DIPLOMARBEIT

Topology Management and Routing in Wireless Networks - An Overview

ausgeführt am
Institut für Automation
Abteilung Automatisierungssysteme
Technische Universität Wien

unter der Anleitung von
ao. Univ. Prof. Dr. Ulrich Schmid

durch

Hannes STRATIL
Grazerstraße 85a
A–8605 Kapfenberg

Matr.Nr. 9625427

Wien, 6. Februar 2002

Abstract

This master thesis contains a survey about topology management and routing in wireless networks. The thesis briefly discusses several important and common problems and then presents some current solutions.

The main task for Topology Management is the exploration of different network structures. Furthermore, it concerned with how and where the different nodes are located and which communication technology is used in the wireless environment.

Routing in such networks deals with the data exchange between different nodes. The problems appear, whenever a direct communication between the source and the destination is impossible. In this case some other nodes have to relay the information from the source to the destination.

This master thesis is part of the W2F[1] (Wireless/Wireline Factory/Facility Fieldbus) project under the aegis of Prof. Ulrich Schmid at the Technical University of Vienna. The W2F project is concerned with the development of a next-generation LAN/Fieldbus which adresses security, fault-tolerance and real-time issues.

Zusammenfassung

Diese Diplomarbeit liefert einen Überblick über Topology Management und Routing in drahtlosen Netzwerken. Es wird versucht, die wichtigsten und häufigsten Probleme aufzuzeigen und ein paar der bis heute gefundenen Lösungen vorzustellen.

Das Topology Management beschäftigt sich vor allem mit der Frage, welche Strukturen von Netzwerken anzutreffen sind. Weiters aber auch damit, wie und wo die einzelnen Knoten angeordnet sind beziehungsweise wie die verschiedenen Knoten untereinander kommunizieren.

Routing in solchen Netzwerken behandelt hingegen das Problem, wie Daten zwischen den einzelnen Knoten ausgetauscht werden. Hierbei tauchen Schwierigkeiten auf wenn die Datenquelle und die Datensenke nicht direkt miteinander kommunizieren können und deshalb andere Knoten als Zwischenstationen mit dem Weiterleiten der Daten beauftragt werden müssen.

Diese Diplomarbeit ist Teil des W2F[1] (Wireless/Wireline Factory/Facility Fieldbus) Forschungsprojektes unter der Leitung von Prof. Ulrich Schmid an der Technischen Universität Wien. Das W2F–Projekt beschäftigt sich mit der Entwicklung eines zukünftigen LANs/Feldbusses mit hoher Sicherheit, hoher Fehlertoleranz und durchgehendem Echtzeit Verhalten.

[1]The W2F-project (*http://www.auto.tuwien.ac.at/Projects/W2F/*) has received funding from the Austrian START-programme Y41-MAT.

Danksagung

An dieser Stelle möchte ich mich bei all jenen bedanken, die mich in den Jahren meines Studiums und während dieser Diplomarbeit begleitet haben.

In besonderer Weise gilt dies für meinen Betreuer Ao. Prof. Dr. Ulrich Schmid, bei dem ich mich für die finanzielle Unterstützung und für die Betreuung während der Diplomarbeit bedanke.

Ebenso danke ich meinen Eltern Heinz und Helga Stratil, die mir dieses Studium ermöglicht haben und immer hinter mir gestanden sind.

Weiters bedanke ich mich bei Manfred Pircher, Gerhard Pircher, Gerald Wimmer, Uwe Bamschabl, Roman Schmidt und bei all meinen anderen Freunden. Sie haben mich oft vom Studium abgelenkt und mit mir viele heitere, schöne und vor allem geistreiche Stunden verbracht.

Abschließend bedanke ich mich bei Kanonikus Simon Poier, der mich seit vielen Jahren begleitet und mir immer mit Rat zur Seite steht.

Wien, 6. Februar 2001 *Hannes STRATIL*

Contents

Chapter 1

Introduction

In the last few years a new type of local area network has appeared. This new type of LAN, which is the wireless LAN, provides an alternative to the traditional LANs based on twisted pair, coaxial cable, and optical fiber. The wireless LAN serves the same purpose as that of a wired or optical LAN: To convey information among the devices attached to the LAN. But with the lack of physical cabling to tie down the location of a node on a network, the network can be much more flexible – moving a wireless node is easy. As opposed to the large amount of labor required to add or move the cabling in any other type of network. Also going wireless may be a better choice where the physical makeup of the building makes it difficult or impossible to run wire in the building.

As their name indicates, wireless networks do not use physical wire links for communication, but instead rely on radio or infrared transmissions. Inherently, wireless links are less reliable and have significantly lower capacity than wired links. In addition, their capacity is not only a function of distance and power levels as in wired links, but also of propagation conditions and cochannel interference. For example, all other parameters being equal, it might require more transmission power to achieve equal link capacity in a hilly region than in a flat region.

On the other hand, unlike a node in wired networks which has a number of outgoing links, a node in wireless networks is usually characterized by a transmission radius, which indicates the area within which other nodes can hear its transmissions. Consequently, interference between channels can occur, and unless addressed properly it can deteriorate the links capacities greatly. In fact, the capacity of a given link might not necessarily work equally well in both directions, due to differing interference levels at the end-nodes.

In addition to all of that, the scarcity in bandwidth and energy of most wireless networks further increases the difficulty level of the problems by adding additional constraints to the design.

Wireless networks are ideal for portable computers. Using wireless connections allows portable computers to still be portable without sacrificing the ad-

Paradigm	Main property
Wireless Computing	Distributed computing via a wireless network
Ubiquitous Computing	Very many cheap computers widely spread
Decoupled Computing	Computing and moving while being disconnected
Nomadic/mobile Computing	Computing and moving while being connected

Table 1.1: *Paradigms related to wireless computing.*

vantages of being connected to a network. These machines can be setup virtually anywhere within the building.

Wireless networks can be used in combination with cabled LANs. In that all the machines that will require relative mobility will be connected wirelessly, while the stations that are for the most part permanant can be connected through cable.

Terms like ubiquitous computing, nomadic computing, mobile computing and decoupled computing are being used to describe computing systems very closely related to wireless systems but each focusing on a certain property. Ubiquitous computing expects a world in which a huge amount of very cheap computers is widely spread among all the domains of everyday life. Connecting that many computing devices by a wired network would be a time consuming and expensive task, leading to wireless computing. Decoupled computing denotes the ability to compute disconnectedly from servers and even communication networks. Frequent disconnections in turn are predicted to be common in wireless computing since the unreliability of wireless channels or the need to save energy may force wireless sites to temporarily disconnect form the wireless network. Decoupled computing also implies to be able to change location while being disconnected. Nomadic or mobile computing, however, assume users to change location while being connected and computing. Consequently, location management is even harder than in decoupled computing. Since mobile computing is not feasible using conventional fixed wired networks we are back to wireless computing. Table 1.1 briefly summarizes this terminology. A wireless information system is an information system based on wireless computing typically serves distibuted users and involves database technology.

Chapter 2

Location Systems

2.1 Location Sensing Techniques

A broad set of issues arises when we discuss and classify location system implementations. These issues are generally independent of the technologies or techniques a system uses. When attempting to determine a given location we can choose form three major techniques. [2]

2.1.1 Triangulation

The triangulation location sensing technique uses the geometric properties of triangles to compute object locations. Triangulation is divisible into the subcategories of lateration, using distance measurements, and angulation, using primarily angle or bearing measurements.

2.1.1.1 Lateration

The term lateration means for distance measurements what angulation means for angles. Lateration computes the position of an object by measuring its distance from multiple reference positions. Calculating an object's position in two dimensions requires distance measurements from 3 non-collinear points. In 3 dimensions, distance measurements from 4 non-coplanar points are required. Domain-specific knowledge may reduce the number of required distance measurements.

There are two interesting approaches to measuring the distance required by the lateration technique.

1. Time-of-Flight

   Measuring distance from an object to some point P using time-of-flight means measuring the time it takes to travel between the object and point P at a known velocity. The object itself may be moving, such as an airplane travelling at a known velocity for a given time interval, or, as is far more typical, the object is approximately stationary and we are instead observing the difference in transmission and arrival time of an emitted signal.

For example, sound waves have a velocity of approximately 344 meters per second in 21°C air. Therefore, an ultrasound pulse sent by an object and arriving at point P 14.5 milliseconds later allows us to conclude that the object is 5 meters away from point P. Measuring the time-of-flight of light or radio is also possible but requires clocks with much higher resolution (by six orders of magnitude) then those used for timing ultrasound since a light pulse emitted by the object has a velocity of 299,792,458 meters per second and will travel the 5 meters to point P in 16.7 nanoseconds. Also, depending on the capabilities of the object and the receiver at point P, it may be necessary to measure a round-trip delay corresponding to twice the distance.

2. Attenuation

   The intensity of an emitted signal decreases as the distance from the emission source increases. The decrease relative to the original intensity is the attenuation. Given a function correlating attenuation and distance for a type of emission and the original strength of the emission, it is possible to estimate the distance from an object to some point P by measuring the strength of the emission when it reaches P. For example, a free space radio signal emitted by an object will be attenuated by a factor proportional to $1/r^2$ when it reaches point P at distance r from the object.

2.1.1.2 Angulation

Angulation is similar to lateration except, instead of distances, angles are used for determining the position of an object. In general, two dimensional angulation requires two angle measurements and one length measurement such as the distance between the reference points. In three dimensions, one length measurement, one azimuth[1] measurement, and two angle measurements are needed to specify a precise position. Angulation implementations sometimes choose to designate a constant reference vector (e.g. magnetic north) as 0°.

Phased antenna arrays are an excellent enabling technology for the angulation technique. Multiple antennas with known separation measure the time of arrival of a signal. Given the differences in arrival times and the geometry of the receiving array, it is then possible to compute the angle from which the emission originated. If there are enough elements in the array and large enough separations, the angulation calculation can be performed.

[1] In astronomy the azimuth is the angle along the horizon, with zero degrees corresponding to North, and increasing in a clockwise fashion. Thus, 90 degrees is East, 180 degrees is South, and 270 degrees is West.

2.1.2 Proximity

A proximity location sensing technique entails determining when an object is "near" a known location. The object's presence is sensed using a physical phenomenon with limited range. There are three general approaches to sensing proximity:

1. Detecting physical contact

   Detecting physical contact with an object is the most basic sort of proximity sensing. Technologies for sensing physical contact include pressure sensors, touch sensors, and capacitive field detectors.

2. Monitoring wireless cellular access points

   Monitoring when a mobile device is in range of one or more access points in a wireless cellular network is another implementation of the proximity location technique.

3. Observing automatic ID systems

   A third implementation of the proximity location sensing technique uses automatic identification systems such as credit card point-of-sale terminals, computer login histories, land-line telephone records, electronic card lock logs, and identification tags. If the device scanning the label, interrogating the tag, or monitoring the transaction has a known location, the location of the mobile object can be inferred.

2.1.3 Scene analysis

The scene analysis location sensing technique uses features of a scene observed from a particular vantage point to draw conclusions about the location of the observer or of objects in the scene. Usually the observed scenes are simplified to obtain features that are easy to represent and compare. In static scene analysis, observed features are looked up in a predefined dataset that maps them to object locations. In contrast, differential scene analysis tracks the difference between successive scenes to estimate location. Differences in the scenes will correspond to movements of the observer and if features in the scenes are known to be at specific positions, the observer can compute its own position relative to them.

The advantage of scene analysis is that the location of objects can be inferred using passive observation and features that do not correspond to geometric angles or distances as we have seen, measuring geometric quantities often requires motion or the emission of signals, both of which can compromise privacy and can require more power. The disadvantage of scene analysis is that the observer needs to have access to the features of the environment against which it will compare its observed scenes. Furthermore, changes to the environment in a way that alters the perceived features of the scenes may necessitate reconstruction of the predefined dataset or retrieval of an entirely new dataset.

Location system implementations generally use one or more of these techniques to locate objects, people, or both.

2.2 Location System Properties

Before we survey some of the location technologies, let us introduce some special terms and properties [1].

2.2.1 Physical position and symbolic location

A location system can provide two kinds of information: Physical and symbolic. GPS provides physical positions. For example, the peak of the Großglockner is situated at 47.07472°North by 12.69417°East, at a 3798-meter elevation. In contrast, symbolic location encompasses abstract ideas of where something is: In the kitchen, on the Vienna Airport, next to a mailbox or a postal address. A system providing a physical position can usually be advanced to provide corresponding symbolic location information with additional information, infrastructure, or both. For example, a laptop equipped with a GPS receiver can access a separate database that contains the positions and geometric service regions of other objects to provide applications with symbolic information. [3]

2.2.2 Absolute versus relative

An absolute location system uses a shared reference grid for all located objects. For example, all GPS receivers use latitude, longitude, and altitude for reporting location. Two GPS receivers placed at the same position will report equivalent position readings, and 47.07472°North by 12.69417°East refers to the same place regardless of GPS receiver. In a relative system, each object can have its own frame of reference. For example, a mountain rescue team searching for avalanche victims can use handheld computers to locate victims' avalanche transceivers. Each rescuer's device reports the victim's position relative to itself. An absolute location can be transformed into a relative location - relative to a second reference point, that is. However, a second absolute location is not always available. In reverse, we can use triangulation to determine an absolute position from multiple relative readings if we know the absolute position of reference points.

2.2.3 Localized versus extern location computation

Some systems provide a location capability and insist that the object being located actually computes its own position. This model ensures privacy by mandating that no other entity may know where the located object is unless the object specifically takes action to publish that information. For example, orbiting GPS satellites have no knowledge about who uses the signals they transmit.

In contrast, some systems require the located object to periodically broadcast, respond with, or otherwise emit telemetry to allow the external infrastructure to locate it. The infrastructure can find objects in his purview without directly involving the objects in its computation. Personal-badge-location systems fit into this category. Placing the burden on the infrastructure decreases the computational and power demands on the objects being located, which makes many more applications possible due to lower costs and smaller form factors.

2.2.4 Accuracy and precision

A location system should report locations accurately and consistently from measurement to measurement. Some inexpensive GPS receivers can locate positions to within 10 meters for approximately 95 percent of measurements. More expensive differential units (DGPS) usually do much better, reaching 1 to 3 meter accuracies 99 percent of the time. These distances (10 meters or 1 to 3 meters) denote the accuracy, or grain size, of the position information GPS can provide. The percentages (95 percent or 99 percent) denote the precision, or how often we can expect to get that accuracy.

2.3 An overview of location systems

In the remaining part of this chapter we present some of the research and commercial location technologies that are representative of the location-sensing field.

2.3.1 Global Positioning System (GPS)

GPS is a satellite-based global navigation system created and operated by the United States Department of Defense (DOD). Originally intended only to enhance military defense capabilities, GPS capabilities have expanded to provide highly accurate position and timing information for many civilian applications [4, 5, 6, 7].

Twenty four satellites in six orbital paths circle the earth twice each day at an inclination angle of approximately 55 degrees to the equator. This constellation of satellites continuously transmit coded positional and timing information at high frequencies in the 1500 Megahertz range. GPS receivers with antennas located in a position to clearly view the satellites pick up these signals and use the coded information to calculate a position in an earth coordinate system.

Part of every GPS receiver is a radio listening for the signals being broadcast by these satellites. Each spacecraft continuously sends a data stream that contains its orbit information, equipment status, and the exact time. All of the information is useful, but the exact time is crucial. GPS receivers have computers that can calculate the difference between the time a satellite sends a signal and the time it is received. The computer multiplies this time of signal travel by the speed of travel to get the distance between the GPS receiver and the satellite (TIME x SPEED = DISTANCE), location in space.

A GPS receiver must be locked on to the signal of at least three satellites to calculate a 2D position (latitude and longitude) and track movement. With four or more satellites in view, the receiver can determine the user's 3D position (latitude, longitude and altitude). Once the user's position has been determined, the GPS unit can calculate other information, such as speed, bearing, track, trip distance, distance to destination, sunrise and sunset time and more.

Position accuracy depends on the receiver's ability to accurately calculate the time it takes for each satellite signal to travel to earth. This is where the problem lies. There are primarily five sources of errors which can affect the receiver's calculation. These errors consist of

1. ionosphere and troposphere delays on the radio signal,
2. signal multi-path,
3. receiver clock biases,
4. orbital errors, also known as ephemeris errors of the satellite's exact location, and
5. the intentional degradation of the satellite signal by the DOD.

Today, GPS units are accurate to within 20 meters; although in good conditions, units should display an error of less than 10 meters. The combination of these errors in conjunction with poor satellite geometry can limit GPS accuracy to 100 meters 95% of the time and up to 300 meters 5% of the time. Fortunately, many of these errors can be reduced or eliminated through a technique known as "Differential".

2.3.2 Differential Global Positioning System (DGPS)

DGPS [4, 5, 6] works by placing a high-performance GPS receiver (reference station) at a known location. Since the receiver knows its exact location, it can determine the errors in the satellite signals. It does this by measuring the ranges to each satellite using the signals received and comparing these measured ranges to the actual ranges calculated from its known position. The difference between the measured and calculated range is the total error. The error data for each tracked satellite is formatted into a correction message and transmitted to GPS users. The correction message format follows the standard established by the Radio Technical Commission for Maritime Services, Special Committee 104 (RTCM-SC104) These differential corrections are then applied to the GPS calculations, thus removing most of the satellite signal error and improving accuracy.

The level of accuracy obtained is a function of the GPS receiver. Sophisticated receivers can achieve accuracy on the order of 1 meter or less.

2.3.3 Active Badge

The Active Badge location system [8] was developed at Olivetti Research Laboratory, now AT&T Cambridge.

Each object or person the system can locate wears a small tag (Active Badge) that emits a globally unique identifier for approximately a tenth of a second every 15 seconds. These periodic signals are picked up by a network of sensors placed around the host building. A master station, also connected to the network, polls the sensors for badge 'sightings', processes the data, and then makes it available to clients that may display it in a useful visual form. For signalling between the badge and sensor pulse-width modulated infrared (IR) signals are used, because IR emitters and detectors can be made very small and very cheaply. They can be made to operate with a 6m range, and the signals are reflected by walls and therefore are not directional when used inside a small room. Moreover, the signals will not travel through walls, unlike radio signals that can penetrate the small walls found in some buildings. Infrared communication has been used in a number of commercial applications ranging from the remote control of domestic appliances to date backup links for programmable calculators and personal organizers. IR technology has already been exploited commercially, it is inexpensive and readily available.

It is warranted, that several people in the same locality be detectable by the system. Because the signals have a duration of only one-tenth of a second, there is approximately a 2/150 chance that two signals will collide when two badges are placed in the same location. For a small number of people, there is a good probability that they will be detected.

To detect Active Badges in transit through a building, a sensor network must provide thorough coverage through adequate placement and density of sensors. Sensors need to be placed high up on walls or ceiling tiles of offices and on the entrances and exits of corridors and other public areas. An ideal system would take advantage of existing computer networks as a means of gathering badge sightings and relaying the data back to a central server for processing. The Active Badge system provides absolute location information. A badge's location is symbolic, representing, for example, the room in which the badge is located.

As with any diffuse infrared system, Active Badges have difficulty in locations with fluorescent lighting or direct sunlight because of the spurious infrared emissions these light sources generate. Diffuse infrared has an effective range of several meters, which limits cell sizes to small- or medium-sized rooms. In larger rooms, the system can only use multiple infrared sensors.

2.3.4 Active Bat

AT&T researchers have developed the Active Bat location system [9], which uses an ultrasound time-of-flight lateration technique to provide more accurate physical positioning then Active Badges.

Small units called Bats are attached to equipment and are carried by personal. Bats consist of a radio transceiver, controlling logic and an ultrasonic transducer. Each Bat has a globally unique identifier. Ultrasound receiver units are placed at known points on the ceiling of the rooms to be instrumented. Receivers are connected by a wired daisy-chain network.

A base station periodically transmits a radio massage containing a single identifier, causing the corresponding Bat to emit a short unencoded pulse of ultrasound. Simultaneously, the ultrasound receivers in the rooms covered by the base station are reset via the wired network. Receivers monitor the incoming ultrasound and record the time of arrival of any signal from the Bat. Using the speed of the sound in air (which can be estimated from the ambient temperature), the times-of-flight of the ultrasound pulse from the Bat to receivers can be converted into corresponding Bat-receiver distances.

If distances from Bat to three or more non-collinear receivers can be found, its position in 3D space may be determined using the process of multilateration. The position of the object to which the Bat is attached can then be derived. Reflections of the ultrasonic signal from objects in the environment are commonplace, and can cause incorrect measurement of Bat-receiver distances. These erroneous distance measurements are eliminated by the use of a statistical outlier rejection algorithm, so as to improve the accuracy of the calculated Bat position.

After the distance-measuring pulse has been emitted, the base station waits for reverberations of the pulse to die out before triggering another Bat, ensuring that receivers can ascribe incoming ultrasonic signals to the correct Bat. This process divides time into timeslots, each of which can be used to locate a single Bat. In typical office spaces, reverberations may take up to 20 ms to die away, implying that there may be up to 50 timeslots (and hence location updates) per second per base station. The System, as reported in 1999, can locate Bats to within 9cm of their true position for 95 percent of the measurements. Using ultrasound time-of-flight this way requires a large fixed-sensor infrastructure throughout the ceiling and is rather sensitive to the precise placement of these sensors. Thus, scalability, ease of deployment, and cost are disadvantages of this approach.

2.3.5 Cricket

Complementing the Active Bat system, the Cricket Location Support System (MIT Laboratory, Cambridge) [10] uses ultrasound emitters to create the infrastructure and embeds receivers in the object being located. This approach forces the objects to perform all their own triangulation computations. Cricket uses a combination of radio frequency (RF) and ultrasound to provide a location-support service to users and applications. Wall- and ceiling mounted beacons are spread through the building, publishing location information on an RF signal.

A beacon is a small device attached to some location within the geographic space it advertises. Cricket does not attached any semantics to the space advertised by the beacon; any short string can be disseminated, such as the name of a server to contact to learn more about the space or a name resolver for the space to discover resources. A randomised algorithm allows multiple uncoordinated beacons to coexist in the same space. Cricket beacons are inexpensive and more than one of them can be used in any space for fault-tolerance and better coverage.

