Faculty of Engineering

Department of Computer Engineering

WIRELESS DATA COMMUNICATION FOR ROBOT CONTROL

(An Application of Mobile Robot)

Graduation Project

COM-400

Student: Murat Küçük (980240)

Supervisor: Prof. Dr. Fakhreddin Mamedov

Lefkoşa - 2001

Table of Contents

ACKNOWLEDGMENT

First of all I would like to thank to Prof. Dr. Fakhreddin Mamedov who was supervisor of my project and to Assist. Prof. Dr. Rahib Abiyev who helped me preparing this project. With their endless knowledge I easily overcome many difficulties and learn a lot of things about Communication and Control Systems. Preparing this project is a nice experience of my life.

Also I would like to thank to all my friends, my family, and all my instructors because they never leave me alone and always try to help me during my education. Without their encouragements I would not be where I am now.

ABSTRACT

Graduation Project is devoted to the investigation of wireless communication and its application in mobile robot control. Problems in data communication, transmission media, different type of modulations, demodulations and shift-keying problems are considered. The structure of wireless LANs is given. The future evolution of wireless communication is given. In second chapter the two types of interfaces - USB and FireWire are considered. Their comparisons with other interfaces are given. In the last chapter the use of wireless communication in intelligent mobile robot control is considered. The structure and operation principle of mobile robot through wireless communication are given. At the end an application of wireless communication to mobile robot control is considered. Animate vision, mapping and navigation, planning, High-speed obstacle avoidance, vector field histogram and global path planning are considered.

INTRODUCTION

In this thesis, I tried to present wireless data communication for robot control systems. The aim of my researches is to implement a mobile robot moving from the source point to destination point. To perform such an approach I took mobile robot hardware and design to add some electronic components on it. As the title of my thesis is wireless data communication for robot control systems a wireless modem is added for the data communication, for autonomous movements distance measuring circuits are added, for navigation feedback a CCD sensor and a DSP is added to get the navigational data array. This is a new topic on mobile robots to control the movements of it. And in addition a temperature sensor and its driving circuit is added just for monitoring the temperature from the remote point onto the monitor. A Camera can be added for taking the frames while the robot is moving instead of monitoring temperature. With this robot my aim is to make the robot move from source to destination point and monitoring its path with navigation feedback and drawing the edges that are detected with its distance sensors and as a result to monitor the temperature values on the computer screen.

At the end of this thesis, a scheme of a 56-kilobaud synchronous RF modem with a 70 kHz bandwidth is given. The modulation in this modem is bandwidth limited MSK generated by a digital state machine driving two digital-to-analog converters, and two double balanced modulators. The carrier phase is shifted plus or minus 90 degrees for each bit. Demodulation is accomplished with a standard quadrature detector chip but various coherent methods can be used for operation at lower signal to noise ratios.

The distance measuring circuit scheme is given. It works with sonar. It sends sounds and receives the signal and calculates the time and generates the distance between obstacle and the robot.

The navigation feedback circuit is given. It has a CCD sensor on the chip and a DSP. As described above it takes 1800 fps. And the DSP processes the images. It generates the navigational information. It is a new topic in robotic applications. The circuit is HDNS-2000 by Agilent Semiconductor.

CHAPTER 1. CURRENT TECHNOLOGIES IN WIRELESS COMMUNICATION AND MOBILE COMPUTING

1.1. Historical Overview

It started with the Telegraph …

> "Electric telegraph is called the most perfect invention of modern times as anything more perfect than this is scarcely conceivable, and it thought what will be left for the next generation, upon which to expend the restless energies of the human mind." [An Australian newspaper, 1853.]

Origins of Coded Transmission:

- 1793, Revolutionary France
  - Aerial Telegraph, invented by Claude Chappe
  - Extensive network throughout France
- 1840s, Samuel F. B. Morse
  - Coded transmission via electronic means
  - Rapidly spread throughout US and Europe
  - International Telegraph Union (ITU) formed in 1865

Submarine Telegraphy: High Tech of the late 19th Century:

- 1850: Dover-to-Calais, first submarine line
- 1858: First transatlantic cable
  - Breaks after 3 months!
  - President Buchanan & Queen Victoria exchange telegrams
- 1866: Relaid with higher quality cable
  - Development of cable materials, technology of laying, repair
- Typical "Performance":
  - 1870: London to Bombay in 4 minutes, 22 seconds
  - 1901: London to British Guiana, 22 minutes
  - 1924: Telegram around the world in 80 seconds!

Radio Telegraphy (also know as "Wireless"):

- Radio technology
  - Communicate with ships and other moving vehicles
  - Messages sprayed into the "ether" crossing wide boundaries
  - Downfall of the nationally supported monopolistic telegraph companies

- 1896: Guglielmo Marconi
  - First demonstration of wireless telegraphy
  - Built on work of Maxwell and Hertz to send and receive Morse Code
  - Based on long wave (>> 1 km), spark transmitter technology, requiring very large, high power transmitters
  - First used by British Army and Navy in the Boer War
  - 1899: Reported to shore America's Cup yacht races

Wireless:

- 1907: Commercial Trans-Atlantic Wireless Service
  - Huge ground stations: 30 x 100m antenna masts
  - Beginning of the end for cable-based telegraphy
- WW I: Rapid development of communications intelligence, intercept technology, cryptography
- 1920: Marconi discovers short-wave (<100 m) radio
  - Long wave follow contour of land
    - Very high transmit power, 200 KW+
  - Short waves reflect, refract, and absorb, like light
    - Bounce off ionosphere
    - Higher frequencies made possible by vacuum tube (1906)
    - Cheaper, smaller, better quality transmitters

Other Important Dates:

- 1915: Wireless voice transmission NY to SF
- 1920: First commercial radio broadcast (Pittsburgh)
- 1921: Police car dispatch radios, Detroit
- 1935: First telephone call around the world
- WW II: Rapid development of radio technology
- 1968: Carter phone decision
- 1974: FCC allocates 40 MHz for cellular telephony
- 1982: European GSM and Inmarsat established
- 1984: Breakup of AT&T
- 1984: Initial deployment of AMPS cellular system

1.2. Radio Fundamentals

Radio Waves! Portable, even hand-held, short wave transmitters can reach thousands of miles beyond the horizon. Tiny microwave transmitters aboard space probes return data from across the solar system. And all at the speed of light. Yet before the late 1800s there was nothing to suggest that telegraphy through empty space would be possible even with mighty dynamos, much less with insignificantly small and inexpensive apparatus. The Victorians could extrapolate from experience to imagine flight aboard a steam-powered mechanical bird or space travel in a scaled-up Chinese skyrocket. But what experience would even have hinted at wireless communication? The key to radio came from theoretical physics. Maxwell consolidated the known laws of electricity and magnetism and added the famous displacement current term, $\partial D/\partial t$. By virtue of this term, a changing electric field produces a magnetic field, just as Faraday had discovered that a changing magnetic field produces an electric field. Maxwell's equations predicted that electromagnetic waves could break away from the electric currents that generate them and propagate independently through space with the electric and magnetic field components of the wave constantly regenerating each other.

Maxwell's equations predict the velocity of these waves to be $1/\sqrt{\varepsilon_0 \mu_0}$ where the constants ε_0 and μ_0 can be determined by simple measurements of the static forces between electric charges and between current-carrying wires. The dramatic result is, of course, the experimentally known speed of light, 3×10^8 m/s. The electromagnetic nature of light is revealed. Hertz conducted a series of brilliant experiments in the 1880s in which he generated and detected electromagnetic waves with wavelengths very long compared to light. The distribution of wavelengths can be seen in Figure 1.1. The utilization of Hertzian waves (the radio waves we now take for granted) to transmit information developed hand-in-hand with the new science of electronics.

Where is radio today? AM radio, the pioneer broadcast service, still exists along with FM, television, and two-way communication. Now radio also includes radar, surveillance, navigation and broadcast satellites, cellular telephones, remote control devices, and wireless data communications. Applications of radio frequency (RF) technology outside radio include microwave heaters, medical imaging systems, and cable television.

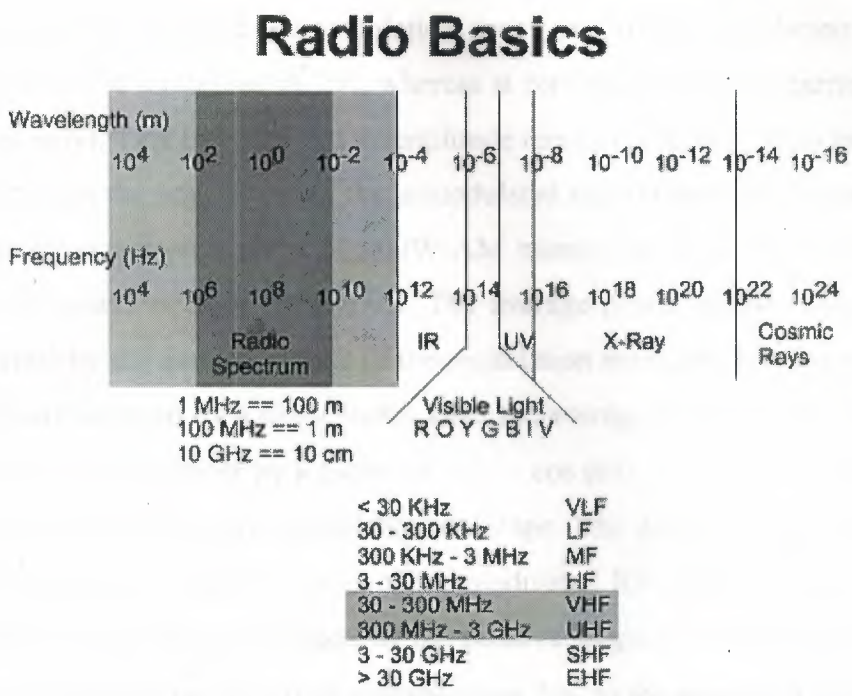


Figure 1.1. Radio Spectrum

1.3. Analogue Modulation Techniques

1.3.1. Amplitude Modulation

Modulation means adding information to an otherwise pure sinusoidal carrier wave by varying the amplitude or the phase (or both). The simplest, amplitude modulation (AM) is on/off keying. This binary AM can be accomplished with just a switch (telegraph key) connected in series with the power source. The earliest voice transmissions used a carbon microphone as a variable resistance in series with the antenna. Amplitude modulation is used in the long-wave, middle-wave, and short wave broadcast bands. Without modulation (when the music or speech is silent) the voltage and current at the antenna are pure sine waves at the carrier frequency. The rated power of a station is defined as the transmitter output power when the modulation is zero. The presence of an audio signal changes the amplitude of the carrier. The audio signal (amplified microphone voltage) has positive and negative excursions, but its average value is zero. The audio voltage is bounded by $+V_m$ and $-V_m$. A dc bias voltage of V_m volts is added to the audio voltage. The sum, $V_m + V_{audio}$, is always positive, and is used to multiply the carrier wave, $sin\ (\omega_c t)$. The resulting product is the AM signal; the

amplitude of the RF sine wave is proportional to the biased audio signal. The simulation in Figure 1.2 shows the various waveforms in the transmitter and receiver. The biased audio waveform is called the modulation envelope. At full modulation where $V_{audio} + V_m$, the carrier is multiplied by $2V_m$ whereas at zero modulation the carrier multiplied by V_m (bias only). This factor of two in amplitude means the fully (100%) modulated signal has four times the peak power of the unmodulated signal (carrier wave alone). It follows that the antenna system for a 50,000W AM transmitter must be capable of handling 200,000W peaks without breakdown. The average power of the modulated signal is determined by the average square of the modulation envelope. For example, in the case of 100% modulation by a single audio tone, the average power of the modulated signal is greater than the carrier by a factor of $< (1 + \cos(\theta))^2 >= 3/2$. Receiver demodulates the signal by detecting the modulation envelope. The detector is just a rectifier diode that eliminates the negative cycles of the modulated RF signal. A simple *RC* low-pass then produces the average voltage of the positive loops. (The average voltage of these sinusoidal loops is just their peak voltage times $2/\pi$, so the average is proportional to the peak, that is, the envelope.) Finally, ac coupling removes the bias, leaving an audio signal identical to the signal from the microphone. Figure 1.2 shows a basic Amplitude Modulation.

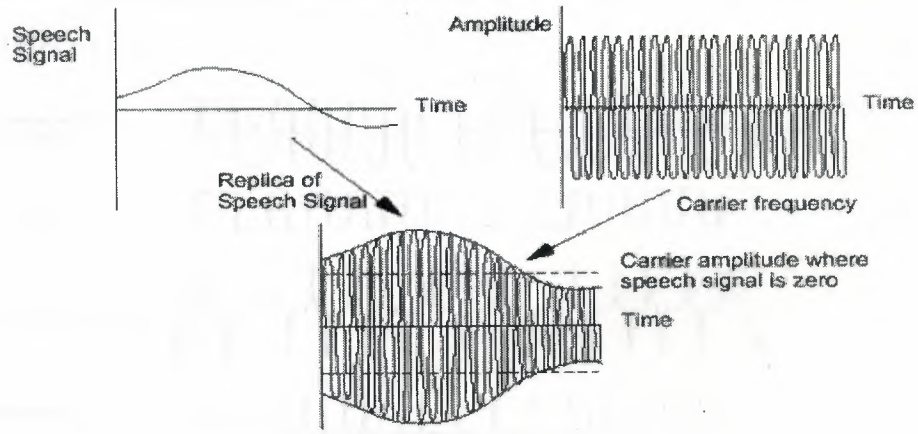


Figure 1.2. Amplitude Modulation

1.3.2. Frequency Modulation

Noise has a greater effect on amplitude than frequency.

Sufficient to detect zero crossings to reconstruct the signal

Easy to eliminate amplitude distortion

Constant envelope, i.e., envelope of carrier wave does not change with changes in modulated signal. This means that more efficient amplifiers can be used, reducing power demands. Transmitted signal can be seen in Figure 1.3

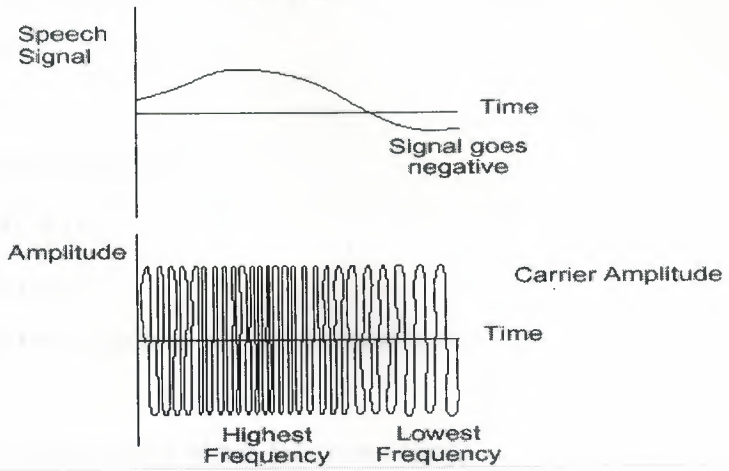


Figure 1.3. Frequency Modulation

Detection of FM Signal:

Noise translates into amplitude changes, and sometimes frequency changes.

Detection based on zero crossings: the limiter.

Alternative schemes to translate limited signal into bit streams. The steps are showed in figure 1.4.

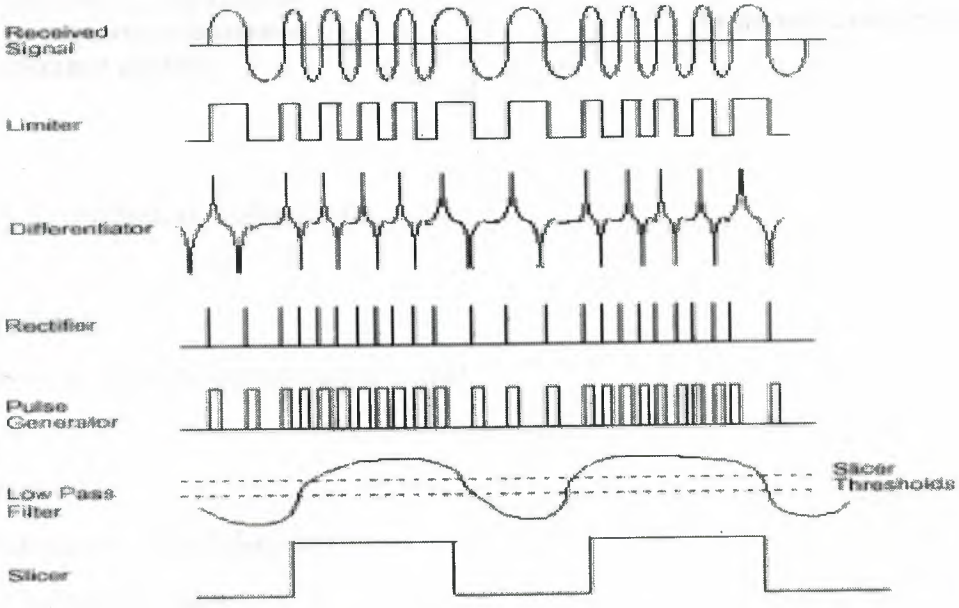


Figure 1.4. Steps of detecting FM signal

1.4. Digital Modulation Techniques

Carrier wave s:

$S(t) = A(t) * \cos[\theta(t)]$

Function of time varying amplitude A and time varying angle θ

Angle θ rewritten as:

$\theta(t) = \omega_0 + \varphi(t)$

ω_0 radian frequency, phase φ (t)

$S(t) = A(t) \cos[\omega_0 t + \varphi(t)]$

ω Radians per second

Relationship between radians per second and hertz

$\omega = 2\pi f$

Modify carrier's amplitude and/or phase (and frequency)

Constellation: Vector notation/polar coordinates. Figure 1.4 describes the technique for the basic digital modulation technique.

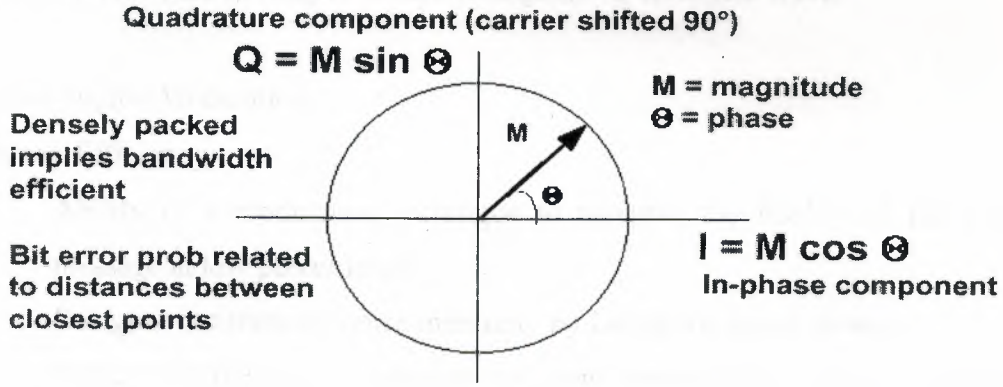


Figure 1.5. Quadrature Components

Demodulation:

Process of removing the carrier signal

Detection:

Process of symbol decision

Coherent detection

Receiver users the carrier phase to detect signal

Cross correlate with replica signals at receiver

Match within threshold to make decision

Noncoherent detection

Does not exploit phase reference information

Less complex receiver, but worse performance

Table 1.1. Coherent and Noncoherent Techniques

Coherent	Noncoherent
Phase shift keying (PSK)	FSK
Frequency shift keying (FSK)	ASK
Amplitude shift keying (ASK)	Differential PSK (DPSK)
Continuous phase modulation (CPM)	CPM
Hybrids	Hybrids

Coherent (aka synchronous) detection: process-received signal with a local carrier of same frequency and phase.

Noncoherent (aka envelope) detection: requires no reference wave.

Metrics for Digital Modulation:

- Power Efficiency
  - Ability of a modulation technique to preserve the fidelity of the digital message at low power levels
  - Designer can increase noise immunity by increasing signal power
  - Power efficiency is a measure of how much signal power should be increased to achieve a particular BER for a given modulation scheme
  - Signal energy per bit / noise power spectral density: E_b / N_0
- Bandwidth Efficiency
  - Ability to accommodate data within a limited bandwidth
  - Tradeoff between data rate and pulse width.
  - Throughput data rate per hertz: R/B bps per Hz
- Shannon Limit: Channel capacity / bandwidth
  - $C/B = \log 2 (1 + S/N)$

Criteria on selecting the right modulation:

- High spectral efficiency

- High power efficiency
- Robust to multi-path effects
- Low cost and ease of implementation
- Low carrier-to-cochannel interference ratio
- Low out-of-band radiation
- Constant or near constant envelope
  - Constant: only phase is modulated
  - Non-constant: phase and amplitude modulated

1.4.1. Amplitude Shift Keying (ASK)

The amplitude of the carrier c (t) is varied to represent binary of 1 or 0.Both frequency and phase remains constant. It is shown in figure 1.6.

The technique in ASK is called "On-Off-Keying" (OOK). In OOK no voltage represents one of the bit values (for example 0). A bit duration Tb is the interval of time that defines one bit. The amplitude of carrier c(t) is switched between two levels depending on the bits (0 or 1). Which voltage represents 1 and which represents 0 is left to the system designers. The speed of transmission using ASKS is limited by the physical characteristics of the transmission medium.

The advantage is a reduction in the amount of energy required to transmit information.

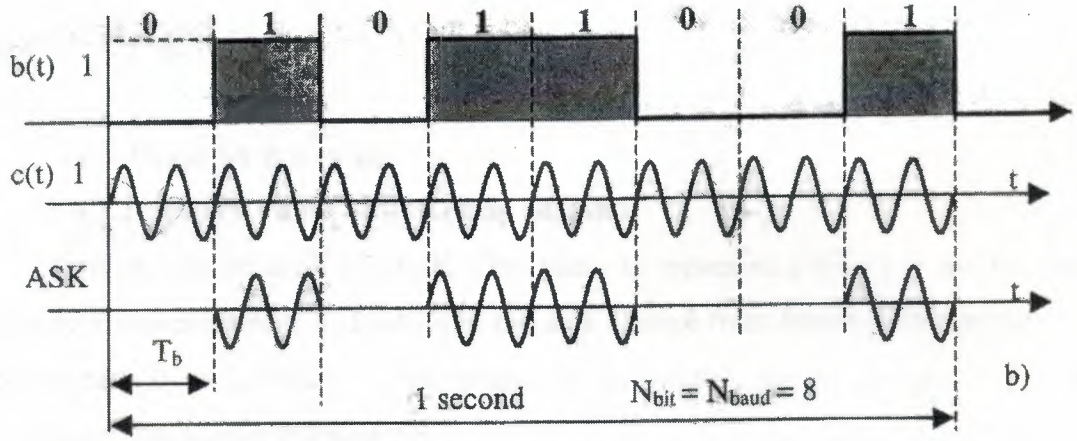


Figure 1.6. ASK signal

1.4.2. Frequency Shift Keying (FSK)

1/0 represented by two different frequencies slightly offset from carrier frequency in FSK. Two fixed amplitude carrier $c_1(t)=\cos 2\pi f_{c1}t$ and $c_2(t)=\cos 2\pi f_{c2}t$ one for binary 0 one for binary 1.The frequency of the signal during each bit duration is constant and its value depends on the bit (0 or 1). Figure 1.7 gives the conceptual view of FSK. FSK avoids most of the noise problem of ASK. As the receiving device is looking for specific frequency changes over a given number of periods, it can ignore voltage spikes.

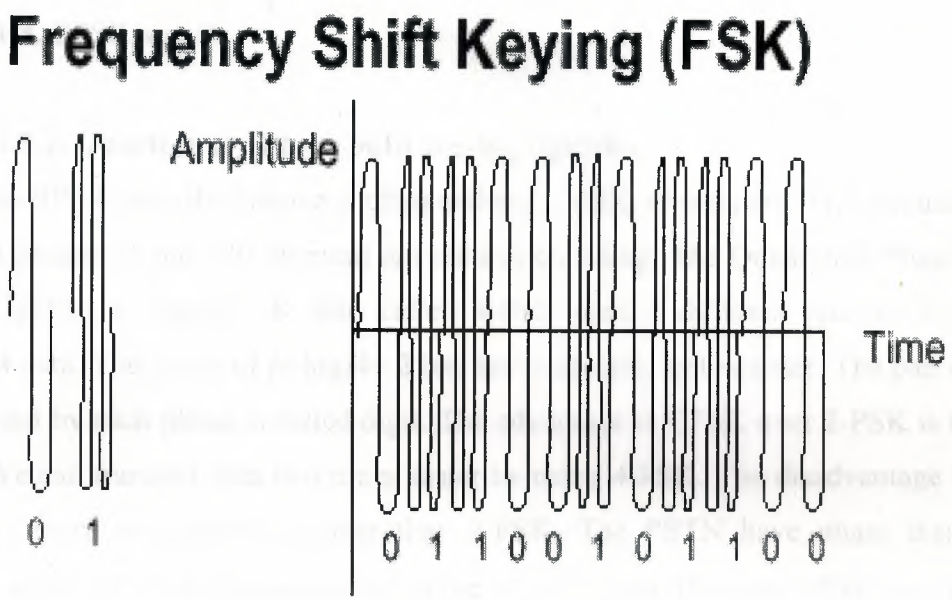


Figure 1.7. FSK signal

1.4.3. Phase Shift Keying

1.4.3.1. Binary Phase Shift Keying (BPSK)

Two phases are used in BPSK. One phase to represent a binary 0 and the other phase to represent binary 1. Each time the data change from binary 1 to a binary 0 or from binary 0 to a binary 1, the phase of transmitted signal changes 180°. Its characteristic is shown in figure 1.8.

- Simple to implement, inefficient use of bandwidth
- Very robust, used extensively in satellite communications

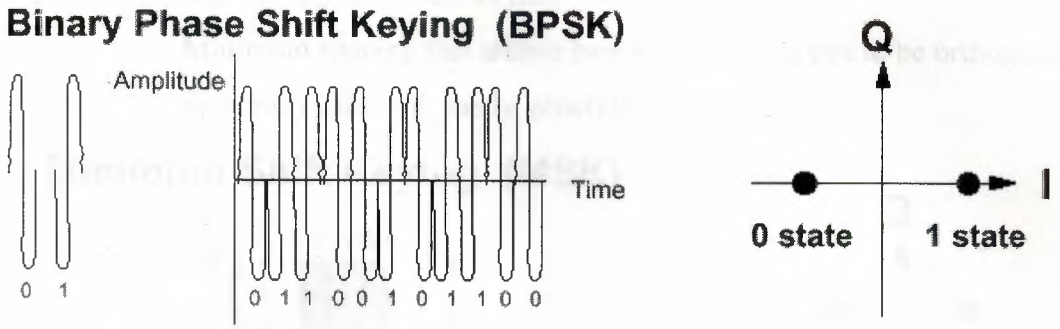


Figure 1.8. BPSK signal

1.4.3.2. Quarternary Phase Shift Keying (QPSK)

The BPSK described above is often called 2 – PSK, or ordinary PSK, because two different phases (0 and 180 degrees) are used in encoding. The Quadrature Phase-Shift Keying QPSK, in figure 1.8, also called 4-PSK uses 4 different phases (M=4) to represent data. The group of $n=\log_2 4= 2$ bits are modulated onto carrier. The pair of bits represented by each phase is called digit. The advantage is QPSK over 2-PSK is higher speed. We can transmit data two times faster by using 4-PSK. The disadvantage is that QPSK is more susceptible to error than 2-PSK. The PSTN have phase distortion (achieve up to 20°), which causes error in the received data. Because 2-PSK uses a 180° phase shift and it can tolerate phase tolerance approaching 90°. QPSK tolerate telephone circuit phase tolerance approaching 45°.

- Multilevel modulation technique: 2 bits per symbol
- More spectrally efficient, more complex receiver.

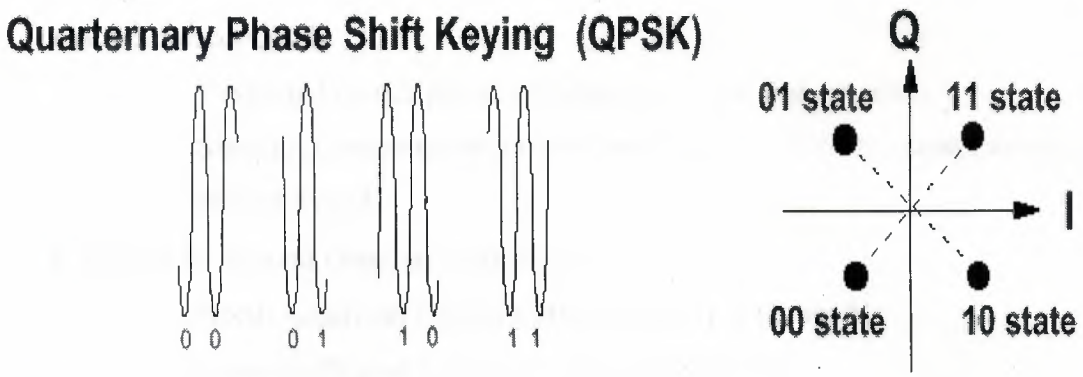


Figure 1.9. QPSK signal

1.4.3.3. Minimum Shift Keying

- Special form of frequency shift keying
  - Minimum spacing that allows two frequencies states to be orthogonal
  - Spectrally efficient, easily generated (Figure 1.10)

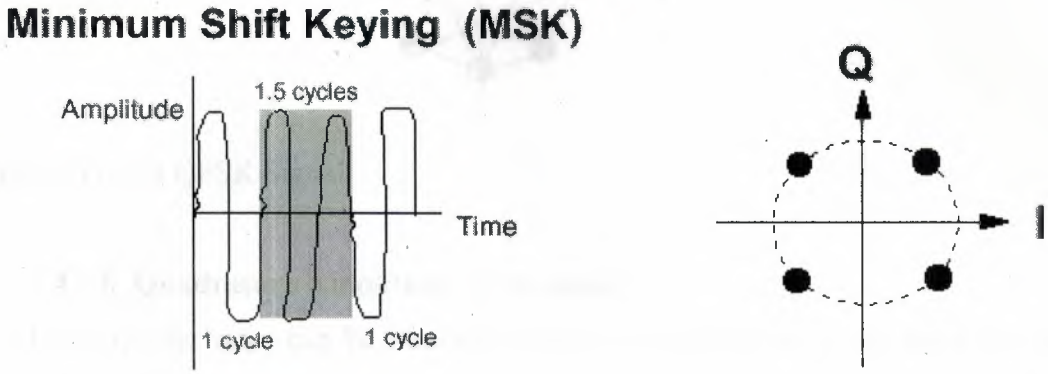


Figure 1.10. MSK Signal

1.4.3.4. Gaussian Minimum Shift Keying (GMSK)

- MSK + premodulation Gaussian low pass filter
- Increases spectral efficiency with sharper cutoff
- Used extensively in second generation digital cellular and cordless telephone applications
  - GSM digital cellular: 1.35 bps/Hz
  - DECT cordless telephone: 0.67 bps/Hz
  - RAM Mobile Data

1.4.3.5. p/4-Shifted QPSK

- Variation on QPSK
  - Restricted carrier phase transition to +/- p/4 and +/- p/4
  - Signaling elements selected in turn from two QPSK constellations, each shifted by p/4
- Popular in Second Generation Systems
  - North American Digital Cellular (IS-54): 1.62 bps/Hz
  - Japanese Digital Cellular System: 1.68 bps/Hz
  - European TETRA System: 1.44 bps/Hz
  - Japanese Personal Handy Phone (PHP)

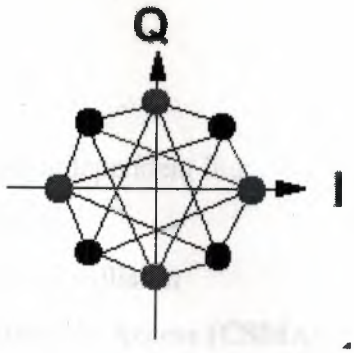


Figure 1.11. p/4 QPSK Signal

1.4.3.6. Quadrature Amplitude Modulation

Data transfer rates can be increased further by decreasing phase angle between two adjacent pharos. Four bits, or a quad bit, for example can be encoded into 16 possible phase changes (M=16). The phase differential between adjacent phasers would amount to 22.5° (360°/16=22.5°). The problem here, however, is that any phase shift 11.25° degree 2ill be within of the phase distortion introduced by PSTN (11.25°<20°). For this reason, 16 phase PSK is generally not used. To avoid the problem of phase jitter, the combination of ASK and PSK called Quadrature Amplitude Modulation (QAM) are used. Possible variation of QAM is numerous. Theoretically any measurable number of changes in amplitude can be combined with any measurable number of changes in phase. In the Figure 1.12 level QAM can be seen.

