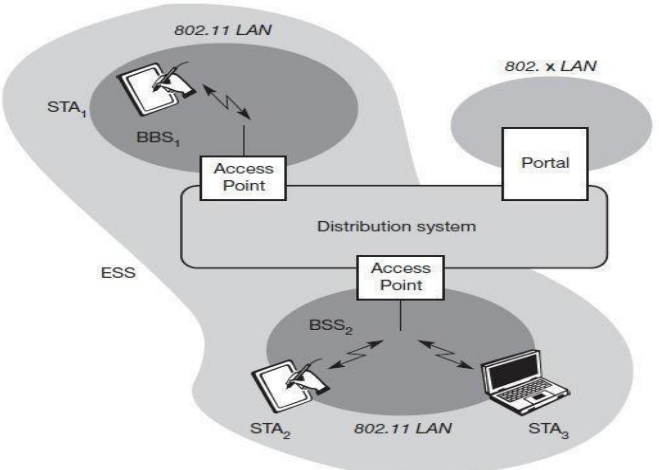
The IEEE standard 802.11 (IEEE, 1999) specifies the most famous family of WLANs in which many products are available. As the standard's number indicates, this standard belongs to the group of 802.x LAN standards, e.g., 802.3 Ethernet or 802.5 Token Ring. This means that the standard specifies the physical and medium access layer adapted to the special requirements of wireless LANs, but offers the same interface as the others to higher layers to maintain interoperability.

The primary goal of the standard was the specification of a simple and robust WLAN which offers time-bounded and asynchronous services. The MAC layer should be able to operate with multiple physical layers, each of which exhibits a different medium sense and transmission characteristic. Candidates for physical layers were infra red and spread spectrum radio transmission techniques.

Additional features of the WLAN should include the support of power management to save battery power, the handling of hidden nodes, and the ability to operate worldwide. The 2.4 GHz ISM band, which is available in most countries around the world, was chosen for the original standard. Data rates envisaged forth standard were 1 Mbit/s mandatory and 2 Mbit/s optional.

The following sections will introduce the system and protocol architecture of the initial IEEE 802.11 and then discuss each layer, i.e., physical layer and medium access. After that, the complex and very important management functions of the standard are presented. Finally, this subsection presents the enhancements of the original standard for higher data rates, 802.11a (up

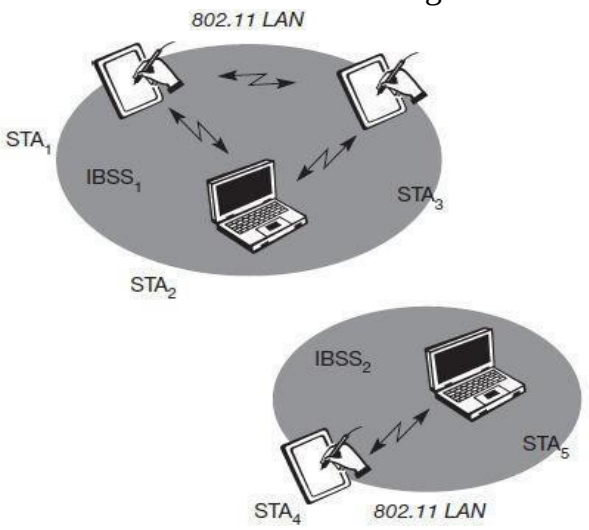
Wireless networks can exhibit two different basic system architectures as shown in infrastructure- based or ad-hoc. Figure shows the components as specified for IEEE 802.11. Several nodes, called **stations (STAi)**, are connected to **access points (AP)**. Stations are terminals with access mechanisms to the wireless medium and radio contact to the AP. The stations and the AP which are within the same radio coverage form a **basic service set (BSSi)**. The example shows two BSSs – BSS1 and BSS2 – which are connected via a **distribution system**. ***Figure:*** *Architecture of an infrastructure-*



A distribution system connects several BSSs via the AP to form a single network and thereby extends the wireless coverage area. This network is now called an **extended service set (ESS)** and has its own identifier, the ESSID. The ESSID is the 'name' of a network and is used to separate different networks. Without knowing the ESSID (and assuming no hacking) it should not be possible to participate in the WLAN. The distribution system connects the wireless networks via the APs with a **portal**, which forms the interworking unit to other LANs. The architecture of the distribution system is not specified further in IEEE 802.11. It could

consist of bridged IEEE LANs, wireless links, or any other networks. However, **distribution system services** are defined in the standard (although, many products today cannot interoperate and needs the additional standard IEEE 802.11f to specify an inter access point protocol. Stations can select an AP and associate with it. The APs support roaming (i.e., changing access points), the distribution system handles data transfer between the different APs. APs provide synchronization within a BSS, support power management, and can control medium access to support time-bounded service. These and further functions are explained in the following sections.

In addition to infrastructure-based networks, IEEE 802.11 allows the building of ad-hoc networks between stations, thus forming one or more independent BSSs (IBSS) as **shown in Figure**. In this case, an IBSS comprises a group of stations using the same radio frequency. Stations STA1, STA2, and STA3 are in IBSS1, STA4 and STA5 in IBSS2. This means for example that STA3 can communicate directly with STA2 but not with STA5. Several IBSSs can either be formed via the distance between the IBSSs or by using different carrier frequencies (then the IBSSs could overlap physically). IEEE 802.11 does not specify any special nodes that support routing, forwarding of data or exchange of topology information as, e.g., HIPERLAN 1 or Bluetooth.



Protocol architecture:

Figure:

Architecture of EEE 802.11 ad-hoc wireless LANs

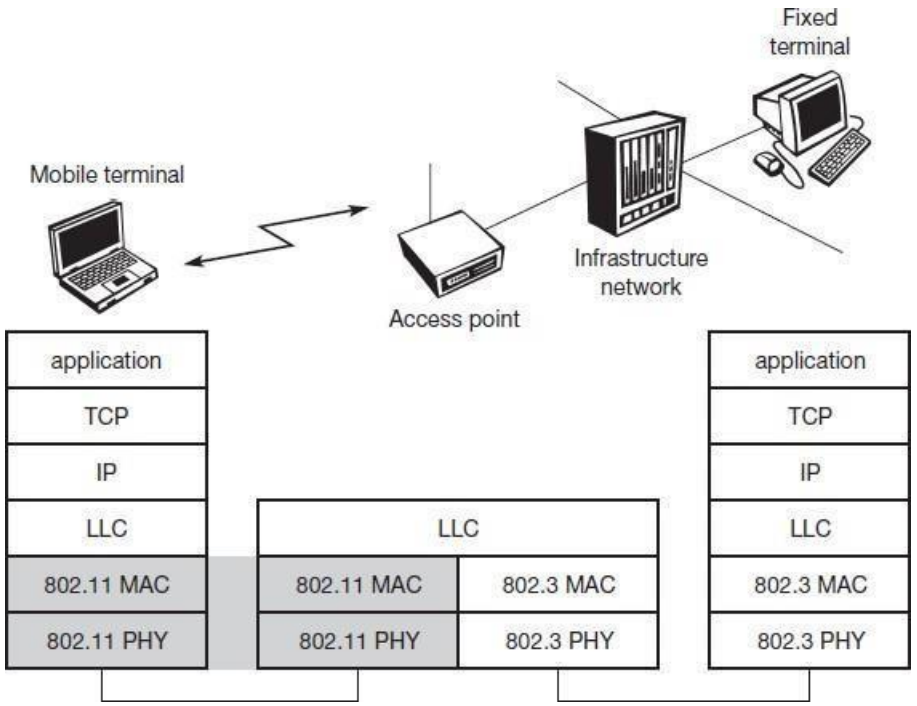


Figure: *IEEE 802.11 protocol architecture and bridging*

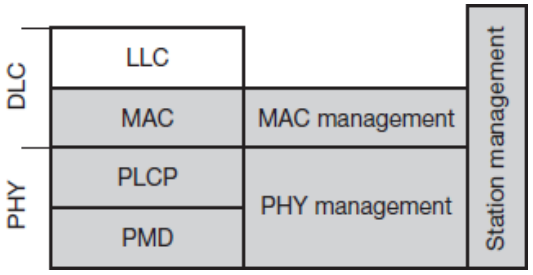


Figure*: Detailed IEEE 802.11 protocol architecture and management*

As indicated by the standard number, IEEE 802.11 fits seamlessly into the other 802.x standards for wired LANs Figure shows the most common scenario: an IEEE 802.11 wireless LAN connected to a switched IEEE 802.3 Ethernet via a bridge. Applications should not notice any difference apart from the lower bandwidth and perhaps higher access time from the wireless LAN. The WLAN behaves like a slow wired LAN. Consequently, the higher layers (application, TCP, IP) look the same for wireless nodes as for wired nodes. The upper part of the data link control layer, the logical link control (LLC), covers the differences of the medium access control layers needed for the different media. In many of today's networks, no explicit LLC layer is visible. Further details like Ether type or sub-network access protocol (SNAP) and bridging technology are explained in, e.g., Perlman (1992).

The IEEE 802.11 standard only covers the physical layer **PHY** and medium access layer **MAC** like the other 802.x LANs do. The physical layer is subdivided into the **physical layer convergence protocol (PLCP)** and the **physical medium dependent** sublayer **PMD**. The basic tasks of the MAC layer comprise medium access, fragmentation of user data, and encryption. The PLCP sublayer provides a carrier sense signal, called clear channel assessment (CCA), and provides a common PHY service access point (SAP) independent of the transmission technology. Finally, the PMD sublayer handles modulation and

encoding/decoding of signals. The PHY layer (comprising PMD and PLCP) and the MAC layer will be explained in more detail in the following sections.

Apart from the protocol sublayers, the standard specifies management layers and the station management. The **MAC anagement** supports the association and re-association of a station to an access point and roaming between different access points. It also controls authentication mechanisms, encryption, synchronization of a station with regard to an access point, and power management to save battery power. MAC management also maintains the MAC management information base (MIB).

The main tasks of the **PHY management** include channel tuning and PHY MIB maintenance. Finally, **station management** interacts with both management layers and is responsible for additional higher layer functions (e.g., control of bridging and interaction with the distribution system in the case of an access point).

2.16. Comparison SDMA/TDMA/FDMA/CDMA

Approach	SDMA	TDMA	FDMA	CDMA
Idea	segment space into cells/sectors	segment sending time into disjoint time-slots, demand driven or fixed patterns	segment the frequency band into disjoint sub-bands	spread the spectrum using orthogonal codes
Terminals	only one terminal can be active in one cell/one sector	all terminals are active for short periods of time on the same frequency	every terminal has its own frequency, uninterrupted	all terminals can be active at the same place at the same moment, uninterrupted
Signal separation	cell structure, directed antennas	synchronization in the time domain	filtering in the frequency domain	code plus special receivers
Advantages	very simple, increases capacity per km²	established, fully digital, flexible	simple, established, robust	flexible, less frequency planning needed, soft handover
Dis-advantages	inflexible, antennas typically fixed	guard space needed (multipath propagation), synchronization difficult	inflexible, frequencies are a scarce resource	complex receivers, needs more complicated power control for senders
Comment	only in combination with TDMA, FDMA or CDMA useful	standard in fixed networks, together with FDMA/SDMA used in many mobile networks	typically combined with TDMA (frequency hopping patterns) and SDMA (frequency reuse)	still faces some problems, higher complexity, lowered expectations; will be integrated with TDMA/FDMA

UNIT 3

DHCP

UNIT-3:

Mobile Network Layer: IP and Mobile IP Network Layers, Packet Delivery and Handover Management, Location Management, Registration, Tunneling and Encapsulation, Route Optimization, DHCP.

Need for Mobile IP

The IP addresses are designed to work with stationary hosts because part of the address defines the network to which the host is attached**.** A host cannot change its IP address without terminating on-going sessions and restarting them after it acquires a new address. Other link layer mobility solutions exist but are not sufficient enough for the global Internet.

Mobility is the ability of a node to change its point-of-attachment while maintaining allexisting communications and using the same IP address.

Nomadicity allows a node to move but it must terminate all existing communicationsand then can initiate new connections with a new address.

Mobile IP is a network layer solution for homogenous and heterogeneous mobility on the global Internet which is scalable, robust, secure and which allows nodes to maintain all ongoing communications while moving.

Design Goals: Mobile IP was developed as a means for transparently dealing with problems of mobile users. Mobile IP was designed to make the size and the frequency of required routing updates as small as possible. It was designed to make it simple to implement mobile node software. It was designed to avoid solutions that require mobile nodes to use multiple addresses.

Requirements: There are several requirements for Mobile IP to make it as a standard. Someof them are:

1. *Compatibility*: The whole architecture of internet is very huge and a new standard cannot introduce changes to the applications or network protocols already in use. Mobile IP is to be integrated into the existing operating systems. Also, for routers also it may be possible to enhance its capabilities to support mobility instead of changing the routers which is highly impossible. Mobile IP must not require special media or MAC/LLC protocols, so it must use the same interfaces and mechanisms to access the lower layers as IP does. Finally, end-systems enhanced with a mobile IP implementation should still be able to communicate with fixed systems without mobile IP.
2. *Transparency*: Mobility remains invisible for many higher layer protocols and applications. Higher layers continue to work even if the mobile computer has

changed its point of attachment to the network and even notice a lower bandwidth and some interruption in the service. As many of today's applications have not been designed to use in mobile environments, the effects of mobility will be higher delay and lower bandwidth.

3. *Scalability and efficiency*: The efficiency of the network should not be affected evenif a new mechanism is introduced into the internet. Enhancing IP for mobility mustnot generate many new messages flooding the whole network. Special care is necessary to be taken considering the lower bandwidth of wireless links. Many mobile systems have a wireless link to an attachment point. Therefore, only some additional packets must be necessary between a mobile system and a node in the network. It is indispensable for a mobile IP to be scalable over a large number of participants in the whole internet, throughout the world.
4. *Security*: Mobility possesses many security problems. A minimum requirement is the authentication of all messages related to the management of mobile IP. It must be sure for the IP layer if it forwards a packet to a mobile host that this host really is the receiver of the packet. The IP layer can only guarantee that the IP address of the receiver is correct. There is no way to prevent faked IP addresses and other attacks.

The goal of a mobile IP can be summarized as: 'supporting end-system mobility while maintaining scalability, efficiency, and compatibility in all respects with existing applications and Internet protocols'.

Entities and terminology

The following defines several entities and terms needed to understand mobile IP as defined in RFC 3344.

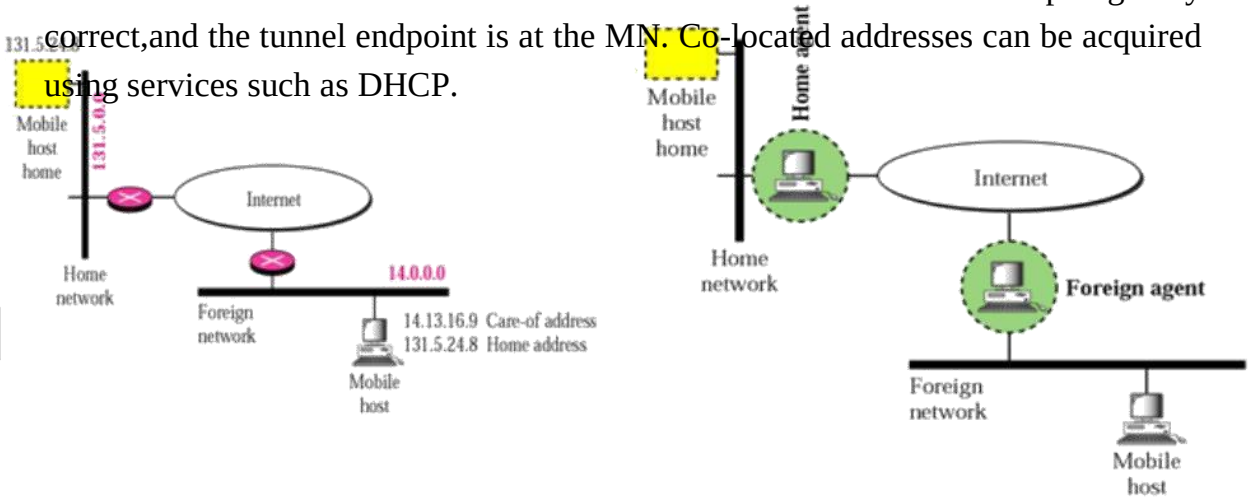
- **Mobile Node (MN):** A mobile node is an end-system or router that can change its point of attachment to the internet using mobile IP. The MN keeps its IP address and can continuously communicate with any other system in the internet as long as link-layer connectivity is given. Examples are laptop, mobile phone, router on an aircraft etc.
- **Correspondent node (CN):** At least one partner is needed for communication. In the following the CN represents this partner for the MN. The CN can be a fixed or mobile node.
- **Home network:** The home network is the subnet the MN belongs to with respect to itsIP address. No mobile IP support is needed within the home network.
- **Foreign network:** The foreign network is the current subnet the MN visits and which isnot the home network.

Foreign agent (FA): The FA can provide several services to the MN during its visit to the foreign network. The FA can have the COA, acting as tunnel endpoint and forwarding packets to the MN. The FA can be the default router for the MN. FAs can also provide security services because they belong to the foreign network as opposed to the MN which is only visiting. FA is implemented on a router for the subnet the MN attaches to.

Care-of address (COA): The COA defines the current location of the MN from an IP point of view. All IP packets sent to the MN are delivered to the COA, not directly to theIP address of the MN. Packet delivery toward the MN is done using a tunnel, i.e., the COAmarks the tunnel endpoint, i.e., the address where packets exit the tunnel. There are twodifferent possibilities for the location of the COA:

- **Foreign agent COA:** The COA could be located at the FA, i.e., the COA is an IP address of the FA. The FA is the tunnel end-point and forwards packets to the MN. Many MN using the FA can share this COA as common COA.

  Co-located COA: The COA is co-located if the MN temporarily acquired an additional IP address which acts as COA. This address is now topologically correct,and the tunnel endpoint is at the MN. Co-located addresses can be acquired using services such as DHCP.



Home agent (HA): The HA provides several services for the MN and is located in the home network. The tunnel for packets toward the MN starts at the HA. The HA maintainsa location registry, i.e., it is informed of the MN's location by the current COA. Three alternatives for the implementation of an HA exist.

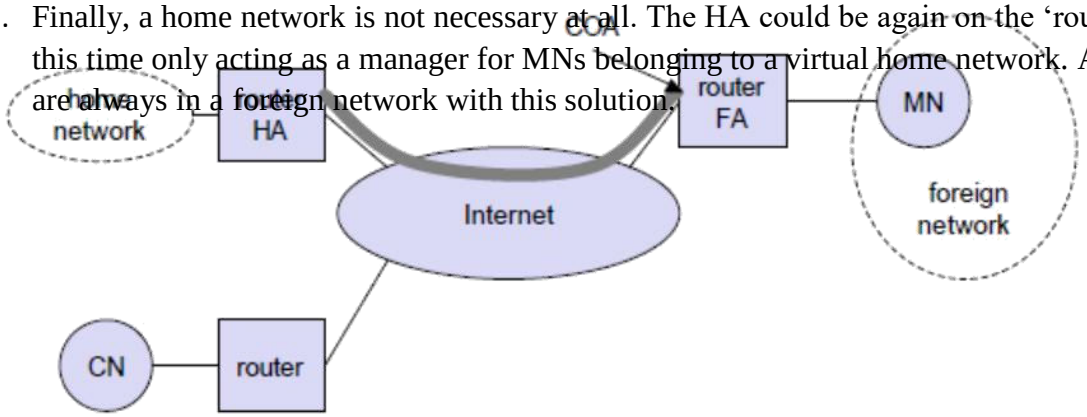
1. The HA can be implemented on a router that is responsible for the home network. This is obviously the best position, because without optimizations to mobile IP, all packets for the

MN have to go through the router anyway.

2. If changing the router's software is not possible, the HA could also be implemented on an arbitrary node in the subnet. One disadvantage of this solution is the double

crossing of the router by the packet if the MN is in a foreign network. A packet forthe MN comes in via the router; the HA sends it through the tunnel which again crosses the router.

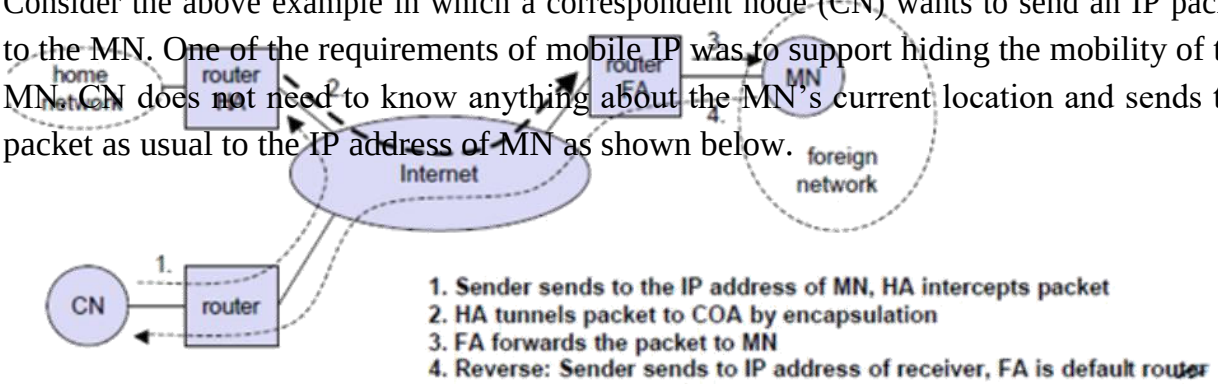
3. Finally, a home network is not necessary at all. The HA could be again on the 'router' but this time only acting as a manager for MNs belonging to a virtual home network. All MNs are always in a foreign network with this solution.



A CN is connected via a router to the internet, as are the home network and the foreign network. The HA is implemented on the router connecting the home network with the internet, an FA is implemented on the router to the foreign network. The MN is currently in the foreign network. The tunnel for packets toward the MN starts at the HA and ends at the FA, for the FA has the COA in the above example.

IP packet delivery and Handover Management

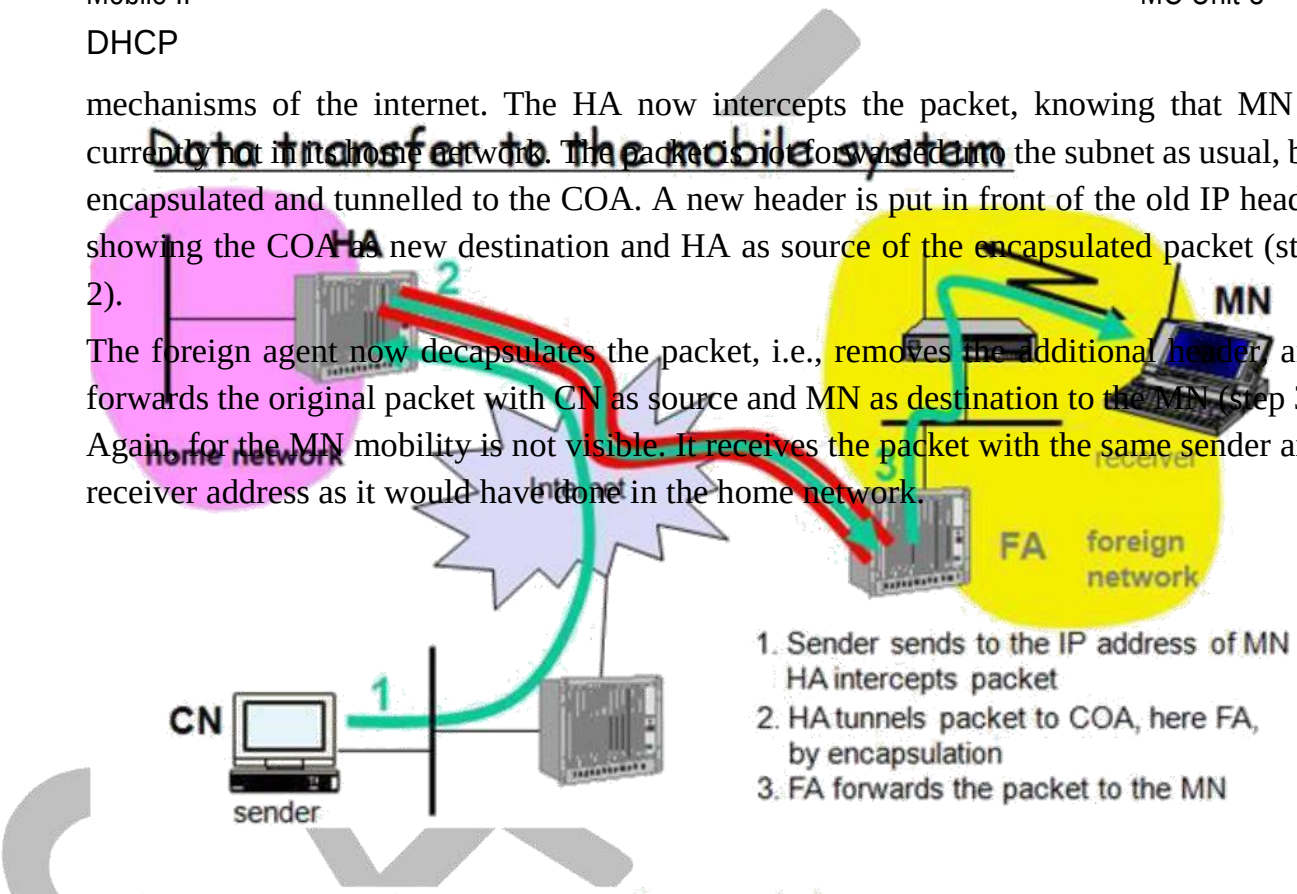
Consider the above example in which a correspondent node (CN) wants to send an IP packet to the MN. One of the requirements of mobile IP was to support hiding the mobility of the MN. CN does not need to know anything about the MN's current location and sends the packet as usual to the IP address of MN as shown below.