To obtain information about a space, every mobile and static node has a listener attached to it. A listener is a small device that listens to messages from beacons, and uses these messages to infer the space it is currently in. The listener provides an API (Application Programming Interface) to programs running on the node that allow them to learn where they are, so that they can use this information to appropriately advertise themselves and their location to a resource discovery service.

With each RF advertisement, the beacon transmits a concurrent ultrasonic pulse. A purely RF-based system, using received signal strength to infer the location of the nearest beacon, does not work well. This mainly because RF propagation within buildings deviates heavily from empirical mathematical models and one needs very expensive RF-radio transmitters to produce a constant corresponding signal behaviour over a long time.

Therefore a combination of RF and ultrasonic hardware enables a listener to determine the distance to beacons, from which the closest beacon can be more unambiguously inferred. This is possible by measuring the one-way propagation time of the ultrasonic signals emitted by a beacon, taking advantage of the fact that the speed of sound in air (about 344 meters per second in 21°C air) is much smaller than the speed of light (RF) in air (299,792,458 meters per second). On each transmission, a beacon concurrently sends information about space over RF, together with an ultrasonic pulse. When the listener hears the RF signal, it uses the first few bits as training information an then turns on its ultrasonic receiver. It then listens for the ultrasonic pulse, which will usually arrive a short time later. The listener uses these time difference between the receipt of the first bit of RF information an the ultrasonic signal to determine the distance to the beacon. Of course, the value of the estimated distance is not as important as the decision of which the closest beacon is.

Cricket, in its currently implemented form, is much less precise than Active Bat in that it can accurately delineate 1.2 x 1.2 square-meter regions within a room, while Active Bat is accurate to 9 cm. However, the fundamental limit of range-estimation accuracy used in Cricket should be no different than that of Active Bat, and future implementations may compete with each other on accuracy. Cricket implements both the lateration and proximity techniques. Receiving multiple beacons lets receivers triangulate their position. Receiving only one beacon still provides useful proximity information when combined with the semantic string the beacon transmits on the radio.

Cricket's advantages include privacy and decentralized scalability, while its disadvantages include a lack of centralized management or monitoring and the computational burden that timing and processing both the ultrasound pulses and RF data place on the mobile receivers.

2.3.6 RADAR

A microsoft research group developed RADAR [11, 12], a building-wide tracking system based on the IEEE 802.11 WaveLAN wireless network technology. RADAR is an radio-frequency(RF) based system for locating and tracking users inside buildings. RADAR uses signal strength information gathered at multiple receiver locations to triangulate the user's coordinates.

The RADAR system is built on a deployment of off-the-shelf wireless LAN technology. Access points (or base stations) are located so as to provide overlapping coverage in the area of interest. The mobile user carries with her a communications device equipped with a wireless LAN card capable of bi-directional communication with the access points.

The fundamental idea in RADAR is that in an RF network, the energy level or signal strength (SS) of a packet is a function of the receiver's (mobile user's) location. Consequently, it provides a means for inferring the user's position. There is a clear trend in SS as a user walks about the building. Not surprisingly, the signal received at the mobile is strongest when the receiver is close to the AP and weakest when it is far away. This strong trend, observed for all neighboring APs independently, is exploited by the system to estimate the mobile's location.

With this as motivation, RADAR takes the following approach to location determination. A Radio Map of the building is created. A Radio Map is a database of locations in the building and the observed (or estimated) signal strength of the beacons emanating from the APs as recorded at these locations. So, for example, an entry in the Radio Map may look like $(x, y, z, ss_{i(i=1..n)})$ where (x, y, z) is the physical coordinates of the location where the signal is recorded and ss_i is the signal strength of the beacon signal emanating from the i^{th} AP.

The major part of the effort in deploying RADAR goes into creating the Radio Map of the building.

The method for creating a Radio Map involves a mobile user walking to different locations in the building, preferably close to one another, and explicitly measuring and recording at each location, both the physical coordinates (using a floor layout map as reference) and the signal strength of the beacon packets from each of the APs within range.

To locate the position of the mobile user in real-time, the mobile measures the signal strength of each of the APs within range. It then searches through the Radio Map database to determine the signal strength tuple that best matches the signal strengths it has measured. The system estimates the location associated with the best-matching signal strength tuple, to be the location of the mobile.

Microsoft has developed two RADAR implementations: one using scene analysis and the other using lateration. RADAR's scene-analysis implementation can place objects to within about 3 meters of their actual position with 50 percent probability, while the signal strength lateration implementation has 4.3-meter accuracy at the same probability level. Although the scene-analysis version provides better accuracy, significant changes in the environment, such as moving metal file cabinets or large groups of people congregating in rooms or hallways, may necessitate reconstructing the predefined signal-strength database or creating an entirely new database.

2.3.7 DC magnatic tracker

Electromagnetic sensing offers a classic position tracking method [13]. The large body of research and products that support virtual reality and motion capture for computer animation often offer modern incarnations of this technology. For example, Ascension (http://www.ascension-tech.com) offers a variety of motion-capture solutions, including Flock of Birds and the MotionStar DC magnetic tracker [14]. Flock of Birds (Flock) is a patented magnetic-transducing technique developed and manufactured by Ascension Technology Corporation for interactive computer graphics applications.

The Flock's major elements are a transmitter, transmitter driver circuit, sensor, and signal processing electronics. The transmitter is typically located within one to eight feet of the sensor(s). It consists of three individual antennas arranged concentrically to generate a multiplicity of DC magnetic fields that are picked up by the sensor. The sensor measures the position and orientation of the object to which it is attached. The sensor consists of three axes of antenna that are sensitive to DC magnetic fields. The transmitter driver provides a controlled amount of DC current to each axis of the transmitter. The computer controls the amount of DC current provided by the driver to the transmitter axis. The signal output from the sensor travels to the signal processing electronics. It controls, conditions and converts the analog sensor signals into a digital format that can be read by the computer. The Flock's algorithm next computes the position and orientation of the sensor with respect to the transmitter and then outputs this information to the user's host computer.

Tracking systems such as MotionStar sense precise physical positions relative to the magnetic transmitting antenna. These systems offer the advantage of very high precision and accuracy, on the order of less than 1 mm spatial resolution, 1 ms time resolution, and 0.1°orientation capability. Disadvantages include steep implementation costs and the need to tether the tracked object to a control unit. Further, the sensors must remain within 1 to 3 meters of the transmitter, and accuracy degrades with the presence of metallic objects in the environment.

2.3.8 Ad-hoc location sensing

This approach to locating objects without drawing on the infrastructure or central control borrows ideas from the ad-hoc networking research community. In a purely ad-hoc location-sensing system, all of the entities become mobile objects with the same sensors and capabilities. To estimate their locations, objects cooperate with other nearby objects by sharing sensor data to factor out overall measurement error [15, 16]. In this way, a cluster of ad-hoc objects converges to an accurate estimate of all nearby objects' positions. Objects in the cluster are located relative to one another or absolutely if some objects in the cluster occupy known locations.

The techniques for building ad-hoc systems include triangulation, scene analysis, or proximity. The work of Lance Doherty and colleagues [15], Nirupama Bulusu and colleagues [16], Dragos Niculescu and Badri Nath [17] explores ad-hoc proximity systems that consider variants of the following question: Given a set S of tiny sensor devices and a proximity model of radio connectivity, such as a sphere or circle with a fixed radius, if we know that $s_0...s_n, s_i \subset S$ are subsets of sensors in proximity to one another, how accurately can we infer the relative location of all sensors in set S?

Doherty [15] present an algorithmic approach to this problem using convex optimization. If one node can communicate with another, a proximity constraint exists between them. As a physical example, if a particular RF system can transmit 20m and two nodes are in communication, their separation must be less than 20m. These constraints restrict the feasible set of unknown node positions.

Formally, the network is a graph with n nodes at the vertices (each node having a Cartesian position) and with bi-directional communication constraints as the edges. Positions of the first m nodes are known $(x_1, y_1, ...x_m, y_m,)$ and the remaining $n - m$ positions are unknown. The feasibility problem is then to find $(x_{m+1}, y_{m+1}...x_n, y_n)$ such that the proximity constraints are satisfied. A solution for the position estimation problem requires centralized computation. Namely, all nodes must communicate their connectivity information to a single computer to solve the location problem. In general, efficient computational methods are available for most convex programming problems. In the case of radially constrained connections, using a variable radius constraint improves performance over a globally fixed radius. The placement of known positions around the perimeter of the network results in small errors with as few as four known positions. While requiring significant computational overhead, rectangles placed around the feasible set not only bound the possible positions of the unknown nodes, but also provide a better estimate.

Another example is the Ad-hoc Positioning System (APS) presented in [17]. In principle, if a graph is sufficiently connected, and the lengths of its edges are all known, then its plane topology may be reconstructed. But what is a sufficient degree of connectivity? If we assimilate the graph with a wire frame, where nodes act as hinges, our goal is to determine which part of the graph has non-moving parts, and those will be the nodes which can determine their location. Once such a wire-frame is fixed, it will have a reference system of its own, that eventually has to be aligned to the global coordinate system of the GPS. In order to fix this wire frame somewhere on the global plane, at least three nodes(called landmarks), that are GPS enhanced, or know their position by some other means, have to be present in the connected graph.

The Devices must have software access to the signal strength of the radio signal or to the round-trip delay measurement. Thus offering a way to estimate distance to immediate neighbors. This measurements however, are affected by errors. One of the aims of such a positioning system is to enhance position accuracy as the fraction of landmarks of the entire population increases. Even if it is theoretically sufficient to have three landmarks, the presence of measurement errors will demand higher fractions of landmarks, depending on the requirements of the application.

It is not desirable to have the landmarks emit with large power to cover the entire network for several reasons: collisions in local communication, high power usage, coverage problems when moving. Also, it is not acceptable to assume some fixed positions for the landmarks, as the applications are either in flight deployments over inaccessible areas, or possibly involving movement and reconfiguration of the network. In this case, one option is to use hop by hop propagation capability of the network to forward distances to landmarks. In general, the same principle as GPS is used, with the difference that the landmarks are contacted in a hop by hop fashion, rather than directly. Once an arbitrary node has estimates to a number (≥ 3) of landmarks, it can compute its own position in the plane, using a similar procedure with the one used in GPS position calculation. The immediate neighbors of the landmark can estimate the distance to the landmark by direct signal strength or round-trip delay measurement. Using some propagation method, the second hop neighbors then are able to infer their distance to the landmark, and the rest of the network follows, in a controlled flood manner, initiated at the landmark. Complexity of signaling is therefore driven by the total number of landmarks, and by the average degree of each node.

Chapter 3

Wireless communication

Mobile Computers require wireless network access, although sometimes – when in meeting rooms or at a user's desk – they may remain stationary long enough to be physically attached to the network for a better or cheaper connection.

Wireless networks communicate by modulating radio waves or pulsing infrared light. Wireless communication is linked to wired network infrastructure by stationary or mobile transceivers. The area covered by an individual transceiver's signal is known as a cell. Cell sizes vary widely; for example, an infrared transceiver can cover a small meeting room, a cellular telephone transceiver has a range of few miles, and a satellite beam can cover an area more than 400 miles in diameter.

Wireless communication faces more obstacles than wired communication because the surrounding environment interacts with the signal, blocking signal paths and introducing noise and echoes. As a result, wireless communication is characterized by lower bandwidths, higher error rates, and more frequent spurious disconnections. These factors can in turn increase communication latency resulting from retransmissions, retransmission time-out delays, error-control protocol processing, and short disconnections.

Mobility can also cause wireless connections to be lost or degraded. Users or mobile terminals may travel beyond the coverage of network transceivers or enter areas of high interference. Unlike typical wired networks, the number of devices in a network cell varies dynamically, and large concentrations of mobile users, say, at conventions and public events, may overload network capacity.

From the technical point of view, the past few years have seen tremendous advances in the research, development, and design of mobile radio systems, and many more are expected in the near future [19, 20, 21, 22, 23]. The purpose of this chapter is to give a brief overview of this exciting field and to provide some basic information about the problems inherent to the wireless environment, the communication technologies of common wireless networks, and the standards (IEEE 802.11, HIPERLAN)

3.1 Technical challenges

The idea of being able to communicate with other people or with machines (e.g., databases or video servers) without being constrained to a fixed location is of course very appealing. Unfortunately, mobility introduces new management issues and the mobile wireless environment is very hostile when compared to classic telecommunications systems. In particular, radio channels are inherently prone to serve impairments (especially when the transmitter and/or the receiver are moving) and have limited bandwidth. In this section, the key features, which make the wireless environment so interesting, are presented.

3.1.1 Channel unreliability

Today's computer systems often depend on a network and may cease to function during connection errors. For example, distributed file systems may block waiting for other servers, and application processes may fail altogether if the network stays down too long.

Network failure is a greater concern in mobile computing than in traditional computing because wireless communication is so susceptible to disconnections. Designers must decide whether to spend available resources on the network, or spend them trying to enable systems to cope with disconnections more gracefully and work around them where possible.

The more autonomous a mobile computer, the better it can tolerate network disconnections. For example, certain applications can reduce communication by running entirely locally on the mobile unit rather than by splitting the application and the user interface across the network. In environments with frequent disconnections, it is better for a mobile device to operate as a stand-alone computer than as a portable terminal.

Due to a number of physical factors, signals propagating through wireless channels are subject to severe impairments. The effect of propagation on the transmitted signal can be adequately described by three phenomena, namely, path loss, shadowing, and multipath fading.

- Path loss accounts for the signal attenuation due to the physical distance r between transmitter and receiver. In free space, the received power is proportional to r^{-2}, whereas in groundwave propagation there is an additional effect due to the combination of the direct (free-space) propagated wave and its replica as reflected by the earth surface. In this case, the path loss behaviour is better described by the function r^{η}, where η depends on a number of properties of the environment, and usually takes values between 2 and 5. More accurate models account for other effects, by introducing some correction factors in the relationship between received power and distance to model, for example, effects similar to guided propagation (as happens along streets with tall buildings). Such models are usually employed in system simulations, and their parameters can be chosen by direct measurements of the environment to be studied.

- Another macroscopic effect, i.e., an effect which does not change much over distances on the order of many wavelengths, is the shadowing effect (sometimes called shadow fading) which accounts for obstructions along the propagation path (e.g., building or trees), which may cause significant signal attenuations. As these attenuations combine with each other, shadowing is often modelled as having Gaussian distribution when path loss is expressed in dB. This model (also called log-normal shadowing), has been found to match field measurements very well.

- On a much finer scale, signals are subject to multipath fading. Due to various objects and reflectors, the signal does not directly propagate to the receiver, but rather it is scattered in many directions so that multiple versions of it, coming from different angles and with different attenuations and delays, are received. If the delay differences among the various reflections are much smaller than the time scale at which the signal is detected (e.g., the bit duration), then the channel impulse response is very short (in time), and the effect of multipath in this case is only an additional attenuation. Such a situation corresponds to the transfer function of the channel being essentially constant over the signal bandwidth, and is called flat fading. If on the other hand the delay distribution on the different paths is spread out (e.g., it exceeds the bit duration), then the channel changes the spectral shape of the transmitted signal, which is therefore distorted. As different frequencies may undergo different attenuations, this case is called frequency-selective fading, and is typical of wideband signalling. For digital communications, some sort of equalization is usually needed in this case. Also, the fact that not all spectral components fade together can be exploited as a form of frequency diversity.

Since in a wireless system terminals are allowed to move, the channel impulse response changes in time in general. Even in the case in which terminals do not move, the environment around them may change (e.g., vehicles or people passing by). Path loss and shadowing are relatively constant over significant distances (tens of meters or more), whereas multipath fading can usually be considered independent at distances on the order of the wavelength (which is equal to the speed of light divided by the carrier frequency, and is (λ=33 cm at 900 MHz), and therefore changes very rapidly. This means that the channel, besides being subject to significant impairments, is also very unstable and sometimes may even suddenly disappear. This makes wireless communications very different from communications over more stable and predictable channels such as cables and fibers or even microwave links. Stabilizing the radio channel and overcoming problems posed by the propagation environment is a major task in the design of wireless communications systems.

Typical techniques which are used in today's systems to counteract the channel behaviour are power control and diversity. Power control is based on the concept of trying to boost the power if the channel is currently undergoing large attenuations and to reduce the power when the channel experiences good conditions. This ideally results in all users transmitting just as much power as they need in order to meet their quality objectives, and therefore minimizes the interference level caused to others, under Quality of Service constraints. Implementation of the power control mechanism requires knowledge of the channel attenuation. In open-loop power control schemes, the transmitter autonomously estimates the channel attenuation (either measuring the received signal in full duplex communications, or through the use of a pilot channel) and adjusts the transmitted power in order to compensate for such attenuation. In closed-loop power control, based on the quality of the received signal the receiver sends feedback to the transmitter indicating whether the power should be increased or decreased.

A classic means of counteracting multipath fading is through diversity, which is based on the fact that multiple copies of a signal received at antennas sufficiently apart from each other (a wavelength is usually considered sufficient, at least at the mobile terminal) experience almost independent fading, so that it is very unlikely that all antennas receive weak signals at the same time. More advanced diversity systems try to optimally combine all received signals rather than just selecting one of the signals.

Robust modulation techniques can also be used to maintain transmission quality even under severe propagation conditions. A well-known example for digital communications is spread spectrum modulation, where signal processing techniques are used at the transmitter to spread the signal power over a wide band and at the receiver to despread it. At the same time, the power spectral density of other spread spectrum users is essentially unchanged, whereas the power of narrowband interferers is spread over the whole bandwidth. Therefore, when filtering over the intended signal bandwidth after despreading, the power interference is greatly reduced. The robustness of this technique to interference comes at the expense of bandwidth expansion. More about spread spectrum later in this chapter.

Finally, another well-established means to fix the channel unreliability in digital communications is through the use of error control techniques, whereby redundant symbols are added to the information stream to be transmitted, and are used at the receiver for enhanced detection. Forward error correction schemes, traditionally designed for channels with additive white Gaussian noise, are applied in this environment jointly with interleaving, whereby the bits are scrambled prior to transmission and de-scrambled at the receiver, thus breaking the error correlation inherently present because of the channel memory. The interleaving/deinterleaving operation introduces some delay which effectively limits the extent to which correlation can be removed: In the presence of fast varying channels interleaving usually works very well, whereas for slow channels it is basically ineffective (but in this case very accurate power control can be performed due to high channel predictability).

3.1.2 Low bandwidth

Mobile computing designs need to reflect a greater concern for bandwidth consumption and constraints than do designs for stationary computing. Wireless networks deliver lower bandwidth than wired networks: Common products for portable wireless communications achieve only 1 megabit per second for infrared communication, 2 Mbps for radio communication, and 9–14 kilobits per second for cellular telephony, while Ethernet provides 10 Mbps, fast Ethernet and FDDI 100 Mbps, and ATM (asynchronous transfer mode) 155 Mbps. Even nonportable wireless networks, such as the Motorola Altair, barely achieve 5.7 Mbps.

Practical communications schemes today operate at UHF (Ultra-high Frequency) bands (typical bands are 900 MHz and 1.8 GHz), and additional frequency bands being considered for future systems are mostly in the 1–6 GHz range (even though 17 GHz has been designated for wireless LANs in Europe, and there are proposals for systems at frequencies up to 60 GHz). The amount of bandwidth available in these ranges is not large, since the radio spectrum itself is an inherently public resource, and is already very crowded due to the presence of other services (e.g., broadcast TV, military communications, and point-to-point radio links). Therefore, wireless systems are always limited by interference, which dictates the amount of bandwidth available and imposes restrictions on how it can be used, whereas interference effects in wireline networks, although present (e.g., crosstalk), are less significant in limiting the capacity. Important consequences of this fact are (i) the limited bandwidth available for wireless services, which constrains both the capacity (i.e., the number of users the system can accommodate) and the user's capabilities (i.e., the bandwidth which may be instantaneously available to a user), and (ii) the need to conform to regulatory constraints. The fact that the transmission medium in radio communications is inherently public also raises concerns about its security and privacy. Means need to be provided to guarantee that telephone conversations or data transactions cannot be eavesdropped (matching at least the security achieved in wireline networks). In addition, in this environment, privacy about the user's location and equipment security needs to be ensured.

Because of the shortage of the radio spectrum, a major research objective towards improved design of wireless networks has been to provide more bandwidth-efficient techniques, in order to increase the number of users who can be accommodated for a given system bandwidth. A major breakthrough in this sense was the invention of the cellular concept, devised at Bell Laboratories in the 1970s. Before then, the number of available channels was very limited, because of the limited spectrum available. Cellular systems are based on the same concept used for many years in broadcasting, where the same channel can be used in non-interfering geographic locations. This allows a capacity increase which is in principle unlimited, even though other issues make it infeasible to reduce the cell size beyond a certain limit.

Typical cellular systems today are based on circuit switching, which has been a very successful paradigm for speech communications. The total system bandwidth is split into a number of channels, and these channels are in turn assigned to cells, according to the frequency reuse plan, if any. Therefore, each cell has a number of channels, and cells using the same channels are separated geographically by a sufficiently large distance, thereby causing small interference to each other. Cells using the same radio channels belong to the same co-channel set, and the number of distinct co-channel sets is sometimes called cluster size. How the system bandwidth is subdivided into channels depends on the multiple access technology of choice. The early systems used Frequency Division Multiple Access (FDMA), in which the system bandwidth is sliced in the frequency domain to achieve orthogonal (non-overlapping in frequency) channels. With the introduction of digital transmission in wireless communications systems, other techniques became possible, most notably Time Division Multiple Access (TDMA) and Code Division Multiple Access (CDMA). In the former, the time axis has a framed structure, and different channels use the same bandwidth but at different times, thereby achieving orthogonality in the time domain. In CDMA, Spread Spectrum modulation is used to allow multiple users to communicate simultaneously in the same radio channel, and the separation is obtained through signal processing techniques: If all signals are synchronously received, and in the absence of multipath, it is possible to achieve perfect orthogonality in the code domain, whereas some residual interference will be present if multipath is present or synchronization is not possible (due, for example, to different propagation delays). Hybrid multiple access schemes can also be envisioned, and are actually used in many practical system implementations. A very important advantage of CDMA with respect to TDMA and FDMA is its inherent robustness to interference, which makes it possible to use the same frequency band in all cells. This eliminates the need for frequency planning (which is a major undertaking, especially in systems with small cells), and achieves higher capacity, since the fact that the same frequency can only be used in distant cells results in a significant capacity penalty in TDMA and FDMA, where the bandwidth assigned to each cell is only a fraction of the total.

