Notes of Mobile Computing

(ECS-087)

By

Associat Prof. Abhishek Pandey

Unit 1

Introduction :-

What is wireless communication?

Wireless communication is basically transmitting and receiving voice and data using electromagnetic waves in open space. They are basically free from wireless

Types of Wireless Communication

1) Mobile
   -Cellular Phones (GSM/cdma)
2) Portable
   - IEEE802.11b(WiFi)
   - IEEE 802.15.3
3) Fixed
   - IEEE802.16 (Wireless MAN)

Why Wireless Communication

1) Freedom from wires
   a. No Cost of installing wires or rewiring
   b. No bunches of wires running here and there " Auto magical" instantaneous communication without physical connection setup.
2) Global Coverage
   a. Communications can reach where wiring is infeasible or costly, e.g. rural areas, battlefield, vehicles, outer space (through communication satellites)
3) Stay Connected
   a. Rooming allows flexibility to stay connected anywhere and anytime.
   b. Rapidly growing market attracts to public need for mobility and uninterrupted access
4) Flexibility
   a. Services reach you wherever you go
   b. Connect to multiple devices simultaneously (no physical communication required
5) Increasing dependence on telecommunication services for business and personal reasons.
6) Consumers and businesses are willing to pay for it

Challenges

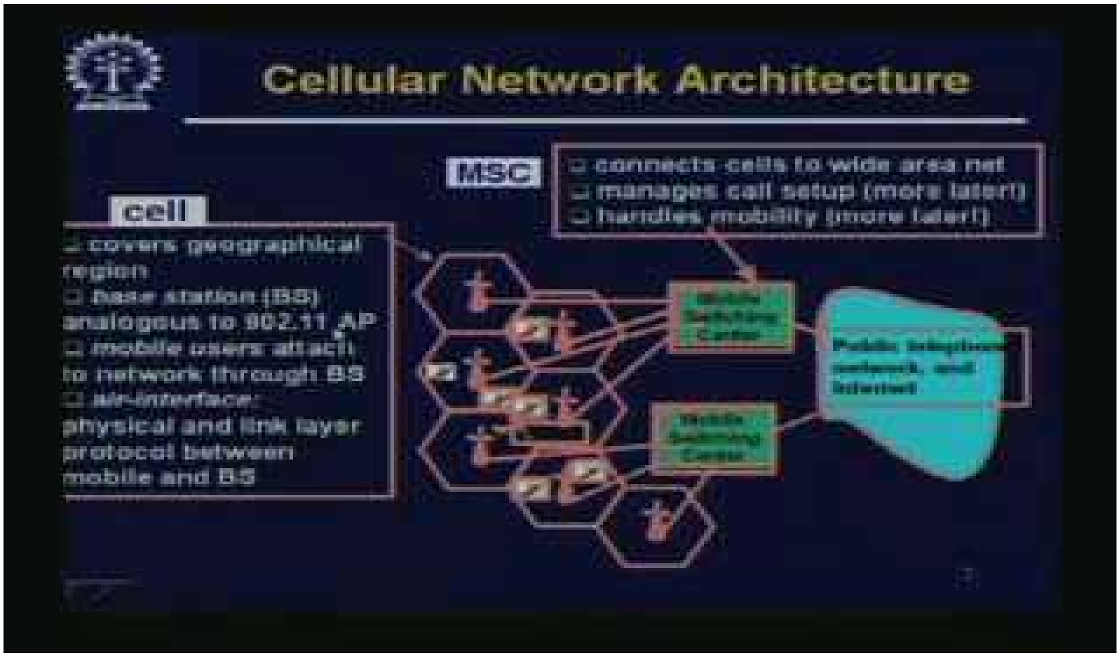
1) Efficient hardware
   a. Low power transmitters Receivers
   b. Low power signal processing tools

2) Efficient use of finite radio spectrum
    a. Cellular frequency reuse
    b. Medium access control protocols
3) Integrated services
    a. Voice, data, multimedia over a single network
    b. Service differentiation, priorities, resource sharing
4) Network support for user mobility (location identification, handover)
5) Maintaining quality of service over unreliable links
6) Connectivity and coverage (internetworking)
7) Cost efficiency
8) Fading
9) Multipath
10) Higher probability of data corruption
11) Need for strong security mechanisms

Current Wireless Systems

1) Cellular System
2) Wireless LANs
3) Satellite Systems
4) PANs (Bluetooth)

Cellular System



Cellular Network Architecture

MSC

- ❑ connects cells to wide area net
- ❑ manages call setup (more later!)
- ❑ handles mobility (more later!)

cell

- ❑ covers geographical region
- ❑ base station (BS) analogous to 802.11 AP
- ❑ mobile users attach to network through BS
- ❑ air-interface: physical and link layer protocol between mobile and BS

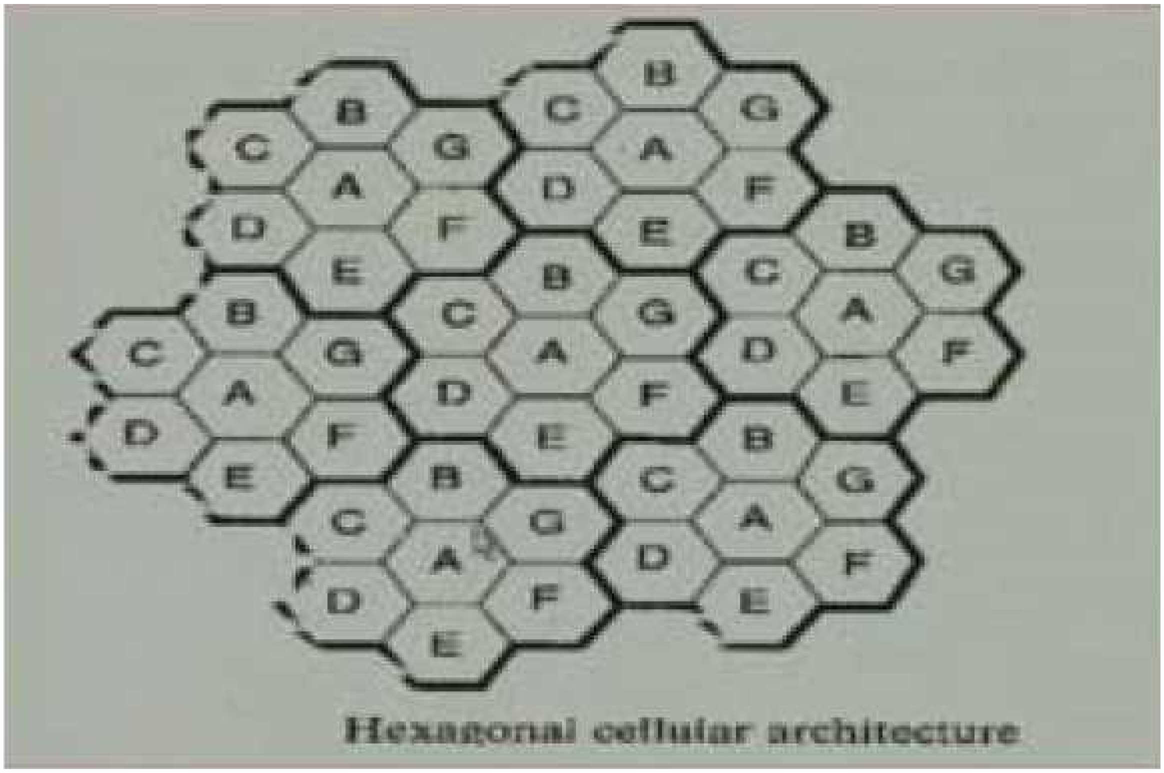
What is a cell?

The cellular network is organized in the form of some cells and it covers a geographical region. It has base station analogous to 802.11 AP. AP is for access point. 802.11 is the wireless LAN technology. The point is that there is a base station and it will have an antenna, some transmitters and receivers. It will be connected to the backbone through a line. May be this could also be a wireless line, but usually this could be a fiber optic line. In a circluar geographical location around this base station, mobiles will communicate with this base station and through this base station to the rest of the network. So mobile users attach to network through BS.

Now all these base stations are connected to the mobile switching center (MSC). The switching is essentially done over here. The MSC connects cells to Wide Area Network. The mobile switching center will be referred to as MSC, Base station as BS, mobile switching center as MSC, MS, by the way, is a mobile station. MSC connects cells to Wide Area Network and manages call setup. More about that later and these MSCs will be connected to each other for a particular service provider. They will also connect to public telephone network and the internet etc. So one service provider, their network would be connected to another service provider's network.

Cell Fundamentals

In practice, cells are may be of arbitrary shape. But they will be close to a circle because usually the kind of antenna used in base stations is omni directional antenna, in the sense that it gives the same power on all sides. It has the same sensitivity on all sides. If that is so, the area of influence would be a circle. But when many circles are put together they are pulling and they will intersect with each other. To solve this problem, we can use a tessellation. There are only three types of tessellations, which are possible – equilateral triangles, squares or regular hexagons. Out of these three, the regular hexagon is the closest to a circle. That is why usually the regular hexagons are used to represent a cellular structure.



Hexagonal cellular architecture

If you notice carefully some of the cells are dark and these cells are marked as A B C D E F G. So, these are seven. There are seven hexagons like this and these are actually different frequency ranges. These frequencies are again reused. For example, you have another A B C D E F G over here. This B and this B – although they use the same frequency ranges – are far apart. So, different groups of people can use it at the same. This really shows you the frequency reuse. A, the set of frequency bands, which are associated with A, will also be reused here, here, and there and so on. That is how a hexagonal cellular structure is constructed and we do this frequency reuse.

Co-channel reuse ratio is given by DL/RL is equal to √3N, where DL is the distance between co-channel cells, that means those who share the same channel. RL is cell radius; N is the cluster size. The number of cells in a cluster N determines the amount of co-channel interference and the number of frequency channels available per cell. This really comes from geometry.

Frequency reuse has its foundations in the attenuation of the signal strength of EM waves with distance. So, if two points are at a distance from each other, this signal gets attenuated and does not interfere significantly with this one, although they are using the same frequency band. Frequency reuse can be done by two waya

a) Fixed Channel Allocation

   means for a particular cell, the channels that means, the frequency band associated with the cell is fixed. So, total number of channels is NC is equal to W/B, where W is the bandwidth of the available for spectrum. B is the bandwidth needed by each channel. The total number of channels per cell is Cc is equal to Nc/N, where N is the cluster size.

Adjacent radio frequency bands are assigned to different cells. In analog each channel corresponds to one user while in digital each RF channel carries several time slots or codes. So, you are doing either TDMA or CDMA. So, if you are doing this, the naturally FDMA TDMA combine or CDMA uses spread spectrum technology. So, it's simple to implement. So, fixed channel allocation is simple to implement if traffic is uniform

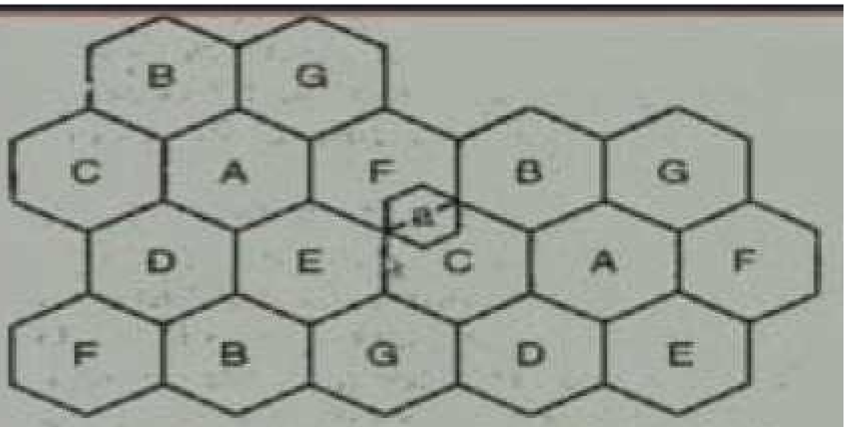
b) Channel Borrowing Technique.

sometimes traffics are not uniform – there may be two cells, which are side by side. So, this has been given one band. One has been given another band of frequencies. They do not interfere with each other. But, we find for each cell.

Now, I find that in one particular cell, the user density is much higher, whereas in adjacent cell, the user density is lower. So, I could use some more bandwidth in this cell and I could do with a little less bandwidth here. So, what could do is that, a part of this frequency band can borrow from the adjacent cell. So, that is called Channel Borrowing technique. High traffic cells borrow channel frequencies from low traffic cells. Temporary channel borrowing and static channel borrowing.

Cell Splitting

When the number of subscribers in a given area increases, allocation of more channels covered by that cell is necessary. What happens is that in one area, say a small town, one base station could satisfy people, who had these cellular phones or mobile phones. Now what happens is that, the number of people who wanted to use mobile phones kept on increasing and now we cannot serve them any longer. The number of requests, which are denied, keeps on increasing. How can we increase? May be break it up into two cells and then break it up into four cells and break it up into many more cells, depending on the clusters of users and the cells. Now, the same area has been divided into smaller cells. May be in the BS, you decrease the transmitter power so that they do not interfere with each other. So when the number of subscribers in a given area increases, allocation of more channels covered by that cell is necessary. This is done by cell splitting. A single small cell midway between two co-channel cells may be introduced.

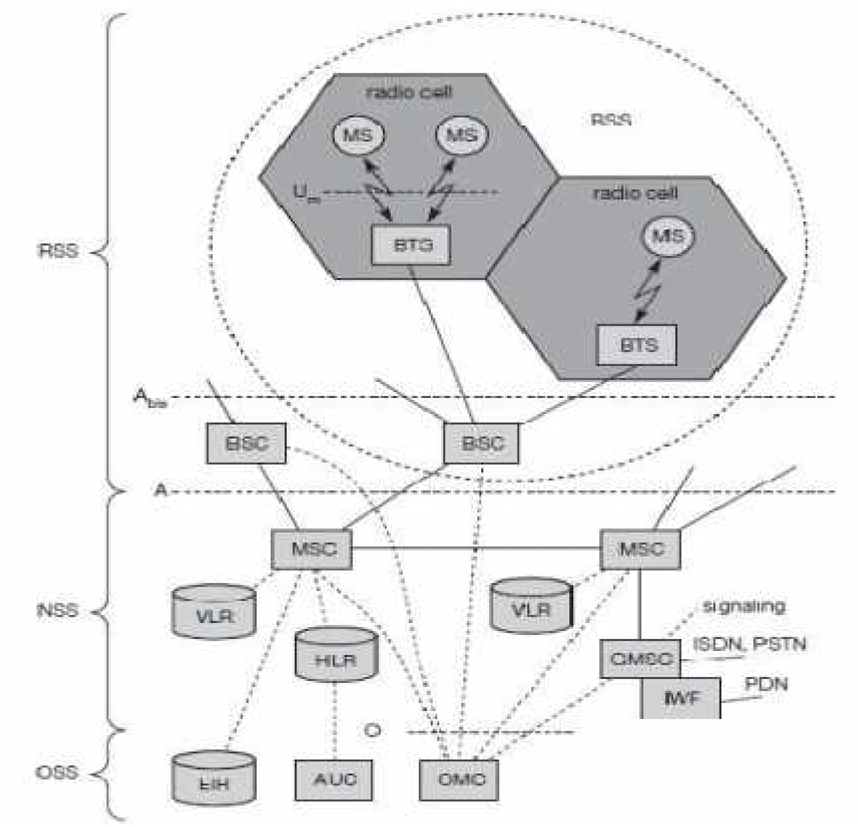


Illustrations of cell splitting for capacity expansion.

For example over here, you had a large number of cells. We created a small cell over here using A, which uses the same frequency bands as the already existing ones. You cannot use E, F, C, or B. But you can use A with other cells. So that is called cell splitting. These are ad hoc solutions, when a particular area has more number of users

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).



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

Home location register (HLR): It is a database used for storage and management of subscriptions. HLR stores permanent data about subscribers, including a subscribers service profile, location information and activity status. When an individual buys a subscription from the PCS provider, he or she is registered in the HLR of that operator.

Visitor location register (VLR): It is a database that contains temporary information about subscribers that is needed by the MSC in order to service visiting subscribers. VLR is always integrated with the MSC. When a MS roams into a new MSC area, the VLR

connected to that MSC will request data about the mobile station from the HLR. Later if the mobile station needs to make a call, VLR will be having all the information needed for call setup.

- ✓ 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 controls calls to and from other telephone and data systems.

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 BTS. A number of BSC's are served by and MSC.

Base transceiver 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 **Abis 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 and Support system: The operations and maintenance center (OMC) is connected to all equipment in the switching system and to the BSC. Implementation of OMC is called operation and support system (OSS). The OSS is the functional entity from which the network operator monitors and controls the system. 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.

Radio Interface:

A GSM system has 124 pairs of simplex channels. They are in pairs because one goes from BS to MS and the other from BS to MS. Each of these is 200 kilo hertz wide and supports 8 separate connections on it, using TDM. So, each active station is assigned to one time slot on one channel pair. 992 channels can be supported in each cell, but many of them are not available to avoid frequency conflicts with neighboring cells.

Transmitting and receiving does not happen in the same time slot because the GSM radios cannot transmit and receive at the same time and it takes time to switch from one to another. That is why different time slots are given. A data frame is transmitted in 547 microseconds, but a transmitter is only allowed to send one data frame every 4.615 milliseconds, since it is sharing the channel with seven other stations. The gross rate of each channel is about 270 or about 271 kbps divided among eight users. This gives about 33 or 34 kbps gross. CC i.e., control channels are used to manage the system if somebody is getting only 33 or 34 kbps. Previously, we have been talking about voice channel requiring 64 kbps.

Now, the 64 kbps happens to be if you are doing a plain vanilla PCM. That means we have explained, how it is encoded by sampling it at eight samples and eight levels for each sample – that gives us 64 kbps. The point is that, it is not the only coding scheme. Actually, there are more advanced coding schemes. We did not find time to discuss those coding schemes. Using those coding schemes, good quality voice transmission can be achieved, using a much lower bandwidth. This 33.854 kbps is actually enough, if you are doing your coding in a smart fashion.