- Quadrature Amplitude Modulation (QAM)
  - Amplitude modulation on both Quadrature carriers
  - 2 n discrete levels, n = 2 same as QPSK
- Extensive use in digital microwave radio links

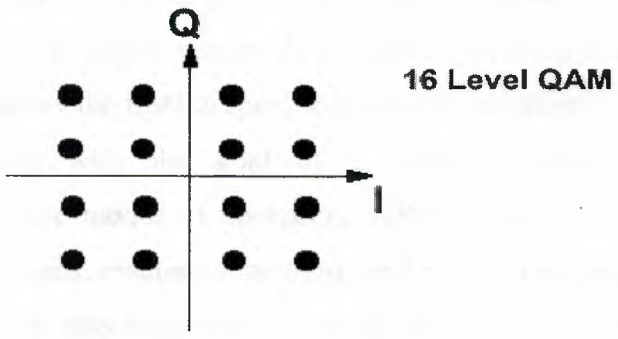


Figure 1.12. Quadrature Amplitude Modulation

1.5. Media Access

1.5.1. ALOHA

Transmit when desired

Positive ACK from receiver on independent link

Back off and retransmit if timeout

Slotted scheme reduces chance of collision

1.5.2. Carrier Sense/Multiple Access (CSMA)

Listen before transmit

Back off and retransmit if collision detected

1.5.3. Inhibit Sense/Multiple Access

Base station transmits busy tone

Transmit when not busy

Back off and retransmit if collision

1.5.4. Time Division Multiple Access

Multiple users share channel through time allocation scheme

Time Division Duplexing (TDD): DECT, PHP

Frequent Division Duplexing (FDD): GSM, IS-54, PACS

TDMA is an extension of AMPS. IS-136 systems are capable of operating with AMPS terminals, dual-mode terminals, and all-digital terminals. The network architecture is a more general version of the AMPS architecture. Corresponding to the AMPS network infrastructure of land stations and mobile telephone switching offices (base stations and switches), TDMA defines a BMI: "Base Station Mobile Switching Center, and Inter-working Function." Because IS-136 is confined to the air interface, it is appropriate to specify, in this general way, the functions performed in the network infrastructure. Each equipment vendor then makes its own decisions on how to allocate functions performed by the BMI to specific pieces of equipment.

In accordance with the goal of a personal communications system to accommodate multiple modes of operation, TDMA specifies three types of external network: public systems, residential systems, and private systems. Thus, a terminal can function as a cellular telephone with access to the base stations of cellular operating companies (public network). It can also be programmed to function as a cordless telephone operating with a specific residential base station (residential network), and as

a business phone operating with a specific wireless private branch exchange (private network).

To deliver mobile telephony, cryptographic authentication, and a wide range of service enhancements relative to AMPS, TDMA defines a large number of identification codes, including all of the AMPS identifiers. A major addition to the set of identification codes is the 64-bit A-key, assigned to each subscriber by her cellular operating company. This encryption key plays a critical role in promoting network security and communication privacy in a dual-mode TDMA system. Another identification code in TDMA is a 12-bit location area identifier, LOCAID. The system can divide its service area into clusters of cells, referred to as location areas. Each base station broadcasts its LOCAID. When a terminal that does not have a call in progress enters a new location area, it sends a registration message to the system. When a call arrives for the terminal, the system pages the terminal in the location area that received the most recent registration message.

The IMSI is a telephone number with up to 15 decimal digits that conforms to an international numbering plan (E.212) published by the International Telecommunication Union. The value of PV reflects the standards document (for example, IS-54 or IS-136) that governs the operation of a base station or terminal. The system operator code (SOC) transmitted by a base station identifies to terminals the company that operates the base station, while BSMC indicates the manufacturer of the base station. The digital verification color code (DVCC) plays the same role in digital traffic channels as the SAT transmitted in analog traffic channels.

Table 1.2 Comparisons of Cellular Systems

	GSM	IS-54	DECT
Bit Rate	270.8 kbps	48.6 kbps	1.152 Mbps
Bandwidth (Carrier Spacing)	200 KHz	30 KHz	1.728 MHz
Time Slot Duration	0,577 ms	6.7 ms	0.417 ms
Upstream slots per frame	8/16	3/6	12
Speech Coding	13 kbps RPE-LTP	7.95 kbps VSELP	32 kbps ADPCM
FDD or TDD	FDD	FDD	TDD
% Payload in Time Slot	73%	80%	67%
Modulation	GMSK	$\pi/4$ DQPSK	GMSK
Coding	Coded/Convol Coded+CRC Uncoded	Coded/Convol Coded+CRC Uncoded	CRC Only
Adaptive Equalizer	Mandatory	Mandatory	None

TDMA Advantages/Disadvantages:

In Table 1.2 the comparisons between GSM, IS-54 and DECT can be seen.

- Advantages
  - Sharing among N users
  - Variable bit rate by ganging slots
  - Less stringent power control due to reduced interuser interference—dedicated frequencies and slots
  - Mobile assisted/controlled handoff enable by available measurement slots
- Disadvantages
  - Pulsating power envelope interference with devices like hearing aids have been detected
  - Complexity inherent in slot/frequency allocation
  - High data rates imply need for equalization

1.5.5. Code Division Multiple Access

- A strategy for multiple users per channel based on orthogonal spreading codes
  - Multiple communicators simultaneously transmitting using direct sequence techniques, yet not conflicting with each other.
  - Pilot tone on BS to mobile unit forward channel used to time synchronize and equalize the channel (coherent detection).
  - Reverse channel is contention based, dynamically power controlled to eliminate the near-far problem.
- Developed by Qualcomm as IS-95
  - Special soft handoff capability
  - "Narrowband CDMA": 1.228 MHz chipping rate, 1.25 MHz spread bandwidth
  - Contrast with Broadband CDMA proposal: 10 MHz spread bandwidth
    - Multipath: Can leverage frequency diversity better
    - Interference tolerance: Can overlay existing analog user

Like a TDMA, IS-95 prescribes dual-mode operation. However, the two systems differ substantially in their relationship to the analog AMPS systems in which they operate. Recall that an A TDMA signal occupies exactly the same bandwidth as an analog AMPS signal. As a consequence system operators can replace individual AMPS channel units in analog base stations with TDMA radios that carry three full-rate physical charnels. By contrast, IS-95 prescribes spread spectrum signals with a bare-

width of 1.23 MHz in each direction. This is approximately 10 percent 0f a company's total spectrum allocation. As a consequence, a cellular operating company that adopts CDMA has to convert frequency bands of at least this size, corresponding to 41 contiguous AMPS channels, from analog to digital operation.

1S-95 contains many innovations relative to earlier cellular systems. One of them is a soft hand off mechanism, in which a terminal establishes contact with a new base station before giving up its radio link to the original base station. When a call is in a soft handoff condition, the terminal transmits coded speech signals to two base stations simultaneously. Both base stations send their demodulated signals to the switch, which estimates the quality of the two signals and sends one of them to a speech decoder. A complementary process takes place in the forward direction. The switch sends coded

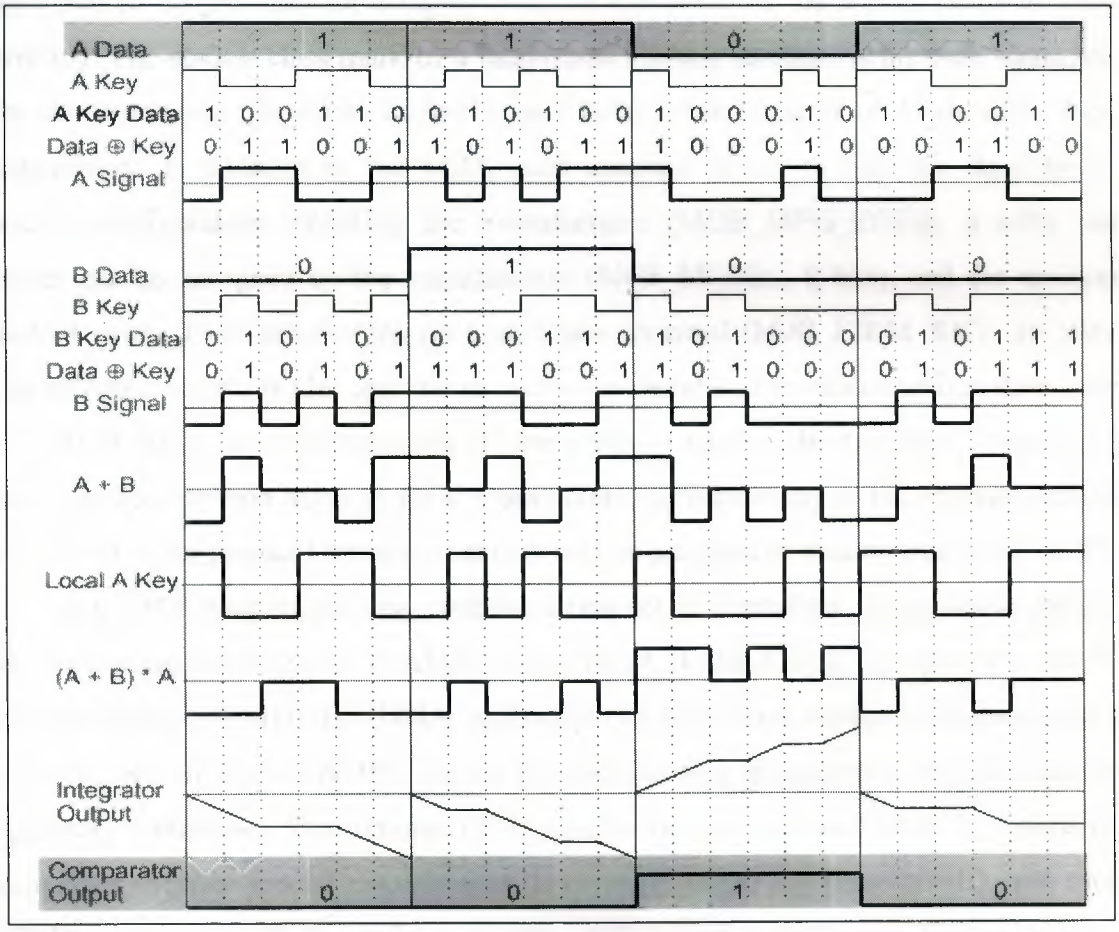


Figure 1.12. Recovery of a channel

speech signals to both base stations, which transmit them simultaneously to the terminal. The terminal combines the signals received from the two base stations and demodulates the result. Thus we have the network architecture illustrated in Figure 1.12,

which shows a vocoder in the switch rather than in base stations, their location in many TDMA implementations.

CDMA soft handoff requires base stations to operate in synchronism with one another. In order to achieve the necessary synchronization, all base stations contain global positioning system (GPS) receivers. A network of GPS satellites transmits signals that enable each GPS receiver to calculate its location in coordinates of latitude, longitude, and elevation. The satellite signals also include precise time information, accurate to within one microsecond, relative to universal coordinated time, an international standard.

In common with AMPS and TDMA, CDMA terminals and base stations employ an extensive set of identification codes that help control various network operations. Note that IS-95 provides for a highly detailed indication of the configuration of each terminal. The station class mark of a dual-mode CDMA terminal is an 8-bit identifier. The corresponding identifiers in AMPS and TDMA have lengths of 4 bits and 5 bits, respectively. In addition to the SCM, each terminal stores 40 bits that describe its precise configuration including the manufacturer (MOB_MFG_CODE, 8 bits), the model number assigned by the manufacturer (M0B_M0DEL, 8 bits), and the revision number of the firmware running on a particular terminal (M0B_FIRM_REV, 16 bits). The revision number is also specific to each manufacturer. The other configuration code is M0B_P_REV, an 8-bit indicator of the protocol run by the terminal. Initially all terminals operate with M0B_P_REV = 00000001, corresponding to the original version of 15-95. Higher protocol revision numbers will be assigned to future versions of 15-95.

A CDMA base station also contains a rich set of identifiers. Augmenting the 15-bit system identifier (SID) in AMPS and TDMA, CDMA systems specify a 16-bit network identifier (NID). In CDMA, a network is a set of base stations contained within a system. Recall that an AMPS system corresponds to a geographical area defined by regulatory authorities. By contrast, CDMA networks can be established by operating companies to meet special requirements. Each base station has its own NID, and each CDMA terminal can be programmed with a SID/NID pair indicating the system and network associated with the terminal's home subscription. Each base station has its own PN_0FFSET. This is a time delay applied to forward direction transmissions that enables the terminals in a cell to decode the desired signal and reject signals from other base stations. The 4-bit BASE_CLASS identifier anticipates terminals that will have

access to a variety of wireless services. In the initial issue of IS-95, the only assigned BASE_CLASS is 0000, corresponding to public macro cellular systems. Future class numbers could be assigned to other public networks or to various types of private networks such as wireless business systems (PBX) and residential cordless telephones.

The CDMA system anticipates a variety of mobility management schemes including location-area registration, as in TDMA and GSM; timer-based registration; and distance based registration. To facilitate location-area registration, IS-95 defines a 12-bit REG_ZONE identifier to be assigned to each base station. REG_ZONE plays the same role as the location area identifier, LOCAID, in TDMA. The identifiers, BASE_LAT (22 bits) and BASE_LONG (23 bits), specify the exact geographic location of the base station, in latitude-longitude coordinates. Terminals can use this information to perform distance-based registration.

1.6. Wireless LANs

Wireless LAN technology is becoming increasingly popular for a wide variety of applications. After evaluating the technology, most users are convinced of its reliability, satisfied with its performance and are ready to use it for large-scale and complex wireless networks. Originally designed for indoor office applications, today's Wireless LANs can be used for both indoor peer-to-peer networks as well as for outdoor point-to-point and point-to-multipoint remote bridging applications. Wireless LANs can be designed to be modular and very flexible. They can also be optimized for different environments. For example, point-to-point outdoor links are less susceptible to interference and can have higher performance if designers increase the "dwell time" and disable the "collision avoidance" and "fragmentation" mechanisms described later in this section.

1.6.1. Topology

Wired LAN Topology: Traditional LANs (Local Area Networks) link PCs and other computers to one another and to file servers, printers and other network equipment using cables or optic fibers as the transmission medium (Figure 1.13).

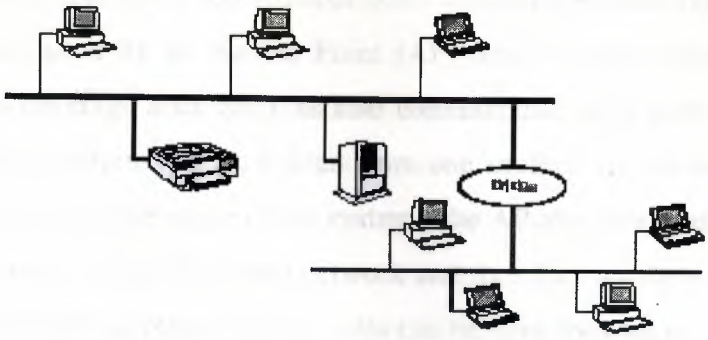


Figure 1.13: Wired LAN Topology

Wireless LAN Topology: Wireless LANs allow workstations to communicate and to access the network using radio propagation as the transmission medium. The wireless LAN can be connected to an existing wired LAN as an extension, or can form the basis of a new network. While adaptable to both indoor and outdoor environments, wireless LANs are especially suited to indoor locations such as office buildings, manufacturing floors, hospitals and universities. The basic building block of the wireless LAN is the Cell. This is the area in which the wireless communication takes place. The coverage area of a cell depends on the strength of the propagated radio signal and the type and construction of walls, partitions and other physical characteristics of the indoor environment. PC-based workstations, notebook and pen-based computers can move freely connected in the cell (Figure 1.13)

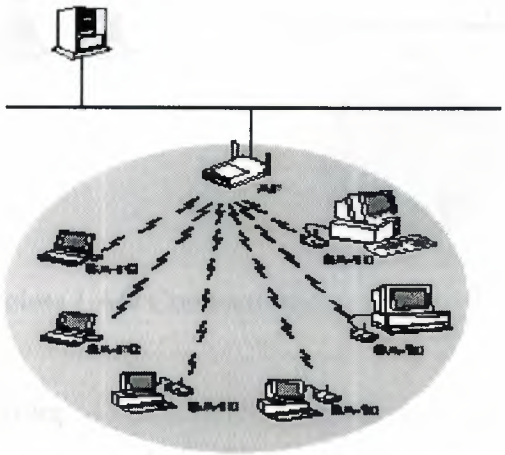


Figure 1.14: The Basic Wireless LAN Cell

Each Wireless LAN cell requires some communications and traffic management. This is coordinated by an Access Point (AP) that communicates with each wireless station in its coverage area. Stations also communicate with each other via the AP, so communicating stations can be hidden from one another. In this way, the AP functions as a relay, extending the range of the system. The AP also functions as a bridge between the wireless stations and the wired network and the other wireless cells. Connecting the AP to the backbone or other wireless cells can be done by wire or by a separate wireless link, using wireless bridges. The range of the system can be extended by cascading several wireless links, one after the other (Figure 1.14).

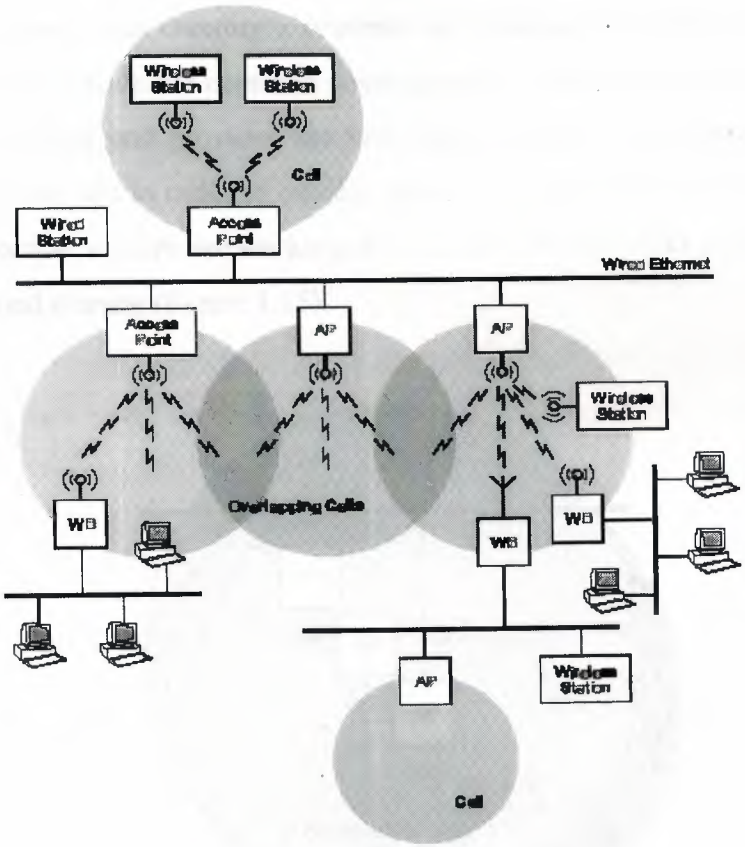


Figure 1.14. Wireless LAN Connectivity

1.6.2. Roaming

When any area in the building is within reception range of more than one Access Point, the cells' coverage is said to overlap. Each wireless station automatically establishes the best possible connection with one of the Access Points. Overlapping coverage areas are an important attribute of the wireless LAN setup, because it enables

seamless roaming between overlapping cells. Roaming allows mobile users with portable stations to move freely between overlapping cells, constantly maintaining their network connection. Roaming is seamless; a work session can be maintained while moving from one cell to another. Multiple access points can provide wireless coverage for an entire building or campus. When the coverage area of two or more APs overlap, the stations in the overlapping area can establish the best possible connection with one of the APs, continuously searching for the best AP. In order to minimize packet loss during switchover, the "old" and "new" APs communicate to coordinate the process. Load Balancing Congested areas with many users and heavy traffic load per unit may require a multi-cell structure. In a multi-cell structure, several co-located APs "illuminate" the same area creating a common coverage area that increases aggregate throughput. Stations inside the common coverage area automatically associate with the AP that is less loaded and provides the best signal quality. The stations are equally divided between the APs in order to equally share the load between all APs. Efficiency is maximized because all APs are working at the same low-level load. Load balancing is also known as load sharing (Figure 1.15).

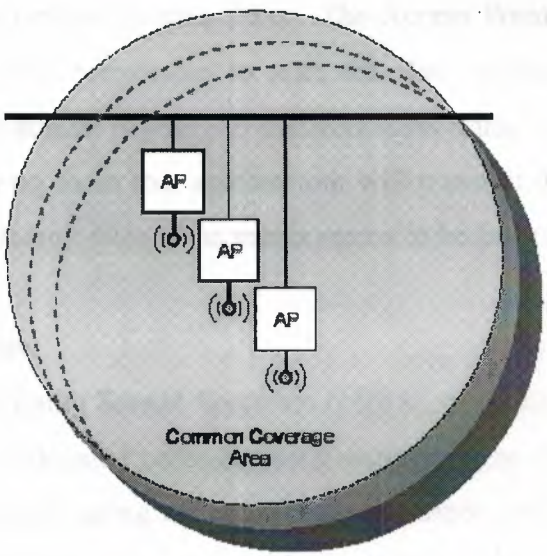


Figure 1.15. The Common Coverage Area of a Multi-cell Structure

1.6.3. Dynamic Rate Switching

The data rate of each station is automatically adjusted according to the received signal quality. Performance (throughput) is maximized by increasing the data rate and

decreasing re-transmissions. This is very important for mobile applications where the signal quality fluctuates rapidly, but less important for fixed outdoor installations where signal quality is stable.

1.6.4. Media Access

When many users are located in the same area, performance becomes an issue. To address this issue, Wireless LANs use the Carrier Sense Multiple Access (CSMA) algorithm with a Collision Avoidance (CA) mechanism in which each unit senses the media before it starts to transmit. If the media is free for several microseconds, the unit can transmit for a limited time. If the media is busy, the unit will back off for a random time before it senses again. Since transmitting units compete for air time, the protocol should ensure equal fairness between the stations. Fragmentation of packets into shorter fragments add protocol overhead and reduce protocol efficiency when no errors are expected, but reduce the time spent on re-transmissions if errors arelikely to occur. No fragmentation or longer fragment length add overhead and reduce efficiency in case of errors and re-transmissions (multi-path).

1.6.5. Collision Avoidance

To avoid collisions with other incoming calls, each station transmits a short RTS (Request To Send) frame before the data frame. The Access Point sends back a CTS (Clear To Send) frame with permission to start the data transmission. This frame includes the time that this station is going to transmit. This frame is received by all the stations in the cell, notifying them that another unit will transmit during the following Xmsec, so they can not transmit even if the media seems to be free (the transmitting unit is out of range).

1.6.6. Channelization

Using Frequency Hopping Spread Spectrum (FHSS), different hopping sequences are assigned to different co-located cells. Hopping sequences are designed so different cells can work simultaneously using different channels. Since hopping sequences and hopping timing of different cells cannot be synchronized (according to FCC regulations), different cells might try to use the same channel occasionally. Then, one cell uses the channel while the other cell backs off and waits for the next hop. In the case of a very noisy environment (multiples and interference), the system must hop quickly. If the link is quiet and clean, it is better to hop slowly, reducing overhead and increasing efficiency.

1.7. Future of Mobile Wireless Communications

3rd Generation Wireless, or 3G, is the generic term used for the next generation of mobile communications systems. 3G systems aim to provide enhanced voice, text and data services to user. The main benefit of the 3G technologies will be substantially enhanced capacity, quality and data rates than are currently available. This will enable the provision of advanced services transparently to the end user (irrespective of the underlying network and technology, by means of seamless roaming between different networks) and will bridge the gap between the wireless world and the computing/Internet world, making inter-operation apparently seamless. The third generation networks should be in a position to support real-time video, high-speed multimedia and mobile Internet access. All this should be possible by means of highly evolved air interfaces, packet core networks, and increased availability of spectrum. Although ability to provide high-speed data is one of the key features of third generation networks, the real strength of these networks will be providing enhanced capacity for high quality voice services. The need for landline quality voice capacity is increasing more rapidly than the current 2nd generation networks will be able to support. High data capacities will open new revenue sources for the operators and bring the Internet more closer to the mobile customer. The use of all-ATM or all-IP based communications between the network elements will also bring down the operational costs of handling both voice and data, in addition to adding flexibility.

On The Way To 3G:

As reflected in the introduction above, the drive for 3G is the need for higher capacities and higher data rates. Whereas higher capacities can basically be obtained by having a greater chunk of spectrum or by using new evolved air interfaces, the data requirements can be served to a certain extent by overlaying 2.5G technologies on the existing networks. In many cases it is possible to provide higher speed packet data by adding few network elements and a software upgrade.

A Look At GPRS, HCSD, and EDGE:

Technologies like GPRS (General Packet Radio Service), High Speed Circuit Switched Data (HSCSD) and EDGE fulfill the requirements for packet data service and increased data rates in the existing GSM/TDMA networks. I'll talk about EDGE separately under the section "Migration To 3G". GPRS is actually an overlay over the existing GSM network, providing packet data sevices using the same air interface by the

addition of two new network elements, the SGSN and GGSN, and a software upgrade. Although GPRS was basically designed for GSM networks, the IS-136 Time Division Multiple Access (TDMA) standard, popular in North and South America, will also support GPRS. This follows an agreement to follow the same evolution path towards third generation mobile phone networks concluded in early 1999 by the industry associations that support these two network types.

The General Packet Radio Service (GPRS):

The General Packet Radio Service (GPRS) is a wireless service that is designed to provide a foundation for a number of data services based on packet transmission. Customers will only be charged for the communication resources they use. The operator's most valuable resource, the radio spectrum, can be leveraged over multiple users simultaneously because it can support many more data users. Additionally more than one time slots can be used by a user to get higher data rates. GPRS introduces two new major network nodes in the GSM PLMN:

Serving GPRS Support Node (SGSN) - The SGSN is the same hierarchical level as an MSC. The SGSN tracks packet capable mobile locations, performs security functions and access control. The SGSN is connected to the BSS via Frame Relay.

Gateway GPRS Support Node (GGSN) - The GGSN interfaces with external packet data networks (PDNs) to provide the routing destination for data to be delivered to the MS and to send mobile originated data to its intended destination. The GGSN is designed to provide inter-working with external packet switched networks, and is connected with SGSNs via an IP based GPRS backbone network.

A packet control unit is also required which may be placed at the BTS or at the BSC. A number of new interfaces have been defined between the existing network elements and the new elements and between the new network elements. Theoretical maximum speeds of up to 171.2 kilobits per second (kbps) are achievable with GPRS using all eight timeslots at the same time. This is about three times as fast as the data transmission speeds possible over today's fixed telecommunications networks and ten times as fast as current Circuit Switched Data services on GSM networks. Actually we may not see speeds greater than 64 kbps however it would be much higher than the

speeds possible in any 2G network. Also, another advantage is the fact that the user is always connected and is charged only for the amount of data transferred and not for the time he is connected to the network. Packet switching means that GPRS radio resources are used only when users are actually sending or receiving data. Rather than dedicating a radio channel to a mobile data user for a fixed period of time, the available radio resource can be concurrently shared between several users. This efficient use of scarce radio resources means that large numbers of GPRS users can potentially share the same bandwidth and be served from a single cell. The actual number of users supported depends on the application being used and how much data is being transferred. Because of the spectrum efficiency of GPRS, there is less need to build in idle capacity that is only used in peak hours.

Already many field trials and also some commercial GPRS implementations have taken place. GPRS is the evolution step that almost all GSM operators are considering. Also, coupled with other technologies like WAP, GPRS can act as a stepping stone towards convergence of cellular service providers and the internet service providers. HSCSD (High speed circuit switched data) is the evolution of circuit switched data within the GSM environment. HSCSD will enable the transmission of data over a GSM link at speeds of up to 57.6kbit/s. This is achieved by concatenating, i.e. adding together, consecutive GSM timeslots, each of which is capable of supporting 14.4kbit/s. Up to four GSM timeslots are needed for the transmission of HSCSD. This allows theoretical speeds of up to 57.6 kbps. This is broadly equivalent to providing the same transmission rate as that available over one ISDN B-Channel. HSCSD is part of the planned evolution of the GSM specification and is included in the GSM Phase 2 development. In using HSCSD a permanent connection is established between the called and calling parties for the exchange of data. As it is circuit switched, HSCSD is more suited to applications such as video conferencing and multimedia than 'busty' type applications such as email, which is more suited to packet switched data. In networks where High Speed Circuit Switched Data (HSCSD) is deployed, GPRS may only be assigned third priority, after voice as number one priority and HSCSD as number two. In theory, HSCSD can be preempted by voice calls- such that HSCSD calls can be reduced to one channel if voice calls are seeking to occupy these channels. HSCSD does not disrupt voice service availability, but it does affect GPRS. Even given preemption, it is difficult to see how HSCSD can be deployed in busy networks and still confer an

agreeable user experience, i.e. continuously high data rate. HSCSD is therefore more likely to be deployed in start up networks or those with plenty of spare capacity since it is relatively inexpensive to deploy and can turn some spare channels into revenue streams.

An advantage for HSCSD could be the fact that while GPRS is complementary for communicating with other packet-based networks such as the Internet, HSCSD could be the best way of communicating with other circuit switched communications media such as the PSTN and ISDN. But one potential technical difficulty with High Speed Circuit Switched Data (HSCSD) arises because in a multi-timeslot environment, dynamic call transfer between different cells on a mobile network (called "handover") is complicated unless the same slots are available end-to-end throughout the duration of the Circuit Switched Data call. Because of the way these technologies are evolving, the market need for high-speed circuit switched data and the market response to GPRS, the mobile infrastructure vendors are not as committed to High Speed Circuit Switched Data (HSCSD) as they are to General Packet Radio Service (GPRS). So, we may only see HSCSD in isolated networks around the world. HSCSD may be used by operators with enough capacity to offer it at lower prices, such as Orange. [1] Believes that every

Table 1.3. The Specifications of the Technologies

GENERATION	2G Technology	2G+ Technology	2.5G Technology	3G Technology
BENEFITS	Capacity, Battery life	Capacity, Cost, Data	Higher speed data	Multimedia
TECHNOLOGIES	GSM	▪ HSCSD ▪ SMS data	▪ GPRS packet radio ▪ EDGE	▪ W-CDMA (part of UMTS)
	cdmaOne	▪ IS95B	▪ 1XRTT ▪ HDR ▪ 1X Plus	▪ 3XRTT ▪ W-CDMA? (Japan, Korea)
	TDMA	▪ IS136+	▪ GPRS ▪ EDGE	▪ UWC 136
	PDC (Japan)	▪ Imode	(skip to 3G)	▪ W-CDMA

GSM operator in Europe will deploy GPRS, and by 2005 GPRS users will almost match the number of voice only users. Right now there are 300 million wireless phones in the world. By 2005 we expect one billion. Before I proceed, a quick look at the table below

would help you appreciate and understand clearly the technology characterizations as 2nd generation, 2.5 generation and 3G. We have looked into 2G and some 2.5G technologies so far.