A major improvement in the spectrum utilization efficiency has been achieved with the introduction of digital communications systems. Digitized speech can be easily compressed, so that, for example, the 30 kHz necessary for a single FM analog voice channel (used in the early systems) can support three digital voice channels with data rate of 16 kbps. The original Pulse Code Modulation technique, where voice signals were sampled and quantized, has been successively refined, based on exploiting the large amount of redundancy of the source, and schemes requiring fewer and fewer bits per second to reliably transmit a single telephone call have been developed. An even more dramatic compression rate can be achieved by vocoders (voice encoders) which, instead of compressing the sampled version of the actual speech waveform, model the mechanics of how the sounds are produced as linear filters, and transmit their parameters which are then used at the receiver to generate the same sounds.

The increased compression rates, along with the use of bandwidth-efficient digital modulation schemes such as QPSK (quadrature phase shift keying) or Gaussian MSK (minimum shift keying), make it possible, for example, the replacement of each 30-kHz analog channel with three TDMA channels, thus tripling the system capacity. An even larger gain may be obtained when using CDMA, because of the possibility of reusing the spectrum everywhere, so that the capacity gain offered by CDMA with respect to the early systems is estimated to be on the order of three to five in circuit switched systems.

Certain software techniques for coping with disconnection can also help cope with low bandwidth. Network usage typically occurs in bursts in that demand temporarily exceeds available bandwidth. For example, delayed write-back and prefetching use the periods of low network activity to reduce demand at the peaks. Delayed write-back can even reduce overall communication if the data to be transmitted is further mutated or deleted before it is transmitted. Prefetching involves knowing or guessing which files will be needed soon and downloading them over the network before they are demanded. Bad guesses can waste network bandwidth, however.

3.1.3 Mobility

The ability to change locations while connected to the network increases the volatility of some information. Certain data considered static for stationary computing becomes dynamic for mobile computing. For example, a stationary computer can be configured statically to prefer the nearest server, but a mobile computer needs a mechanism for determining which server to use.

Mobility introduces several problems: A mobile computer's network address changes dynamically, it's current location affects configuration parameters as well as answers to user queries, and the communication path grows as it wanders away from a nearby server.

As people move, their mobile computers will use different network access points, or "addresses". Today's networking is not designed for dynamically changing addresses. Active network connections usually cannot be moved to a new address. Once an address for a host name is known to a system, it is typically cached with a long expiration time and with no way to invalidate out-of-date entries. In the Internet Protocol, for example, a host IP name is inextricably bound with its network address; moving to a new location means acquiring a new IP name. Human intervention is commonly required to coordinate the use of addresses.

To communicate with a mobile computer, messages must be sent to its most recent address. Four basic mechanisms determine a mobile computer's current address: broadcast [25, 26], central services [27], home base [28], and forwarding pointers [25].

- Selective broadcast: With the broadcast method, a message is sent to all network cells asking the mobile computer sought to reply with its current address. This becomes too expensive for frequent use in a large network, but if the mobile computer is known to be in some small set of cells, selectively broadcasting in just those cells is workable. The selective broadcast method can be used to obtain the current address when only approximate location information is known. For example, a slightly out-of-date cell address may suffice if adjacent cells are known.

- Central services: With the central services method, the current address for each mobile computer is maintained in a logically centralized database. Each time a mobile computer changes its address, it sends a message to update the database. Even with the database's centralized location, the common techniques of distribution, replication, and caching can be used to improve availability and response time. In current cellular systems, the network is divided into service areas, each composed of a number of cells managed by a central unit, called Mobile Switching Centre (MSC). Each MSC has two databases, called Home Location Register (HLR) and Visitor Location Register (VLR). The HLR contains information regarding users who are registered as subscribers in the area, whereas the VLR contains information regarding users who are registered as subscribers somewhere else but happen to be roaming in the area. Every time a user registers as a roamer in a new area, some signalling takes place between the MSCs involved, and the databases are appropriately updated. Basically, the information contained in those registers allows the network to locate the user, i.e., to identify the set of cells where the user can be found at any given time. In fact, every time the network needs to locate a user, the HLR at the user's home MSC will be looked up. If the user is not in that area, the HLR contains information on where the user can be found. If an incoming call for that user arrives, it may be redirected to the MSC of the area in which the user can be found. The MSC will then instruct the central transmitters, called Base Stations (BS), serving all cells in that area to send a paging message for the user, who will reply to the BS of the appropriate cell. This response completes the location procedure, and the call setup can be completed at the designated BS.

- Home base: The home base method is essentially the limiting case of distributing a central service; that is, the location of a given mobile computer is known by a single server. This aggressive distribution without replication can lead to low availability of information. For example, if a home base is down or inaccessible, the mobile computers it tracks cannot be contacted. If users sometimes change home base, the address migration problem arises again, albeit with much lower volatility.

- Forwarding pointers: With the forwarding pointer method, each time a mobile computer changes its address, a copy of the new address is deposited at the old location. Each message is forwarded along the chain of pointers until it reaches the mobile computer. To avoid the inefficient routing that can result from long chains, pointer at message forwarders can update gradually to reflect more recent addresses.

  Although the forwarding pointer method is among the fastest, it is prone to failures anywhere along the trail of pointers, and in its simplest form it does not allowed forwarding pointers to be deleted unless all possible message sources have been updated. Hence, forwarding pointers are often used only to speed up the common case, and another method is used to fall back on for failures and to allow reclamation of old pointers.

  The forwarding pointer method requires an active entity at the old address to receive and forward messages. This does not fit standard networking models, where either a network address is a passive entity, such as an Ethernet cable, or it's specific to the mobile computer, which cannot remain to forward its own message. This mismatch introduces subtle difficulties in implementing forwarding efficiently(for example, with intracell traffic or when multiple gateways are attached to a network address).

3.2 Wireless networks and communication techniques

3.2.1 Current wireless cellular (mobile telephone) systems

The following are the main elements of all cellular systems(see Figure 3.1):

- Mobile terminal, which contains a radio transceiver and a processor;

- Base Station(BS), which manages communications to and from the users belonging to its own cells, and is composed of a radio part and a Base Station Controller;
- Mobile Switching Centre(MSC), which controls a number of BS (typically a few tens), provides additional functionalities such as call management and diagnostics and fault recovery, and is the interface to and from the public switched telephone network;
- Databases, such as the already mentioned Home Location Register(HLR) and Visitor Location Register(VLR), plus the Equipment Identity Register and Authentication Centre, used for security and authentication purposes.

  Also, the following control channels are present in all systems:
- Broadcast channels (base-to-mobile), where general control information is transmitted to all mobiles;
- Paging channels (base-to-mobile), used to notify a mobile user of an incoming call;
- Random access channels (mobile-to-base), used for mobile-initiated call requests.

These elements provide essential communications services and network management functionalities, which enable proper system operation. In particular, they provide location and mobility management. Therefore, although in different forms and possibly with different names, they are present in all cellular systems.

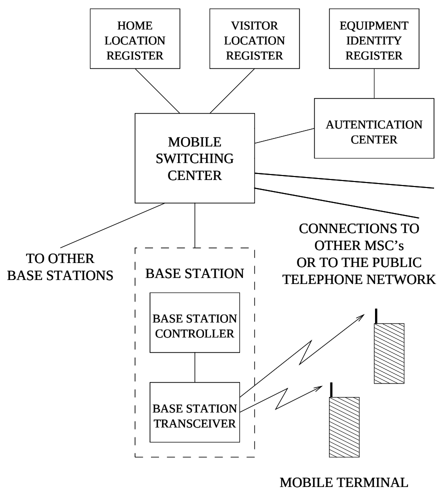

Figure 3.1: *Main components of a cellular system and their relationships.*

3.2.1.1 First generation cellular systems

The early cellular systems were based on analog FM. Examples of such systems are the Advanced Mobile Phone Service (AMPS) in the United States and the Total Access Communications System (TACS) in Europe. As an example, the AMPS analog cellular standard was developed by Bell Labs in the 1970's and deployed in 1983. The AMPS system operates in the vicinity of 850 MHz. Two competing operators are allocated 25 MHz of bandwidth each, corresponding to 833 30-kHz radio channels. Since in these systems the modulation format for the voice channels is analog, FDMA must be used for multiple access. With a save-cell frequency reuse plan (which guarantees adequately low interference), this results in about 60 full-duplex voice channels per cell.

First generation systems provide supervisory signalling during voice transmission, to verify proper connection between Base Station and Mobile Terminal. This is achieved by sending tones at frequencies above the audio bands. In addition, a wideband signalling capability is provided (e.g., for use during handoffs) by means of the blank-and-burst technique, in which bursts of data are sent for short (almost unnoticeable) durations while speech is muted. During a call, the BS makes signal strength measurements of its own mobiles. Also, a receiver is used to monitor signal strength of users in other cells. Based on the measurements forwarded by the BS, the MSC makes decisions about call handoffs whenever needed.

For data transmissions, the cellular telephone is connected to a standard modem and used to dial an Internet Service Provider(ISP), making a low speed connection [30]. Typical speeds are 2,4 kbps to 9,6 kbps.

A major drawback of first generation standards is their incompatibility, due for example to different channel spacing or frequency band of operation. This problem was especially severe in Europe, where many different countries had different standards, preventing users from roaming abroad. This is one of the reasons that the second-generation pan-European GSM has been so successful both in the standardization process and commercially. Another disadvantage is the lack of security. The Signals are not encrypted and both voice and data can easyily be intercepted.

3.2.1.2 Second generation cellular systems

The main difference between first and second generation cellular systems is that the modulation format used in the latter is digital rather than analog. This introduces considerable flexibility in handling the information signals, allowing the use of error control coding, source coding and compression, TDMA and CDMA. Also, digital modulation is inherently more robust to channel impairments than frequency modulation, even though channel errors may result in artifacts which are less tolerated by human perception. Also, digital transmission increases the compatibility with the fixed portion of the telephone network (which is mostly digital), allows greater switching and routing flexibility, and lends itself to implementation in VLSI. Most importantly, the digital transmission format is more suitable for data communications, which is expected to be a major player in future systems and is already supported by some second generation systems. The coverage provided by second generation systems is not yet universal, but roaming is fully supported at least within Europe (this was a major goal of GSM), within North America and within Japan.

There are basically two different types of second generation systems, depending on whether they are TDMA-based or CDMA-based.

- TDMA-based standards:

  IS-54/136(D-AMPS): Digital AMPS. An upgrade to the analogue AMPS in the United States. Designed to address the problem of using existing channels more efficiently, DAMPS (IS-54) employs the same 30 kHz channel spacing and frequency bands (824–849 and 869–894 MHz) as AMPS. By using TDMA instead of FDMA, IS-54 increases the number of users from 1 to 3 per channel (up to 10 with enhanced TDMA). The successor of IS-54, which provided a only TDMA voice channel, were IS-136 which also uses TDMA on the control channel. An AMPS/D-AMPS infrastructure can support use of either analogue AMPS phone or digital D-AMPS phones. This was because the Federal Communications Commission mandated only that digital cellular in the U.S. must act in a dual-mode capacity with analogue. Both operate in the 800 MHz band. These telephones would support Internet access in a similar way as the AMPS telephones, often falling back to the AMPS protocol to deliver the services.

  GSM: Global System for Mobile Communications [29]. The first European digital standard, developed to establish cellular compatibility throughout Europe. It's success has spread to all parts of the world and over 80 GSM networks are now operational. It operates at 900 and 1800 MHz. One interesting feature built into GSM is the Subscriber Identity Module (SIM), a small chip that can be moved from device to device and identifies the owner and the service they subscribe to. GSM used encryption to secure communications and has native data support built in for 9,600 BPS rates.

  PDC: Personal Digital Cellular is a TDMA-based Japanese standard operating in the 800 and 1500 MHz bands.

- CDMA-based standards:

  IS-95: The only second generation CDMA-based standard is IS-95. It uses the same frequency bands as AMPS and supports AMPS operation, employing spread-spectrum technology and a special coding scheme. It was adopted by the Telecommunications Industry Association (TIA) in 1993.

TDMA-based systems

In GSM [29], which is an example of TDMA-based second generation system, the voice signal consists of a 13 kbps stream produced by the RPE (Regular Pulse Excited) speech coder. This bit stream is subdivided into two classes. The bits which are more critical in the speech reconstruction at the receiver are given protection, and encoded with a rate 1/2 convolutional code, whereas the others are transmitted unprotected. Additional protection is provided by the interleaving mechanism which breaks the channel error correlation and makes the use of the error correction code more effective. This coding scheme introduces an additional 10 kbps to the original bit stream which, with the further addition of synchronization bits, training sequences for equalization and guard times, sums up to a total of about 33.8 kbps, which is the channel rate for each user (a 156-bit time slot every 4.6 ms).

In addition to the common control channels already mentioned (broadcast, paging and random access), GSM provides two additional channels, called Associated Control Channels (ACCH). The Slow ACCH (SACCH) is assigned as part of a connection, and is implemented by using dedicated frames in the GSM multiframe structure. The data rate on the SACCH sums up to a little less than 1 kbps. A typical use of the SACCH is the transmission of information about the signal strength received by the mobile from a number of surrounding BSs, to be used in the event that a handoff is needed. The Fast ACCH is not continuously present, and is activated only under exceptional circumstances (e.g., while performing a handoff), when the transmission capacity normally assigned to the voice signal is preempted and reassigned to control traffic. Therefore, unlike in the analog systems, when a call is initiated a user is assigned a two-way control channel (SACCH) in addition to the two-way voice channel.

Mobile Assisted Handoff (MAHO) is implemented in GSM. Each base station has a broadcast channel which sends a pilot signal. Via its SACCH, a user learns the identities of the broadcast channels of the neighboring cells. During idle time, taking advantage of the TDMA structure where a user is active only 1/8th of the time in either direction, each user listens to the broadcast channels of the closest 32 cells and reports data about the 6 strongest to its own BS via the SACCH. This information is periodically updated and used to determine candidate BS if the need for a handoff arises.

The various Base Stations are interconnected through a backbone network, which also includes Mobile Switching Centres and their databases. In order to manage a complex network with a variety of features, GSM proponents have chosen to rely on some already defined procedures and interfaces, possibly with some adaptation for the mobile environment. GSM uses the CCITT signalling system number 7, originally defined for signalling in wireline ISDN.

	Europe (GSM)	North America (IS-54/136)	Japan (PCD)
access method	TDMA	TDMA	TDMA
carrier spacing	200 kHz	30 kHz	25 kHz
users per carrier	8(16)	3(6)	3(tdb)
modulation	GMSK	Π/4–DQPSK	Π/4–DQPSK
voice codec	RPE 13 kbps	VSLEP 7.95 kbps	VSLEP 6.7 kbps
voice frame	20 ms	20 ms	20 ms
channel code	convolutional	convolutional	convolutional
coded bit rate	22.8 kbps	13 kbps	11.2 kbps
TDMA frame duration	4.6 ms	20 ms	20 ms
interleaving	40 ms	27 ms	27 ms
ACCH	extra slot	in slot	in slot
Handoff	MAHO	MAHO	MAHO

Table 3.1: *Summary and comparison of second generation TDMA-based system parameters.*

An evolution of GSM is the DCS1800 system, whose architecture is very similar to what has been described so far. Differences with GSM are the frequency range (1800 MHz rather than 900 MHz), the amount of spectrum allocated (150 MHz rather than 50 MHz), the cell size (microcells rather than macrocells). The DCS1800 is targeted towards the so-called PCS services (which potentially include data, fax, e-mail and various other services), whereas GSM has as its primary goal cellular telephony.

Other TDMA-based standards include IS-54 and PDC. Similar to GSM, they use speech compression and channel coding, organize transmissions in frames, provide associated control channels and support MAHO. Specific differences among the three systems are detailed in Table 3.1.

CDMA-based systems

As already mentioned, IS-95 is based on CDMA, which takes advantage of spread spectrum modulation in order to guarantee reliable communication in the presence of interference and multipath. The particular interference rejection properties of spread spectrum allow in principle the use of the same frequency band everywhere throughout the system, so that user channel separation is achieved by the appropriate use of quasi-orthogonal spreading sequences. Due to the fact that the bandwidth of the transmitted signal is much larger than the bandwidth of the information signal (a factor of 128 is specified in IS-95), CDMA can enjoy the benefit of frequency diversity, since different spectral components fade independently.

In the time domain, this corresponds to being able to discriminate among the various replicas of the signal arriving at the receiver with different delays, which can therefore be constructively added rather than randomly interfere. The particular structure which is able to combine signals in this way is called RAKE receiver [24].

Another advantage of spread-spectrum systems is the possibility of trading bandwidth expansion for coding redundancy. For example, if the channel bandwidth is 100 times larger than the user's signal bandwidth, it is possible to transmit such signal uncoded with a spreading factor of 100, or coded with a rate-1/2 code and a spreading factor of 50. This allows the use of coding (which is essential in certain applications) without further increase in the bandwidth of the transmitted signal.

Due to the fact that all users share the same bandwidth and that interferers are essentially seen by the intended receiver as Gaussian noise, any techniques to reduce interference or improve detection result automatically into a capacity gain. That is, if one can find a means of reducing the interference by 50%, twice as many users can be admitted to the system. Similarly, if a detection scheme achieves the same performance with a Signal-to-Interference Ratio 3 dB lower, then twice as many users can be admitted. Also, this characteristic provides graceful performance degradation as users join the system, so that CDMA systems have "soft capacity", in the sense that a user can always be admitted to the system if a slight performance degradation for all users is accepted. This is not possible in static resource assignment systems (such as TDMA and FDMA with fixed channel allocation), where there is a constant number of physical channels per cell, and one more cannot be added.

Typical techniques to enhance detection are the use of antenna diversity and of error correction coding (convolutional codes with rate 1/2 on the base-to-mobile link and rate 1/3 on the mobile-to-base link are specified in IS-95). Interleaving is also used to improve the error correction coding performance. Typical techniques to reduce interference are cell sectorization (120°antennas are used, which pick up 1/3 as much interference as omnidirectional antennas) and silence suppression (during periods when the user is silent, which amount to more than half the time in typical telephone conversations, the transmitter switches to a low-rate mode used for signalling and synchronization purposes only).

A key requirement for spread-spectrum systems to work properly is that all signals reach the receiver with the same power, as power imbalances may greatly reduce the system capacity. To this purpose, power control is to be used in the mobile-to-base direction, in order to compensate for the channel attenuation. If perfect power control were possible, all signals coming from mobiles in the same cell would have exactly the same power at the Base Station. On the other hand, power control errors do occur and interference from other cells cannot be controlled, and these effects degrade the performance of the system. In any event, accurate power control is a critical requirement of CDMA-based systems for proper operation, unlike in other systems where power control may increase the performance to some extent but is by no means critical. Open-loop power control is implemented based on the measurements of the pilot signal transmitted by the BS. Closed-loop power control is also specified in IS-95, with the BS sending one-bit power control commands every 1.25 ms, indicating whether the mobile should turn up or down the transmitted power. A power dynamic range of 80 dB is specified. In the other direction (base-to-mobile), a simpler power control scheme is used, whose main goals are improved transmission quality and minimized interference.

As in other systems, paging and random access channels are present. The broadcast channels are used by the Base Stations to transmit a pilot signal, which is used by the mobiles as a power control reference and also as a coherent reference to be used for signal detection. A synchronization channel is also provided, which is used in the acquisition phase as a timing reference.

MAHO is specified in IS-95, where a mobile monitors the pilot signals coming from a number of cells and reports the results of these measurements to the controller. A unique feature of IS-95 is the so-called "soft handoff", where during the handoff procedure a mobile may be connected to more than one BS simultaneously. This allows the use of macro-diversity reception (the network can combine signals received by multiple BS), and provides a means to avoid interruptions as, unlike in other systems, the connection to the new Base Station is set up before the connection with the old BS is torn down ("make before break").

3.2.1.3 Data transmission over cellular networks

In 1993, the US cellular industry developed the Cellular Digital Packet Data standard (CDPD) [31], intended to coexist with the first and second generation cellular systems in the US, and to use the 30 kHz radio channels on a shared basis to provide packet-switched data communications. CDPD takes advantage of the fact that some radio channels in cellular systems are often left unused during normal operation, and in this case the corresponding radio resources can be devoted to data transmission. By doing so, CDPD is able to provide packet data services without additional bandwidth. CDPD uses the existing cellular base stations and air interface, but not the MSCs, as it has its own routing capabilities.

CDPD uses the 30 kHz AMPS radio channels to provide 19.2 kbps data. The use of the cellular radio channels is only allowed when such channels are not supporting telephone calls, which have preemptive priority over the data traffic. In the standard specifications, when a channel used by CDPD is assigned to a telephone call, another empty channel is sought, so that the CDPD connections "hop" among various channels according to the channel assignment dynamics of the cellular network. In practice, however, in many cases a voice channel per cell is permanently reserved for CDPD use, as this is much simpler to manage (although less efficient). CDPD service requires a CDPD capable device like a PCMCIA modem card. The user gets a static IP address and can use the service just as with any ISP.

The multiple access protocol in CDPD is DSMA/CD, which belongs to the class of carrier-sense random access schemes described in the next section. Data-link functions are provided by the Mobile Data Link Protocol, whereas the Radio Resource Management Protocol is responsible for finding and managing unused cellular radio channels. Services provided by CDPD include broadcast, e-mail, monitoring and dispatch. Error control capability is also provided, through the use of a Reed Solomon code.

Another standard for cellular data transmission is IS-99 [32], which uses the CDMA-based physical layer of IS-95. IS-99 supports TCP on top of a radio link protocol (RLP) which partially recovers from channel errors via Selective Repeat ARQ. Other wireless data systems include the General Packet Radio Service (GPRS) supported in GSM; ARDIS, by Motorola, which provides a very good in-building coverage in urban areas; and RAM, based on Ericsson's MOBITEX standard.