Control Channels

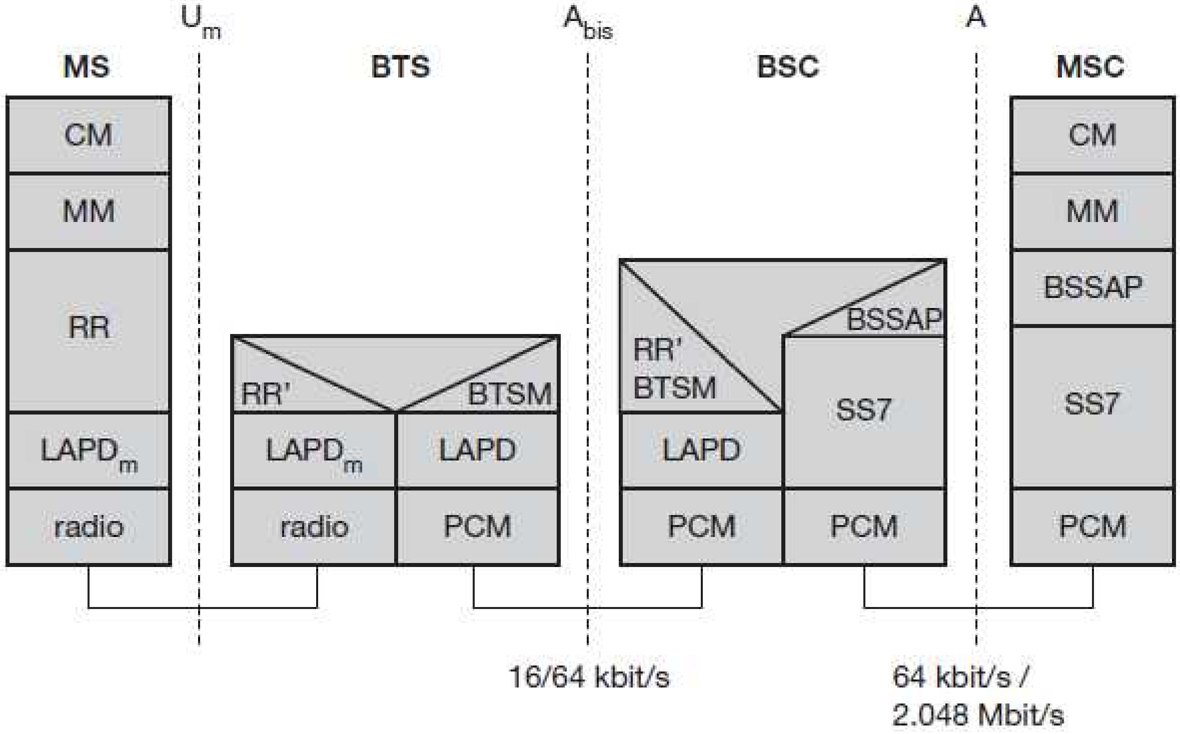
1) The Broadcast Control Channel (BCC) is a continuous stream of output from the base station containing the BS's identity and the channel status. All mobile stations monitor their signal strength to see when they moved into a new cell. The point is that the mobile station, when it gets these broadcasts from BS, by just sensing how much transmitter power it is getting, it can identify whether it is near this particular BCC, what this particular BS is, or what its identity is, or whether it is near some other BS. In some systems like CDMA, this power is very crucial even for decoding purposes.
2) dedicated control channel is used for location updating, registration, and call setup; in particular, each BS maintains a database of mobile stations, which are in its area. So, information needed to maintain this database is sent on the dedicated control channel. So, the point is that these mobile stations are moving.

Common Control Channel

1) **The paging sub channel-** In which the BS uses to announce incoming calls. Each MS monitors it continuously to watch for the call it should answer. The point is that, if there is a call, and MS is in the area of some BS and then somebody wants to call to this MS, which one particular MS has to be alerted. So, there is a paging for that MS from the BS and the MS is always listening to it.
2) **The random access channel-** This allows users to request a slot on the dedicated control channel. If two requests collide, they are garbled and have to be retried later on. So, this is the part of the call set-up. So, it is the first part of call set-up. It tries to put a request in the random access channel for a slot in the dedicated control channel. When it gets a slot in the dedicated control channel, it can go away with the further steps of call set-up. Next is the access grant channel.
3) **The access grant channel-** The announced assigned slot.

GSM Protocols

The signalling 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 qualit**y on the downlink. The physical layer at Um uses GMSK for digital **modulation** and performs **encryption/decryption** of data, i.e., encryption is not performed end-to-end, but only between MS and BSS over the air interface



The main tasks of the physical layer comprise **channel coding** and **error detection/correction**, which is directly combined with the coding mechanisms. Channel coding makes extensive use of different **forward error correction (FEC)** schemes. Signaling between entities in a GSM network requires higher layers. For this purpose, the **LAPDm** protocol has been defined at the Um interface for **layer two**. LAPDm has been derived from link access procedure for the D-channel (**LAPD**) in ISDN systems, which is a version of HDLC. LAPDm is a lightweight LAPD because it does not need synchronization flags or checksumming for error detection. LAPDm offers reliable data transfer over connections, re- sequencing of data frames, and flow control.

The network layer in GSM, layer three, comprises several sublayers. The lowest sublayer is the radio resource management (RR). Only a part of this layer, RR', is implemented in the BTS, the remainder is situated in the BSC. The functions of RR' are supported by the BSC via the BTS management (BTSM). The main tasks 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. This layer also provides functions to send in-band tones, called dual tone multiple frequency (DTMF), over the GSM network. These tones are used, e.g., for the remote control of answering machines or the entry of PINs in electronic banking and are, also used for dialing in traditional analog telephone systems.

Additional protocols are used at the Abis and A interfaces. 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)**.

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:

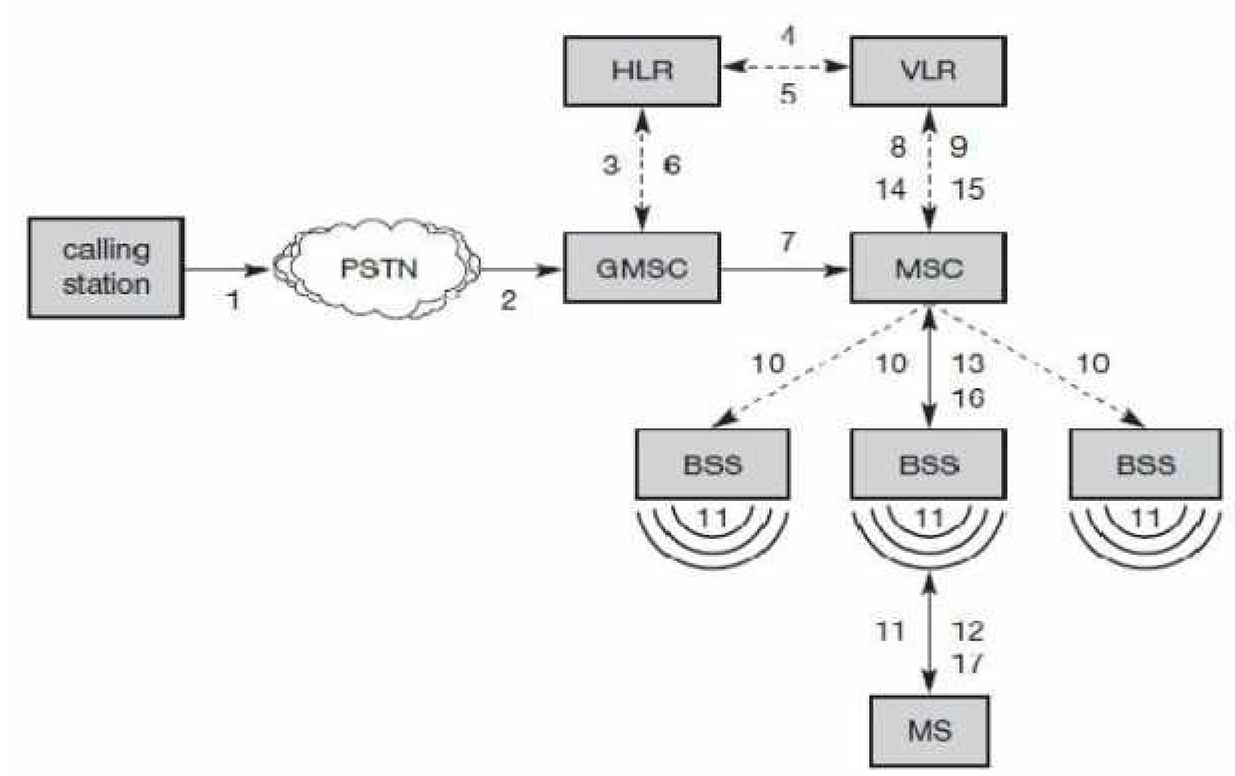
Mobile station international ISDN number (MSISDN):- The only important number for a user of GSM is the phone number. This number consists of the country code (CC), the national destination code (NDC) and the subscriber number (SN).

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 roaming number (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)*, the following figure shows the different steps that take 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 subscriber and signals the call setup to 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

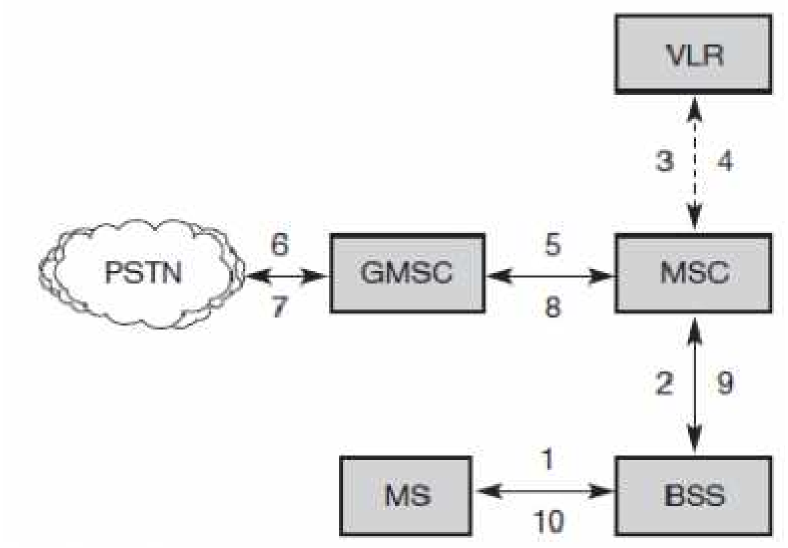
step 10: If MS is available, the MSC initiates paging in all cells it is responsible for.

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

For a **mobile originated call (MOC),** the following steps take place:

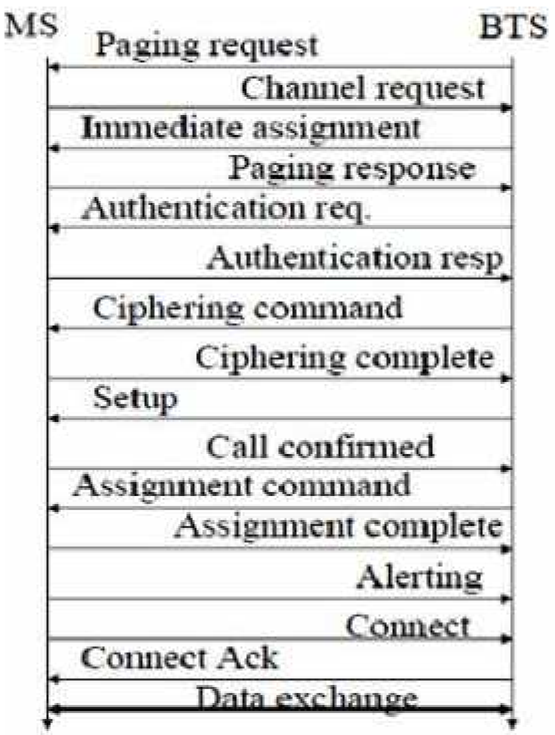


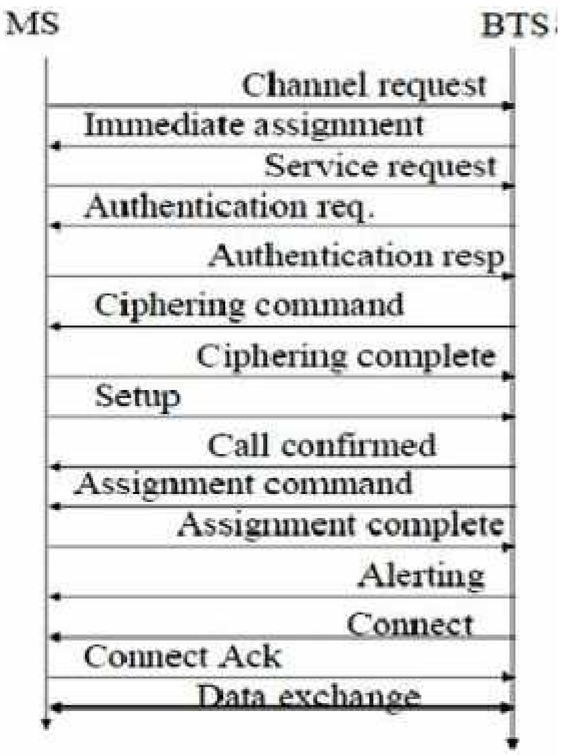
The MS transmits a request for a new connection

step 2: The BSS forwards this request to the MSC

step 3: The MSC then checks if this user is allowed to set up a call with the requested and checks the availability of resources through the GSM network and into the PSTN. If all resources are available, the MSC sets up a connection between the MS and the fixed network.

In addition to the steps mentioned above, other messages are exchanged between an MS and BTS during connection setup (in either direction).





Mobility Management:-

1) Location Management
    - Access point of a mobile station(MS) changes as it moves around the network coverage area.
    - Important for effective delivery of incoming calls

2) Handoff management
    - Access point of a mobile station changes as it moves around the network coverage area and important for effective delivery of incoming cells and other is handoff management.
3) Location Updates
    - Each time the MS makes an update to its location a database in the fixed part of the network has to be updated to reflect the new location information. So, that for a particular MS, if you go to the data base and find out what is the last point, where he said, that he was. Of course what he might have done is that he might have switched off his mobile and then moved to somewhere else and then put it on or something.
4) Paging
    - Required to deliver an incoming message to the MS
    - Response from the paged terminal enables the network to locate the MS
5) Location Information

- Procedures to store and distribute the location information related to MSs serviced by the network.

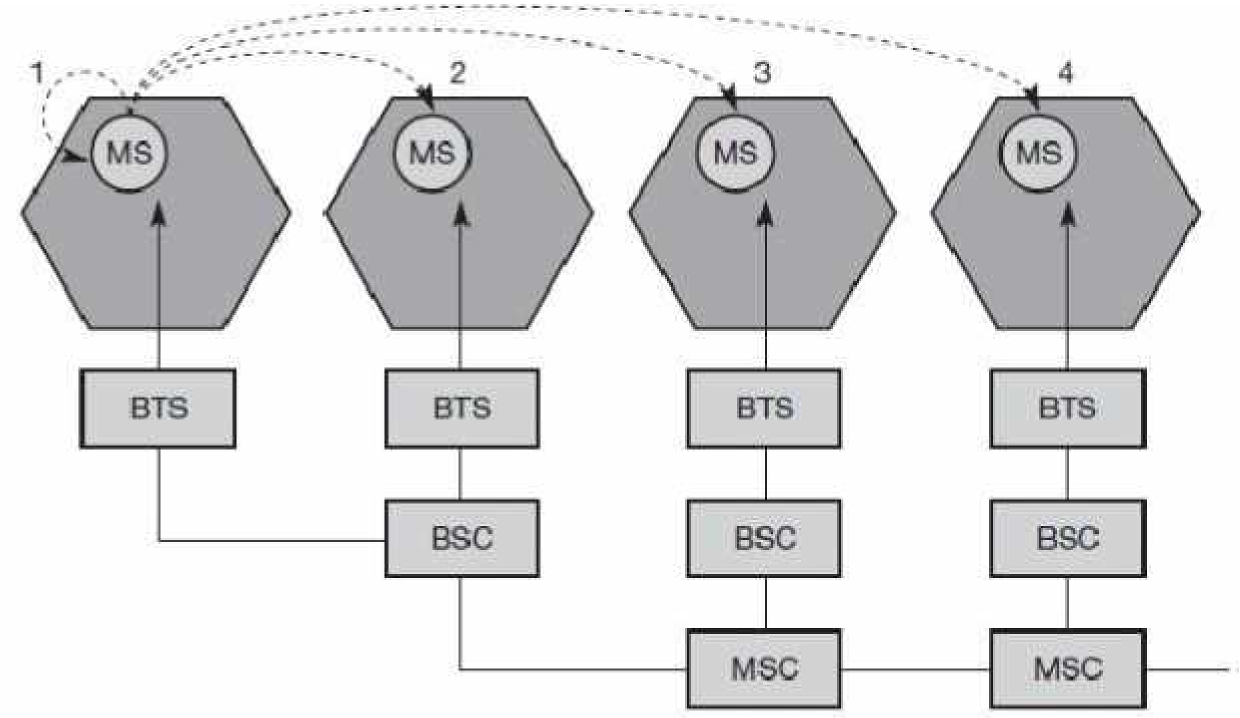
Handover:-

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 a cell, narrow-band interference could make transmission at a certain frequency impossible. The BSC could then decide to change the carrier frequency (scenario 1).

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, assigns a new radio channel in the new cell and releases the old one (scenario 2).

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 (scenario 3).

Inter MSC handover: A handover could be required between two cells belonging to different MSCs. Now both MSCs perform the handover together (scenario 4).

GPRS (General Packet Radio Service)

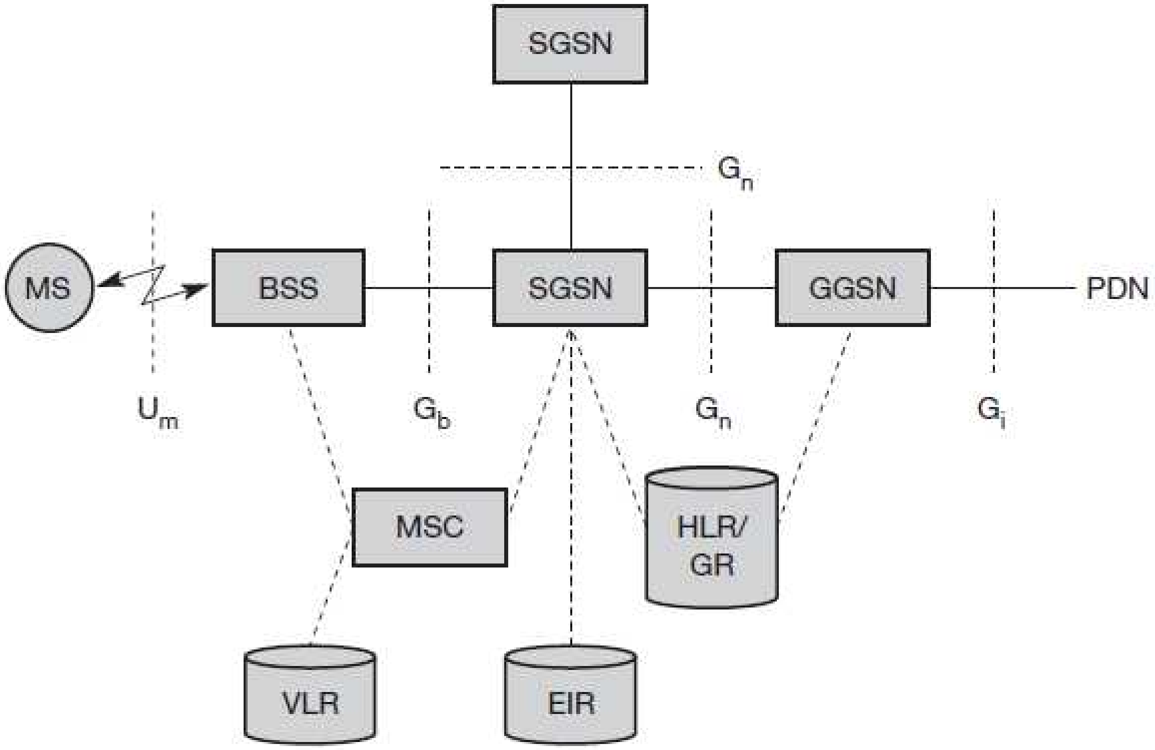
Is a packet based communication service for mobile devices that allows data to be sent and received across a mobile telephone network. GPRS is a step towards 3G and is often referred to as 2.5G. Here are some key benefits of GPRS:

Speed :-GPRS is packet switched. Higher connection speeds are attainable at around 56–118 kbps, a vast improvement on circuit switched networks of 9.6 kbps. By combining standard GSM time slots theoretical speeds of 171.2 kbps are attainable. However in the very short term, speeds of 20-50 kbps are more realistic.