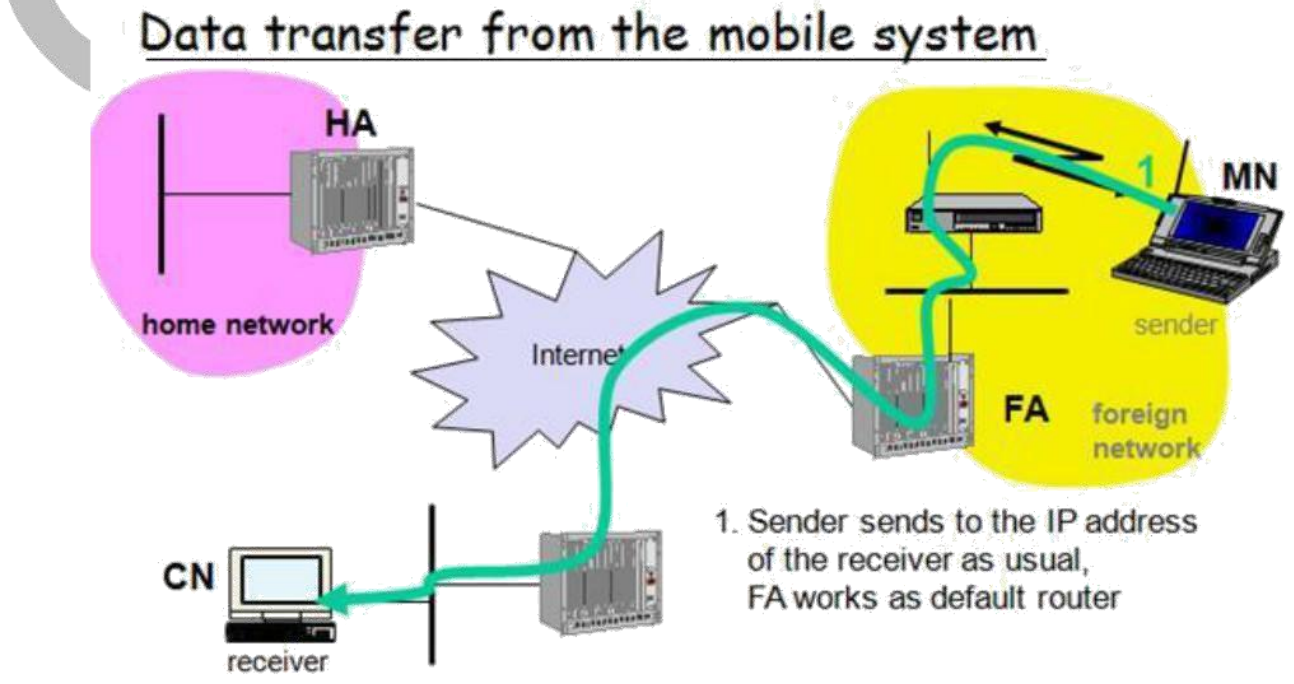


CN sends an IP packet with MN as a destination address and CN as a source address. Theinternet, not having information on the current location of MN, routes the packet to the router responsible for the home network of MN. This is done using the standard routing 4

mechanisms of the internet. The HA now intercepts the packet, knowing that MN is currently not in its home network. The packet is not forwarded into the subnet as usual, but encapsulated and tunnelled to the COA. A new header is put in front of the old IP header showing the COA as new destination and HA as source of the encapsulated packet (step 2).

The foreign agent now decapsulates the packet, i.e., removes the additional header and forwards the original packet with CN as source and MN as destination to the MN (step 3). Again, for the MN mobility is not visible. It receives the packet with the same sender and receiver address as it would have done in the home network.





DHCP

Sending packets from the mobile node (MN) to the CN is comparatively simple. The MN sends the packet as usual with its own fixed IP address as source and CN's address as destination (step 4). The router with the FA acts as default router and forwards the packet in the same way as it would do for any other node in the foreign network. As long as CN is a fixed node the remainder is in the fixed internet as usual. If CN were also a mobile node residing in a foreign network, the same mechanisms as described in steps 1 through 3 would apply now in the other direction.

Working of Mobile IP:- Mobile IP has two addresses for a mobile host: one home address and one care- of address. The home address is permanent; the care-of address changes as the mobile host movesfrom one network to another. To make the change of address transparent to the rest of the Internet requires a home agent and a foreign agent. The specific function of an agent is performed in the application layer. When the mobile host and the foreign agent are the same, the care-of address is called a co-located care-of address. To communicate with a remote host, a mobile host goes through

three phases: agent discovery, registration, and data transfer.

Agent Discovery

A mobile node has to find a foreign agent when it moves away from its home network. To solve this problem, mobile IP describes two methods: agent advertisement and agent solicitation.

Agent advertisement

For this method, foreign agents and home agents advertise their presence periodically using special **agent advertisement** messages, which are broadcast into the subnet. Mobile IP does not use a new packet type for agent advertisement; it uses the router advertisement packet of ICMP, and appends an agent advertisement message. The agent advertisement packet according to RFC 1256 with the extension for mobility is shown below:

DHCP

The TTL field of the IP packet is set to 1 for all advertisements to avoid forwarding them. The **type** is set to 9, the **code** can be 0, if the agent also routes traffic from non-mobile nodes, or 16, if it does not route anything other than mobile traffic. The number of addresses advertised with this packet is in **#addresses** while the **addresses** themselves follow as shown. **Lifetime** denotes the length of time this advertisement is valid. **Preference** levels for each address help a node to choose the router that is the most eager one to get a new node.

The extension for mobility has the following fields defined: **type** is set to 16, **length** depends on the number of COAs provided with the message and equals 6 + 4*(number of addresses). The **sequence number** shows the total number of advertisements sent since initialization by the agent. By the **registration lifetime** the agent can specify the maximum lifetime in seconds a node can request during registration. The following bits specify the characteristics of an agent in detail.

The **R** bit (registration) shows, if a registration with this agent is required even when using a colocated COA at the MN. If the agent is currently too busy to accept new registrations it can set the **B** bit. The following two bits denote if the agent offers services as a home agent (**H**) or foreign agent (**F**) on the link where the advertisement has been sent. Bits M and G specify the method of encapsulation used for the tunnel. While IP-in-IP encapsulation is the mandatory standard, **M** can specify minimal encapsulation and **G** generic routing encapsulation. In the first version of mobile IP (RFC 2002) the **V** bit specified the use of header compression according to RFC 1144. Now the field **r** at the same bit position is set to zero and must be ignored. The new field **T** indicates that reverse tunnelingis supported by the FA. The following fields contain the **COAs** advertised. A foreign agent setting the F bit must advertise at least one COA. A mobile node in a subnet can now receive agent advertisements from either its home agent or a foreign agent. This is one way for the MN to discover its location.

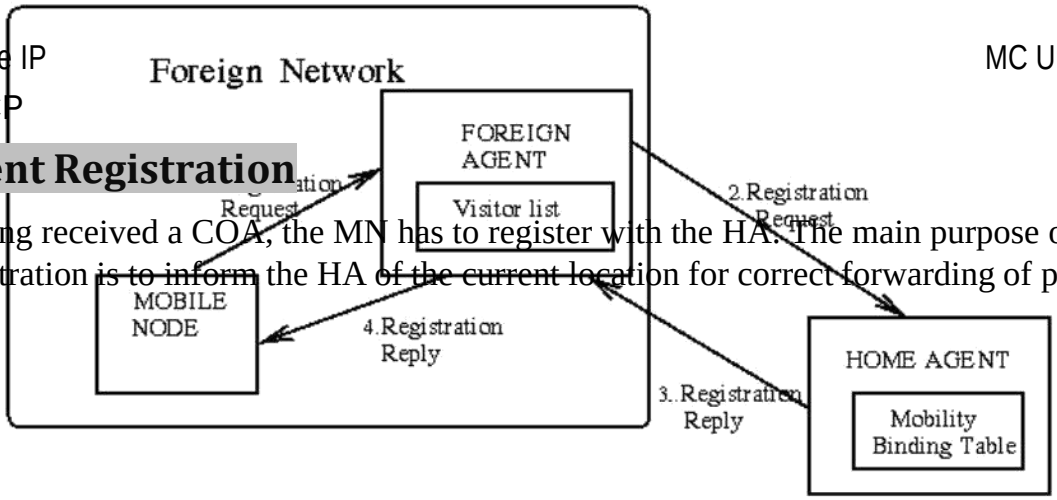
Agent Solicitation

If no agent advertisements are present or the inter-arrival time is too high, and an MN has not received a COA by other means, the mobile node must send **agent solicitations**. Care must be taken to ensure that these solicitation messages do not flood the network, but basically an MN can search for an FA endlessly sending out solicitation messages. If a node does not receive an answer to its solicitations it must decrease the rate of solicitations exponentially to avoid flooding the network until it reaches a maximum interval between

solicitations (typically one minute). Discovering a new agent can be done anytime, not just if the MN is not connected to one.

After these steps of advertisements or solicitations the MN can now receive a COA,either one for an FA or a co-located COA.

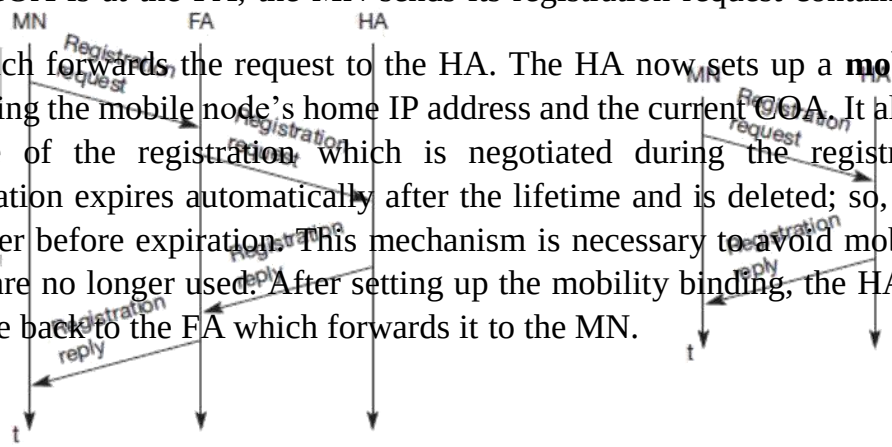
Agent Registration



Having received a COA, the MN has to register with the HA. The main purpose of the registration is to inform the HA of the current location for correct forwarding of packets.

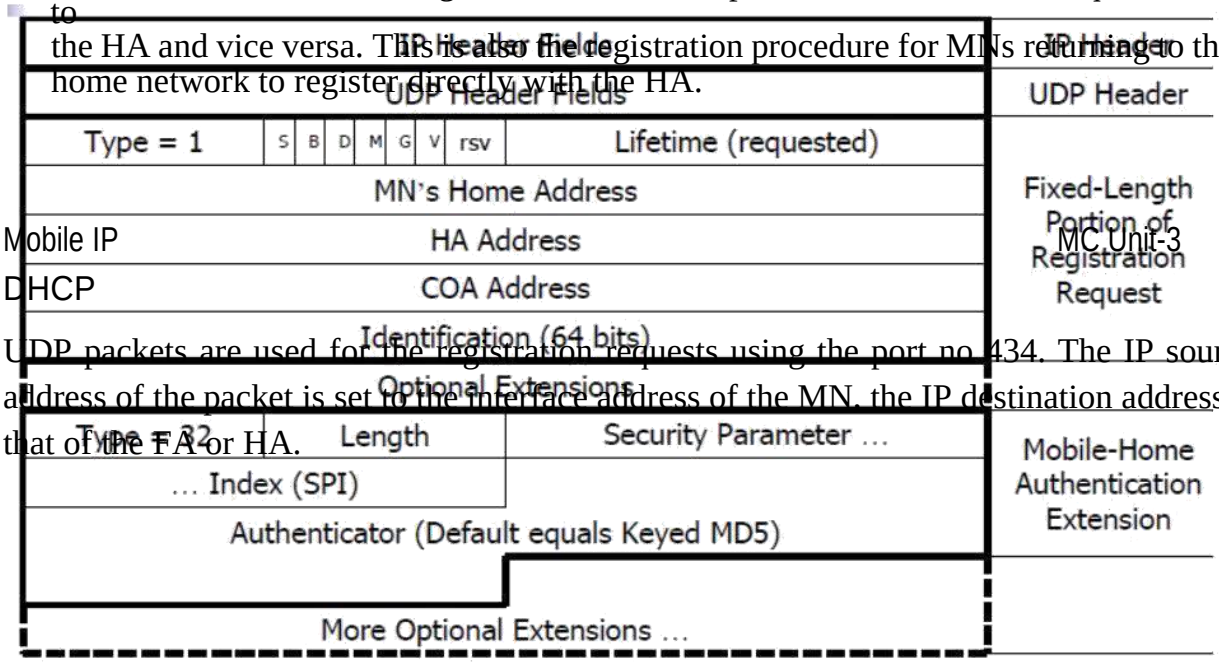
Registration can be done in two different ways depending on the location of the COA.

- If the COA is at the FA, the MN sends its registration request containing the COA to the FA which forwards the request to the HA. The HA now sets up a **mobility binding,** containing the mobile node's home IP address and the current COA. It also contains the lifetime of the registration which is negotiated during the registration process. Registration expires automatically after the lifetime and is deleted; so, an MN should reregister before expiration. This mechanism is necessary to avoid mobility binding s which are no longer used. After setting up the mobility binding, the HA sends a reply message back to the FA which forwards it to the MN.

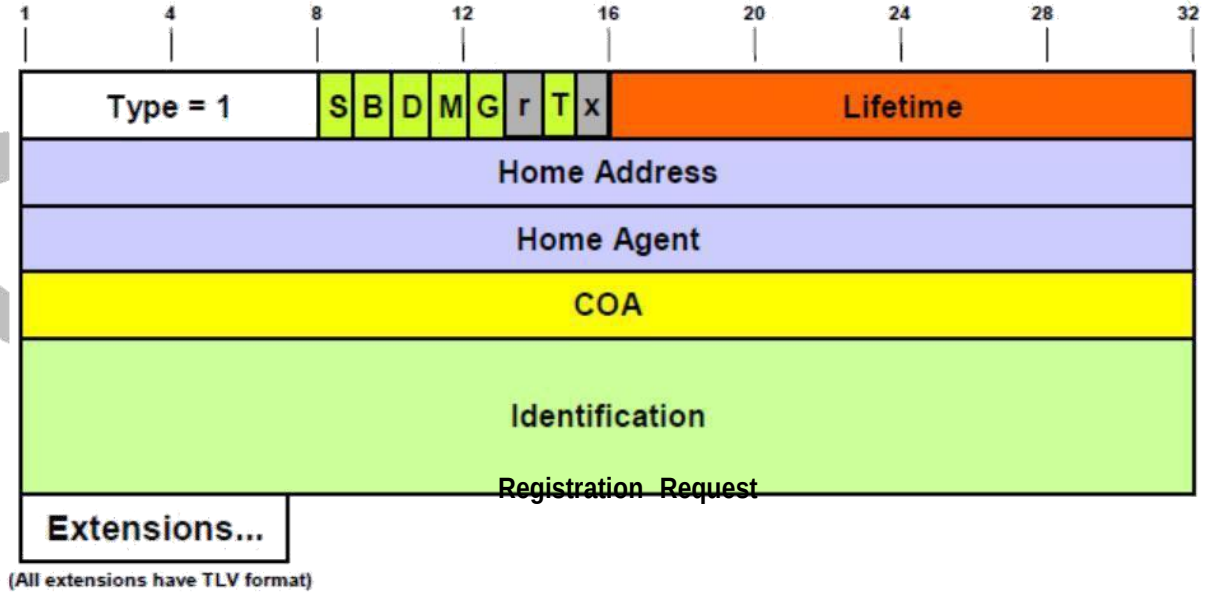


Registration of a mobile node via the FA or directly with the HA

If the COA is co-located, registration can be simpler, the MN sends the request directly to the HA and vice versa. This is also the registration procedure for MNs returning to their home network to register directly with the HA.

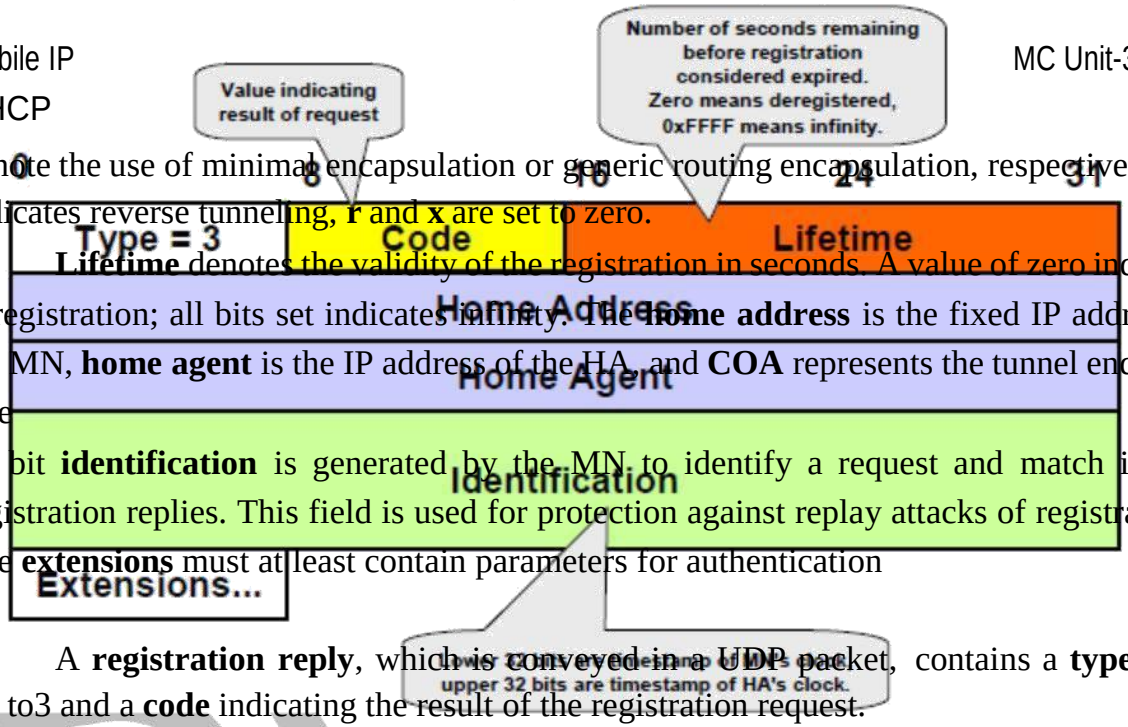


UDP packets are used for the registration requests using the port no 434. The IP source address of the packet is set to the interface address of the MN, the IP destination address is that of the FA or HA.



Registration Request

The first field **type** is set to 1 for a registration request. With the **S** bit an MN can specify if it wants the HA to retain prior mobility bindings. This allows for simultaneous bindings. Setting the **B** bit generally indicates that an MN also wants to receive the broadcastpackets which have been received by the HA in the home network. If an MN uses a co- located COA, it also takes care of the decapsulation at the tunnel endpoint. The **D** bit indicates this behavior. As already defined for agent advertisements, the bits **M** and **G**



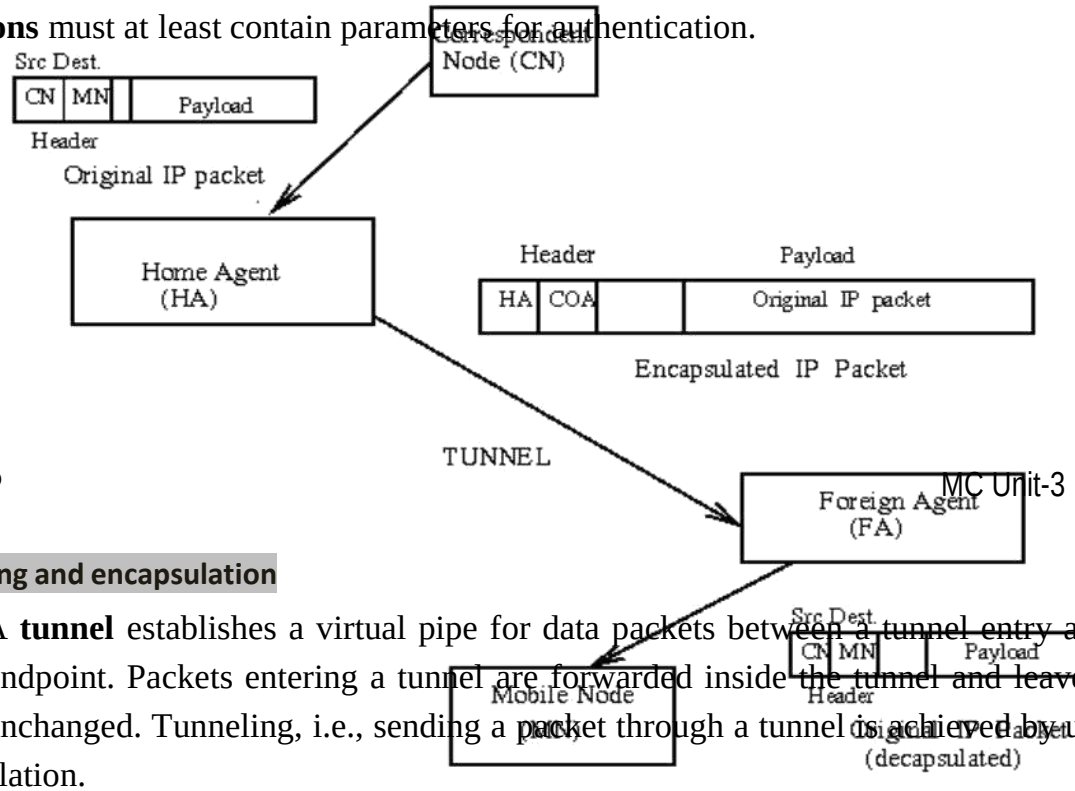
denote the use of minimal encapsulation or generic routing encapsulation, respectively. **T** indicates reverse tunneling, **r** and **x** are set to zero.

Lifetime denotes the validity of the registration in seconds. A value of zero indicates deregistration; all bits set indicates infinity. The **home address** is the fixed IP address of the MN, **home agent** is the IP address of the HA, and **COA** represents the tunnel endpoint. The 64 bit **identification** is generated by the MN to identify a request and match it with registration replies. This field is used for protection against replay attacks of registrations. The **extensions** must at least contain parameters for authentication

A **registration reply**, which is conveyed in a UDP packet, contains a **type** field set to3 and a **code** indicating the result of the registration request.

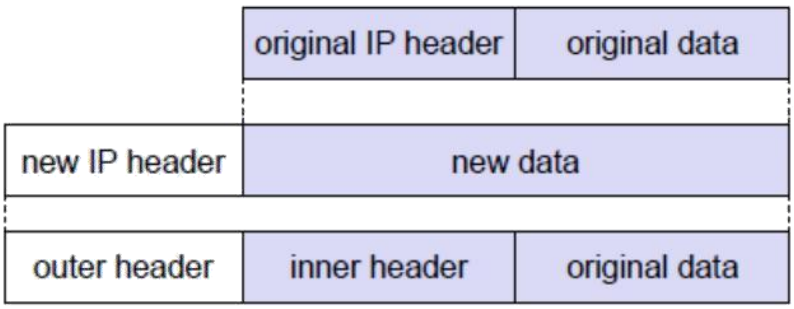
Registration Reply

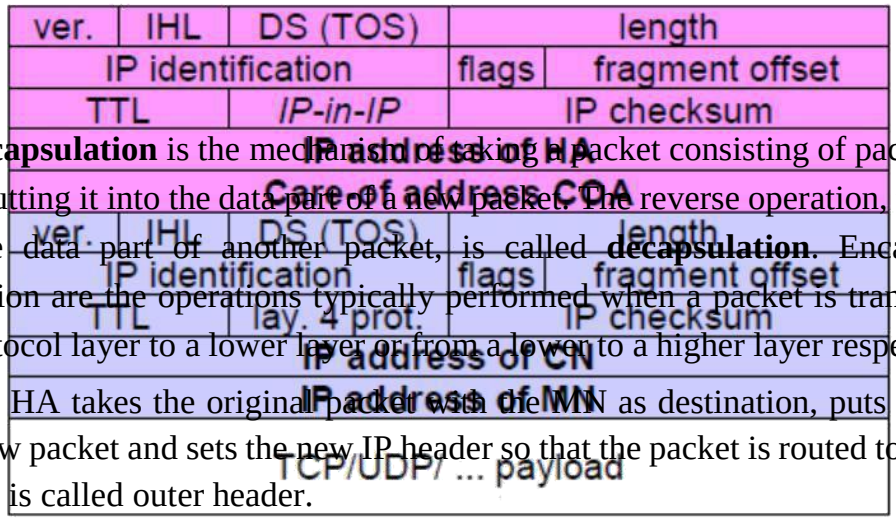
The **lifetime** field indicates how many seconds the registration is valid if it was successful. **Home address** and **home agent** are the addresses of the MN and the HA, respectively. The 64-bit **identification** is used to match registration requests with replies. The value is based on the identification field from the registration and the authentication method. Again, the **extensions** must at least contain parameters for authentication.



Tunnelling and encapsulation

A **tunnel** establishes a virtual pipe for data packets between a tunnel entry and a tunnel endpoint. Packets entering a tunnel are forwarded inside the tunnel and leave the tunnel unchanged. Tunneling, i.e., sending a packet through a tunnel is achieved by using encapsulation.





Encapsulation is the mechanism of taking a packet consisting of packet header and data and putting it into the data part of a new packet. The reverse operation, taking a packet out of the data part of another packet, is called **decapsulation**. Encapsulation and decapsulation are the operations typically performed when a packet is transferred from a higher protocol layer to a lower layer or from a lower to a higher layer respectively.

The HA takes the original packet with the MN as destination, puts it into the data partof a new packet and sets the new IP header so that the packet is routed to the COA. The newheader is called outer header.

IP-in-IP encapsulation

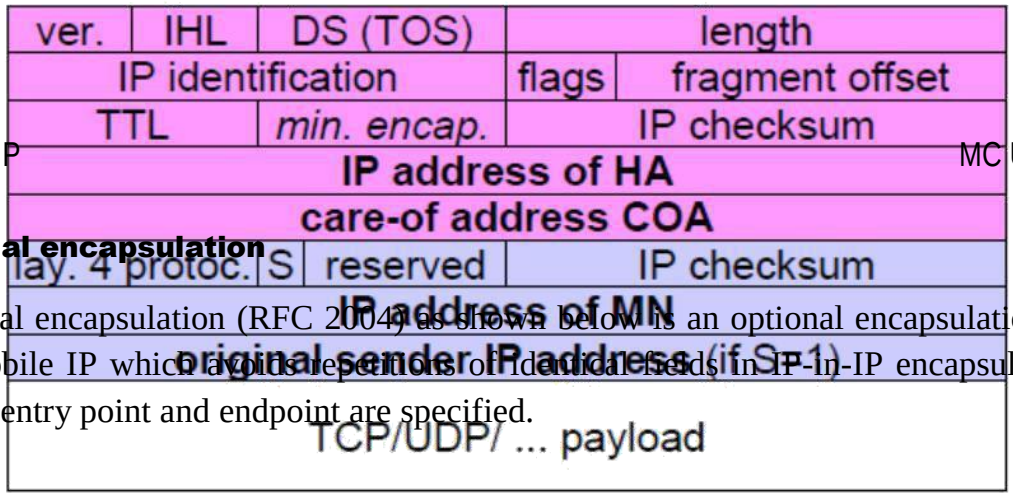
There are different ways of performing the encapsulation needed for the tunnel between HA and COA. Mandatory for mobile IP is **IP-in-IP encapsulation** as specified in RFC 2003. The following fig shows a packet inside the tunnel.

The version field **ver** is 4 for IP version 4, the internet header length (**IHL**) denotes the length of the outer header in 32 bit words. **DS(TOS)** is just copied from the inner header, the **length** field covers the complete encapsulated packet. The fields up to TTL have no special meaning for mobile IP and are set according to RFC 791. **TTL** must be high enough so the packet can reach the tunnel endpoint. The next field, here denoted with **IP- in-IP**, is the type of the protocol used in the IP payload. This field is set to 4, the protocol type for IPv4 because again an IPv4 packet follows after this outer header. IP **checksum** is calculated as usual. The next fields are the tunnel entry as source address (the **IP address of the HA**) and the tunnel exit point as destination address (the **COA**).