3.2.2 Wireless LAN

A wireless LAN (WLAN) is typically an extension of a wired LAN. WLAN components convert data packets into radio waves or infrared (IR) light pulses and send them to other wireless devices or to an access point that serves as a gateway to the wired LAN. Most WLANs today are based on the IEEE 802.11 and 802.11b standards for wireless communication between devices and a LAN. These standards permit data transmissions at 1 to 2 Mbps or 5 to 11 Mbps, respectively, and specify a common architecture, transmission methods, and other aspects of wireless data transfer to improve interoperability among products.

With the need for mobility and the positive experience with wireless telephony, it is natural for users to be attracted by wireless capabilities in the computer communications domain, so that the demand for wireless LANs has fuelled a considerable amount of research and a number of commercial products.

Wireless LANs have many advantages over wired LANs, including quick and simple installation (no need for wiring), reconfigurability, application in environments where wiring is difficult or impractical (e.g., historical buildings), support for ad-hoc and temporary facilities (e.g., during a convention). Moreover, they can be used in environments where mobility is desired (e.g., inventory of stocks, car rental check-in).

The main differences between wireless LANs and cellular telephone systems are as follows. Cellular systems today are targeted towards circuit-switched low-rate voice communications, provide wide-area coverage, support mobility of terminals, have switching capabilities. On the other hand, wireless LANs are targeted towards computer and data applications, support high transmission speeds (up to several Mbps), are based on packet-switching technology, have limited switching capability and support no or limited mobility, and are intended as local area facilities. It is therefore not surprising that a completely separate set of technological solutions has been devised for cellular systems and wireless LANs (as happened for wireline telephone and computer networks). On the other hand, it must be said that the current trend towards integration of communications and computing will bring the two fields closer and closer together, and future wireless systems will grow out of a combination of these technologies, as already testified by many proposals about third generation systems which share many aspects of computer communications protocols.

The major difference between wireline and wireless LANs is in the data rates and quality they can support. Providing high transmission speeds in the wireless environment is very challenging because of the very hostile propagation environment. The IEEE 802.11 and HIPERLAN standards offer channel speeds of 1–2 Mbps and 25 Mbps, respectively (to be compared with 10 Mbps of Ethernet and 100 Mbps of fast Ethernet), and can operate over limited distances (tens of meters or less, much smaller than those allowed in wireline). Higher data rates can be achieved by means of Infrared communications (as opposed to radio transmission, implicitly considered so far), which on the other hand operates on shorter distances and usually requires line-of-sight propagation.

3.2.2.1 Standard technologies

There are two standardization activities for WLAN:

- IEEE standard (802.11)
- ETSI standard (HIPERLAN)

We discuss these in this section in some detail.

IEEE 802.11

The IEEE standard for WLANs started in 1988 as IEEE 802.4L, a part of the IEEE 802.4 Token Bus wired LAN standard. In 1990 the IEEE 802.4L changed its name to IEEE 802.11 to form a stand alone WLAN standard in the IEEE 802 LAN standards organization. The technical aspects of this standard were completed in the year 1997. Throughout this unexpectedly long endeavour, the 802.11 group developed a framework to incorporate wireless specific issues such as power control, frequency management, roaming, and authentication in a LAN standard. The IEEE 802.11 standard was developed based on existing products in the market and so it addresses both technical and marketing issues.

The IEEE 802.11 standard [33] specifies data rates up to 2 Mbps using spread spectrum technology in the 2.4 GHz ISM bands. As wider frequency bands with no restriction on modulation become available it is possible to update the physical layer of the standard to support higher data rates and other modulation techniques. The 802.11 standard considers two network topologies: infrastructure based and ad-hoc. In an infrastructure network, mobile terminals communicate with the backbone network through an Access Point (AP). The AP is a bridge interconnecting the 802.11 network to the backbone wired infrastructure. In this configuration, a distribution system interconnects multiple Basic Service Sets (BSS) through access points to form a single infrastructure network named an Extended Service Set (ESS). A mobile terminal can roam among different BSSs in one ESS without losing the connectivity to the backbone. In ad-hoc configuration, the mobile terminals communicate with each other in an independent BSS without connectivity to the wired backbone network. In this case some of the functions of the AP, such as release of a beacon with a defined ID and timing reference, that are needed to form and maintain a BSS are provided by one of the mobile terminals. The cells in both configurations can overlap with one another. In addition to the contention based CSMA/CA access method suited for asynchronous data applications, the IEEE 802.11 also supports a contention free prioritized point coordination function (PCF) mechanism to support time bounded isochronous[1] applications. The PCF mechanism supports limited assurance for providing quality of service.

The IEEE 802.11 network uses an 83 MHz wide band divided into 79 channels. Frequency hopping is used with 22 hop patterns. This means that as data is transmitted the channel is constantly changing among the 79 choices using one of the predefined 22 patterns, with the receiver matching the changes to assemble the data [33].

The MAC services in the IEEE 802.11 support authentication, frequency management, power conservation mechanisms and encryption that are not available in the MAC layer of other 802 standards such as 802.3. Additionally, someone trying to intercept a transmission would need to duplicate the rapid channel changes that are present in 802.11 networks. These changes happen several times per second and follow one of 22 different patterns within 79 different channels. The combination of this technique with encryption makes decoding information quite difficult [34].

HIPERLAN

The HIPERLAN standard [35] was developed by the RES-10 group of the ETSI as a pan-European standard for high speed wireless local networks. The so called HIPERLAN I, the first defined technology by this standard group, was started in 1992 and completed in the year 1997. It supports data rates of 2–23 Mbps using traditional radio modulation techniques in the 5.2 GHz band. The HIPERLAN I network supports a multi-hop ad-hoc configuration. The multi-hop routing extends the HIPERLAN communication beyond the radio range of a single node. Each HIPERLAN node is either a forwarder or a non-forwarder. A non-forwarder node simply accepts the packet that is intended for it. A forwarder node retransmits the received packet, if the packet does not have its own node

[1]Isochronous pertains to processes that require timing coordination to be successful. Isochronous data transfer ensures that data flows continously and at a steady rate in close timing. Note: “Isochronous” and “anisochronous” are characteristics, while “synchronous” and “asynchronous” are relationships.

address, to other terminals in its neighborhood. Each non-forwarder node should select at least one of its neighbors as a forwarder. Inter-HIPERLAN forwarding needs bilateral cooperation and agreement between two HIPERLANs. To support routing and maintain the operation of a HIPERLAN, the forwarder and non-forwarder nodes need to periodically update six and four databases respectively. The Non-Preemptive Multiple Access (NPMA) protocol used in HIPERLAN is a listen before talk protocol that supports both asynchronous and isochronous transmissions. The HIPERLAN defines a priority scheme and a lifetime for each packet which facilitates the control of the quality of service. In addition to the routing, the MAC layer also handles the encryption and power conservation.

3.2.3 Communication Technologies

Manufacturers of wireless LANs have a range of technologies to choose from when designing a wireless LAN solution. Each technology comes with its own set of advantages and limitations.

3.2.3.1 Narrowband Technology

A narrowband radio system transmits and receives user information on a specific radio frequency. Narrowband radio keeps the radio signal frequency as narrow as possible just to pass the information. Undesirable crosstalk between communications channels is avoided by carefully coordinating different users on different channel frequencies. A private telephone line is much like a radio frequency. When each home in a neighborhood has its own private telephone line, people in one home cannot listen to calls made to other homes. In a radio system, privacy and noninterference are accomplished by the use of separate radio frequencies. The radio receiver filters out all radio signals except the ones on its designated frequency.

3.2.3.2 Spread Spectrum

Most wireless LAN systems use spread-spectrum technology [24], a wideband radio frequency technique developed by the military for use in reliable, secure, mission-critical communications systems. Spread-spectrum is designed to trade off bandwidth efficiency for reliability, integrity, and security. Spread spectrum "spreads" a signal's power over a wider band of frequencies, sacrificing bandwidth in order to gain signal-to-noise performance (referred to as process gain). This contradicts the desire to conserve frequency bandwidth, but the spreading process makes the data signal much less susceptible to electrical noise than conventional radio modulation techniques. Other transmission and electrical noise, typically narrow in bandwidth, will only interfere with a small portion of the spread spectrum signal, resulting in much less interference and less errors when the receiver

USA	Europe
900 MHz	2.4 GHz
2.4 GHz	5,15 GHz
5.7 GHz	19,3 GHz

Table 3.2: *frequencies for wireless LAN's.*

demodulates the signal. The authorized frequencies for wireless networks are specified at Table 3.2.

Each data call is assigned a unique code that permits it to be distinguished from the multitude of calls simultaneously transmitted over the same broadcast spectrum. As long as the receiving device has the right code, it can pick its conversation out from all the others.

The situation is analogous to two people conversing over dinner at a crowded restaurant in a foreign country. Although surrounded by people carrying on conversations in many different languages, these people are able to "tune out" the other conversations and understand each other because they are speaking the same language. The code used by CDMA is similar to the language shared by the two diners. It allows many conversations to occur simultaneously, while each conversation is understood only by its intended recipients.

Spread spectrum modulators use one of two methods to spread the signal over a wider area: Frequency hopping or direct sequence.

- Frequency Hopping Spread Spectrum

  Frequency hopping works very much like its name implies. It takes the data signal and modulates it with a carrier signal that hops from frequency to frequency as a function of time over a wide band of frequencies. With frequency hopping spread spectrum, the carrier frequency changes periodically. The frequency hopping technique reduces interference because an interfering signal from a narrowband system will only affect the spread spectrum signal if both are transmitting at the same frequency at the same time. Thus, the aggregate interference will be very low, resulting in little or no bit errors. A frequency hopping radio, for example, will hop the carrier frequency over the 2.4 GHz frequency band between 2.4 GHz and 2.483 GHz.

  In a frequency hopping system, a transmitter "hops" between available frequencies according to a specified algorithm, which can be either pseudo-random or preplanned. The transmitter operates in synchronization with a receiver, which remains tuned to the same center frequency as the transmitter. A short burst of data is transmitted on a narrowband. Then, the transmitter tunes to another frequency and transmits again. The receiver thus is capable of hopping its frequency over a given bandwidth several times a second, transmitting on one frequency for a certain period of time, then hopping to another frequency and transmitting again. Frequency hopping requires a much wider bandwidth than is needed to transmit the same information using only one carrier frequency. FCC regulations require manufacturers to use 75 or more frequencies per transmission channel with a maximum residence time (the time spent at a particular frequency during any single hop) of 400 ms. If the radio encounters interference on one frequency, then the radio will retransmit the signal on a subsequent hop on another frequency. Because of the nature of its modulation technique, frequency hopping can achieve up to 2 Mbps data rates. Faster data rates are susceptable to an enormous number of errors.

It is possible to operate radios use spread spectrum within the same frequency band and not interfere, assuming they each use a different hopping pattern. While one radio is transmitting at one particular frequency, the other radio is using a different frequency. A set of hopping codes that never use the same frequencies at the same time are considered orthogonal. The FCC's requirement for the number of different transmission frequencies allows frequency hopping radios to have many non-interfering channels.

When making a decision on which physical layer to use, consider the following characteristics of frequency hopping:

  - Lower cost
  - Lowest power consumption
  - Most tolerant to signal interference
  - Lowest potential data rates from individual physical layers
  - Highest aggregate capacity using multiple physical layers
  - Less range than direct sequence, but greater range than infrared

- Direct Sequence Spread Spectrum

  Direct sequence spread spectrum combines a data signal at the sending station with a higher data rate bit sequence, which many refer to as a chipping code. A high processing gain[2] increases the signals resistance to interference. The minimum linear processing gain that the FCC allows is 10, and most commercial products operate under 20. The IEEE 802.11 Working Group has set their minimum processing gain requirements at 11. In comparison to frequency hopping, direct sequence can achieve much higher than 2 Mbps data rates.

[2]processing gain = chip rate / bit rate

The data signal is multiplied by a Pseudo Random Noise Code (PN code). A PNcode is a sequence of chips valued -1 and 1 (polar) or 0 and 1 (non-polar) and has noise-like properties. This results in low cross-correlation values among the codes and the difficulty to jam or detect a data message.

In the receiver, the received signal is multiplied again by the same (synchronized) PN code. Since the code existed of +1s and -1s, this operation completely removes the code from the signal and the original data-signal is left. Another observation is that the despread operation is the same as the spread operation. The consequence is that a possible jamming-signal in the radio channel will be spread before data-detection is performed. So jamming effects are reduced.

When deciding which physical layer to use, consider the following characteristics of DSSS:

- Highest cost
- Highest power consumption
- Highest potential data rates from individual physical layers as compared to frequency hopping. (The current version of 802.11 specifies the same data rates for both frequency hopping and direct sequence; however, future versions of the standard are likely to support higher data rates for direct sequence.)
- Lower aggregate capacity using multiple physical layers than frequency hopping.
- Smallest number of geographically separate radio cells due to a limited number of channels.
- More range than frequency hopping and infrared physical layers

The main problem with applying Direct Sequence spreading is the so-called Near-Far effect. This effect is present when an interfering transmitter is much closer to the receiver than the intended transmitter. Although the cross-correlation between codes A and B is low, the correlation between the received signal from the interfering transmitter and code A can be higher than the correlation between the received signal from the intended transmitter and code A. The result is that proper data detection is not possible. Frequency-Hopping is less effected by this Near-Far effect.

In most cases, frequency hopping is the most cost-effective type of wireless LAN to deploy if needs for network bandwidth is 2 Mbps or less. Direct sequence, having higher potential data rates, would be best for bandwidth intensive applications.

3.2.3.3 Narrowband Microwave

Microwave technology is not really a LAN technology. It's main use is to interconnect LANs between buildings. This requires microwave dishes on both ends of the link. The dishes must be in line-of-sight to transmit and collect the microwave signals. Microwave is used to bypass the telephone company when connecting LANs between buildings.

One major drawback to the use of microwave technology is that the frequency band used requires licensing by the national telecommunication commissions. Once a license is granted for a particular location, that frequency band cannot be licensed to anyone else, for any purpose, within a 17.5 mile radius. Data rates range from 2 Mbps at 2.5 GHz to 274 Mbps at frequencies of 18 GHz.

3.2.3.4 Infrared Technology

Infrared LANs use infrared signals to transmit data. This is the same technology used in products like remote controls for televisions and VCRs. Infrared (IR) systems use very high frequencies, just below visible light in the electromagnetic spectrum, to carry data. Like light, IR cannot penetrate opaque objects; it is either directed (line-of-sight) or diffuse technology.

Diffuse (or reflective) IR LAN's provide moderate data rates and coverage, and are suitable for moderate-size offices, short-distance battery-oriented applications, and environments with significant amounts of radio interference. Diffuse IR LAN systems do not require line-of-sight, but cells are limited to individual rooms.

Inexpensive directed systems provide very limited range (1m) and typically are used for LANs but occasionally are used in specific WLAN applications. High performance directed IR is impractical for mobile users and is therefore used only to implement fixed subnetworks. This directed-beam IR is suitable for large-file transfers between main frames and servers as well as large open offices with many fixed terminals. Wireless LAN's at tens of gigahertz are suitable high-speed communication in large partitioned areas such as larger offices or similar open areas.

The major advantage of infrared is its ability to carry a high bandwidth, but its major disadvantage is that they can easily be obstructed, since light cannot pass through solid objects.

3.2.4 Multiple access schemes

Channel-access methods and methods for sharing a channel among multiple users are essential ingredients in achieving efficient operation and good performance in a wireless network. Users in a wireless network seldom have need to access a channel for a long period of time, and thus what is needed are schemes for providing multiuser access, usually referred to simply as multiple access, to the frequency-and-time resources of the network in an orderly manner and in a way that minimizes transmission overhead and maximizes overall network capacity. In general, the access methods fall into two categories — fixed-access or dynamic-access schemes, where the schemes in either category are fundamentally the same as those commonly used in wired networks, with some variations that are important for operation in the wireless environment. Detailed description of multiple access protocols can be found in a number of references, including [20, 19, 36].

3.2.4.1 Fixed assignment protocols

In fixed assignment protocols, the channel is statically partitioned into a fixed number of units. In Time Division Multiple Access (TDMA), sets of round-robin time slots are assigned, and in Frequency Division Multiple Access (FDMA), the frequency spectrum is partitioned into separate bands. Bandwidth is pre-allocated for a fixed set of stations data transmission. Each data unit station is identified by its position in the data stream, thus reducing control overhead and providing simplicity and efficiency. Fair channel allocation is obtained, with a separate (possibly logical) channel assigned to each station. These synchronous systems assure control over average and worst case delay for voice and video traffic. Throughput is optimal during heavy traffic when all capacity is being utilized, due to low overhead and contention-free transmission. Obviously, fixed assignment systems provide less flexibility for data traffic, as long bursts of data must be stored waiting for some channels, while other channels are unused. In particular, for non-uniform traffic conditions, capacity that is not used is wasted, while long queues develop for access to busy channels. Results from queueing theory state that systems with multiple queues for multiple servers, as in fixed assignment protocols, engender longer delays and lower utilization than single queue systems. These methods require careful synchronization with some cost in bandwidth (perhaps 30–50 microsecond guard bands in TDMA and 600Hz guard bands in FDMA). An additional disadvantage of fixed assignment schemes is that they are not scalable; new stations cannot be added dynamically to the station set — it might not be feasible to add them even if a network were brought down temporarily.

3.2.4.2 Dynamic allocation protocols

In an environment where data users are present and generate traffic in packet bursts, it is well known that circuit switching and fixed resource assignment techniques (e.g., TDMA or FDMA) may be very inefficient. For this reason, other multiple access techniques have been devised, which are more suitable for this environment. The range of possible solutions has at one extreme complete coordination among users, and at the other extreme no coordination at all. An example of complete coordination may be polling systems, where a central controller tells users when they should transmit, based on the knowledge of their communications needs. This kind of structure may be more flexible than fixed resource assignment, but requires coordination among users and controller, and in particular some signalling is needed and some overhead is involved. On the other hand, the best example of non-coordinated access is the ALOHA protocol, in which users transmit whenever they have packets, regardless of what other users are doing. Overlapping transmissions (collisions) may lead to packet losses, which are recovered by retransmission.

The idea underlying the ALOHA protocol, and in particular its extreme simplicity, has been very successful, and has originated the whole research field of random access protocols. Several modifications have been proposed to the original idea, trying to solve some problems that random access presents. The first improvement was to slot the time axis, so that packets can be transmitted only at some fixed instants of time, thereby reducing the probability of overlapping transmissions. The recognition of the fact that the throughput performance greatly depends on the occurrence of collisions led to more advanced schemes, where transmitters listen to the channel before attempting transmission, and refrain from sending packets if the channel is sensed busy. In the presence of zero signal propagation delay, this would lead to perfect sensing and would completely eliminate collisions. On the other hand, since signals take time to propagate, the sensing information may be incorrect. For example, just after a transmission is initiated a transmitter would sense the channel as idle and also start transmitting, thus causing a collision. Another improvement led to schemes where overlapping transmissions can be detected and aborted, thus minimizing the waste of communications resources caused by collisions. This last idea is implemented in the Carrier Sense Multiple Access with Collision Detection (CSMA/CD) protocol, which is the scheme used in Ethernet. Other more elaborate improvements have also been proposed, addressing other more subtle issues such as protocol stability.

Given the large amount of research performed in the 1970s and 1980s on multiple access schemes for LANs, trying to extend those protocols as they are to the wireless domain seems a reasonable thing to do. However, this gives rise to problems which are not present in wireline LANs, and which effectively prevent from using the well-established LAN protocols in the wireless environment. A typical issue goes under the name of hidden terminal problem. Basically, in wireline communications, all users are attached to a cable, and therefore they can all hear each other's signals. This condition is at the basis of the carrier sensing protocols. Unfortunately, in radio communications this is not necessarily the case. In fact, the erratic behaviour of the radio channel may not guarantee connectivity. Moreover, the fact that two mobile users A and B are within range of the same Base Station does not necessarily mean that they are within range of each other (e.g., if they are located at the extremes of a diameter of the BS's coverage region). Therefore, CSMA protocols as they are fail to work in the wireless environment, and other provisions must be made to guarantee proper operation. One possibility is the use of a busy tone, which is activated by the Base Station when some transmission is active, so that all mobiles which could potentially start communications to that Base Station (and are therefore within its range) will be informed about the channel being busy. The multiple access protocol specified in CDPD (called DSMA/CD) uses a busy/idle flag in the base-to-mobile direction to notify mobiles about the mobile-to-base channel occupancy, thus overcoming the hidden terminal problem.

In the two recent standards for wireless LANs, the IEEE 802.11 standard and the European HIPERLAN, another protocol is used, which is called CSMA with Collision Avoidance (CSMA/CA). The basic concept relies on the observation that collisions are most likely to occur at the end of a transmission, since during that transmission communications requests may have accumulated and they are likely to be colliding as soon as the channel becomes idle. To avoid this effect, additional backoff mechanisms are introduced, in order to limit the number of terminals which will eventually attempt transmission.

In IEEE 802.11, the concept of Inter-Frame Space (IFS) has been introduced. IFS is a time delay that a terminal has to wait before attempting transmission. Different lengths for IFS allow the introduction of different priorities. In particular, three IFS are specified, i.e., short IFS (SIFS) for high-priority control packets such as acknowledgements, Point-coordination-function IFS (PIFS) for time-sensitive traffic, and Distributed-coordination-function IFS (DIFS, the longest) for asynchronous traffic. If at the end of its IFS a terminal still senses the channel as idle (i.e., there is no higher priority traffic), it chooses a random time uniformly in a Contention Window (which is dynamically updated), and if after that time still no transmission is sensed, then the terminal sends its message. The hidden terminal problem can be solved by using some handshake procedure, where prior to message transmission there is a dialogue between mobile terminal and Base Station, which improves reliability at the expense of some overhead.