Always on connectivity GPRS:- Is an always-on service. There is no need to dial up like you have to on a home PC for instance. This feature is not unique to GPRS but is an important standard that will no doubt be a key feature for migration to 3G. It makes services instantaneously available to a device.

New and Better applications:- Due to its high-speed connection and always-on connectivity GPRS enables full Internet applications and services such as video conferencing straight to your desktop or mobile device. Users are able to explore the Internet or their own corporate networks more efficiently than they could when using GSM. There is often no need to redevelop existing applications.

GSM operator Costs:- GSM network providers do not have to start from scratch to deploy GPRS. GPRS is an upgrade to the existing network that sits along side the GSM network. This makes it easier to deploy, there is little or no downtime of the existing GSM network whilst implementation takes place, most updates are software so they can be administered remotely and it allows GSM providers to add value to their business at relatively small costs. The GSM network still provides voice and the GPRS network handles data, because of this voice and data can be sent and received at the same time.



As mentioned earlier GPRS is not a completely separate network to GSM. Many of the devices such as the base transceiver stations and base transceiver station controllers are still used. Often devices need to be upgraded be it software, hardware or both. When deploying GPRS many of the software changes can be made remotely.

There are however two new functional elements which play a major role in how GPRS works. The Serving GPRS Support Node (SGSN) and the Gateway GPRS support node (GGSN). These 2 nodes are new to the network with the other changes being small if any.

In simple terms there are in practice two different networks working in parallel, GSM and GPRS. In any GSM network there will be several BSC's (Base Station Controllers). When implementing GPRS, a software and hardware upgrade of this unit is required. The hardware upgrade consists of adding a Packet Control Unit (PCU). This extra piece of hardware differentiates data destined for the standard GSM network or Circuit Switched Data and data destined for the GPRS network or Packet Switched Data. In some cases a PCU can be a separate entity.

From the upgraded BSC there is a fast frame relay connection that connects directly to the newly introduced SGSN.

SGSN

The Serving GPRS Support Node, or SGSN for short, takes care of some important tasks, including routing, handover and IP address assignment.

The SGSN has a logical connection to the GPRS device. As an example, if you where in a car travelling up the M1 on a long journey and were browsing the Internet on a GPRS device, you will pass through many different cells. One job of the SGSN is to make sure the connection is not interrupted as you make your journey passing from cell to cell. The SGSN works out which BSC to “route” your connection through.

If the user moves into a segment of the network that is managed by a different SGSN it will perform a handoff of to the new SGSN, this is done extremely quickly and generally the user will not notice this has happened. Any packets that are lost during this process are retransmitted. The SGSN converts mobile data into IP and is connected to the GGSN via a tunneling protocol.

GGSN

The Gateway GPRS Support Node is the “last port of call” in the GPRS network before a connection between an ISP or corporate network’s router occurs. The GGSN is basically a gateway, router and firewall rolled into one. It also confirms user details with RADIUS servers for security, which are usually situated in the IP network and outside of the GPRS network.

Connectivity between the SGSN & GGSN

The connection between the two GPRS Support Nodes is made with a protocol called GPRS Tunnelling Protocol (GTP). GTP sits on top of TCP/IP and is also responsible for the collection of mediation and billing information. GPRS is billed on per megabyte basis unlike GSM.

HLR

The HLR or Home Location Register is a database that contains subscriber information, when a device connects to the network their MSISDN number is associated with services, account status information, preferences and sometimes IP addresses.

GPRS Handset Classes GPRS devices are not as straightforward as you may think. There are in fact 3 different classes of device.

Class A

Class A terminals have 2 transceivers which allow them to send / receive data and voice at the same time. This class of device takes full advantage of GPRS and GSM. You can be taking a call and receiving data all at the same time.

Class B

Class B devices can send / receive data or voice but not both at the same time. Generally if you are using GPRS and you receive a voice call you will get an option to answer the call or carry on.

Class C

This device only allows one means of connectivity. An example would be a GPRS PCMCIA card in a laptop.

Unit -2

Wireless Networks

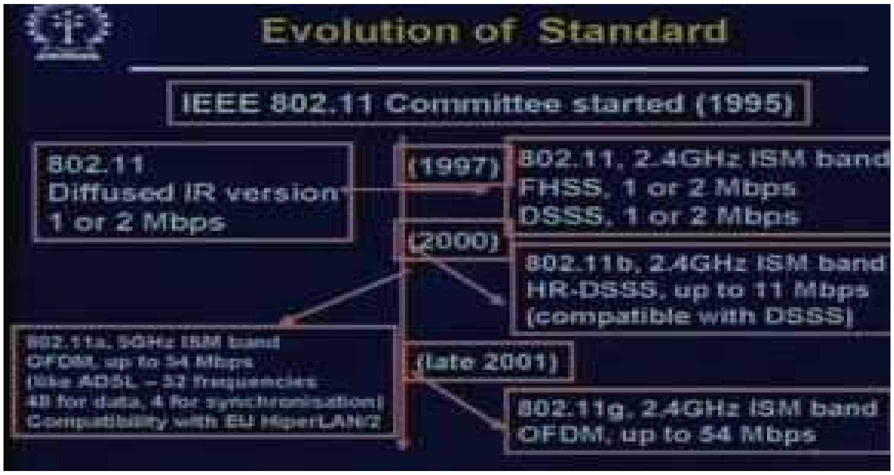
Wireless LAN:-

A LAN means a local area network that works without wires, which means you do not have to wire up the whole place; you do not have to have a wire coming into your system; you can walk into a room with a laptop and you are already on the network. But this has some peculiar problems. This is not as easy since signals are of limited range. Unlike wired LAN, if A can hear B and B can hear C, it is not necessarily true that A can hear C. So this is a problem which we have to handle; secondly in many of the cases, these wireless LANs use unlicensed frequencies and low power. Low power is important because you want to have a small-sized cell so that in another part of the building there may be another cell just giving services to another group of users. As we know that this way, by doing space division multiplexing, we can increase the number of users who are on the network. One of the most important LAN standards today, wireless LAN standard, is 802.11 and there are various versions of 802.11. The speed varies from 2 mbps to 54 mbps. We will also talk a little bit about Bluetooth, which is a personal area network. We will talk a little bit about wireless MAN, which is 802.16 and just mention of few other emerging technologies.

WLAN requirement This is some kind of a wish list actually – what all we would want from wireless LAN. Good use of bandwidth is we want – high throughput – everybody uses a number of nodes; it should be large, may be in the hundreds. A good connection to LAN backbone is required because nowadays just a local network by itself is of limited utility since everybody is getting use to be connected to the entire network meaning the internet even all the time.

so the backbone connectivity is also important; good service coverage, or range; minimal battery power consumption this an important issue in any kind of mobile system because if the battery consumption becomes high, either you have to carry heavier batteries or you have to charge them often so that is not good so we want minimal battery power consumption; transmission security and robustness – this may be an issue in many cases – because you know so in a wireless system the medium is of course open to everybody alright including snoopers if any so but you would like your communication to remain somewhat private or protected and in some cases that may even become crucial so we want security and robustness ok and some collocated network operation.

Evolution of Standard



Evolution of Standard

IEEE 802.11 Committee started (1995)

802.11
Diffused IR version
1 or 2 Mbps

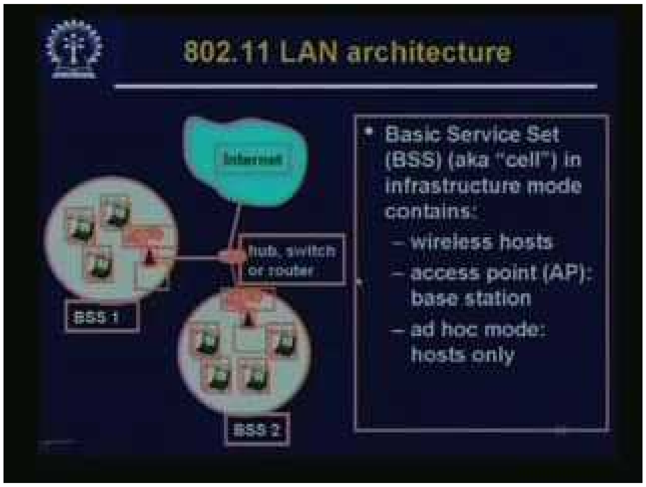
(1997)
802.11, 2.4GHz ISM band
FHSS, 1 or 2 Mbps
DSSS, 1 or 2 Mbps

(2000)
802.11b, 2.4GHz ISM band
HR-DSSS, up to 11 Mbps
(compatible with DSSS)

802.11a, 5GHz ISM band
OFDM, up to 54 Mbps
(like ADSL – 52 frequencies
48 for data, 4 for synchronisation)
Compatibility with EU HiperLAN/2

(late 2001)
802.11g, 2.4GHz ISM band
OFDM, up to 54 Mbps

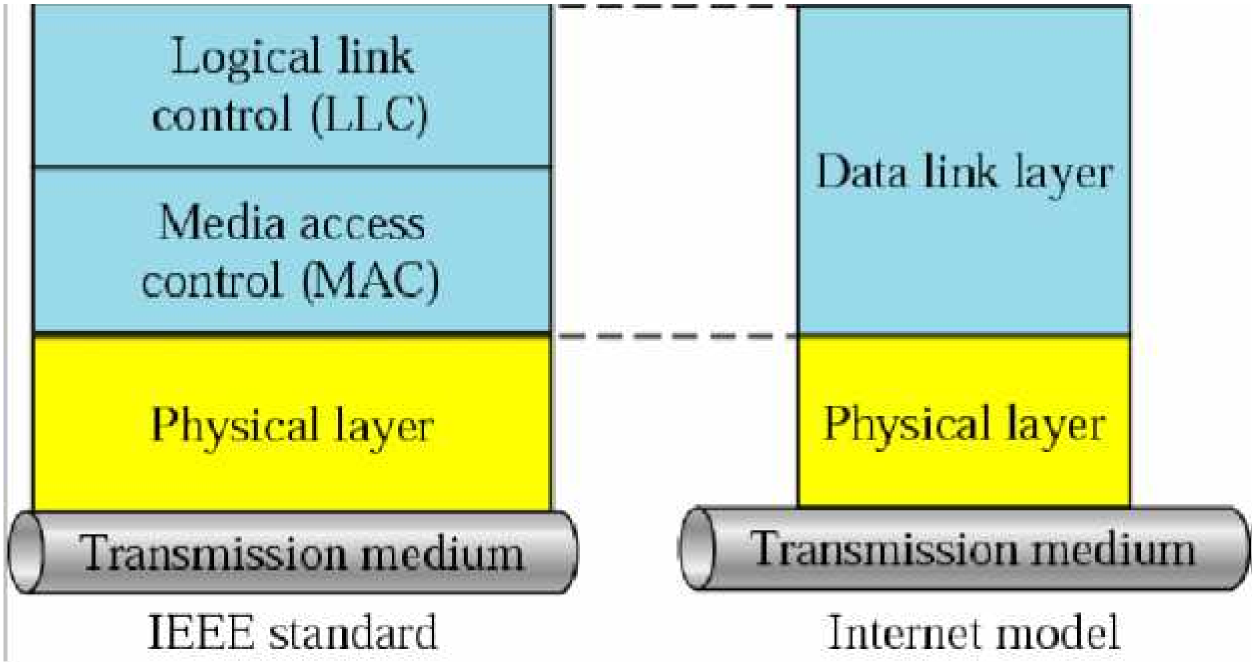
WLAN (802.11) architecture



802.11 LAN architecture

- Basic Service Set (BSS) (aka "cell") in infrastructure mode contains:
  - wireless hosts
  - access point (AP): base station
  - ad hoc mode: hosts only

Media Access Control (**MAC**) data communication protocol sub-layer, also known as the Medium Access Control, is a sublayer 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) sublayer 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.

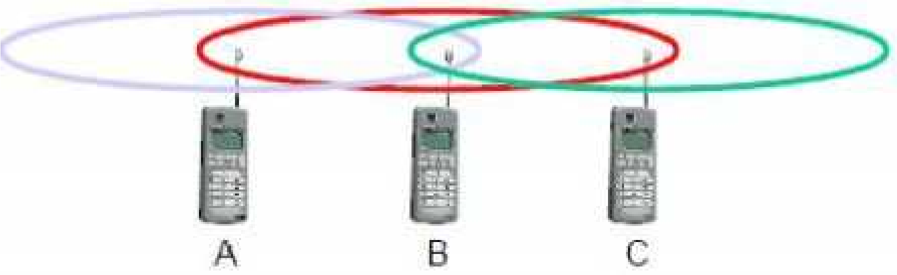


Motivation for a specialized MAC

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, CS might not work, if for e.g., a terminal is “hidden”

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 hear A
- C wants to send to B, C senses a "free" medium (CS 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

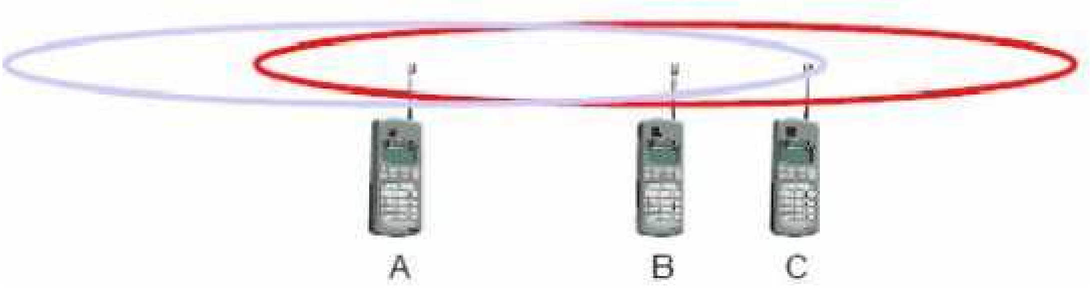
Exposed terminals

- 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.

Near and far terminals

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



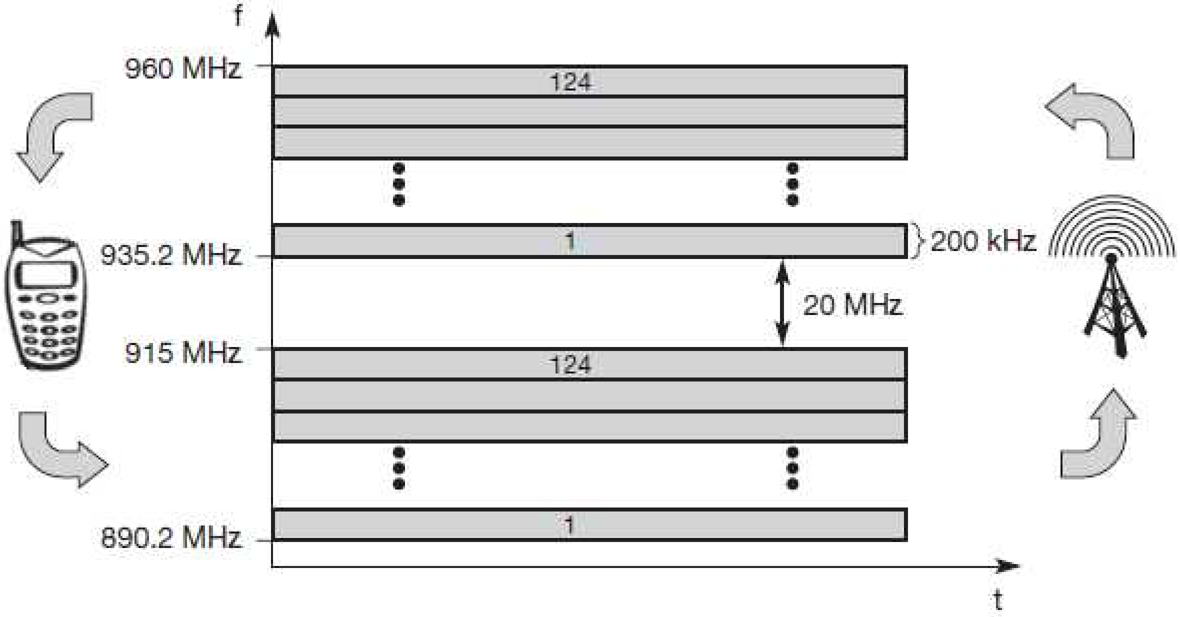
- Signal strength decreases proportional to the square of the distance
- So, B's signal drowns out A's signal making C unable to receive A's transmission
- If C is an arbiter for sending rights, B drown out A's signal on the physical layer making C unable to hear out A.

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.

FDMA

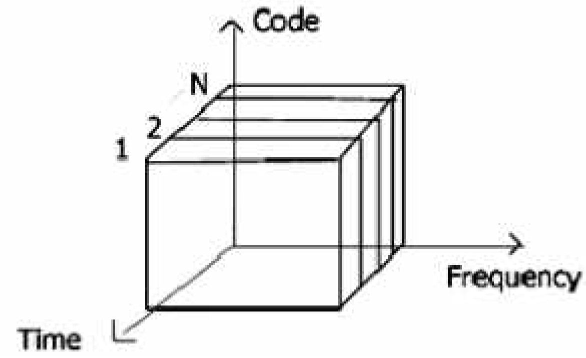
Frequency division multiplexing (FDM) describes schemes to subdivide the frequency dimension into several non-overlapping frequency bands.

Frequency 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.