If no options follow the outer header, the inner header starts with the same fields as above. This header remains almost unchanged during encapsulation, thus showing the original sender CN and the receiver MN of the packet. The only change is TTL which is decremented by 1. This means that the whole tunnel is considered a single hop from the original packet's point of view. This is a very important feature of tunneling as it allows

the MN to behave as if it were attached to the home network. No matter how many real hops the packet has to take in the tunnel, it is just one (logical) hop away for the MN. Finally, the payload follows the two headers.

Minimal encapsulation

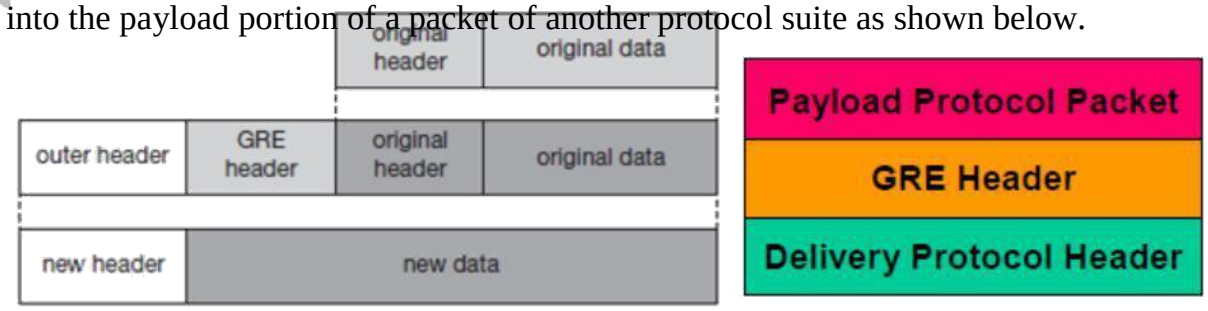


Minimal encapsulation (RFC 2004) as shown below is an optional encapsulation method for mobile IP which avoids repetitions of identical fields in IP-in-IP encapsulation. The tunnel entry point and endpoint are specified.

The field for the type of the following header contains the value 55 for the minimal encapsulation protocol. The inner header is different for minimal encapsulation. The type of the following protocol and the address of the MN are needed. If the **S** bit is set, the original sender address of the CN is included as omitting the source is quite often not an option. No field for fragmentation offset is left in the inner header and minimal encapsulation does not work with already fragmented packets.

Generic Routing Encapsulation

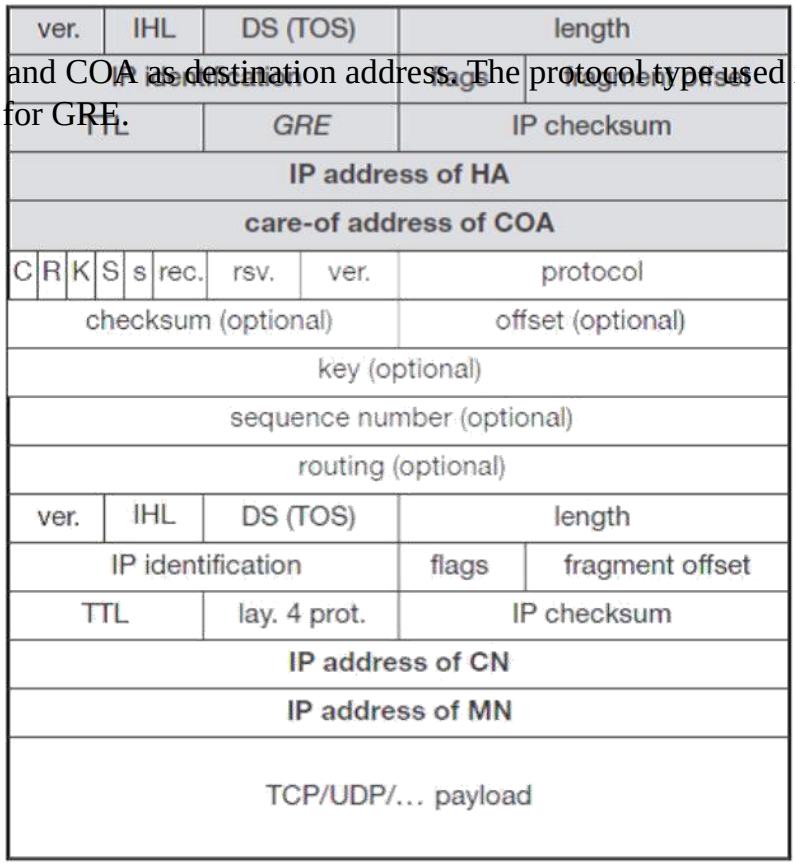
Unlike IP-in-IP and Minimal encapsulation which work only for IP packets, **Generic routing encapsulation** (GRE) allows the encapsulation of packets of one protocol suite into the payload portion of a packet of another protocol suite as shown below.



The packet of one protocol suite with the original packet header and data is taken and a new GRE header is prepended. Together this forms the new data part of the new packet. Finally, the header of the second protocol suite is put in front.The following figure shows the fields of a packet inside the tunnel between HA and COA using GRE as an encapsulation scheme according to RFC 1701. The outer header is the standard IP header with HA as

DHCP

source address and COA as destination address. The protocol type used in this outer IPheader is 47 for GRE.



The GRE header starts with several flags indicating if certain fields are present or not. A minimal GRE header uses only 4 bytes. The **C** bit indicates if the checksum field is present and contains valid information. If **C** is set, the **checksum** field contains a valid IP checksum ofthe GRE header and the payload. The **R** bit indicates if the offset and routing fields are present and contain valid information. The **offset** represents the offset in bytes for the first source **routing** entry. The routing field, if present, has a variable length and contains fieldsfor source routing. GRE also offers a **key** field which may be used for authentication. If this field is present, the **K** bit is set. The sequence number bit **S** indicates if the **sequence** numberfield is present, if the s bit is set, strict source routing is used.

The **recursion control** field (rec.) is an important field that additionally distinguishes GRE from IP-in-IP and minimal encapsulation. This field represents a counter that shows the number of allowed recursive encapsulations. The default value of this field should be 0, thus allowing only one level of encapsulation. The following **reserved** fields must be zero and are ignored on reception. The **version** field contains 0 for the GRE version.

The following 2 byte **protocol** field represents the protocol of the packet following the GRE header. The standard header of the original packet follows with the source address of the correspondent node and the destination address of the mobile node.

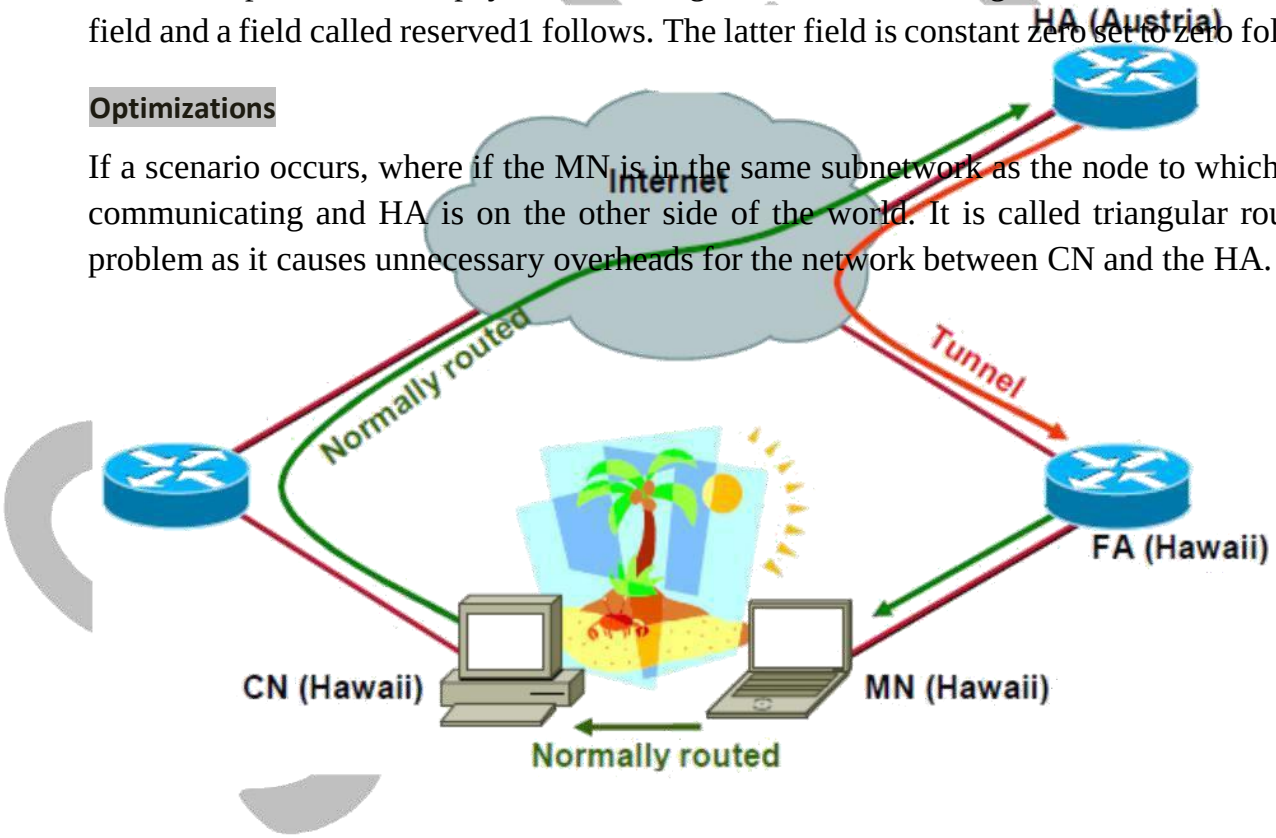
C	reserved0	ver.	protocol
checksum (optional)			reserved1 (=0)

A simplified header of GRE following RFC 2784 is shown below.

The field **C** indicates again if a checksum is present. The next 5 bits are set to zero, then 7 reserved bits follow. The **version** field contains the value zero. The **protocol** type, again, defines the protocol of the payload following RFC 3232. If the flag C is set, then **checksum** field and a field called reserved1 follows. The latter field is constant zero set to zero follow.

Optimizations

If a scenario occurs, where if the MN is in the same subnetwork as the node to which it is communicating and HA is on the other side of the world. It is called triangular routing problem as it causes unnecessary overheads for the network between CN and the HA.



A solution to this problem is to inform the CN of the current location of the MN. The CN can learn the location by caching it in a binding cache, which is a part of the routing table for the CN. HA informs the CN of the location. It needs four additional messages:

Binding Request: It is sent by the node that wants to know the current location of an MN
to the HA. HA checks if it is allowed to reveal the location and then sends back a bindingupdate

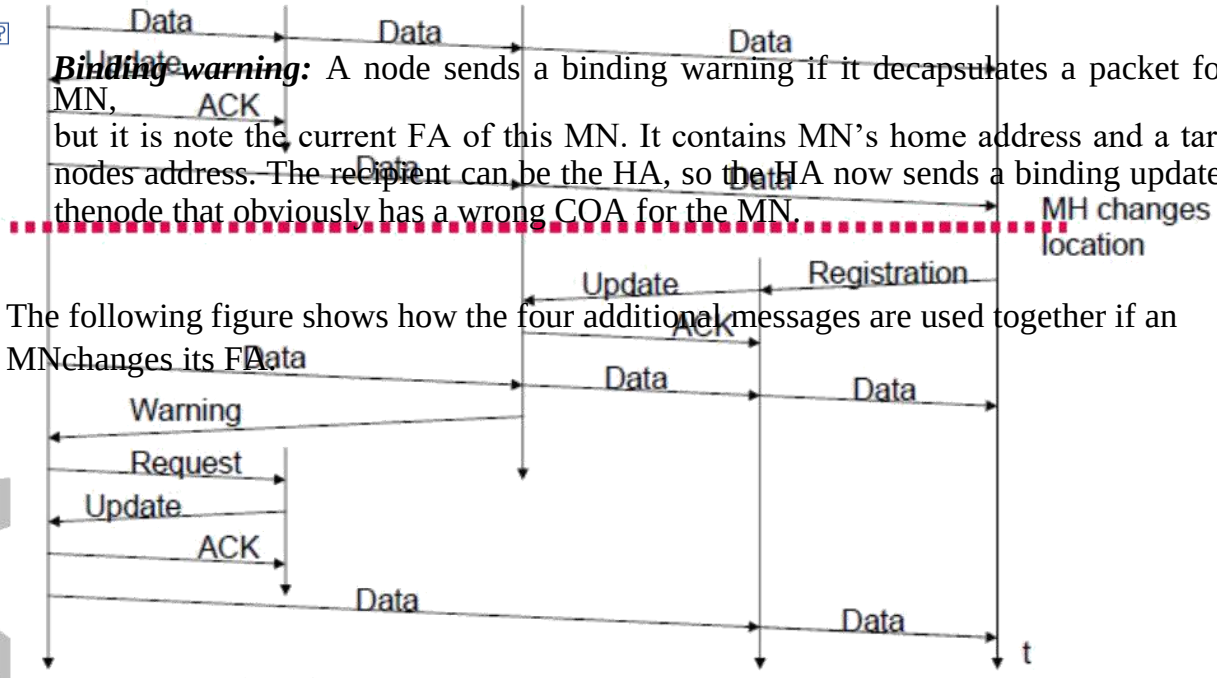
Binding update: It is sent by the HA to the CN revealing the current location of an MN. It
contains the fixed IP address of the MN and the COA. This message can request anacknowledgement.

DHCP

➢

Binding acknowledgement: If requested, a node returns this acknowledgement after receiving a binding update message

Binding warning: A node sends a binding warning if it decapsulates a packet for an MN,
but it is note the current FA of this MN. It contains MN's home address and a target nodes address. The recipient can be the HA, so the HA now sends a binding update to thenode that obviously has a wrong COA for the MN.



The following figure shows how the four additional messages are used together if an MNchanges its FA.

The CN can request the current location from the HA. If allowed by the MN, the HA returns the COA of the MN via an update message. The CN acknowledges this update message and stores the mobility binding. Now the CN can send its data directly to the current foreign agent FA_{old}. FA_{old} forwards the packets to the MN. This scenario shows a COA located at an FA. Encapsulation of data for tunneling to the COA is now done by the CN, not the HA.

The MN might now change its location and register with a new foreign agent, FA_{new}. This registration is also forwarded to the HA to update its location database. Furthermore, FA_{new} informs FA_{old} about the new registration of MN. MN's registration message contains the address of FA_{old} for this purpose. Passing this information is achieved via an update message, which is acknowledged by FA_{old}.

DHCP

Without the information provided by the new FA, the old FA would not get to know anything about the new location of MN. In this case, CN does not know anything about the new location, so it still tunnels its packets for MN to the old FA, FA_{old}. This FA now notices packets with destination MN, but also knows that it is not the current FA of MN. FA_{old} might now forward these packets to the new COA of MN which is FA_{new} in this example. This forwarding of packets is another optimization of the basic Mobile IP providing **smooth handovers**. Without this optimization, all packets in transit would be lost while the MN moves from one FA to another.

To tell CN that it has a stale binding cache, FA_{old} sends, a binding warning message to CN. CN then requests a binding update. (The warning could also be directly sent to the HA triggering an update). The HA sends an update to inform the CN about the new location, which is acknowledged. Now CN can send its packets directly to FA_{new}, again avoiding triangular routing. Unfortunately, this optimization of mobile IP to avoid triangular routing causes several security problems

Reverse Tunnelling

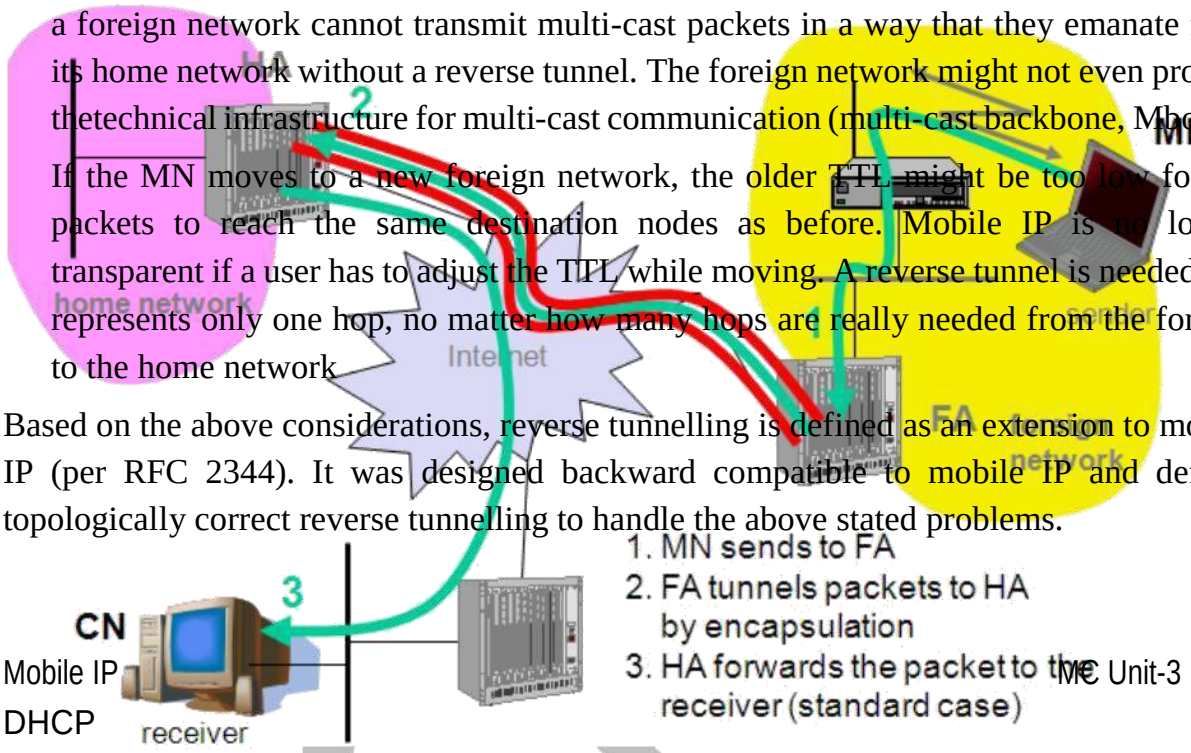
The reverse path from MS to the CN looks quite simple as the MN can directly send its packets to the CN as in any other standard IP situation. The destination address in the packets is that of CN. But it has some problems explained below:-

Quite often firewalls are designed to only allow packets with topologically correct addresses to pass to provide simple protection against misconfigured systems of unknown addresses. However, MN still sends packets with its fixed IP address as source which is not topologically correct in a foreign network. Firewalls often filter packets coming from outside containing a source address from computers of the internal network. This also implies that an MN cannot send a packet to a computer residing in

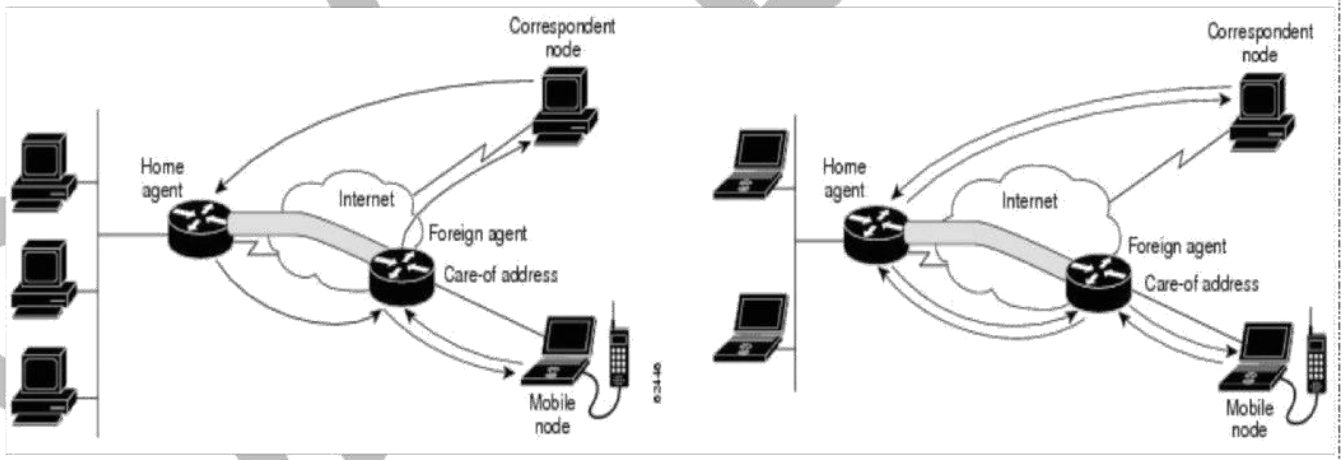
its home network.

While the nodes in the home network might participate in a multi-cast group, an MN in a foreign network cannot transmit multi-cast packets in a way that they emanate from its home network without a reverse tunnel. The foreign network might not even provide thetechnical infrastructure for multi-cast communication (multi-cast backbone, Mbone).

If the MN moves to a new foreign network, the older TTL might be too low for the packets to reach the same destination nodes as before. Mobile IP is no longer transparent if a user has to adjust the TTL while moving. A reverse tunnel is needed that represents only one hop, no matter how many hops are really needed from the foreign to the home network



Based on the above considerations, reverse tunnelling is defined as an extension to mobile IP (per RFC 2344). It was designed backward compatible to mobile IP and defines topologically correct reverse tunnelling to handle the above stated problems.



Reverse Tunnelling

Packet Forwarding **Reverse Tunnel**

Reverse tunneling does not solve

- problems with *firewalls*, the reverse tunnel can be abused to circumvent securitymechanisms (tunnel hijacking)
- optimization of data paths, i.e. packets will be forwarded through the tunnel via the HAto a sender (double triangular routing)

DHCP

IPv6

The design of Mobile IP support in IPv6 (Mobile IPv6) benefits both from the experiences gained from the development of Mobile IP support in IPv4, and from the opportunities provided by IPv6. Mobile IPv6 thus shares many features with Mobile IPv4, but is integrated into IPv6 and offers many other improvements. This section summarizes the major differences between Mobile IPv4 and Mobile IPv6:

There is no need to deploy special routers as "foreign agents", as in Mobile IPv4. Mobile IPv6 operates in any location without any special support required from the local router.

Support for route optimization is a fundamental part of the protocol, rather than a nonstandard set of extensions.

Mobile IPv6 route optimization can operate securely even without pre-arranged security associations. It is expected that route optimization can be deployed on a global scale between all mobile nodes and correspondent nodes.

Support is also integrated into Mobile IPv6 for allowing route optimization to coexist efficiently with routers that perform "ingress filtering"

The IPv6 Neighbor Unreachability Detection assures symmetric reachability between the mobile node and its default router in the current location.

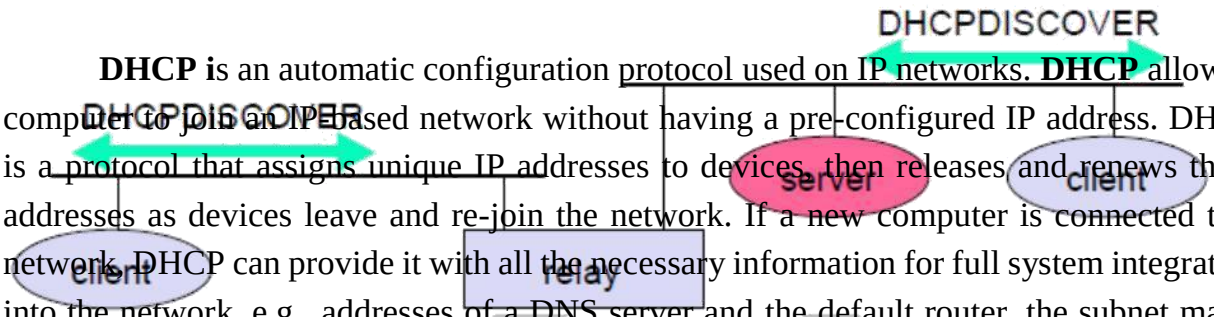
Most packets sent to a mobile node while away from home in Mobile IPv6 are sent using an IPv6 routing header rather than IP encapsulation, reducing the amount of resulting overhead compared to Mobile IPv4.

Mobile IPv6 is decoupled from any particular link layer, as it uses IPv6 NeighborDiscovery instead of ARP. This also improves the robustness of the protocol.

The use of IPv6 encapsulation (and the routing header) removes the need in Mobile IPv6to manage "tunnel soft state".

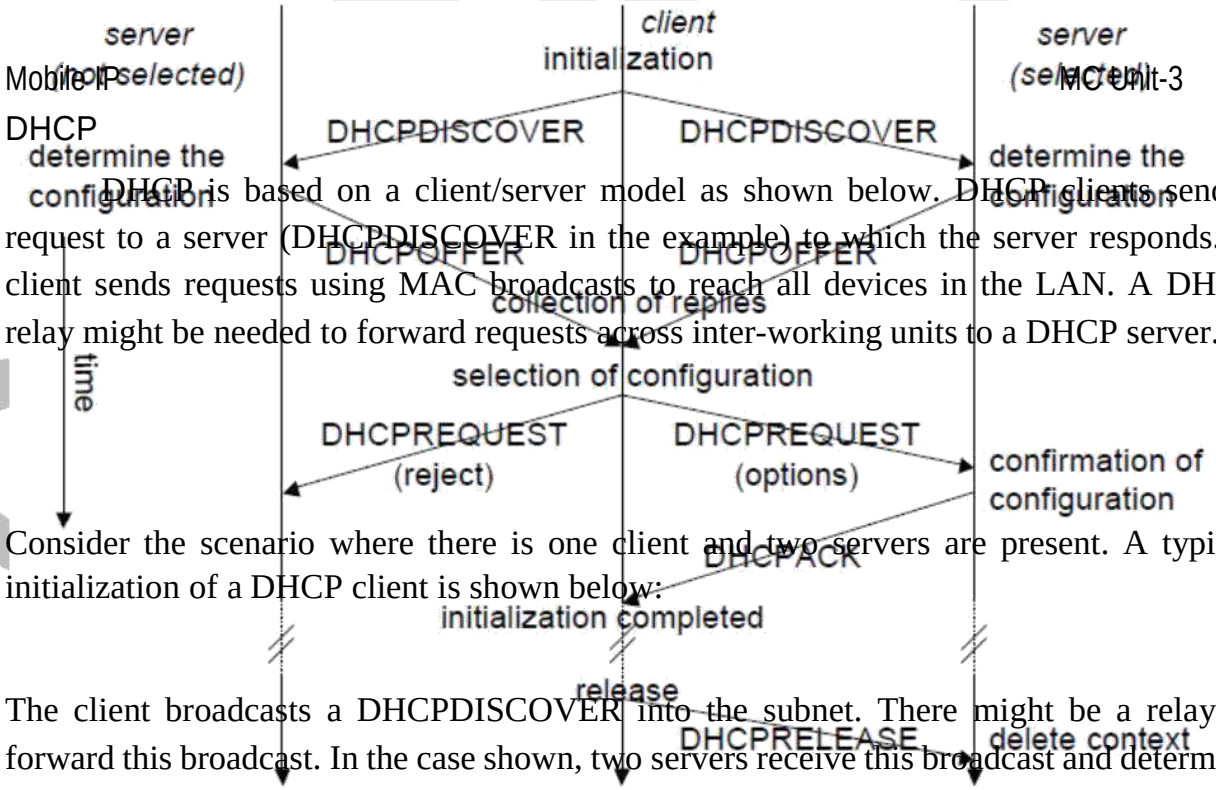
The dynamic home agent address discovery mechanism in Mobile IPv6 returns a single reply to the mobile node. The directed broadcast approach used in IPv4 returns separate replies from each home agent.