A more elaborate version of CSMA/CA is used in HIPERLAN. If upon message generation a terminal senses the channel as idle for a given amount of time (1700 bits in this case), then transmission can immediately take place. Otherwise, a three-phase collision avoidance mechanism is performed. In the prioritization phase there are up to 5 time slots, one for each priority class $p = 0, 1, \cdots, 4$ (with 0 the highest priority), and a terminal with priority p transmits a signal burst in slot $p + 1$. This phase terminates upon the first transmission, so that only the terminals belonging to the highest priority class will survive, whereas all the others will yield. In the elimination phase, each node transmits a signal burst for a geometrically distributed number of slots, and then listens for one more slots: if any signal is detected in that slot, the terminal drops out, and only those which chose the highest number survive. Finally, during the yield phase, each of the surviving nodes defers transmission for a geometrically distributed number of slots, and listens to the channel in the meantime, dropping out if any signal is detected. This whole procedure has as its objective to guarantee with high probability that only one terminal will eventually transmit, as it is very unlikely that collisions occur after all these random choices.

Spread Spectrum communication techniques provide a new method of multiple access, known as CDMA (Code Division Multiple Access) [24], that is proposed as an interesting alternative with respect to the traditional methods, i.e.: the well-known TDMA (Time Division Multiple Access) and FDMA (Frequency Division Multiple Access). CDMA is a dynamic allocation protocol, because the maximal number of stations is only restricted by the number of codes and the channel behaviour and not on the number of time slots(TDMA) or the number of frequency bands(FDMA). In CDMA systems, each user access the common channel by its pseudorandom spreading pattern: The PN sequence in the case of DS modulation, the hopping sequence if FH modulation is employed. Hence, the CDMA methods are distinguished in DS/CDMA and FH/CDMA. For more information about Spread Spectrum, see Section 3.2.3.2.

Chapter 4

Different network topologies

The basic topologies commonly used in local wireless networks [37] are centralized, distributed and multihop configurations, all three of which are shown in Figure 4.1.

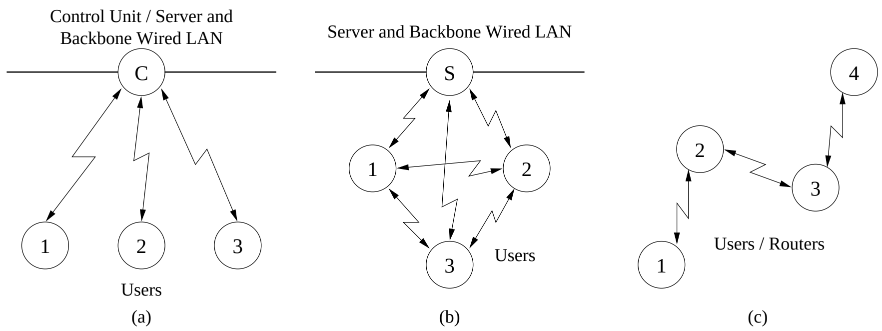

Figure 4.1: *Three common topologies in wireless networks: (a) centralized, (b) distributed and (c) multihop.*

4.1 Centralized topology

In the centralized configuration in Figure 4.1(a), one station (C) serves as a hub of the network and the user stations are located at the ends of the spokes. Any communication form one user station to another, that is, between peers, goes through the hub. The hub station controls the user stations and monitors what each user station is transmitting. Thus the hub station is involved in managing the access by user stations to the network's allocated bandwidth. There is no provision in the centralized topology for direct peer-to-peer communication. The centralized topology is a common one in the realm of telephone communications, where all the subscriber in a local exchange area are connected through the central office switch. Cellular mobile telephone systems use a centralized network configuration to serve the mobile terminals currently operating within the coverage area of any one cell site. The architecture of larger networks requires a cellular structure to provide frequency reuse and extend the coverage. The star topology for wired LAN's is another example of a centralized configuration. In the LAN industry, the Altair, ARLAN, and FreePort products use the centralized, or hub-based, network configuration. Optical wireless LAN's using IR radiation are typically cellular networks, in which each cell is served by a central base station, usually ceiling-mounted, providing coverage throughout the cell [37].

One important advantage of the centralized topology is that the network can be designed to operate with relatively efficient use of signal transmission power. For example, if compared with a fully connected peer-to-peer network, user stations in the centralized network can reach stations at twice the distance with the same signal power. A related advantage is that the hub station can be placed in a location that is optimised with respect to the set of user stations. For example, in a wireless application, the central hub might be placed in an appropriate location to facilitate unobstructed propagation between the user stations and the hub. For IR networks, the central unit is often installed on a ceiling, while for radio wireless Networks operating at frequencies on the order of 10 GHz, it is usually recommended that the central unit be located high on a wall. In the centralized configuration, the central hub also provides a natural point at which connection is made to a backbone network, as indicated in Figure 4.1(a). For this latter reason, many wireless LAN's have this centralized configuration. Another advantage of this topology is that the user terminals can be made functionally simple, which the more sophisticated control functionality is concentrated in the single hub station. The hub-based wireless LAN configuration also offers convenience in network setup, since each user node need only communicate with the hub.

Power control is an effective means of minimizing the radiated power of individual user terminals, which in turn conserves terminal power and controls interference. Where battery-operated portable units are to be used, power control serves an important function of maximizing battery life. If power control is to be employed in a radio network, a centralized configuration is necessary.

One disadvantage of the centralized topology is the existence of a single failure point. If the central control module fails, the entire network is disabled. Another disadvantage is the delay characteristic of the network. Since all user-to-user communication is passed through the hub, the store-and-forward delay is twice that of a fully connected peer-to-peer network. Another measure of communication efficiency is channel occupancy measured in hertz-seconds (bandwidth * time). This parameter measure the channel occupancy in assessing the fair sharing of a common bandwidth among multiple users. By this measure, the occupancy for store-and-forward communication between users in the centralized network is twice that of direct user-to-user communication in the fully connected network.

For limited numbers of users and none real-time systems, the single failure point, the delay and bandwidth-time usage characteristics are not as important as single coverage and the ease of connection to the backbone network. So, the centralized topology is more popular than other topologies in today's market (especially in cellular telephone networks).

4.2 Distributed topology

In a distributed network as shown in Figure 4.1(b), all terminals can communicate with one another directly. This topology is an ideal solution for the evolving ad-hoc network applications where several portable terminals such as lap-tops intend to form a network in a quickly arranged setup. The peer-to-peer capability of this topology provides an instant connectivity without any need for a centralized controller. However, in the existing wireless LAN's, distributed (peer-to-peer) network configurations are not as widely used as are hub-based configurations. Many small very-high-frequency (VHF) and ultra-high-frequency (UHF) mobile radio networks, such as those used for fleet dispatch and public safety communications, are distributed networks. However, distributed configurations are generally not used in mobile data networks.

In a distributed topology, connection to a backbone network provided by a server node that acts as a bridge or gateway. In Figure 4.1(b) the server node (S) provides the connection between the terminals and the wired backbone. An important advantage of the fully connected topology is that it has no single point of failure. An other advantage is that the peer-to-peer message do not suffer the store-and-forward delay of the centralized topology, and thus time delay and channel occupancy, as measured in hertz-seconds, are both halved.

In a distributed wireless network the routing functions must be implemented in all of the stations, and so the complexity of equipment in this design is grater then the complexity in a centralized wireless network. Another disadvantage of the fully connected peer-to-peer network is that we do not have a central unit to control the power or timing. A related issue is the presence of a near-far problem, owning to the fact that transmitters needing to operate at high power level can interfere with unintended receivers in close proximity to the transmitting station. However, the fully connected topology offers an attractive alternative for small real-time networks where reliable connectivity can be assured between all pairs of user stations.

4.3 Multihop topology

In a multihop topology, shown in Figure 4.1(c), all users can reach another via transmissions traversing multiple links in the network. This topology provides the best user-to-user connectivity of all presented topologies. A major advantage of the multihop topology is power efficiency, which derives form the fact that message transmissions between widely separated users is accomplished with multiple shorter hops. In some applications, such as military tactical communication over a wide operational area, multihop networks provide the only practical approach to reliable connectivity among mobile users. As with a fully connected network, connection to a backbone or other networks is provided by equipping one or more nodes with the appropriate connection capability.

One disadvantage of the multihop topology is the added complexity needed in user terminals to implement efficient message routing and control algorithms. Associated with the store-and-forward capability is a considerable amount of transmission overhead carried with transmitted messages. While multihop topologies have found important applications in military radio and public safety communications networks, as well as other packet radio networks, they have until today not been adopted in the wireless LAN industry.

Chapter 5

Routing in Wireless Networks

Routing [40, 41, 39, 43] is the act of moving information across a network from a source to a destination. Along the way, at least one intermediate node is typically encountered. A routing algorithm involves two main activities: Determining (optimal) routes and transporting packets throughout the network. While the latter of these, also referred to as switching in the routing context, is a relatively straightforward task in most routing protocols, path determination can be more complex and is discussed in more detail in this chapter.

Typically, a routing algorithm's goal is to achieve optimum network performance (e.g., delay and throughput) while minimizing costs (e.g., power consumption). However, the measures of performance and the costs associated with a network are not the same for every network. In fact, depending on the network's function and environment, some performance parameters are usually more important than others and some costs more expensive than others. As a result, different types of routing algorithms have emerged to serve the requirements of the different types of networks and environments. The following section presents a general classification for the existing routing algorithms.

5.1 Algorithm Types

Different factors affect the choice of a routing algorithm for a given network. Network environment, size, rate of change, and predictability of flow are some of the parameters that can play a role in the determination of the optimal routing method. Since the early study of the routing problem, a vast number of routing algorithms have been developed. All of them, however, can be classified using three key differentiators: adaptability, degree of centralization, and connectivity.

5.1.1 Adaptability: Static versus Dynamic

One major classification for routing algorithms is based on the response of the routing protocol to changes in the state of the network (e.g, load change, topology change, etc).

At one extreme are static routing algorithms whose routing remains fixed, independent of the current state of the network. Algorithms that use static routing are usually simple to design and are suitable for small networks with relatively predictable traffic and known topologies. In fact, given knowledge of links capacity, expected flow, and network topology, designing a static routing procedure for a given network becomes a straightforward task. This type of routing design problem is also known as Flow-Based Routing.

As simple as it is for small predictable networks, static routing design can be extremely difficult for large size networks. Moreover static routing can be highly inefficient and could even become impossible if it were to be used in a network with changing traffic or topology.

Fortunately, a different class of routing algorithms exists for this type of networks. Dynamic routing algorithms change their routing decisions based on the current state information of the network. The state information can be estimated or measured, and usually consists of information used by the routing algorithm to determine optimal routes (e.g, number of hops between nodes, average link delay, transmission power, etc).

As a price for their adaptive aspect, dynamic algorithms require much more computational overhead than equivalent static algorithms. Nonetheless, they still are the most dominant algorithms for wireless networks.

5.1.2 Degree of centralization: Centralized versus Distributed

Another classification for routing algorithms, which is mainly but not exclusively applicable to dynamic routing, is the way the algorithm divides its computational process (calculating and determining optimal routes) among the nodes of the network.

At one end, there is the centralized approach where all computations are done at one or few central nodes and then communicated to the other nodes of the network. The only remaining job for the non-central nodes is then to forward the packets as indicated in their routing tables.

This method might work for networks with light loads, but has some major drawbacks in highly loaded networks or networks with changing topology. A possible scenario could be one of a highly dynamic network where the computational resources of the central nodes cannot keep up with the changes in the network causing a slower response to state changes. As a result, the network nodes may end up using old state information or network topology. An even worse scenario can occur in the case of the failure of one or more of the central nodes (single point of failure !!). In that situation, the nodes that depend on these central nodes for routing information will keep on using old or even non-existing routes until the central nodes come back on again.

Fortunately, most of these problems can be usually prevented by adopting a distributed routing algorithm. In a distributed system, the computation of routes is shared among all the nodes of the network. Typically, each node determines its own routes to every other node in its region which sometimes can be the entire network.

It is true that distributed algorithms require some computational power for every node in the network, but unlike centralized algorithms, they are fast to react to network changes and less prone to failures since the routing functionality is spread over the entire network. To capture these advantages, the majority of modern communication networks adopt distributed protocols, encouraged by the decreasing costs of computational power due to the recent advances in that technology.

5.1.3 Connectivity: Flat versus Hierarchical

A third classification for routing algorithms can be based on network connectivity.

Some routing algorithms operate in a flat space, while others use routing hierarchies. In flat-routed networks, all the nodes are peers of all others and communication between any two nodes is allowed by the protocol when the propagation conditions allow it. At the other end, hierarchical systems usually have at least two levels; on the lower-level networks, nodes in the same geographical region use a flat-routing protocol for their communication. In addition, each lower-level network has at least one node to serve as a gateway node to the higher levels. These gateway nodes form the higher-level network. Whenever a packet needs to travel from one lower-level network to another, it has to go through at least one gateway node. Also, it is not uncommon to have a gateway node be shared between two or more lower-level networks.

In comparing the two approaches, we first note that a hierarchical routing system would require at least two different routing procedures depending on the number of hierarchies in the system. On the other hand, routing algorithms in hierarchical networks can be simplified greatly due to the fact that each node needs to only know the topology information of its own hierarchical zone. This, in turn, can reduce the amount of routing information traffic that needs to be transmitted across the network.

However, as in centralized systems, the performance of hierarchical systems can suffer greatly in the case of the failure of one or more gateway nodes. Furthermore, as more work is usually demanded from the gateway nodes than the other nodes, their power consumption is much higher. In fact, "pure" hierarchical routing would very likely fail in environments where power is a scarce resource (e.g, microsensor networks).

Although many existing networks actually use "pure" hierarchical networks, a lot of the modern networks, especially highly dynamic ones, tend to slide to the other side of the spectrum or use some kind of hybrid algorithm.

5.2 Design goals

Different networks operate in different environments and serve different goals, and consequently have different performance priorities. For some, short routes might be the top priority, while for others, it might be fast packet delivery. The following section explains some of the main performance parameters and design goals.

Throughput: Throughput is the measure of the amount of data transmitted over time. Increasing the network's throughput while keeping the average delay low, is often one of the main design goals in most routing algorithms.

In addition to selecting optimal routes, there are other factors that can help increase the throughput of a given communication network. Results in [38] indicate that coordination and scheduling among neighboring nodes can help achieve that goal.

On a different level, it is important to maintain the link capacities high enough to be able to handle the network's throughput. This problem becomes especially interesting in wireless communication due to the broadcast nature of the environment. According to Shannon's theorem, the capacity C of a communication channel is bounded by a function of the average signal power S, the average interfering noise power N, and the bandwidth W:

$$C \leq W \log_2(1 + \frac{S}{N})$$

As, in most cases, the bandwidth W is fixed, increasing the capacity turns into a mixed power control-interference management problem, which will be returned to in more detail in the context of wireless networks in Section 5.5.

Average packet delay: As a packet travels from a source to a destination, it usually traverses one or more intermediate nodes. The total packet delay is the accumulated delay over all the links traversed by the packet. At each node, the delay is composed of the processing delay (time between the moment the packet is fully received and the time it is queued out), the queueing delay (time spent inside the outgoing queue before transmission), the transmission delay (time between the transmission of the first bit and the transmission of the last bit of the packet), and the propagation delay (time between the transmission of the last bit in the packet and its reception at the other end node).

A good routing algorithm, through congestion control, works on maintaining constant average packet delay under high traffic conditions, and decreasing it under low traffic conditions. Furthermore, the routes selected by the routing algorithm directly affect the packet delay. In fact, routes with shorter distances and less hops incur less packet delay. However, this condition alone cannot be used as the rule for route selection, since routes with longer hops (less hops) have less capacity due to noise, and consequently less feasible throughput. This is especially a problem to consider in the wireless environment.

Energy-efficiency: Energy-efficiency is one of the main design goals in a routing protocol, and could become a top priority in energy-constrained environments. Nodes in a network consume power for communication as well as computation purposes. As computation is much cheaper energy-wise than communication [42], the latter tends to be the essence of this energy optimization. Apart from efficiently reducing the communication overhead of the protocol, there are many other factors to consider depending on the network environment. In fact, energy-efficiency is not a one-dimensional problem and there is not a unique approach to tackle it. This design problem will be returned to in more detail in the discussion on wireless networks in Section 5.5.

Optimal routes: Optimality refers to the capability of the routing algorithm to select the best routes based on the routing metrics. In many cases, there exist more than one route between a given source and destination. The optimal route is the route that minimizes a cost function. For example, where energy is the metric used, the optimal route is the one that requires the minimum energy consumption. Section 5.3 discusses the most common metrics used in algorithms, as well as well-known methods to calculate optimal routes.

Convergence: Routes in communication networks, especially in dynamic ones, are often subject to changes or failures. Most routing algorithms use some sort of routing update process to inform the nodes of such events, which prompts re-calculation of new routes. Convergence is the process of agreement by all nodes on optimal routes after a routing update process. A good routing algorithm design must insure rapid convergence, especially in highly changing networks. Slow convergence can cause routing loops or even network disconnection.

Communication overhead: The communication overhead of a routing protocol consists of all transmissions required for the operation of the algorithm. It generally includes routing updates, control information, and receipt acknowledgments when supported by the protocol. As bandwidth is in most cases a scarce resource, it is important that the communication overhead of a protocol be as efficient as possible leaving enough bandwidth for data packets.

Robustness: Routing algorithms must be robust, which means that they should perform correctly in face of unusual or unexpected conditions. Such conditions may include high loads, node or link failure, or even infrastructure damage. In general, dynamic, distributed, and flat-routed algorithms are more robust than static, centralized, and hierarchical algorithms respectively.

This is clearly the case when high loads are the problematic condition: Dynamic algorithms can adapt to changing conditions, while static algorithms run on expected flows and cannot change their routing procedures. Similarly, link or node failure can affect centralized algorithms much more than distributed algorithms. This is especially true when the failure occurs at a central node. Finally, the complexity of the infrastructure in a hierarchical system is more complex than in a flat-routed system and therefore the effect of an infrastructural damage can be larger in the former type.

The ultimate goal in that respect is a rapidly deployable network that does not require the preexistence of any fixed infrastructure, i.e., an ad-hoc network.

5.3 Routing metrics

A metric is a standard of measurement that is used by the routing algorithm to determine the optimal route to a destination. The metric associated with a given path contains information about the desirability of that path based on some performance parameter (e.g, delay). In most routing algorithms however, metrics are not assigned directly to the routes but to the individual links. A given route's metric is then calculated by adding the metrics of its component links.

Routing algorithms have used many different metrics to determine the best routes. It is not uncommon, however, that multiple metrics are combined with different weightings to form a single (hybrid) metric or cost function. One routing algorithm, for example, may use the number of hops and delays, but may weight delay more heavily. Naturally, the network's goal together with its environment, play an important role in the determination of the metrics. In what follows, some of the most commonly used metrics are presented:

Path length: Path length is the most common routing metric. Some routing algorithms compute the path length as the sum of the costs associated with each link traversed. Other routing protocols define it as the number of hops from source to destination. Either way, the importance of this metric comes from the fact that it encompasses other factors such as delay, energy, throughput. For example, if the path length is defined in terms of the hop count, a route with more hops generally implies a route that requires a longer delay.

Delay: Route delay refers to the length of time required to move a packet from source to destination. Delay depends on many factors, including the bandwidth and the congestion of the intermediate links, the queue lengths at nodes along the path, as well as the physical distance to be traveled. Being an aggregate of several important variables, makes path delay a common and useful metric.

Bandwidth: Bandwidth refers to the available traffic capacity of a link. All other conditions being equal, a 10-Mbps link would be preferable to a 64-kbps line. Although bandwidth is a rating of the maximum achievable throughput on a link, routes using links with greater bandwidth do not necessarily provide better routes than routes using slower links. For example, it is quite possible that the delay over a fast but busy link is greater than the delay over a slower link.

Bit-Energy: Bit-energy refers to the amount of energy needed to transmit a bit over a given link. Several factors affect bit-energy, including physical distance, propagation conditions, and the geography of the environment. In energy-constrained environments, optimal routes might be those that require least energy to traverse. Minimum-energy routing, presented in [38], is one example of algorithms that use the energy metric.

Load: Load refers to the degree to which a network node is busy. It can be calculated in a variety of ways, such as CPU utilization at the node, packets processed per time, or queue (incoming and/or outgoing) length. A routing protocol taking load in consideration opt for routes with less loaded intermediate nodes.

Reliability: In the context of routing, the reliability of a link refers to how dependable the link is. When assigned to a given link, a reliability metric can take into account one or more factors, including the link's bit-error rate, the number of packets lost over the link, or the number of failures of the link's end nodes.

5.4 Dynamic routing solutions

The demand for large, scalable, changing networks in the past decade necessitated the development of routing algorithms capable of servicing the needs of such networks. Clearly, protocols based on static routing could not provide the solution, and research moved towards other types of algorithms. The result were dynamic protocols that adapt to changes in the state (topology and traffic) of the network.

This section presents two different classes of dynamic protocols. While the first constitutes the conventional approach in dynamic routing, the second is a non-traditional one, which deserves special attention particularly in the wireless environment.

5.4.1 Conventional solutions: Proactive protocols

The fundamental concept of conventional dynamic routing is that every node maintains up-to-date information on how to optimally reach any other destination in the network. For that purpose, every node in the network continuously monitors changes in topology and traffic, and informs other nodes of such events using periodical or event-driven route updates. Consequently, every node in the network will always have the most current picture of the network topology, and thus will be able to determine the optimal routes at any time. The conventional approach is the basis of several dynamic algorithms. Here is a brief description of some of its main features:

- Nodes maintain routing tables, which contain information on how to reach, at any time, every other destination in the network.
- Nodes continuously monitor state information, and adapt to it by changing their routing decisions. Furthermore, nodes communicate their routing information with other nodes in the network either periodically or when changes take place. Conventional algorithms differ, however, in what each node needs to monitor, and with who it needs to communicate its routing information. Yet, most of them belong to one of the following major approaches:
    - Distance vector: By the distance vector approach, each node maintains a table giving the distance (metric) from itself to all other destinations it is aware of. Each node periodically broadcasts this information to each of its neighboring nodes, and uses similar broadcasts from its neighbors to update the values of its own table. By comparing the distances received for each destination from each of its neighbors and knowing its distances to its neighbors, a node can determine the neighbor through which the best route can be obtained to any destination. In fact, this is all the information a node needs to maintain in its routing table, also known as the destination/next hop information.
    - Link state: By the link state approach, each node maintains a complete picture of the topology of the entire network. Each node monitors the cost of the links to each of its neighbors, and periodically broadcasts an update (usually using flooding) of this information to all other nodes in the network. Given this information of the cost of each link in the network, each router can compute the best route to each possible destination. Unlike distance vector, in link state the node knows the complete path to the destination although it does not necessarily need to maintain it in its routing table.