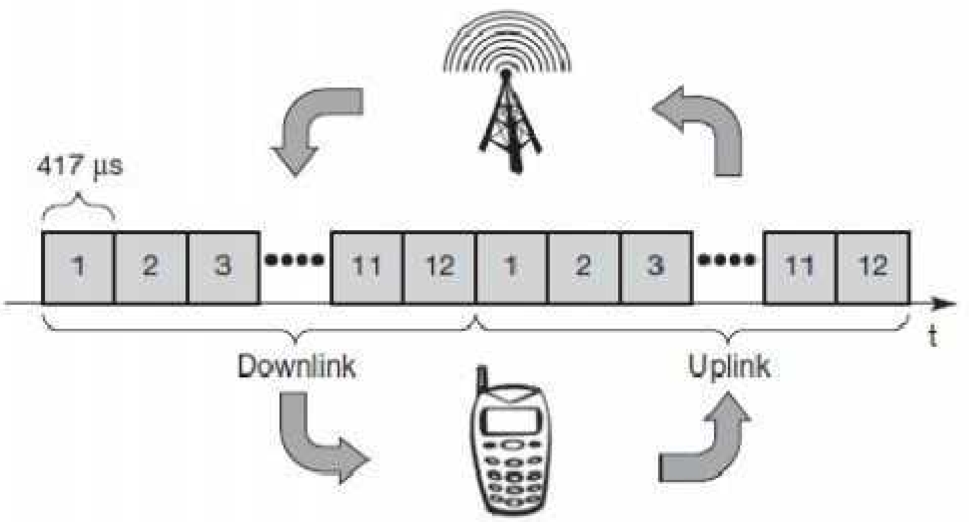
TDMA

A more flexible multiplexing scheme for typical mobile communications is time division multiplexing (TDM). Compared to FDMA, time division multiple access (TDMA) offers a much more flexible scheme, which comprises all technologies that allocate certain time slots for communication. Now synchronization between sender and receiver has to be achieved in the time domain. Again this can be done by using a fixed pattern similar to FDMA techniques, i.e., allocating a certain time slot for a channel, or by using a dynamic allocation scheme



Fixed TDM

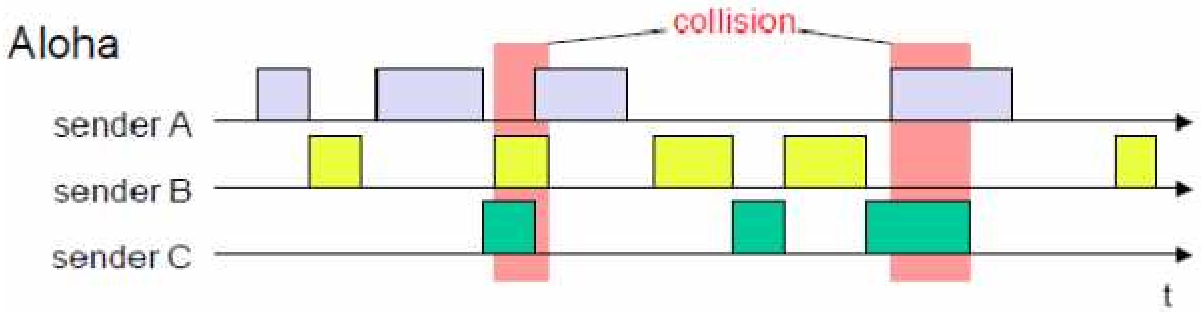
The simplest algorithm for using TDM is allocating time slots for channels in a fixed pattern. This results in a fixed bandwidth and is the typical solution for wireless phone systems. MAC is quite simple, as the only crucial factor is accessing the reserved time slot at the right moment. If this synchronization 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 above 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

Classical Aloha

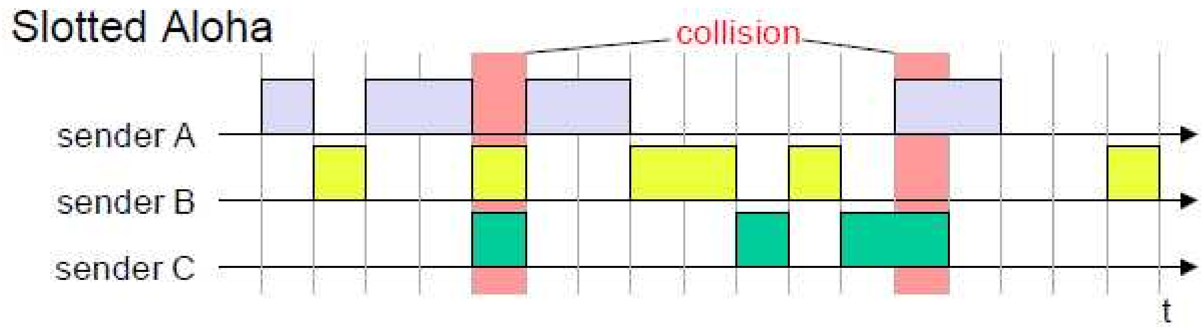
In this scheme, TDM is applied without controlling medium access. Here each station can access the medium at any time as shown below:



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.

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.

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 a 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

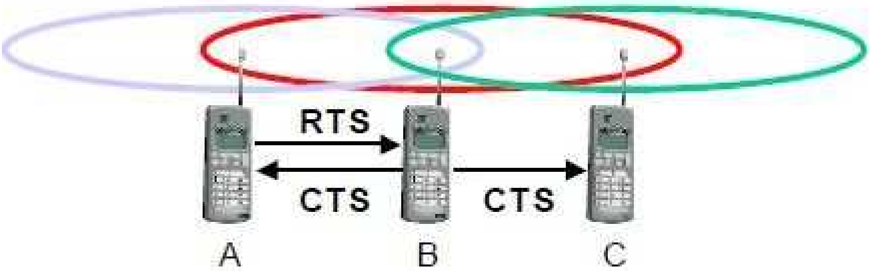
CSMA with collision avoidance (CSMA/CA) is one of the access schemes used in wireless LANs following the standard IEEE 802.11. Here sensing the carrier is combined with a back-off scheme in case of a busy medium to achieve some fairness among competing stations.

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.

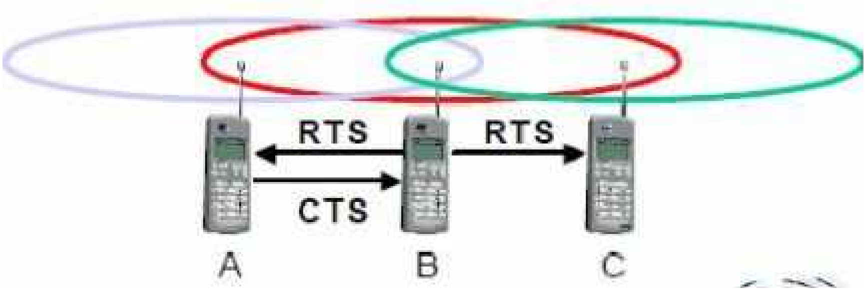
A starts sending to B, C does not receive this transmission. C also wants to send something to B and senses the medium. The medium appears to be free, the carrier sense fails. C also starts sending causing a collision at B. But A cannot detect this collision at B and continues with its transmission. 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 an 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 transmissionassuming it will not cause a collision at A. The problem with exposed terminals is solved without fixed access patterns or a base station.

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.

Example: Randomly Addressed Polling

- 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

Mobile IP

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.

Entities and terminology

The following defines several entities and terms needed to understand mobile IP

□ **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 its IP 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 is not 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.

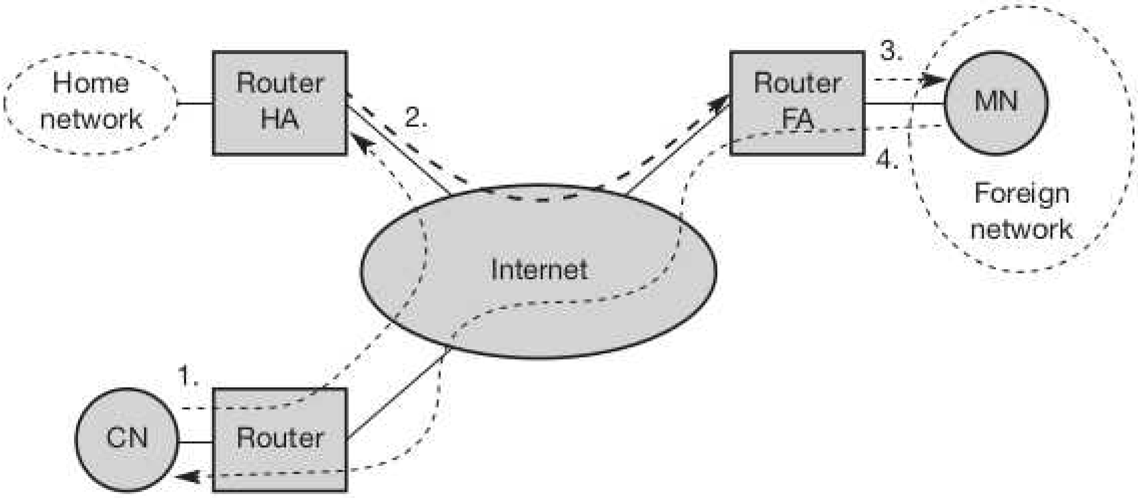
□ **Care-of address (COA):** The COA defines the current location of the MN from an IP point of view. All IP packets sent to the MN are delivered to the COA, not directly to the IP address of the MN. Packet delivery toward the MN is done using a tunnel, i.e., the COA marks the tunnel endpoint, i.e.,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

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.

CN sends an IP packet with MN as a destination address and CN as a source address. The internet, 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 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.

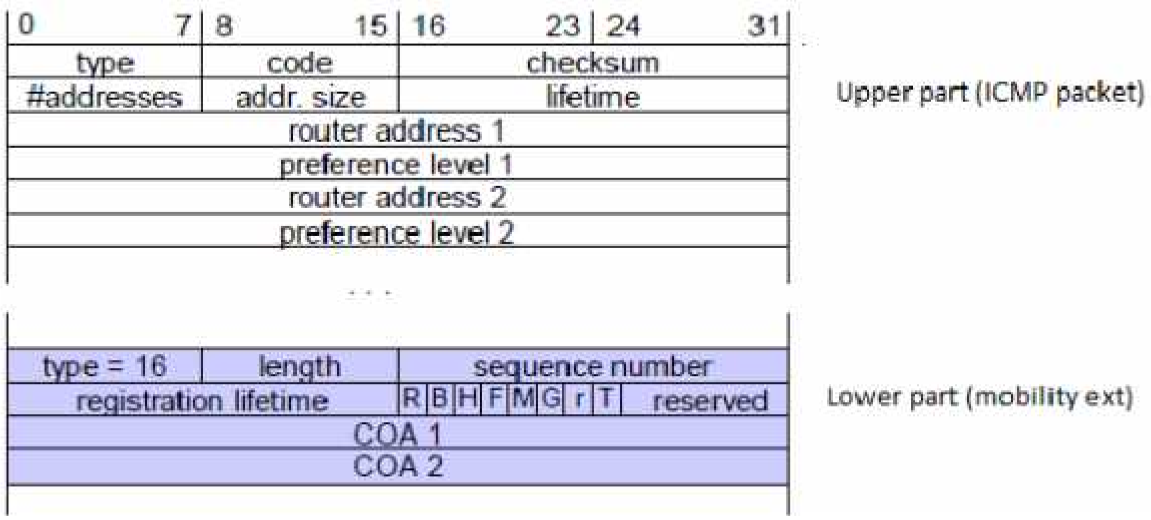
Working of Mobile IP:- Mobile IP has two addresses for a mobile host: one home address and one care- of address. The home address is permanent; the care-of address changes as the mobile host moves from one network to another. To make the change of address transparent to the rest of the Internet requires 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:



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 bindings 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.

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.

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**.

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.

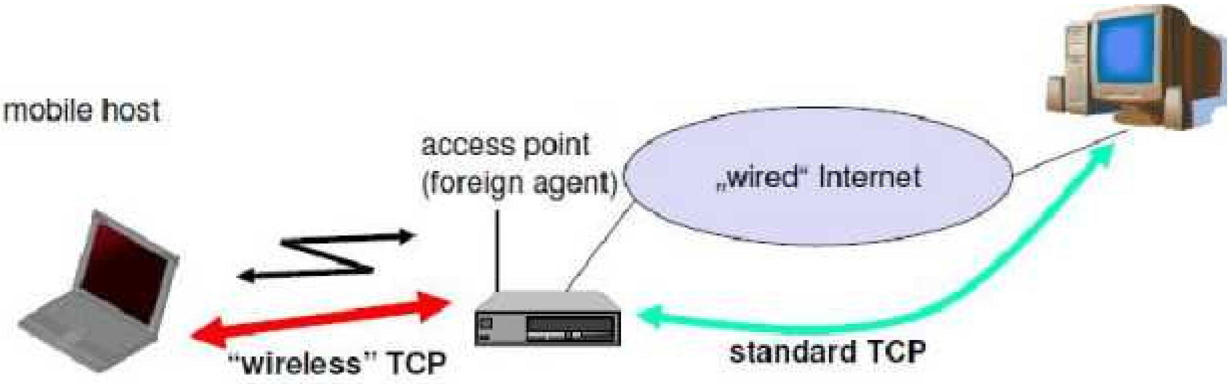
Problems with Traditional TCP in wireless environments

- ➢ 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

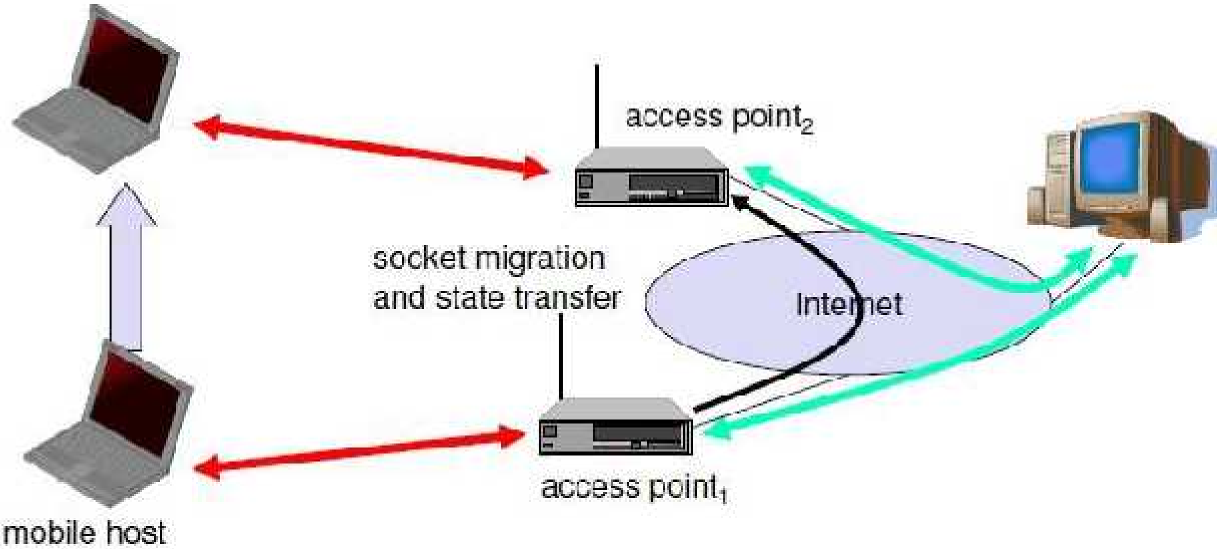
Indirect TCP (I-TCP)

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



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.



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

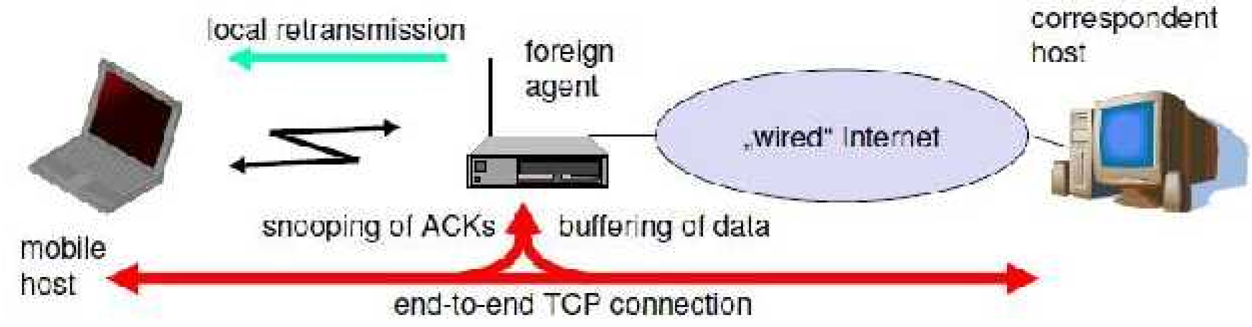
□ Loss of end-to-end semantics:- an acknowledgement to a sender no longer means that a receiver really has received a packet, foreign agents might crash.

□ 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.

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
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.

Unit 3

Mobile Data base

- Recent advances in portable and wireless technology led to **mobile computing**, a new dimension in data communication and processing.
- Portable computing devices coupled with wireless communications allow clients to access data from virtually anywhere and at any time.
- There are a number of hardware and software problems that must be resolved before the capabilities of mobile computing can be fully utilized.
- Some of the software problems – which may involve data management, transaction management, and database recovery – have their origins in distributed database systems.
- In mobile computing, the problems are more difficult, mainly:
    - The limited and intermittent connectivity afforded by wireless communications.
    - The limited life of the power supply(battery).
    - The changing topology of the network.
    - In addition, mobile computing introduces new architectural possibilities and challenges.

Mobile Database Management Issue

Building an integrated platform to manage complexities demands a scalable, robust environment providing the following fundamental services: data management, connection management, integration management, mobility management, handoff management

Data Management:

Today’s mobile applications require more than simple data synchronization. They require a complete set of data management services, including strong data modeling, mobile and server-side support for schema deployment and versioning, rules-based data distribution, bi-directional data transfers that are fast and secure, mobile device-based database services, and tight transaction-level integration with multiple enterprise information sources. The asymmetric nature of the wireless communication link is another challenge for data management in wireless links to ensure low consumption and data access latency.

Connection Management:

Today, mobile connection management is technically complex and esoteric, and it varies widely as travelling across the globe. Newcomers to mobile computing must wrestle with the plethora of emerging communication protocols, standards, and low-level operational aspects of wireless connectivity. However, a mobile platform should provide the ability to seamlessly service multiple connection methods, wireless connectivity service options, and handheld device types at the same time. Load balancing and scalability options should be provided to handle volume and frequency spikes as they occur, connections between mobile devices and the enterprise should be secure, efficient and extremely reliable.

Integration Management:

In a mobile platform approach, integration management services provide flexible and robust methods for tying into multiple back-end information sources. The requirement for data transformation and business data processing before entry into the back-end source is a key issue. Perhaps the most important aspect of integration management from the mobile platform perspective is the ability to extend the investment made in large corporate information systems to the mobile workforce in an efficient, transparent and meaningful way.

Mobility management:

Location management on mobile devices will become increasingly important in the new future, considering the increasing number of location-enabled mobile devices and location-based services.. However, there are two issues, one is, how to make location information openly available on the Web, and the second is, how to provide users with privacy control in such an environment. Location management is a two-stage process that enables the network to discover the current attachment point of the mobile user for call delivery. The first stage is location registration (or location update, the mobile terminal periodically notifies the network of its new access point, allowing the network to authenticate the user and revise the user's location file. The second stage is call delivery. Here, the network is queried for the user location profile and the current position of the mobile host is found.