Dynamic Host Configuration Protocol (DHCP)



DHCP is an automatic configuration protocol used on IP networks. **DHCP** allows a computer to join an IP-based network without having a pre-configured IP address. DHCP is a protocol that assigns unique IP addresses to devices, then releases and renews these addresses as devices leave and re-join the network. If a new computer is connected to a network, DHCP can provide it with all the necessary information for full system integration into the network, e.g., addresses of a DNS server and the default router, the subnet mask, the domain name, and an IP address. Providing an IP address makes DHCP very attractive for mobile IP as a source of care-of-addresses.

DHCP



DHCP is based on a client/server model as shown below. DHCP clients send a request to a server (DHCPDISCOVER in the example) to which the server responds. A client sends requests using MAC broadcasts to reach all devices in the LAN. A DHCP relay might be needed to forward requests across inter-working units to a DHCP server.

Consider the scenario where there is one client and two servers are present. A typical initialization of a DHCP client is shown below:

The client broadcasts a DHCPDISCOVER into the subnet. There might be a relay to forward this broadcast. In the case shown, two servers receive this broadcast and determine the configuration they can offer to the client. Servers reply to the client's request with DHCPOFFER and offer a list of configuration parameters. The client can now choose one of the configurations offered. The client in turn replies to the servers, accepting one of the configurations and rejecting the others using DHCPREQUEST. If a server receives a DHCPREQUEST with a rejection, it can free the reserved configuration for other possible

clients. The server with the configuration accepted by the client now confirms the configuration with DHCPACK. This completes the initialization phase. If a client leaves a subnet, it should release the configuration received by the server using DHCPRELEASE. Now the server can free the context stored for the client and offer the configuration again. The configuration a client gets from a server is only leased for a certain amount of time, it has to be reconfirmed from time to time. Otherwise the server will free the configuration. This timeout of configuration helps in the case of crashed nodes or nodes moved away without releasing the context.

DHCP is a good candidate for supporting the acquisition of care -of addresses for mobile nodes. The same holds for all other parameters needed, such as addresses of the default router, DNS servers, the timeserver etc. A DHCP server should be located in the subnet of the access point of the mobile node, or at least a DHCP relay should provide forwarding of the messages. RFC 3118 specifies authentication for DHCP messages so as to provide protection from malicious DHCP servers. Without authentication, a DHCP server cannot trust the mobile node and vice versa…

Location Management

Introduction

Location management is an important problem in mobile computing since wireless mobile computers can change location while connected to the network. New strategies must be introduced to deal with the dynamic changes of a mobile computer`s network address. A few problems associated with mobility will be discussed in this article.

Mobility and Location Management

The ability to change locations while connected to the network creates a dynamic computing environment. This means that data which is static for stationary computing becomes dynamic for mobile computing. An example is that a stationary computer is permanently attached to the nearest server while mobile computers need a mechanism to determine which server to use.

As people move, their mobile computers will use different network addresses. The

networking used today has to be changed to deal with dynamically changing addresses. If we, for example, look at how the Internet Protocol (IP) is designed for fixed computing, a host IP is bound with its network address so moving to a new location means that it needs a new IP name.

There are a few questions that must be answered when looking at a location management scheme. What happens when a mobile user changes location? Who should know about the change? How can you contact a mobile host? Should you search the whole network or does anyone know about the mobile users moves?

A few basic mechanisms to determine a mobile computer's current location has been discussed to modify the IP-based protocols. We will look at four of them in this article; broadcast, central services, home base and forwarding pointers.

Selective Broadcast

With this method a message is sent to all network cells asking the mobile computer to replywith its current address. This scheme may be too expensive in large networks. However, if the mobile computer is known to be in one of a few cells a message is sent out to the selected cells. A disadvantage with selective broadcast is that it can only be used when we have enough information about current location.

Central Services

The current address for each mobile user is kept in a centralized database. When a mobile computer changes its address it also updates the central database by sending a message containing its new address.

Home Bases

With this method the location of a given mobile computer is known by a single server (MSS), often called the *Home Location Server*. The user is permanently registered under this server and it keeps track of where the mobile computer is. To send a message to a mobile user , thehome location server has to be contacted first to obtain the users' current address.

The main disadvantage with this scheme is that the way a message must travel may be much

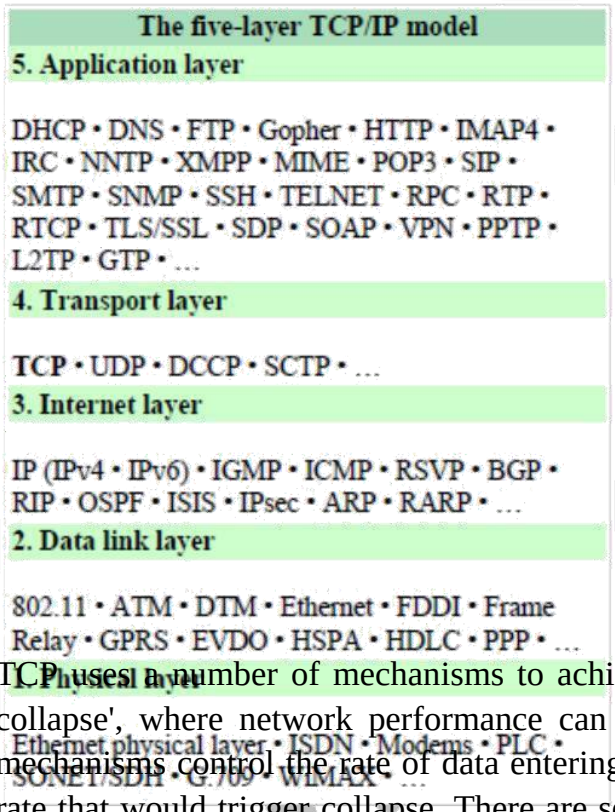
longer than the real distance. For example, two mobile computers, A and B, which are registered under two different home location servers in two different areas, may be currently in the same area. For A to contact B it has to first contact B's home location server which then contacts B. If A and B are likely to be in the same area, this scheme could be modified to first broadcast a message to all MSSs in that local area. If B is not currently located there a message is then sent to B's home location server.
This scheme can also lead to low availability of information. The home location server maybe down or inaccessible which makes it impossible to track the requested mobile user.

Forwarding Pointers

This method is probably one of the fastest. Each time a mobile computer changes its address, a copy of the new address is added at the old location. The message sent is then forwarded along the chain of pointers until the mobile computer is reached. The pointer chain will be made longer every time the mobile computer changes location and this may lead to inefficient routing. To solve these pointers at the message forwarders can be updated to contain more recent addresses.
Even though this method is among the fastest it suffers from failure anywhere along the chain of pointers. Another problem is associated with deleting pointers which cannot be done before all message sources have been updated.
The forwarding pointer method can be hard to implement. It does not fit standard networking models since it must have an active entity at the old address to receive and forward messages. Today’s network address is usually a passive entity.
There has not yet been done much work on comparing different locating and addressing schemes. The problem is difficult because it involves several dimensions. An issue introduced by these locating and addressing schemes is the cost of search. The less information the sender has about the mobile computer the more it will cost to search. This must also be considered when choosing for a location management scheme.

Traditional TCP

The **Transmission Control Protocol** (**TCP**) is one of the core protocols of the Internet protocol suite, often simply referred to as TCP/IP. TCP is reliable, guarantees in-order delivery of data and incorporates congestion control and flow control mechanisms.



TCP supports many of the Internet's most popular application protocols and resulting applications, including the World Wide Web, e-mail, File Transfer Protocol and Secure Shell. In the Internet protocol suite, TCP is the intermediate layer between the Internet layer and application layer.

The major responsibilities of TCP in an active session are to:

- **Provide reliable in-order transport of data**: to not allow losses of data.
- **Control congestions in the networks**: to not allow degradation of the network performance,
- **Control a packet flow between the transmitter and the receiver**: to not exceed the receiver's capacity.

TCP uses a number of mechanisms to achieve high performance and avoid 'congestion collapse', where network performance can fall by several orders of magnitude. These mechanisms control the rate of data entering the network, keeping the data flow below a rate that would trigger collapse. There are several mechanisms of TCP that influence the efficiency of TCP in a mobile environment. Acknowledgments for data sent, or lack of acknowledgments, are used by senders to implicitly interpret network conditions between the TCP sender and receiver.

Congestion Control

A transport layer protocol such as TCP has been designed for fixed networks with fixed end- systems. Congestion may appear from time to time even in carefully designed networks. The packet buffers of a router are filled and the router cannot forward the packets fast enough because the sum of the input rates of packets destined for one output link is higher than the capacity of the output link. The only thing a router can do in this situation is to drop packets. A dropped packet is lost for the transmission, and the receiver notices a gap in the packet stream. Now the receiver does not directly tell the sender which packet is missing, but continues to acknowledge all in- sequence packets up to the missing one.

The sender notices the missing acknowledgement for the lost packet and assumes a packet loss due to congestion. Retransmitting the missing packet and continuing at full sending rate would now be unwise, as this might only increase the congestion. To mitigate congestion, TCP slows down the transmission rate dramatically. All other TCP connections experiencing the same congestion do exactly the same so the congestion is soon resolved.

Slow start

TCP's reaction to a missing acknowledgement is quite drastic, but it is necessary to get rid of congestion quickly. The behavior TCP shows after the detection of congestion is called **slow start.** The sender always calculates a **congestion window** for a receiver. The start size of the congestion window is one segment (TCP packet). The sender sends one packet and waits for acknowledgement. If this acknowledgement arrives, the sender increases the congestion window by one, now sending two packets (congestion window = 2). This scheme doubles the congestion window every time the acknowledgements come back, which takes one round trip time (RTT). This is called the exponential growth of the congestion window in the slow start mechanism.

But doubling the congestion window is too dangerous. The exponential growth stops at the **congestion threshold**. As soon as the congestion window reaches the congestion threshold, further increase of the transmission rate is only linear by adding 1 to the congestion window each time the acknowledgements come back.

Linear increase continues

until a time-out at the sender occurs due

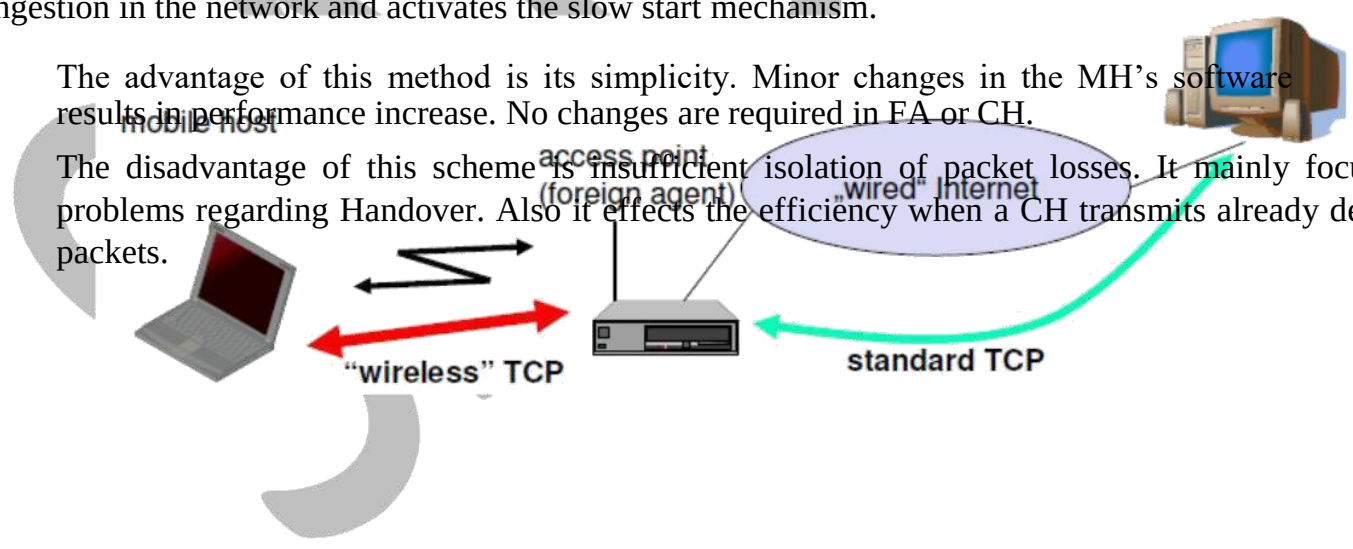
to a missing acknowledgement, or until the sender detects a gap in transmitted data because of continuous acknowledgements for the same packet. In either case the sender sets the congestion threshold to half of the current congestion window. The congestion window itself is set to one segment and the sender starts sending a single segment. The exponential growth starts once more up to the new congestion threshold, then the window grows in linear fashion.

Fast retransmit/fast recovery

The congestion threshold can be reduced because of two reasons. First one is if the sender receives continuous acknowledgements for the same packet. It informs the sender that the receiver has got all the packets upto the acknowledged packet in the sequence and also the receiver is receiving something continuously from the sender. The gap in the packet stream is not due to congestion, but a simple packet loss due to a transmission error. The sender can now retransmit the missing packet(s) before the timer expires. This behavior is called **fast retransmit**. It is an early enhancement for preventing slow-start to trigger on losses not caused by congestion. The receipt of acknowledgements shows that there is no congestion to justify a slow start. The sender can continue with the current congestion window. The sender performs a **fast recovery** from the packet loss. This mechanism can improve the efficiency of TCP dramatically. The other reason for activating slow start is a time-out due to a missing acknowledgement. TCP using fast retransmit/fast recovery interprets this congestion in the network and activates the slow start mechanism.

The advantage of this method is its simplicity. Minor changes in the MH's software results in performance increase. No changes are required in FA or CH.

The disadvantage of this scheme is insufficient isolation of packet losses. It mainly focuses on problems regarding Handover. Also it effects the efficiency when a CH transmits already delivered packets.



Problems with Traditional TCP in wireless environments

Slow Start mechanism in fixed networks decreases the efficiency of TCP if used with mobile receivers or senders.

Error rates on wireless links are orders of magnitude higher compared to fixed fiber or copper links. This makes compensation for packet loss by TCP quite difficult.

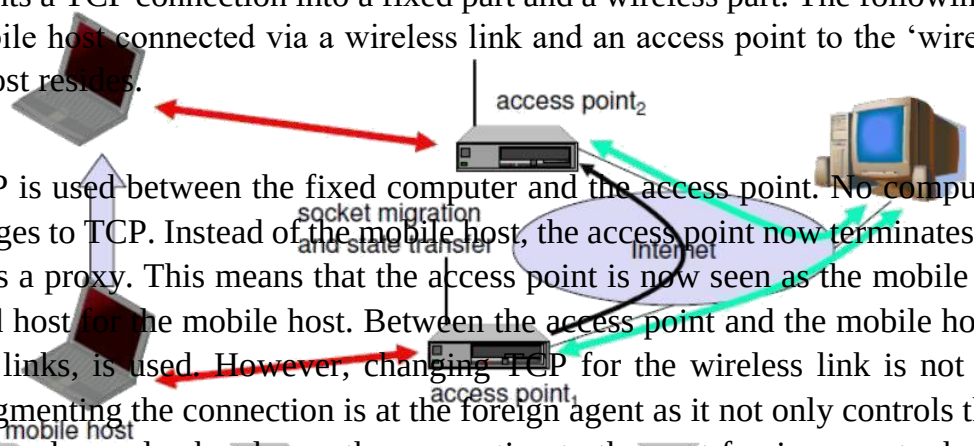
Mobility itself can cause packet loss. There are many situations where a soft handover from one access point to another is not possible for a mobile end-system.

Standard TCP reacts with slow start if acknowledgements are missing, which does not help in the case of transmission errors over wireless links and which does not really help during handover. This behavior results in a severe performance degradation of an unchanged TCP if used together with wireless links or mobile nodes

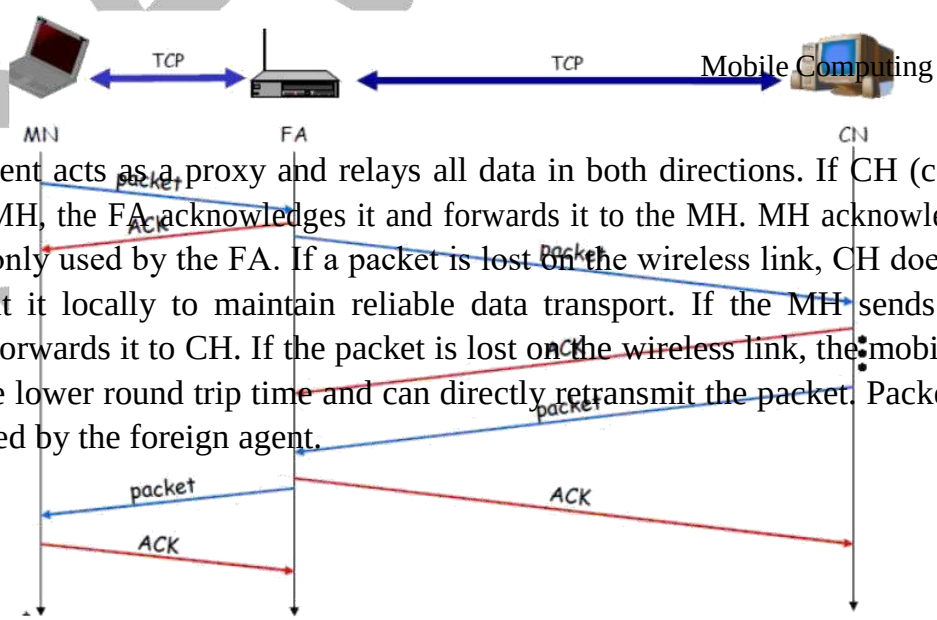
Classical TCP Improvements

Indirect TCP (I-TCP)

Indirect TCP segments a TCP connection into a fixed part and a wireless part. The following figure shows an example with a mobile host connected via a wireless link and an access point to the 'wired' internet where the correspondent host resides.



Standard TCP is used between the fixed computer and the access point. No computer in the internet recognizes any changes to TCP. Instead of the mobile host, the access point now terminates the standard TCP connection, acting as a proxy. This means that the access point is now seen as the mobile host for the fixed host and as the fixed host for the mobile host. Between the access point and the mobile host, a special TCP, adapted to wireless links, is used. However, changing TCP for the wireless link is not a requirement. A suitable place for segmenting the connection is at the foreign agent as it not only controls the mobility of the mobile host anyway and can also hand over the connection to the next foreign agent when the mobile host moves on.



The foreign agent acts as a proxy and relays all data in both directions. If CH (correspondent host) sends a packet to the MH, the FA acknowledges it and forwards it to the MH. MH acknowledges on successful reception, but this is only used by the FA. If a packet is lost on the wireless link, CH doesn't observe it and FA tries to retransmit it locally to maintain reliable data transport. If the MH sends a packet, the FA acknowledges it and forwards it to CH. If the packet is lost on the wireless link, the mobile hosts notice this much faster due to the lower round trip time and can directly retransmit the packet. Packet loss in the wired network is now handled by the foreign agent.

Socket and state migration after handover of a mobile host

During handover, the buffered packets, as well as the system state (packet sequence number, acknowledgements, ports, etc), must migrate to the new agent. No new connection may be established for the mobile host, and the correspondent host must not see any changes in connection state. Packet delivery in I-TCP is shown below:

Advantages of I-TCP

- No changes in the fixed network necessary, no changes for the hosts (TCP protocol) necessary, all current optimizations to TCP still work
- Simple to control, mobile TCP is used only for one hop between, e.g., a foreign agent and mobile host
  1. transmission errors on the wireless link do not propagate into the fixed network
  2. therefore, a very fast retransmission of packets is possible, the short delay on the mobile hop s known
- It is always dangerous to introduce new mechanisms in a huge network without knowing exactly how they behave.
  - New optimizations can be tested at the last hop, without jeopardizing the stability of the Internet.
- It is easy to use different protocols for wired and wireless networks.

Disadvantages of I-TCP

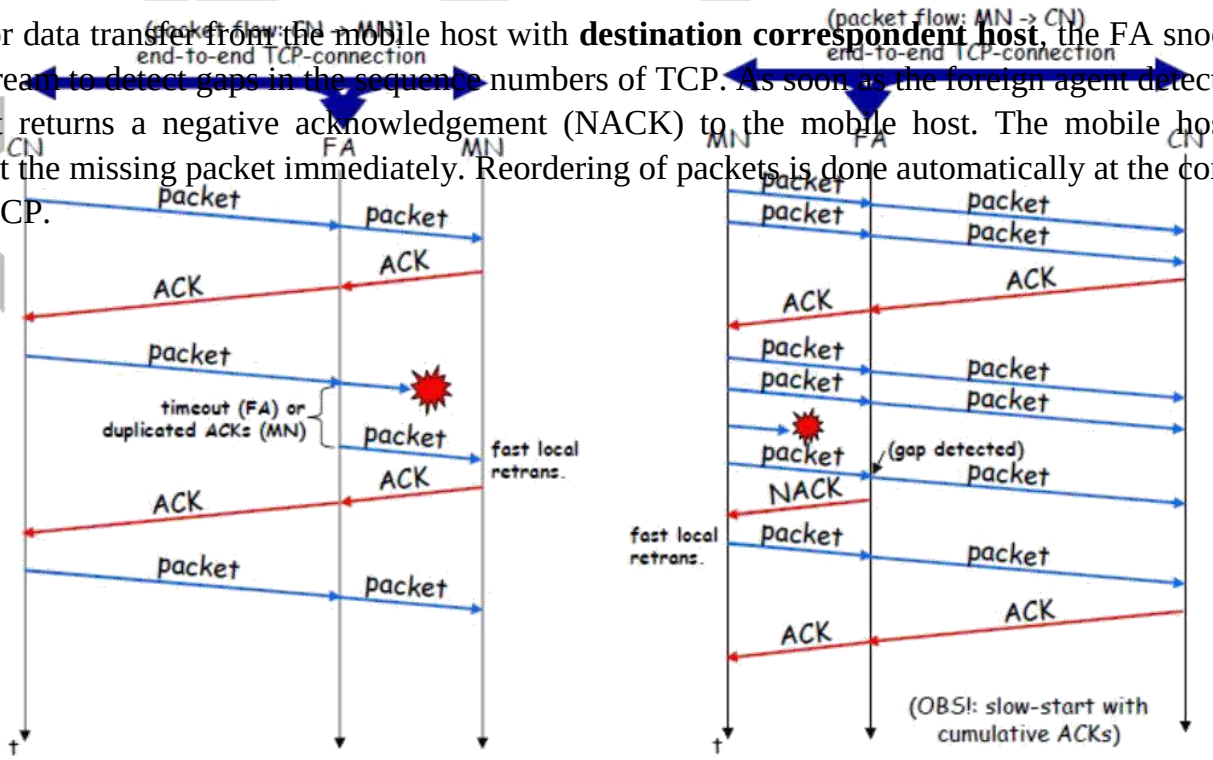
- Loss of end-to-end semantics:- an acknowledgement to a sender no longer means that a receiver really has received a packet, foreign agents might crash.
- Higher latency possible:- due to buffering of data within the foreign agent and forwarding to a new foreign agent
- Security issue:- The foreign agent must be a trusted entity

Snooping TCP

The main drawback of I-TCP is the segmentation of the single TCP connection into two TCP connections, which loses the original end-to-end TCP semantic. A new enhancement, which leaves the TCP connection intact and is completely transparent, is Snooping TCP. The main function is to buffer data close to the mobile host to perform fast local retransmission in case of packet loss.

Here, the foreign agent buffers all packets with **destination mobile host** and additionally 'snoops' the packet flow in both directions to recognize acknowledgements. The foreign agent buffers every packet until it receives an acknowledgement from the mobile host. If the FA does not receive an acknowledgement from the mobile host within a certain amount of time, either the packet or the acknowledgement has been lost. Alternatively, the foreign agent could receive a duplicate ACK which also shows the loss of a packet. Now, the FA retransmits the packet directly from the buffer thus performing a faster retransmission compared to the CH. For transparency, the FA does not acknowledge data to the CH, which would violate end-to-end semantic in case of a FA failure. The foreign agent can filter the duplicate acknowledgements to avoid unnecessary retransmissions of data from the correspondent host. If the foreign agent now crashes, the time-out of the correspondent host still works and triggers a retransmission. The foreign agent may discard duplicates of packets already retransmitted locally and acknowledged by the mobile host. This avoids unnecessary traffic on the wireless link.

For data transfer from the mobile host with **destination correspondent host**, the FA snoops into the packet stream to detect gaps in the sequence numbers of TCP. As soon as the foreign agent detects a missing packet, it returns a negative acknowledgement (NACK) to the mobile host. The mobile host can now retransmit the missing packet immediately. Reordering of packets is done automatically at the correspondent host by TCP.



Snooping TCP: Packet delivery

Advantages of snooping TCP:

- The end-to-end TCP semantic is preserved.
- Most of the enhancements are done in the foreign agent itself which keeps correspondent host unchanged.
- Handover of state is not required as soon as the mobile host moves to another foreign agent. Even though packets are present in the buffer, time out at the CH occurs and the packets are transmitted to the new COA.
- No problem arises if the new foreign agent uses the enhancement or not. If not, the approach automatically falls back to the standard solution.

Disadvantages of snooping TCP

- Snooping TCP does not isolate the behavior of the wireless link as well as I -TCP. Transmission errors may propagate till CH.
- Using negative acknowledgements between the foreign agent and the mobile host assumes additional mechanisms on the mobile host. This approach is no longer transparent for arbitrary mobile hosts.
- Snooping and buffering data may be useless if certain encryption schemes are applied end-to- end between the correspondent host and mobile host. If encryption is used above the transport layer, (eg. SSL/TLS), snooping TCP can be used.

Mobile TCP

Both I-TCP and Snooping TCP does not help much, if a mobile host gets disconnected. The **M-TCP (mobile TCP)** approach has the same goals as I-TCP and snooping TCP: to prevent the sender window from shrinking if bit errors or disconnection but not congestion cause current problems. M-TCP wants to improve overall throughput, to lower the delay, to maintain end-to- end semantics of TCP, and to provide a more efficient handover. Additionally, M-TCP is especially adapted to the problems arising from lengthy or frequent disconnections. M-TCP splits the TCP connection into two parts as I-TCP does. An unmodified TCP is used on the standard host-**supervisory host (SH)** connection, while an optimized TCP is used on the SH-MH connection.