  In general, link state protocols converge more quickly than distance vector, but require more computational power and network bandwidth.
- Major advantages of proactive protocols:
    - Routes are immediately available when needed.

– Every node has an up-to-date picture of the network topology.

- Major disadvantages of proactive protocols:

  – Substantial amounts of bandwidth and power are used for large update information, which may not be used.

  – In highly changing networks, proactive protocols suffer from slow convergence, particularly with distance vector algorithms.

5.4.1.1 Count-to-infinity Problem

All distance vector protocols are susceptible to the well known count-to-infinity problem.

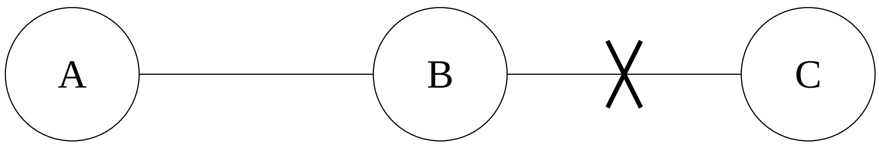

Figure 5.1: *Count-to-infinity Problem.*

In Figure 5.1, node B will store a distance of 1-hop to get to C and A will store a distance of 2-hops to get to C.

The count to infinity problem occurs when the link to node C crashes. Node B notices the timeout, and assumes that the cost to get to C is infinity. However, node A tells B that it can get to C in a cost of 2 (but A does not know, that this requires going through the broken B to C link that is now infinity). So now B thinks it can get to C in a cost of 3 (the cost to get to A plus the cost for A to get to C). B now tells A that it can get to C in a cost of 3. A then updates its table and the cost for A to C is now 4. A tells this to B, who updates his cost to C to 5. The process will repeat forever, with nodes A and B counting up to infinity.

5.4.2 Non-conventional solutions: Reactive protocols

The increasing interest in wireless mobile networks led to the development of new routing approaches to satisfy the requirements of such networks.

Although conventional protocols have been applied in these types of environments, they suffered from serious problems. In fact, the considerable amounts of bandwidth and power required for routing updates posed a big problem in the wireless environment due to the scarcity of resources. Furthermore, it was very inefficient to have each node maintain routes to every other node in the network, when it might possibly never attempt to use a large number of them. Furthermore it is neither wise nor possible to keep monitoring the network topology when it is changing at rates larger than the rate of convergence of the protocol.

The newly developed approach, also known as reactive, directly tackled the problems faced by the conventional approach, but fell into problems of its own as will be discussed later. Reactive protocols discarded the concept of continuous monitoring of state information. Consequently, the resource problems associated with routing updates were solved. In addition, since the nodes did not need to receive routing updates anymore, considerable power could be saved by allowing nodes to go into sleep mode when they are not part of any communication. Furthermore, the protocol does not need to chase the fast changing network topology anymore.

Reactive protocols are based on two main ideas: On-demand route discovery and route maintenance [39]. In general, the operation of reactive protocols is as described below (node names are only used for simplification):

- Whenever source node A is presented with a routing decision to destination node D, it runs the algorithm below:

    1. Checks its routing table for an existing valid route to the destination. If one is found, then it is used. Otherwise it continues with step 2.
    2. When no existing route is found, the route discovery procedure is launched by A. The route discovery process consists of a global search algorithm that works as follows:
        - Node A floods a route query packet on the network.
        - Every node X (other than the destination) that receives the query either responds back with a route to the destination if it has one in its routing table and stops forwarding the packet, or forwards the packet once again after recording its address in it. As a result, every forwarded packet will have the addresses of all the intermediate node that it traversed.
        - If no route is found to the destination the query will continue to be forwarded until it reaches the neighbors of the destination (one hop away). These nodes are usually aware of the destination, and have a one hop route to it. Eventually, a route will be found.
        - Once a node B detects that it has a route to D (the destination), it has to inform A. First, it stops forwarding the query packet further. It then examines the query packet, and forms the return route to A by reversing the order of the intermediate node addresses recorded in the query. Finally, it builds a reply packet containing the full route from itself to D and sends it on the route just formed.

    - The reply packet will contain its complete route from B to A. Every node C receiving it will only need to forward it to the next hop. However, C will also be able to see the route from B to D contained in the packet. Using that and the route of the reply, C will be able to calculate routes from itself to both B and D. If C did not have any existing valid routes in its routing table, then the routes just learned are added. If other better routes already existed in C's routing table, then the new routes are disregarded. Similarly, any intermediate node flooding the query packet can learn new routes to the source. This way, intermediate nodes will be able to benefit from forwarding packets.

  3. Once the source node receives the reply packet, it forms its new route to the destination and stores it into its routing table. It can then use that route immediately to that destination.

- Route maintenance: After a predetermined amount of time, a discovered route expires, and when needed again it will have to be rediscovered. This is a precautionary step used by reactive protocols to prevent the use of invalid routes. Additionally, some reactive protocols might rely on hop-by-hop acknowledgements to detect invalid links on an old route. More interestingly however, protocols used in wireless networks take advantage of the broadcast feature of the environment. To detect whether a link is still working after sending a packet on it, a node would try to listen to the other end-node, further forwarding the packet. An acknowledgment however, would still be needed for the last hop on the route.

The main advantages and disadvantages of the reactive approach can be summarized as follows:

- Major advantages:
  - Routes are determined on-demand only.
  - Route updates are eliminated: no bandwidth or power is consumed sending information that might not be needed.
  - Power-saving schemes can be used by allowing nodes to go into sleep mode.

- Major disadvantages:
  - Large amounts of bandwidth and power are consumed by the global search algorithm (because it uses flooding).
  - Nodes have to wait for routes to be discovered, adding additional packet delays.

5.4.3 Conclusion

The choice between the proactive and the reactive approach for a given network is not an easy one. As mentioned earlier, the new reactive method did in effect solve the major performance issues of the conventional protocols, but paid its price by considerable problems of its own. In dealing with that choice though, it is important to consider the "dynamism" required of the network and the expected amount of traffic through it.

In highly dynamic networks, it might be very inefficient to try to keep track of the topology. Therefore, a reactive approach might be preferable. On the other hand, a network with high traffic would be too costly for a reactive protocol, due to the enormous costs (power and bandwidth) associated with every single route discovery. Additionally, as far as delay is concerned, the designer must determine which will be a greater problem, the convergence of a proactive protocol or the route discovery delay of a reactive one.

In the end, it might be safe to claim that neither a pure reactive nor a pure proactive protocol can be too efficient in a highly-loaded, highly dynamic environment. Better solutions will probably be found using hybrid (combination of the two) protocols or even new approaches.

5.5 Wireless routing protocols

5.5.1 Routing issues in wireless networks

In this section, some of the main characteristics of routing in a wireless environment are briefly considered.

Several characteristics distinguish routing in the wireless environment from routing in traditional wired networks. Scarce resources, interference between transmissions, unreliable channels, and non-uniform propagation are only some of the difficulties encountered in such networks. These issues greatly affect major decisions in a routing algorithm by creating multi-dimensional tradeoffs between the various performance criteria (delay, throughput, energy, etc).

Consider, for example, the tradeoffs associated with the decision on route selection. In a multi-hop network, the transmission of a packet can be achieved by direct transmission from source to destination (no intermediate hops), by using few hops, or by stopping at all intermediate nodes.

If direct transmission is used, more power will be needed on each transmission due to longer hops. It will also mean greater co-channel interference, since the transmission radius of a given node will now be larger. So, while higher signal power is helping the maximum throughput on a given link by improving the SNR (signal-to-noise ratio), it is at the same time causing higher co-channel interference decreasing the SNR, and consequently deteriorating throughput.

On another level, while direct transmission results in more power consumption for the source node, it conserves the energy of the intermediate nodes by not having them participate in receiving and forwarding packets. Determining which routing decision is more energy-efficient requires the knowledge of the number of hops in the multi-hop route, and the individual transmission and receive energy of each node on the path. In general, the receive energy is lower than the transmission energy, but when taken to be on the same order, simulations in [42] have shown that direct transmission will be more energy-efficient than multi-hop routing. It is always the case, that in any given network an optimal number of hops exists for which minimum energy consumption is achieved.

Returning to feasible throughput, it is inherently true that short hops have higher capacity than longer ones, and consequently afford more throughput. As a result, the more the hops in a route, the more the throughput it can handle since routes with more hops usually mean routes with shorter hops. On the other hand, more hops means that packets need to be received and re-transmitted at each node, resulting in higher delays.

These are some of the tradeoffs associated with a particular routing decision in the wireless environment. Many other situations exist in a routing protocol, where such tradeoffs need to be carefully considered. In conclusion, a good routing algorithm for the wireless environment should strike a balance between throughput and packet delay, while using the energy and bandwidth resourceswisely.

In the remaining part of this chapter we present two groups of ad-hoc routing protocols. We start with non location-aware routing protocols. Those protocols are based on a graph model, in the sense that a mobile host only knows the connectivity relation with respect to its neighbors, but not its relative location to its neighbors [43]. In the second subsection, we discuss location-aware routing protocols, which are based on a geographic model. It is assumed that a mobile host knows its current physical location (see Chapter 2 for details about common location techniques and location systems), such that location information can be exploited to facilitate routing.

5.5.2 Non location-aware routing protocols

5.5.2.1 Destination-Sequenced Distance-Vector Routing Algorithm (DSDV)

Destination-Sequenced Distance-Vector Routing [44] is an adaptation of a conventional routing protocol to ad-hoc networks. DSDV is based on the Distributed Bellman-Ford (DBF) Routing Algorithm. Consequently, DSDV only makes use of bi-directional links. DSDV is one of the earlier ad-hoc routing protocols developed.

In DSDV, packets are routed between nodes of an ad-hoc network using routing tables stored at each node. Each routing table, at each node, contains a list of the addresses of every other node in the network. Along with each node's address, the table contains the address of the next hop for a packet to take in order to reach the node.

The bigger part of the DSDV protocol does not involve routing at all. Rather, the problem of DSDV is the generation and maintenance of the routing tables. Every time the network topology changes, the routing table in every node needs to be updated. As one might expect, this is no trivial task. The situation is further complicated by the fact that when routing tables are out of synchronization (i.e., the routing protocol has not converged), routing loops may form. To facilitate routing table maintenance, several additional pieces of information are stored in the routing tables. In addition to the destination address and next hop address, routing tables maintain the route metric and the route sequence number.

Periodically, or immediately when network topology changes are detected, each node will broadcast a routing table update packet. The update packet starts out with a metric of one. This signifies to each receiving neighbor they are one hop away from the node. The neighbors will increment this metric (in this case, to two) and then retransmit the update packet. This process repeats until every node in the network has received a copy of the update packet with a corresponding metric. If a node receives duplicate update packets, the node will only pay attention to the update packet with the smallest metric and ignore the rest. To distinguish stale update packets from valid ones, the original node tags each update packet with a sequence number. The sequence number is a monotonically increasing number that uniquely identifies each update packet from a given node. Consequently, if a node receives an update packet from another node, the sequence number must be equal to or greater than the sequence number already in the routing table; otherwise the update packet is stale and ignored. If the sequence number matches the sequence number in the routing table, then the metric is compared and updated as previously discussed. Update packets with higher sequence numbers are always entered into the routing table, regardless of whether they have a higher metric or not.

Each time an update packet is forwarded, the packet not only contains the address of the eventual destination, but it also contains the address of the transmitting node. The address of the transmitting node is entered into the routing table as the next hop.

Each node must periodically transmit its entire routing table to its neighbors using update packets. The neighbors will update their tables based on this information, if required. Likewise each node will listen to its neighbor's update packets and update its own routing table.

The routing table updates can be sent in two ways. One, called *full dump*, will carry all the available routing information and can span many packets. The other type, called *incremental*, will carry only information changed since the last dump and must fit in one network protocol data unit (NPDU). *Full dumps* can be transmitted relatively infrequently when no movement of mobile hosts is occurring. When movement becomes frequent, and the size of an *incremental* approaches the size of NPDU, then a *full dump* can be sent (so that the next *incremental* will be smaller).

It is conceibably, that a mobile host always receive two routes to the same destination, with the same newer sequence number. The host receive the routes one after another (via different neighbors), but always get the route with the worse metric first. Unless care is taken, this will lead to a continuing burst of new route transmissions upon every new sequence number from the destination. A solution is to delay the advertisement of such routes when a mobile host can determine that a route with a better metric is likely to show up soon. The route with the later sequence number must be available for use, but it does not have to be advertised immediately unless it is a route to a destination which was previously unreachable. Thus, there will be two routing tables kept at each mobile host; one for use with forwarding packets, and another to be advertised via incremental routing information packets. To determine the probability of imminent arrival of routing information showing a better metric, the mobile host has to keep a history of the weighted average time that routes to a particular destination fluctuate until the route with the best metric is received. A parameter must be selected which indicates how long a route has to remain stable before it is counted as truly stable.

Mobile nodes cause broken links as they move from place to place. The broken link may be detected by the communication hardware, or it may be inferred if no broadcasts have been received for a while from a former neighbor. A broken link is described as having a metric of infinity. Since this qualifies as a substantial route change, the detecting node will broadcast an update message for the lost destination. This update message will have a new sequence number and a metric of infinity. This will essentially cause the routing table entries for the lost node to be flushed from the network. Routes to the lost node will be established again when the lost node sends out its next broadcast.

Building information to describe broken links is the only situation where the sequence number is generated by a node other than the destination. To avoid that nodes and their neighbors generate conflicting sequence numbers when the topology changes, nodes only generate even sequence numbers for themselves, and neighbors responding to link changes only generate odd sequence numbers.

When the number of nodes in the network grows, the size of the routing tables and the bandwidth required also grows. Furthermore, whenever the topology changes, DSDV is unstable until update packets propagate throughout the network. DSDV also has a lot of troubles in dealing with high node mobility.

5.5.2.2 Clusterhead Gateway Switch Routing (CGSR)

The Clusterhead Gateway Switch Routing (CGSR) protocol differs from the previous protocol in the type of addressing and network organization scheme employed. Instead of a flat network, CGSR is a clustered multihop mobile wireless network with several heuristic routing schemes [45]. The authors state that by having a cluster head controlling a group of ad-hoc nodes, a framework for code separation (among clusters), channel access, routing and bandwidth allocation can be achieved. A cluster head selection algorithm is utilized to elect a node as the cluster head using a distributed algorithm within the cluster. All nodes within the transmission range of the cluster head belong to this cluster, i.e., all nodes in a cluster can communicate with a cluster head and possibly with each other. The disadvantage of having a cluster head scheme is that frequent cluster head changes can adversely affect routing protocol performance since nodes are busy with cluster head selection rather than packet relaying. Hence, instead of invoking cluster head reselection every time the cluster membership changes, a Least Cluster Change (LCC) clustering algorithm is introduced. Using LCC, only two conditions cause the cluster head to change. One is when two cluster heads come within range of each other, and the other is when a node becomes disconnected from all clusters. This is an improvement over existing algorithms which select the cluster head every time the cluster membership changes.

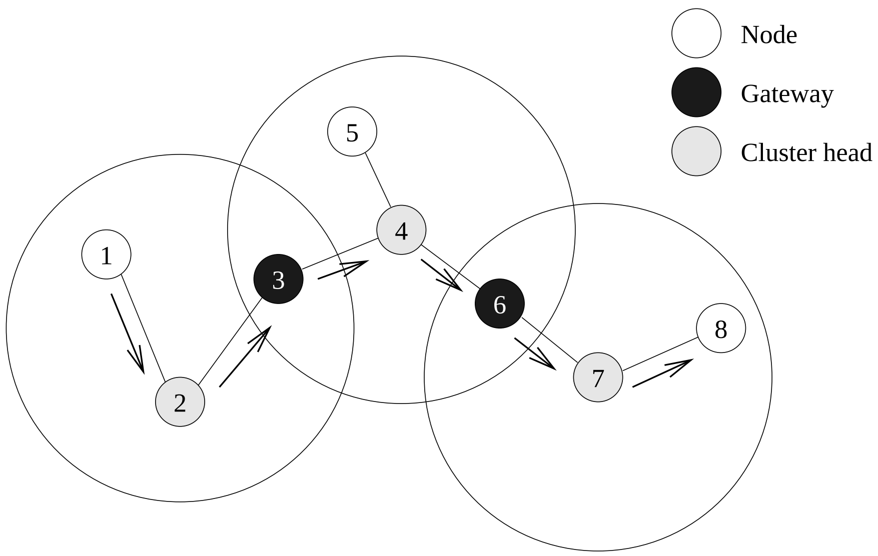

Figure 5.2: *CGSR: Routing from Node 1 to Node 8.*

CGSR uses DSDV as the underlying routing scheme. It modifies DSDV by using a hierarchical cluster head-to-gateway routing approach to route traffic from source to destination. Gateway nodes are nodes that are within communication range of two or more cluster heads. A packet sent by a node is first routed to its cluster head, and then the packet is routed from the cluster head to a gateway to another cluster head, and so on until the cluster head of the destination node is reached. The packet is then transmitted to the destination. Figure 5.2 illustrates an example of this routing scheme. Using this method, each node must keep a *cluster member table* where it stores the destination cluster head for each mobile node in the network. These cluster member tables are broadcast by each node periodically using the DSDV algorithm. Nodes update their cluster member tables on the reception of such a table from a neighbor.

In addition to the cluster member table, each node must also maintain a routing table, which is used to determine the next hop in order to reach the destination. On receiving a packet, a node will consult its cluster member table and routing table to determine the nearest cluster head along the route to the destination. Next the node will check its routing table to determine the next node in order to reach the selected cluster head. It then transmits the packet to this node.

5.5.2.3 The Wireless Routing Protocol (WRP)

The Wireless Routing Protocol (WRP) described in [46] is a table-based protocol with the goal of maintaining routing information among all nodes in the network. Each node in the network is responsible for maintaining four tables:

- distance table
- routing table
- link-cost table
- message retransmission list (MRL) table

The distance table of a node x contains the distance of each destination node y via each neighbor z of x. It also contains the predecessor (second-to-last hop) of z through which this path is realized. Using its distance table, a node creates a routing table by choosing the neighbor which offers the smallest cost to a given destination when used as the next hop. The routing table of node x contains the distance of each destination node y from node x, the predecessor and the successor of node x on this path. It also contains a tag to identify if the entry is a simple path, a loop or invalid. Storing predecessor and successor in the table is beneficial in detecting loops and avoiding count-to-infinity problems (Subsection 5.4.1.1 Count-to-infinity Problem). The link-cost table contains the cost of the link to each neighbor of the node and the number of timeouts since an error-free message was received from that neighbor. Each entry of the MRL contains the sequence number of the update message, a retransmission counter, an acknowledgement-required flag vector with one entry per neighbor, and a list of updates sent in the update message. The MRL records which updates in an update message need to be retransmitted and which neighbors should acknowledge the retransmission [46].

Mobiles inform each other of link changes through the use of update messages. An update message is sent only between neighboring nodes and contains a list of updates (the destination, the distance to the destination, and the predecessor of the destination), as well as a list of responses indicating which mobiles should acknowledge (ACK) the update. Mobiles send update messages after processing updates from neighbors or after detecting a change in a link to a neighbor. In the event of the loss of a link between two nodes, the nodes send update messages to their neighbors. The neighbors then update their distance table entries and check for new possible paths through other nodes. Any new paths are relayed back to the original nodes so that they can update their tables accordingly. On receiving an ACK, the node updates its MRL. A unique feature of this algorithm is that it checks the consistency of all its neighbors every time it detects a change in the link of any of its neighbors. The consistency check helps eliminate looping situations in a better way and also has fast convergence.

Nodes learn of the existence of their neighbors through the reception of acknowledgements and other messages. If a node is not sending messages, it must send a hello message within a specified time period to ensure connectivity. Otherwise, the lack of messages from the node indicates the failure of that link and may cause a false alarm. When a mobile receives a hello message from a new node, that new node is added to the mobile's routing table, and the mobile sends the new node a copy of its routing table information.

Part of the novelty of WRP stems from the way in which it achieves loop freedom. In WRP, routing nodes communicate the distance and second-to-last hop information for each destination in the wireless network. WRP belongs to the class of path finding algorithms with an important exception. It avoids the "count-to-infinity" problem by forcing each node to perform consistency checks of predecessor information reported by all its neighbors. This ultimately (though not instantaneously) eliminates looping situations and provides faster route convergence when a link failure event occurs.

In WRP, as in all proactive protocols, each node constantly maintains full routing information to all destinations in the network, requiring significant communication overhead.

5.5.2.4 Ad-hoc On Demand Distance Vector Routing Algorithm (AODV)

The Ad-hoc On Demand Distance Vector (AODV) routing algorithm [47] is a reactive routing protocol for ad-hoc mobile networks. It is capable of both unicast and multicast routing. It is an on demand algorithm, meaning that it builds routes between nodes only as desired by source nodes. It maintains these routes as long as they are needed by the sources. Additionally, AODV forms trees which connect multicast group members. The trees are composed of the group members and the nodes needed to connect the members.

AODV builds routes using a route request / route reply query cycle. When a source node desires a route to a destination for which it does not already have a route, it broadcasts a route request (RREQ) packet across the network. Nodes receiving this packet update their information for the source node and set up backward pointers to the source node in their route tables. In addition to the source node's IP address, current sequence number, and broadcast ID, the RREQ also contains the most recent sequence number for the destination of which the source node is aware.

The source sequence number is used to maintain freshness information about the reverse route to the source. This is to make sure that packets from the destination to the source follow a path, which was set up by the route request packet for this session. The destination sequence number specifies how fresh a route to the destination must be before it can be accepted by the source.