Destination: Third Generation:

Standardization of 3G mobile systems is based on ITU (International Telecom Union) recommendations for IMT 2000. IMT 2000 specifies a set of requirements that must be achieved 100% for a network to be called 3G. By providing multimedia capacities and higher data rates, these systems will enhance the range and quality of services provided by 2G systems. The main contenders for 3G systems are wideband CDMA (W-CDMA) and cdma2000. The ETSI/ GSM players including infrastructure vendors such as Nokia and Ericsson backed W-CDMA. Cdma2000 was backed by the North American CDMA community, led by the CDMA Development Group (CDG) including infrastructure vendors such as Qualcomm and Lucent Technologies. Universal Mobile Telephone System (UMTS) is the widely used European name for 3G. The proposed IMT-2000 standard for third generation mobile networks globally is a CDMA-based standard that encompasses THREE OPTIONAL modes of operation, each of which should be able to work over both GSM MAP and IS-41 network architectures.

Mode Title Origin Supporters 1 Direct Sequence FDD (Frequency Division Duplex) based on the first operational mode of ETSI's UTRA (UMTS Terrestrial Radio Access) RTT proposal. Japan's ARIB and GSM network operators and vendors. 2 Multi-Carrier FDD (Frequency Division Duplex) Based on the cdma2000 RTT proposal from the US Telecommunications Industry Association (TIA). Cdma One operators and members of the CDMA Development Group (CDG). 3 Time Division Duplex (TDD) The second operational mode of ETSI's UTRA (UMTS Terrestrial Radio Access) RTT proposal. An unpaired band solution to better facilitate indoor cordless communications. Harmonized with China's TD-SCDMA RTT proposal.

UMTS is the European designation for 3G systems. The UMTS frequency bands selected by the ITU are 1,885 MHz - 2,025 MHz (Tx) and 2,110 MHz - 2,2,20 MHz (Rx). Higher frequency bands could be added in future if need be, for stationary data. There is still some confusion about all the frequency options, as FCC has not given clear indications so far. The following table should briefly give an idea about the 3G system specifications.

3rd Generation Initiatives:

3GPP (Third Generation Partnership Project) and 3GPP2 are the two alliances working towards the specification for the 3G systems. 3GPP partners are ETSI, TTC, ARIB, TTA, T1 and the 3GPP2 includes TIA, TTC, ARIB, and TTA. Although both have chosen CDMA as the technology behind the 3G systems, the systems advocated by these two groups are different. The 3GPP organizational partners have agreed to co-operate for the production of Technical Specifications for a 3rd Generation Mobile System based on the evolved GSM core networks and the radio access technologies that the Organizational Partners support (i.e. UTRA both FDD and TDD modes). 3GPP2 provides global specifications for ANSI/TIA/EIA-41 network evolution to 3G and global specifications for the RTTs (Radio transmission technologies) supported by ANSI/TIA/EIA-41. Yet another group, the Operators Harmonization Group, is dedicated to achieving the maximum possible level of commonality of technologies to maximize interworking of different versions. It was as a result of pushing by OHG that led to ITU's mixed solution to 3G air interfaces with ANSI-41 and GSM MAP networking.

3G Timeframes:

The actual deployment of 3G will not be a homogeneous occurrence. Japan will lead with the service in early 2001, followed by Western Europe in mid to late 2003. U.S. is expected to wait for some time at 2.5G and 2.75G before going in to true 3G. As I have mentioned earlier, with TDMA based networks like GSM and IS-136, increased capacity will be the initial driving factors. Therefore these networks will take a comparatively longer time to true 3G.

Evolving Today's Networks Towards 3G:

The 3rd Generation Mobile System will most likely grow out of the convergence of enhanced 2nd generation mobile systems with greater data transfer speed and capacity and 1st generation satellite mobile systems. Evolution to the current generation mobile networks to 3G doesn't necessarily mean seamless upgradation to the existing infrastructure to the 3G. Evolution should also be seen in context of coexistence of the 2G and 3G networks for some time, with users able to roam across the new and the old networks, able to access 3G services wherever 3G coverage is available. As mentioned before, a 3G network can have one of the 3 optional air interfaces supporting one of the two GSM MAP and IS-41 network architectures. This results in a range of choices for

the existing networks to evolve/migrate towards 3G. Possible convergence of TDMA and GSM networks with EDGE adds another variable to the overall migration paths. Another variable that adds complexity to this already complex list of options is the time frames involved. By the time some of the 2.5 or 2.75G technologies go to field, we may see the emergence of 3G technologies also. So, a lot of thought regarding the costs involved, and/or the viability of 2.5G technologies like EDGE could be questioned. The same is true about the time frames of the so called "4G".

Before I talk about evolution/migration paths of all the existing 2G mobile wireless technologies, let me briefly discuss the 3G-network architecture and other technology factors involved in the migration to 3G.

3G Architecture:

The 3G networks will have a layered architecture, which will enable the efficient delivery of voice and data services. A layered network architecture, coupled with standardized open interfaces, will make it possible for the network operators to introduce and roll out new services quickly. These networks will have a connectivity layer at the bottom providing support for high quality voice and data delivery. Using IP or ATM or a combination of both, this layer will handle all data and voice info. The layer consists of the core network equipment like routers, ATM switches and transmission equipment. Other equipment provides support for the core bit stream of voice or data, providing QOS etc. Note that in 3G networks, voice and data will not be treated separately which could lead to a reduction in operational costs of handling data separately from voice. The application layer on top will provide open application service interfaces enabling flexible service creation. This user application layer will contain services for which the end user will be willing to pay. These services will include eCommerce, GPS and other differentiating services. In between the application layer and the connectivity layer, will run the control layer with MSC servers, support servers, HLR etc. These servers are needed to provide any service to a subscriber.

Migration Strategies:

The migration to 3G is not just based on evolving core networks and the radio interface to IMT 2000 compliant systems. Migration towards 3G would also be based on the following steps/technologies:

Network upgrades in the form of EDGE, GPRS, HSCSD, CDPD, IS-136+ and HDR. Evolution to 2.5G basically will provide support for high-speed packet data. Though

these technologies are extensions to 2G rather than precursors to 3G these will have a major impact either by proving (or not) demand for specific services. Service trials to test infrastructure, handsets and applications etc

EDGE! Will TDMA and GSM ever meet:

EDGE is a new time division multiplexing based radio access technology that gives GSM and TDMA an evolutionary path towards 3G in 400, 800, 900, 1800 and 1900 MHz bands. It was proposed to ETSI in 1997 as an evolution to GSM. Although EDGE reuses GSM carrier bandwidth and time slot structures, it is not restricted to use in GSM cellular systems only. In fact, it can provide a generic air interface for higher data rates. It provides an evolutionary path to 3G. Some call it 2.5G. It can be introduced smoothly into the existing systems without altering the cell planning. But as with GPRS, EDGE doesn't provide any additional voice capacity. The initial EDGE standard promised mobile data rates of 384 kbps. It allows data transmission speeds of 384 kbps to be achieved when all eight timeslots are used. In fact, EDGE was formerly called GSM384. This means a maximum bit rate of 48 kbps per timeslot. Even higher speeds may be available in good radio conditions. Actual rates will be lower with rates falling as one goes away from the cell site. EDGE can also provide an evolutionary migration path from GPRS to UMTS by implementing now, the changes in modulation that will be necessary for implementing UMTS later. Both High Speed Circuit Switched Data (HSCSD) and GPRS are based on something called Gaussian minimum-shift keying (GMSK) which only yields a moderate increase in data bit rates per time slot. EDGE, on the other hand, is based on a new modulation scheme that allows a much higher bit rate across the air interface. This modulation technique is called eight-phase-shift keying (8 PSK). It automatically adapts to radio circumstances and thereby offers its highest rates in good propagation conditions close to the site of base stations. This shift in modulation from GMSK to 8 PSK is the central change with EDGE that prepares the GSM world (and TDMA in general) for UMTS.

Only one EDGE transceiver unit will need to be added to each cell. With most vendors, it is envisioned that software upgrades to the BSCs and Base Stations can be carried out remotely. The new EDGE-capable transceiver can also handle standard GSM traffic and will automatically switch to EDGE mode when needed. EDGE capable terminals will also be needed - existing GSM terminals do not support the new modulation techniques and will need to be upgraded to use EDGE network functionality.

EDGE is currently being developed in two modes: compact and classic. Compact employs a new 200 kHz control channel structure. Synchronized base stations are used to maintain a minimum spectrum deployment of 1 MHz in a 1/3-frequency reuse pattern. EDGE Classic on the other hand employs the traditional GSM 200 kHz control structure with a 4/12 frequency reuse pattern on the first frequency.

How Can GSM and TDMA Converge With EDGE:

While developing the 3G wireless technology for TDMA, the Universal Wireless Communication Consortium (UWCC) proposed the 136 High-Speed (136 H-S) radio interface as a means of satisfying requirements for IMT-200 radio transmission technology (RTT). After evaluating various proposals, UWCC adopted EDGE (Actually EGPS, EDGE+GPRS) as the outdoor component of 136HS to provide 384 kbps data services. Since GSM networks can also have an evolutionary path via EDGE, this presents an interesting opportunity where the air interfaces of TDMA and GSM can converge and then evolve together. EDGE is being developed concurrently in ETSI and UWCC. The phase one of EDGE emphasizes enhanced circuit-switched data (ECSD) and enhanced GPRS (EGPS).

The TDMA terminals that support 30 kHz circuit switched services scan for a 30 kHz control channel (DCCH) according to TIA/EIA 136 procedures. If an acceptable 200 kHz EGPRS carrier exists, a pointer to this system will be available on the DCCH. On finding this, the terminal will leave the 30KHz system and start scanning of the 200 kHz systems. When it finds it, it starts behaving as if it is a GSM/GPRS terminal. To answer a circuit switched page, the mobile suspends packet data traffic and starts looking for a 30 kHz control channel. Mobile terminals that only support 200 kHz carriers immediately start looking for 200 kHz packet data system.

Will this happen? While EDGE provides a common air interface for TDMA and GSM to converge, there is one possible problem. GSM operators may decide to skip EDGE altogether in their migration path to 3G. By the time EDGE will be commercially available for GSM systems, 3G will already be in sight with W-CDMA and since W-CDMA will need an entirely new air interface, the additional investments in EDGE, only to be replaced by another system seems a bit unjustified. EDGE has lost favor in Europe with some wireless operators and vendors that are not convinced it will actually be adopted in force once carriers move to GPRS. As described above, the belief is that wireless service providers may be more inclined to move straight to WCDMA

from GPRS. On the other hand, some North American operators have taken the position that they may not need to upgrade to WCDMA after EDGE because it doesn't offer increased speeds in the mobile environment (the ITU/UMTS definition of G3G is 384 Kbps mobile, 2 Mbps low mobility/fixed wireless). This is an especially strong point when one considers that the market demand for high-speed wireless data has yet to be fully proven. The convergence of TDMA and GSM can't be ruled out also. Particularly in the US, operators may have more interest in moving on to EDGE to get compatibility with the TDMA networks. According to a study [1], EDGE should be available in the North American markets by 2002.

Individual Technology Evolution Paths:

A variety of technologies/standards exist and therefore, so do the number of paths that can be taken. The table below briefly summarizes these standards (Table 1.4).

Table 1.4. Cellular Standards

Standard Name	Other Names (Aliases)	Upgrade Path for . . .	Expected Availability
Code Division Multiple Access (CDMA)	IS-95, IS-95A, cdmaOne	N/A	Current
Global System Mobile (GSM)	N/A	N/A	Current
1XRTT	G3G-MC-CDMA-1X, also called 2.5G step for CDMA	CDMA	End of 2001
General Packet Radio System (GPRS)	Also called 2.5G for GSM	GSM and potentially TDMA	End of 2000
Enhanced Data for GSM Efficiency (EDGE)	Also called 2.5G for GSM and TDMA	GSM or TDMA	End of 2001
Wideband CDMA (WCDMA)	WCDMA, FDD Mode 1 (Direct Sequence), G3G-DS-CDMA	GSM or TDMA, and in rare cases CDMA	2002 Europe, later for North America depending upon spectrum availability
cdma2000	3XRTT, FDD Mode 2 (Multicarrier), G3G-MC-CDMA-3X	CDMA	2003
High Data Rate (HDR)	HDR	Not a true 3G upgrade; a network extension using a CDMA base system	End of 2001

GSM and TDMA To 3G:

GSM and TDMA systems have more or less the same set of options for migrating to 3G. The path to 3G is not as simple in case of GSM/TDMA as is in the case of CDMA. The main evolutionary standards are GPRS, EDGE and, finally, W-CDMA. Vendors are positioning each of these standards as a step to the next, but operators are

not so sure. For an operator moving from GSM to GPRS to EDGE and then to W-CDMA, he'll have to make investments 3 times which won't be pleasing to any operator. As [1] suggests, at this time, there seem to be four basic options that GSM and TDMA operators are considering:

Install GPRS, then move straight to WCDMA;

Install EDGE, then move straight to WCDMA;

Install GPRS, then move to EDGE, then to WCDMA; or

Install EDGE, skip move to WCDMA, and wait for the next generation (4G) (see Figure 1.16)

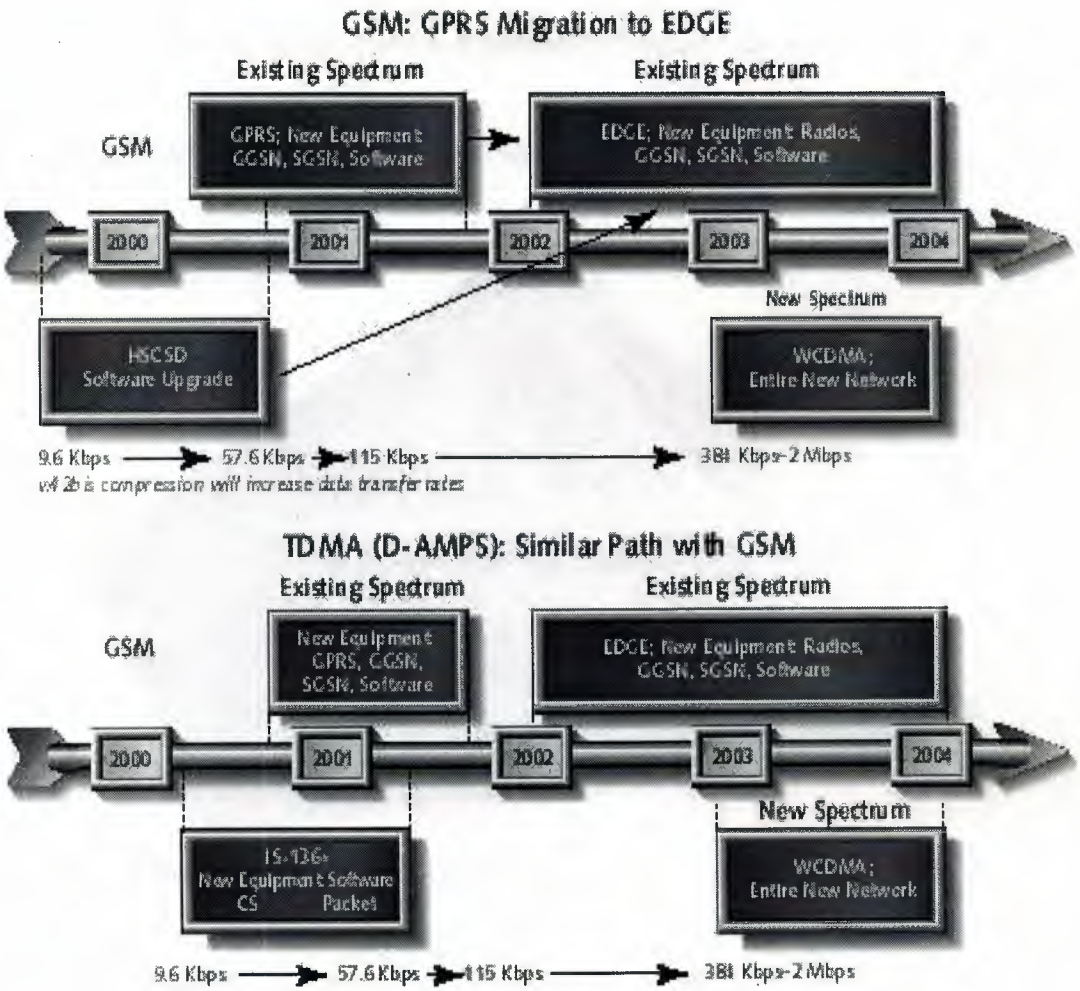


Figure 1.16. Technology path for the GSM operators

CDMA To 3G

While GSM and TDMA operators have multiple choices ahead for progressing to the next-generation networks, CDMA operators have a single path that truly builds upon itself. Currently all North American CDMA networks are based on IS-95 (cdmaOne), which can be setup to provide data rates upto 14.4 kbps. The next step is to have a software upgrade from IS-95A to IS-95B, which provides additional voice efficiencies giving additional capacity, and allows for up to 84-Kbps packet data. (We might not see 84kbps but instead 64kbps, initially.) While this migration does not need any additional hardware but as brought out by [1] most operators may decide not to move to IS-95B because of two reasons.

1. IS-95A in itself is relatively new and carriers have just launched their IS-95A data services.
2. By the time IS-95B becomes available, 1XRTT will be ready.

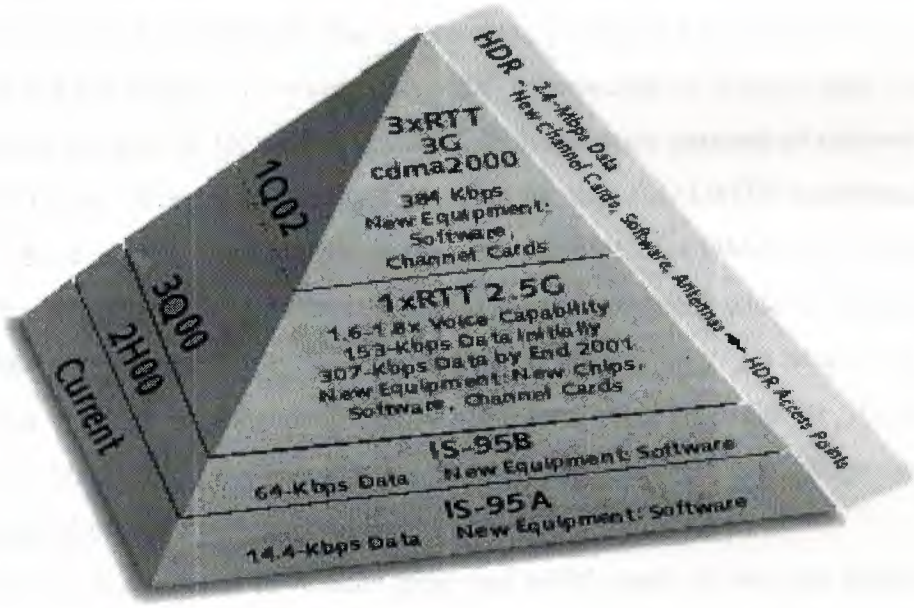


Figure 1.17. Options for the GSM operators

What Are The Costs?

In the shorter term, TDMA and GSM have a much more cost-effective upgrade option by means of moving to GPRS to be in a position to provide data services. As mentioned earlier, an upgrade to GPRS doesn't require substantial investments and existing GSM/TDMA service providers can upgrade to GPRS at around 28% cost of their initial 2G investments. The IS-95 upgrade path to 1xRTT is comparatively costly

at around 40% investments on the existing 2G networks. It should also be noted that IS-95A in itself has also not been in existence for long. However, in the final run to truly 3G networks, GSM/TDMA operators may have to incur much higher investments as shown in the figure below. The cost equations for TDMA or GSM may vary depending on the exact path taken (EDGE or no EDGE or only EDGE). CDMA has the unique advantage of having the same air interface in 2G as in 3G (same underlying technology).

Therefore, it is very probable that most CDMA carriers in North America will move straight to 1XRTT. 1XRTT is the first step in moving to the full ITU/UMTS-defined 3G standard. It has many features that make it completely different from IS-95B. It will provide more than double the data speeds available from IS-95B (153 Kbps vs. 64 Kbps); but, more importantly, in the spectrum-constrained market of North America, 1X will almost double the voice capacity. Additionally, the software and chip boards necessary for 1X are also an essential step to continue the upgrade to 3XRTT, which is also called G3G-MC-3X, but is also more popularly known by the trade name of cdma2000 (307 Kbps). However, cdma2000 is expected to provide only moderate voice capacity gains over 1X, and as such, 1X is the primary concern of carriers for the immediate future. Besides 1X and 3X paths to the ITU/UMTS-sanctioned G3G standards, there is also the Qualcomm-defined offshoot of CDMA--High Data Rate (HDR). This standard, which is proprietary to Qualcomm, sets aside a standard 1.25-MHz CDMA carrier specifically for data, and offers rates of up to 2.4 Mbps in a mobile environment. Though this standard achieves the data rates required for 3G, it is not considered a 3G standard because it is a data-only standard and has not been opened up for the approval of any standards bodies.

Several new standards have been proposed which don't fit into this classification of 2, 2.5 or 3G. These standards either provide only data services and/or provide much higher data rates than those specified by 3G systems. Examples are 1Xplus and 1XTREME. Since they use a single CDMA carrier they may be called 2.5G but then they provide much higher data rates than 3G. According to Motorola, 1XTREME will not require additional antennas as HDR will, and it will also keep data on the same spectrum as the voice services, meaning carriers won't have to devote any spectrum specifically to data services. 1XTREME is proposed to deliver the same voice capacity increases as standard 1X, and provide data rates approaching 1.4 Mbps. The second

iteration, expected to be in trials by the first quarter of 2001, is expected to deliver data rates as high as 5.2 Mbps. Motorola expects 1XTREME to be market-ready in the same time frame as HDR: by the end of 2001 to the first half of 2002.

Another interesting thing is that these so called 4G technologies may start appearing almost at the same time when 3G comes. It is not very clear as to how these developments will influence an already very complex set of equations.

Concluding Remarks:

Mobile communications are really poised to see major improvements in terms of capabilities of mobile networks. The next generation of wireless services, besides improving the overall capacity, will create new demand and usage patterns, which will in turn, drive the development and continuous evolution of services and infrastructure. While development of 3G networks will continue and pick up pace in the near future, the 2nd generation networks will keep evolving in terms of continuous enhancements and towards convergence of existing 2G standards. The initial 3G solutions should coexist with the 2G networks while slowly evolving to all 3G networks. While 3G in its true sense should have transparent roaming across all networks through out the world, given the penetration and the investments in the 2nd generation, true roaming (consistent service availability, across networks, independent of networks) looks to be to a very distant proposition!

CHAPTER 2. INTERFACES

2.1. Universal Serial Bus (USB)

USB, or Universal Serial Bus is a connectivity specification developed by computer and telecommunication industry members for attaching peripherals to computers. USB is designed to free all the troubles when installing external peripherals. It eliminates the hassle to open computer case for installing cards needed for certain devices. It is designed to meet Microsoft Plug and Play (PnP) specification, meaning users can install, and hot-swap devices without long installation procedures and reboots. Furthermore, it allows 127 devices to run at the same time on the bus. USB bus provides two types of data transfer speed 1.5Mbps and 12Mbps and it can provide a maximum of 500mA of current to devices attached on the bus. All these features will only need one interrupt to operate on a computer equipped with USB ports. Universal means all peripherals share the same connector. Serial simply defines devices can daisy chain together. We will now look at the different parts of USB.

The goal of the Universal Serial Bus (USB) design is to sweep the plethora of I/O ports on the PC into one serial channel. The bus runs at a base data rate of 12 Mbps but offers a 1.5-Mbps option that helps keep down the cost of low-performance devices, such as keyboards. USB's physical configuration is a tiered star. The PC acts as the host and root hub to which the user can attach devices or additional hubs. These additional hubs can, in turn, connect to a combination of devices and another hub layer. The bus supports as many as five hub layers and 127 devices on one host. The bus also can provide 5V power to attached devices. The control of system setup, device initialization, and data flow all reside with the host system. Upon system power-up, the host performs enumeration on each USB device. The host queries the device for a description, assigns the device a unique address for subsequent transactions, and sets the device's operating configuration. The host also identifies and loads hardware-specific drivers for the USB device into the operating system during device enumeration. Every millisecond (the USB frame time), the USB host initiates a series of data transactions. The USB offers two types of data transactions for high-volume data. Isochronous transactions offer guaranteed bandwidth, with the host system allocating 1 to 1023 bytes/frame for the transaction. Bulk transactions have no bandwidth guarantees. The host allocates to bulk data as much bandwidth as is left over after all other transactions are accounted for.

Bulk transactions have guaranteed delivery with resend ability, ensuring that all data arrives eventually.

All USB data transactions are host-initiated. Based on application program requests for data transfers, the host signals the USB device to start the data flow, regardless of the data's source or destination. USB devices cannot initiate a transfer; they can only respond to the host's command.

2.1.1. Inside USB

Essentially, there are three pieces to this USB technology -- host, hub and function. Host is actually the central point for all connections in the USB topology. It serves as the exchange point between each of the components of USB. The hardware implementation of host is called USB host controller, which is either integrated into the south bridge of the motherboards or included in USB add-on solutions. Hub allows multiple USB devices to share a single output to the USB host controller. Hubs on the back of computers are called root hubs. External USB hubs are available for users to connect more USB devices to the computer. Function is actually the USB device. Each USB device provides a function. Compound device provides multiple functions on the USB bus.

2.1.2. How fast is USB?

USB 1.1 provides high-speed and low-speed mode. In high speed mode, the host allows USB device to communicate and transfer data at 12Mbps; in low speed mode,

Table 2.1. Bandwidth of the other interfaces

Technology	Theoretical Maximum Throughput
Apple Desktop Bus (ADB)	10 kbps
Serial Port	230 kbps
Geoport Port	2 Mbps
USB at low speed	**1.5 Mbps**
USB at high speed	**12 Mbps**
SCSI	1 - 40 MBps
FireWire	400 Mbps
USB 2.0 at full speed	**360 - 480 Mbps**
Fast SCSI	8 - 80 MBps
Ultra SCSI-3	18 - 160 MBps

the host allows USB device to operate at 1.5Mbps. People tend to mix up between the capitalized "B" and small "b." Basically, small "b" stands for bits where capitalized "B" stands for bytes. 1 byte contains 8 bits. So, when you hear 1.5Mbps (Mega bits per second), you can determine the bytes per second by dividing 1.5Mbps by 8 to convert the unit to mega bytes per second. As you may have guessed, keyboards, mice and joysticks are among the low speed devices, using USB low-speed chips. Zip drives, scanners and printers are named as high-speed devices. USB host manages the bandwidth of each pipe used by different USB devices; 4 types of data transfer methods serve different kinds of USB devices. They are isochroous, interrupt, bulk, and control data transfers. As seen above, USB falls somewhere in the middle. So, do not expect USB to be a replacement of SCSI. USB will probably wipe out the ADB, serial, parallel ports in the next few years.

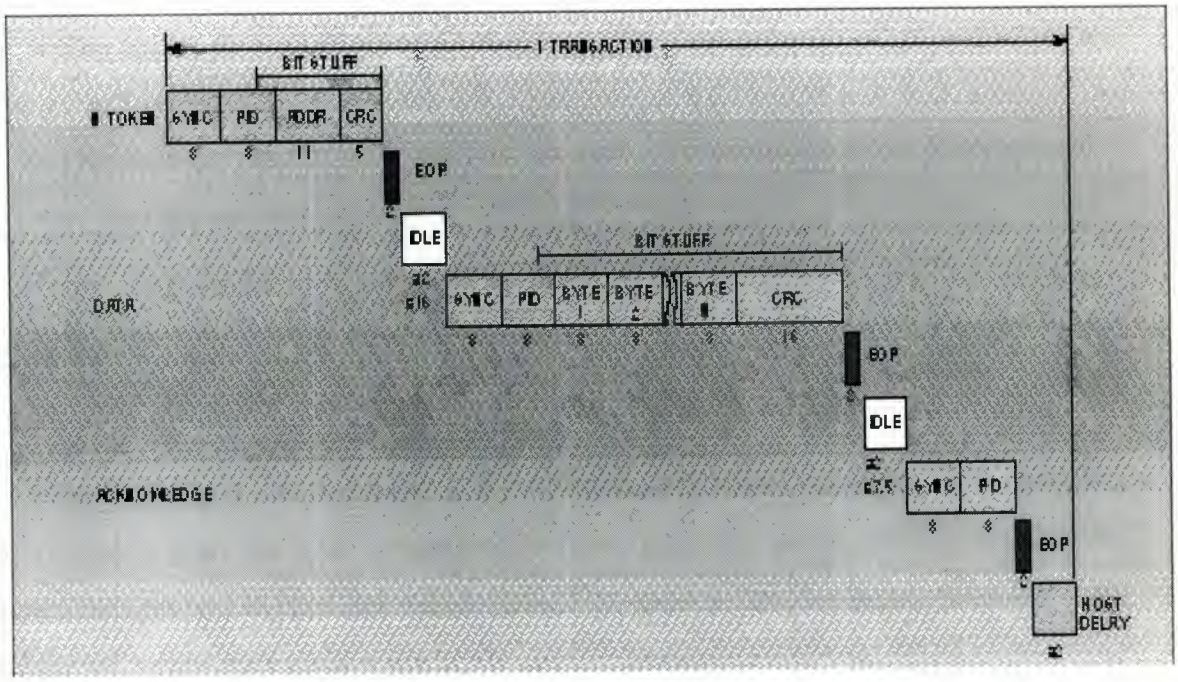


Figure 2.1. Scheme of USB Data Transmission

2.1.3. Architecture of USB

A USB bulk read-data transfer has three parts. First, the host requests data. Next, the device sends the data. Then, the host acknowledges the successful transfer. Delays are part of the protocol, but the host delay between transfers is unbounded (Figure 2.1)

2.2. IEEE 1394 (Firewire)

The emergence of digital video and multimedia applications has brought with it the need to move large amounts of data quickly between peripherals and PCs. And as audio/video products migrate to digital technology, both consumers and professionals alike stand to benefit from a simple high-speed connection that would make this transmission more efficient. Enter 1394: the digital cable. The IEEE 1394 serial bus is the industry-standard implementation of Apple Computer, Inc.'s FireWire digital I/O system. It is a versatile, high-speed, low-cost method of interconnecting a variety of personal computer peripherals and consumer electronics devices. Developed by the industry's leading technology companies, the specification was accepted as an industry standard by the IEEE Standards Board on December 12,1995.FireWire offers several advantages over other technologies. These benefits include:

- Guaranteed delivery of multiple data streams through isochronous data transport.
- The ability to connect up to 63 devices without the need for additional hardware, such as hubs
- Data transfer rates of up to 400 Mb/sec with 1.2 Gb / sec speeds in development.
- A flexible, six-wire cable.
- Complete plug-and-play operation, including the hot swapping of live devices.
- Acceptance by over 40 leading manufacturers in the computer and electronics consumer industries.

2.2.1. How does FireWire work?

Using special integrated circuits, FireWire multiplexes a variety of different types of digital signals such as compressed video, digitized audio, and device control commands on two twisted-pair conductors. The result is that FireWire's standard, 6-pin cables and connectors replace the myriad of I/O connectors currently found in consumer electronics equipment, PCs and peripherals. FireWire also employs isochronous data transfer to guarantee the delivery of multiple time-critical multimedia data streams. And, the protocol uses a "fairness" arbitration scheme to ensure that all nodes having information to send get a chance to use the bus.

2.2.2. ADVANTAGES OF IEEE-1394

Speed: up to 400 Mb/sec with 1.2 Gb/sec speeds in development.

Expandability Up to 63 devices supported.

Convenience Easy-to-use cable and connectors for plug-and-play and “hot swapping”.

Guaranteed data transfer Isochronous transport of multiple time-critical data streams. Low-cost flexible, six-pin cable for use in high-volume commercial markets.