The SH monitors all packets sent to the MH and ACKs returned from the MH. If the SH does not receive an ACK for some time, it assumes that the MH is disconnected. It then chokes the sender by setting the sender's window size to 0. Setting the window size to 0 forces the sender to go into **persistent mode**, i.e., the state of the sender will not change no matter how long the receiver is disconnected. This means that the sender will not try to retransmit data. As soon as the SH (either the old SH or a new SH) detects connectivity again, it reopens the window of the sender to the old value. The sender can continue sending at full speed.

This mechanism does not require changes to the sender's TCP. The wireless side uses an adapted 8

TCP that can recover from packet loss much faster. This modified TCP does not use slow start, thus, M-TCP needs a **bandwidth manager** to implement fair sharing over the wireless link.

Advantages of M-TCP:

- It maintains the TCP end-to-end semantics. The SH does not send any ACK itself but forwards the ACKs from the MH.
- If the MH is disconnected, it avoids useless retransmissions, slow starts or breaking connections by simply shrinking the sender's window to 0.
- As no buffering is done as in I-TCP, there is no need to forward buffers to a new SH. Lost packets will be automatically retransmitted to the SH.

Disadvantages of M-TCP:

- As the SH does not act as proxy as in I-TCP, packet loss on the wireless link due to bit errors is propagated to the sender. M-TCP assumes low bit error rates, which is not always a valid assumption.
- A modified TCP on the wireless link not only requires modifications to the MH protocol software but also new network elements like the bandwidth manager.

Transmission/time-out freezing

Often, MAC layer notices connection problems even before the connection is actually interrupted from a TCP point of view and also knows the real reason for the interruption. The MAC layer can inform the TCP layer of an upcoming loss of connection or that the current interruption is not caused by congestion. TCP can now stop sending and 'freezes' the current state of its congestion window and further timers. If the MAC layer notices the upcoming interruption early enough, both the mobile and correspondent host can be informed. With a fast interruption of the wireless link, additional mechanisms in the access point are needed to inform the correspondent host of the reason for interruption. Otherwise, the correspondent host goes into slow start assuming congestion and finally breaks the connection.

As soon as the MAC layer detects connectivity again, it signals TCP that it can resume operation at exactly the same point where it had been forced to stop. For TCP time simply does not advance, so no timers expire.

Advantages:

- It offers a way to resume TCP connections even after long interruptions of the connection.
- It can be used together with encrypted data as it is independent of other TCP mechanisms such as sequence no or acknowledgements

Disadvantages:

- Lots of changes have to be made in software of MH, CH and FA.

Selective retransmission

A very useful extension of TCP is the use of selective retransmission. TCP acknowledgements are cumulative, i.e., they acknowledge in-order receipt of packets up to a certain packet. A single acknowledgement confirms reception of all packets upto a certain packet. If a single packet is lost, the sender has to retransmit everything starting from the lost packet (go-back-n retransmission). This obviously wastes bandwidth, not just in the case of a mobile network, but for any network.

Using selective retransmission, TCP can indirectly request a selective retransmission of packets. The receiver can acknowledge single packets, not only trains of in-sequence packets. The sender can now determine precisely which packet is needed and can retransmit it. The **advantage** of this approach is obvious: a sender retransmits only the lost packets. This lowers bandwidth requirements and is extremely helpful in slow wireless links. The disadvantage is that a more complex software on the receiver side is needed. Also more buffer space is needed to resequence data and to wait for gaps to be filled.

Transaction-oriented TCP

Assume an application running on the mobile host that sends a short request to a server from time to time, which responds with a short message and it requires reliable TCP transport of the packets. For it to use normal TCP, it is inefficient because of the overhead involved. Standard TCP is made up of three phases: setup, data transfer and release. First, TCP uses a three-way handshake to establish the connection. At least one additional packet is usually needed for transmission of the request, and requires three more packets to close the connection via a three-way handshake. So, for sending one data packet, TCP may need seven packets altogether. This kind of overhead is acceptable for long sessions in fixed networks, but is quite inefficient for short messages or sessions in wireless networks. This led to the development of transaction-oriented TCP (T/TCP).

Approach	Mechanism	Advantages	Disadvantages
Indirect TCP	splits TCP connection into two connections	isolation of wireless link, simple	loss of TCP semantics, higher latency at handover
Snooping TCP	"snoops" data and acknowledgements, local retransmission	transparent for end-to-end connection, MAC integration possible	problematic with encryption, bad isolation of wireless link
M-TCP	splits TCP connection, chokes sender via window size	Maintains end-to-end semantics, handles long term and frequent disconnections	Bad isolation of wireless link, processing overhead due to bandwidth management
Fast retransmit/ fast recovery	retransmission initiated by roaming	simple and efficient	mixed layers, not transparent
Transmission/ time-out freezing	freezes TCP state at disconnect, resumes after reconnection	independent of content, applicable also for encryption, works for longer interrupts	TCP must be informed about disconnect/reconnect, MAC dependant
Selective retransmission	retransmit only lost data	very efficient	slightly more complex receiver, sender must buffer needed
Transaction oriented TCP	combine connection setup/release and data transmission	Efficient for certain applications	changes in TCP required, not transparent

T/TCP can combine packets for connection establishment and connection release with user data packets. This can reduce the number of packets down to two instead of seven. The obvious **advantage** for certain applications is the reduction in the overhead which standard TCP has for connection setup and connection release. Disadvantage is that it requires changes in the software in mobile host

and all correspondent hosts. This solution does not hide mobility anymore. Also, T/TCP exhibits several security problems.

Classical Enhancements to TCP for mobility: A comparison

Databaseissues: Hoarding techniques, caching invalid ation mech anisms. client server computing with adapt ion , power-aware and context-aware computing, transactional models, query processing, recovery, and quality of service issues

A database is a collection of systematically stored records or information. Databases store data in a particular logical manner. A mobile device is not always connected to the server or network; neither does the device retrieve data from a server or a network for each computation. Rather, the device caches some specific data, which may be required for future computations, during the interval in which the device is connected to the server or network. Caching entails saving a copy of select data or a part of a database from a connected system with a large database. The cached data is hoarded in the mobile device database. Hoarding of the cached data in the database ensures that even when the device is not connected to the network, the data required from the database is available for computing.

Database Hoarding

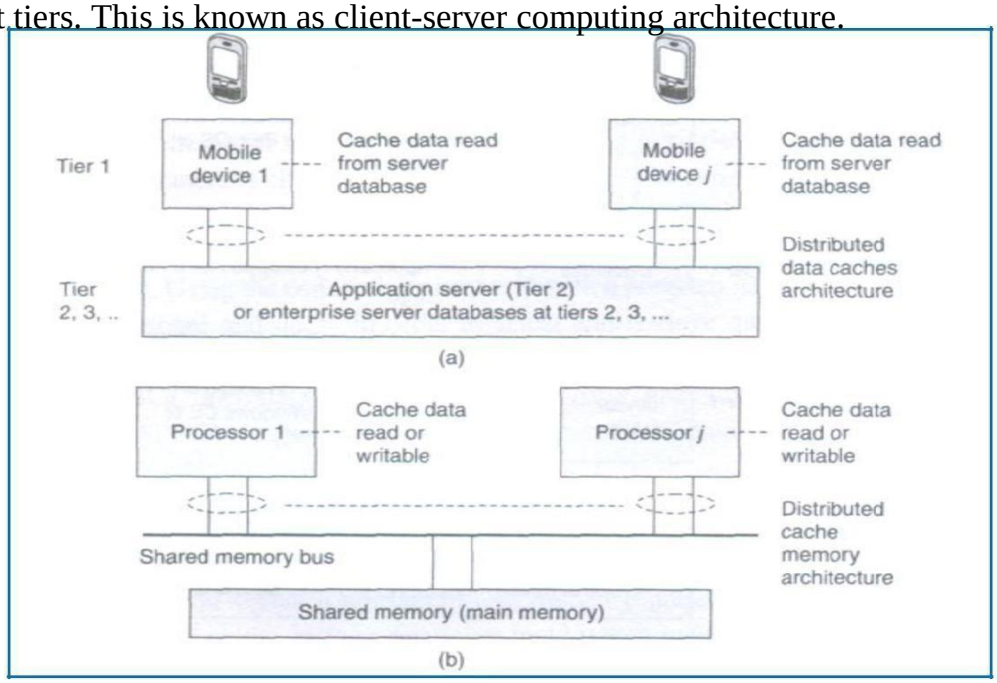
Database hoarding may be done at the application tier itself. The following figure shows a simple architecture in which a mobile device API directly retrieves the data from a database. It also shows another simple architecture in which a mobile device API directly retrieves the data from a database through a program, for ex: IBM DB2 Everyplace (DB2e).

(a) API at mobile device sending queries and retrieving data from local database (Tier 1)

(b) API at mobile device retrieving data from database using DB2e (Tier 1)

Both the two architectures belong to the class of one-tier database architecture because the databases are specific to a mobile device, not meant to be distributed to multiple devices, not synchronized with the new updates, are stored at the device itself. Some examples are downloaded ringtones, music etc. **IBM DB2 Everyplace (DB2e)** is a relational database engine which has been designed to reside at the device. It supports J2ME and most mobile device operating systems. DB2e synchronizes with DB2 databases at the synchronization, application, or enterprise server

The database architecture shown below is for two-tier or multi-tier databases. Here, the databases reside at the remote servers and the copies of these databases are cached at the client tiers. This is known as client-server computing architecture.



(a) Distributed data caches in mobile devices
(b) Similar architecture for a distributed cache memory in multiprocessor systems

A cache is a list or database of items or records stored at the device. Databases are hoarded at the application or enterprise tier, where the database server uses business logic and connectivity for retrieving the data and then transmitting it to the device. The server provides and updates local copies of the database at each mobile device connected to it. The computing API at the mobile device (first tier) uses the cached local copy. At first tier (tier 1), the API uses the cached data records using the computing architecture as explained above. From tier 2 or tier 3, the server retrieves and transmits the data records to tier 1 using business logic and synchronizes the local copies at the device. These local copies function as device caches.

The advantage of hoarding is that there is no access latency (delay in retrieving the queried record from the server over wireless mobile networks). The client device API has instantaneous data access to hoarded or cached data. After a device caches the data distributed by the server, the data is hoarded at the device. The disadvantage of hoarding is that the consistency of the cached data with the database at the server needs to be maintained.

Data Caching

Hoarded copies of the databases at the servers are distributed or transmitted to the mobile devices from the enterprise servers or application databases. The copies cached at the devices are equivalent to the cache memories at the processors in a multiprocessor system with a shared main memory and copies of the main memory data stored at different locations.

Cache Access Protocols: A client device caches the pushed (disseminated) data records from a server. Caching of the pushed data leads to a reduced access interval as compared to the pull (on- demand) mode of data fetching. Caching of data records can be-based on pushed 'hot records' (the most needed database records at the client device). Also, caching can be based on the ratio of two parameters—access probability (at the device) and pushing rates (from the server) for each record. This method is called cost-based data replacement or caching.

Pre-fetching: Pre-fetching is another alternative to caching of disseminated data. The process of pre-fetching entails requesting for and pulling records that may be required later. The client device can pre-fetch instead of caching from the pushed records keeping future needs in view. Pre-fetching reduces server load. Further, the cost of cache-misses can thus be reduced. The term 'cost of cache-misses' refers to the time taken in accessing a record at the server in case that record is not found in the device database when required by the device API.

Caching Invalidation Mechanisms

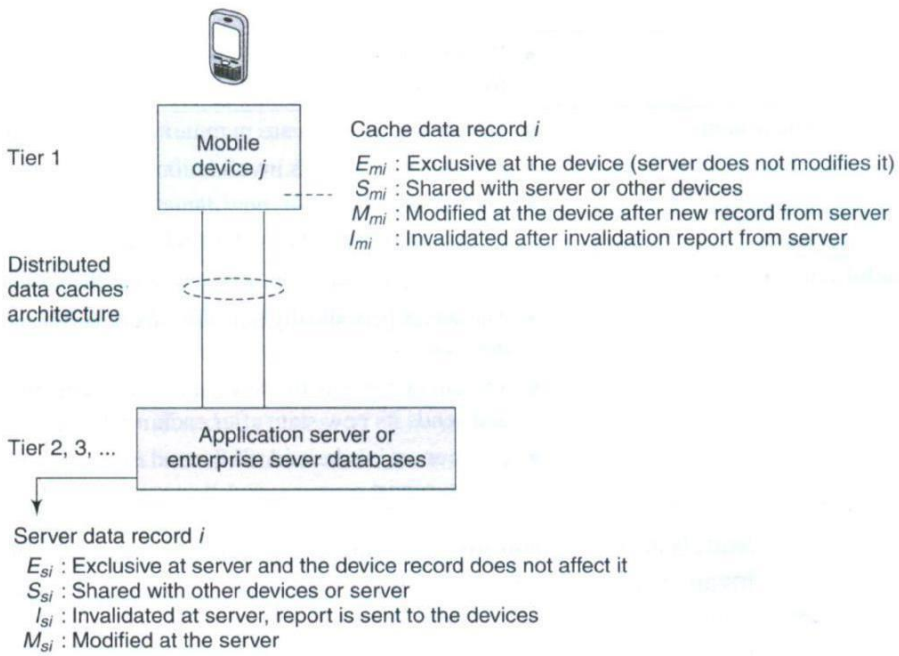
A cached record at the client device may be invalidated. This may be due to expiry or modification of the record at the database server. Cache invalidation is a process by which a cached data item or record becomes invalid and thus unusable because of modification, expiry, or invalidation at another computing system or server. Cache invalidation mechanisms are used to synchronize the data at other processors whenever the cache-data is written (modified) by a processor in a multiprocessor system, cache invalidation mechanisms are also active in the case of mobile devices having distributed copies from the server.

A cache consists of several records. Each record is called a cache-line, copies of which can be stored at other devices or servers. The cache at the mobile devices or server databases

given time can be assigned one of four possible tags indicating its state—modified (after rewriting), exclusive, shared, and invalidated (after expiry or when new data becomes available) at any given instance. These four states are indicated by the letters M, E, S, and I, respectively (MESI). The states indicated by the various tags are as follows:

a) The ***E*** tag indicates the *exclusive* state which means that the data record is for internal use and cannot be used by any other device or server.
b) The ***S*** tag indicates the *shared* state which indicates that the data record can be used by others.
c) The ***M*** tag indicates the *modified* state which means that the device cache
d) The ***I*** tag indicates the *invalidated state* which means that the server database no longer has a copy of the record which was shared and used for computations earlier.

The following figure shows the four possible states of a data record *i* at any instant in the server database and its copy at the cache of the mobile device *j*.



Four possible states (M, E, S, or /) of a data record /at any instance at the server database and device j cache

Another important factor for cache maintenance in a mobile environment is *cache consistency* (also called *cache coherence).* This requires a mechanism to ensure that a database record is identical at the server as well as at the device caches and that only the valid cache records are used for computations.

Cache invalidation mechanisms in mobile devices are triggered or initiated by the server. There are four possible invalidation mechanisms – Stateless asynchronous, stateless synchronous, stateful asynchronous and stateful synchronous.

Stateless Asynchronous: A stateless mechanism entails broadcasting of the invalidation of the cache to all the clients of the server. The server does not keep track of the records stored at the device caches. It just uniformly broadcasts invalidation reports to all clients irrespective of whether the device cache holds that particular record or not. The term 'asynchronous' indicates that the invalidation information for an item is sent as soon as its value changes. The server does not keep the information of the present state (whether E_{mi}, M_{mi}, S_{mi}, or I_{mi}) of a data-record in cache for broadcasting later. The server advertises the invalidation information only. The client can either request for a modified copy of the record or cache the relevant record when data is pushed from the server. The server advertises as and when the corresponding data-record at the server is invalidated and modified (deleted or replaced).

The advantage of the asynchronous approach is that there are no frequent, unnecessary transfers of data reports, thus making the mechanism more bandwidth efficient. The disadvantages of this approach are—(a) every client device gets an invalidation report, whether that client requires that copy or not and (b) client devices presume that as long as there is no invalidation report, the copy is valid for use in computations. Therefore, even when there is link failure, the devices may be using the invalidated data and the server is unaware of state changes at the clients after it sends the invalidation report.

Stateless Synchronous This is also a stateless mode, i.e., the server has no information regarding the present state of data records at the device caches and broadcasts to all client devices. However, unlike the asynchronous mechanism, here the server advertises invalidation information at periodic intervals as well as whenever the corresponding data-record at server is invalidated or modified. This method ensures synchronization because even if the in-between period report is not detected by the device due to a link failure, the device expects the period-end report of invalidation and if that is not received at the end of the period, then the device sends a request for the same (deleted or replaced). In case the client device does not get the periodic report due to link failure, it requests the server to send the report.

The advantage of the synchronous approach is that the client devices receive periodic information regarding invalidity (and thus validity) of the data caches. The periodic invalidation reports lead to greater reliability of cached data as update requests for invalid data can be sent to the server by the device-client. This also helps the server and devices maintain cache consistency through periodical exchanges. The disadvantages of this mode of cache invalidation are—(a) unnecessary transfers of data invalidation reports take place, (b) every client device gets an advertised invalidation report periodically, irrespective of whether that client has a copy of the invalidated data or not, and (c) during

the period between two invalidation reports, the client

devices assume that, as long as there is no invalidation report, the copy is valid for use in computations. Therefore, when there are link failures, the devices use data which has been invalidated in the in-between period and the server is unaware of state changes at the clients after it sends the invalidation report.

Stateful Asynchronous The stateful asynchronous mechanism is also referred to as the AS (asynchronous stateful) scheme. The term 'stateful' indicates that the cache invalidation reports are sent only to the affected client devices and not broadcasted to all. The server stores the information regarding the present state (a record I can have its state as E_{mi}, M_{mi}, S_{mi}, or I_{mi}) of each data-record at the client device caches. This state information is stored in the home location cache (HLC) at the server. The HLC is maintained by an HA (home agent) software. This is similar to the HLR at the MSC in a mobile network. The client device informs the HA of the state of each record to enable storage of the same at the HLC. The server transmits the invalidation information as and when the records are invalidated and it transmits only to the device-clients which are affected by the invalidation of data. Based on the invalidation information, these device-clients then request the server for new or modified data to replace the invalidated data. After the data records transmitted by the server modify the client device cache, the device sends information about the new state to the server so that the record of the cache-states at the server is also modified.

The advantage of the stateful asynchronous approach is that the server keeps track of the state of cached data at the client device. This enables the server to synchronize with the state of records at the device cache and keep the HLC updated. The stateful asynchronous mode is also advantageous in that only the affected clients receive the invalidation reports and other devices are not flooded with irrelevant reports. The disadvantage of the AS scheme is that the client devices presume that, as long as there is no invalidation report, the copy is valid for use in computations. Therefore, when there is a link failure, then the devices use invalidated data.

Stateful Synchronous: The server keeps the information of the present state (E_{mi}, M_{mi}, S_{mi}, or I_{mi}) of data-records at the client-caches. The server stores the cache record state at the home location cache (HLC) using the home agent (HA). The server transmits the invalidation information at periodic intervals to the clients and whenever the data-record relevant to the client is invalidated or modified (deleted or replaced) at the server. This method ensures synchronization because even if thein-between period report is not detected by the device due to a link failure, the device expects the period-end report of invalidation and if it is not received at the end of the period, then the device requests for the same.

The advantage of the stateful synchronous approach is that there are reports identifying invalidity (and thus, indirectly, of validity) of data caches at periodic intervals and that the server also periodically updates the client-cache states stored in the HLC. This enables to synchronize with the client device when invalid data gets modified and becomes valid. Moreover, since the invalidation report is sent periodically, if a device does not receive an invalidation report after

the specified period of time, it can request the server to send the report. Each client can thus be periodically updated of any modifications at the server. When the invalidation report is not received after the designated period and a link failure is found at the device, the device does not use the invalidated data. Instead it requests the server for an invalidation update. The disadvantage of the stateful synchronous approach is the high bandwidth requirement to enable periodic transmission of invalidation reports to each device and updating requests from each client device.

Data Cache Maintenance in Mobile Environments

Assume that a device needs a data-record during an application. A request must be sent to the server for the data record (this mechanism is called pulling). The time taken for the application software to access a particular record is known as *access latency*. Caching and hoarding the record at the device reduces access latency to zero. Therefore, data cache maintenance is necessary in a mobile environment to overcome access latency.

Data cache inconsistency means that data records cached for applications are not invalidated at the device when modified at the server but not modified at the device. Data cache consistency can be maintained by the three methods given below:

I. *Cache invalidation mechanism* (server-initiated case): the server sends invalidation reports on invalidation of records (asynchronous) or at regular intervals (synchronous).

II. *Polling mechanism* (client-initiated case): Polling means checking from the server, the state of data record whether the record is in the valid, invalid, modified, or exclusive state. Each cached record copy is polled whenever required by the application software during computation. If the record is found to be modified or invalidated, then the device requests for the modified data and replaces the earlier cached record copy.

III. *Time-to-live mechanism* (client-initiated case): Each cached record is assigned a TTL (time-to-live). The TTL assignment is adaptive (adjustable) previous update intervals of that record. After the end of the TTL, the cached record copy is polled. If it is modified, then the device requests the server to replace the invalid cached record with the modified data. When TTL is set to 0, the TTL mechanism is equivalent to the polling mechanism.

Web Cache Maintenance in Mobile Environments

The mobile devices or their servers can be connected to a web server (e.g., traffic information server or train information server). Web cache at the device stores the web server data and maintains it in a manner similar to the cache maintenance for server data described above. If an application running at the device needs a data record from the web which is not at the web cache, then there is access latency. Web cache maintenance is necessary in a mobile environment to overcome access latency in downloading from websites due to disconnections. Web cache consistency can be maintained by two methods. These are:

Time-to-live (TTL) mechanism (client-initiated case): The method is identical to the one discussed for data cache maintenance.

I. Power-aware computing mechanism (client-initiated case): Each web cache maintained at the device can also store the CRC (cyclic redundancy check) bits. Assume that there are N cached bits and *n* CRC bits. N is much greater than *n.* Similarly at the server, *n* CRC bits are stored. As long as there is consistency between the server and device records, the CRC bits at both are identical. Whenever any of the records cached at the server is modified, the corresponding CRC bits at the server are also modified. After the TTL expires or on-demand for the web cache records by the client API, the cached record CRC is polled and obtained from the website server. If the *n* CRC bits at server are found to be modified and the change is found to be much higher than a given threshold (i.e., a significant change), then the modified part of the website hypertext or database is retrieved by the client device for use by the API. However, if the change is minor, then the API uses the previous cache. Since $N \gg n$, the power dissipated in the web cache maintenance method (in which invalidation reports and all invalidated record bits are transmitted) is much greater than that in the present method (in which the device polls for the significant change in the CRC bits at server and the records are transmitted only when there is a significant change in the CRC bits).

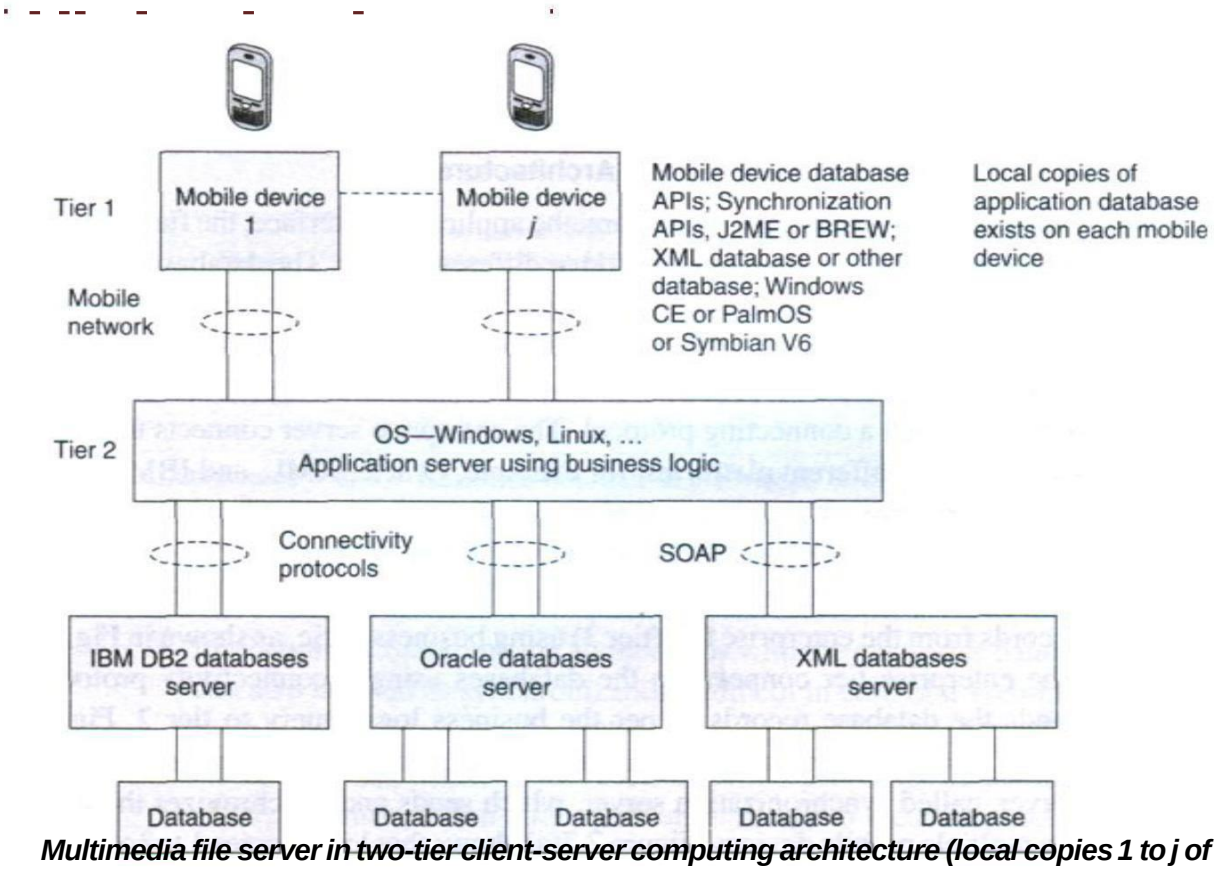
Client-Server Computing

Client-server computing is a distributed computing architecture, in which there are two types of nodes, i.e., the clients and the servers. A server is defined as a computing system, which responds to requests from one or more clients. A client is defined as a computing system, which requests the server for a resource or for executing a task. The client can either access the data records at the server or it can cache these records at the client device. The data can be accessed either on client request or through broadcasts or distribution from the server.

The client and the server can be on the same computing system or on different computing systems. Client-server computing can have N-tier architecture *(N*= 1, 2 ...). When the client and the server are on the same computing system then the number of tiers, $N = 1$. When the client and the server are on different computing systems on the network, then $N = 2$. A command interchange protocol (e.g., HTTP) is used for obtaining the client requests at the server or the server responses at the client.

The following subsections describe client-server computing in 2, 3, or N-tier architectures. Each tier connects to the other with a connecting, synchronizing, data, or command interchange protocol.