A node receiving the RREQ may send a route reply (RREP) if it is either the destination or if it has a route to the destination with corresponding sequence number greater than or equal to that contained in the RREQ. If this is the case, it unicasts a RREP back to the source. Otherwise, it rebroadcasts the RREQ. Nodes keep track of the RREQ's source IP address and broadcast ID. If they receive a RREQ which they have already processed, they discard the RREQ and do not forward it.

As the RREP propagates back to the source, nodes set up forward pointers to the destination. Once the source node receives the RREP, it may begin to forward data packets to the destination. If the source later receives a RREP containing a greater sequence number or the same sequence number with a smaller hop count, it may update its routing information for that destination and begin using the better route.

As long as the route remains active, it will continue to be maintained. A route is considered active as long as there are data packets periodically traveling from the source to the destination along that path. Once the source stops sending data packets, the links will time out and eventually be deleted from the intermediate node routing tables. If a link break occurs while the route is active, the node upstream (the node on the source site) of the break propagates a route error (RERR) message to the source node to inform it of the now unreachable destination(s). After receiving the RERR, if the source node still desires the route, it can reinitiate route discovery.

Multicast routes are set up in a similar manner. A node wishing to join a multicast group broadcasts a RREQ with the destination IP address set to that of the multicast group and with the "J"(join) flag set to indicate that it would like to join the group. Any node receiving this RREQ that is a member of the multicast tree with a fresh enough sequence number for the multicast group may send a RREP. As the RREPs propagate back to the source, the nodes forwarding the message set up pointers in their multicast route tables. As the source node receives the RREPs, it keeps track of the route with the freshest sequence number and the smallest hop count to the next multicast group member. After the specified discovery period, the source node will unicast a Multicast Activation (MACT) message to its selected next hop. This message serves the purpose of activating the route. A node that does not receive this message and has set up a multicast route pointer will timeout and delete the pointer. If the node receiving the MACT was not already a part of the multicast tree, it will also have been keeping track of the best route from the RREPs it received. Hence it must also unicast a MACT to its next hop, and so on until a node that was previously a member of the multicast tree is reached.

AODV maintains routes for as long as the route is active. This includes maintaining a multicast tree for the life of the multicast group. It uses bandwidth efficiently (by minimizing the network load for control and data traffic), is responsive to changes in topology, is scalable and ensures loop free routing.

5.5.2.5 Dynamic Source Routing Protocol (DSR)

DSR [48] is a source initiated on-demand routing protocol. Each mobile host participating in the ad-hoc network maintains a route cache in which it caches source routes that it has learned.

When a source node desires a route to a destination for which it does not already have a route in its cache, it broadcasts a route request packet across the network. In addition to the address of the original initiator of the request and the target of the request, each route request packet contains a route record in which is accumulated a record of the sequence of hops taken by the route request packet as it propagated through the network. Each route request packet also contains a unique identification number (request id), set by the initiator from a locally maintained sequence number. An intermediate node discards the request when it has recently seen the route request with the same initiator address and the same request id. This ensures that an intermediate node removes later copies of the request that arrive at it by a different route. An intermediate node also discards the request if the node's address is already listed in the route record to guarantee that no single copy of the request can propagate around a loop.

If an intermediate node has a cache route to the destination then it will discard the request and will send the route reply back to the source. Otherwise, the destination will send the route reply. As the route request packet propagates through the network, it records the route record. If the route reply is generated by the destination, then it places the route record from the route request into the route reply packet. On the other hand, if the intermediate node generates the route reply, then it appends its cached route to the destination to the route record in the route request packet and puts that into the route reply packet.

To return the route reply, the responding node must have a route to the initiator. If it has a route to the initiator in its route cache, it may use that route. Otherwise, if symmetric links are supported, the node may reverse the route in the route record. If symmetric links are not supported, the node must initiate its own route discovery and piggyback the route reply on the new route request.

The route maintenance procedure monitors the operation of the route and informs the sender of any routing errors. A host detects transmission of corrupted or lost packets by means of passive acknowledgements, i.e., after sending the packet to next hop, the sender tries to hear the host transmitting the packet again. If the route breaks due to a link failure or any other such problem, then the detecting host sends a route error packet to the source. When a route error packet is received, the hop in error is removed from this host's cache, and all the routes which contain this hop, must be deleted

One tradeoff between source routing and distance vector routing is the handling of partitioned networks. Under dynamic source routing, if a host wishes to communicate with an unreachable host, then the rate at which route request are made will be reduced by a back off mechanism, but the protocol continues to make periodic efforts to find a route to the unreachable host, consuming some network resources. With distance vector routing, under the assumption that the nodes know that a particular host became unreachable, no network resources are used for trying to send packets to the unreachable host, as none of the hosts in the sender's partition of the network have a routing table entry for the destination.

DSR has the advantage that no routing tables must be kept at the nodes to route a given packet, since the entire route is contained in the packet header. The primary disadvantages are two-fold. DSR is not scalable to large networks and the large packet header requires significantly more bandwidth than most other protocols.

5.5.2.6 Temporally Ordered Routing Algorithm (TORA)

Temporally Ordered Routing Algorithm (TORA) [49] is a distributed routing protocol based on a "link reversal" algorithm. It is designed to discover routes on demand, provide multiple routes to a destination, establish routes quickly, and minimize communication overhead by localizing the reaction to topological changes when possible. Route optimality (shortest-path routing) is considered of secondary importance, and longer routes are often used to avoid the overhead of discovering newer routes. It is also not necessary (nor desirable) to maintain routes between every source/destination pair at all times.

The actions taken by TORA can be described in terms of water flowing downhill towards a destination node through a network of tubes that model the routing state of the network. The tubes represent links between nodes in the network, the junctions of the tubes represent the nodes, and the water in the tubes represents the packets flowing towards the destination. Each node has a height with respect to the destination that is computed by the routing protocol. If a tube between two nodes becomes blocked such that water can no longer flow through it, the height of the nodes are set to a height greater than that of any neighboring nodes, such that water will now flow back out of the blocked tube and find an alternate path to the destination.

A logically separate version of TORA is run for each destination to which routing is required. When a node needs a route to a particular destination, it broadcasts a "route query packet" containing the address of the destination. This packet propagates through the network until it reaches either the destination or an intermediate node having a route to the destination. The recipient of the query packet then broadcasts an "update packet" listing its height with respect to the destination (if the recipient is the destination, this height is 0). As this packet propagates back through the network, each node that receives the update sets its height to a value greater than the height of the neighbor from which the update was received. This has the effect of creating a series of directed links from the original sender of the query to the node that initially generated the update.

When a node loses its last downstream link due to a link failure, it selects a new height (global maximum) by defining a new reference level. This action results in link reversals which may cause other nodes to lose their last downstream link. Any such node executes a partial reversal by choosing a height which is smaller than the height of the origination node but greater than the height of all other nodes. This propagation will only extend through nodes which have lost all routes to the destination (thus propagation is limited). Any node which prior to the initial link failure had only downstream links and now experiences link reversals from all its neighbors selects a local maximum height by reflecting back the reference level (by setting the reflection bit). This node could either be the source node or a part of the network which is partitioned or both. If this reflected level is propagated back to the originating node from all of its neighbors, then no route to the destination exists. The originating node (which generated the new reference level) has then detected a new partition and can begin the process of erasing the invalid routes.

An advantage of TORA is that it supports multiple routes to any source/destination pair. Failure or removal of one node is quickly resolved without source intervention by switching to an alternate route. Unfortunately, there are drawbacks to TORA as well. The biggest is that it relies on synchronized clocks among nodes in the network. While external time sources are available (GPS for example), it makes the hardware more costly and introduces a single point of failure if the time source becomes unavailable. Since it uses internodal co-ordination it exhibits instability behavior similar to the "count-to-infinity" problem of distance vector routing protocols. There is a potential for oscillations, especially when multiple sets of coordinating nodes are concurrently detecting partitions, erasing routes, and building new routes based on each other. However such oscillations are temporary and route convergence will ultimately occur.

5.5.2.7 Associativity-Based Routing (ABR)

A totally different approach in mobile routing is proposed in [50]. The Associativity-Based Routing (ABR) protocol is free from loops, deadlocks, and packet duplicates, and defines a new routing metric for ad-hoc mobile networks. This metric is known as the degree of association stability. In ABR, a route is selected based on the degree of association stability of mobile nodes. Each node periodically generates a beacon to signify its existence. When received by neighboring nodes, this beacon causes their associativity tables to be updated. For each beacon received, the associativity tick of the current node with respect to the beaconing node is incremented. Association stability is defined by connection stability of one node with respect to another node over time and space. A high degree of association stability may indicate a low state of node mobility, while a low degree may indicate a high state of node mobility. Associativity ticks are reset when the neighbors of a node or the node itself move out of proximity. A fundamental objective of ABR is to derive longer-lived routes for ad-hoc mobile networks.

The three phases of ABR are:

- Route discovery
- Route reconstruction (RRC)
- Route deletion

The route discovery phase is accomplished by a broadcast-query and await-reply cycle (BQ-REPLY). A node desiring a route broadcasts a BQ message in search of mobiles that have a route to the destination. All nodes receiving the query that are not the destination append their addresses and their associativity ticks with their neighbors along with QoS information to the query packet. A successor node erases its upstream node neighbors' associativity tick entries and retains only the entry concerned with itself and its upstream node. In this way, each resultant packet arriving at the destination will contain the associativity ticks of the nodes along the route to the destination. The destination is then able to select the best route by examining the associativity ticks along each of the paths. When multiple paths have the same overall degree of association stability, the route with the minimum number of hops is selected. The destination then sends a REPLY packet back to the source along this path. Nodes propagating the reply mark their routes as valid. All other routes remain inactive, and the possibility of duplicate packets arriving at the destination is avoided.

RRC may consist of partial route discovery, invalid route erasure, valid route updates, and new route discovery, depending on which node(s) along the route move. Movement by the source results in a new BQ-REPLY process. If the destination moves, the immediate upstream node erases its route and determines if the node is still reachable by a localized query (LQ[H]) process, where H refers to the hop count from the upstream node to the destination. When the destination receives the LQ packet, it REPLYs with the best partial route. If no response is received from the destination, a time-out occurs and the partial route discovery will be delegated to the next upstream node. If this process results in backtracking more than halfway to the source, the LQ process is discontinued and a new BQ process is initiated at the source. If any intermediate node fails then the upstream node performs the localized query and downstream the node sends a notification for all the nodes downstream to delete invalid routes.

After communication, the source initiates a route delete broadcast which cause all the nodes up to the destination to delete the route entry from their tables. It is propagated by a full broadcast since the source may not know of any route change that occurred during the route reconstruction phase. Even though many nodes may move concurrently, the protocol ensures that only one route reconstruction cycle is valid.

In ABR, the route selection is primarily based on the aggregated associativity ticks of nodes along the path. Hence, although the resulting path does not necessarily result in the smallest possible number of hops, the path tends to be longer-lived than other routes. A long-lived route requires fewer route reconstructions and therefore yields higher throughput. Another benefit of ABR is that, like the other protocols, it is guaranteed to be free of packet duplicates. The reason is that only the best route is marked valid, while all other possible routes remain passive. ABR, however, relies on the fact that each node is beaconing periodically. The beaconing interval must be short enough to accurately reflect the spatial, temporal, and connectivity state of the mobile hosts. This beaconing requirement may result in additional power consumption. When a route is broken due to an intermediate link failure, the backtracking done in partial route discovery could entail a high latency, though at times it could provide quicker route maintenance. So, there is a trade-off involved. In case of partitioned networks, it will take a lot of time for a source to ultimately detect it and will also result in waste of resources (bandwidth).

5.5.2.8 Signal Stability Algorithm (SSA)

Signal Stability Algorithm (SSA) [51] is basically an ABR with the additional property of selecting routes based on signal strength of the link. This route selection criteria has the effect of choosing routes that have "stronger" connectivity. Functionally, the SSA protocol consists of two protocols: the Forwarding Protocol (FP) and the Dynamic Routing Protocol (DRP), which utilize the extended device driver interface. This interface is responsible for making the signal strength information from the device available to the routing protocols.

The Forwarding Protocol (FP) and the Dynamic Routing Protocol (DRP) work together to route packets in the ad-hoc network. Each node maintains two tables, the Signal Stability Table (SST), and the Routing Table (RT).

Each host sends out a link layer beacon ("I am alive" messages which are exchanged between wireless devices) at regular intervals to maintain connectivity. The receiving neighbors record the signal strength at which the beacon is received (signal stability) and also the number of times it is received (location stability). Based on this a link is assigned a status of strong or weak channel. The DRP uses the SST to maintain routes to neighboring hosts in the RT and the non-neighbor routes via information provided by route-search, route-reply, error and erase messages.

The FP forwards a packet to the next hop by looking in the RT. If no entry exists for that destination then the FP initiates a route-search to find a route to the destination. This packet has a hop-list to record the path taken by the message. Each intermediate DRP uses this list to prevent loops and adds its own address to the hop-list. Route-search packets arriving at the destination have necessarily arrived on a path of strong signal stability since the packets arriving over a weak channel are dropped at intermediate nodes. The destination chooses the route recorded in the first arriving request, since this route is probably shorter and less jammed than routes for slower arriving requests. The destination returns the route reply along the selected route, and each intermediate node includes the entry next-hop and destination in its routing table. If the source times out before receiving a reply then it changes the header of the request to indicate that weak channels are acceptable, since these may be the only links over which the packet can be forwarded.

A route may become unavailable due to migration of the hosts along the route. When a host moves out of range of its neighbors or shuts down, the neighbors will recognize that the host is unreachable since they no longer receive beacons from the host. The DRP will modify the SST and RT to reflect the changes. Any routes that have this unreachable host as the next hop will become invalid. When the host receives a packet to forward along an invalid route, the FP will determine the lack of a route and will notify the source via an error message. The source FP will initiate a new route discovery to find an available route, and it will send a message to erase the invalid route.

The SSR algorithm is a logical descendant of ABR. It utilizes a new technique of selecting routes based on the signal strength and location stability of nodes along the path. As in ABR, while the paths selected by this algorithm are not necessarily shortest in terms of hop count, they be more stable and longer-lived, resulting in fewer route reconstructions. One of the major drawbacks of the SSR protocol is that intermediate nodes cannot reply to route requests sent towards a destination; this results in potentially long delays before a route can be discovered. Additionally, when a link failure occurs along a path, the route discovery algorithm must be reinvoked from the source to find a new path to the destination. Unlike ABR, no attempt is made to use partial route recovery (unlike ABR), that is, to allow intermediate nodes to attempt to rebuild the route themselves.

5.5.2.9 Zone Routing Protocol (ZRP)

The Zone Routing Protocol [52] attempts to balance the tradeoff between proactive and reactive routing using a dynamically maintained zone topology. The objective of ZRP is to make ad-hoc routing both scalable and robust in the face of significant node mobility, without saturating the network with routing update traffic or excessive route setup latency. ZRP makes several key contributions to ad-hoc routing, most important of which is the specification of a novel technique for improving the efficiency of the reactive route-search process using a hybrid routing scheme based upon proactive routing zones. Routing within a zone is managed by a proactive routing protocol, whereas routing to destinations beyond a source node's routing zone is achieved via a reactive routing protocol. The two routing mechanisms are referred to as the Intrazone Routing Protocol (IARP) and Interzone Routing Protocol (IERP), respectively.

Each node in an ad-hoc network maintains its own routing zone, a cluster of nodes which can be reached along paths that are less than or equal to r_{zone}. The parameter r_{zone} is a predefined value which is measured in hops. Since each node is the center of its own routing zone, the zone topology consists of a set of fully overlapping dynamic clusters. The zone organization does not reflect any measure of topological stability, and although zone membership is maintained dynamically, zone size is not adaptive either to node mobility, transmission characteristics or network traffic.

Peripheral nodes refer to nodes whose minimum distance from the node in question is the zone radius. Nodes use their peripheral nodes when issuing a network-wide search for a route. By targeting queries to peripheral nodes, route requests tend to diverge away from the source, thereby reducing redundancy.

The strategy of using peripheral nodes for reactive route discoveries is called bordercasting. Bordercasting can be implemented as a series of unicast transmissions or as a single multicast transmission.

As the zone radius approaches the radius of the network, ZRP behaves like a proactive routing protocol. At the other extreme, where the zone radius is one, ZRP behaves like a reactive protocol. When the zone radius lies in between these two extremes, ZRP merges the benefits of each to provide a robust, efficient, hybrid protocol.

IARP is the proactive component of ZRP. Nodes proactively maintain routing information for all nodes within their zone. The routing information learned for each node within the zone includes its identity and the minimum distance to it. IARP can be implemented as either a Distance Vector or Link State algorithm. The essential feature of IARP is that its control packets are restricted to the zone radius. The overhead of proactively maintaining routing information in ZRP varies with the zone radius, not with the size of the network.

IERP is the reactive component of ZRP. It is used to discover routes to destinations outside of a node's zone. IERP can be implemented as a reactive Distance Vector algorithm, where nodes cache route information and packet routing is performed on a hop-by-hop basis. Alternatively, it can require no caching, in which case it relies on source routing. When a node needs to send packets to a destination, it first checks to see if the destination is in its own zone. If so, then it already knows the route (through IARP) and delivers them accordingly. If the destination is not within the same zone, then it bordercasts a route query to all of its border nodes. If the destination is not in the border node's zone, the border node adds its ID to the request and re-bordercasts it. When the request reaches a border node whose zone includes the destination, it sends a route reply to the source indicating the route to the destination.

IERP must maintain route information during the propagation of route requests that the discovering node may transmit the route reply message back to the source node. The route reply message must also contain the newly discovered route from the source to the destination. Each node may place its address into the header of the route request packet as it propagates towards the destination. The discovering node may then reverse the route to send the route reply packet to the original source. If nodes do not cache any route information, strict source routing must be used. If some route information is cached, then loose source routing may be used. If all nodes cache route information, then next-hop routing may be used.

To reduce redundant route request propagation through zones that have already been queried, ZRP may incorporate a query detection and early termination mechanism. In such a scheme, nodes keep track of the queries they hear on the network and record query source address and unique identifier (i.e., sequence number) in a Detected Queries table. If a redundant query enters a region, then the detecting node may drop the packet and discontinue its propagation in that area.

Route failure is detected proactively, in conjunction with the IARP. Failures may be repaired locally, in which case it may not even be necessary to inform the source node. If necessary, a hop-limited local request can be used to repair the route, or a route error message can be sent to re-initiate the route discovery at the source.

An advantage of ZRP is that information about topological changes are propagated only to the neighborhood, as opposed to a fully proactive method which would flood the entire network when there is a change in topology occurred. The major drawback is that a change in topology will affect several routing zones. Furthermore a bordercast is more expansive than the broadcast flooding used in other reactive protocols. Nodes generally have many more border nodes than neighbors; in addition, each bordercast message has to traverse zone-radius hops to the border.

5.5.3 Location-aware routing protocols

5.5.3.1 Location-Aided Routing Algorithm (LAR)

The LAR protocol [53] is an approach to utilize location information for mobile hosts with the goal of decreasing routing-related overhead. The basic idea of LAR is that routing discovery cost can be dramatically reduced by limiting the search space for desired route to a destination. LAR is basically the same as flooding algorithms, but with some modifications. The modification at the source is to define a limited search space, named "request zone", so that only nodes in the request zone are allowed to forward a route request message to their neighbors. The intended destination is modified to reply with its location information so that this information can be used for a future route discovery at the source. With these small modifications on flooding algorithms, LAR results in much lower route discovery overhead in wireless ad-hoc networks.

In the LAR approach two algorithms exist that differ in how they define the request zone: LAR scheme 1 and LAR scheme 2. As an example, let us consider LAR scheme 1 in the case that a source host S needs to find a route to a destination host D at time t_1. Figure 5.3 illustrates the basic LAR scheme 1. If we assume that host S knows about the previous location L of host D at time t_0 ($t_1 \geq t_0$), then S may be able to determine some region (the so-called "expected zone") in which D is expected to exist at time t_1. Based on this expected zone, S can define the request zone for the route discovery phase. In LAR scheme 1, the request zone is the smallest rectangle including the current location of the source and the expected zone for the destination. The sides of the rectangular request zone are parallel to the X and Y axes (see Figure 5.3). Any intermediate host receiving the route request message can make a decision whether to forward it or not by using the explicitly specified request zone on the route request. For instance, if node I receives the route request from another node, node I forwards the request to its neighbors because I determines that it is within the rectangular request zone. However, when node J receives the route request, node J discards the request, as node J is not within the request zone.

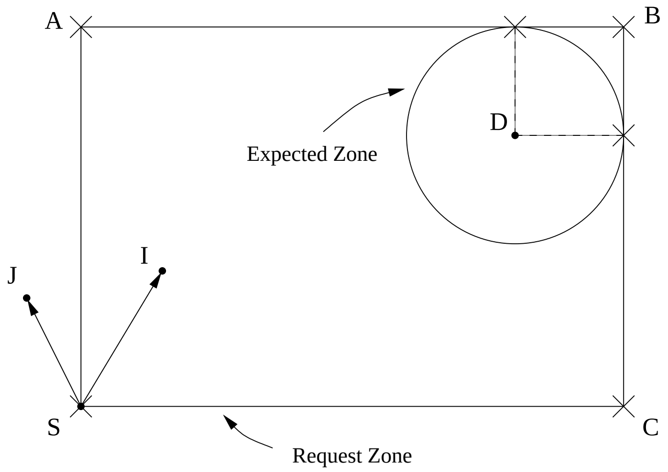

Figure 5.3: *An example of LAR Scheme 1.*

When node D receives the route request message, it replies by sending a route reply message (as in the flooding algorithm). However, in case of LAR, node D includes its current location and current time in the route reply message. When node S receives this route reply message (ending its route discovery), it records the location of node D. Node S can use this information to determine the request zone for a future route discovery. The request zone in the LAR scheme 1 is not modified by any intermediate hosts, whereas that in the LAR scheme 2 is adapted as the route request packet is propagated to various nodes [53].