2.2.3. Architecture

The 1394 standard defines two bus categories: backplane and cable. The backplane bus is designed to supplement parallel bus structures by providing an alternate serial communication path between devices plugged into the backplane. The cable bus is a "non-cyclic network with finite branches," consisting of bus bridges and nodes (cable devices). *Non-cyclic* means that you can't plug devices together so as to create loops. 16-bit addressing provide for up to 64K nodes in a system. Up to 16 cable hops are allowed between nodes, thus the term finite branches. A bus bridge serves to connect busses of similar or different types; a 1394-to-PCI interface within a PC constitutes a bus bridge, which ordinarily serves as the root device and provides bus master (controller) capability. A bus bridge also would be used to interconnect a 1394 cable and a 1394 backplane bus. Six-bit Node_IDs allow up to 63 nodes to be connected to a single bus bridge; 10 bit Bus_IDs accommodate up to 1,023 bridges in a system. This means, as an example, that the limit is 63 devices connected to a conventional 1394 adapter card in a PC. Each node usually has three connectors, although the standard provides for 1 to 27 connector per a device's physical layer or PHY. Up to 16 nodes can be daisy-chained through the connectors with standard cables up to 4.5 m in length for a total standard cable length of 72 m. (Using higher-quality "fatter" cables permits longer interconnections.) Additional devices can be connected in a leaf-node configuration, as shown in figure 1. All 1394 consumer electronic devices announced as of early 1997 have only a single connector; there are no currently are digital camcorders or VCRs that correspond to the devices with ID 3 or ID 5 shown in figure 1. Physical addresses are assigned on bridge power up (bus reset) and whenever a node is added or removed from the system, either by physical connection/disconnection or power up/down. No device ID switches are required and hot plugging of nodes is supported. Thus 1394 truly qualifies as a plug-and-play bus. The 1394 cable standard defines three signaling rates: 98.304, 196.608, and 393.216 Mbps (megabits per second; MBps in this thesis refers to megabytes per second.) These rates are rounded to 100, 200, and 400 Mbps, respectively, in this paper and are referred to in the 1394 standard as S100, S200 and S400. Consumer DV gear uses S100 speeds, but most 1394 PC adapter cards support the S200 rate. The slowest active node ordinarily governs the signaling rate for

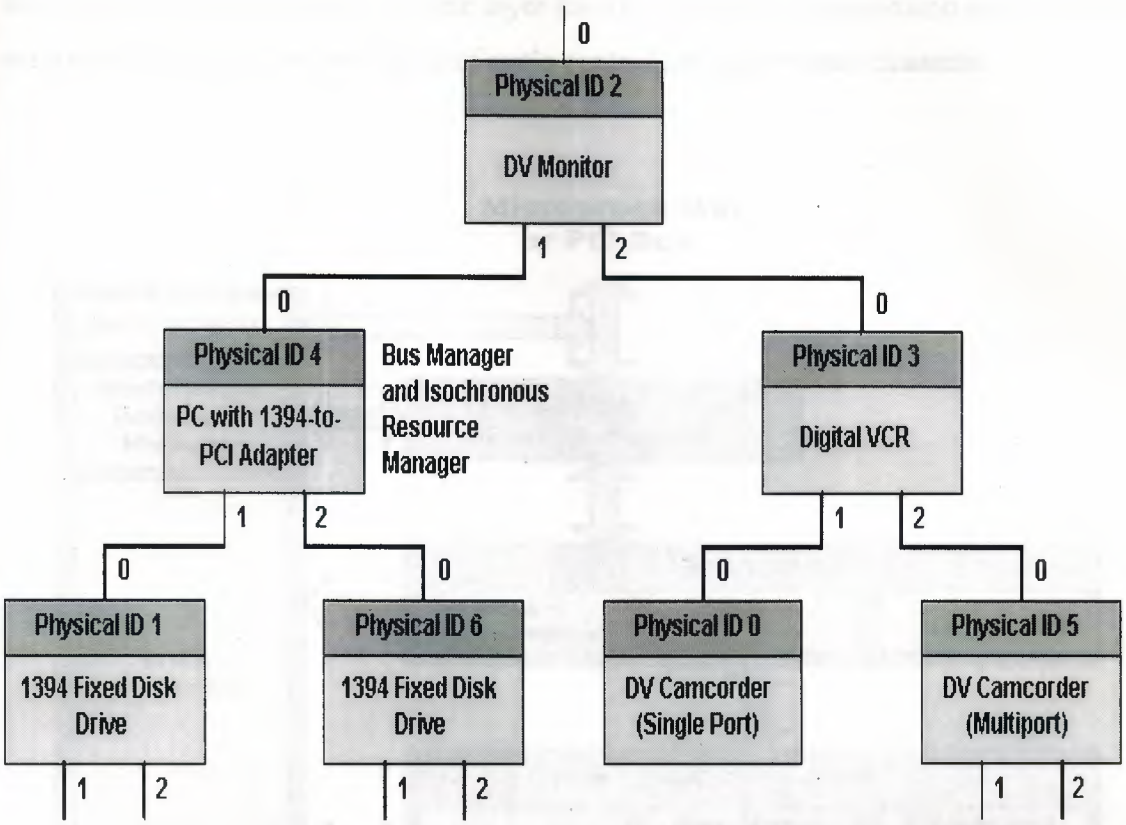


Figure 2.2. Topology of a typical PC-based 1394 bus system for DV applications.

the entire bus; however, if a bus master (controller) implements a Topology Map and a Speed Map for specific node pairs, the bus can support multiple signaling speeds between individual pairs. The 1394 Trade Association's 1394.1 working group presently are refining and clarifying the setup requirements for handling interconnected devices with multiple signaling speeds.

2.2.4. Physical, Link, and Transaction Layers

The three-stacked layers shown in figure 2 implement the 1394 protocol. The three layers perform the following functions:

The transaction layer implements the request-response protocol required to conform to the ISO/IEC 13213:1994 [ANSI/IEEE Std 1212, 1994 Edition] standard Control and Status Register (CSR) Architecture for Microcomputer Buses (read, write and lock). Conformance to ISO/IEC 13213:1994 minimizes the amount of circuitry required by 1394 ICs to interconnect with standard parallel buses. The link layer supplies an acknowledged datagram to the transaction layer. (A datagram is a one-way data transfer

with request confirmation.) The link layer handles all packet transmission and reception responsibilities, plus the provision of cycle control for isochronous channels.

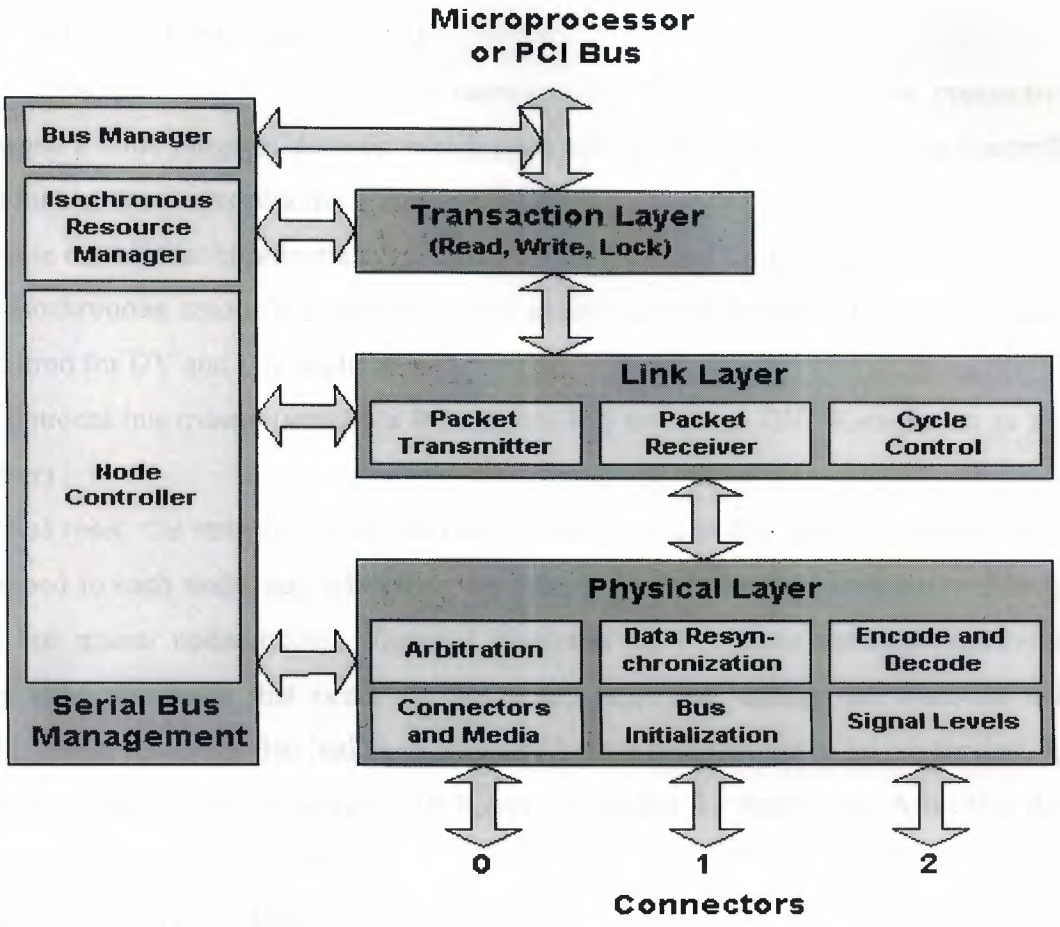


Figure 2.3. The 1394 Protocol Stack and Serial Bus Management Controller

The physical layer provides the initialization and arbitration services necessary to assure that only one node at a time is sending data and to translate the serial bus data stream and signal levels to those required by the link layer. Galvanic isolation may be implemented between the physical layer and the link layer using optical isolators; with isolation, the chip implementing the physical layer is powered by the bus conductors. Isolation should be provided where three-wire power cords are used to prevent ground loops through the green-wire ground; consumer devices, which use two-wire power cords or wall-wart power supplies, ordinarily don't require galvanic isolation.

The physical (PHY) layer is the bottleneck in 1394 systems. Historically, commercial PHY chips operated at half the potential data rate of link layer (LINK) chips (100 Mbps

vs. 200 Mbps, later 200 Mbps vs. 400 Mbps.) Texas Instruments announced in fall 1998 a set of 400-bps PHY chips that conform to the updated 1394a tentative specification and support the Open Host Controller Interface (OHCI) in conjunction with an OHCI-compliant link Chip.

2.2.5. 1394 Bus Management

1394 provides a flexible bus management system that provides connectivity between a wide range of devices, which need not include a PC or other bus controller. Bus management involves the following three services:

A cycle master that broadcasts cycle start packets (required for isochronous operation)

An isochronous resource manager, if any nodes support isochronous communication (required for DV and DA applications)

An optional bus master (usually a PC adapter, but an editing DVCR might act as a bus master)

On bus reset, the structure of the bus is determined, node IDs (physical addresses) are assigned to each node, and arbitration for cycle master, isochronous resource manager, and bus master nodes occurs. Figure 4 illustrates on a timeline the identification and arbitration processes that occur on bus reset. Note that during the 1-second delay isochronous resources that had been allocated before the reset are to be reallocated. Any resources that are not reclaimed will become available for future use. After that delay new resources may be allocated.

Isochroous Data Transport:

The isochronous data transport of the 1394 bus provides the guaranteed bandwidth and latency required for high-speed data transfer over multiple channels. The isochronous resource manager includes a BANDWIDTH_AVAILABLE register that specifies the remaining bandwidth available to all nodes with isochronous capability. On bus reset or when an isochronous node is added to the bus, the node requests a bandwidth allocation. As an example, a DV device would request approximately 30 Mbps of bandwidth, representing the 25+ Mbps DV data rate plus 3-4 Mbps for digital audio, time code, and packet overhead. Bandwidth is measured in bandwidth allocation units, 6,144 in a 125 ms cycle. (A unit is about 20 ns, the time required to send one data quadlet at 1,600 Mbps, called the S1600 data rate; the S1600 data rate is unlikely be supported in future implementations. A quadlet is a 32-bit word; all bus data is transmitted in quadlets.) 25 m s of the cycle is reserved for asynchronous traffic on the bus, so the default value of

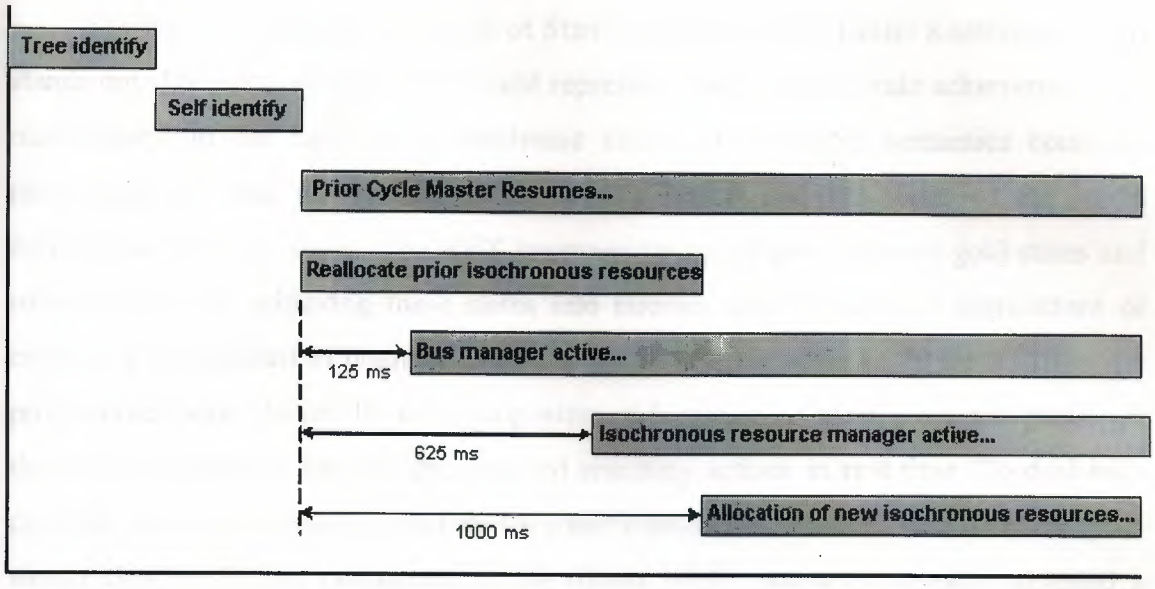


Figure 2.4. Bus reset timeline

the BANDWIDTH_AVAILABLE register on bus reset is 4915 units. In a 100-Mbps system, a DV device would request about 1,800 units; in a 200-Mbps system, about 900 units would be sufficient. If adequate bandwidth is not available, the requesting device is expected to repeat its request periodically.

The isochronous resource manager assigns a channel number (0 to 63) to nodes that request isochronous bandwidth based on values in the manager's CHANNELS_AVAILABLE register. The assigned channel number identifies all isochronous packets. When a node no longer requires isochronous resources, it is expected to release its bandwidth and channel number. As an example, the bus manager sends signals to cause a camcorder to commence talking on its channel and a record deck to commence listening on its channel for video data from the bus manager application. Device control is managed by asynchronous communication. Video acquisition for non-linear digital editing is simpler than the camcorder-DVCR example, because it requires only a single isochronous channel, plus an asynchronous path for device control. Timecode is built into the DV data, but asynchronous timecode transmission over the bus is useful when in camcorder or DVCR shuttle mode.

CHAPTER 3. INTELLIGENT ROBOTS

For the AI formalists, the work of Stan Rosenschein and Leslie Kaelbling (1986) stands out. They proved that if one could represent robot goals of state achievement and maintenance in the form of an electronic circuit, a consistent semantics could be maintained between the memory states of the circuit and the states of the world represented by these states. The REX language compiled propositional goal states and robot actions for achieving these states into circuits (actually c-based simulations of circuits) that executed in bounded time and usually on the order of 10 hertz. Thus, the programmer was allowed to use a propositional language to specify desired goals, yet the robot was able to execute the required resulting actions in real time. To deal with multiple goals that would contend for the robot's sensors or actuators (the REX compiler would flag conflicting commands to the robot), REX programs typically included a scheme to arbitrate among active circuits.

3.1. Animate Vision

What of the use of cameras for robots? Human and animal visions are the most powerful perception systems. However, we have seen how the processing of the low-level data alone, much less the addition of rapid control of a pan-tilt head, seemed to have little chance of fitting into the new paradigm. Fortunately, in the late 1980s, an analogous paradigm shift was taking place in the way researchers were approaching vision for agents. Again using ethology, several researchers began using animal visual behaviors as models for computational counterparts. For example, a frog primarily used its motion detection to catching flying food. Other animals keyed on specific aspects of the color spectrum for certain tasks. Indeed, psychophysical studies showed the human visual system to be, not surprisingly, even more adept. The human retina is arranged in such a manner as to have a higher concentration of receptors in the center and decreasing numbers radiating outward. Thus, humans don't process square arrays of data, for example, 512 x 512 x 8 bits, in a time step; rather, they use lower-resolution peripheral vision to watch for indications of motion or looming objects while they concentrate the higher-resolution center of the retina -the fovea- to reason about a specific object or part of an object in great detail. Humans don't take in everything at once in all its color and motion dimensions; instead, they concentrate on a narrow portion of their visual field.

Moreover, humans move this portion rapidly about the environment in patterns dictated by the task at hand and the last time step of visual information that was produced. For example, when looking at a picture of a group of people, if one is asked what the ages of the people are, one's eyes move in a pattern that concentrates on the faces of the people, with a few scans to determine the height of the people. To determine where a cup is in the picture, the eyes dart quickly about the picture for a table, then move to the objects on the table. Only when told that they must remember as many objects in the picture as possible will the eye-scanning machinery move in a pattern resembling the scan of a full image as found in the classic algorithms of computer vision (Ballard 1991).

Now computer vision paradigms were being recast into small, quick behaviors that not only dealt with a given field of view more efficiently but also where to next point the pan-tilt head of the camera. These well defined, compact routines, such as tracking a given color or attending to peripheral motion, were much like the behaviors being developed by the nouveau planning community and could now be incorporated as another part of the paradigm shift in programming robots.

3.2. A New Kind of Mapping

Just as AI had much to learn from the attempts to make robots intelligent, so did the robotics community stand to gain from the same endeavor? A good example was the use of maps for robot navigation. Early maps in the robotics community were geometric in nature, often as grids with each cell representing some amount of space in the real world. The grid had a single-coordinate system in which elements were represented. These maps became sophisticated at representing the spatial structure of the world (Moravec and Elfes 1985). It also was easy to do path planning and obstacle avoidance with geometric maps (for example, Lozano-Perez and Wesley [1979] and Brooks [1982]). However, geometric maps, as a part of the traditional world model of the robot, can require vast amounts of memory for large areas; in addition, the robot must know precisely where it is so that it can reason from the map or add to it. Just as with computer vision, trying to maintain accurate geometric maps was computationaly intensive and extremely difficult in real-world situations.

The solution, in keeping with the paradigm shift in vision, was the use of topological maps (Kuipers and Byun 1987; Brooks 1985). Patterned after how humans represent space, topological maps represent the world as a graph of places connected by

arcs, thus using no metric or geometric information, only the notions of proximity and order. With a topological map, the robot navigates locally from place to place, minimizing movement errors. Moreover, topological maps are clearly much more compact in their representation of space.

This notion was rapidly adopted by the AI and robotics communities. Kuipers and Byun (1991) continued their work on topological maps, producing a representation of space called the spatial semantic hierarchy. Another implementation of topological maps, by Kortenkamp and Weymouth (1994), used, both sonar and vision to determine places in a topological representation. The first part of this book introduces several other topological-based map representations and also some initial attempts at integrating topological and grid-based map representations.

3.3. Planning

Although the movement away from general representations was considered healthy, the resulting degree of specialization was viewed with some alarm. As Chuck Thorpe of Carnegie Mellon University once remarked about Brooks's Robots: "I wouldn't want one to be my chauffeur." In point of fact, many researchers exploring the new paradigm had no intention of throwing out the classic planning baby with the bath water. However, it was clear that planning in both its form and its function had to be rethought.

Two researchers involved in this rethinking by looking at the psychophysical aspects of human activity were Phil Agre and David Chapman (1987). Their research pointed to evidence that humans somehow put together plans for action based on the set of routine behaviors they can carry out. Moreover, logical decomposition planning is rarely invoked in the course of human affairs, and when it is, it serves primarily as a guide to the general direction in which one should head rather than a production of rigid sets of action.

During the late 1980s and early 1990s, several approaches along these lines were being pursued at once for intelligent robots. There were attempts to expand on the mobot approach (Maes 1990); others went further in the direction of enumerating all possible actions using planning prior to run time (for example, Kaelbling [1988] and Schoppers [1987]). Still others tried a combination of these approaches (for example, Bonasso [1991]).

One of the most important of these efforts was the work by Jim Firby (1989) on reactive action packages (RAPs). In his dissertation, Firby described a three-layered architecture with classic planning at the top, a reactive layer of behaviors at the bottom, and a middle layer with the goals of the resulting plan executed as dynamic sequences of these behaviors (that is, RAPs). When this framework was significantly expanded (Bonasso et al. 1995; Gat 1992), it became possible to program a large variety of robots -or any group of computer controlled machines for that matter- to carry out a variety of tasks over long duration in the vicinity of, and in concert with, human counterparts. Erann Gat's chapter in this book on the three layered approach explains why it has become a popular approach for the design and implementation of intelligent robots. We might call this approach P-SA; that is, the robot plans based on initial conditions and common knowledge (P) and then executes this plan using senseact (SA) behaviors, replanning only when the reactive behaviors run out of routine solutions. In this architecture, simple representations are tailored to specific tasks. Layered software allows behaviors such as obstacle avoidance to coordinate smoothly with behaviors such as path following. A new level of routines -cached plans- execute between the reactive behaviors and the central brain, and planning and other deliberate reasoning guide the procedures and behaviors in accomplishing the primary task and interacting with humans. In addition, there have been some remarkable advances in hardware. Plug-and-play subsystems that combine sensors and effectors are much more common.

Already the technical successes we have seen in the robot competitions over the years are finding their way into practical applications. Today unattended mobots fetch and carry in semi structured hospital environments. Vacuum mobots are used regularly in North American industry to clean large storage and staging spaces during off-work hours. Mobots are also used to semiautonomous explore uninhabitable venues such as Terran volcanoes and Martian Landscapes. The case studies in this book are ample proof that mobots have technically evolved to a point where, today, they are poised to help humankind in broad ranging tasks from mapping the ocean floor and long-term nursing home care to planetary colonization.

For mobots to move to the next level of competence necessary to complete such tasks, however, they need a broader base of technical support. It is our hope that, from this book, AI researchers will be inspired to expend additional effort in mobile robot research. One of the editors is fond of saying that "acting and sensing are still the

hardest parts." So naturally, new developments in robot perception and low-level control will always be necessary to advance the state of the art and meet the challenge of applications in difficult environments such as under water or outer space.

Mobots can benefit from all artificial intelligence disciplines, however, and, as we have previously explained, the robot architectures that support more traditional AI research are already in place. Mobots need to reason about their acts, both for feasibility and for rationality. Thus they can benefit from advances in planning and logical theories of sensing and acting. Most future robot tasks will involve working with humans. Consequently, spoken language generation and understanding must be developed for them to be effective team members. For missions of long duration -such as those involving deep sea or planetary exploration-mobots must adapt their behavior, and even their preferences over time. This requirement involves machine learning at all levels of competency.

The time has thus come, and the technology is here, for artificial intelligence and robotics to more closely join forces in improving the quality of life on earth and in establishing new civilizations in the cosmos.

3.4. Mapping And Navigation

While it is possible for a robot to be mobile and not do mapping and navigation, sophisticated tasks require that a mobile robot build maps and use them to move around. Levitt and Lawton (1990) posit three basic questions that de6ne mobile robot mapping and navigation:

- Where am I?
- How do I get to other places from here?
- Where are other places relative to me?

Each of the case studies in part one of this book address one or more of these questions. Each uses a different approach to representing and using spatial information. As such, they span the spectrum of options for mapping and navigation

On one side of the spectrum are purely metric maps. The robot's environment is defined by a single, global coordinate system in which all mapping and navigation takes place. Typically, the map is a grid with each cell of the grid representing some amount of space in the real world. These grids became quite sophisticated at representing the spatial structure of the world (see, for example, Moravec and Elfes [1985]). The case study of CARMEL Kortenkamp and his colleagues describes a mobile robot that uses a

grid-based approach to mapping and navigation. These approaches typically work in bounded environments, with little consistent structure and where the robot has opportunities to realign itself with the global coordinate system using external markers.

On the other side of the spectrum are qualitative maps, in which the robot's environment is represented as places and connections between places. Indeed, the idea of a map that contains no metric or geometric information, but only the notions of proximity and order, is enticing because such an approach eliminates the inevitable problems of dealing with movement uncertainty in mobile robots. Movement errors do not accumulate globally in qualitative maps as they do in maps with a global coordinate system since the robot only navigates locally, between places. Qualitative maps can also be more compact in their representation of space, in that they represent only interesting places and not the entire environment. Qualitative maps (also referred to as topological maps) have become increasingly popular in mobile robotics (see, for example, Brooks 1985; Kuipers and Byun 1991; and Kortenkamp and Weymouth 1994). The case studies by Nourbakhsh and by Koenig and Simmons describe the current state-of-the-art in qualitative mapping. These techniques work well in structured environments (i.e., office buildings) where there are distinctive places that are goals for the robot.

There have been efforts to combine metric and qualitative maps so that the strengths of both representations can be used (Asada et al. 1988; Kuipers and Levitt, 1988). The first case study in this part, by Thrun and his colleagues, gives an overview of both metric and topological mapping and describes their approach to integrating these two representations.

3.5 High Speed Obstacle Avoidance, Map Planning, Navigation Feedback

3.5.1 Obstacle Avoidance

Obstacle avoidance is performed a mobile robot called CARMEL (in this thesis it is taken as an example). A key to CARMEL's success was its ability to deal with sonar sensor noise. This robot used a novel algorithm called EERUF (error-eliminating ultrasonic firing) to allow for rapid firing and sampling of ultrasonic sonar sensors, which means faster obstacle avoidance. EERUF allows the robot to detect and reject ultrasonic noise, including crosstalk. The sources of ultrasonic noise may be classified as external sources, such as ultrasonic sensors used on another mobile robot operating in the same environment; or internal sources, such as stray echoes from other on-board

Faculty of Engineering

Department of Computer Engineering

WIRELESS DATA COMMUNICATION FOR ROBOT CONTROL

(An Application of Mobile Robot)

Graduation Project

COM-400

Student: Murat Küçük (980240)

Supervisor: Prof. Dr. Fakhreddin Mamedov

Lefkoşa - 2001

Table of Contents

ACKNOWLEDGMENT

First of all I would like to thank to Prof. Dr. Fakhreddin Mamedov who was supervisor of my project and to Assist. Prof. Dr. Rahib Abiyev who helped me preparing this project. With their endless knowledge I easily overcome many difficulties and learn a lot of things about Communication and Control Systems. Preparing this project is a nice experience of my life.

Also I would like to thank to all my friends, my family, and all my instructors because they never leave me alone and always try to help me during my education. Without their encouragements I would not be where I am now.

ABSTRACT

Graduation Project is devoted to the investigation of wireless communication and its application in mobile robot control. Problems in data communication, transmission media, different type of modulations, demodulations and shift-keying problems are considered. The structure of wireless LANs is given. The future evolution of wireless communication is given. In second chapter the two types of interfaces - USB and FireWire are considered. Their comparisons with other interfaces are given. In the last chapter the use of wireless communication in intelligent mobile robot control is considered. The structure and operation principle of mobile robot through wireless communication are given. At the end an application of wireless communication to mobile robot control is considered. Animate vision, mapping and navigation, planning, High-speed obstacle avoidance, vector field histogram and global path planning are considered.

INTRODUCTION

In this thesis, I tried to present wireless data communication for robot control systems. The aim of my researches is to implement a mobile robot moving from the source point to destination point. To perform such an approach I took mobile robot hardware and design to add some electronic components on it. As the title of my thesis is wireless data communication for robot control systems a wireless modem is added for the data communication, for autonomous movements distance measuring circuits are added, for navigation feedback a CCD sensor and a DSP is added to get the navigational data array. This is a new topic on mobile robots to control the movements of it. And in addition a temperature sensor and its driving circuit is added just for monitoring the temperature from the remote point onto the monitor. A Camera can be added for taking the frames while the robot is moving instead of monitoring temperature. With this robot my aim is to make the robot move from source to destination point and monitoring its path with navigation feedback and drawing the edges that are detected with its distance sensors and as a result to monitor the temperature values on the computer screen.

At the end of this thesis, a scheme of a 56-kilobaud synchronous RF modem with a 70 kHz bandwidth is given. The modulation in this modem is bandwidth limited MSK generated by a digital state machine driving two digital-to-analog converters, and two double balanced modulators. The carrier phase is shifted plus or minus 90 degrees for each bit. Demodulation is accomplished with a standard quadrature detector chip but various coherent methods can be used for operation at lower signal to noise ratios.

The distance measuring circuit scheme is given. It works with sonar. It sends sounds and receives the signal and calculates the time and generates the distance between obstacle and the robot.

The navigation feedback circuit is given. It has a CCD sensor on the chip and a DSP. As described above it takes 1800 fps. And the DSP processes the images. It generates the navigational information. It is a new topic in robotic applications. The circuit is HDNS-2000 by Agilent Semiconductor.

CHAPTER 1. CURRENT TECHNOLOGIES IN WIRELESS COMMUNICATION AND MOBILE COMPUTING

1.1. Historical Overview

It started with the Telegraph …

> "Electric telegraph is called the most perfect invention of modern times as anything more perfect than this is scarcely conceivable, and it thought what will be left for the next generation, upon which to expend the restless energies of the human mind." [An Australian newspaper, 1853.]

Origins of Coded Transmission:

- 1793, Revolutionary France
  - Aerial Telegraph, invented by Claude Chappe
  - Extensive network throughout France
- 1840s, Samuel F. B. Morse
  - Coded transmission via electronic means
  - Rapidly spread throughout US and Europe
  - International Telegraph Union (ITU) formed in 1865

Submarine Telegraphy: High Tech of the late 19th Century:

- 1850: Dover-to-Calais, first submarine line
- 1858: First transatlantic cable
  - Breaks after 3 months!
  - President Buchanan & Queen Victoria exchange telegrams
- 1866: Relaid with higher quality cable
  - Development of cable materials, technology of laying, repair
- Typical "Performance":
  - 1870: London to Bombay in 4 minutes, 22 seconds
  - 1901: London to British Guiana, 22 minutes
  - 1924: Telegram around the world in 80 seconds!

Radio Telegraphy (also know as "Wireless"):

- Radio technology
  - Communicate with ships and other moving vehicles
  - Messages sprayed into the "ether" crossing wide boundaries
  - Downfall of the nationally supported monopolistic telegraph companies

- 1896: Guglielmo Marconi
  - First demonstration of wireless telegraphy
  - Built on work of Maxwell and Hertz to send and receive Morse Code
  - Based on long wave (>> 1 km), spark transmitter technology, requiring very large, high power transmitters
  - First used by British Army and Navy in the Boer War
  - 1899: Reported to shore America's Cup yacht races

Wireless:

- 1907: Commercial Trans-Atlantic Wireless Service
  - Huge ground stations: 30 x 100m antenna masts
  - Beginning of the end for cable-based telegraphy
- WW I: Rapid development of communications intelligence, intercept technology, cryptography
- 1920: Marconi discovers short-wave (<100 m) radio
  - Long wave follow contour of land
    - Very high transmit power, 200 KW+
  - Short waves reflect, refract, and absorb, like light
    - Bounce off ionosphere
    - Higher frequencies made possible by vacuum tube (1906)
    - Cheaper, smaller, better quality transmitters

Other Important Dates:

- 1915: Wireless voice transmission NY to SF
- 1920: First commercial radio broadcast (Pittsburgh)
- 1921: Police car dispatch radios, Detroit
- 1935: First telephone call around the world
- WW II: Rapid development of radio technology
- 1968: Carter phone decision
- 1974: FCC allocates 40 MHz for cellular telephony
- 1982: European GSM and Inmarsat established
- 1984: Breakup of AT&T
- 1984: Initial deployment of AMPS cellular system

1.2. Radio Fundamentals

Radio Waves! Portable, even hand-held, short wave transmitters can reach thousands of miles beyond the horizon. Tiny microwave transmitters aboard space probes return data from across the solar system. And all at the speed of light. Yet before the late 1800s there was nothing to suggest that telegraphy through empty space would be possible even with mighty dynamos, much less with insignificantly small and inexpensive apparatus. The Victorians could extrapolate from experience to imagine flight aboard a steam-powered mechanical bird or space travel in a scaled-up Chinese skyrocket. But what experience would even have hinted at wireless communication? The key to radio came from theoretical physics. Maxwell consolidated the known laws of electricity and magnetism and added the famous displacement current term, $\partial D/\partial t$. By virtue of this term, a changing electric field produces a magnetic field, just as Faraday had discovered that a changing magnetic field produces an electric field. Maxwell's equations predicted that electromagnetic waves could break away from the electric currents that generate them and propagate independently through space with the electric and magnetic field components of the wave constantly regenerating each other.