Two-tier Client-Server Architecture

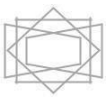


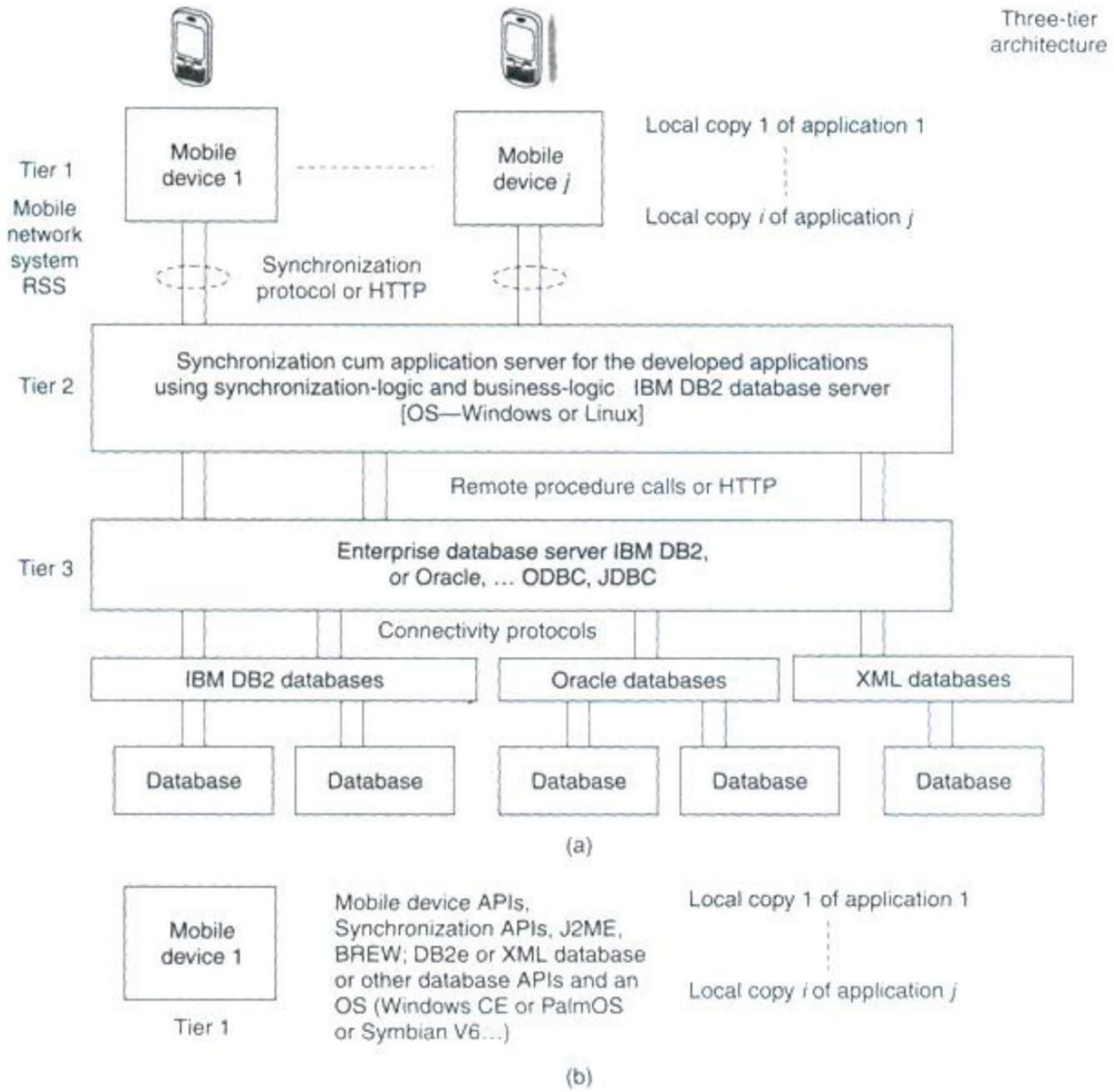
Multimedia file server in two-tier client-server computing architecture (local copies 1 to j of image and voice hoarding at the mobile devices)

The following figure shows the application server at the second tier. The data records are retrieved using business logic and a synchronization server in the application server synchronizes with the local copies at the mobile devices. Synchronization means that when copies of records at the server-end are modified, the copies cached at the client devices should also be accordingly modified. The APIs are designed independent of hardware and software platforms as far as possible as different devices may have different platforms.

Three-tier Client-Server Architecture

In a three-tier computing architecture, the application interface, the functional logic, and the database are maintained at three different layers. The database is associated with the enterprise server tier (tier 3) and only local copies of the database exist at mobile devices. The database connects to the enterprise server through a connecting protocol. The enterprise server connects the complete databases on different platforms, for example, Oracle, XML, and IBM DB2.



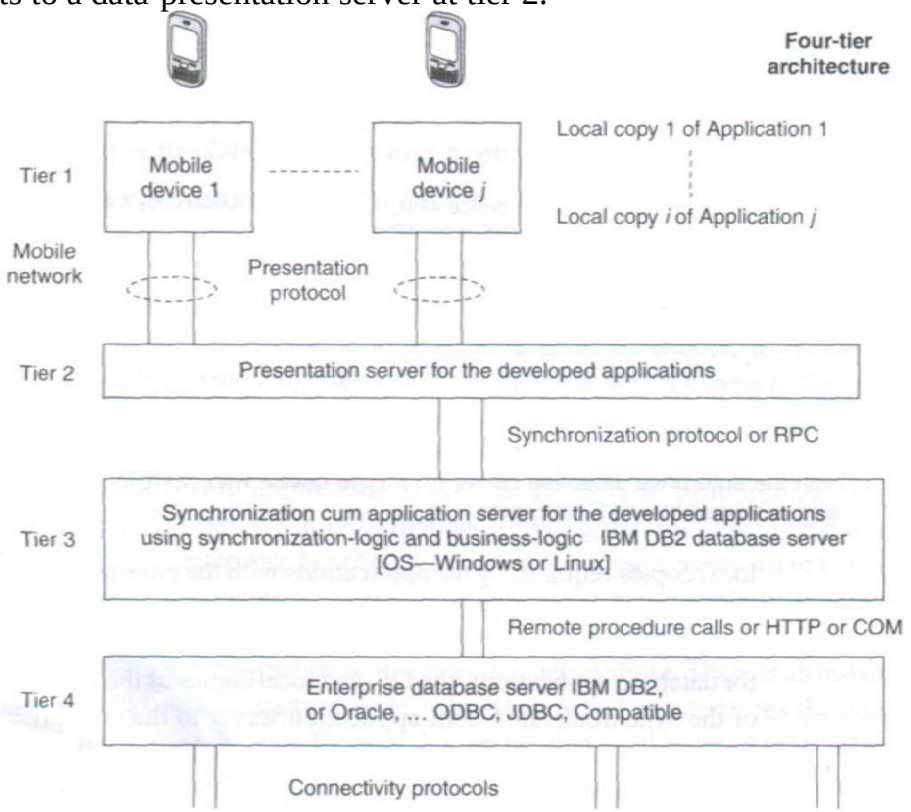


(a) Local copies 1 to j of database hoarded at the mobile devices using an enterprise database connection synchronization server, which synchronizes the required local copies for application with the enterprise database server (b) Mobile device with J2ME or BREW platform, APIs an OS and database having local copies

Data records at tier 3 are sent to tier 1 as shown in the figure through a synchronization-cum- application server at tier 2. The synchronization-cum-application server has synchronization and server programs, which retrieves data records from the enterprise tier (tier 3) using business logic. There is an in-between server, called synchronization server, which sends and synchronizes the copies at the multiple mobile devices. The figure shows that local copies 1 to j of databases are hoarded at the mobile devices for the applications 1 to j.

~~*N-tier Client-Server Architecture*~~

When *N* is greater than 3, then the database is presented at the client through in-between layers. For example, the following figure shows a four-tier architecture in which a client device connects to a data-presentation server at tier 2.



4-tier architecture in which a client device connects to a data-presentation server

The presentation server then connects to the application server tier 3. The application server can connect to the database using the connectivity protocol and to the multimedia server using Java or XML API at tier 4. The total number of tiers can be counted by adding 2 to the number of in-between servers between the database and the client device. The presentation, application, and enterprise servers can establish connectivity using RPC, Java RMI, JNDI, or HOP. These servers may also use HTTP or HTTPS in case the server at a *tier j* connects to tier j+1 using the Internet.

Client-Server Computing with Adaptation

The data formats of data transmitted from the synchronization server and those required for the device database and device APIs are different in different cases, there are two adapters at a

mobile device—an adapter for standard data format for synchronization at the mobile device and another adapter for the backend database copy, which is in a different data format for the API at the mobile device. An adapter is software to get data in one format

or data governed by one protocol and convert it to another format or to data governed by another protocol.

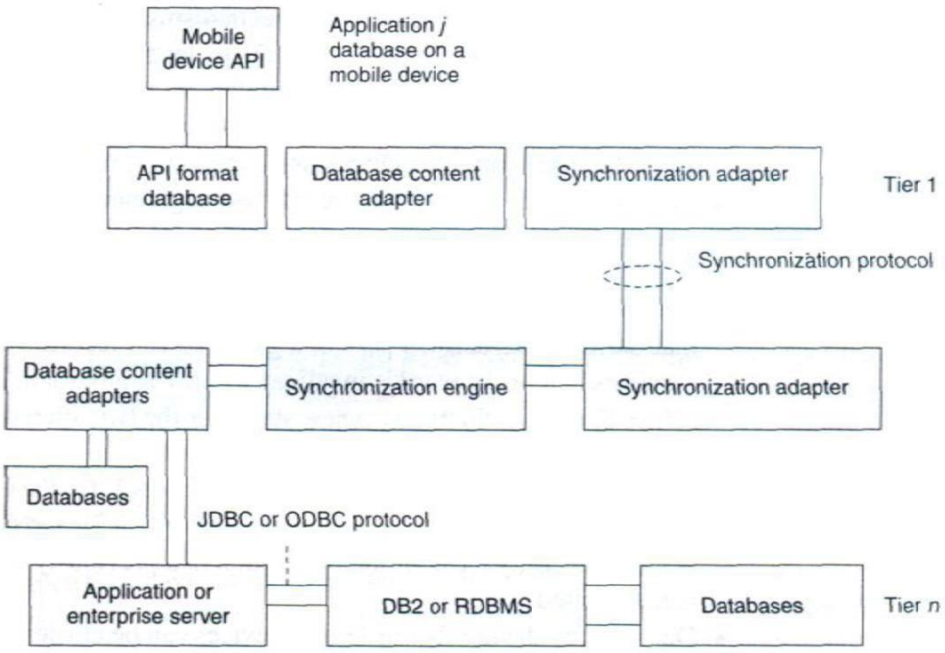


Figure shows an API, database, and adapters at a mobile device and the adapters at the synchronization, application, or enterprise servers. Here the adapters are an addition used for interchange between standard data formats and data formats for the API.

Transaction Models

A transaction is the execution of interrelated instructions in a sequence for a specific operation on a database. Database transaction models must maintain data integrity and must enforce a set of rules called ACID rules. These rules are as follows:

- ***Atomicity****:* All operations of a transaction must be complete. In case, a transaction cannot be completed; it must be undone (rolled back). Operations in a transaction are assumed to be one indivisible unit (atomic unit).
- ***Consistency****:* A transaction must be such that it preserves the integrity constraints and follows the declared consistency rules for a given database. Consistency means the data is not in a contradictory state after the transaction.
- ***Isolation****:* If two transactions are carried out simultaneously, there should not be any interference between the two. Further, any intermediate results in a transaction should be invisible to any other transaction.
- ***Durability:*** After a transaction is completed, it must persist and cannot be aborted or discarded. For example, in a transaction entailing transfer of a balance from account A to account B, once the transfer is completed and finished there should be no roll back.

Consider a base class library included in Microsoft.NET. It has a set of computer software components called ADO.NET (ActiveX Data Objects in .NET). These can be used to access the data and data services including for access and modifying the data stored in relational database systems. The ADO.NET transaction model permits three transaction commands:

1. `BeginTransaction`**: It is used to begin a transaction. Any operation after** `BeginTransaction` **is assumed to be a part of the transaction till the** `CommitTransaction` **command or the** `RollbackTransaction` **command. An example of a command is as follows:**

   ```
   connectionA.open();

   transA = connectionA.BeginTransaction();
   ```

   Here `connectionA` **and** `transA` **are two distinct objects.**

2. `Commit`**: It is used to commit the transaction operations that were carried out after the** `BeginTransaction` **command and up to this command. An example of this is**

   ```
   transA.Commit();
   ```

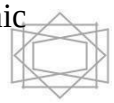
   All statements between `BeginTransaction` **and** `commit` **must execute automatically.**

3. `Rollback`**: It is used to rollback the transaction in case an exception is generated after the** `BeginTransaction` **command is executed.**

A DBMS may provide for auto-commit mode. *Auto-commit mode* means the transaction finished automatically even if an error occurs in between.

Query Processing

Query processing means making a correct as well as efficient execution strategy by *query decomposition* and *query-optimization*. A relational-algebraic equation defines a set of operations needed during query processing. Either of the two equivalent relational-algebraic equations given below can be used.

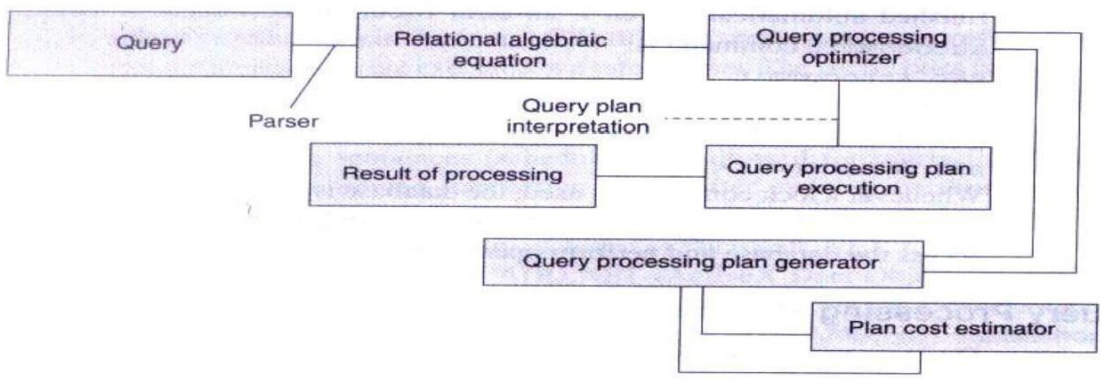


$$\pi_{\text{cName, cTelNum}}\ (\sigma_{\text{Contacts.firstChar = "R"}}\ (\sigma_{\text{Contacts.cTelNum = DialledNumbers.dTelNum}}\ (\text{Contacts}) \times \text{DialledNumbers}))$$

This means first select a column `Contacts.cTelNum` in a row in `Contacts` in which `Contacts.cTelNum` column equals a column `DialledNumbers.dTelNum` by crosschecking and matching the records of a column in `Contacts` with all the rows of `DialledNumbers`. Then in the second step select the row in which `Contacts.firstChar = "R"` and the selected `cTelNum` exists. Then in the third step project `cName` and `CTelNum`.

$$\pi_{\text{cName, cTelNum}}\ (\sigma_{\text{Contacts.firstChar = "R" } \wedge \text{ Contacts.cTelNum = DialledNumbers.dTelNum}}\ (\text{Contacts} \times \text{DialledNumbers}))$$

This means that in first series of step, crosscheck all rows of `Contacts` and `DialledNumbers` and select, after AND operation, the rows in which `Contacts.firstchar = "R"` and `Contacts.cTelNum = DialledNumbers.dTelNum`. Then in the next step project `cName` and `cTelNum` form the selected records.



Query processing architecture

Π represents the projection operation, σ the *selection* operation, and Λ, the AND operation. It is clear that the second set of operations in query processing is less efficient than the first. Query decomposition of the first set gives efficiency. Decomposition is done by (i) analysis, (ii) conjunctive and disjunctive normalization, and (iii) semantic analysis.

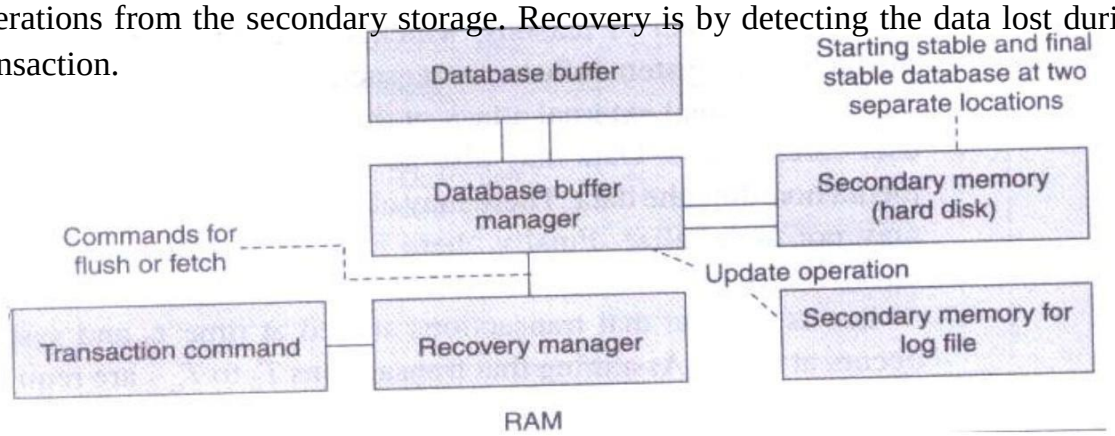
Efficient processing of queries needs optimization of steps for query processing. Optimization can be based on cost (number of micro-operations in processing) by evaluating the costs of sets of equivalent expressions. Optimization can also be based on a heuristic approach consisting of

the following steps: perform the selection steps and projection steps as early as possible and eliminate duplicate operations.

The query optimizer employs (a) query processing plan generator and (b) query processing cost estimator to provide an efficient plan for query processing.

Data Recovery Process

Data is non-recoverable in case of media failure, intentional attack on the database and transactions logging data, or physical media destruction. However, data recovery is possible in other cases. Figure below shows recovery management architecture. It uses a recovery manager, which ensures atomicity and durability. Atomicity ensures that an uncommitted but started transaction aborts on failure and aborted transactions are logged in log file. Durability ensures that a committed transaction is not affected by failure and is recovered. Stable state databases at the start and at the end of transactions reside in secondary storage. Transaction commands are sent to the recovery manager, which sends fetch commands to the database manager. The database manager processes the queries during the transaction and uses a database buffer. The recovery manager also sends the flush commands to transfer the committed transactions and database buffer data to the secondary storage. The recovery manager detects the results of operations. It recovers lost operations from the secondary storage. Recovery is by detecting the data lost during the transaction.



Recovery Management Architecture

The recovery manager uses a log file, which logs actions in the following manner:

1. Each instruction for a transaction for update (insertion, deletion, replacement, and addition) must be logged.
2. Database read instructions are not logged
3. Log files are stored at a different storage medium.
4. Log entries are flushed out after the final stable state database is stored.

Each logged entry contains the following fields.

- transaction type (begin, commit, or rollback transaction)
- transaction ID
- operation-type
- object on which the operation is performed
- pre-operation and post-operation values of the object.

A procedure called the Aries algorithm is also used for recovering lost data. The basic steps of the algorithm are:

I. Analyse from last checkpoint and identify all dirty records (written again after operation restarted) in the buffer.
II. Redo all buffered operations logged in the update log to finish and make final page.
III. Undo all write operations and restore pre-transaction values.

The recovery models used in data recovery processes are as follows:

I. The *full recovery model* creates back up of the database and incremental backup of the changes. All transactions are logged from the last backup taken for the database.
II. The *bulk logged recovery model* entails logging and taking backup of bulk data record operations but not the full logging and backup. Size of bulk logging is kept to the minimum required. This improves performance. We can recover the database to the point of failure by restoring the database with the bulk transaction log file backup. This is unlike the full recovery model in which all operations are logged.
III. The *simple recovery model* prepares full backups but the incremental changes are not logged. We can recover the database to the most recent backup of the given database.

QoS Issues:

Quality of service (QoS) mechanism controls the performance, reliability and usability of a telecommunications service. Mobile cellular service providers may offer **mobile QoS** to customers just as the fixed line PSTN services providers and Internet service providers may offer QoS. QoS mechanisms are always provided for circuit switched services, and are essential for non-elastic services, for example streaming multimedia. It is also essential in networks dominated by such services, which is the case in today's mobile communication networks, but not necessarily tomorrow.

Mobility adds complication to the QoS mechanisms, for several reasons:

- A phone call or other session may be interrupted after a handover, if the new base station is overloaded. Unpredictable handovers make it impossible to give an absolute QoS guarantee during a session initiation phase.
- The pricing structure is often based on per-minute or per-megabyte fee rather than flat rate, and may be different for different content services.
- A crucial part of QoS in mobile communications is grade of service, involving outage probability (the probability that the mobile station is outside the service coverage area, or affected by co-channel interference, i.e. crosstalk) blocking probability (the probability that the required level of QoS cannot be offered) and scheduling starvation. These performance measures are affected by mechanisms such as mobility management, radio resource management, admission control, fair scheduling, channel- dependent scheduling etc.

Types

- Factors affecting QoS
- Measurement of QoS
- Cellular GoS
- Cellular audio quality

Factors affecting QoS

Many factors affect the quality of service of a mobile network.[It is correct to look at QoS mainly from the customer's point of view, that is, QoS as judged by the user. There are standard metrics of QoS to the user that can be measured to rate the QoS. These metrics are: the **coverage**, **accessibility** (includes GoS), and the **audio quality**. In **coverage** the strength of the signal is measured using test equipment and this can be used to estimate the size of the cell. **Accessibility** is about determining the ability of the network to handle successful calls from mobile-to-fixed networks and from mobile-to-mobile networks. The **audio quality** considers monitoring a successful call for a period of time for the clarity of the communication channel. All these indicators are used by the telecommunications industry to rate the quality of service of a network.

Measurement of QoS

The QoS in industry is also measured from the perspective of an expert (e.g. teletraffic engineer). This involves assessing the network to see if it delivers the quality that the network planner has been

required to target. Certain tools and methods (protocol analysers, drive tests and Operation and Maintenance measurements), are used for this QoS measurement:

- Protocol analysers are connected to BTSs, BSCs, and MSCs for a period of time to check for problems in the cellular network. When a problem is discovered the staff can record it and it can be analysed.
- Drive tests allow the mobile network to be tested through the use of a team of people who take the role of users and take the QoS measures discussed above to rate the QoS of the network. This test does not apply to the entire network, so it is always a statistical sample.
- In the Operation and Maintenance Centres (OMCs), counters are used in the system for various events which provide the network operator with information on the state and quality of the network.
- Finally, customer complaints are a vital source of feedback on the QoS, and must not be ignored.

Cellular GoS

In general, grade of service (GoS) is measured by looking at traffic carried, traffic offered and calculating the traffic blocked and lost. The proportion of lost calls is the measure of GoS. For cellular circuit groups an acceptable GoS is 0.02. This means that two users of the circuit group out of a hundred will encounter a call refusal during the busy hour at the end of the planning period. The grade of service standard is thus the acceptable level of traffic that the network can lose. GoS is calculated from the Erlang-B formula, as a function of the number of channels required for the offered traffic intensity.

Cellular audio quality

The audio quality of a cellular network depends on, among other factors, the modulation scheme (e.g., FSK, QPSK) in use, matching to the channel characteristics and the processing of the received signal at the receiver using DSPs.

UNIT V

Data Dissemination

Data Dissemination and Synchronization : Communications Asymmetry, Classification of Data Delivery Mechanisms, Data Dissemination, Broadcast Models, Selective Tuning and Indexing Methods, Data Synchronization – Introduction, Software, and Protocols.

Ongoing advances in communications including the proliferation of internet, development of mobile and wireless networks, high bandwidth availability to homes have led to development of a wide range of new-information centered applications. Many of these applications involve data dissemination, i.e. delivery of data from a set of producers to a larger set of consumers.

Data dissemination entails distributing and pushing data generated by a set of computing systems or broadcasting data from audio, video, and data services. The output data is sent to the mobile devices. A mobile device can select, tune and cache the required data items, which can be used for application programs.

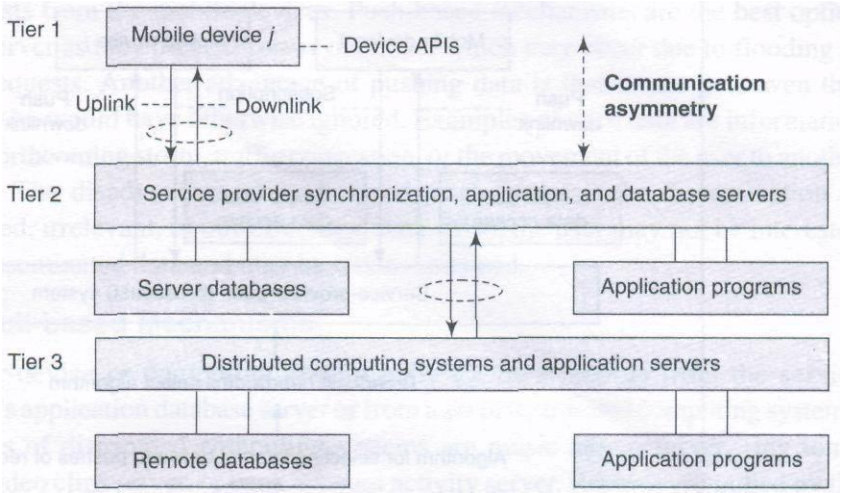
Efficient utilization of wireless bandwidth and battery power are two of the most important problems facing software designed for mobile computing. Broadcast channels are attractive in tackling these two problems in wireless data dissemination. Data disseminated through broadcast channels can be simultaneously accessed by an arbitrary number of mobile users, thus increasing the efficiency of bandwidth usage.

Communications Asymmetry

One key aspect of dissemination-based applications is their inherent communications asymmetry. That is, the communication capacity or data volume in the downstream direction (from servers-to-clients) is much greater than that in the upstream direction (from clients-to- servers). Content delivery is an asymmetric process regardless of whether it is performed over a symmetric channel such as the internet or over an asymmetric one, such as cable television (CATV) network. Techniques and system architectures that can efficiently support asymmetric applications will therefore be a requirement for future use.

Mobile communication between a mobile device and a static computer system is intrinsically asymmetric. A device is allocated a limited bandwidth. This is because a large number of devices access the network. Bandwidth in the downstream from the server to the device is much larger than the one in the upstream from the device to the server. This is because mobile devices have limited power resources and also due to the fact that faster data transmission rates for long intervals of time need greater power dissipation from the devices. InGSM networks data transmission rates go up to a maximum of 14.4 kbps for

both uplink and downlink. The communication is symmetric and this symmetry can be maintained because GSMis only used for voice communication.



Communication asymmetry in uplink and downlink and participation of device APIs and distributed computing systems when an application runs

The above figure shows communication asymmetry in uplink and downlink in a mobile network.The participation of device APIs and distributed computing systems in the running of an application is also shown.