Handoff Management:

Handover management enables the network to maintain a user’s connection as the mobile terminal continues. Mobility Management in Next-Generation wireless systems moves and changes its access point to the network. The three-stage process for handoff first involves initiation- where the user, a network agent, or changing network conditions identify the need for handoff. The second stage is new connection generation- where the network must find new resources for handoff connection and perform any additional routing operations. Under network-controlled handoff, or mobile-assisted handoff, the network generates a new connection, finding new resources for the handoff and performing any additional routing operations. For mobile-controlled handoff, the mobile terminal finds the new resources and the network approves. The final stage is data-flow control- where the delivery of the data from the old connection path to the new connection path is maintained according to agreed upon service, mobile terminal finds the new resources and the network approves. The final stage is data-flow control- where the delivery of the data from the old connection path to the new connection path is maintained according to agreed-upon service guarantees.

CHALLENGES OF MOBILE DATABASE SYSTEM

- Limited Resources
- Power consumption:
- Disconnection
- Insufficient bandwidth:
- Limited storage
- Limited battery power:

Database Replication

- Functionality of DDBMS is attractive. However, implementations of required protocols and algorithms are complex and can cause problems that may outweigh advantages.
- Alternative and more simplify approach to data distribution is provided by a replication server.
- Every major database vendor has replication solution.
- Database Replication is the process of copying and maintaining database objects, such as relations, in multiple databases that make up a distributed database system.

Benefits of Database Replication

- Availability
- Reliability
- Performance
- Load reduction
- Disconnected computing
- Supports many users
- Supports advanced applications

Applications of Replication

- Replication supports a variety of applications that have very different requirements.
- Some applications are supported with only limited synchronization between the copies of the database and the central database system.
- Other applications demand continuous synchronization between all copies of the database.

Basic Components of Database Replication

- Replication object is a database object such as a relation, index, view, procedure, or function existing on multiple servers in a distributed database system.
- In a replication environment, any updates made to a replication object at one site are applied to the copies at all other sites.
- Replication objects are managed using replication groups.

- A replication group is a collection of replication objects that are logically related.
- A replication group can exist at multiple replication sites.
- Replication environments support two basic types of sites: *master sites* and *slave sites*.
- A replication group can be associated with one or more master sites and with one or more slave sites.
- One site can be both master site for one replication group and slave site for different replication group.
- However, one site cannot be both the master site and slave site for same replication group.
- A master site controls a replication group and the objects in that group.
- This is achieved by maintaining a complete copy of all objects in a replication group and by propagating any changes to a replication group to any slave sites.
- A slave site can contain all or a subset of objects from a replication group. However, slave sites only contain a snapshot of a replication group.
- Typically, a snapshot site is refreshed periodically to synchronize it with its master site.
- For a replication environment with many master sites, all of those sites communicate directly with one another to continually propagate data changes in the replication group.

Synchronous Versus Asynchronous Replication

- *Synchronous* – updates to replicated data are part of enclosing transaction.
    - If one or more sites that hold replicas are unavailable transaction cannot complete.
    - Large number of messages required to coordinate synchronization.
- *Asynchronous* - target database updated after source database modified. Delay in regaining consistency may range from few seconds to several hours or even days.

Replication Servers Functionality

- Basic function is copy data from one database to another (using synch. or asynch. replication).
- Other functions include:
    - Scalability
    - Mapping and Transformation
    - Object Replication
    - Specification of Replication Schema
    - Subscription mechanism
    - Initialization mechanism
    - Easy Administration

Implementation Issues

- Issues associated with the provision of data replication by the replication server include:
    - transactional updates;
    - snapshots and database triggers;
    - conflict detection and resolution.

Conflict Detection and Resolution

- When multiple sites are allowed to update replicated data, need to detect conflicting updates and restore data consistency.
- For a single table, source site could send both old and new values for any rows updated since last refresh.
- At target site, replication server can check each row in target database that has also been updated against these values.
- Also want to detect other types of conflict such as violation of referential integrity.
- Some of most common mechanisms are:
    - Earliest and latest timestamps.
    - Site Priority.
    - Additive and average updates.
    - Minimum and maximum values.
    - User-defined.
    - Hold for manual resolution**.**

adaptive clustering for mobile wireless networks

Formation of Cluster

Clustering is defined as a partitioning a network into several virtual groups (known as clusters) based on certain predefined criteria. The following algorithm which contains steps are used to define cluster formation.

1. Assign a common ID to each node.

2. Each node broadcasts its own ID to its neighbours.

3. A node can be a cluster head if all IDs of nodes, that it can hear are larger than its own.

4. Remaining neighboring nodes become its cluster members.

5. If the node can hear two or more cluster heads becomes a gateway. In a cluster, the number of hops between any two nodes is no more than two. In the whole network, there is no direct connection between cluster heads. After the formation of the clustering architecture, frequent changes of cluster heads will cause the clustering architecture to be unstable and will increase traffic overhead. In addition, the instability of clusters will degrade the performance of protocols based on the clustering architecture. The algorithm enhances by making the cluster head change as little as possible. When a node enters a cluster, it will never challenge the status of the current cluster head in the algorithm. In contrast, in the algorithm, this node will become the cluster head if it has a lower contact probability than the current cluster head. Therefore, in this algorithm, the changes in cluster heads can be greatly reduced.

Fractional Clusters

Because of possible errors in the estimation of contact probabilities and unpredictable sequence of the meetings among mobile nodes, many unexpected small size clusters may be formed. To deal with this problem, a merging process is employed that allows a node to join a “better” cluster, where the node has a higher stability as to be discussed in the next section. The merging process is effective to avoid fractional clusters.

Cluster Member

A node with a very low nodal contact probability may still appear in the member list of another node. The main reason is that a mobile node may change its mobility pattern in real life applications. For example, a student may have his/her regular mobility pattern in a semester (i.e., visiting certain class rooms, libraries, dormitories, cafeterias, etc.). When entering a new semester, his/her mobility may change due to the new/rescheduled classes and after-school activities. Similar scenarios will happen at holidays, summer/winter breaks, or when the student changes his/her major. When mobility pattern changes, but the member list is yet updated, the problem stated above may happen. A possible solution for this problem is to use timeout for membership binding.

Distributed Adaptive Clustering Algorithm

The key part of the algorithm lies on the meeting event between any pair of nodes. A node then decides its actions subsequently. Specifically, a node will enter into a new cluster if it is qualified to be a member. Similarly, when a non-cluster head node moves out its clusters and doesn't enter into any existing cluster, it becomes a new cluster head, forming a new Cluster. When two member nodes meet, they trigger the synchronization process to update their information. During initialization, Node which creates a cluster that consists of itself only and two empty tables. Its cluster ID is set to be its node ID appended with a sequence number. Each node maintains its own sequence number, which increases by one whenever the node creates a new cluster, to avoid duplication. The Algorithm steps are given below.

- If a cluster member leaves from a cluster and come into another cluster, it will not change the status of the existing cluster head even if it has a lower communicate probability than the cluster head.

- Iftwo cluster heads travel within communication range, the node with the higher communicate probability will give up its status as a cluster head.

- If a cluster member becomes separated from any cluster include better contact with other members, it will become a cluster head, and a new cluster is formed.

- A group of nodes which travels out of a cluster will form a new cluster according to the algorithm.

File systems

Goal

– efficient and transparent access to shared files within a mobile environment while maintaining data consistency

Problems

– limited resources of mobile computers (memory, CPU, ...)

– low bandwidth, variable bandwidth, temporary disconnection

– high heterogeneity of hardware and software components (no standard PC architecture)

– wireless network resources and mobile computer are not very reliable

– standard file systems (e.g. NFS) are very inefficient, almost unusable • Solutions

– replication of data (copying, cloning, caching)

– data collection in advance (hoarding, pre-fetching)

File systems - consistency problems

- A central problem of distributed, loosely coupled systems

  – are all views on data the same?

  – how and when should changes be propagated to what users?

- Strong consistency

  – many algorithms offering strong consistency like in database systems (via atomic updates) cannot be used in mobile environments

  – invalidation of data located in caches through a server is very problematic if the mobile computer is currently not connected to the network

- Weak consistency

  – occasional inconsistencies have to be tolerated, but conflict resolution strategies must be applied afterwards to reach consistency again

- Conflict detection

– content independent: version numbering, time-stamps

– content dependent: dependency graphs

File system variables

- Client/Server or Peer-to-Peer relations
- Support in the fixed network and/or mobile computers
- One file system (or namespace) or several file systems
- Transparency

  – hide the mobility support, applications on mobile computers should not notice the mobility

  – user should not notice additional mechanisms needed
- Optimistic or pessimistic consistency model
- Caching and Prefetching

  – bytes, paragraphs, single files, directories, subtrees, partitions, ...

  – permanent or only at certain points in time
- Data management
- Conflict solving

Coda

- Application transparent extensions of client and server

  – changes in the cache manager of a client

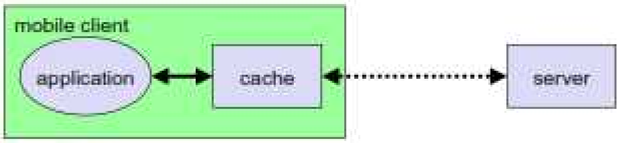
  – applications use cache replicates of files

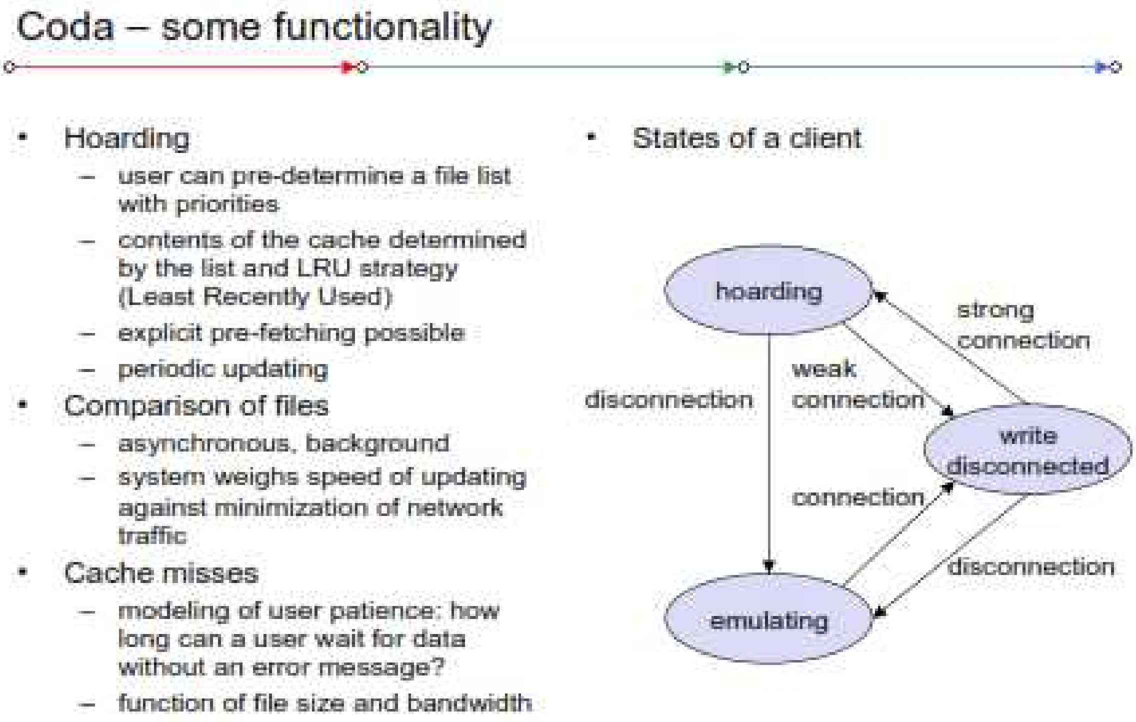
  – extensive, transparent collection of data in advance for possible future use („hoarding“)
- Consistency

– system keeps a record of changes in files and compares files after reconnection

– if different users have changed the same file a manual reintegration of the file into the system is necessary

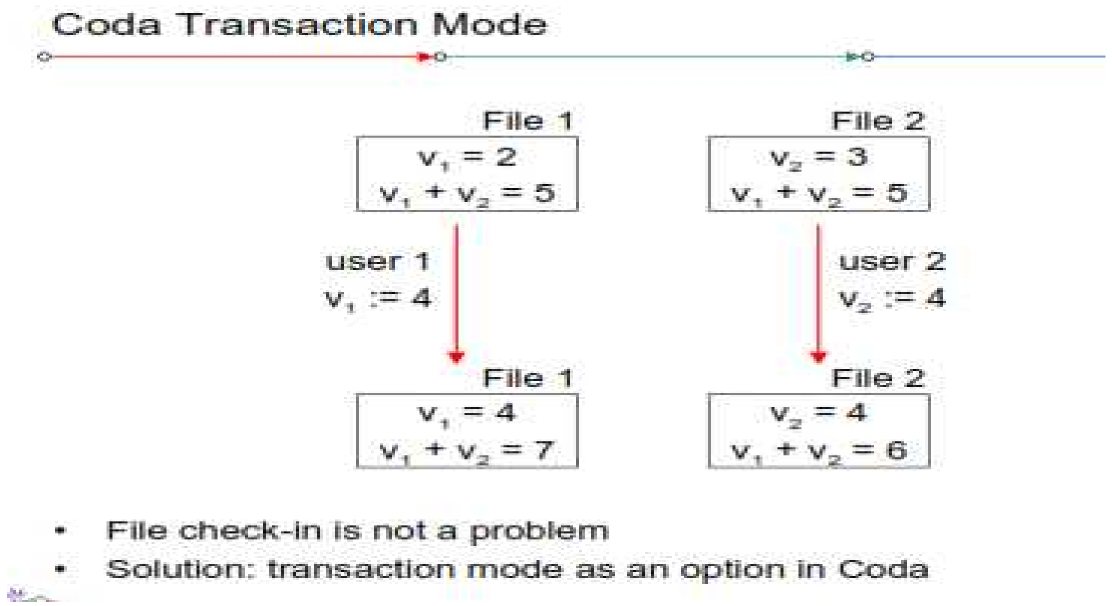
– optimistic approach, coarse-grained (file size)





Coda – some functionality

- Hoarding
  - user can pre-determine a file list with priorities
  - contents of the cache determined by the list and LRU strategy (Least Recently Used)
  - explicit pre-fetching possible
  - periodic updating
- Comparison of files
  - asynchronous, background
  - system weighs speed of updating against minimization of network traffic
- Cache misses
  - modeling of user patience: how long can a user wait for data without an error message?
  - function of file size and bandwidth
- States of a client



Coda Transaction Mode

File 1
$v_1 = 2$
$v_1 + v_2 = 5$

user 1
$v_1 := 4$

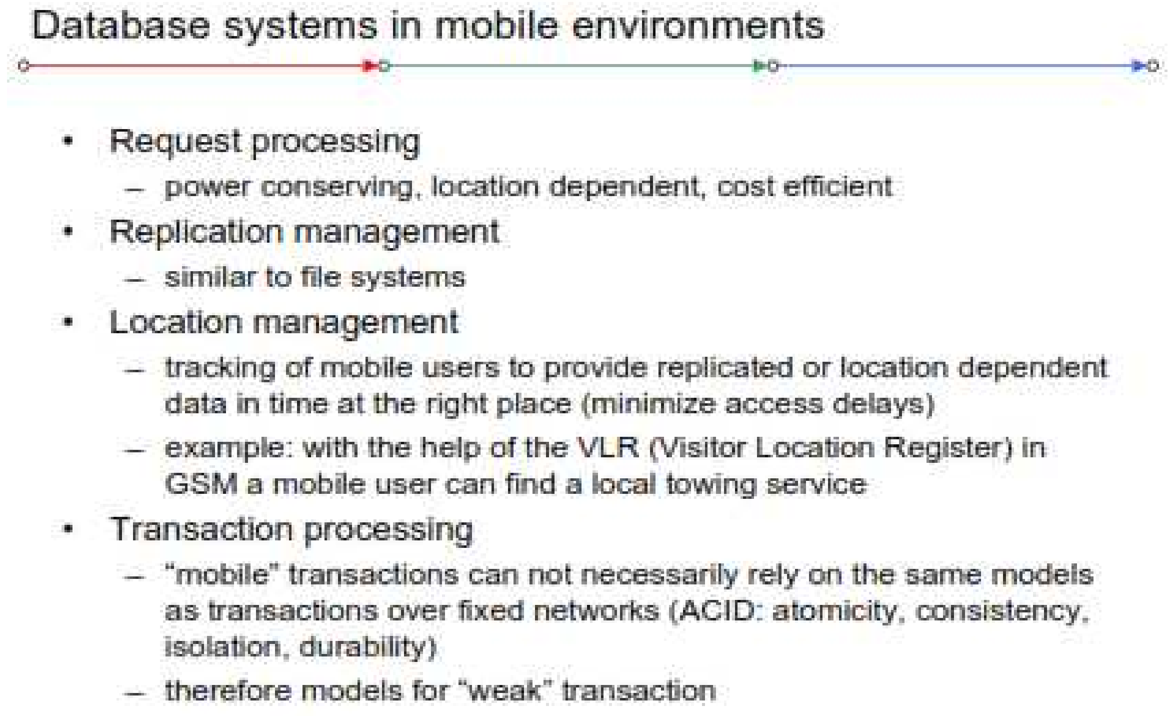
File 1
$v_1 = 4$
$v_1 + v_2 = 7$

File 2
$v_2 = 3$
$v_1 + v_2 = 5$

user 2
$v_2 := 4$

File 2
$v_2 = 4$
$v_1 + v_2 = 6$

- File check-in is not a problem
- Solution: transaction mode as an option in Coda



Database systems in mobile environments

- Request processing
  - power conserving, location dependent, cost efficient
- Replication management
  - similar to file systems
- Location management
  - tracking of mobile users to provide replicated or location dependent data in time at the right place (minimize access delays)
  - example: with the help of the VLR (Visitor Location Register) in GSM a mobile user can find a local towing service
- Transaction processing
  - “mobile” transactions can not necessarily rely on the same models as transactions over fixed networks (ACID: atomicity, consistency, isolation, durability)
  - therefore models for “weak” transaction

File systems – more examples

• Ficus

– not a client/server approach

– use of gossip“ protocols: a mobile computer does not necessarily need to have direct connection to a server, with the help of other mobile computers updates can be propagated through the network

– optimistic approach based on replicates

– detection of write conflicts, conflict resolution on directory level

- **MIo-NFS (Mobile Integration of NFS)**

  – NFS extension

  – pessimistic approach: only token holder can write

  – Three modes: connected, loosely connected, disconnected

Database systems in mobile environments

- Request processing

  – power conserving, location dependent, cost efficient

- Replication management

  – similar to file systems

- Location management

  – tracking of mobile users to provide replicated or location dependent data in time at the right place (minimize access delays)

  – example: with the help of the VLR (Visitor Location Register) in GSM a mobile user can find a local towing service

- Transaction processing

  – “mobile” transactions can not necessarily rely on the same models as transactions over fixed networks (ACID: atomicity, consistency, isolation, durability)

  – therefore models for “weak” transaction

Disconnected Operation

- no connection to home file server
  - users optimistically *hoard* replicas of desired files prior to disconnection
  - all file operations processed in the cache
- read misses are fatal

- updates to file system are *logged* at the client
- upon reconnection, replay of logged events reintegrates changes with home file system

Weakly Connected Operation

- a low-bandwidth connection is available
- read misses are no longer fatal
- asynchronous write backs provide for reintegration of logged changes with home file system, but must share the bandwidth available with reads
- reads should have priority

Strongly Connected Operation

- a high-bandwidth connection is available, over which read and write operations are serviced
- file caching can improve performance (by reducing latency)
- the conventional distributed file system

Unit 4

1 INTRODUCTION(Mobile Agent)

Mobile agents are autonomous programs that can travel from computer to computer in a net- work, at times and to places of their own choosing. The state of the running program is saved, by being transmitted to the destination. The program is resumed at the destination continuing its processing with the saved state. They can provide a convenient, efficient, and robust framework for implementing distributed applications and smart environments for several reasons, including improvements to the latency and bandwidth of client-server applications and reducing vulnerability to network disconnection. In fact, mobile agents have several advantages in the development of various services in smart environments in addition to distributed applications.