Maxwell's equations predict the velocity of these waves to be $1/\sqrt{\varepsilon_0 \mu_0}$ where the constants ε_0 and μ_0 can be determined by simple measurements of the static forces between electric charges and between current-carrying wires. The dramatic result is, of course, the experimentally known speed of light, 3×10^8 m/s. The electromagnetic nature of light is revealed. Hertz conducted a series of brilliant experiments in the 1880s in which he generated and detected electromagnetic waves with wavelengths very long compared to light. The distribution of wavelengths can be seen in Figure 1.1. The utilization of Hertzian waves (the radio waves we now take for granted) to transmit information developed hand-in-hand with the new science of electronics.

Where is radio today? AM radio, the pioneer broadcast service, still exists along with FM, television, and two-way communication. Now radio also includes radar, surveillance, navigation and broadcast satellites, cellular telephones, remote control devices, and wireless data communications. Applications of radio frequency (RF) technology outside radio include microwave heaters, medical imaging systems, and cable television.

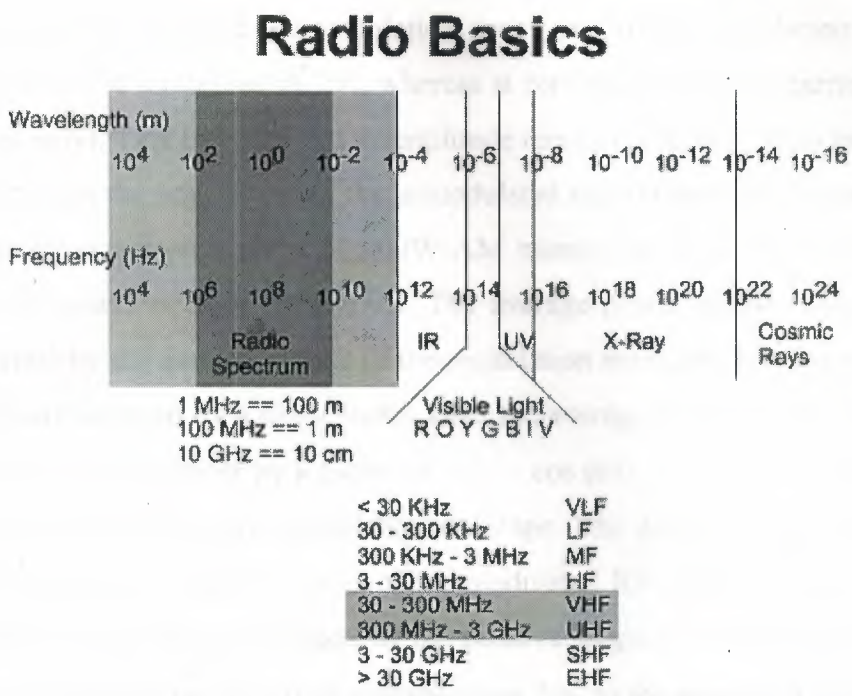


Figure 1.1. Radio Spectrum

1.3. Analogue Modulation Techniques

1.3.1. Amplitude Modulation

Modulation means adding information to an otherwise pure sinusoidal carrier wave by varying the amplitude or the phase (or both). The simplest, amplitude modulation (AM) is on/off keying. This binary AM can be accomplished with just a switch (telegraph key) connected in series with the power source. The earliest voice transmissions used a carbon microphone as a variable resistance in series with the antenna. Amplitude modulation is used in the long-wave, middle-wave, and short wave broadcast bands. Without modulation (when the music or speech is silent) the voltage and current at the antenna are pure sine waves at the carrier frequency. The rated power of a station is defined as the transmitter output power when the modulation is zero. The presence of an audio signal changes the amplitude of the carrier. The audio signal (amplified microphone voltage) has positive and negative excursions, but its average value is zero. The audio voltage is bounded by $+V_m$ and $-V_m$. A dc bias voltage of V_m volts is added to the audio voltage. The sum, $V_m + V_{audio}$, is always positive, and is used to multiply the carrier wave, $sin\ (\omega_c t)$. The resulting product is the AM signal; the

amplitude of the RF sine wave is proportional to the biased audio signal. The simulation in Figure 1.2 shows the various waveforms in the transmitter and receiver. The biased audio waveform is called the modulation envelope. At full modulation where $V_{audio} + V_m$, the carrier is multiplied by $2V_m$ whereas at zero modulation the carrier multiplied by V_m (bias only). This factor of two in amplitude means the fully (100%) modulated signal has four times the peak power of the unmodulated signal (carrier wave alone). It follows that the antenna system for a 50,000W AM transmitter must be capable of handling 200,000W peaks without breakdown. The average power of the modulated signal is determined by the average square of the modulation envelope. For example, in the case of 100% modulation by a single audio tone, the average power of the modulated signal is greater than the carrier by a factor of $< (1 + \cos(\theta))^2 >= 3/2$. Receiver demodulates the signal by detecting the modulation envelope. The detector is just a rectifier diode that eliminates the negative cycles of the modulated RF signal. A simple *RC* low-pass then produces the average voltage of the positive loops. (The average voltage of these sinusoidal loops is just their peak voltage times $2/\pi$, so the average is proportional to the peak, that is, the envelope.) Finally, ac coupling removes the bias, leaving an audio signal identical to the signal from the microphone. Figure 1.2 shows a basic Amplitude Modulation.

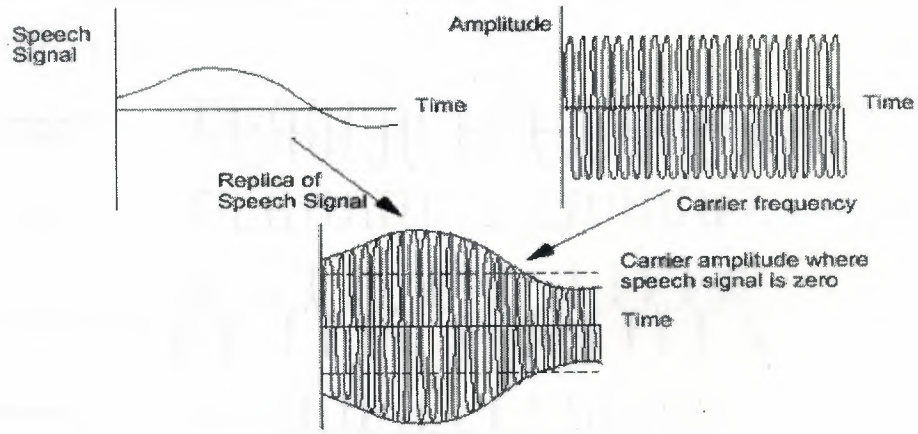


Figure 1.2. Amplitude Modulation

1.3.2. Frequency Modulation

Noise has a greater effect on amplitude than frequency.

Sufficient to detect zero crossings to reconstruct the signal

Easy to eliminate amplitude distortion

Constant envelope, i.e., envelope of carrier wave does not change with changes in modulated signal. This means that more efficient amplifiers can be used, reducing power demands. Transmitted signal can be seen in Figure 1.3

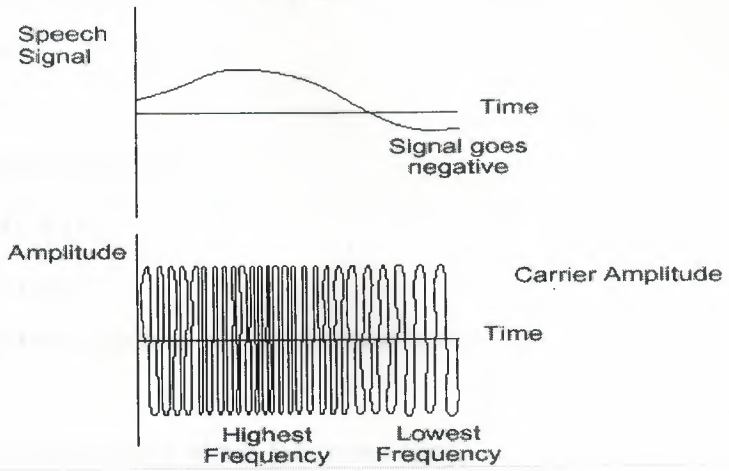


Figure 1.3. Frequency Modulation

Detection of FM Signal:

Noise translates into amplitude changes, and sometimes frequency changes.

Detection based on zero crossings: the limiter.

Alternative schemes to translate limited signal into bit streams. The steps are showed in figure 1.4.

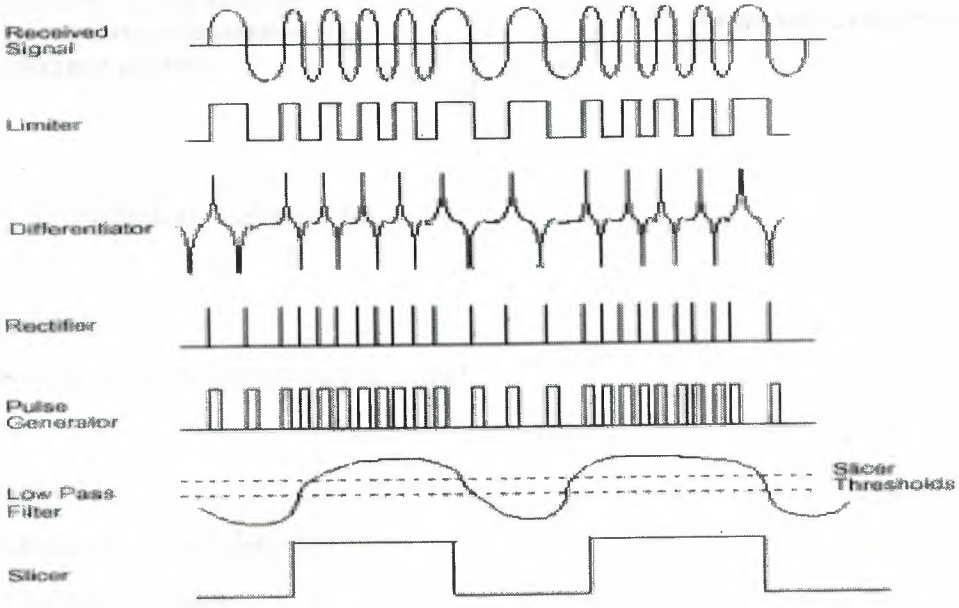


Figure 1.4. Steps of detecting FM signal

1.4. Digital Modulation Techniques

Carrier wave s:

$S(t) = A(t) * \cos[\theta(t)]$

Function of time varying amplitude A and time varying angle θ

Angle θ rewritten as:

$\theta(t) = \omega_0 + \varphi(t)$

ω_0 radian frequency, phase φ (t)

$S(t) = A(t) \cos[\omega_0 t + \varphi(t)]$

ω Radians per second

Relationship between radians per second and hertz

$\omega = 2\pi f$

Modify carrier's amplitude and/or phase (and frequency)

Constellation: Vector notation/polar coordinates. Figure 1.4 describes the technique for the basic digital modulation technique.

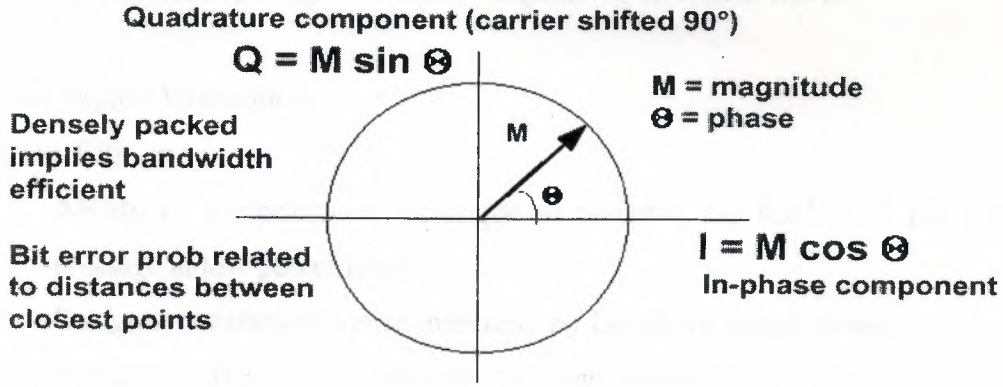


Figure 1.5. Quadrature Components

Demodulation:

Process of removing the carrier signal

Detection:

Process of symbol decision

Coherent detection

Receiver users the carrier phase to detect signal

Cross correlate with replica signals at receiver

Match within threshold to make decision

Noncoherent detection

Does not exploit phase reference information

Less complex receiver, but worse performance

Table 1.1. Coherent and Noncoherent Techniques

Coherent	Noncoherent
Phase shift keying (PSK)	FSK
Frequency shift keying (FSK)	ASK
Amplitude shift keying (ASK)	Differential PSK (DPSK)
Continuous phase modulation (CPM)	CPM
Hybrids	Hybrids

Coherent (aka synchronous) detection: process-received signal with a local carrier of same frequency and phase.

Noncoherent (aka envelope) detection: requires no reference wave.

Metrics for Digital Modulation:

- Power Efficiency
  - Ability of a modulation technique to preserve the fidelity of the digital message at low power levels
  - Designer can increase noise immunity by increasing signal power
  - Power efficiency is a measure of how much signal power should be increased to achieve a particular BER for a given modulation scheme
  - Signal energy per bit / noise power spectral density: E_b / N_0
- Bandwidth Efficiency
  - Ability to accommodate data within a limited bandwidth
  - Tradeoff between data rate and pulse width.
  - Throughput data rate per hertz: R/B bps per Hz
- Shannon Limit: Channel capacity / bandwidth
  - $C/B = \log 2 (1 + S/N)$

Criteria on selecting the right modulation:

- High spectral efficiency

- High power efficiency
- Robust to multi-path effects
- Low cost and ease of implementation
- Low carrier-to-cochannel interference ratio
- Low out-of-band radiation
- Constant or near constant envelope
  - Constant: only phase is modulated
  - Non-constant: phase and amplitude modulated

1.4.1. Amplitude Shift Keying (ASK)

The amplitude of the carrier c (t) is varied to represent binary of 1 or 0.Both frequency and phase remains constant. It is shown in figure 1.6.

The technique in ASK is called "On-Off-Keying" (OOK). In OOK no voltage represents one of the bit values (for example 0). A bit duration Tb is the interval of time that defines one bit. The amplitude of carrier c(t) is switched between two levels depending on the bits (0 or 1). Which voltage represents 1 and which represents 0 is left to the system designers. The speed of transmission using ASKS is limited by the physical characteristics of the transmission medium.

The advantage is a reduction in the amount of energy required to transmit information.

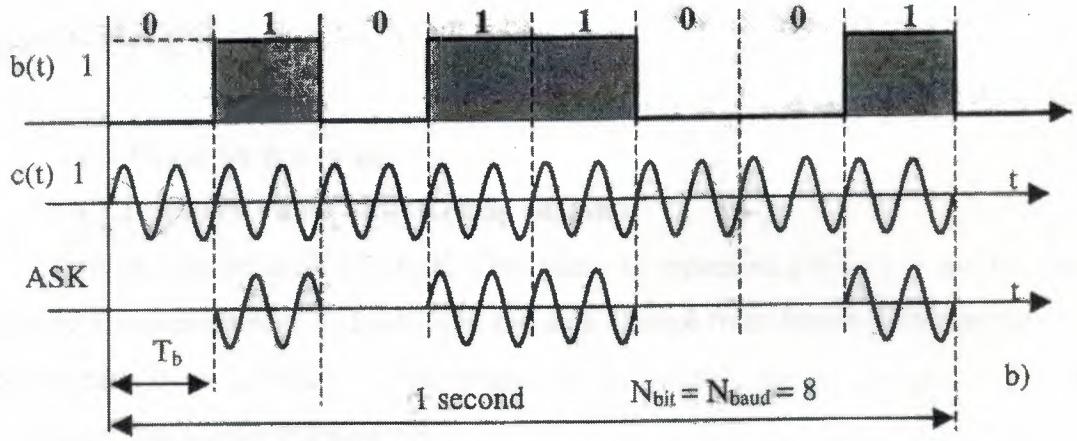


Figure 1.6. ASK signal

1.4.2. Frequency Shift Keying (FSK)

1/0 represented by two different frequencies slightly offset from carrier frequency in FSK. Two fixed amplitude carrier $c_1(t)=\cos 2\pi f_{c1}t$ and $c_2(t)=\cos 2\pi f_{c2}t$ one for binary 0 one for binary 1.The frequency of the signal during each bit duration is constant and its value depends on the bit (0 or 1). Figure 1.7 gives the conceptual view of FSK. FSK avoids most of the noise problem of ASK. As the receiving device is looking for specific frequency changes over a given number of periods, it can ignore voltage spikes.

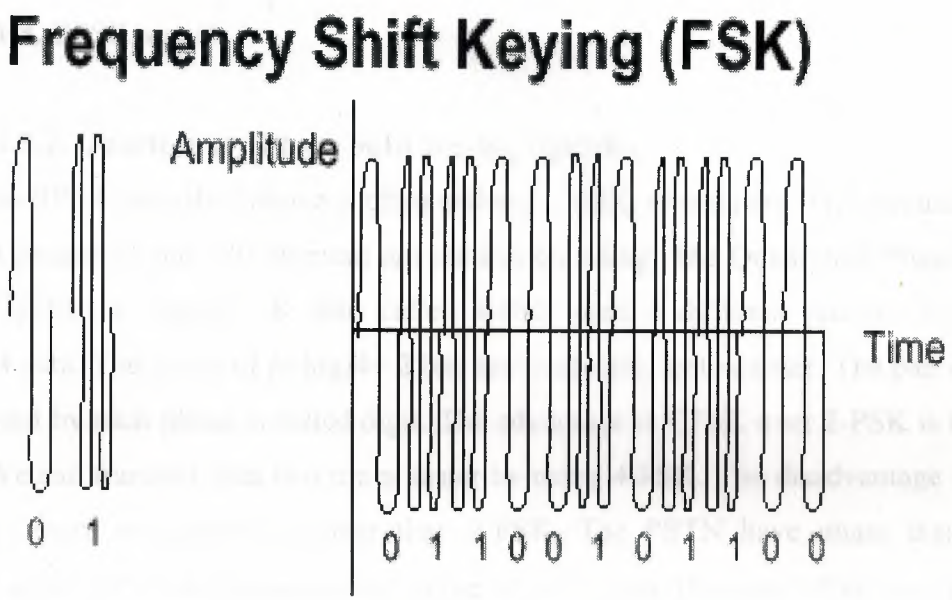


Figure 1.7. FSK signal

1.4.3. Phase Shift Keying

1.4.3.1. Binary Phase Shift Keying (BPSK)

Two phases are used in BPSK. One phase to represent a binary 0 and the other phase to represent binary 1. Each time the data change from binary 1 to a binary 0 or from binary 0 to a binary 1, the phase of transmitted signal changes 180°. Its characteristic is shown in figure 1.8.

- Simple to implement, inefficient use of bandwidth
- Very robust, used extensively in satellite communications

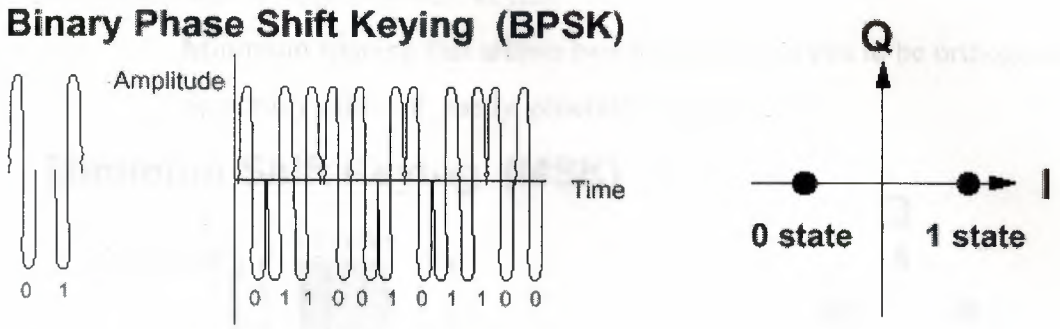


Figure 1.8. BPSK signal

1.4.3.2. Quarternary Phase Shift Keying (QPSK)

The BPSK described above is often called 2 – PSK, or ordinary PSK, because two different phases (0 and 180 degrees) are used in encoding. The Quadrature Phase-Shift Keying QPSK, in figure 1.8, also called 4-PSK uses 4 different phases (M=4) to represent data. The group of $n=\log_2 4= 2$ bits are modulated onto carrier. The pair of bits represented by each phase is called digit. The advantage is QPSK over 2-PSK is higher speed. We can transmit data two times faster by using 4-PSK. The disadvantage is that QPSK is more susceptible to error than 2-PSK. The PSTN have phase distortion (achieve up to 20°), which causes error in the received data. Because 2-PSK uses a 180° phase shift and it can tolerate phase tolerance approaching 90°. QPSK tolerate telephone circuit phase tolerance approaching 45°.

- Multilevel modulation technique: 2 bits per symbol
- More spectrally efficient, more complex receiver.

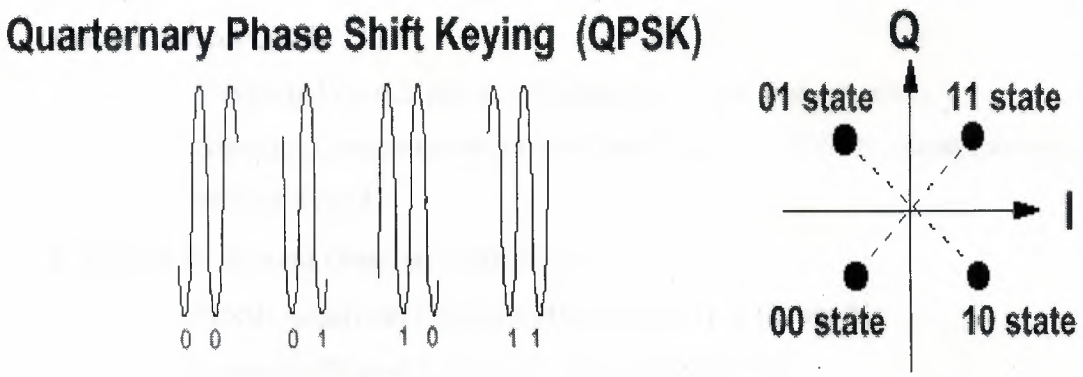


Figure 1.9. QPSK signal

1.4.3.3. Minimum Shift Keying

- Special form of frequency shift keying
  - Minimum spacing that allows two frequencies states to be orthogonal
  - Spectrally efficient, easily generated (Figure 1.10)

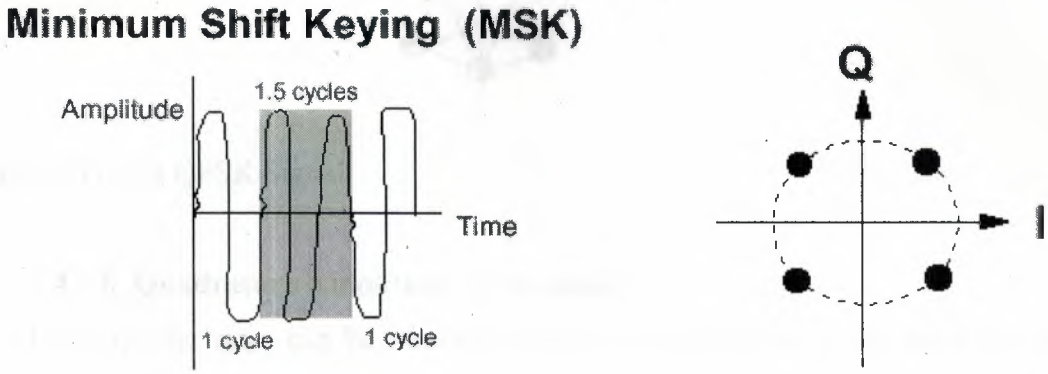


Figure 1.10. MSK Signal

1.4.3.4. Gaussian Minimum Shift Keying (GMSK)

- MSK + premodulation Gaussian low pass filter
- Increases spectral efficiency with sharper cutoff
- Used extensively in second generation digital cellular and cordless telephone applications
  - GSM digital cellular: 1.35 bps/Hz
  - DECT cordless telephone: 0.67 bps/Hz
  - RAM Mobile Data

1.4.3.5. p/4-Shifted QPSK

- Variation on QPSK
  - Restricted carrier phase transition to +/- p/4 and +/- p/4
  - Signaling elements selected in turn from two QPSK constellations, each shifted by p/4
- Popular in Second Generation Systems
  - North American Digital Cellular (IS-54): 1.62 bps/Hz
  - Japanese Digital Cellular System: 1.68 bps/Hz
  - European TETRA System: 1.44 bps/Hz
  - Japanese Personal Handy Phone (PHP)

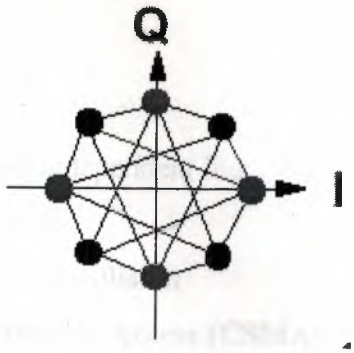


Figure 1.11. p/4 QPSK Signal

1.4.3.6. Quadrature Amplitude Modulation

Data transfer rates can be increased further by decreasing phase angle between two adjacent pharos. Four bits, or a quad bit, for example can be encoded into 16 possible phase changes (M=16). The phase differential between adjacent phasers would amount to 22.5° (360°/16=22.5°). The problem here, however, is that any phase shift 11.25° degree 2ill be within of the phase distortion introduced by PSTN (11.25°<20°). For this reason, 16 phase PSK is generally not used. To avoid the problem of phase jitter, the combination of ASK and PSK called Quadrature Amplitude Modulation (QAM) are used. Possible variation of QAM is numerous. Theoretically any measurable number of changes in amplitude can be combined with any measurable number of changes in phase. In the Figure 1.12 level QAM can be seen.

- Quadrature Amplitude Modulation (QAM)
  - Amplitude modulation on both Quadrature carriers
  - 2 n discrete levels, n = 2 same as QPSK
- Extensive use in digital microwave radio links

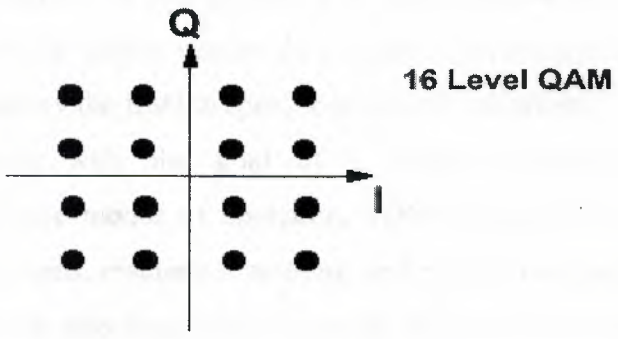


Figure 1.12. Quadrature Amplitude Modulation

1.5. Media Access

1.5.1. ALOHA

Transmit when desired

Positive ACK from receiver on independent link

Back off and retransmit if timeout

Slotted scheme reduces chance of collision

1.5.2. Carrier Sense/Multiple Access (CSMA)

Listen before transmit

Back off and retransmit if collision detected

1.5.3. Inhibit Sense/Multiple Access

Base station transmits busy tone

Transmit when not busy

Back off and retransmit if collision

1.5.4. Time Division Multiple Access

Multiple users share channel through time allocation scheme

Time Division Duplexing (TDD): DECT, PHP

Frequent Division Duplexing (FDD): GSM, IS-54, PACS

TDMA is an extension of AMPS. IS-136 systems are capable of operating with AMPS terminals, dual-mode terminals, and all-digital terminals. The network architecture is a more general version of the AMPS architecture. Corresponding to the AMPS network infrastructure of land stations and mobile telephone switching offices (base stations and switches), TDMA defines a BMI: "Base Station Mobile Switching Center, and Inter-working Function." Because IS-136 is confined to the air interface, it is appropriate to specify, in this general way, the functions performed in the network infrastructure. Each equipment vendor then makes its own decisions on how to allocate functions performed by the BMI to specific pieces of equipment.

In accordance with the goal of a personal communications system to accommodate multiple modes of operation, TDMA specifies three types of external network: public systems, residential systems, and private systems. Thus, a terminal can function as a cellular telephone with access to the base stations of cellular operating companies (public network). It can also be programmed to function as a cordless telephone operating with a specific residential base station (residential network), and as

a business phone operating with a specific wireless private branch exchange (private network).

To deliver mobile telephony, cryptographic authentication, and a wide range of service enhancements relative to AMPS, TDMA defines a large number of identification codes, including all of the AMPS identifiers. A major addition to the set of identification codes is the 64-bit A-key, assigned to each subscriber by her cellular operating company. This encryption key plays a critical role in promoting network security and communication privacy in a dual-mode TDMA system. Another identification code in TDMA is a 12-bit location area identifier, LOCAID. The system can divide its service area into clusters of cells, referred to as location areas. Each base station broadcasts its LOCAID. When a terminal that does not have a call in progress enters a new location area, it sends a registration message to the system. When a call arrives for the terminal, the system pages the terminal in the location area that received the most recent registration message.

The IMSI is a telephone number with up to 15 decimal digits that conforms to an international numbering plan (E.212) published by the International Telecommunication Union. The value of PV reflects the standards document (for example, IS-54 or IS-136) that governs the operation of a base station or terminal. The system operator code (SOC) transmitted by a base station identifies to terminals the company that operates the base station, while BSMC indicates the manufacturer of the base station. The digital verification color code (DVCC) plays the same role in digital traffic channels as the SAT transmitted in analog traffic channels.

Table 1.2 Comparisons of Cellular Systems

	GSM	IS-54	DECT
Bit Rate	270.8 kbps	48.6 kbps	1.152 Mbps
Bandwidth (Carrier Spacing)	200 KHz	30 KHz	1.728 MHz
Time Slot Duration	0,577 ms	6.7 ms	0.417 ms
Upstream slots per frame	8/16	3/6	12
Speech Coding	13 kbps RPE-LTP	7.95 kbps VSELP	32 kbps ADPCM
FDD or TDD	FDD	FDD	TDD
% Payload in Time Slot	73%	80%	67%
Modulation	GMSK	$\pi/4$ DQPSK	GMSK
Coding	Coded/Convol Coded+CRC Uncoded	Coded/Convol Coded+CRC Uncoded	CRC Only
Adaptive Equalizer	Mandatory	Mandatory	None

TDMA Advantages/Disadvantages:

In Table 1.2 the comparisons between GSM, IS-54 and DECT can be seen.

- Advantages
  - Sharing among N users
  - Variable bit rate by ganging slots
  - Less stringent power control due to reduced interuser interference—dedicated frequencies and slots
  - Mobile assisted/controlled handoff enable by available measurement slots
- Disadvantages
  - Pulsating power envelope interference with devices like hearing aids have been detected
  - Complexity inherent in slot/frequency allocation
  - High data rates imply need for equalization

1.5.5. Code Division Multiple Access

- A strategy for multiple users per channel based on orthogonal spreading codes
  - Multiple communicators simultaneously transmitting using direct sequence techniques, yet not conflicting with each other.
  - Pilot tone on BS to mobile unit forward channel used to time synchronize and equalize the channel (coherent detection).
  - Reverse channel is contention based, dynamically power controlled to eliminate the near-far problem.
- Developed by Qualcomm as IS-95
  - Special soft handoff capability
  - "Narrowband CDMA": 1.228 MHz chipping rate, 1.25 MHz spread bandwidth
  - Contrast with Broadband CDMA proposal: 10 MHz spread bandwidth
    - Multipath: Can leverage frequency diversity better
    - Interference tolerance: Can overlay existing analog user

Like a TDMA, IS-95 prescribes dual-mode operation. However, the two systems differ substantially in their relationship to the analog AMPS systems in which they operate. Recall that an A TDMA signal occupies exactly the same bandwidth as an analog AMPS signal. As a consequence system operators can replace individual AMPS channel units in analog base stations with TDMA radios that carry three full-rate physical charnels. By contrast, IS-95 prescribes spread spectrum signals with a bare-

width of 1.23 MHz in each direction. This is approximately 10 percent 0f a company's total spectrum allocation. As a consequence, a cellular operating company that adopts CDMA has to convert frequency bands of at least this size, corresponding to 41 contiguous AMPS channels, from analog to digital operation.