Classification of Data-Delivery Mechanisms

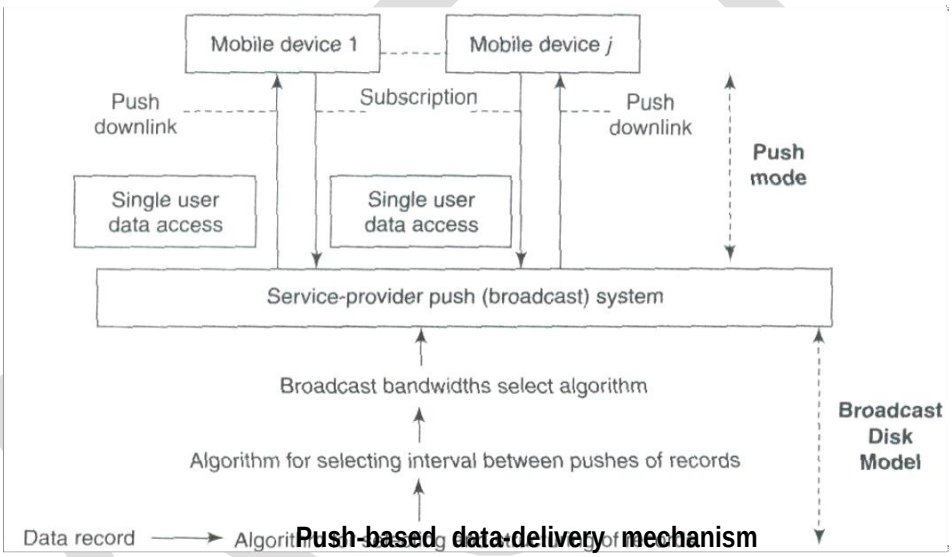
There are two fundamental information delivery methods for wireless data applications:Point-to-Point access and Broadcast. Compared with Point-to-Point access, broadcast is a more attractive method. A single broadcast of a data item can satisfy all the outstanding requests for that item simultaneously. As such, broadcast can scale up to an arbitrary number of users.

There are three kinds of **broadcast models**, namely *push-based* broadcast, *On-demand* (or *pull-based*) broadcast, and *hybrid* broadcast. In push based broadcast, the server disseminates information using a periodic/aperiodic broadcast program (generally without any intervention of clients). In on demand broadcast, the server disseminates information based on the outstanding requests submitted by clients; In hybrid broadcast, push based broadcast and on demand data deliveries are combined to complement each other. In addition, mobile computers consume less battery power on monitoring broadcast channels to receive data than accessing data through point-to-point communications.

Data-delivery mechanisms can be classified into three categories, namely, push-based mechanisms (publish-subscribe mode), pull-based mechanisms (on-demand mode), and hybrid mechanisms (hybrid mode).

Push-based Mechanisms

The server pushes data records from a set of distributed computing systems. Examples are advertisers or generators of traffic congestion, weather reports, stock quotes, and news reports.The following figure shows a push-based data-delivery mechanism in which a server or computing system pushes the data records from a set of distributed computing systems. The data records are pushed to mobile devices by broadcasting without any demand. The push mode is also known as ***publish-subscribe*** mode in which the data is pushed as per the subscription for a push service by a user. The subscribed query for a data record is taken as perpetual query till the user unsubscribe to that service. Data can also be pushed without user subscription.



Push-based data-delivery mechanism

Push-based mechanisms function in the following manner:

1. A structure of data records to be pushed is selected. An algorithm provides an adaptable multi-level mechanism that permits data items to be pushed uniformly or non-uniformly after structuring them according to their relative importance.
2. Data is pushed at selected time intervals using an adaptive algorithm. Pushing only once saves bandwidth. However, pushing at periodic intervals is important because it provides the devices that were disconnected at the timeof previous push with a chance to cache the data when it is pushed again.
3. Bandwidths are adapted for downlink (for pushes) using an algorithm. Usually higher bandwidth is allocated to records having higher number of subscribers or to those with higher access probabilities.

4. A mechanism is also adopted to stop pushes when a device is handed over to another cell.

The application-distribution system of the service provider uses these algorithms and adopts bandwidths as per the number of subscribers for the published data records. On the device handoff, the subscription cancels or may be passed on to new service provider system.

Advantages of Push based mechanisms:

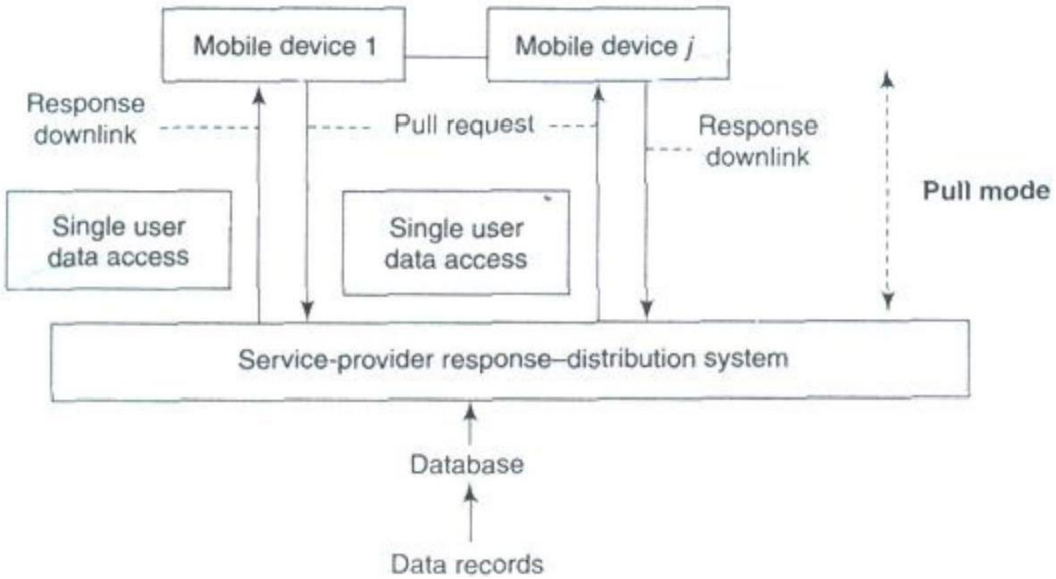
- Push-based mechanisms enable broadcast of data services to multiple devices.
- The server is not interrupted frequently by requests from mobile devices.
- These mechanisms also prevent server overload, which might be caused by flooding of device requests
- Also, the user even gets the data he would have otherwise ignored such as traffic congestion, forthcoming weather reports etc

Disadvantages:

- Push-based mechanisms disseminate of unsolicited, irrelevant, or out-of-context data,which may cause inconvenience to the user.

Pull based Mechanisms

The user-device or computing system pulls the data records from the service provider's application database server or from a set of distributed computing systems. Examples are musicalbum server, ring tones server, video clips server, or bank account activity server. Records are pulled by the mobile devices on demand followed by the selective response from the server. Selective response means that server transmits data packets as response selectively, for example, after client-authentication, verification, or subscription account check. The pull mode is also known as the on-demand mode. The following figure shows a pull-based data-delivery mechanism in which a device pulls (demands) from a server or computing system, the data records generated by a set of distributed computing systems.



Pull based Delivery Mechanism

Pull-based mechanisms function in the following manner:

1. The bandwidth used for the uplink channel depends upon the number of pull requests.
2. A pull threshold is selected. This threshold limits the number of pull requests in a given period of time. This controls the number of server interruptions.
3. A mechanism is adopted to prevent the device from pulling from a cell, which has handed over the concerned device to another cell. On device handoff, thesubscription is cancelled or passed on to the new service provider cell

In pull-based mechanisms the user-device receives data records sent by server on demand only.

Advantages of Pull based mechanisms:

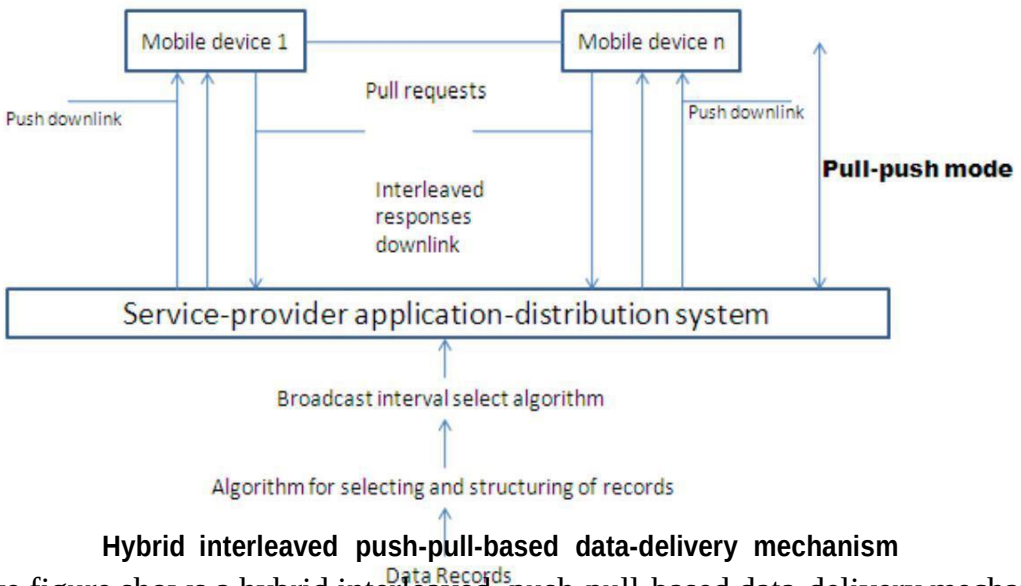
- With pull-based mechanisms, no unsolicited or irrelevant data arrives at the device and the relevant data is disseminated only when the user asks for it.
- Pull-based mechanisms are the best option when the server has very little contention and is able to respond to many device requests within expected time intervals.

Disadvantages:

- The server faces frequent interruptions and queues of requests at the server may cause congestion in cases of sudden rise in demand for certain data record.
- In on-demand mode, another disadvantage is the energy and bandwidth required forsending the requests for hot items and temporal records

Hybrid Mechanisms

A hybrid data-delivery mechanism integrates pushes and pulls. The hybrid mechanism is also known as interleaved-push-and-pull (IPP) mechanism. The devices use the back channel to send pull requests for records, which are not regularly pushed by the front channel. The front channel uses algorithms modeled as broadcast disks and sends the generated interleaved responses to the pull requests. The user device or computing system pulls as well receives the pushes of the data records from the service provider's application server or database server or from a set of distributedcomputing systems. Best example would be a system for advertising and selling music albums. The advertisements are pushed and the mobile devices pull for buying the album.



Hybrid interleaved push-pull-based data-delivery mechanism

The above figure shows a hybrid interleaved, push-pull-based data-delivery mechanism in which a device pulls (demands) from a server and the server interleaves the responses along with the pushes of the data records generated by a set of distributed computing systems. Hybrid mechanisms function in the following manner:

1. There are two channels, one for pushes by front channel and the other for pulls by back channel.
2. Bandwidth is shared and adapted between the two channels depending upon the number of active devices receiving data from the server and the number of devices requesting data pulls from the server.
3. An algorithm can adaptively chop the slowest level of the scheduled pushes successively The data records at lower level where the records are assigned lower priorities can have long push intervals in a broadcasting model.

Advantages of Hybrid mechanisms:

The number of server interruptions and queued requests are significantly reduced.

Disadvantages:

- IPP does not eliminate the typical server problems of too many interruptions and queued requests.
- Another disadvantage is that adaptive chopping of the slowest level of scheduled pushes.

Selective Tuning and Indexing Techniques

The purpose of pushing and adapting to a broadcast model is to push records of greater interest with greater frequency in order to reduce access time or average access latency. A mobile device does not have sufficient energy to continuously cache the broadcast records and hoard them in its memory. A device has to dissipate more power if it gets each pushed item andcaches it. Therefore, it should be activated for listening and caching only when it is going to receive the selected data records or buckets of interest. During remaining time intervals, that is, when the broadcast data buckets or records are not of its interest, it switches to idle or power down mode.

Selective tuning is a process by which client device selects only the required pushed buckets or records, tunes to them, and caches them. Tuning means getting ready for caching at those instants and intervals when a selected record of interest broadcasts. Broadcast data has astructure and overhead. Data broadcast from server, which is organized into buckets, is interleaved. The server prefixes a directory, hash parameter (from which the device finds the key), or index to the buckets. These prefixes form the basis of different methods of selective

tuning. Access time (**t_{access}**) is the time interval between pull request from device and reception of response from broadcasting or data pushing or responding server. Two important factors affect t_{access} – (i) number and size of the records to be broadcast and (ii) directory- or cache- miss factor (if there is a miss then the response from the server can be received only in subsequent broadcast cycle or subsequent repeat broadcast in the cycle).

Directory Method

One of the methods for selective tuning involves broadcasting a directory as overhead at the beginning of each broadcast cycle. If the interval between the start of the broadcast cycles is T, then directory is broadcast at each successive intervals of T. A directory can be provided which

specifies when a specific record or data item appears in data being broadcasted. For example, adirectory (at header of the cycle) consists of directory start sign, 10, 20, 52, directory end sign. It means that after the directory end sign, the 10th, 20th and 52nd buckets contain the data items in response to the device request. The device selectively tunes to these buckets from the broadcast data.

A device has to wait for directory consisting of start sign, pointers for locating buckets orrecords, and end sign. Then it has to wait for the required bucket or record before it can get

tuned to it and, start caching it. Tuning time t_{tune} is the time taken by the device for selection ofrecords. This includes the time lapse before the device starts receiving data from the server. In other words, it is the sum of three periods—time spent in listening to the directory signs and pointers for the record in order to select a bucket or record required by the device, waiting for the buckets of interest while actively listening (getting the incoming record wirelessly), and caching the broadcast data record or bucket.

The device selectively tunes to the broadcast data to download the records of interest. When a directory is broadcast along with the data records, it minimizes t_{tune} and t_{access}. The device saves energy by remaining active just for the periods of caching the directory and the data buckets. For rest of the period (between directory end sign and start of the required bucket), it remains idle or performs application tasks. Without the use of directory for tuning,

$t_{tune} = t_{access}$ and the device is not idle during any time interval.

Hash-Based Method

Hash is a result of operations on a pair of key and record. Advantage of broadcasting a hash is that it contains a fewer bits compared to key and record separately. The operations are done bya hashing function. From the server end the hash is broadcasted and from the device end a key is extracted by computations from the data in the record by operating the data with a function called hash function (algorithm). This key is called hash key.

Hash-based method entails that the hash for the hashing parameter (hash key) is broadcasted. Each device receives it and tunes to the record as per the extracted key. In this method, the records that are of interest to a device or those required by it are cached from the broadcast cycle by first extracting and identifying the hash key which provides the location of the record. This helps in tuning of the device. Hash-based method can be described as follows:

1. A separate directory is not broadcast as overhead with each broadcast cycle.
2. Each broadcast cycle has hash bits for the hash function H, a shift function S,

and the data that it holds. The function S specifies the location of the

record or remaining part of the record relative to the location of hash and, thus, the time interval for wait before the record can be tuned and cached.

3. Assume that a broadcast cycle pushes the hashing parameters $\mathbf{H(R_i)}$ [H and S] and record R_i. The functions H and S help in tuning to the H(R_i) and hence to R_i as follows—H gives a key which in turn gives the location of H(R_i) in the broadcast data. In case H generates a key that does not provide the location of H(R_i) by itself, then the device computes the location from S after the location of H(R_i). That location has the sequential records R_i and the devices tunes to the records from these locations.
4. In case the device misses the record in first cycle, it tunes and caches that in next or some other cycle.

Index-Based Method

Indexing is another method for selective tuning. Indexes temporarily map the location of the buckets. At each location, besides the bits for the bucket in record of interest data, an offset value may also be specified there. While an index maps to the absolute location from the beginning of a broadcast cycle, an offset index is a number which maps to the relative location after the end of present bucket of interest. Offset means a value to be used by the device along with the present location and calculate the wait period for tuning to the next bucket. Allbuckets have an offset to the beginning of the next indexed bucket or item.

Indexing is a technique in which each data bucket, record, or record block of interest is assigned an index at the previous data bucket, record, or record block of interest to enable the device to tune and cache the bucket after the wait as per the offset value. The server transmits this index at the beginning of a broadcast cycle as well as with each bucket corresponding to data of interest to the device. A disadvantage of using index is that it extends the broadcast cycle and hence increases t_{access}.

The index I has several offsets and the bucket type and flag information. A typical index may consist of the following:

1. $I_{offset}(1)$ which defines the offset to first bucket of nearest index.
2. Additional information about T_b, which is the time required for caching the bucket bits in full after the device tunes to and starts caching the bucket. This enables transmission of buckets of variable lengths.

3. I_{offset} (next) which is the index offset of next bucket record of interest.

4. $I_{offset}(end)$ which is the index offset for the end of broadcast cycle and the start of next cycle. This enables the device to look for next index I after the time interval as per $I_{offset}(end)$. This also permits a broadcast cycle to consist of variable number of buckets.
5. I_{type}, which provides the specification of the type of contents of next bucket to be tuned, that is, whether it has an index value or data.
6. A flag called dirty flag which contains the information whether the indexed buckets defined by $I_{offset}(1)$ and $I_{offset}(next)$ are dirty or not. An indexed bucket being dirty means that it has been rewritten at the server with new values. Therefore, the device should invalidate the previous caches of these buckets and update them by tuning to and caching them.

The advantage of having an index is that a device just reads it and selectively tunes to the data buckets or records of interest instead of reading all the data records and then discarding those which are not required by it. During the time intervals in which data which is not of interest is being broadcast, the device remains in idle or power down mode.

Transmission of an index I only once with every broadcast cycle increases access latency of a record as follows: This is so because if an index is lost during a push due to transmission loss, then the device must wait for the next push of the same index-record pair. The data tuningtime now increases by an interval equal to the time required for one broadcast cycle. An index assignment strategy ***(I, m)*** is now described. *(I, m)* indexing means an index I is transmitted *m* times during each push of a record. An algorithm is used to adapt a value of **m** such that it minimizes access (caching) latency in a given wireless environment which may involve frequent or less frequent loss of index or data. Index format is adapted to (I, *m)* with a suitable value of m chosen as per the wireless environment. This decreases the probability of missing I and hence the caching of the record of interest

Indexing reduces the time taken for tuning by the client devices and thus conserves their powerresources. Indexing increases access latency because the number of items pushed is more (equals m times index plus *n* records).

Distributed Index Based Method

Distributed index-based method is an improvement on the (I, *m)* method. In this method, there is no need to repeat the complete index again and again. Instead of replicating the whole index *m* times, each index segment in a bucket describes only the offset I' of data items which immediately follow. Each index I is partitioned into two parts—I' and I''. I'' consists of

unrepeated *k* levels (sub-indexes), which do not repeat and I' consists of top I repeated levels (sub-indexes).

Assume that a device misses I(includes I' and I' once) transmitted at the beginning of the broadcast cycle. As I' is repeated *m* - I times after this, it tunes to the pushes by using I', The access latency is reduced as I' has lesser levels.

Flexible Indexing Method

Assume that a broadcast cycle has number of data segments with each of the segments having a variable set of records. For example, let *n* records, R_o to R_{n-1}, be present in four data segments, $R_{()}$ to R_{i-1}, R_i to R_{j-1} , R_j to R_{j-1} and R_k to R_{n-1}. Some possible index parameters are (i) I_{seg}, having just 2 bits for the offset, to specify the location of a segment in a broadcast cycle, (ii) I_{rec}, having just 6 bits for the offset, to specify the location of a record of interest within a segment of the broadcast cycle, (iii) I_b, having just 4 bits for the offset, to specify the location ofa bucket of interest within a record present in one of the segments of the broadcast cycle.

Flexible indexing method provides dual use of the parameters (e.g., use of I_{seg} or I_{rec} in an indexsegment to tune to the record or buckets of interest) or multi-parameter indexing (e.g., use of

$I_{seg,}$ I_{rec}, or I_b in an index segment to tune to the bucket of interest).

Assume that broadcast cycle has *m* sets of records (called segments). A set of binary bitsdefines the index parameter $I_{seg,}$. A local index is then assigned to the specific record (or

bucket). Only local index (I_{rec} or I_b) is used in (I_{loc}, m) based data tuning which corresponds to the case of flexible indexing method being discussed. The number of bits in a local index ismuch smaller than that required when each record is assigned an index. Therefore, the flexible indexing method proves to be beneficial.

Alternative Methods

Temporal Addressing Temporal addressing is a technique used for pushing in which instead of repeating I several times, a temporal value is repeated before a data record istransmitted. When temporal information contained in this value is used instead of address,there can be effective synchronization of tuning and caching of the record of interest in case ofnon-uniform time intervals between the successive bits. The device remains idle and startstuning by synchronizing as per the temporal (time)-information for the pushed record.Temporal information gives the time at which cache is scheduled. Assume that temporaladdress is 25675 and each address corresponds to wait of 1 ms, the device waits and startssynchronizing the record after 25675 ms.

Broadcast Addressing: Broadcast addressing uses a broadcast address similar to IP or multicast address. Each device or group of devices can be assigned an address. The devices cache the records which have this address as the broadcasting address in a broadcast cycle. Thisaddress can be used along with the pushed record. A device uses broadcast address in place of the index I to select the data records or sets. Only the addressed device(s) caches the pushed record and other devices do not select and tune to the record. In place of repeating I several times, the broadcast address can be repeated before a data record is transmitted. The advantage of using this type of addressing is that the server addresses to specific device or specific group of devices.

Use of Headers: A server can broadcast a data in multiple versions or ways. An index or address only specifies where the data is located for the purpose of tuning. It does not specify the details of data at the buckets. An alternative is to place a header or a header with an extension with a data object before broadcasting. Header is used along with the pushed record.The device uses header part in place of the index / and in case device finds from the header that the record is of interest, it selects the object and caches it. The header can be useful, for example it can give information about the type, version, and content modification data or application for which it is targeted.

Notes for Indexing Techniques (Prepared by Kancherla Yasesvi, 08071A0522)

(1, m) Index

The (1, m) indexing scheme is an index allocation method where a complete index is broadcast m times during a broadcast. All buckets have an offset to the beginning of the next index segment. The first bucket of each index segment has a tuple containing two fields. The first field contains the key value of the object that was broadcast last and the second field is an offset pointing to the beginning of the next broadcast. This tuple guides required object in the current broadcast so that they can tune to the next broadcast.

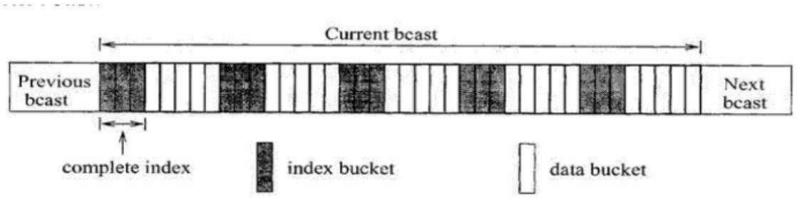


Figure 4.3. Bcast organization in the (1 ,*m*) indexing method.

Data Dissemination

The client's access protocol for retrieving objects with key value k is as follows:

1. Tune into the current bucket on the broadcast channel. Get the offset to the next index segment.
2. Go to the doze mode and tune in at the broadcast of the next index segment.
3. Examine the tuple in the first bucket of the index segment. If the target object has been missed, obtain the offset to the beginning of the next bcast and goto 2; otherwise goto 4.
4. Traverse the index and determine the offset to the target data bucket. This may be accomplished by successive probes, by following the pointers in the multi-level index. The client may doze off between two probes.
5. Tune in when the desired bucket is broadcast, and download it(and subsequent ones as long as their key is k).

Advantage:

1. This scheme has good tuning time.

Disadvantage:

1. The index is entirely replicated m times; this increases the length of the broadcast cycleand hence the average access time.

The optimal m value that gives minimal average access time is $(\text{data file size}/\text{index size})^{1/2}$.

There is actually no need to replicate the complete index between successive data blocks. It is sufficient to make available only the portion of index related to the data buckets which follow it. This is the approach adopted in all the subsequent indexing schemes.

Tree-based Index/Distributed indexing scheme

In this scheme a data file is associated with a B^+-tree index structure. Since the broadcast medium is a sequential medium, the data file and index must be flattened so that thedata are index are broadcast following a preorder traversal of the tree. The index comprises
two portions: the first k levels of the index will be partially replicated in the broadcast, and the remaining levels will not be replicated. The index nodes at the $(k+1)^{th}$ level are called the non-replicated roots.

Essentially, each index subtree whose root is a non-replicated root will appear once in the whole bcast just in front of the set of data segments it indexes. On the other hand, the nodes at the replicated levels are replicated at the beginning of the first broadcast of each of itschildren nodes.

To facilitate selective tuning, each node contains meta-data that help in the traversal of the trees. All non-replicated buckets contain pointers that will direct the search to the nextcopy of its replicated ancestors. On the other hand, all replicated index buckets contain two tuples that can direct the search to continue in the appropriate segments. The first tuple is a

pair(x, ptr_{begin}) that indicates that key values less than x have been missed and so search must continue from the beginning of the next bcast(which is ptr_{begin} buckets away). The second pair (y, ptr) indicates that key values greater than or equal to y can be found ptr offset away. Clearly,if the desired object has key value between x and y, the search can continue as in conventional search operation.

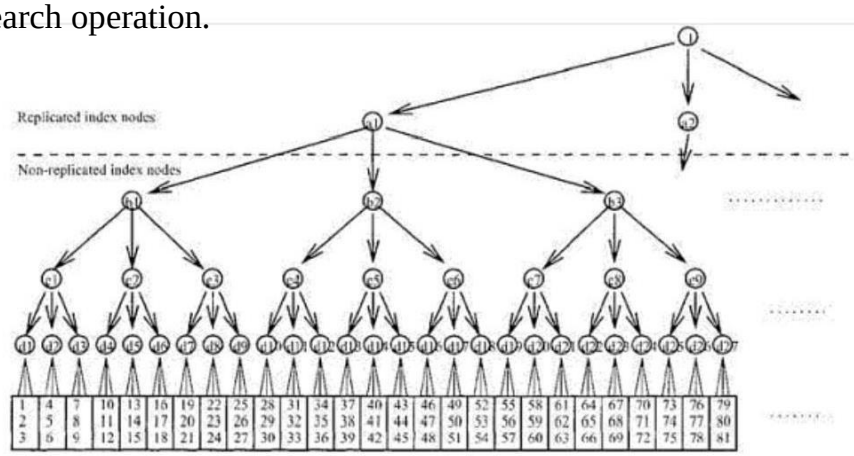


Figure 4.4. A partial data file and its index tree.

The client's access protocol for retrieving objects with key value k is as follows:

1. Tune to the current bucket of the bcast. Get the offset to the next index bucket, and doze off.
2. Tune to the beginning of the designated bucket and examine the meta-data.
   - If the desired object has been missed, doze off till the beginning of the nextbcast. Goto 2.
   - If the desired object is not within the data segment covered by the index bucket,doze off to the next higher level index bucket. Goto 3.

     If the desired object is within the data segment covered by the index bucket,goto 3.