- **Reduced communication costs:** Distributed computing needs interactions between different computers through a network. The latency and network traffic of interactions often seriously affect the quality and coordination of two programs running on different computers. As we can see from Figure 1, if one of the programs is a mobile agent, it can migrate to the computer the other is running on communicate with it locally. That is, mobile agent technology enables remote communications to operate as local communications.

- **Asynchronous execution** After migrating to the destination-side computer, a mobile agent does not have to interact with its source-side computer. Therefore, even when the source can be shut down or the network between the destination and source can be disconnected, the agent can continue processing at the destination. This is useful within unstable communications, including wireless communication, in smart environments.

- **Direct manipulation** A mobile agent is locally executed on the computer it is visiting. It can directly access and control the equipment for the computer as long as the computer allows it to do so. This is helpful in network management, in particular in detecting and removing device failures. Installing a mobile agent close to a real-time system may prevent delays caused by network congestion.

- **Dynamic-deployment of software** Mobile agents are useful as a mechanism for the de- ployment of software, because they can decide their destinations and their code and data can be dynamically deployed there, only while they are needed. This is useful in smart en- vironments, because they consist of computers whose computational resources are limited.

- **Easy-development of distributed applications** Most distributed applications consist of at least two programs, i.e., a client-side program and a server side program and often spare codes for communications, including exceptional handling. However, since a mobile agent itself can carry its information to another computer, we can only write a single program to define distributed computing. A mobile agent program does not have to define communications with other computers. Therefore, we can easily modify standalone programs as mobile agent programs.

As we can see from Figure 2, mobile agents can save themselves through persistent storage, duplicate themselves, and migrate themselves to other computers under their own control so that they can support various types of processing in distributed systems.

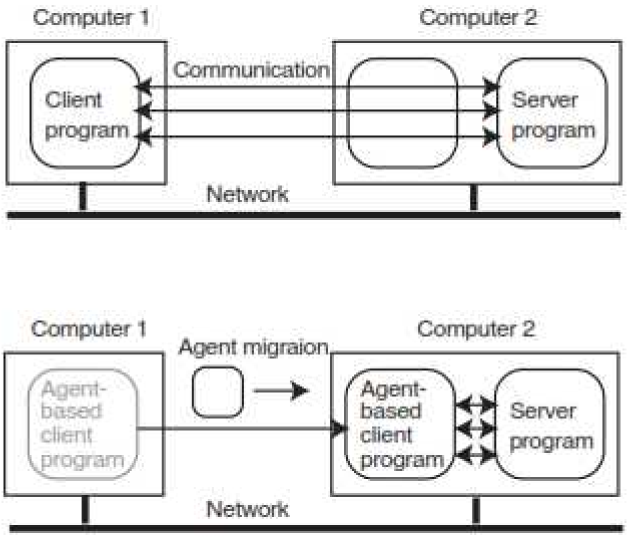


Figure 1: Reduced communication

Although not all applications for distributed systems will need mobile agents, there are many other applications that will find mobile agents the most effective technique for implementing all or part of their tasks. Mobile agent technology can be treated as a type of software agent technology, but it is not always required to offer intelligent capabilities, e.g., reactive, pro-active, and social behaviors that are features of existing software agent technologies.   This is because these capabilities tend to be large in terms of scale and processing, and no mobile agent should consume

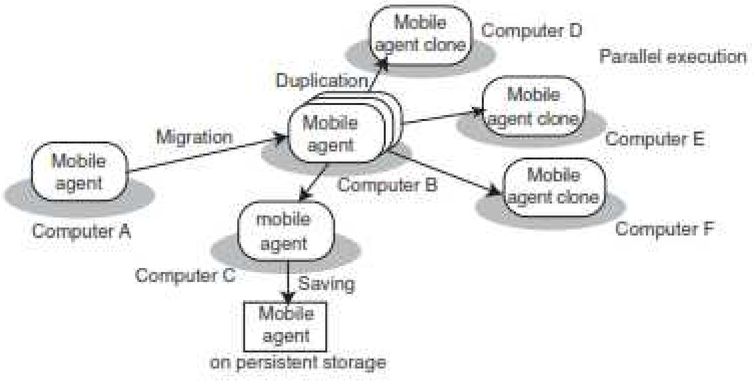


Figure 2: Functions of mobile agents in distributed system

excessive computational resources, such as processors, memory, files, and networks, at its destina- tions. Also, the technology is just an implementation approach of distributed systems rather than intelligent systems.

1.1 Mobility and Distribution

Fuggetta provided a description of mobile software paradigms for distributed applications. These are classified as client/server (CS), remote evaluation (REV), code on demand (COD), and mobile agent (MA) approaches. By decompiling distributed applications into code, data, and execution, most distributed executions can be modeled as primitives of these approaches .

- The client-server approach is widely used in traditional and modern distributed systems (Figure 3 a)). The code, data, and execution remain fixed at computer A. Computer B requests a service from the server with some data arguments of the request. The code and remaining data to provide the service are resident within computer B. As a response, computer B provide the service requested by accessing computational resources provided in it. Computer B returns the results of the execution to computer A.

- The remote evaluation approach assumes that the code to perform the execution is stored at computer A (Figure 3 b)). Both the code and data are sent to computer B. As a response, computer B executes the code and data by accessing computational resources, including data, provided in them. An additional interaction returns the results from computer B to computer A.

- The codeon-demand approach is an inversion of the remote evaluation approach. The code and data are stored at computer A and execution is done at computer B. Computer A fetches code and data from computer B and then executes the code with its local data as well as the imported data. An example of this is Java applets, which are Java codes that web-browsers download from remote HTTP servers to execute locally.

- The mobile agent approach assume that the code and data are initially hosted by computer
A (Figure 3 d)). Computer A migrates the data and code it need to computer B. After it has moved to computer B, the code is executed with the data and the resources available on computer B.

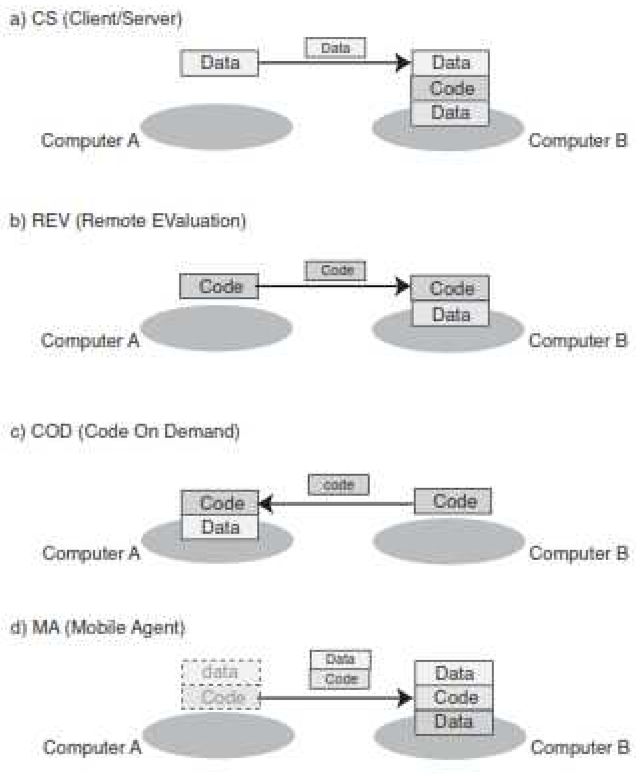


Figure 3: Client/server, remote evaluation, code on demand, and mobile agent

2 MOBILE AGENT PLATFORM

Mobile agent platforms consist of two parts: mobile agents and runtime systems. The former defines the behavior of software agents. The latter are called agent platforms, agent systems, and agent servers, and support their execution and migration. The same architecture exists on all computers at which agents are reachable. That is, each mobile agent runs within a runtime systems on its current computer. When an agent requests the current runtime system to migrate itself, the runtime system can migrate the agent to a runtime system on the destination computer, carrying its state and code with it. Each runtime system itself runs on top of the operating system as a middleware. It provides interpreters or virtual machines for executing agent programs, or the system themselves are provided on top of virtual machines, e.g., the Java virtual machine (JVM).

2.1 Remote procedure call

Agent migration is similar to RPC (Remote Procedure Calling) or RMI (Remote Method Invoca- tion). RPC enables a client program to call a procedure for server programs running in separate processes, generally in different computers from the client. RMI is an extension of local method invocation that allows an object to invoke the methods of the object on a remote com- puter. RPC or RMI can pass arguments to a procedure or method of a program on the server and receives a return value from the server. The mechanism for passing arguments and results between two computers through RPC or RMI correspond to that for agent migration between two computers. Figure 4 shows how for the basic mechanism of RPC between two computers.

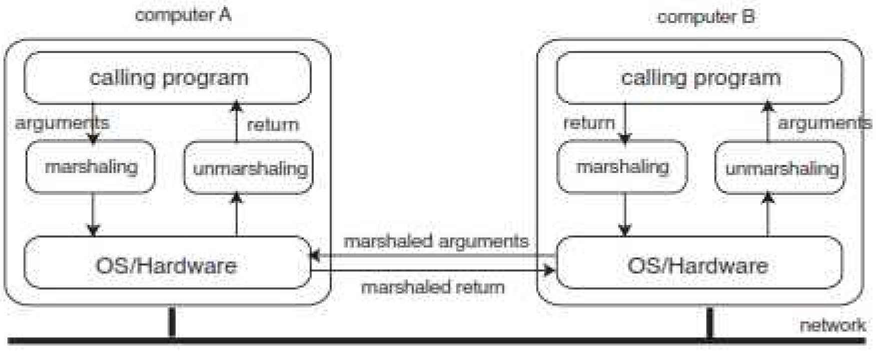


Figure 4: Remote procedure call between two computers

2.1.2 Agent migration

Figure 6 shows the basic mechanism for agent migration between two computers.

Step.1 The runtime system on the sender-side computer suspends the execution of the agent.

Step.2 It marshals the agent into a bit-chunk that can be transmitted over a network.

Step.3 It transmits the chunk to the destination computer through the underlying network protocol.

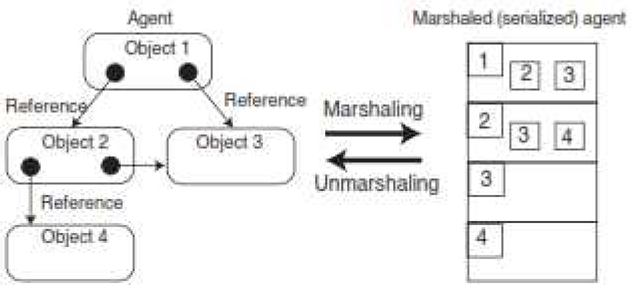


Figure 5: Marshaling agent

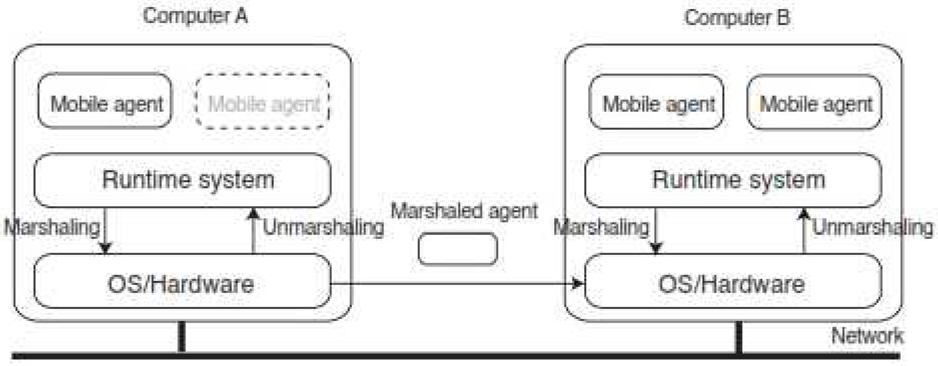


Figure 6: Agent migration between two computers

Step.4 The runtime system on the receiver-side computer receives the chunk.

Step.5 It unmarshals the chunk into the agent and resumes the agent.

Most existing mobile agent systems use TCP channels, SMTP, or HTTP as their underlying com- munication protocols. Mobile agents themselves are separated from the underlying communica- tion protocols.

2.1.3 Strong migration vs. weak migration

The state of execution is migrated with the code so that computation can be resumed at the des- tination. According to the amount of detail captured in the state, we can classify agent migration into two types: strong and weak.

• **Strong migration:** is the ability of an agent to migrate over a network, carrying the code and execution state, where the state includes the program counter, saved processor registers, and local variables, which correspond to variables allocated in the stack frame of the agent's memory space, global variables. These correspond to variables allocated in the heap frame. The agent is suspended, marshaled, transmitted, unmarshaled and then restarted at the exact position where it was previously suspended on the destination node without loss of data or execution state.

• **Weak migration:** is the ability of an agent to migrate over a network, carrying the code and partial execution state, where the state is variables in the heap frame, e.g., instance variables in object oriented programs, instead of its program counter and local variables declared in methods or functions. The agent is moved to and restarted on the destination with its global variables. The runtime system may explicitly invoke specified agent methods.

Strong migration can cover weak migration, but it is a minority. This is because the execution state of an agent tends to be large and the marshaling and transmitting of the state over a network need heavy processing. Moreover, like the latter, the former cannot migrate agents that access the computational resources only available in current computers, e.g., input-and-output equipment and networks. The former unfortunately has no significant advantages in the development and operation of real distributed applications.

The program code for an agent needs to be available at the destination where the agent is running. The code must to be deployed at the source at the time of creation and at the destination to which it moves. Therefore, existing runtime systems offer a facility for statically deploying program code that is needed to execute the agent, for loading the program code on demand, or for transferring the program code along with the agent.

2.3 Agent execution management

The runtime system manages execution and monitoring of all agents on a computer. It allows several hundred agents to be present at any one time on a computer. It also provide these agents with an execution environment and executes them independently of one another. It manages the life-cycle of its agents, e.g., creation, termination, and migration.

Each agent program can access basic functions provided by its runtime system by invoking APIs (Table 1). The agent uses the go command to migrate from one computer to another with the destination system address (and its target agent''s identifier) and does not need to concern itself with any other details of migration. Instead, the runtime system supports the migration of the agent. It stops the agent's execution and then marshals the agent's data items to the destination via the underlying communication protocol, e.g., TCP channel, HTTP (hyper text transfer protocol), and SMTP (simple mail transfer protocol). The agent is unpacked and reconstituted on the destination.

2.4 Inter-agent communication

Mobile agents can interact with other agents residing within the same computer or with agents on remote computers as other multi-agents. Existing mobile agent systems provide various inter- agent communication mechanisms, e.g., method invocation, publish/subscribe-based event pass- ing, and stream-based communications.

2.5 Locating mobile agents

Since mobile agents can autonomously travel from computer to computer, a mechanism for track- ing the location of agents is needed by the users to control their agents and for agents to commu- nicate with other agents. Several mobile agent systems provide such mechanisms, which can be classified into three schemes:

- A name server multicasts query messages about the location of an agent the to computers and receives a reply message from a computer hosting the agent (Figure 7 (a)).

- An agent registers its current loation at a predefined name server whenever it arrives at another computer (Figure 7 (b)).

- An agent leaves a footprint specifying its destination at its current computer whenever it migrates to another computer to track the trails of the agent (Figure 7 (c)).

In many cases, locating agents is application specific. For example, the first scheme is suitable for an agent moving within a local region. It is not suitable for agents visiting distant nodes. The second scheme is suitable for an agent migrating within a far away region; in the case of a large number of nodes, registering nodes are organized hierarchically. However, it is not suitable for a large number of migrations. The third scheme is suitable for a small number of migrations; it is not appropriate for long chains.

2.6 Security

Security is one of the most important issues with mobile agent systems. Most security issues in mobile agents are common to existing computer security problems in communication and the downloading of software. There are two problems in mobile agent security: the protection of hosts from malicious mobile agents and the protection of mobile agents from malicious hosts. It is difficult to verify with complete certainty whether an incoming agent is malicious or not. However, there are two solutions to protecting hosts from malicious mobile agents. The first is to provide access-control mechanisms, e.g., Java's security manager. They explicitly specify the permission of agents and restrict any agent behaviors that are beyond their permissions. The sec- ond is to provide authentication mechanisms by using digital signatures or authentication systems. They explicitly permit runtime systems to only receive agents that have been authenticated, have been sent from authenticated computers, or that have originated from authenticated computers.

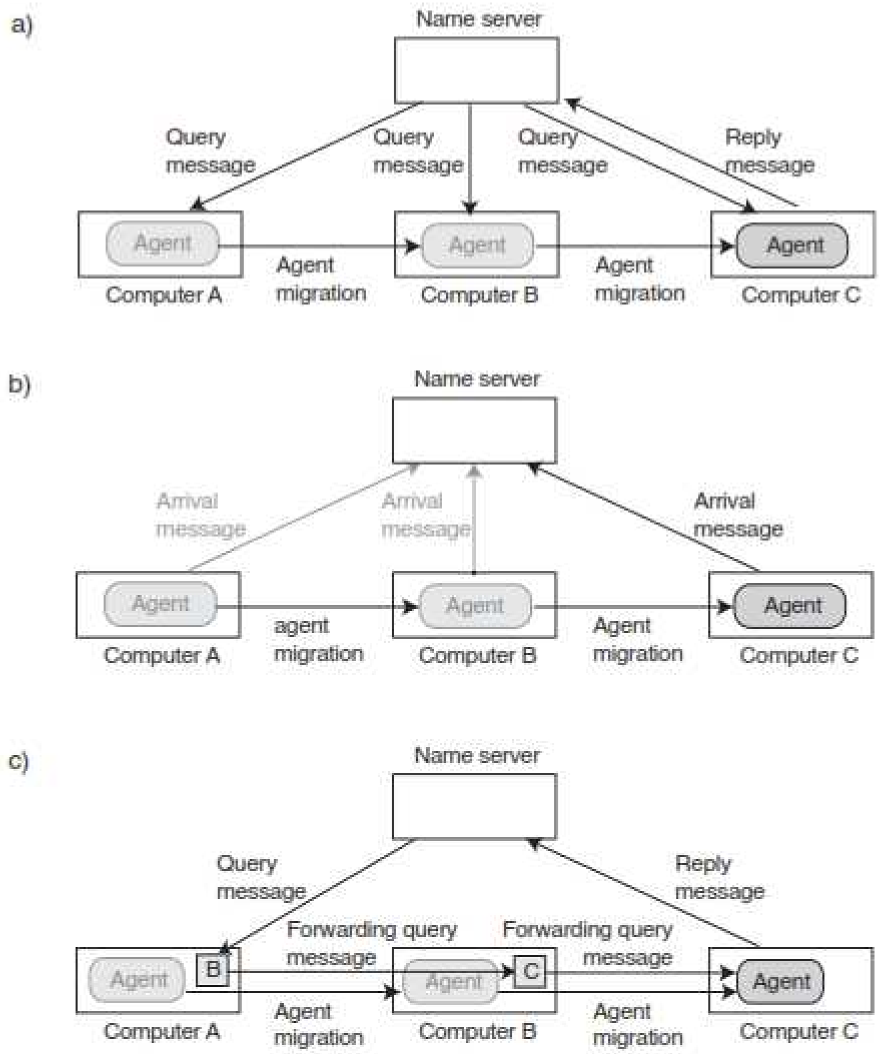


Figure 7: Discovery for migrating agents

before they migrate the agents to these destinations. While strong security features would not immediately make mobile agents appealing, the absence of security would certainly make mobile agents unattractive and unpractical.