1S-95 contains many innovations relative to earlier cellular systems. One of them is a soft hand off mechanism, in which a terminal establishes contact with a new base station before giving up its radio link to the original base station. When a call is in a soft handoff condition, the terminal transmits coded speech signals to two base stations simultaneously. Both base stations send their demodulated signals to the switch, which estimates the quality of the two signals and sends one of them to a speech decoder. A complementary process takes place in the forward direction. The switch sends coded

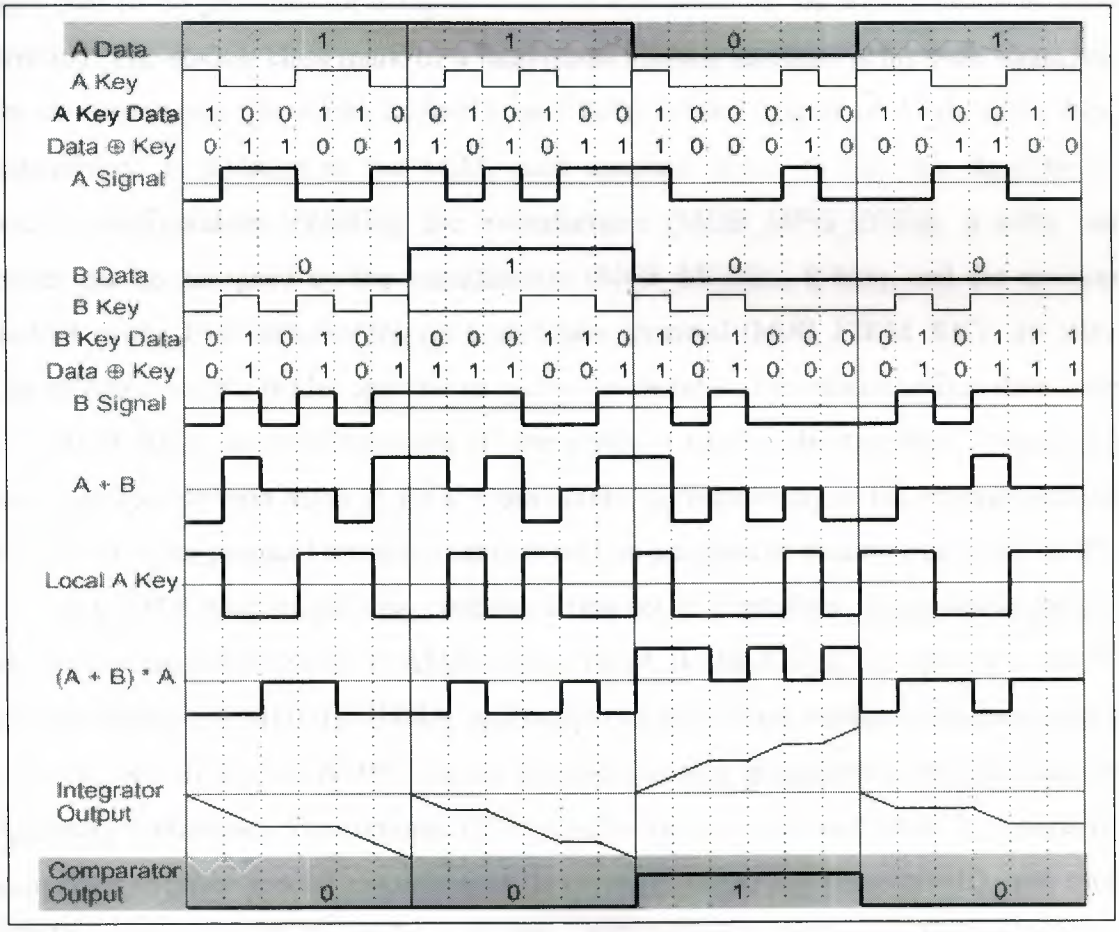


Figure 1.12. Recovery of a channel

speech signals to both base stations, which transmit them simultaneously to the terminal. The terminal combines the signals received from the two base stations and demodulates the result. Thus we have the network architecture illustrated in Figure 1.12,

which shows a vocoder in the switch rather than in base stations, their location in many TDMA implementations.

CDMA soft handoff requires base stations to operate in synchronism with one another. In order to achieve the necessary synchronization, all base stations contain global positioning system (GPS) receivers. A network of GPS satellites transmits signals that enable each GPS receiver to calculate its location in coordinates of latitude, longitude, and elevation. The satellite signals also include precise time information, accurate to within one microsecond, relative to universal coordinated time, an international standard.

In common with AMPS and TDMA, CDMA terminals and base stations employ an extensive set of identification codes that help control various network operations. Note that IS-95 provides for a highly detailed indication of the configuration of each terminal. The station class mark of a dual-mode CDMA terminal is an 8-bit identifier. The corresponding identifiers in AMPS and TDMA have lengths of 4 bits and 5 bits, respectively. In addition to the SCM, each terminal stores 40 bits that describe its precise configuration including the manufacturer (MOB_MFG_CODE, 8 bits), the model number assigned by the manufacturer (M0B_M0DEL, 8 bits), and the revision number of the firmware running on a particular terminal (M0B_FIRM_REV, 16 bits). The revision number is also specific to each manufacturer. The other configuration code is M0B_P_REV, an 8-bit indicator of the protocol run by the terminal. Initially all terminals operate with M0B_P_REV = 00000001, corresponding to the original version of 15-95. Higher protocol revision numbers will be assigned to future versions of 15-95.

A CDMA base station also contains a rich set of identifiers. Augmenting the 15-bit system identifier (SID) in AMPS and TDMA, CDMA systems specify a 16-bit network identifier (NID). In CDMA, a network is a set of base stations contained within a system. Recall that an AMPS system corresponds to a geographical area defined by regulatory authorities. By contrast, CDMA networks can be established by operating companies to meet special requirements. Each base station has its own NID, and each CDMA terminal can be programmed with a SID/NID pair indicating the system and network associated with the terminal's home subscription. Each base station has its own PN_0FFSET. This is a time delay applied to forward direction transmissions that enables the terminals in a cell to decode the desired signal and reject signals from other base stations. The 4-bit BASE_CLASS identifier anticipates terminals that will have

access to a variety of wireless services. In the initial issue of IS-95, the only assigned BASE_CLASS is 0000, corresponding to public macro cellular systems. Future class numbers could be assigned to other public networks or to various types of private networks such as wireless business systems (PBX) and residential cordless telephones.

The CDMA system anticipates a variety of mobility management schemes including location-area registration, as in TDMA and GSM; timer-based registration; and distance based registration. To facilitate location-area registration, IS-95 defines a 12-bit REG_ZONE identifier to be assigned to each base station. REG_ZONE plays the same role as the location area identifier, LOCAID, in TDMA. The identifiers, BASE_LAT (22 bits) and BASE_LONG (23 bits), specify the exact geographic location of the base station, in latitude-longitude coordinates. Terminals can use this information to perform distance-based registration.

1.6. Wireless LANs

Wireless LAN technology is becoming increasingly popular for a wide variety of applications. After evaluating the technology, most users are convinced of its reliability, satisfied with its performance and are ready to use it for large-scale and complex wireless networks. Originally designed for indoor office applications, today's Wireless LANs can be used for both indoor peer-to-peer networks as well as for outdoor point-to-point and point-to-multipoint remote bridging applications. Wireless LANs can be designed to be modular and very flexible. They can also be optimized for different environments. For example, point-to-point outdoor links are less susceptible to interference and can have higher performance if designers increase the "dwell time" and disable the "collision avoidance" and "fragmentation" mechanisms described later in this section.

1.6.1. Topology

Wired LAN Topology: Traditional LANs (Local Area Networks) link PCs and other computers to one another and to file servers, printers and other network equipment using cables or optic fibers as the transmission medium (Figure 1.13).

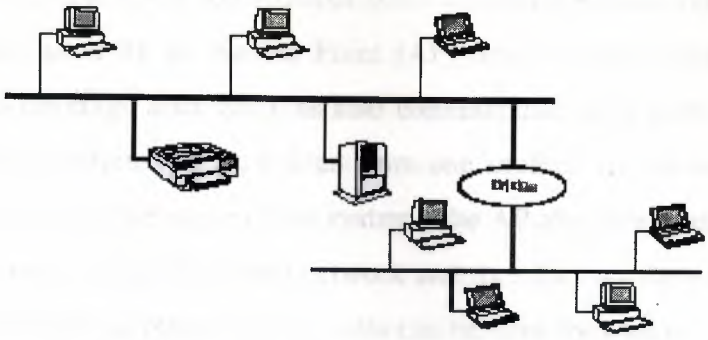


Figure 1.13: Wired LAN Topology

Wireless LAN Topology: Wireless LANs allow workstations to communicate and to access the network using radio propagation as the transmission medium. The wireless LAN can be connected to an existing wired LAN as an extension, or can form the basis of a new network. While adaptable to both indoor and outdoor environments, wireless LANs are especially suited to indoor locations such as office buildings, manufacturing floors, hospitals and universities. The basic building block of the wireless LAN is the Cell. This is the area in which the wireless communication takes place. The coverage area of a cell depends on the strength of the propagated radio signal and the type and construction of walls, partitions and other physical characteristics of the indoor environment. PC-based workstations, notebook and pen-based computers can move freely connected in the cell (Figure 1.13)

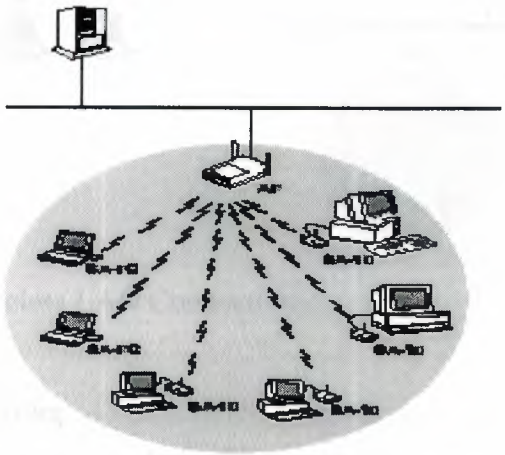


Figure 1.14: The Basic Wireless LAN Cell

Each Wireless LAN cell requires some communications and traffic management. This is coordinated by an Access Point (AP) that communicates with each wireless station in its coverage area. Stations also communicate with each other via the AP, so communicating stations can be hidden from one another. In this way, the AP functions as a relay, extending the range of the system. The AP also functions as a bridge between the wireless stations and the wired network and the other wireless cells. Connecting the AP to the backbone or other wireless cells can be done by wire or by a separate wireless link, using wireless bridges. The range of the system can be extended by cascading several wireless links, one after the other (Figure 1.14).

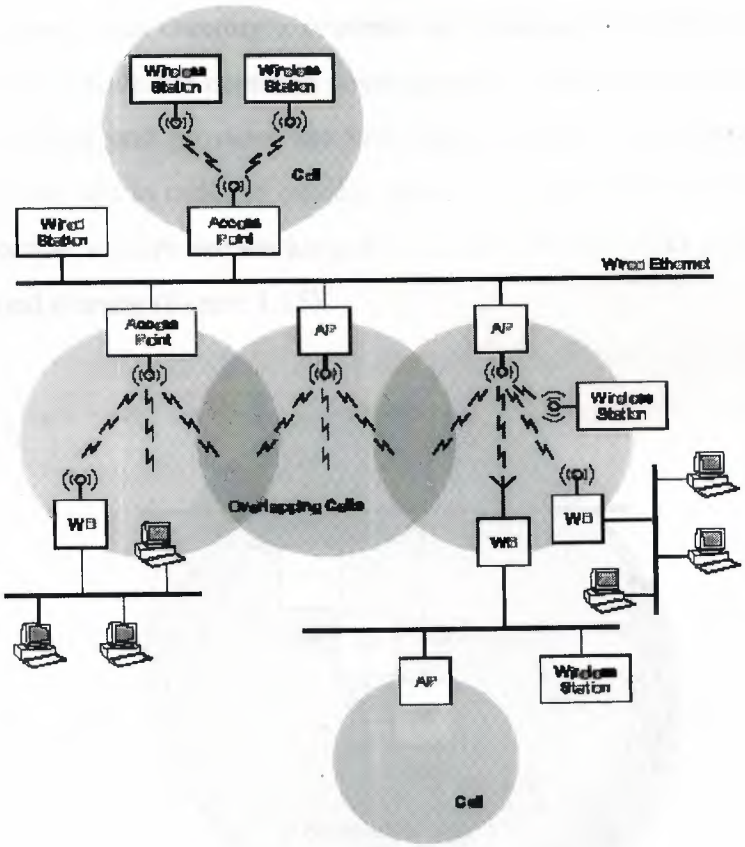


Figure 1.14. Wireless LAN Connectivity

1.6.2. Roaming

When any area in the building is within reception range of more than one Access Point, the cells' coverage is said to overlap. Each wireless station automatically establishes the best possible connection with one of the Access Points. Overlapping coverage areas are an important attribute of the wireless LAN setup, because it enables

seamless roaming between overlapping cells. Roaming allows mobile users with portable stations to move freely between overlapping cells, constantly maintaining their network connection. Roaming is seamless; a work session can be maintained while moving from one cell to another. Multiple access points can provide wireless coverage for an entire building or campus. When the coverage area of two or more APs overlap, the stations in the overlapping area can establish the best possible connection with one of the APs, continuously searching for the best AP. In order to minimize packet loss during switchover, the "old" and "new" APs communicate to coordinate the process. Load Balancing Congested areas with many users and heavy traffic load per unit may require a multi-cell structure. In a multi-cell structure, several co-located APs "illuminate" the same area creating a common coverage area that increases aggregate throughput. Stations inside the common coverage area automatically associate with the AP that is less loaded and provides the best signal quality. The stations are equally divided between the APs in order to equally share the load between all APs. Efficiency is maximized because all APs are working at the same low-level load. Load balancing is also known as load sharing (Figure 1.15).

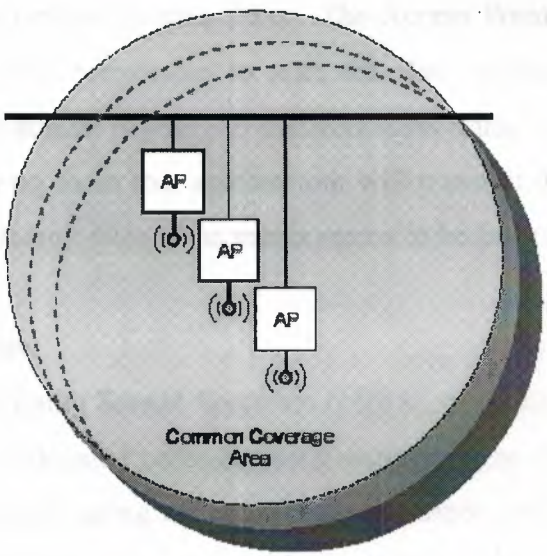


Figure 1.15. The Common Coverage Area of a Multi-cell Structure

1.6.3. Dynamic Rate Switching

The data rate of each station is automatically adjusted according to the received signal quality. Performance (throughput) is maximized by increasing the data rate and

decreasing re-transmissions. This is very important for mobile applications where the signal quality fluctuates rapidly, but less important for fixed outdoor installations where signal quality is stable.

1.6.4. Media Access

When many users are located in the same area, performance becomes an issue. To address this issue, Wireless LANs use the Carrier Sense Multiple Access (CSMA) algorithm with a Collision Avoidance (CA) mechanism in which each unit senses the media before it starts to transmit. If the media is free for several microseconds, the unit can transmit for a limited time. If the media is busy, the unit will back off for a random time before it senses again. Since transmitting units compete for air time, the protocol should ensure equal fairness between the stations. Fragmentation of packets into shorter fragments add protocol overhead and reduce protocol efficiency when no errors are expected, but reduce the time spent on re-transmissions if errors arelikely to occur. No fragmentation or longer fragment length add overhead and reduce efficiency in case of errors and re-transmissions (multi-path).

1.6.5. Collision Avoidance

To avoid collisions with other incoming calls, each station transmits a short RTS (Request To Send) frame before the data frame. The Access Point sends back a CTS (Clear To Send) frame with permission to start the data transmission. This frame includes the time that this station is going to transmit. This frame is received by all the stations in the cell, notifying them that another unit will transmit during the following Xmsec, so they can not transmit even if the media seems to be free (the transmitting unit is out of range).

1.6.6. Channelization

Using Frequency Hopping Spread Spectrum (FHSS), different hopping sequences are assigned to different co-located cells. Hopping sequences are designed so different cells can work simultaneously using different channels. Since hopping sequences and hopping timing of different cells cannot be synchronized (according to FCC regulations), different cells might try to use the same channel occasionally. Then, one cell uses the channel while the other cell backs off and waits for the next hop. In the case of a very noisy environment (multiples and interference), the system must hop quickly. If the link is quiet and clean, it is better to hop slowly, reducing overhead and increasing efficiency.

1.7. Future of Mobile Wireless Communications

3rd Generation Wireless, or 3G, is the generic term used for the next generation of mobile communications systems. 3G systems aim to provide enhanced voice, text and data services to user. The main benefit of the 3G technologies will be substantially enhanced capacity, quality and data rates than are currently available. This will enable the provision of advanced services transparently to the end user (irrespective of the underlying network and technology, by means of seamless roaming between different networks) and will bridge the gap between the wireless world and the computing/Internet world, making inter-operation apparently seamless. The third generation networks should be in a position to support real-time video, high-speed multimedia and mobile Internet access. All this should be possible by means of highly evolved air interfaces, packet core networks, and increased availability of spectrum. Although ability to provide high-speed data is one of the key features of third generation networks, the real strength of these networks will be providing enhanced capacity for high quality voice services. The need for landline quality voice capacity is increasing more rapidly than the current 2nd generation networks will be able to support. High data capacities will open new revenue sources for the operators and bring the Internet more closer to the mobile customer. The use of all-ATM or all-IP based communications between the network elements will also bring down the operational costs of handling both voice and data, in addition to adding flexibility.

On The Way To 3G:

As reflected in the introduction above, the drive for 3G is the need for higher capacities and higher data rates. Whereas higher capacities can basically be obtained by having a greater chunk of spectrum or by using new evolved air interfaces, the data requirements can be served to a certain extent by overlaying 2.5G technologies on the existing networks. In many cases it is possible to provide higher speed packet data by adding few network elements and a software upgrade.

A Look At GPRS, HCSD, and EDGE:

Technologies like GPRS (General Packet Radio Service), High Speed Circuit Switched Data (HSCSD) and EDGE fulfill the requirements for packet data service and increased data rates in the existing GSM/TDMA networks. I'll talk about EDGE separately under the section "Migration To 3G". GPRS is actually an overlay over the existing GSM network, providing packet data sevices using the same air interface by the

addition of two new network elements, the SGSN and GGSN, and a software upgrade. Although GPRS was basically designed for GSM networks, the IS-136 Time Division Multiple Access (TDMA) standard, popular in North and South America, will also support GPRS. This follows an agreement to follow the same evolution path towards third generation mobile phone networks concluded in early 1999 by the industry associations that support these two network types.

The General Packet Radio Service (GPRS):

The General Packet Radio Service (GPRS) is a wireless service that is designed to provide a foundation for a number of data services based on packet transmission. Customers will only be charged for the communication resources they use. The operator's most valuable resource, the radio spectrum, can be leveraged over multiple users simultaneously because it can support many more data users. Additionally more than one time slots can be used by a user to get higher data rates. GPRS introduces two new major network nodes in the GSM PLMN:

Serving GPRS Support Node (SGSN) - The SGSN is the same hierarchical level as an MSC. The SGSN tracks packet capable mobile locations, performs security functions and access control. The SGSN is connected to the BSS via Frame Relay.

Gateway GPRS Support Node (GGSN) - The GGSN interfaces with external packet data networks (PDNs) to provide the routing destination for data to be delivered to the MS and to send mobile originated data to its intended destination. The GGSN is designed to provide inter-working with external packet switched networks, and is connected with SGSNs via an IP based GPRS backbone network.

A packet control unit is also required which may be placed at the BTS or at the BSC. A number of new interfaces have been defined between the existing network elements and the new elements and between the new network elements. Theoretical maximum speeds of up to 171.2 kilobits per second (kbps) are achievable with GPRS using all eight timeslots at the same time. This is about three times as fast as the data transmission speeds possible over today's fixed telecommunications networks and ten times as fast as current Circuit Switched Data services on GSM networks. Actually we may not see speeds greater than 64 kbps however it would be much higher than the

speeds possible in any 2G network. Also, another advantage is the fact that the user is always connected and is charged only for the amount of data transferred and not for the time he is connected to the network. Packet switching means that GPRS radio resources are used only when users are actually sending or receiving data. Rather than dedicating a radio channel to a mobile data user for a fixed period of time, the available radio resource can be concurrently shared between several users. This efficient use of scarce radio resources means that large numbers of GPRS users can potentially share the same bandwidth and be served from a single cell. The actual number of users supported depends on the application being used and how much data is being transferred. Because of the spectrum efficiency of GPRS, there is less need to build in idle capacity that is only used in peak hours.

Already many field trials and also some commercial GPRS implementations have taken place. GPRS is the evolution step that almost all GSM operators are considering. Also, coupled with other technologies like WAP, GPRS can act as a stepping stone towards convergence of cellular service providers and the internet service providers. HSCSD (High speed circuit switched data) is the evolution of circuit switched data within the GSM environment. HSCSD will enable the transmission of data over a GSM link at speeds of up to 57.6kbit/s. This is achieved by concatenating, i.e. adding together, consecutive GSM timeslots, each of which is capable of supporting 14.4kbit/s. Up to four GSM timeslots are needed for the transmission of HSCSD. This allows theoretical speeds of up to 57.6 kbps. This is broadly equivalent to providing the same transmission rate as that available over one ISDN B-Channel. HSCSD is part of the planned evolution of the GSM specification and is included in the GSM Phase 2 development. In using HSCSD a permanent connection is established between the called and calling parties for the exchange of data. As it is circuit switched, HSCSD is more suited to applications such as video conferencing and multimedia than 'busty' type applications such as email, which is more suited to packet switched data. In networks where High Speed Circuit Switched Data (HSCSD) is deployed, GPRS may only be assigned third priority, after voice as number one priority and HSCSD as number two. In theory, HSCSD can be preempted by voice calls- such that HSCSD calls can be reduced to one channel if voice calls are seeking to occupy these channels. HSCSD does not disrupt voice service availability, but it does affect GPRS. Even given preemption, it is difficult to see how HSCSD can be deployed in busy networks and still confer an

agreeable user experience, i.e. continuously high data rate. HSCSD is therefore more likely to be deployed in start up networks or those with plenty of spare capacity since it is relatively inexpensive to deploy and can turn some spare channels into revenue streams.

An advantage for HSCSD could be the fact that while GPRS is complementary for communicating with other packet-based networks such as the Internet, HSCSD could be the best way of communicating with other circuit switched communications media such as the PSTN and ISDN. But one potential technical difficulty with High Speed Circuit Switched Data (HSCSD) arises because in a multi-timeslot environment, dynamic call transfer between different cells on a mobile network (called "handover") is complicated unless the same slots are available end-to-end throughout the duration of the Circuit Switched Data call. Because of the way these technologies are evolving, the market need for high-speed circuit switched data and the market response to GPRS, the mobile infrastructure vendors are not as committed to High Speed Circuit Switched Data (HSCSD) as they are to General Packet Radio Service (GPRS). So, we may only see HSCSD in isolated networks around the world. HSCSD may be used by operators with enough capacity to offer it at lower prices, such as Orange. [1] Believes that every

Table 1.3. The Specifications of the Technologies

GENERATION	2G Technology	2G+ Technology	2.5G Technology	3G Technology
BENEFITS	Capacity, Battery life	Capacity, Cost, Data	Higher speed data	Multimedia
TECHNOLOGIES	GSM	▪ HSCSD ▪ SMS data	▪ GPRS packet radio ▪ EDGE	▪ W-CDMA (part of UMTS)
	cdmaOne	▪ IS95B	▪ 1XRTT ▪ HDR ▪ 1X Plus	▪ 3XRTT ▪ W-CDMA? (Japan, Korea)
	TDMA	▪ IS136+	▪ GPRS ▪ EDGE	▪ UWC 136
	PDC (Japan)	▪ Imode	(skip to 3G)	▪ W-CDMA

GSM operator in Europe will deploy GPRS, and by 2005 GPRS users will almost match the number of voice only users. Right now there are 300 million wireless phones in the world. By 2005 we expect one billion. Before I proceed, a quick look at the table below

would help you appreciate and understand clearly the technology characterizations as 2nd generation, 2.5 generation and 3G. We have looked into 2G and some 2.5G technologies so far.

Destination: Third Generation:

Standardization of 3G mobile systems is based on ITU (International Telecom Union) recommendations for IMT 2000. IMT 2000 specifies a set of requirements that must be achieved 100% for a network to be called 3G. By providing multimedia capacities and higher data rates, these systems will enhance the range and quality of services provided by 2G systems. The main contenders for 3G systems are wideband CDMA (W-CDMA) and cdma2000. The ETSI/ GSM players including infrastructure vendors such as Nokia and Ericsson backed W-CDMA. Cdma2000 was backed by the North American CDMA community, led by the CDMA Development Group (CDG) including infrastructure vendors such as Qualcomm and Lucent Technologies. Universal Mobile Telephone System (UMTS) is the widely used European name for 3G. The proposed IMT-2000 standard for third generation mobile networks globally is a CDMA-based standard that encompasses THREE OPTIONAL modes of operation, each of which should be able to work over both GSM MAP and IS-41 network architectures.

Mode Title Origin Supporters 1 Direct Sequence FDD (Frequency Division Duplex) based on the first operational mode of ETSI's UTRA (UMTS Terrestrial Radio Access) RTT proposal. Japan's ARIB and GSM network operators and vendors. 2 Multi-Carrier FDD (Frequency Division Duplex) Based on the cdma2000 RTT proposal from the US Telecommunications Industry Association (TIA). Cdma One operators and members of the CDMA Development Group (CDG). 3 Time Division Duplex (TDD) The second operational mode of ETSI's UTRA (UMTS Terrestrial Radio Access) RTT proposal. An unpaired band solution to better facilitate indoor cordless communications. Harmonized with China's TD-SCDMA RTT proposal.

UMTS is the European designation for 3G systems. The UMTS frequency bands selected by the ITU are 1,885 MHz - 2,025 MHz (Tx) and 2,110 MHz - 2,2,20 MHz (Rx). Higher frequency bands could be added in future if need be, for stationary data. There is still some confusion about all the frequency options, as FCC has not given clear indications so far. The following table should briefly give an idea about the 3G system specifications.

3rd Generation Initiatives:

3GPP (Third Generation Partnership Project) and 3GPP2 are the two alliances working towards the specification for the 3G systems. 3GPP partners are ETSI, TTC, ARIB, TTA, T1 and the 3GPP2 includes TIA, TTC, ARIB, and TTA. Although both have chosen CDMA as the technology behind the 3G systems, the systems advocated by these two groups are different. The 3GPP organizational partners have agreed to co-operate for the production of Technical Specifications for a 3rd Generation Mobile System based on the evolved GSM core networks and the radio access technologies that the Organizational Partners support (i.e. UTRA both FDD and TDD modes). 3GPP2 provides global specifications for ANSI/TIA/EIA-41 network evolution to 3G and global specifications for the RTTs (Radio transmission technologies) supported by ANSI/TIA/EIA-41. Yet another group, the Operators Harmonization Group, is dedicated to achieving the maximum possible level of commonality of technologies to maximize interworking of different versions. It was as a result of pushing by OHG that led to ITU's mixed solution to 3G air interfaces with ANSI-41 and GSM MAP networking.

3G Timeframes:

The actual deployment of 3G will not be a homogeneous occurrence. Japan will lead with the service in early 2001, followed by Western Europe in mid to late 2003. U.S. is expected to wait for some time at 2.5G and 2.75G before going in to true 3G. As I have mentioned earlier, with TDMA based networks like GSM and IS-136, increased capacity will be the initial driving factors. Therefore these networks will take a comparatively longer time to true 3G.

Evolving Today's Networks Towards 3G:

The 3rd Generation Mobile System will most likely grow out of the convergence of enhanced 2nd generation mobile systems with greater data transfer speed and capacity and 1st generation satellite mobile systems. Evolution to the current generation mobile networks to 3G doesn't necessarily mean seamless upgradation to the existing infrastructure to the 3G. Evolution should also be seen in context of coexistence of the 2G and 3G networks for some time, with users able to roam across the new and the old networks, able to access 3G services wherever 3G coverage is available. As mentioned before, a 3G network can have one of the 3 optional air interfaces supporting one of the two GSM MAP and IS-41 network architectures. This results in a range of choices for

the existing networks to evolve/migrate towards 3G. Possible convergence of TDMA and GSM networks with EDGE adds another variable to the overall migration paths. Another variable that adds complexity to this already complex list of options is the time frames involved. By the time some of the 2.5 or 2.75G technologies go to field, we may see the emergence of 3G technologies also. So, a lot of thought regarding the costs involved, and/or the viability of 2.5G technologies like EDGE could be questioned. The same is true about the time frames of the so called "4G".

Before I talk about evolution/migration paths of all the existing 2G mobile wireless technologies, let me briefly discuss the 3G-network architecture and other technology factors involved in the migration to 3G.

3G Architecture:

The 3G networks will have a layered architecture, which will enable the efficient delivery of voice and data services. A layered network architecture, coupled with standardized open interfaces, will make it possible for the network operators to introduce and roll out new services quickly. These networks will have a connectivity layer at the bottom providing support for high quality voice and data delivery. Using IP or ATM or a combination of both, this layer will handle all data and voice info. The layer consists of the core network equipment like routers, ATM switches and transmission equipment. Other equipment provides support for the core bit stream of voice or data, providing QOS etc. Note that in 3G networks, voice and data will not be treated separately which could lead to a reduction in operational costs of handling data separately from voice. The application layer on top will provide open application service interfaces enabling flexible service creation. This user application layer will contain services for which the end user will be willing to pay. These services will include eCommerce, GPS and other differentiating services. In between the application layer and the connectivity layer, will run the control layer with MSC servers, support servers, HLR etc. These servers are needed to provide any service to a subscriber.