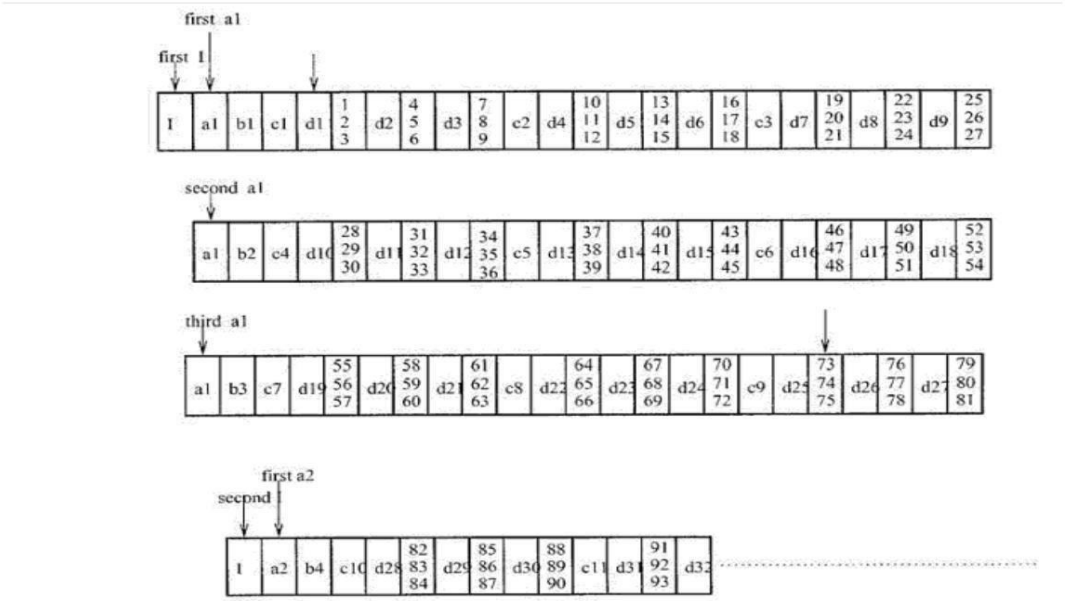


Figure 4.5. The data broadcast for the distributed indexing scheme with partial path replication.

3. Probe the designated index bucket and follow a sequence of pointers to determine when the data bucket containing the target object will be broadcast. The client may doze off in between two probes.
4. Tune in again when the bucket containing objects with key k is broadcast, and download the bucket (and all subsequent buckets as long as they contain objects with key k).

Advantage:

1. Compared to (1, m) index scheme this scheme has lower access time and its tuning timeis also comparable to that of (1, m) index scheme.

Flexible Indexing Scheme

This scheme splits a sorted list of objects into equal-sized segments, and provides indexes to navigate through the segments. At the beginning of each segment, there is a control index which comprises of two components: a global index and a local index. The global index is used to determine the segment which object may be found, while the local index provides the offset to the portion within the segment where the object may be found.

Suppose the file is organized into p segments. Then the global index at a segment, says, has $[\log_2 i]$ (key, ptr) pairs, where i the number of segments in front of and including segment s, key is an object key, and ptr is an offset. For the first entry, key is the key value of the first data item in segment s and ptr is the offset to the beginning of the next version. Bold examining this pair, the client will know if it has missed the data and if so wait till the next bcast. For the j^{th} entry (j>1), key is the key value of the first data item in the $([\log_2 i/2^{j-1}] +1)^{th}$ segment following segment s and ptr is the offset to the first data bucket of that segment.

The local index consists of m(key, ptr) pairs that essentially partition each segment further into m+1 sections. For the first entry, key is the key value of the first data item of section m+1 and ptr is the offset to that section. For the jth pair, key is the key value of the first data item of section (m+1-j) and ptr is the offset to the first bucket of that section.

Hence, it is clear that the number of segments and the number of sections per segment can affect the performance of the scheme. Increasing the number of segments or sections will increase the length of the broadcast cycle and reduce the tuning time, and vice versa. Thus, the scheme is flexible in the sense it can be tuned to fit an application's needs.

The client's access protocol for retrieving objects with key value k is as follows:

1. Tune into the channel for a bucket, obtain the offset to the next index segment. Doze off until the next index segment is broadcast.
2. Examine the global index entries. If the target object belongs to another segment, get the offset; doze off for appropriate amount of time and goto 2.
3. Examine the local index entries. Obtain the offset to the section where the target data is stored. Switch to doze mode for appropriate amount of time.
4. Examine objects in the data bucket for the desired object, and download the object.

Introduction to SYNCHRONIZATION

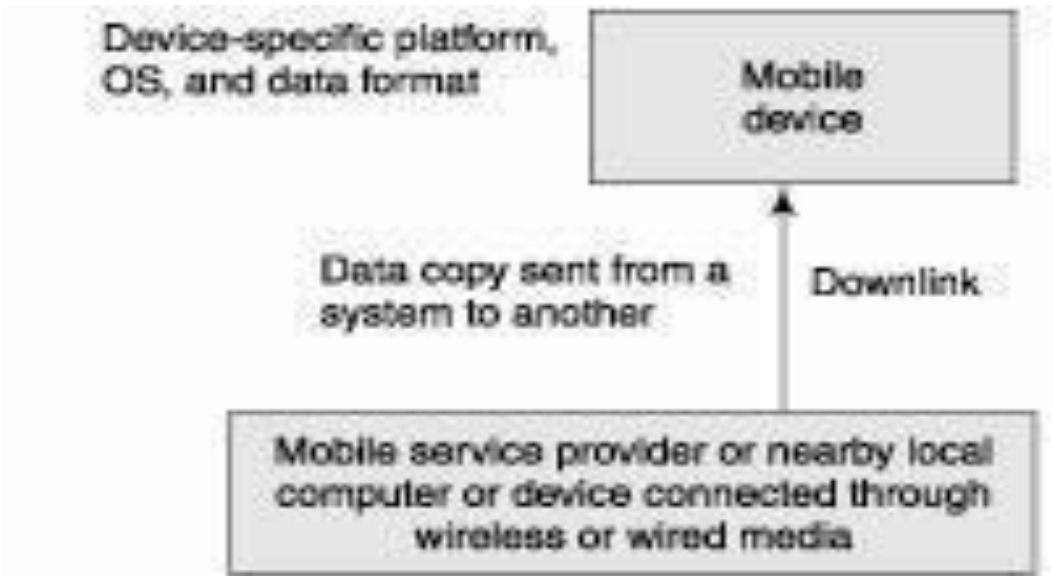
Syllabus: Introduction to Synchronization, What is Synchronization in Mobile Computing Systems, Usage models for synchronization n mobile application, Domain-dependent specific Rules for data synchronization, Personal information manager, Synchronization & conflict resolution strategies, About Synchronizer Mobile Agent, Mobile Agent Design, Application server.

Data Replication and Synchronization in Mobile Computing Systems

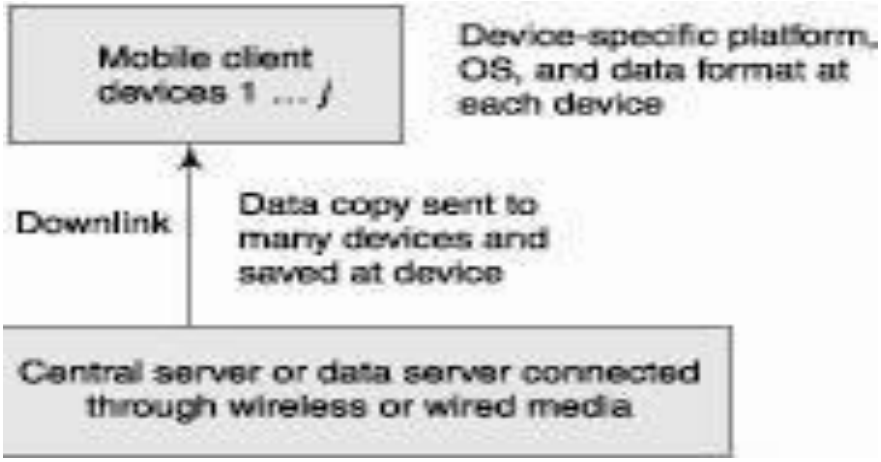
Data dissemination and replication

- Data replication at either remote or local location (s) may entail copying of data at one place after copying from another (i.e., recopying), copying from one to many others or from many to many others
- For example, videos of faculty lectures or music files get replicated at a mobile phone

Data replication from data source and device



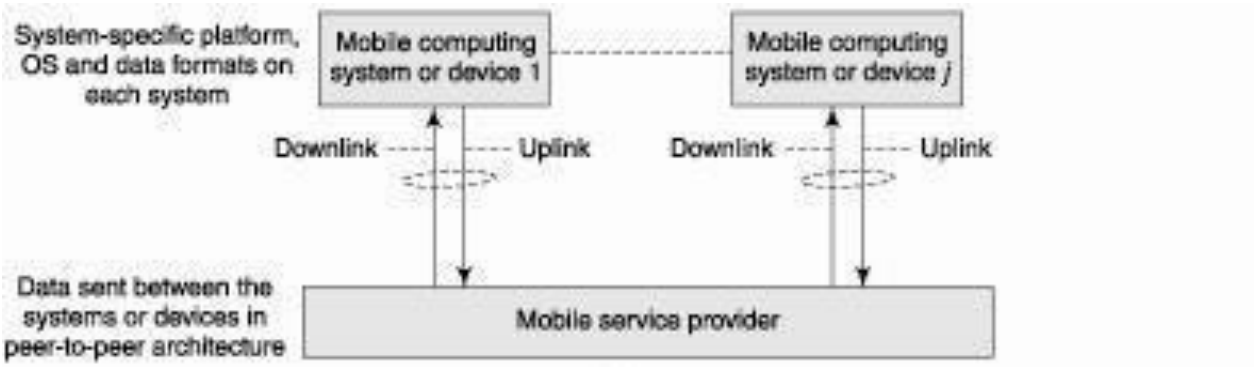
Data replication from data source server to many clients (devices)



One to many synchronization

- Each system or device caches the data pushed from the server or sends a pull request to the central server and gets a response

Data replication among systems and devices in peer to peer architecture



Many-to-many synchronization

- Employs peer-to-peer architecture where each system is capable of sending pull requests and of pushing responses

Full copy from a source

- Means that the full set of data records replicates according to certain domain-specific data format rules at the replicating devices or systems
- A server having a set of 8 images with resolution 640 × 640 pixels
- In the domain of a mobile device, it can replicate and hoard with 160 × 160 pixels
- When all 8 images copied, though with the different resolution, then it is known as full copy replication

Full copy from a source Example

- A server having a set of 8 images with resolution 640 × 640 pixels
- In the domain of a mobile device, it can replicate and hoard with 160 × 160 pixels

- Full copy replication─ when all 8 images copied, though with the different resolution

Partial copying of data from the source

- A subset of the data set copied according to certain domain-specific rules at the devices or systems
- Assume that a server has a hourly data set of 24 temperature records with ± 0.1°C
- Partial copy replication In mobile device domain, assume that it replicates and hoards three hourly records with ± 1°C

Data Synchronization

- Data replication precedes data synchronization
- The synchronization refers to maintaining data consistency among the disseminated or distributed data
- Data consistency─ if there is data modification at the server then that should reflect in the data with the device within a defined period

Data Synchronization in Mobile computing systems

- Defined as the process of maintaining the availability of data generated from the source and maintaining consistency between the copies pushed from the data source and local cached or hoarded data at different computing systems without discrepancies or conflicts among the distributed data.

Consistent copy of data

- A copy which may not be identical to the present data record at the data-generating source, but must satisfy all the required functions and domain-dependent specific rules

Domain Specific rules for consistency

- In terms of resolution, precision, data format, and time interval permitted for replication
- A consistent copy should not be in conflict with the data at the data-generating source

Data synchronization for accessing data from server

- Helps mobile users in accessing data and using it for computing on mobile devices
- When a device not connected to a source or server, the user may employ data that is not in conflict with the present state of data at the source

Data synchronization for caching data in and from personal area computer

- Helps mobile users in hoarding the device data at the personal area computer
- Also helps mobile users in hoarding the personal area computer data

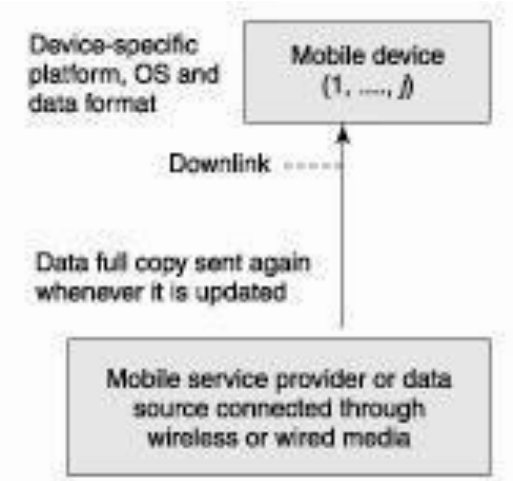
Data synchronization enhancing device mobility

- When initiated at frequent intervals enhances device mobility
- Ensures that device applications use the latest updated data from the source, even when the device is disconnected

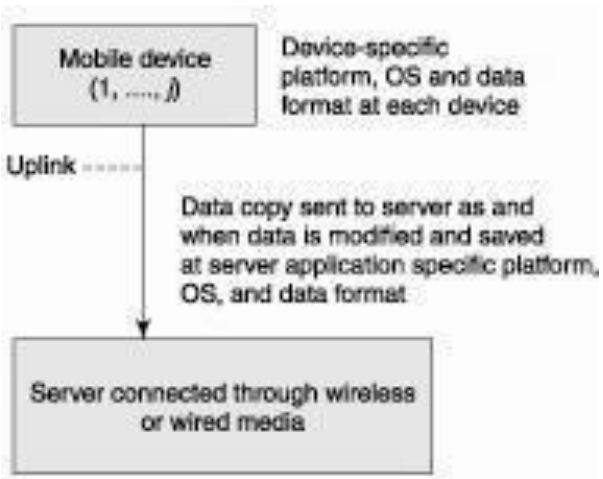
Data synchronization with enterprise server

- Helps storage at enterprise server a large chunks of information for the many devices connected to it and update partial copies of data at frequent intervals

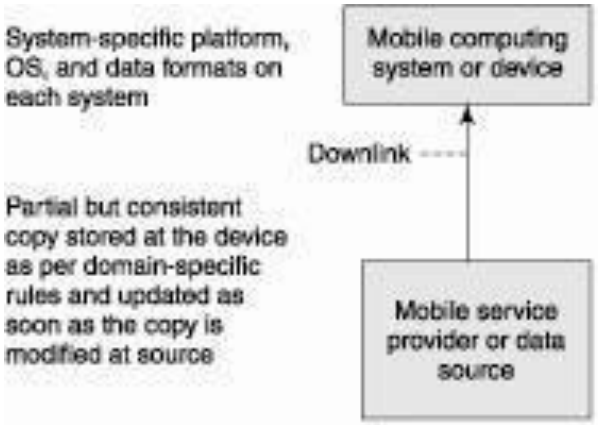
Full copy Synchronization at the device when the server sends data



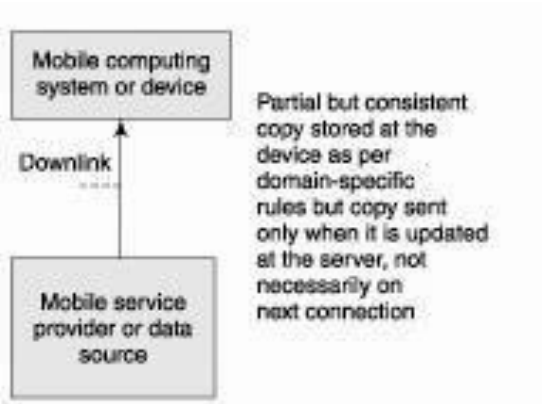
Full copy Synchronization at the server when the device sends data



Partial copy Synchronization of consistent copy without the delays



Partial copy Synchronization of consistent copy but after delays



Data Synchronization Types, Formats and Usage Models in Mobile Computing Systems

Data synchronization Needs

- Required between the mobile device and service provider
- Between the device and personal area computer
- With nearby wireless access point (in WiFi connection)
- Another nearby device

Two-way synchronization of partial or full copies of data

- **Between mobile-device and personal-area computer**
  - **For example, whenever the list of contacts and personal information manager data is modified at any of them, it is made consistent after synchronization**

2. Server-alerted synchronization

- **`The server alerts the client the data modification or additions**
- **The client synchronizes the modified or new data by pull request**
- **For example, alerting new e-mail and the device pulls that**

3. One-way server-initiated synchronization

- **Server initiates synchronization of any new modification since communication of last modification**
- **Sends modified data copies to the client**
- **When a new email arrives at a server, it initiates the synchronization as and when the device connects to the server and pushes the mail**

4. Client initiated refresh synchronization

- The client initiates synchronization with the server for refreshing its existing data copies
- For refreshing the configuration parameters saved at the server for it
- For example, a computer or mobile device initiates refreshing of the hoarded contacts and personal information data either at periodic intervals or as and when it connects
- if the device configuration changes or a new device connects to a server, then the configuration parameters sent earlier refresh at the server

5. Client-initiated synchronization

- With the server for sending its modifications, for example, device configuration for the services
- For example, a client mobile device initiates synchronization of the mails or new ring tones or music files available at the server either at periodic intervals or as and when it connects to it

6. Refresh from client for backup and update synchronization

- The client initiates synchronization
- Sends backup to the server for updating its data
- For example, a computer or mobile device initiates refreshing of the hoarded contacts and personal information data either at periodic intervals or as and when it connects to the server

7. Slow (full data copy and thorough) synchronization

- Client and server data compared for each data field and are synchronized as per conflict resolution rules
- Full copy synchronization usually takes place in idle state of the device
- Not immediately on connecting to the server, that's why called slow

Formats of Synchronized Data Copies

- Can be different from each other at client and server
- When the data at a source synchronizes with the data at other end, it does so as per the format specified at that end

Formats of Database records

- The records indexed enabling search by querying using the indexes, for example, the relational database records
- The database record retrieved by sending a query specifying the entries in these indexes
- Format DB2 at server and DB2e Every place at the mobile device

Flat file Synchronization

- Data can be interpreted only if the file is read from beginning to end and that data cannot be picked from anywhere within the file
- For example, an XML or html file at the server synchronizes with the file at the device which is in text format or is a binary file depending upon the information format
- Information format in mobile computing XML document format
- For transmission it is WBXML (WAP Binary XML) content format
- Address book data at a mobile device with the data transmitted in WBXML format

Device-specific storage Format

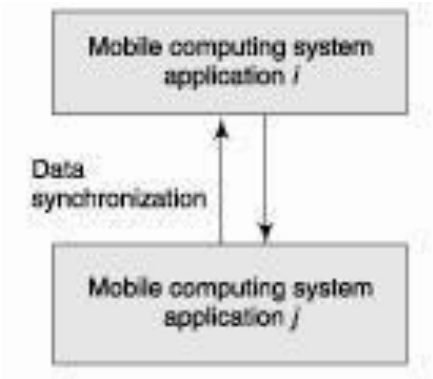
- AAC (Apple Audio Communication) files used for audio communication with an Apple iPhone
- A file in AAC format synchronizes with music files in some other format at a computer or remote website serving the music files
- At a mobile device the *Contacts* information in vCard format
- Calendar, tasks-to-do list, and journal information are in vCalendar, vToDo, and vJournal formats, respectively

Usage Models for Synchronization in Mobile Applications

- Four usage models employed for synchronization in mobile computing systems

1.Synchronization between two APIs within a mobile computing system

Synchronization between two APIs



Synchronization between two APIs

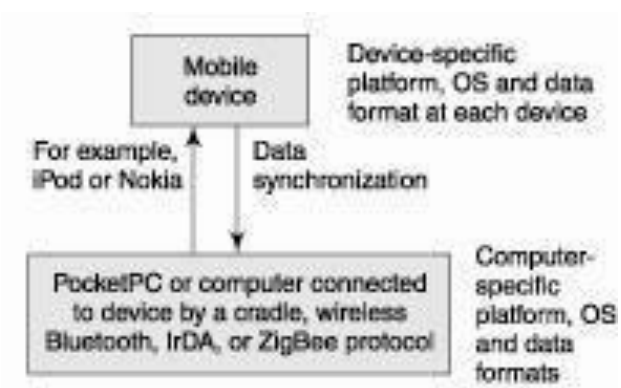
The data generated by an application synchronized and used in another application

- An API running at the device synchronizes data with another application on the same or another device or computer

Example of synchronization between APIs

- Data records at personal information manager (PIM) API synchronized with the email API
- When email from a new source retrieves at the email API in the device, the name and email address data fields at the application saved as new data record at PIM API
- When an email is to be sent to the same person, the email API uses the same data record from the PIM API

2.Synchronization between the device and nearby device

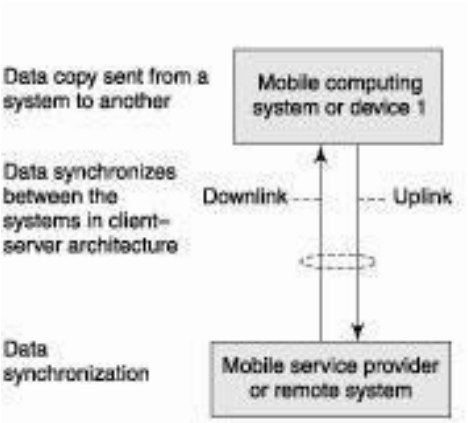


- Device and computer synchronize their data
- Also called personal area synchronization (PAS)
- Using PAS software, for example, HotSync or ActiveSync

Examples of PAS

- Synchronization with nearby PC through a serial port using a cradle and wired connection to PC through the cradle
- Synchronization with the nearby computer through a wireless personal area network (WPAN) using ZigBee or Bluetooth

3. Synchronization between remote systems and device



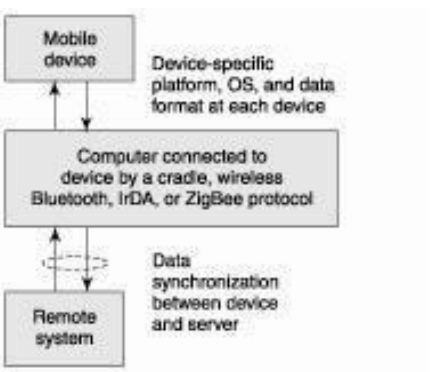
Synchronization between remote systems and device

- The device data records synchronize with the mobile service provider server records
- The remote server or systems synchronize their data with the mobile device
- The device connects to remote systems on Internet through the wired, wireless mobile service provider, or WiFi network

Example of Synchronization between remote systems and device

- Wireless email synchronization using Intellisync between the device and remote server using SyncML language

4. Synchronization through a local pass-through system



5.

Using local pass through computer or system

- Device data records synchronize with the records of remote system, for example, an enterprise server, through a local computer system

Example of Using local pass through computer or system

- The device first synchronizes through ActiveSync or HotSync or Intellisync or Bluetooth to local computer connected by personal area synchronizer
- Then the computer synchronizes to Internet through WLAN, WiFi, or wired network

Domain dependent Specific Rules and Conflict resolution Strategies

1. **Data synchronization in domain-specific platforms and data formats**

Data synchronization between data-generating domain and destined domain, both having different platform and data formats

Examples of synchronization in domain-specific platforms and data formats

- **A copy of database record at the device structured text or XML format and the device OS platform Symbian**
  - **The record synchronized with the database record at the server where it is in DB2 or Oracle database format and the OS is Windows**

2. **Domain-specific data-property-dependent synchronization**
   - **Data synchronization between one domain with one property of data and another domain having different property**

Examples of Domain-specific data-property-dependent synchronization

- **A data record at a device having an ID specified by a byte synchronizes with the record, which hasan ID specified by 16-bit word at the server**
- **A device using 8-bit ASCII characters for an ID while the server using 16-bit Unicode characters**

3. **Synchronization up to the last successful act of synchronization**
   - **A domain-specific rule that data record considered to be synchronized if it was updated at the last connection**

Example of Synchronization up to the last successful act

- A phonebook records of missed calls, dialled numbers, and received calls
- Data record at the device synchronized with the record in the phonebook
- If it updated at the last connection, then it eventually updates again on the next connection

4. **Memory-infrastructure-dependent based synchronization at the domains**

- A domain-specific rule that data records synchronized up to the allotted memory

Example of Memory-infrastructure-dependent based synchronization

- A remote server maintaining full address book with allotted memory of 8 MB and a device allocated 128 kB for the address book
- Only a part of e-mail database, only 100 new email addresses synchronizes and saves in the device PIM (personal information manager)

5. **Synchronization with temporal properties of data**

Domain-specific rule that data records synchronized with data generated at source within specific timeinterval and at time specified at the domain

Example of Synchronization with temporal properties of data

- The flight time table data set of device synchronized every week and weather report once every day
- At the device weather report updated and synchronized up to the last day
- Eventually updates on a day if available at the server

Synchronization with temporal properties of data

- May be periods of inconsistency when temporal properties of data being used for synchronization
- However, mobile applications remain unaffected if there are no temporal conflicts and unaccountable discrepancies

Conflict in synchronization

- Arises when a data copy changed at one end but not simultaneously modified at other ends
- Therefore, the same data item at two ends, *P* and *Q*, in conflict during computation in the time interval between *t*1 and *t*2, where *t*1 and *t*2─ the instants when *P* and *Q* get the modified data copy

Synchronization and Conflict Resolution Strategies

- A conflict resolution strategy adopted in such cases to resolve conflicts

- The strategy specifies the rules that need to be applied for conflict resolution

1. **Priority-based resolution rule**

Data-server can be specified as dominant higher priority entity for conflict resolution of synchronized datarecords

Example of Priority-based resolution rule

- Mobile-service-provider server S having a list of missed, dialled, and received calls for the device D
- D has a synchronised list of missed, dialled, and received calls
- When the list at D in conflict with the list at S, priority-based resolution rule specifies that the server priority is higher

2. **Time-based resolution rule**

- Data node P specified as dominant entity when P always receives copies first from the server S

Time-based resolution rule – Example

- S having the emails disseminated to the device D at an instant $t1$
- D connects to a personal area computer (PC) to which the device always synchronizes the mails at a later instant $t2$
- Time-based resolution rule— D dominant because it receives the mails earlier than the PC

3. **Information-based resolution rule**

a. Data node can be specified as dominant entity when information specific to it is synchronized with other nodes

Example of Information-based resolution rule

- Server S having the device configuration record disseminated from the device D
- Information-based resolution rule specifies that since the information is for the device D hence D is dominant node
- For device-specific information, the device data accepted rather than the server data

4. **Time-stamp-based resolution rule Device-specific storage Format**

- Time-stamp-based resolution rule necessitates that a time-stamp must be used while sending a data copy The copy found to be latest resolves the conflict

- **Example of Time stamping rule for conflict resolution**

- Server S having the flight information which it always disseminates at regular intervals with a time stampover it to the device D and as well as to a PC
- Time-stamp-based resolution rule specifies that the node with flight information with latest time stampdominant

5. **User-interaction-based resolution rule**

- **An API at a device allows a user to interact with the device**
- **Interaction resolves the conflict arising out of the duplicate or multiple entries**
- **The duplicate data entries permitted at the node when a receiver API later on resolves the conflictafter interaction with user**

Example of User-interaction-based resolution rule

- Two phone number entries found for same name and address, the device prompts user to resolve theconflict
- User resolves the conflict by opting for one of it