3 MOBILE AGENT APPLICATIONS

Many researchers have stated that there are no killer applications for mobile agent technology, because almost everything you can do with MAs can be done with more traditional technologies. However, mobile agents make it easier, faster, and more effective to develop, manage, and execute distributed applications than other technologies. We describe typical applications of mobile agents as follows:

3.1 Remote information retrieval

This is one of the most traditional applications of mobile agents. If all information were stored in relational databases, a client could send a message containing SQL commands to database servers. However, given that most of the world's data is in fact maintained in free text files on different computers, remote searching and filtering require the ability to open, read, and filter files. Since mobile agents can perform most of their tasks locally at the destination, Client can send its agents to database servers so that they locally perform a sequence of query or update tasks on the servers. Communications between the client and server can be minimized, i.e., the migration of a search agent to the server and the migration of an agent to the client. Since agents contain program codes for filtering information that is of interest to their users from databases, they only need to carry wanted information back to the client to reduce communication traffic. Furthermore, agents can migrate among multiple database servers to retrieve and gather the interesting data from the servers. They can also determine the destinations based on information they have acquired from the database servers that they have thus far visited.

3.2 Network management

Mobile agent technology provides a solution to the fiexible management of network systems. Mobile agents can locally observe and control equipment at each node by migrating among nodes. Mobile agent-based network management has several advantages in comparison with traditional approaches, such as the client/server one.

- As code is very often smaller than the data it processes, the transmission of mobile agents to sources of data creates less traffic than transferring the data itself. Deploying a mobile agent close to the network nodes that we want to monitor and control prevents delays caused by network congestion.

- Since a mobile agent is locally executed on the node it is visiting, it can easily access the functions of devices on this node.

- The dynamic deployment and configuration of new or existing functionalities into a network system are extremely important tasks, especially as they potentially allow outdated systems to be updated in an efficient manner.

- Network management systems must often handle networks that may have various malfunctions and disconnections and whose exact topology may not be known. Since mobile agents are autonomous entities, they may be able to detect proper destinations or routings on such networks.

Adopting mobile agent technology eliminates the need for administrators to constantly monitor many network management activities, e.g., the installation and upgrading of software and periodic network auditing. There have been several attempts to apply this technology to network manage- ment tasks. Karmouch presented typical mobile agent approaches to network management. Satoh proposed a framework for building and operating agent itineraries for network management systems and constructed domain-specific languages for describing agent migration for network management.

3.3 Mobile computing

Mobile agents use the capabilities and resources of remote servers to process their tasks. When a user wants to do tasks beyond the capabilities of his or her computers, the agents that perform the tasks can migrate to and be executed at a remote server. Mobile agents can also mask temporal disconnections in networks. Mobile computers are not always connected to networks, because their wired networks are disconnected before they are moved to other locations or wireless net- works become unstable or non-available due to deteriorating radio conditions or are not uncovered by the area at all. A stable connection is only requested at the beginning to send the agent, and to take the agent back at the end of the task, but this is not requested during the execution of the whole application execution. Several researchers have explored mechanisms for migrating agents through unstable networks. When a mobile agent requests a runtime system to migrate itself, the system tries to transmit the moving agent to the destination. If the destination cannot be reached, the system automatically stores the moving agent in a queue and then periodically tries to transmit the waiting agent to either the destination or another runtime system on a reachable intermediate node as close to the destination as possible.

3.4 Software testing

Mobile agents are useful in the development of software as well as the operation of software in distributed and mobile computing settings. An example of these applications is testing methodology for software running on mobile computers, called *Flying Emulator*. Wireless LANs or 4G-networks incorporate wireless LAN technologies, and mobile terminals can access the services provided by LANs, as well as global network services. Therefore, software running on mobile terminals may depend on not only its application-logic but also on services within the LANs that the terminals are connected to.

3.5 Active networking

There are two approaches to implementing active networks. The active packet approach replaces destination addresses in the packets of existing architectures with miniature programs that are interpreted at nodes on arrival. The active node approach enables new protocols to be dynamically deployed at intermediate and end nodes using mobile code techniques. Mobile agents are very similar to active networks, because a mobile agent can be regarded as a specific type of active packet, and an agent platform in traditional networks can be regarded as a specific type of active node. There have been a few attempts to incorporate mobile agent technology with active network technology. Of these, the MobileSpaces system provides a mobile agent-based framework for integrating the both approaches. The frame- work enables us to implement network processing of mobile agents with mobile agent-based components, where the components are still mobile agents so that they can be dynamically deployed at computers to customize network processing.

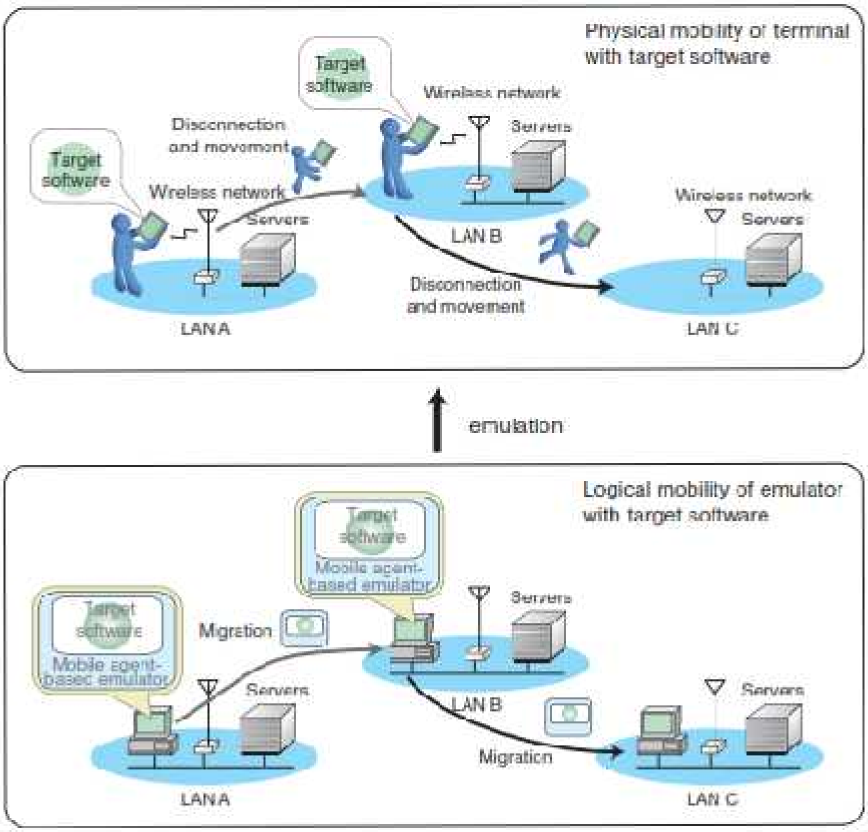


Figure 8: Corelation between the movement of target mobile computer and migration of mobile agent-based emulator

Unit 5

Ad Hoc Networks

- Setting up of fixed access points and backbone infrastructure is not always viable
  - Infrastructure may not be present in a disaster area or war zone
  - Infrastructure may not be practical for short-range radios; Bluetooth (range ~ 10m)
- Ad hoc networks:
  - Do not need backbone infrastructure support
  - Are easy to deploy
  - Useful when infrastructure is absent, destroyed or impractical

Applications

- Personal area networking
  - cell phone, laptop, ear phone, wrist watch
- Military environments
  - soldiers, tanks, planes
- Civilian environments
  - taxi cab network
  - meeting rooms
  - sports stadiums
  - boats, small aircraft
- Emergency operations

- search-and-rescue
- policing and fire fighting

Routing Protocols

- Proactive protocols
  - Traditional distributed shortest-path protocols
  - Maintain routes between every host pair at all times
  - Based on periodic updates; High routing overhead
  - Example: DSDV (destination sequenced distance vector)
- Reactive protocols
  - Determine route if and when needed
  - Source initiates route discovery
  - Example: DSR (dynamic source routing)
- Hybrid protocols
  - Adaptive; Combination of proactive and reactive
  - Example : ZRP (zone routing protocol)

Reactive Routing Protocols

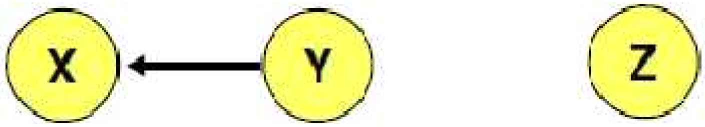
Destination sequence distance vector (DSDV):-

Destination sequence distance vector (DSDV) routing is an example of proactive algorithms and an enhancement to distance vector routing for ad-hoc networks. Distance vector routing is used as routing information protocol (RIP) in wired networks. It performs extremely poorly with certain network changes due to the count-to-infinity problem. Each node exchanges its neighbor table periodically with its neighbors. Changes at one node in the network propagate slowly through the network. The strategies to avoid this problem which are used in fixed networks do not help in the case of wireless ad-hoc networks, due to the rapidly changing topology. This might create loops or unreachable regions within the network.

DSDV adds the concept of sequence numbers to the distance vector algorithm. Each routing advertisement comes with a sequence number. Within ad-hoc networks, advertisements may propagate along many paths. Sequence numbers help to apply the advertisements in correct order. This avoids the loops that are likely with the unchanged distance vector algorithm.

Each node maintains a routing table which stores next hop, cost metric towards each destination and a sequence number that is created by the destination itself. Each node periodically forwards routing table to neighbors. Each node increments and appends its sequence number when sending its local routing table. Each route is tagged with a sequence number; routes with greater sequence numbers are preferred. Each node advertises a monotonically increasing even sequence number for itself. When a node decides that a route is broken, it increments the sequence number of the route and advertises it with infinite metric. Destination advertises new sequence number.

When X receives information from Y about a route to Z,



Let destination sequence number for Z at X be S(X), S(Y) is sent from Y

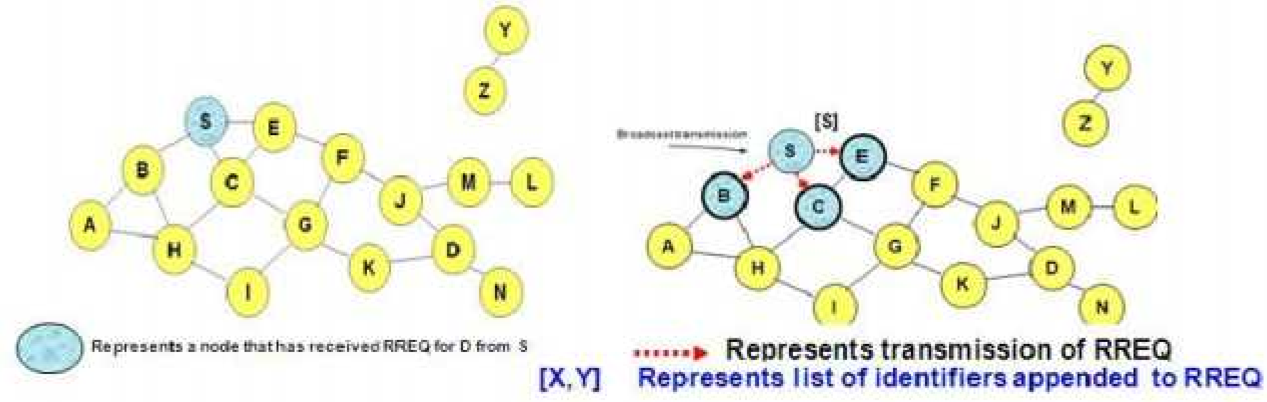
- If S(X) > S(Y), then X ignores the routing information received from Y
- If S(X) = S(Y), and cost of going through Y is smaller than the route known to X, then X sets Y as the next hop to Z

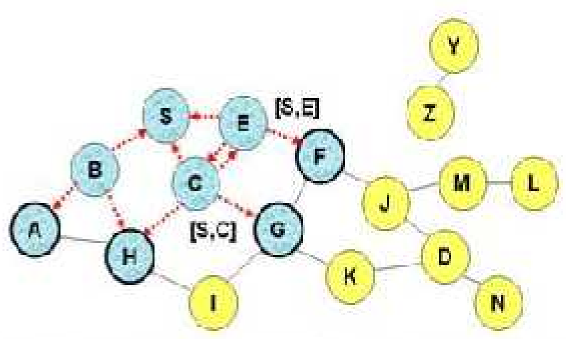
- If S(X) < S(Y), then X sets Y as the next hop to Z, and S(X) is updated to equal S(Y)

Besides being loop-free at all times, DSDV has low memory requirements and a quick convergence via triggered updates. Disadvantages of DSDV are, large routing overhead, usage of only bidirectional links and suffers from count to infinity problem.

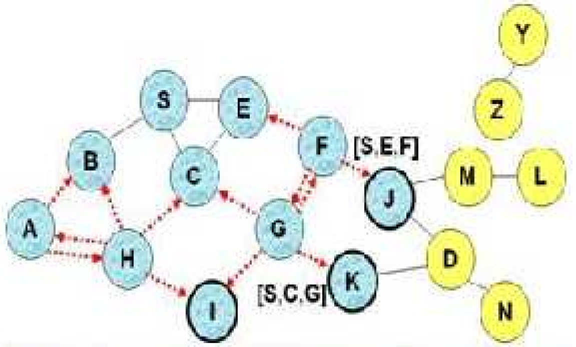
Dynamic Source Routing (DSR)

- When node S wants to send a packet to node D, but does not know a route to D, node S initiates a route discovery
- Source node S floods Route Request (RREQ)
- Each node *appends own identifier* when forwarding RREQ

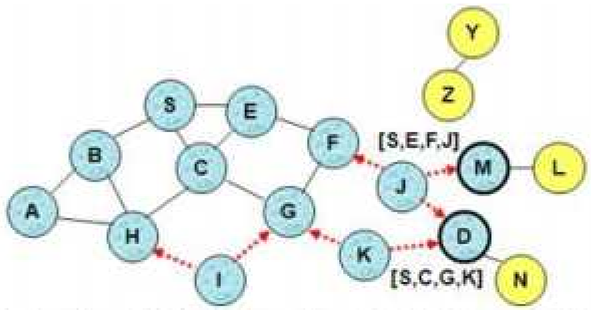

Y
Z
S
E
B
C
F
J
M
L
A
H
G
D
K
I
N
Represents a node that has received RREQ for D from S
Broadcast transmission
[S]
Represents transmission of RREQ
[X,Y]
Represents list of identifiers appended to RREQ



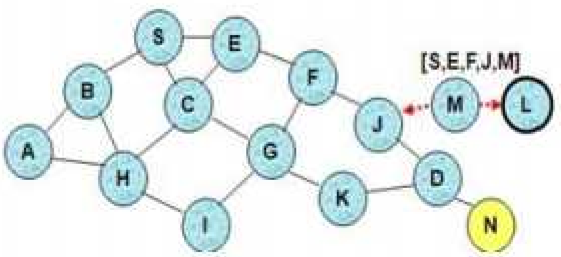
Node H receives packet RREQ from two neighbors: potential for collision



Node C receives RREQ from G and H, but does not forward it again, because node C has already forwarded RREQ once



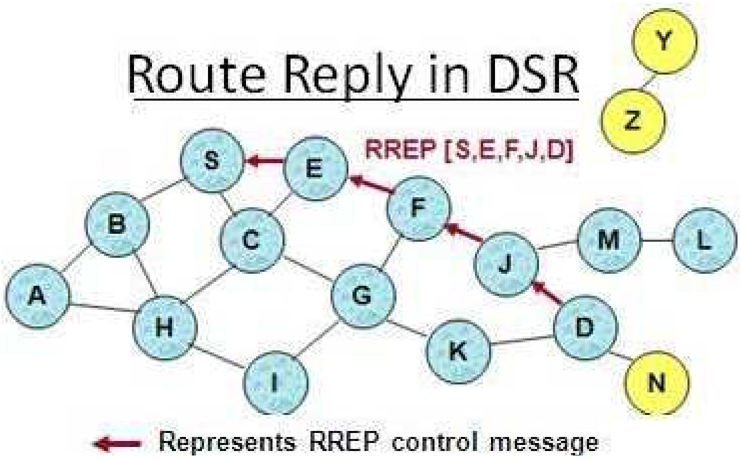
- Nodes J and K both broadcast RREQ to node D
- Since nodes J and K are hidden from each other, their transmissions may collide



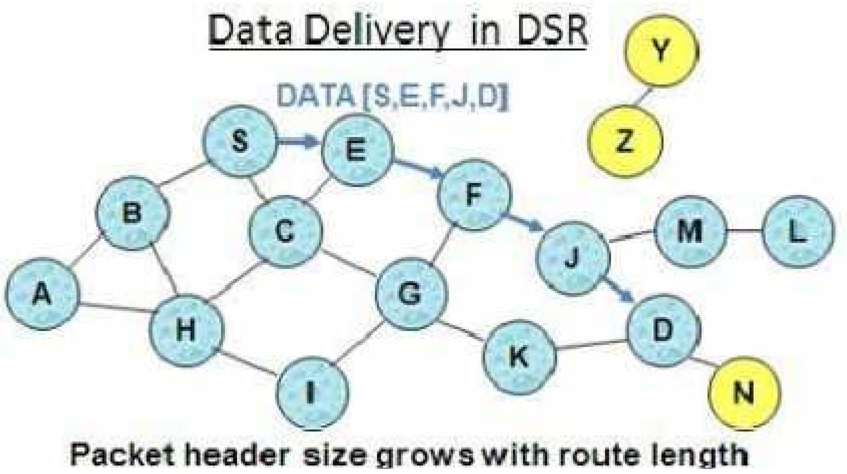
Node D does not forward RREQ, because node D is the intended target of the route discovery

Route Discovery in DSR

- Destination D on receiving the first RREQ, sends a Route Reply (RREP)
- RREP is sent on a route obtained by reversing the route appended to received RREQ
- RREP includes the route from S to D on which RREQ was received by node D