Migration Strategies:

The migration to 3G is not just based on evolving core networks and the radio interface to IMT 2000 compliant systems. Migration towards 3G would also be based on the following steps/technologies:

Network upgrades in the form of EDGE, GPRS, HSCSD, CDPD, IS-136+ and HDR. Evolution to 2.5G basically will provide support for high-speed packet data. Though

these technologies are extensions to 2G rather than precursors to 3G these will have a major impact either by proving (or not) demand for specific services. Service trials to test infrastructure, handsets and applications etc

EDGE! Will TDMA and GSM ever meet:

EDGE is a new time division multiplexing based radio access technology that gives GSM and TDMA an evolutionary path towards 3G in 400, 800, 900, 1800 and 1900 MHz bands. It was proposed to ETSI in 1997 as an evolution to GSM. Although EDGE reuses GSM carrier bandwidth and time slot structures, it is not restricted to use in GSM cellular systems only. In fact, it can provide a generic air interface for higher data rates. It provides an evolutionary path to 3G. Some call it 2.5G. It can be introduced smoothly into the existing systems without altering the cell planning. But as with GPRS, EDGE doesn't provide any additional voice capacity. The initial EDGE standard promised mobile data rates of 384 kbps. It allows data transmission speeds of 384 kbps to be achieved when all eight timeslots are used. In fact, EDGE was formerly called GSM384. This means a maximum bit rate of 48 kbps per timeslot. Even higher speeds may be available in good radio conditions. Actual rates will be lower with rates falling as one goes away from the cell site. EDGE can also provide an evolutionary migration path from GPRS to UMTS by implementing now, the changes in modulation that will be necessary for implementing UMTS later. Both High Speed Circuit Switched Data (HSCSD) and GPRS are based on something called Gaussian minimum-shift keying (GMSK) which only yields a moderate increase in data bit rates per time slot. EDGE, on the other hand, is based on a new modulation scheme that allows a much higher bit rate across the air interface. This modulation technique is called eight-phase-shift keying (8 PSK). It automatically adapts to radio circumstances and thereby offers its highest rates in good propagation conditions close to the site of base stations. This shift in modulation from GMSK to 8 PSK is the central change with EDGE that prepares the GSM world (and TDMA in general) for UMTS.

Only one EDGE transceiver unit will need to be added to each cell. With most vendors, it is envisioned that software upgrades to the BSCs and Base Stations can be carried out remotely. The new EDGE-capable transceiver can also handle standard GSM traffic and will automatically switch to EDGE mode when needed. EDGE capable terminals will also be needed - existing GSM terminals do not support the new modulation techniques and will need to be upgraded to use EDGE network functionality.

EDGE is currently being developed in two modes: compact and classic. Compact employs a new 200 kHz control channel structure. Synchronized base stations are used to maintain a minimum spectrum deployment of 1 MHz in a 1/3-frequency reuse pattern. EDGE Classic on the other hand employs the traditional GSM 200 kHz control structure with a 4/12 frequency reuse pattern on the first frequency.

How Can GSM and TDMA Converge With EDGE:

While developing the 3G wireless technology for TDMA, the Universal Wireless Communication Consortium (UWCC) proposed the 136 High-Speed (136 H-S) radio interface as a means of satisfying requirements for IMT-200 radio transmission technology (RTT). After evaluating various proposals, UWCC adopted EDGE (Actually EGPS, EDGE+GPRS) as the outdoor component of 136HS to provide 384 kbps data services. Since GSM networks can also have an evolutionary path via EDGE, this presents an interesting opportunity where the air interfaces of TDMA and GSM can converge and then evolve together. EDGE is being developed concurrently in ETSI and UWCC. The phase one of EDGE emphasizes enhanced circuit-switched data (ECSD) and enhanced GPRS (EGPS).

The TDMA terminals that support 30 kHz circuit switched services scan for a 30 kHz control channel (DCCH) according to TIA/EIA 136 procedures. If an acceptable 200 kHz EGPRS carrier exists, a pointer to this system will be available on the DCCH. On finding this, the terminal will leave the 30KHz system and start scanning of the 200 kHz systems. When it finds it, it starts behaving as if it is a GSM/GPRS terminal. To answer a circuit switched page, the mobile suspends packet data traffic and starts looking for a 30 kHz control channel. Mobile terminals that only support 200 kHz carriers immediately start looking for 200 kHz packet data system.

Will this happen? While EDGE provides a common air interface for TDMA and GSM to converge, there is one possible problem. GSM operators may decide to skip EDGE altogether in their migration path to 3G. By the time EDGE will be commercially available for GSM systems, 3G will already be in sight with W-CDMA and since W-CDMA will need an entirely new air interface, the additional investments in EDGE, only to be replaced by another system seems a bit unjustified. EDGE has lost favor in Europe with some wireless operators and vendors that are not convinced it will actually be adopted in force once carriers move to GPRS. As described above, the belief is that wireless service providers may be more inclined to move straight to WCDMA

from GPRS. On the other hand, some North American operators have taken the position that they may not need to upgrade to WCDMA after EDGE because it doesn't offer increased speeds in the mobile environment (the ITU/UMTS definition of G3G is 384 Kbps mobile, 2 Mbps low mobility/fixed wireless). This is an especially strong point when one considers that the market demand for high-speed wireless data has yet to be fully proven. The convergence of TDMA and GSM can't be ruled out also. Particularly in the US, operators may have more interest in moving on to EDGE to get compatibility with the TDMA networks. According to a study [1], EDGE should be available in the North American markets by 2002.

Individual Technology Evolution Paths:

A variety of technologies/standards exist and therefore, so do the number of paths that can be taken. The table below briefly summarizes these standards (Table 1.4).

Table 1.4. Cellular Standards

Standard Name	Other Names (Aliases)	Upgrade Path for . . .	Expected Availability
Code Division Multiple Access (CDMA)	IS-95, IS-95A, cdmaOne	N/A	Current
Global System Mobile (GSM)	N/A	N/A	Current
1XRTT	G3G-MC-CDMA-1X, also called 2.5G step for CDMA	CDMA	End of 2001
General Packet Radio System (GPRS)	Also called 2.5G for GSM	GSM and potentially TDMA	End of 2000
Enhanced Data for GSM Efficiency (EDGE)	Also called 2.5G for GSM and TDMA	GSM or TDMA	End of 2001
Wideband CDMA (WCDMA)	WCDMA, FDD Mode 1 (Direct Sequence), G3G-DS-CDMA	GSM or TDMA, and in rare cases CDMA	2002 Europe, later for North America depending upon spectrum availability
cdma2000	3XRTT, FDD Mode 2 (Multicarrier), G3G-MC-CDMA-3X	CDMA	2003
High Data Rate (HDR)	HDR	Not a true 3G upgrade; a network extension using a CDMA base system	End of 2001

GSM and TDMA To 3G:

GSM and TDMA systems have more or less the same set of options for migrating to 3G. The path to 3G is not as simple in case of GSM/TDMA as is in the case of CDMA. The main evolutionary standards are GPRS, EDGE and, finally, W-CDMA. Vendors are positioning each of these standards as a step to the next, but operators are

not so sure. For an operator moving from GSM to GPRS to EDGE and then to W-CDMA, he'll have to make investments 3 times which won't be pleasing to any operator. As [1] suggests, at this time, there seem to be four basic options that GSM and TDMA operators are considering:

Install GPRS, then move straight to WCDMA;

Install EDGE, then move straight to WCDMA;

Install GPRS, then move to EDGE, then to WCDMA; or

Install EDGE, skip move to WCDMA, and wait for the next generation (4G) (see Figure 1.16)

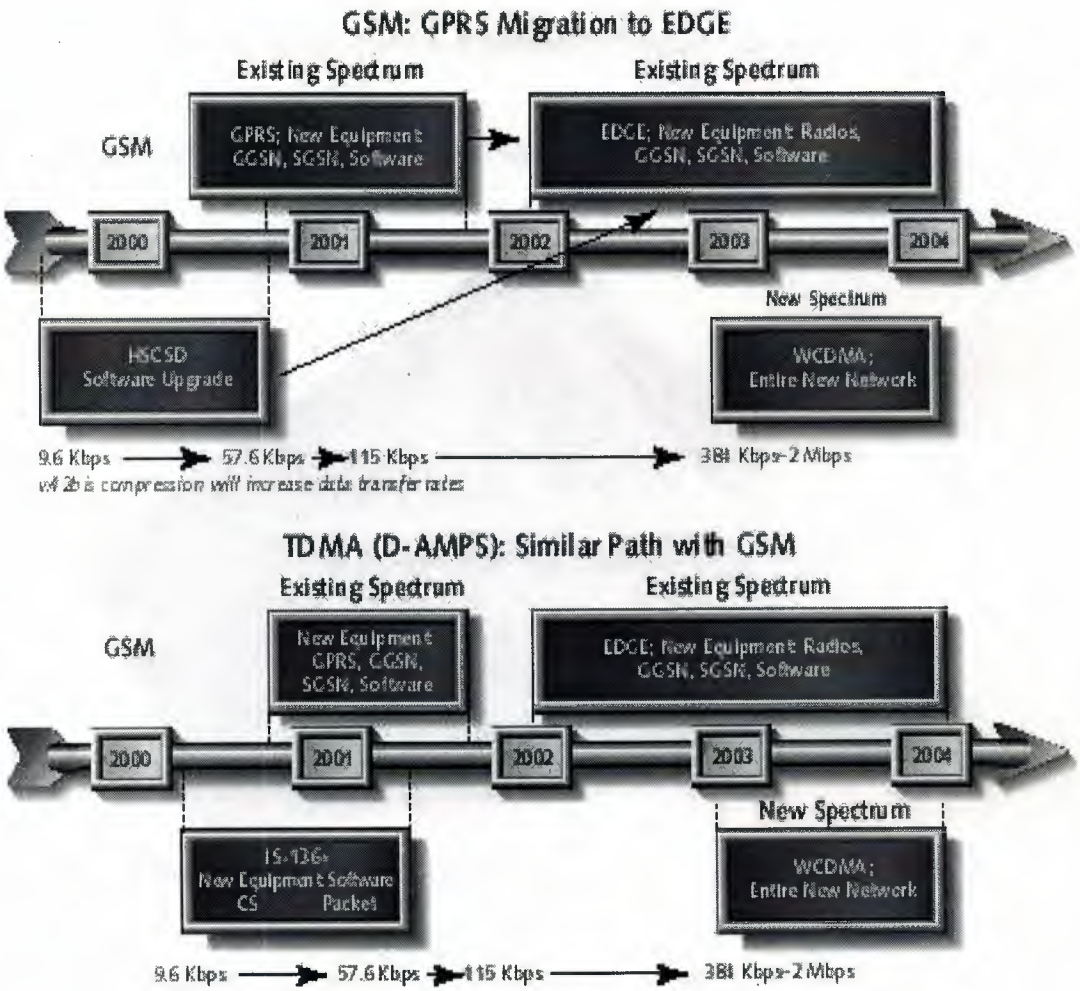


Figure 1.16. Technology path for the GSM operators

CDMA To 3G

While GSM and TDMA operators have multiple choices ahead for progressing to the next-generation networks, CDMA operators have a single path that truly builds upon itself. Currently all North American CDMA networks are based on IS-95 (cdmaOne), which can be setup to provide data rates upto 14.4 kbps. The next step is to have a software upgrade from IS-95A to IS-95B, which provides additional voice efficiencies giving additional capacity, and allows for up to 84-Kbps packet data. (We might not see 84kbps but instead 64kbps, initially.) While this migration does not need any additional hardware but as brought out by [1] most operators may decide not to move to IS-95B because of two reasons.

1. IS-95A in itself is relatively new and carriers have just launched their IS-95A data services.
2. By the time IS-95B becomes available, 1XRTT will be ready.

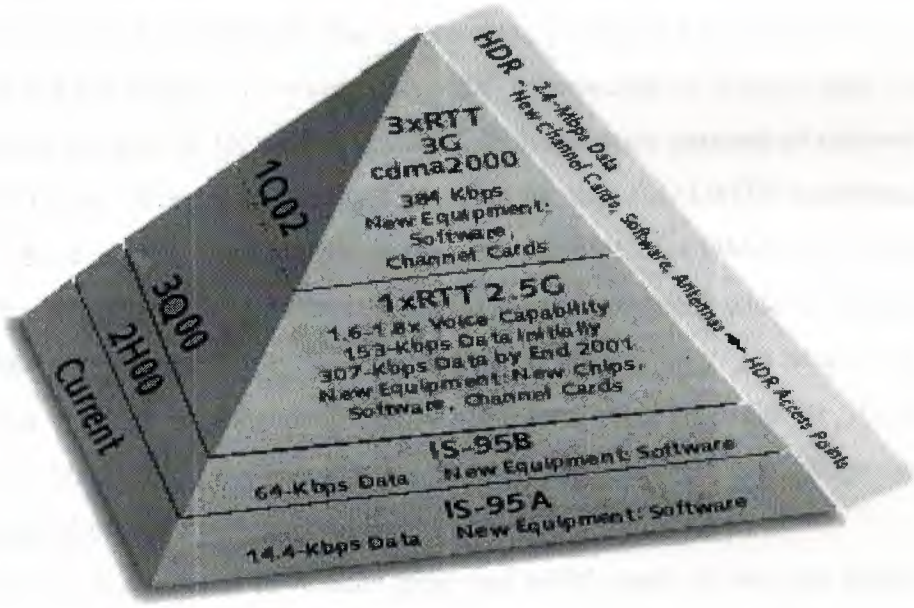


Figure 1.17. Options for the GSM operators

What Are The Costs?

In the shorter term, TDMA and GSM have a much more cost-effective upgrade option by means of moving to GPRS to be in a position to provide data services. As mentioned earlier, an upgrade to GPRS doesn't require substantial investments and existing GSM/TDMA service providers can upgrade to GPRS at around 28% cost of their initial 2G investments. The IS-95 upgrade path to 1xRTT is comparatively costly

at around 40% investments on the existing 2G networks. It should also be noted that IS-95A in itself has also not been in existence for long. However, in the final run to truly 3G networks, GSM/TDMA operators may have to incur much higher investments as shown in the figure below. The cost equations for TDMA or GSM may vary depending on the exact path taken (EDGE or no EDGE or only EDGE). CDMA has the unique advantage of having the same air interface in 2G as in 3G (same underlying technology).

Therefore, it is very probable that most CDMA carriers in North America will move straight to 1XRTT. 1XRTT is the first step in moving to the full ITU/UMTS-defined 3G standard. It has many features that make it completely different from IS-95B. It will provide more than double the data speeds available from IS-95B (153 Kbps vs. 64 Kbps); but, more importantly, in the spectrum-constrained market of North America, 1X will almost double the voice capacity. Additionally, the software and chip boards necessary for 1X are also an essential step to continue the upgrade to 3XRTT, which is also called G3G-MC-3X, but is also more popularly known by the trade name of cdma2000 (307 Kbps). However, cdma2000 is expected to provide only moderate voice capacity gains over 1X, and as such, 1X is the primary concern of carriers for the immediate future. Besides 1X and 3X paths to the ITU/UMTS-sanctioned G3G standards, there is also the Qualcomm-defined offshoot of CDMA--High Data Rate (HDR). This standard, which is proprietary to Qualcomm, sets aside a standard 1.25-MHz CDMA carrier specifically for data, and offers rates of up to 2.4 Mbps in a mobile environment. Though this standard achieves the data rates required for 3G, it is not considered a 3G standard because it is a data-only standard and has not been opened up for the approval of any standards bodies.

Several new standards have been proposed which don't fit into this classification of 2, 2.5 or 3G. These standards either provide only data services and/or provide much higher data rates than those specified by 3G systems. Examples are 1Xplus and 1XTREME. Since they use a single CDMA carrier they may be called 2.5G but then they provide much higher data rates than 3G. According to Motorola, 1XTREME will not require additional antennas as HDR will, and it will also keep data on the same spectrum as the voice services, meaning carriers won't have to devote any spectrum specifically to data services. 1XTREME is proposed to deliver the same voice capacity increases as standard 1X, and provide data rates approaching 1.4 Mbps. The second

iteration, expected to be in trials by the first quarter of 2001, is expected to deliver data rates as high as 5.2 Mbps. Motorola expects 1XTREME to be market-ready in the same time frame as HDR: by the end of 2001 to the first half of 2002.

Another interesting thing is that these so called 4G technologies may start appearing almost at the same time when 3G comes. It is not very clear as to how these developments will influence an already very complex set of equations.

Concluding Remarks:

Mobile communications are really poised to see major improvements in terms of capabilities of mobile networks. The next generation of wireless services, besides improving the overall capacity, will create new demand and usage patterns, which will in turn, drive the development and continuous evolution of services and infrastructure. While development of 3G networks will continue and pick up pace in the near future, the 2nd generation networks will keep evolving in terms of continuous enhancements and towards convergence of existing 2G standards. The initial 3G solutions should coexist with the 2G networks while slowly evolving to all 3G networks. While 3G in its true sense should have transparent roaming across all networks through out the world, given the penetration and the investments in the 2nd generation, true roaming (consistent service availability, across networks, independent of networks) looks to be to a very distant proposition!

CHAPTER 2. INTERFACES

2.1. Universal Serial Bus (USB)

USB, or Universal Serial Bus is a connectivity specification developed by computer and telecommunication industry members for attaching peripherals to computers. USB is designed to free all the troubles when installing external peripherals. It eliminates the hassle to open computer case for installing cards needed for certain devices. It is designed to meet Microsoft Plug and Play (PnP) specification, meaning users can install, and hot-swap devices without long installation procedures and reboots. Furthermore, it allows 127 devices to run at the same time on the bus. USB bus provides two types of data transfer speed 1.5Mbps and 12Mbps and it can provide a maximum of 500mA of current to devices attached on the bus. All these features will only need one interrupt to operate on a computer equipped with USB ports. Universal means all peripherals share the same connector. Serial simply defines devices can daisy chain together. We will now look at the different parts of USB.

The goal of the Universal Serial Bus (USB) design is to sweep the plethora of I/O ports on the PC into one serial channel. The bus runs at a base data rate of 12 Mbps but offers a 1.5-Mbps option that helps keep down the cost of low-performance devices, such as keyboards. USB's physical configuration is a tiered star. The PC acts as the host and root hub to which the user can attach devices or additional hubs. These additional hubs can, in turn, connect to a combination of devices and another hub layer. The bus supports as many as five hub layers and 127 devices on one host. The bus also can provide 5V power to attached devices. The control of system setup, device initialization, and data flow all reside with the host system. Upon system power-up, the host performs enumeration on each USB device. The host queries the device for a description, assigns the device a unique address for subsequent transactions, and sets the device's operating configuration. The host also identifies and loads hardware-specific drivers for the USB device into the operating system during device enumeration. Every millisecond (the USB frame time), the USB host initiates a series of data transactions. The USB offers two types of data transactions for high-volume data. Isochronous transactions offer guaranteed bandwidth, with the host system allocating 1 to 1023 bytes/frame for the transaction. Bulk transactions have no bandwidth guarantees. The host allocates to bulk data as much bandwidth as is left over after all other transactions are accounted for.

Bulk transactions have guaranteed delivery with resend ability, ensuring that all data arrives eventually.

All USB data transactions are host-initiated. Based on application program requests for data transfers, the host signals the USB device to start the data flow, regardless of the data's source or destination. USB devices cannot initiate a transfer; they can only respond to the host's command.

2.1.1. Inside USB

Essentially, there are three pieces to this USB technology -- host, hub and function. Host is actually the central point for all connections in the USB topology. It serves as the exchange point between each of the components of USB. The hardware implementation of host is called USB host controller, which is either integrated into the south bridge of the motherboards or included in USB add-on solutions. Hub allows multiple USB devices to share a single output to the USB host controller. Hubs on the back of computers are called root hubs. External USB hubs are available for users to connect more USB devices to the computer. Function is actually the USB device. Each USB device provides a function. Compound device provides multiple functions on the USB bus.

2.1.2. How fast is USB?

USB 1.1 provides high-speed and low-speed mode. In high speed mode, the host allows USB device to communicate and transfer data at 12Mbps; in low speed mode,

Table 2.1. Bandwidth of the other interfaces

Technology	Theoretical Maximum Throughput
Apple Desktop Bus (ADB)	10 kbps
Serial Port	230 kbps
Geoport Port	2 Mbps
USB at low speed	**1.5 Mbps**
USB at high speed	**12 Mbps**
SCSI	1 - 40 MBps
FireWire	400 Mbps
USB 2.0 at full speed	**360 - 480 Mbps**
Fast SCSI	8 - 80 MBps
Ultra SCSI-3	18 - 160 MBps

the host allows USB device to operate at 1.5Mbps. People tend to mix up between the capitalized "B" and small "b." Basically, small "b" stands for bits where capitalized "B" stands for bytes. 1 byte contains 8 bits. So, when you hear 1.5Mbps (Mega bits per second), you can determine the bytes per second by dividing 1.5Mbps by 8 to convert the unit to mega bytes per second. As you may have guessed, keyboards, mice and joysticks are among the low speed devices, using USB low-speed chips. Zip drives, scanners and printers are named as high-speed devices. USB host manages the bandwidth of each pipe used by different USB devices; 4 types of data transfer methods serve different kinds of USB devices. They are isochroous, interrupt, bulk, and control data transfers. As seen above, USB falls somewhere in the middle. So, do not expect USB to be a replacement of SCSI. USB will probably wipe out the ADB, serial, parallel ports in the next few years.

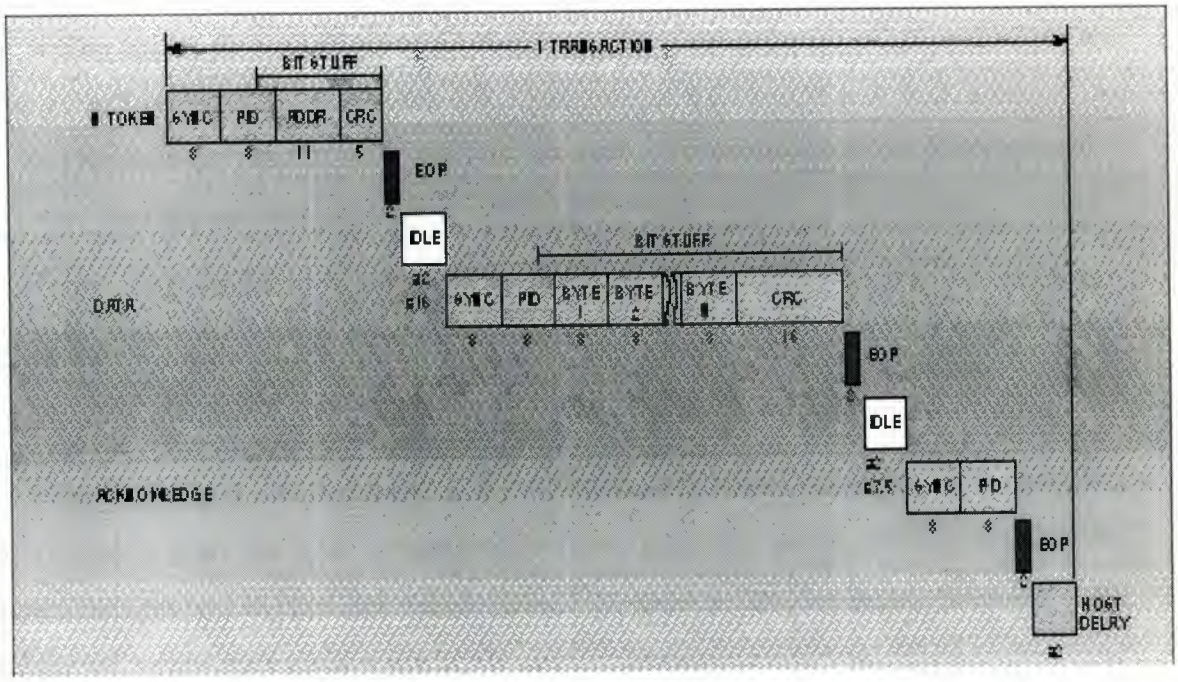


Figure 2.1. Scheme of USB Data Transmission

2.1.3. Architecture of USB

A USB bulk read-data transfer has three parts. First, the host requests data. Next, the device sends the data. Then, the host acknowledges the successful transfer. Delays are part of the protocol, but the host delay between transfers is unbounded (Figure 2.1)

2.2. IEEE 1394 (Firewire)

The emergence of digital video and multimedia applications has brought with it the need to move large amounts of data quickly between peripherals and PCs. And as audio/video products migrate to digital technology, both consumers and professionals alike stand to benefit from a simple high-speed connection that would make this transmission more efficient. Enter 1394: the digital cable. The IEEE 1394 serial bus is the industry-standard implementation of Apple Computer, Inc.'s FireWire digital I/O system. It is a versatile, high-speed, low-cost method of interconnecting a variety of personal computer peripherals and consumer electronics devices. Developed by the industry's leading technology companies, the specification was accepted as an industry standard by the IEEE Standards Board on December 12,1995.FireWire offers several advantages over other technologies. These benefits include:

- Guaranteed delivery of multiple data streams through isochronous data transport.
- The ability to connect up to 63 devices without the need for additional hardware, such as hubs
- Data transfer rates of up to 400 Mb/sec with 1.2 Gb / sec speeds in development.
- A flexible, six-wire cable.
- Complete plug-and-play operation, including the hot swapping of live devices.
- Acceptance by over 40 leading manufacturers in the computer and electronics consumer industries.

2.2.1. How does FireWire work?

Using special integrated circuits, FireWire multiplexes a variety of different types of digital signals such as compressed video, digitized audio, and device control commands on two twisted-pair conductors. The result is that FireWire's standard, 6-pin cables and connectors replace the myriad of I/O connectors currently found in consumer electronics equipment, PCs and peripherals. FireWire also employs isochronous data transfer to guarantee the delivery of multiple time-critical multimedia data streams. And, the protocol uses a "fairness" arbitration scheme to ensure that all nodes having information to send get a chance to use the bus.

2.2.2. ADVANTAGES OF IEEE-1394

Speed: up to 400 Mb/sec with 1.2 Gb/sec speeds in development.

Expandability Up to 63 devices supported.

Convenience Easy-to-use cable and connectors for plug-and-play and “hot swapping”.

Guaranteed data transfer Isochronous transport of multiple time-critical data streams. Low-cost flexible, six-pin cable for use in high-volume commercial markets.

2.2.3. Architecture

The 1394 standard defines two bus categories: backplane and cable. The backplane bus is designed to supplement parallel bus structures by providing an alternate serial communication path between devices plugged into the backplane. The cable bus is a "non-cyclic network with finite branches," consisting of bus bridges and nodes (cable devices). *Non-cyclic* means that you can't plug devices together so as to create loops. 16-bit addressing provide for up to 64K nodes in a system. Up to 16 cable hops are allowed between nodes, thus the term finite branches. A bus bridge serves to connect busses of similar or different types; a 1394-to-PCI interface within a PC constitutes a bus bridge, which ordinarily serves as the root device and provides bus master (controller) capability. A bus bridge also would be used to interconnect a 1394 cable and a 1394 backplane bus. Six-bit Node_IDs allow up to 63 nodes to be connected to a single bus bridge; 10 bit Bus_IDs accommodate up to 1,023 bridges in a system. This means, as an example, that the limit is 63 devices connected to a conventional 1394 adapter card in a PC. Each node usually has three connectors, although the standard provides for 1 to 27 connector per a device's physical layer or PHY. Up to 16 nodes can be daisy-chained through the connectors with standard cables up to 4.5 m in length for a total standard cable length of 72 m. (Using higher-quality "fatter" cables permits longer interconnections.) Additional devices can be connected in a leaf-node configuration, as shown in figure 1. All 1394 consumer electronic devices announced as of early 1997 have only a single connector; there are no currently are digital camcorders or VCRs that correspond to the devices with ID 3 or ID 5 shown in figure 1. Physical addresses are assigned on bridge power up (bus reset) and whenever a node is added or removed from the system, either by physical connection/disconnection or power up/down. No device ID switches are required and hot plugging of nodes is supported. Thus 1394 truly qualifies as a plug-and-play bus. The 1394 cable standard defines three signaling rates: 98.304, 196.608, and 393.216 Mbps (megabits per second; MBps in this thesis refers to megabytes per second.) These rates are rounded to 100, 200, and 400 Mbps, respectively, in this paper and are referred to in the 1394 standard as S100, S200 and S400. Consumer DV gear uses S100 speeds, but most 1394 PC adapter cards support the S200 rate. The slowest active node ordinarily governs the signaling rate for

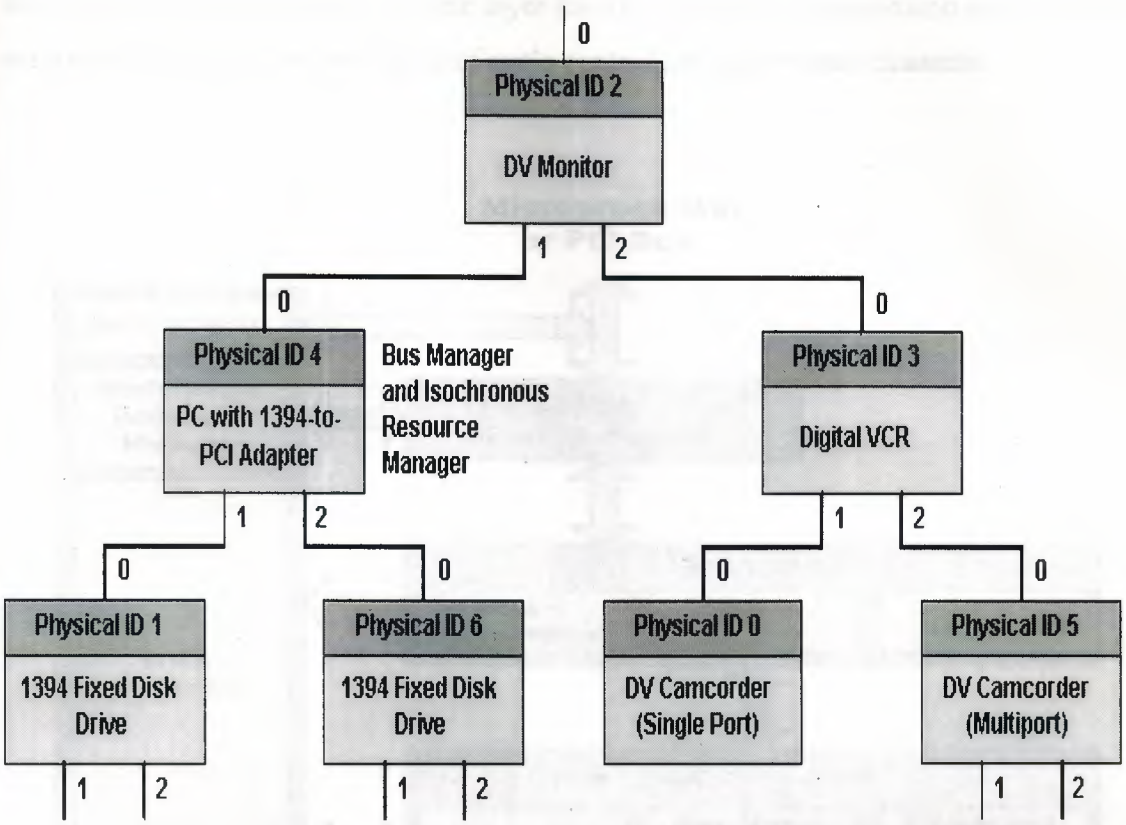


Figure 2.2. Topology of a typical PC-based 1394 bus system for DV applications.

the entire bus; however, if a bus master (controller) implements a Topology Map and a Speed Map for specific node pairs, the bus can support multiple signaling speeds between individual pairs. The 1394 Trade Association's 1394.1 working group presently are refining and clarifying the setup requirements for handling interconnected devices with multiple signaling speeds.

2.2.4. Physical, Link, and Transaction Layers

The three-stacked layers shown in figure 2 implement the 1394 protocol. The three layers perform the following functions:

The transaction layer implements the request-response protocol required to conform to the ISO/IEC 13213:1994 [ANSI/IEEE Std 1212, 1994 Edition] standard Control and Status Register (CSR) Architecture for Microcomputer Buses (read, write and lock). Conformance to ISO/IEC 13213:1994 minimizes the amount of circuitry required by 1394 ICs to interconnect with standard parallel buses. The link layer supplies an acknowledged datagram to the transaction layer. (A datagram is a one-way data transfer

with request confirmation.) The link layer handles all packet transmission and reception responsibilities, plus the provision of cycle control for isochronous channels.