5.5.3.2 Geographic Distance Routing Algorithm (GEDIR)

Lin and Stojmenovic introduced the Geographic Distance Routing (GEDIR) algorithm [55]. When node A wants to send a message m to node D, it uses the location information for D and for all its 1-hop neighbors to determine the neighbor C that is closest to D among all neighbors of A. The message is forwarded to C, and the same procedure is repeated until D, if possible, is eventually reached. In this basic GEDIR method, a node will stop forwarding the message if the best choice for it is the node from which the message arrived. Such a node is referred to as a concave node. The GEDIR algorithm is inherently loop-free [55]. The proof is based on the observation that the distances of nodes towards the destination are decreasing.

The 2-hop variant of the GEDIR algorithm was also proposed in [55]. The delivery rate of the GEDIR algorithm can be improved if each node is aware of its 2-hop neighbors (i.e., the neighbors of its neighbors). In this case, the node A currently holding the message may choose the node C that is closest to the destination among all its direct (1-hop) and 2-hop neighbors. A then forwards the message to C if it is a 1-hop neighbor, or to its neighbor B that is connected to C if C is a 2-hop neighbor of A. If their is more than one neighbor connected to C, A chooses the one that is closest to the destination. Maybe B find a better choice than C in the next step of the routing process. There are no multiple copies of the message in the network at any transmission step by the 2-hop method.

5.5.3.3 GRID Protocol

The GRID protocol's core concept calls for partitioning the geographic area into several squares in Euclidean space called grids [54]. Each host still has a unique ID (such as the IP address or the MAC address) and each host knows its current location by some location technique (see Chapter 2). Given any physical location, there should be a predefined mapping from the location to its grid coordinates.

Routing is performed in a grid-by-grid manner. In each grid, one host will be elected as the gateway of the grid. The responsibility of gateway hosts include: (i) forwarding route discovery requests to neighboring grids, (ii) propagating data packets to neighboring grids, and (iii) maintaining routes which pass the grid. All non-gateway hosts are not responsible for these jobs unless they are destinations of (i) or (ii) or sources/destinations of (iii). For maintaining the quality of routes, the gateway host of a grid should be the one nearest to the physical center of the grid.

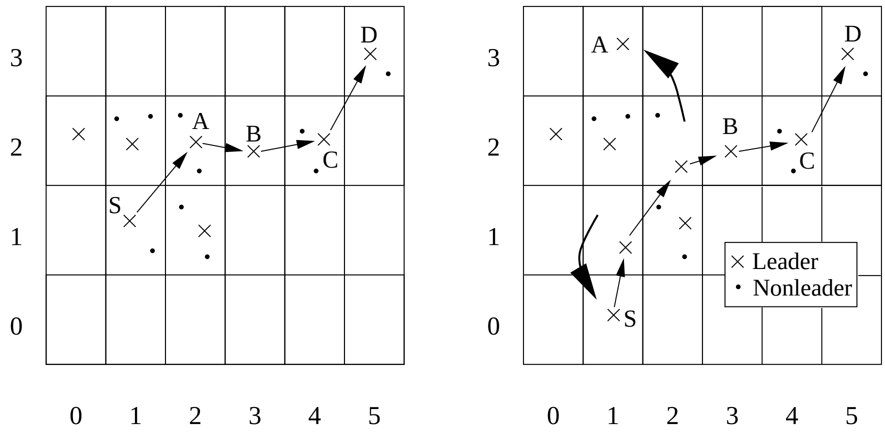

Figure 5.4: *A routing example in the GRID protocol.*

In the GRID protocol a grid ID, rather than a host ID, represents a route. Specifically, each entry in a routing table records the next grid that will lead to the destination. For example, in Figure 5.4a host A will register grid 3,2 instead of B's address as the next hop to reach D. Because routing tables use grid IDs, they offer stronger, more resilient route maintenance. In Figure 5.4b, when host A roams away, the nodes elect another host in grid 2,2 as the new leader to take over the packet-relaying job. This is a big advantage to other routing protocols such as DSR [48], ZRP [52], AODV [47], and LAR [53], where the route breaks if an intermediate host in a route roams outside its neighbors' range. Longer routes are especially vulnerable to this. However, with the GRID protocol, even if S roams out of its original grid, the route can persist. For example, in Figure 5.4b, when S moves to grid 1,0, another host in grid 1,1 can serve as the new leader.

The maintenance jobs this protocol requires include leader election and routing table handoff. In each grid, the hosts must run a leader election protocol to maintain their leader. Also, when a leader roams off its original grid, a simple handoff procedure, implemented via broadcast, must take place to pass the routing table to the next leader. GRID can maintain a much longer route life than other protocols while remaining less vulnerable to host mobility. Because it identifies leaders and nonleaders, GRID can handle high volumes of control traffic while remaining insensitive to host density, making it scalable to larger networks.

5.5.3.4 Zone-Based Hierarchical Link State Routing Protocol

This Zone-Based routing protocol [56] is (like ZRP) a hybrid protocol which exhibits both reactive and proactive behavior. In this scheme the network is divided into non-overlapping zones. A node knows its physical position by some location technique and thus can determine its zone ID by mapping its physical location to a zone map, which has to be worked out at design stage.

Each node receives two types of link state packets (LSPs): node LSPs and zone LSPs. The node LSP of a particular node contains a list of its connected neighbors and is propagated locally within its zone. The zone LSP contains a list of its connected zones and is propagated globally throughout the network.

Each node asynchronously broadcasts a link request and nodes within its communication range in turn reply with link responses containing the node ID and the zone ID. After all link responses are received, the node generates its node LSP that contains the node IDs of its neighbors in the same zone and the zone IDs of its neighbors in different zones. The node LSP is then broadcast locally throughout the zone. However, node LSPs from other zones will not be stored. After receiving all node LSPs, each node will know the node level topology of that zone.

The nodes which receive link responses from the nodes of their neighbor zones are called gateway nodes. Since node LSPs contain the zone IDs of the connected zones, each node will know which zones are connected to its zone. This information is propagated to all the nodes within the zone and each node of the same zone broadcasts the same zone LSP. Care is taken so that the gateway nodes do not forward duplicate copies of zone LSPs.

So, each node has full knowledge of the node connectivity of the nodes in its zone and only zone connectivity information about other zones in the network. If the destination is within the zone then the source can send the packet specifying the node ID only. If the destination is outside the zone, then its zone ID is found by unicasting one location request to each zone. The destination sends back (unicasts) its zone ID. Routing is then done by specifying zone ID and node ID.

It is possible that more than one cluster exists within a zone due to a large obstacle such as a hill which blocks radio communication. Every node will receive two zone LSPs from the same zone. To identify them, one additional field, the smallest node ID, is added to the zone LSP. Thus, now a node may receive two zone LSPs with different zone connectivity information from same the zone, i.e., The zone is effectively split into two sub-zones.

5.6 Gossiping

Gossiping and broadcasting are two problems of information dissemination described for a group of individuals connected by a communication network. In gossiping, every person in the network knows a unique item of information and needs to communicate it to everyone else. In broadcasting, one individual has an item of information which needs to be communicated to everyone else [57].

A popular formulation assumes there are n people, each one of whom knows a scandal which is not known to any of the others. They communicate by telephone, and whenever two people place a call, they pass on to each other as many scandals as they know. How many calls are needed before everyone knows about all the scandals? Denoting the scandal-spreaders as A, B, C, D, a solution for $n = 4$ is given by $\{A, B\}$, $\{C, D\}$, $\{A, C\}$, $\{B, D\}$. The $n = 4$ solution can then be generalized to $n < 4$ by adding the pair $\{A, X\}$ to the beginning and end of the previous solution, i.e., $\{A, E\}$, $\{A, B\}$, $\{C, D\}$, $\{A, C\}$, $\{B, D\}$, $\{A, E\}$.

Let $f(n)$ be the number of minimum calls necessary to complete gossiping among n people, where any pair of people may call each other. Then $f(1) = 0$, $f(2) = 1$, $f(3) = 3$ and $f(n) = 2n - 4$ for $n \geq 4$.

In the case of one-way communication ("polarized telephones"), e.g., where communication is done by letters or telegrams, the graph becomes a directed graph and the minimum number of calls becomes $f(n) = 2n - 2$ for $n \geq 4$.

Gossiping (which is also called total exchange or all-to-all communication) was originally introduced in discrete mathematics as a combinatorial problem in graph theory, but it also has applications in communications and distributed memory multiprocessor systems.

Gossip-based algorithms (e.g. [58, 59, 60]) have recently attained popularity in the research community as a technique for solving several large-scale distributed computing problems.

5.6.1 Gossip-Based Ad-hoc Routing

Due to the changing and high variability nature of ad-hoc networks, routing traffic an be a significant overhead and impediment to application performance. Many ad-hoc routing protocols are based on (some variant of) flooding. Despite various optimizations, many routing messages are propagated unnecessarily. In [61], Haas et al. propose a gossiping-based approach to reduce the overhead of the routing protocols. The major fingings are as follows. In large networks, there is a threshold effect: In almost every execution of the gossiping protocol, almost every node gets the message or almost no node does. The fraction of executions in which almost every node gets the message depends on the gossiping propability and the topology of the network. In the networks they have considered, using gossiping probability between 0.6 and 0.8 suffices to ensure that almost every node gets the message in almost every execution. For large networks, a simple gossiping protocol uses up to 35% fewer messages than flooding, with improved performance to routing protocols without gossiping. Gossiping can also be combined with various optimizations of flooding to yield further benefits. The simulations in [61] show that adding gossiping to AODV results in significant performance improvement, even in networks as small as 150 nodes. The improvement should be even more significant in larger networks.

Vahdat and Becker [62] apply gossiping to ad-hoc unicast routing. In their work, they try to ensure that messages are eventually delivered even if there is no connected path between the source and the destination at any given point in time. As long as there exists a path using communication links at some point in time, messages can be delivered through random point-wise exchanges among mobile hosts.

Chandra et al. [63] use a gossiping mechanism to improve multicast reliability in ad-hoc networks; they do not use gossiping to reduce the number of messages sent. Indeed, they start with an arbitrary, possibly unreliable, multicast protocol to multicast a message. They then use gossiping (under the assumption that routes are known) to randomly exchange messages between nodes in order to recover lost messages.

Bibliography

[1] J. Hightower, G. Borriello
Location systems for ubiquitous computing
IEEE Computer magazine, Volume: 34 Issue: 8, Page(s): 57–66, Aug. 2001

[2] J. Hightower, G. Borriello
Location sensing techniques
http://portolano.cs.washington.edu/papers/UW–CSE–01–07–01.pdf

[3] B. Brumitt, J. Krumm, B. Meyers, S. Shafer
Ubiquitous computing and the role of geometry
IEEE Personal Communications, Volume: 7, Issue: 5, Page(s): 41–43, Oct. 2000

[4] I.A. Getting
Perspective/navigation-The Global Positioning System
IEEE Spectrum, Volume: 30, Issue: 12, December 1993, Page(s): 36–38, 43–47

[5] GARMIN: What is GPS?
http://www.garmin.com/aboutGPS/

[6] Trimble - All About GPS
http://www.trimble.com/gps/

[7] Peter H. Dana
Global Positioning System Overview
http://www.colorado.edu/geography/gcraft/notes/gps/gps.html

[8] R. Want, A. Hopper, V. Falcao, J. Gibbons
The active badge location system
ACM Transactions on Information Systems, Volume: 10, Page(s) 91–102, Jan. 1992.

[9] Andy Harter, Andy Hopper, Pete Steggles, Andy Ward, Paul Webster
The anatomy of a context-aware application
In Proceedings of the Fifth Annual ACM/IEEE International Conference

on Mobile Computing and Networking, Page(s) 59–68, Seattle, WA, August 1999. ACM Press.

[10] Nissanka B. Priyantha, Anit Chakraborty, Hari Balakrishnan
The Cricket location-support system
In Proceedings of the Sixth Annual ACM International Conference on Mobile Computing and Networking, Boston, MA, August 2000. ACM Press.

[11] P. Bahl, V. N. Padmanabhan
Radar: An in-building rf-based user location and tracking system
In Proceedings of the IEEE Infocom 2000, Tel-Aviv, Israel, Volume: 2, Mar. 2000, Page(s) 775–784.

[12] Paramvir Bahl, Venkata N. Padmanabhan, Anand Balachandran
Enhancements to the RADAR User Location and Tracking System
http://www.research.microsoft.com/~padmanab/papers/msr-tr-2000-12.ps

[13] F. Rabb, E. Blood, R. Steiner, H. Jones
Magnetic position and orientation tracking system
IEEE Transaction on Aerospace and Electronic Systems, 15(5), Page(s) 709–718, September 1979.

[14] Technical Description of DC Magnetic Trackers
Ascension Technology Corporation, Burlington, VT 2001
http://www.5dt.com/downloads/fobtechnicalspec.pdf

[15] L. Doherty, L. El Ghaoui, K. S. J. Pister
Convex Position Estimation in Wireless Sensor Networks
Infocom 2001, Anchorage, AK, April 2001.

[16] N. Bulusu, J. Heidemann, D. Estrin
GPS-less LowCost Outdoor Localization for Very Small Devices
IEEE Personal Communications Magazine, Volume: 7, Number: 5, Oct 2000, Page(s) 28–34.

[17] Dragos Niculescu and Badri Nath
Ad-hoc Positioning System
Technical Report DCS-TR-435, Rutgers University, April 2001

[18] Specification of the Bluetooth System
http://www.bluetooth.com

[19] Kaveh Pahlavan, Allen H. Levesque
Wireless Data Communications
Proceedings of the IEEE, Volume: 82, Number: 9, Pages(s) 1398–1430, September 1994

[20] Antonie DeSimone, Sanjiv Nanda
Wireless Data: Systems, Standards, Services
Wireless Networks, Volume: 1, Number: 3, Page(s) 241–253, 1995.

[21] A. Joshi, S. Weerawarana, R. A. Weerasinghe, T. T. Drashansky, N. Ramakrishnan, E. N. Houstis
A Survey of Mobile Computing Technologies and Applications
Tech. Report TR-95-050, Dept. of Computer Sciences, Purdue University, October 1995.

[22] Matthias Nicola
Wireless Information Systems
O. Spaniol (ed.): Forschungsbeiträge des Aachener Graduiertenkollegs "Informatik und Technik", Augustinus, Aachen, Page(s) 157–176, 1996.

[23] Dan Duchamp
Issues in wireless mobile computing
Proceedings of the Third Workshop on Workstation Operating Systems, Page(s): 2 –10, 1992

[24] Alois M.J. Goiser
Handbuch der Spread-Spectrum Technik (in German)
ISBN 3-211-83080-4, SpringerTechnik.

[25] J. Ioanidis, D. Duchamp, G. Q. Maguire
IP-based protocols for mobile internetworking
Proceedings of ACM SIGCOMM 1991, Page(s) 235–245, September 1991.

[26] Tomasz Imielinski, B.R. Badrinath
Data Management for Mobile Computing
SIGMOD Record, Volume: 22, Number: 1, Page(s) 34–39, 1993.

[27] Chaoying Ma
On building very large naming systems
In 5th SIGOPS Workshop on Models and Paradigms for Distributed Systems Structuring, September 1992.

[28] F. Teraoka, M. Tokoro
Host Migration Transparency in IP Networks: The VIP Approach
SIGCOMM, Volume: 23, Number: 1, January 1993, Page(s) 45–65.

[29] M. Rahnema
Overview of the GSM System and Protocol Architecture
IEEE Communication Magazine, Volume: 31, Number: 4, Page(s) 92–100, April 1993

[30] M. Kojo, K. Raatikainen, T. Alanko
Connecting Mobile Workstations to the Internet over a Digital Cellular Telephone Network
In Proceedings of the Workshop on Mobile and Wireless Information Systems (MOBIDATA), Rutgers University, NJ, November 1994.

[31] R. Frank Quick Jr., Kumar Balachandran
An overview of the cellular digital packet data (CDPD) system
In Proceedings of the Fourth International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Yokohama, Japan, September 1993

[32] E. Tiedemann
Data Services for the IS-95 CDMA Standard
Fourth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '93), Yokohama, Japan, September 1993

[33] The IEEE 802.11 Wireless LAN Standard
http://www.wlana.org/learn/80211.htm

[34] Wireless LAN security
http://www.wlana.org/learn/security.htm

[35] A. Nix, M. Li, J. Marvill, T. Wilkinson, I. Johnson, S. Barton
Modulation and equalisation considerations for high performance radio LANs (HIPERLAN)
Personal, Indoor and Mobile Radio Communications, 1994. 5th IEEE International Symposium on Wireless Networks – Catching the Mobile Future, Volume: 3, Page(s): 964–968, 1994

[36] Ian F. Akyildiz, Janise McNair, L. C. Martorell, Roman Puigjaner, Yelena Yesha
Medium Access Control protocols for multimedia traffic in wireless networks
IEEE Network Magazine, Volume: 13, Number 4, Page(s) 39–47, July/August 1999.

[37] W. N. Schilit
A System Architecture for Context-Aware Mobile Computing
PhD-Thesis Columbia University, 1995

[38] Timothy Jason Shepard
Decentralized Channel Management in Scalable Multihop Spread Spectrum Packet Radio Networks
Ph.D. Dissertation, Massachusetts Institute of Technology, 1995.

[39] David B. Johnson
Routing in ad hoc networks of mobile hosts.
In Proceedings of the IEEE Workshop on Mobile Computing Systems and Applications, Page(s): 158–163, December 1994.

[40] Yasser Mourtada
Routing Basics and Protocols
12 August 2000.

[41] Amit Jain
Routing protocols for mobile ad-hoc networks
Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur

[42] W. R. Heinzelman, A. Chandrakasan, H. Balakrishnan
Energy-efficient communication protocol for wireless microsensor networks
Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, Page(s): 3005–3014, 2000.

[43] Elizabeth M. Royer, Chai-Keong Toh
A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks
IEEE Personal Communications, Volume: 6, Page(s): 46–55, April 1999.

[44] Charles E. Perkins and Pravin Bhagwat
Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers
In Proceedings of the SIGCOMM '94 Conference on Communications Architectures, Protocols and Applications, Page(s) 234–244, August 1994.

[45] C. C. Chiang, H. K. Wu, W. Liu, and M. Gerla
Routing in Clustered Multihop, Mobile Wireless Networks with Fading Channels
Proceedings of the IEEE Singapore International Conference on Networks (SICON), Singapore, Page(s) 197–211, April 1997.

[46] S. Murthy and J. J. Garcia-Luna-Aceves
An Efficient Routing Protocol for Wireless Networks
ACM/Baltzer Mobile Networks and Applications, special issue on Routing in Mobile Communications Networks, Volume: 1, Number: 2, Page(s) 183–197, October 1996.

[47] C. E. Perkins, E. M. Royer
Ad-hoc on-demand distance vector routing
Proceedings of the Second IEEE Workshop on Mobile Computing Systems and Applications WMCSA '99, Page(s): 90 -100, 1999

[48] David B. Johnson, David A. Maltz
Dynamic Source Routing in Ad Hoc Wireless Networks
In Mobile Computing, edited by Tomasz Imielinski and Hank Korth, Chapter 5, Page(s) 153–181, Kluwer Academic Publishers, 1996.

[49] Vincent D. Park and M. Scott Corson
A highly adaptive distributed routing algorithm for mobile wireless networks
In Proceedings of INFOCOM'97, Page(s) 1405–1413, April 1997.

[50] Chai-Keong Toh
A novel distributed routing protocol to support ad-hoc mobile computing
Conference Proceedings of the 1996 IEEE Fifteenth Annual International Phoenix Conference on Computers and Communications, Page(s) 480–486, 1996

[51] R. Dube, C. D. Rais, Kuang-Yeh Wang, S. K. Tripathi
Signal stability-based adaptive routing (SSA) for ad hoc mobile networks
IEEE Personal Communications, Volume: 4, Issue: 1, Page(s) 36–45, February 1997

[52] Z. J. Haas and M. R. Pearlman
The zone routing protocol (ZRP) for ad hoc networks.
IETF Internet Draft. http://www.ietf.org/internet-drafts/draft-ietf-manet-zone-ierp-01.txt, 2001.

[53] Young-Bae Ko, N. H. Vaidya
Using location information in wireless ad hoc networks
IEEE 49th Vehicular Technology Conference, Volume: 3, Page(s) 1952–1956, 1999

[54] W.-H. Liao, Y.-C. Tseng, and J.-P. Sheu
GRID: A Fully Location-Aware Routing Protocol for Mobile Ad Hoc Networks
Telecommunication Systems, Volume: 18, Number: 1, Page(s) 37–60, 2001.

[55] I. Stojmenovic, X. Lin
GEDIR: Loop-free location based routing in wireless networks
IASTED Int. Conf. on Parallel and Distributed Computing and Systems, November 3-6, 1999, Boston, MA, USA.

[56] M. Joa-Ng, I-Tai Lu
A peer-to-peer zone-based two-level link state routing for mobile ad hoc networks
IEEE Journal on Selected Areas in Communications, Volume: 17, Issue: 8, Page(s): 1415–1425, August 1999.

[57] S.M. Hedetniemi, S.T. Hedetniemi, A.L. Liestman
A Survey of Gossiping and Broadcasting in Communication Networks.
Networks Volume: 18, Page(s): 319–349, 1988.

[58] N. Gupta and D. Manjunath
Gossiping in Multihop Radio Networks
Proceedings of the International Conference on Personal and Wireless Communication (ICPWC-99), Jaipur India, Feb 17–19, 1999.

[59] A. Bagchi, S.L. Hakimi, E.F. Schmeichel
Gossiping in a distributed network
IEEE Transactions on Computers, Volume: 42, Issue: 2, Page(s): 253 -256, February 1993.

[60] T.F. Gonzalez
Gossiping in the multicasting communication environment
Proceedings of the 15th International Parallel and Distributed Processing Symposium, Page(s): 6 pp., 2001.

[61] Z. Haas, J. Y. Halpern, L. Li
Gossip-Based Ad Hoc Routing
To appear in Proc. IEEE INFOCOM, June 2002.

[62] A. Vahdat, D. Becker
Epidemic routing for partially-connected ad hoc networks.
Duke Technical Report CS-2000-06, July 2000

[63] R. Chandra, V. Ramasubramanian, K. Birman
Anonymous gossip: Improving multicast reliability in mobile ad-hoc networks
In Proc. 21st International Conference on Distributed Computing Systems (ICDCS), Page(s): 275–283, 2001.