- Node S on receiving RREP, caches the route included in the RREP
- When node S sends a data packet to D, the entire route is included in the packet header
  - hence the name source routing
  - Intermediate nodes use the source route included in a packet to determine to whom a packet should be forwarded



DSR Optimization: Route Caching

- Each node caches a new route it learns by *any means*

- When node S finds route [S,E,F,J,D] to node D, node S also learns route [S,E,F] to node F
- When node K receives Route Request [S,C,G] destined for node, node K learns route [K,G,C,S] to node S
- When node F forwards Route Reply RREP [S,E,F,J,D], node F learns route [F,J,D] to node D
- When node E forwards Data [S,E,F,J,D] it learns route [E,F,J,D] to node D
- A node may also learn a route when it overhears Data
- Problem: Stale caches may increase overheads

Dynamic Source Routing: Advantages

- Routes maintained only between nodes who need to communicate

  reduces overhead of route maintenance

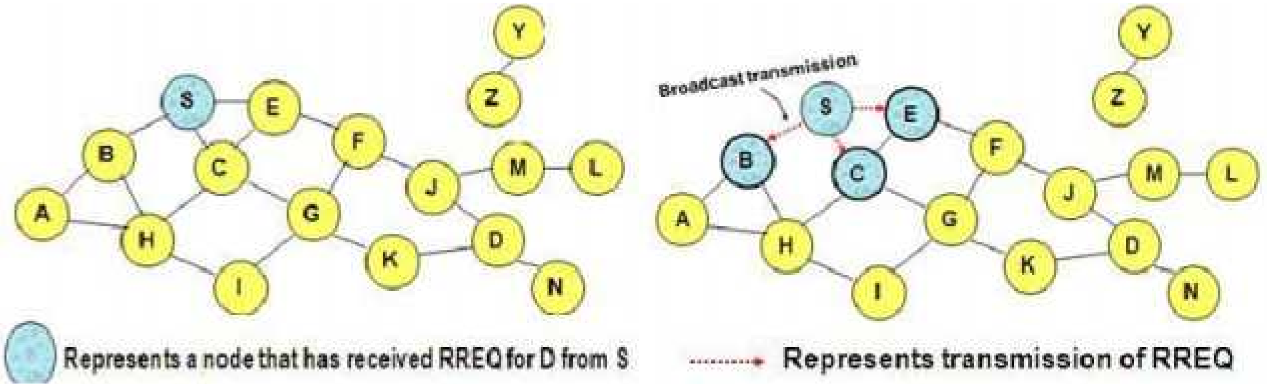
- Route caching can further reduce route discovery overhead
- A single route discovery may yield many routes to the destination, due to intermediate nodes replying from local caches

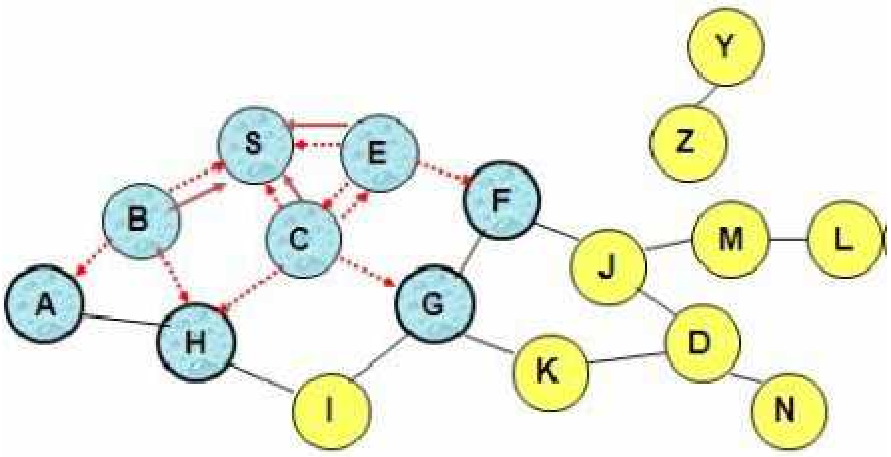
Dynamic Source Routing: Disadvantages

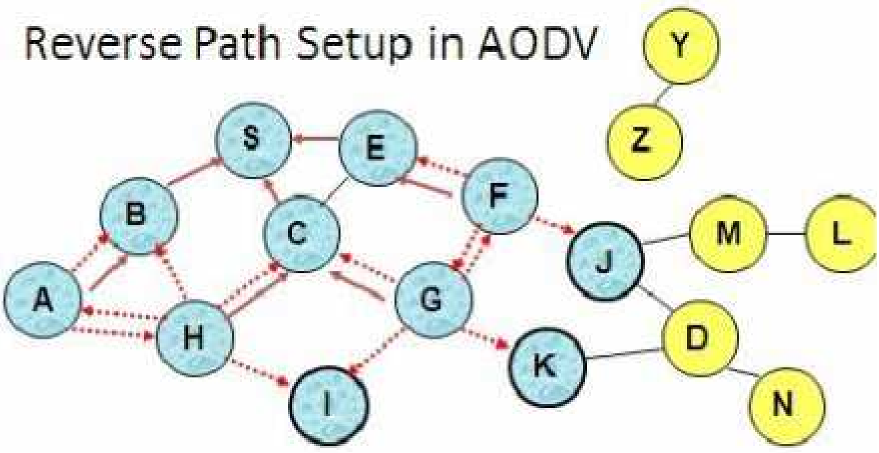
- Packet header size grows with route length due to source routing
- Flood of route requests may potentially reach all nodes in the network
- Potential collisions between route requests propagated by neighboring nodes
  - insertion of random delays before forwarding RREQ
  - Increased contention if too many route replies come back due to nodes replying using their local cache
  - Route Reply *Storm* problem
- Stale caches will lead to increased overhead

Ad Hoc On-Demand Distance Vector Routing (AODV)

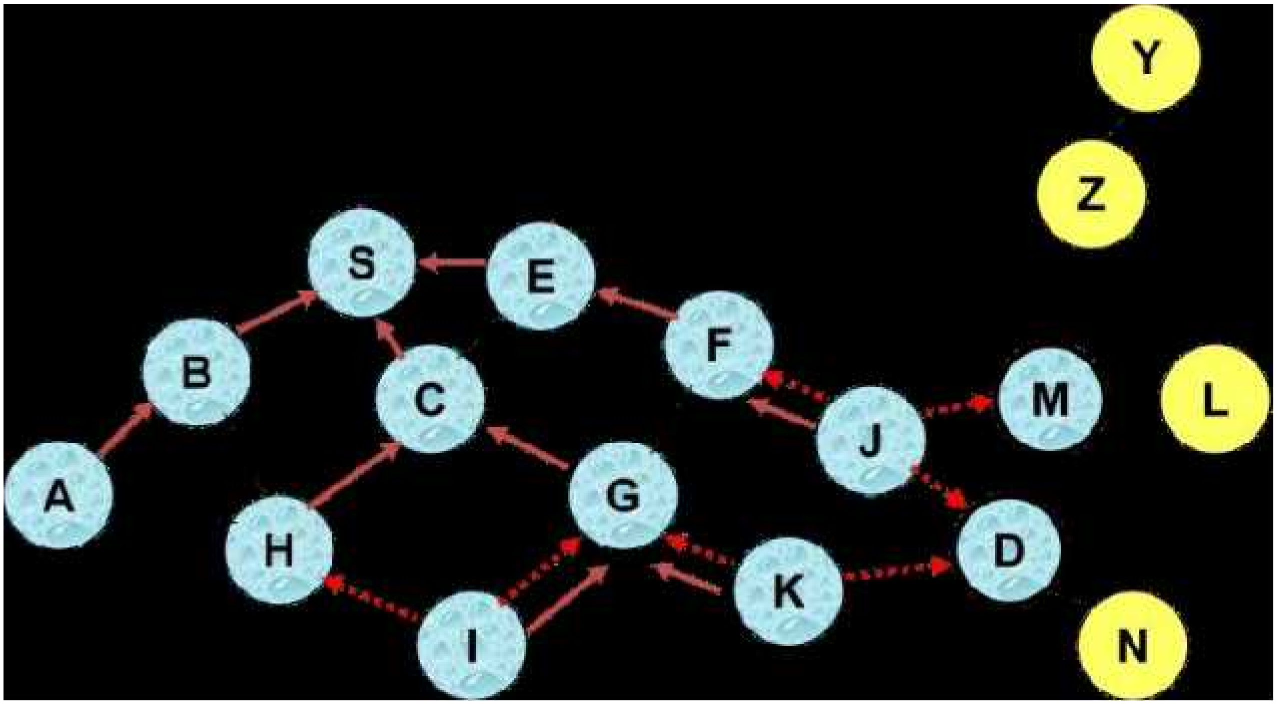
- DSR includes source routes in packet headers
- Resulting large headers can sometimes degrade performance
  - particularly when data contents of a packet are small
  - AODV attempts to improve on DSR by maintaining routing tables at the nodes, so that data packets do not have to contain routes
- AODV retains the desirable feature of DSR that routes are maintained only between nodes which need to communicate
- Route Requests (RREQ) are forwarded in a manner similar to DSR
- When a node re-broadcasts a Route Request, it sets up a reverse path pointing towards the source
  - AODV assumes symmetric (bi-directional) links
- When the intended destination receives a Route Request, it replies by sending a Route Reply (RREP)
- Route Reply travels along the reverse path set-up when Route Request is forwarded

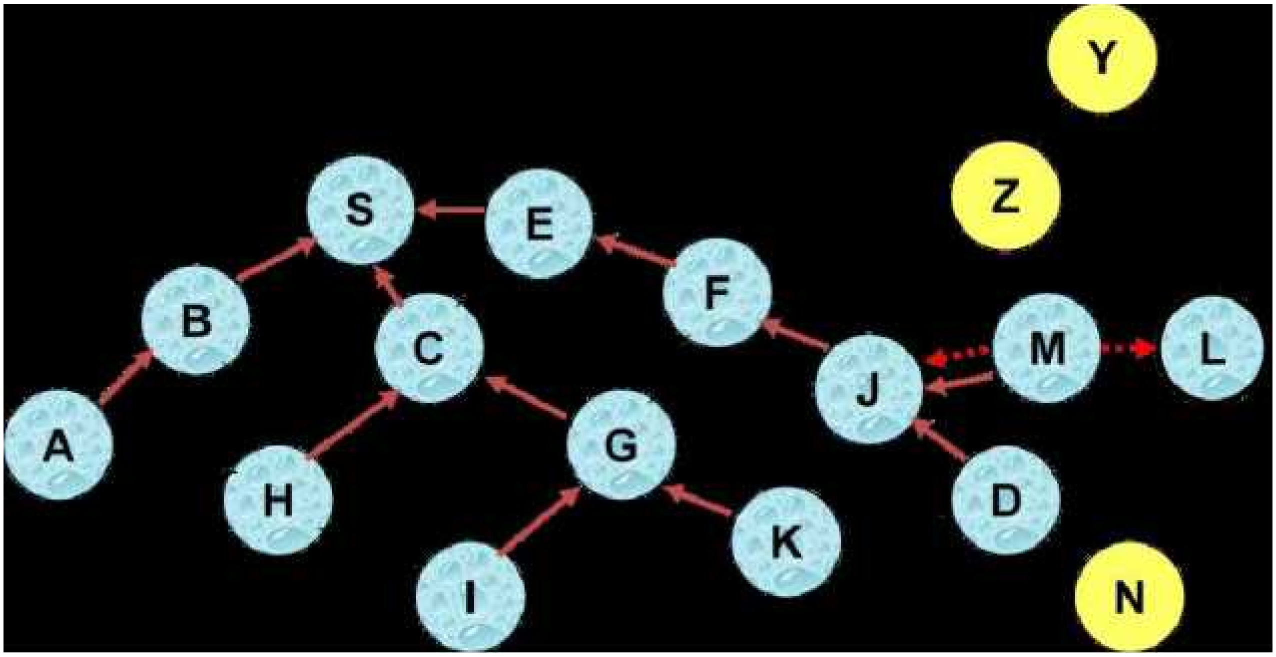

Y
S
E
Z
B
C
F
M
L
A
G
J
H
K
D
I
N
Represents a node that has received RREQ for D from S
Broadcast transmission
Y
S
E
Z
B
C
F
M
L
A
G
J
H
K
D
I
N
Represents transmission of RREQ

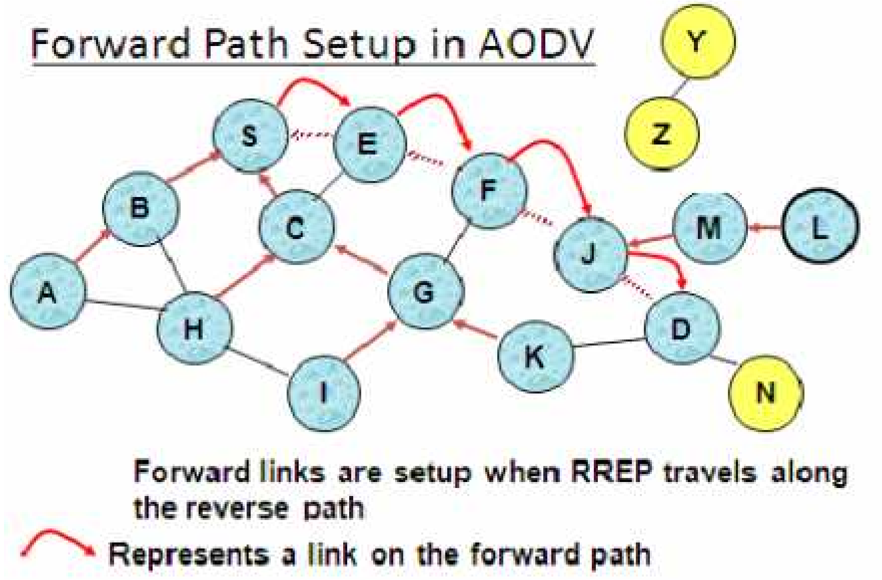



Reverse Path Setup in AODV

Node C receives RREQ from G and H, but does not forward it again, because node C has already forwarded RREQ once


Y
Z
S
E
F
B
C
M
L
J
A
G
H
D
K
I
N


Y
Z
S
E
F
B
C
M
L
J
A
G
H
D
K
I
N


Forward Path Setup in AODV
Y
Z
S
E
F
B
C
M
L
J
A
G
H
D
K
I
N
Forward links are setup when RREP travels along the reverse path
Represents a link on the forward path

Route Request and Route Reply

- Route Request (RREQ) includes the last known sequence number for the destination
- An intermediate node may also send a Route Reply (RREP) provided that it knows a more recent path than the one previously known to sender
- Intermediate nodes that forward the RREP, also record the next hop to destination
- A routing table entry maintaining a reverse path is purged after a timeout interval
- A routing table entry maintaining a forward path is purged if *not used* for a *active_route_timeout* interval

Link Failure

- A neighbor of node X is considered active for a routing table entry if the neighbor sent a packet within *active_route_timeout* interval which was forwarded using that entry
- Neighboring nodes periodically exchange hello message
- When the next hop link in a routing table entry breaks, all active neighbors are informed
- Link failures are propagated by means of Route Error (RERR) messages, which also update destination sequence numbers

Route Error

- When node X is unable to forward packet P (from node S to node D) on link (X,Y), it generates a RERR message
- Node X increments the destination sequence number for D cached at node X
- The incremented sequence number *N* is included in the RERR

- When node S receives the RERR, it initiates a new route discovery for D using destination sequence number at least as large as *N*
- When node D receives the route request with destination sequence number N, node D will set its sequence number to N, unless it is already larger than N

AODV: Summary

- Routes need not be included in packet headers
- Nodes maintain routing tables containing entries only for routes that are in active use
- At most one next-hop per destination maintained at each node
  - DSR may maintain several routes for a single destination
- Sequence numbers are used to avoid old/broken routes
- Sequence numbers prevent formation of routing loops
- Unused routes expire even if topology does not change

Temporally Ordered Routing Algorithm (TORA)

Temporally Ordered Routing Algorithm (TORA) (Park and Corson, 1997a; 1997b) is a distributed protocol designed to be highly adaptive so it can operate in a dynamic network. For a given destination, TORA uses a somewhat arbitrary “height” parameter to determine the direction of a link between any two nodes. As a consequence of this multiple routes are often present for a given destination, but none of them are necessarily the shortest route The TORA routing protocol is based on the LMR protocol. It uses similar link reversal and route repair procedure as in LMR and also the creation of a DAGs, which is similar to the query/reply process used in LMR. Therefore, it also has the same benefits as LMR. The advantage of TORA is that it has reduced the far-reaching control messages to a set of neighboring nodes, where the topology change has occurred. Another advantage of TORA is that it also supports multicasting; however this is not incorporated into its basic operation. TORA can be used in conjunction with Lightweight Adaptive Multicast Algorithm (LAM) to provide multicasting. TORA as its name suggest, is a routing algorithm. It is mainly used in MANETs to enhance scalability. TORA is layered over Internet MANET Encapsulation Protocol (IMEP). This is to ensure reliability in the delivery of control messages and notifications about link status.

Advantages: TORA supports multiple routes. It retains multiple route possibilities for a single source/destination pair. Bandwidth is conserved because of the fever route rebuilding. TORA also supports multicasts.

Disadvantages: TORA’S reliance on synchronized clocks limits in applicability. If the external time source fails, the algorithm ceases to operate. Also route rebuilding may not occur as quickly due to oscillations. During this period this can lead to lengthy delays while for the new routes to be determined.

Global State Routing (GSR)

Global State Routing (GSR) is similar to DSDV. It takes the idea of link state routing but improves it by avoiding flooding of routing messages.

In this algorithm, each node maintains a Neighbor list, a Topology table, a Next Hop table and a Distance table. Neighbor list of a node contains the list of its neighbors (here all nodes that can be heard by a node are assumed to be its neighbors.). For each destination node, the Topology table contains the link state information as reported by the destination and the timestamp of the information. For each destination, the Next Hop table contains the next hop to which the packets for this destination must be forwarded. The Distance table contains the shortest distance to each destination node.

The routing messages are generated on a link change as in link state protocols. On receiving a routing message, the node updates its Topology table if the sequence number of the message is newer than the sequence number stored in the table. After this the node reconstructs its routing table and broadcasts the information to its neighbors.

Quality of Service Challenge

"Providing complex functionality with limited available resources in a dynamic environment" „

Supporting QoS requires knowledge of „

Link delays

Bandwidth

Loss rates

Error rates

Problem with ad hoc networks

Hard to obtain this information

Links constantly changing: node mobility, environmental affects, etc.

QoS Services

Hard QoS

- Guarantee parameters such as delay, jitter, bandwidth
- Required for mission-critical applications
- E.g., air traffic control, nuclear reactor control

Soft QoS

- Aim to meet QoS goals
- Loss in QoS degrades application but does not have disastrous consequences
- E.g., voice, video
- Most research focuses on providing soft QoS

QoS Parameters

Bandwidth

Delay jitter

Delay

Security

Network availability

Battery life