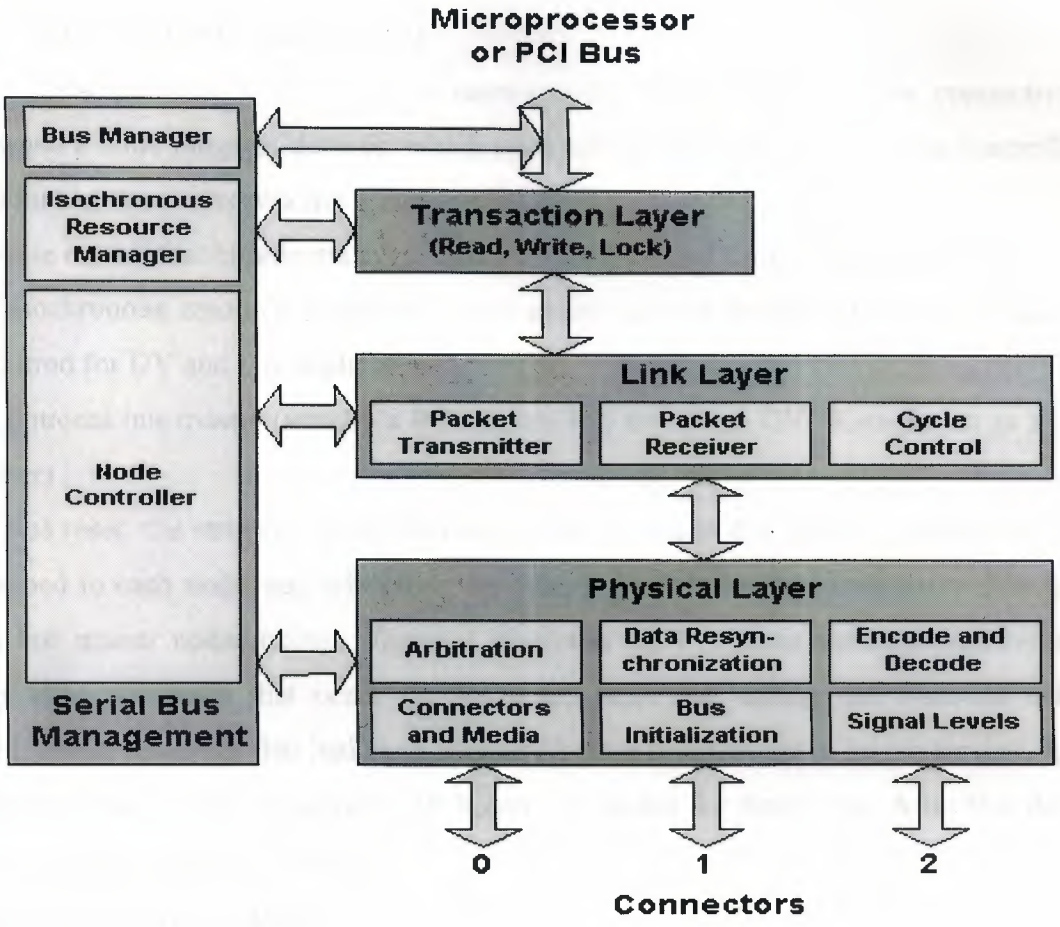


Figure 2.3. The 1394 Protocol Stack and Serial Bus Management Controller

The physical layer provides the initialization and arbitration services necessary to assure that only one node at a time is sending data and to translate the serial bus data stream and signal levels to those required by the link layer. Galvanic isolation may be implemented between the physical layer and the link layer using optical isolators; with isolation, the chip implementing the physical layer is powered by the bus conductors. Isolation should be provided where three-wire power cords are used to prevent ground loops through the green-wire ground; consumer devices, which use two-wire power cords or wall-wart power supplies, ordinarily don't require galvanic isolation.

The physical (PHY) layer is the bottleneck in 1394 systems. Historically, commercial PHY chips operated at half the potential data rate of link layer (LINK) chips (100 Mbps

vs. 200 Mbps, later 200 Mbps vs. 400 Mbps.) Texas Instruments announced in fall 1998 a set of 400-bps PHY chips that conform to the updated 1394a tentative specification and support the Open Host Controller Interface (OHCI) in conjunction with an OHCI-compliant link Chip.

2.2.5. 1394 Bus Management

1394 provides a flexible bus management system that provides connectivity between a wide range of devices, which need not include a PC or other bus controller. Bus management involves the following three services:

A cycle master that broadcasts cycle start packets (required for isochronous operation)

An isochronous resource manager, if any nodes support isochronous communication (required for DV and DA applications)

An optional bus master (usually a PC adapter, but an editing DVCR might act as a bus master)

On bus reset, the structure of the bus is determined, node IDs (physical addresses) are assigned to each node, and arbitration for cycle master, isochronous resource manager, and bus master nodes occurs. Figure 4 illustrates on a timeline the identification and arbitration processes that occur on bus reset. Note that during the 1-second delay isochronous resources that had been allocated before the reset are to be reallocated. Any resources that are not reclaimed will become available for future use. After that delay new resources may be allocated.

Isochroous Data Transport:

The isochronous data transport of the 1394 bus provides the guaranteed bandwidth and latency required for high-speed data transfer over multiple channels. The isochronous resource manager includes a BANDWIDTH_AVAILABLE register that specifies the remaining bandwidth available to all nodes with isochronous capability. On bus reset or when an isochronous node is added to the bus, the node requests a bandwidth allocation. As an example, a DV device would request approximately 30 Mbps of bandwidth, representing the 25+ Mbps DV data rate plus 3-4 Mbps for digital audio, time code, and packet overhead. Bandwidth is measured in bandwidth allocation units, 6,144 in a 125 ms cycle. (A unit is about 20 ns, the time required to send one data quadlet at 1,600 Mbps, called the S1600 data rate; the S1600 data rate is unlikely be supported in future implementations. A quadlet is a 32-bit word; all bus data is transmitted in quadlets.) 25 m s of the cycle is reserved for asynchronous traffic on the bus, so the default value of

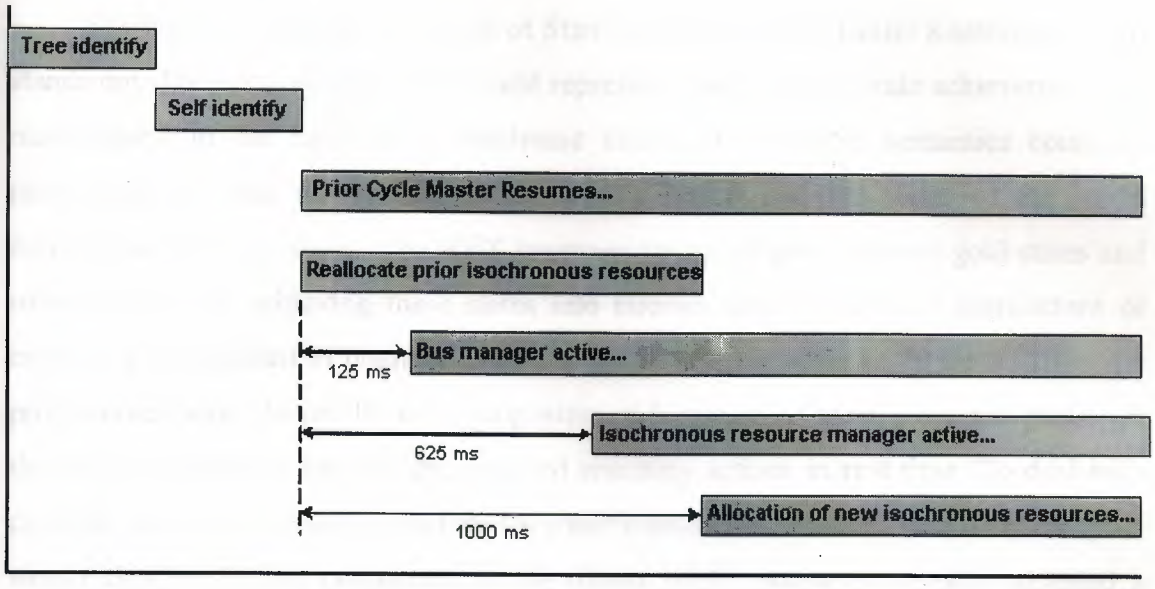


Figure 2.4. Bus reset timeline

the BANDWIDTH_AVAILABLE register on bus reset is 4915 units. In a 100-Mbps system, a DV device would request about 1,800 units; in a 200-Mbps system, about 900 units would be sufficient. If adequate bandwidth is not available, the requesting device is expected to repeat its request periodically.

The isochronous resource manager assigns a channel number (0 to 63) to nodes that request isochronous bandwidth based on values in the manager's CHANNELS_AVAILABLE register. The assigned channel number identifies all isochronous packets. When a node no longer requires isochronous resources, it is expected to release its bandwidth and channel number. As an example, the bus manager sends signals to cause a camcorder to commence talking on its channel and a record deck to commence listening on its channel for video data from the bus manager application. Device control is managed by asynchronous communication. Video acquisition for non-linear digital editing is simpler than the camcorder-DVCR example, because it requires only a single isochronous channel, plus an asynchronous path for device control. Timecode is built into the DV data, but asynchronous timecode transmission over the bus is useful when in camcorder or DVCR shuttle mode.

CHAPTER 3. INTELLIGENT ROBOTS

For the AI formalists, the work of Stan Rosenschein and Leslie Kaelbling (1986) stands out. They proved that if one could represent robot goals of state achievement and maintenance in the form of an electronic circuit, a consistent semantics could be maintained between the memory states of the circuit and the states of the world represented by these states. The REX language compiled propositional goal states and robot actions for achieving these states into circuits (actually c-based simulations of circuits) that executed in bounded time and usually on the order of 10 hertz. Thus, the programmer was allowed to use a propositional language to specify desired goals, yet the robot was able to execute the required resulting actions in real time. To deal with multiple goals that would contend for the robot's sensors or actuators (the REX compiler would flag conflicting commands to the robot), REX programs typically included a scheme to arbitrate among active circuits.

3.1. Animate Vision

What of the use of cameras for robots? Human and animal visions are the most powerful perception systems. However, we have seen how the processing of the low-level data alone, much less the addition of rapid control of a pan-tilt head, seemed to have little chance of fitting into the new paradigm. Fortunately, in the late 1980s, an analogous paradigm shift was taking place in the way researchers were approaching vision for agents. Again using ethology, several researchers began using animal visual behaviors as models for computational counterparts. For example, a frog primarily used its motion detection to catching flying food. Other animals keyed on specific aspects of the color spectrum for certain tasks. Indeed, psychophysical studies showed the human visual system to be, not surprisingly, even more adept. The human retina is arranged in such a manner as to have a higher concentration of receptors in the center and decreasing numbers radiating outward. Thus, humans don't process square arrays of data, for example, 512 x 512 x 8 bits, in a time step; rather, they use lower-resolution peripheral vision to watch for indications of motion or looming objects while they concentrate the higher-resolution center of the retina -the fovea- to reason about a specific object or part of an object in great detail. Humans don't take in everything at once in all its color and motion dimensions; instead, they concentrate on a narrow portion of their visual field.

Moreover, humans move this portion rapidly about the environment in patterns dictated by the task at hand and the last time step of visual information that was produced. For example, when looking at a picture of a group of people, if one is asked what the ages of the people are, one's eyes move in a pattern that concentrates on the faces of the people, with a few scans to determine the height of the people. To determine where a cup is in the picture, the eyes dart quickly about the picture for a table, then move to the objects on the table. Only when told that they must remember as many objects in the picture as possible will the eye-scanning machinery move in a pattern resembling the scan of a full image as found in the classic algorithms of computer vision (Ballard 1991).

Now computer vision paradigms were being recast into small, quick behaviors that not only dealt with a given field of view more efficiently but also where to next point the pan-tilt head of the camera. These well defined, compact routines, such as tracking a given color or attending to peripheral motion, were much like the behaviors being developed by the nouveau planning community and could now be incorporated as another part of the paradigm shift in programming robots.

3.2. A New Kind of Mapping

Just as AI had much to learn from the attempts to make robots intelligent, so did the robotics community stand to gain from the same endeavor? A good example was the use of maps for robot navigation. Early maps in the robotics community were geometric in nature, often as grids with each cell representing some amount of space in the real world. The grid had a single-coordinate system in which elements were represented. These maps became sophisticated at representing the spatial structure of the world (Moravec and Elfes 1985). It also was easy to do path planning and obstacle avoidance with geometric maps (for example, Lozano-Perez and Wesley [1979] and Brooks [1982]). However, geometric maps, as a part of the traditional world model of the robot, can require vast amounts of memory for large areas; in addition, the robot must know precisely where it is so that it can reason from the map or add to it. Just as with computer vision, trying to maintain accurate geometric maps was computationaly intensive and extremely difficult in real-world situations.

The solution, in keeping with the paradigm shift in vision, was the use of topological maps (Kuipers and Byun 1987; Brooks 1985). Patterned after how humans represent space, topological maps represent the world as a graph of places connected by

arcs, thus using no metric or geometric information, only the notions of proximity and order. With a topological map, the robot navigates locally from place to place, minimizing movement errors. Moreover, topological maps are clearly much more compact in their representation of space.

This notion was rapidly adopted by the AI and robotics communities. Kuipers and Byun (1991) continued their work on topological maps, producing a representation of space called the spatial semantic hierarchy. Another implementation of topological maps, by Kortenkamp and Weymouth (1994), used, both sonar and vision to determine places in a topological representation. The first part of this book introduces several other topological-based map representations and also some initial attempts at integrating topological and grid-based map representations.

3.3. Planning

Although the movement away from general representations was considered healthy, the resulting degree of specialization was viewed with some alarm. As Chuck Thorpe of Carnegie Mellon University once remarked about Brooks's Robots: "I wouldn't want one to be my chauffeur." In point of fact, many researchers exploring the new paradigm had no intention of throwing out the classic planning baby with the bath water. However, it was clear that planning in both its form and its function had to be rethought.

Two researchers involved in this rethinking by looking at the psychophysical aspects of human activity were Phil Agre and David Chapman (1987). Their research pointed to evidence that humans somehow put together plans for action based on the set of routine behaviors they can carry out. Moreover, logical decomposition planning is rarely invoked in the course of human affairs, and when it is, it serves primarily as a guide to the general direction in which one should head rather than a production of rigid sets of action.

During the late 1980s and early 1990s, several approaches along these lines were being pursued at once for intelligent robots. There were attempts to expand on the mobot approach (Maes 1990); others went further in the direction of enumerating all possible actions using planning prior to run time (for example, Kaelbling [1988] and Schoppers [1987]). Still others tried a combination of these approaches (for example, Bonasso [1991]).

One of the most important of these efforts was the work by Jim Firby (1989) on reactive action packages (RAPs). In his dissertation, Firby described a three-layered architecture with classic planning at the top, a reactive layer of behaviors at the bottom, and a middle layer with the goals of the resulting plan executed as dynamic sequences of these behaviors (that is, RAPs). When this framework was significantly expanded (Bonasso et al. 1995; Gat 1992), it became possible to program a large variety of robots -or any group of computer controlled machines for that matter- to carry out a variety of tasks over long duration in the vicinity of, and in concert with, human counterparts. Erann Gat's chapter in this book on the three layered approach explains why it has become a popular approach for the design and implementation of intelligent robots. We might call this approach P-SA; that is, the robot plans based on initial conditions and common knowledge (P) and then executes this plan using senseact (SA) behaviors, replanning only when the reactive behaviors run out of routine solutions. In this architecture, simple representations are tailored to specific tasks. Layered software allows behaviors such as obstacle avoidance to coordinate smoothly with behaviors such as path following. A new level of routines -cached plans- execute between the reactive behaviors and the central brain, and planning and other deliberate reasoning guide the procedures and behaviors in accomplishing the primary task and interacting with humans. In addition, there have been some remarkable advances in hardware. Plug-and-play subsystems that combine sensors and effectors are much more common.

Already the technical successes we have seen in the robot competitions over the years are finding their way into practical applications. Today unattended mobots fetch and carry in semi structured hospital environments. Vacuum mobots are used regularly in North American industry to clean large storage and staging spaces during off-work hours. Mobots are also used to semiautonomous explore uninhabitable venues such as Terran volcanoes and Martian Landscapes. The case studies in this book are ample proof that mobots have technically evolved to a point where, today, they are poised to help humankind in broad ranging tasks from mapping the ocean floor and long-term nursing home care to planetary colonization.

For mobots to move to the next level of competence necessary to complete such tasks, however, they need a broader base of technical support. It is our hope that, from this book, AI researchers will be inspired to expend additional effort in mobile robot research. One of the editors is fond of saying that "acting and sensing are still the

hardest parts." So naturally, new developments in robot perception and low-level control will always be necessary to advance the state of the art and meet the challenge of applications in difficult environments such as under water or outer space.

Mobots can benefit from all artificial intelligence disciplines, however, and, as we have previously explained, the robot architectures that support more traditional AI research are already in place. Mobots need to reason about their acts, both for feasibility and for rationality. Thus they can benefit from advances in planning and logical theories of sensing and acting. Most future robot tasks will involve working with humans. Consequently, spoken language generation and understanding must be developed for them to be effective team members. For missions of long duration -such as those involving deep sea or planetary exploration-mobots must adapt their behavior, and even their preferences over time. This requirement involves machine learning at all levels of competency.

The time has thus come, and the technology is here, for artificial intelligence and robotics to more closely join forces in improving the quality of life on earth and in establishing new civilizations in the cosmos.

3.4. Mapping And Navigation

While it is possible for a robot to be mobile and not do mapping and navigation, sophisticated tasks require that a mobile robot build maps and use them to move around. Levitt and Lawton (1990) posit three basic questions that de6ne mobile robot mapping and navigation:

- Where am I?
- How do I get to other places from here?
- Where are other places relative to me?

Each of the case studies in part one of this book address one or more of these questions. Each uses a different approach to representing and using spatial information. As such, they span the spectrum of options for mapping and navigation

On one side of the spectrum are purely metric maps. The robot's environment is defined by a single, global coordinate system in which all mapping and navigation takes place. Typically, the map is a grid with each cell of the grid representing some amount of space in the real world. These grids became quite sophisticated at representing the spatial structure of the world (see, for example, Moravec and Elfes [1985]). The case study of CARMEL Kortenkamp and his colleagues describes a mobile robot that uses a

grid-based approach to mapping and navigation. These approaches typically work in bounded environments, with little consistent structure and where the robot has opportunities to realign itself with the global coordinate system using external markers.

On the other side of the spectrum are qualitative maps, in which the robot's environment is represented as places and connections between places. Indeed, the idea of a map that contains no metric or geometric information, but only the notions of proximity and order, is enticing because such an approach eliminates the inevitable problems of dealing with movement uncertainty in mobile robots. Movement errors do not accumulate globally in qualitative maps as they do in maps with a global coordinate system since the robot only navigates locally, between places. Qualitative maps can also be more compact in their representation of space, in that they represent only interesting places and not the entire environment. Qualitative maps (also referred to as topological maps) have become increasingly popular in mobile robotics (see, for example, Brooks 1985; Kuipers and Byun 1991; and Kortenkamp and Weymouth 1994). The case studies by Nourbakhsh and by Koenig and Simmons describe the current state-of-the-art in qualitative mapping. These techniques work well in structured environments (i.e., office buildings) where there are distinctive places that are goals for the robot.

There have been efforts to combine metric and qualitative maps so that the strengths of both representations can be used (Asada et al. 1988; Kuipers and Levitt, 1988). The first case study in this part, by Thrun and his colleagues, gives an overview of both metric and topological mapping and describes their approach to integrating these two representations.

3.5 High Speed Obstacle Avoidance, Map Planning, Navigation Feedback

3.5.1 Obstacle Avoidance

Obstacle avoidance is performed a mobile robot called CARMEL (in this thesis it is taken as an example). A key to CARMEL's success was its ability to deal with sonar sensor noise. This robot used a novel algorithm called EERUF (error-eliminating ultrasonic firing) to allow for rapid firing and sampling of ultrasonic sonar sensors, which means faster obstacle avoidance. EERUF allows the robot to detect and reject ultrasonic noise, including crosstalk. The sources of ultrasonic noise may be classified as external sources, such as ultrasonic sensors used on another mobile robot operating in the same environment; or internal sources, such as stray echoes from other on-board

ultrasonic sensors. The latter phenomenon, known as crosstalk, is the reason for the slow firing rates in many conventional mobile robot applications: most mobile robots are designed to avoid crosstalk by waiting long enough between firing individual sensors to allow each echo to dissipate before the next sensor is fired. EERUF, on the other hand, is able to detect and reject about ninety-seven percent of all erroneous readings caused by external or internal noise.

In general, external noise is random and can be detected simply by comparison of consecutive readings. Crosstalk, however, is mostly a systematic error, which may cause similar (albeit erroneous) consecutive errors. EERUF overcomes this problem by firing each sensor after individual, alternating delays that disrupt the repetitiveness of crosstalk errors, rendering these errors detectable by the method of comparison of consecutive readings. Because EERUF practically eliminates the problem of crosstalk, it allows for very fast firing rates. See Boren"in and Koren (1992) for more details on EERUF.

3.5.2. Vector Field Histogram

To map the obstacles in the environment, it uses another innovative approach called a vector field histogram (VFH). The VFH method uses a two-dimensional Cartesian grid, called the histogram grid, to represent data from ultrasonic range sensors. Each cell in the histogram grid holds a certainty value that represents the confidence of the algorithm in the existence of an obstacle at that location. This representation was derived from the certainty grid concept originally presented in Moravec and Elfes (1985).

The central idea behind a certainty grid is to fuse sonar readings over time to eliminate errors in individual sonar readings. In its case, each cell of the grid array represents a ten-centimeter square, and the array covers the entire environment space. As the robot travels through the environment, its sonar sensors are continuously being fired and returning range readings to objects. Since the approximate location of the robot at any time is known through odometry and the direction of each sonar sensor is known, the location of objects in the grid array can be estimated. Each time an object is detected at a particular cell location, the value of the cell is increased and the values of all the cells between the robot and this cell are decremented (since they must be empty). The cells have minimum and maximum values, which have been chosen arbitrarily for

computational convenience. Greater detail of the histogram method can be found in Borenstein and Koren (1991).

To perform the actual obstacle avoidance using the histogram grid, it creates an intermediate data representation called the polar histogram. The purpose of the polar histogram is to reduce the amount of data that needs to be handled for real-time analysis while at the same time retaining the statistical information of the histogram grid, which compensates for the inaccuracies of the ultrasonic sensors. In this way, the VFH algorithm produces a sufficiently detailed spatial representation of the robot's environment for travel among densely cluttered obstacles, without compromising the system's real-time performance. The spatial representation in the polar histogram can be visualized as a mountainous panorama around the robot, where the height and size of the peaks represent the proximity of obstacles, and the valleys represent possible travel directions. The VFH algorithm steers the robot in the direction of the valleys, based on the direction of the target location. The details of its algorithm for creating and using the polar histogram is detailed in Borenstein and Koren (1991A). A high-level description of the algorithm follows:

1.Collect current sonar sensor readings.

2.Create the histogram grid. For each sonar reading do the following;

a) Given the direction of the sensor, the distance returned, and the robot's current location, determine the cell in the histogram grid in which the obstacle lies. If the sonar did not detect an obstacle, then determine the cell that lies at the maximum range of the sonar and skip the next step.

b) Increment the value in that cell by a fixed amount (we use three in our implementation) up to some maximum (we use fifteen).

c) For each cell lying on a straight line between the robot and the cell determined in (a) above, decrement its value by a fixed amount (we use one) down to a minimum of zero.

3) Calculate the polar histogram as follows:

a) Define an active window surrounding the center of the robot (we use a window of size 33 x 33).

b) For each direction around the robot, in five-degree increments, total the value in all cells in that direction within the active window.

c) Smooth the histogram by averaging neighboring directions.

4) Set a threshold such that a polar histogram value below that threshold means that direction is "free," while a polar histogram value above that threshold means that direction is "occupied." This threshold can vary greatly between robots and even between environments.

5) Search the polar histogram for a free direction of travel as follows:

a) If the direction to the target is free and enough neighboring histogram directions are free (to assure that the path is wide enough for the robot), then the target direction is the free direction.

b) Otherwise, begin checking to the left and right of the target direction (alterrnating) for a free segment of the histogram that is wide enough for the robot to pass. The last direction in such a free segment can be returned as the free direction.

c) If no free segment is found, the robot is trapped (i.e., surrounded on all sides by obstacles).

The combination of EERUF and VFH is uniquely suited to high-speed obstacle dance; it has been demonstrated to perform obstacle avoidance in the most diffiicult obstacle courses at speeds of up to 1.0 meter per second. All of its experimental runs were at speeds of 780 millimeters per second or less. Its maximum speed was 500 millimeters per second although it would run more slowly when in tight spaces or when approaching an object. A graceful motion around obstacles in open terrain distinguished it.

3.5.3. Global Path Planning

In addition to VFH, CARMEL uses a global path planner that searches the histogram grid created during obstacle avoidance and creates a list of via points (intermediary goal points) that represent the shortest path to the goal. The algorithm was developed at the Oak Ridge National Laboratory (Andersen et al., 1992) and reimplemented on CARMEL. The algorithm is simple and fast:

1. Initialize a planning grid of the same size and granularity as the histogram grid.
2. Threshold the histogram grid to determine occupied cells for planning purposes. Mark these cells as occupied in the planning grid.
3. Expand each occupied cell by the radius of the robot in the planning grid.
4. If the start or target locations fall inside an obstacle, move them to the nearest edge of the obstacle.
5. Create a linked list of structures of the type:

```
struct NODE {
int x,y;            /* coordinates of grid cell */
int distance;       /* distance from the target */
struct NODE *next;  /* forward link */
};
```

6. Initialize the head of the linked list to the coordinates of the target cell and initialize the distance to zero.

7. While there is something in the list and the start cell has not yet been processed:

a) Pop the head of the list.

b) Place the distance of that item into its corresponding cell (x, y) in the planning grid.

c) Put all neighboring cells that are still 0 and are not obstacles at the end of the linked list with a distance equal to one plus the distance of the popped Item.

8. Starting at the start cell, follow the minimum path (including diagonal elements) to the target cell. If all the neighbors of the start cell are zero then there is no path.

9. Select those cells where the slope of the path changes and store their coordinates as via points that the robot should try to follow.

If a path is not found, it is possible to iterate the path planner using higher and higher thresholds for occupied cells. This algorithm is very quick, taking less than one second to run on a grid size of 256 by 256 cells (each cell represents 20 centimeters in the environment). The speed of the algorithm lets it run on the fly to extract it from potential] trap situations. With an appropriate threshold, the algorithm rarely, tails to find a path- in cases where a path is not found. It is simply instructed to head in the direction of the goal. One characteristic of this algorithm is that the path is not always the shortest path; however, VFH will cut corners while following via points along the way to a goal, allowing a path to be straightened out.

Grid-based systems have many limitations; including the amount of memory they require and the need to always know very precisely the location of the robot. In the competition, a grid representation was sufficient, as the arena was of a fixed (and known) size and we had mechanisms for determining the location of the robot. Given a larger environment, or an environment of unknown extent, a different representation scheme might be better.

3.5.4. The Supervising Execution System

The design of our execution system was motivated by several goals. First, and most importantly, we wanted to keep the executive very simple. The hierarchical design eliminated many problems, such as contention for resources between submodules and lack of coordination between sub modules. All scheduling was done through the top-level planner. Second, we wanted the overall system to be extremely robust. Because the competition rules prohibited outside interference with the robots after they had started the task, the executive needed to handle even catastrophic errors during run-time. Third, we wanted to use existing research as sub modules. For example, the obstacle avoidance routines we used had been developed over many years as stand-alone processes. We wanted to integrate them within the framework of the task without substantial modifications. Finally, we want to develop and test each sub module independently of the others. This allowed several groups to work in parallel in developing the software system.

What emerged from these design goals was a strict hierarchy of modules with each sub module and information exchange between each sub module being completely controlled by a supervising executive. No sub module was allowed to run by itself and all sub modules were guaranteed to return, even if they returned an error condition.

3.5.5. Overview of a mobile robot system

The executive has three primary tasks: First, it has to decide what to do (i.e.. "Which sub module to call). Second, it has to decide when to do it; and finally, it has to decide where to do it. The basic strategy for performing the task previously described in the task description is for the executive to: (1) tell the obstacle avoidance sub module to move the robot to a particular location where objects might be seen. (2) Tell the object detection sub module to find all objects within a certain sweep area (e.g., the 180 degrees in front of the robot). (3) Tell the obstacle avoidance sub module to approach the nearest object or item and tell the object detection sub module to verify the location of the object. (4) Tell the obstacle avoidance sub module to approach nearer to the object if the robot is still too far away from it.

These steps are repeated until all of the objects have been visited. In some environments (such as the competition), steps (1) and (2) are repeated before the robot begins to visit objects. This allows many objects to be placed in the map before dead reckoning errors accumulate.

At any time during this cycle, the planner can decide that CARMEL has moved far enough that its dead reckoning is probably in error. The executive then interrupts the cycle and directs the obstacle avoidance sub module to move the robot to a location from which it can see at least three objects in the proper configuration (as was previously described in the section on localization). The executive then points the camera at each of the three objects and asks the object detection sub module to determine new headings to the three objects. These headings are passed to the localization sub module. That sub module returns the new position and orientation of CARMEL, which the executive uses to update its internal position and orientation.

3.5.6. Error Recovery

There are two major causes of errors that the executive has to deal with. First, there are errors in movement, whereby the obstacle avoidance sub module cannot move the robot to the position requested by the executive. The second major cause of errors is errors in object location, whereby the object detection sub module cannot verify that the robot has moved to an object. Recovery from these two types of errors is discussed in the following subsections.

3.5.7. Errors in Movement

The obstacle avoidance sub module can detect two types of movement errors: trap situations, such as U-shaped obstacles, and failures to attain the goal location. Traps are detected automatically by the obstacle avoidance algorithm. The executive's solution to this error is to call the path planning sub module to plot a set of via points that will lead the robot out of the trap and to the goal location. The obstacle avoidance sub module is then called with this list of via points instead of a single goal location. The obstacle avoidance sub module guides the robot through each via point. Failure to attain the goal location is not detected automatically by the obstacle avoidance algorithm. The algorithm was designed to attempt to reach its goal location, running forever if necessary. Obviously, this was not desirable, since the goal location could be located inside an obstacle or behind a long wall. The obstacle avoidance sub module guarantees termination by estimating how long it should take the robot to reach the goal location and stopping the robot and returning to the executive when this time limit is exceeded (or it has reached the goal location).

When the executive is informed of a failure to attain the goal location, it first determines if it is close enough to the goal location so that no further actionis necessary.

"Close enough" varies depending on the situation. For example it is just trying to take an image, a few meters away may be close enough. But when it is approaching within two robot diameters of an object, fractions of a meter are important. If the executive determines that it is not close enough, the executive asks the obstacle avoidance sub module to try again. After repeated failures, the executive chooses a new goal location that will also be suitable (such as the other side of the object) and starts the process all over.

3.5.8. Errors in Object Location

In the course of visiting an object, the planner will move it to a spot approximately three meters from the object. At this point, the executive will request another image to verify the location of the object. It is this new heading –and distance that is used to make the final approach to within two robot diameters of the object; using this new, robot-relative information reduces the reliance on dead reckoning. However, there may be situations when the object cannot be seen from the verification location. In this case, the executive rotates the camera left and right in an ever-widening arc, searching for the object. If this fails, the executive assumes that it is too close to the object to detect it (the vision has a minimum range of one meter) and so it backs the robot up. If this fails, the object is assumed to be occluded and the executive chooses a new goal location on the opposite side of the object and starts the verification process over again.

3.5.9. Analysis

An integrated system to perform a specific task has been described. In practice, this system performed above expectations. The decision to dedicate specific sub modules to different tasks paid off handsomely at execution time. Since sub modules are encapsulated and guaranteed to return, they can have the entire resources of the computer at their disposal, without needing to sacrifice CPU time for the top-level planner. For the obstacle avoidance sub module, this results in movement that is fast and continuous, displaying none of the move-stop-move behavior of many other robots at the competition. For the object detection Slit module, this results in analysis of images that takes only a few seconds.

It also became apparent during the course of implementing the system that the performance of each sub module affected the design of the supervising executive. In particular, the object detection sub module evolved over time and became much more

accurate and could extract objects at much greater ranges. This reduced the need for dealing with uncertainty and error.

The system was tested over many runs in three different environments, with the largest being 22 meters by 22 meters. Error recovery routines for movement errors were tested by forcing it into traps. Error recovery routines for object location errors were tested by manually moving objects after their positions had been recorded by the planner. In practice, the error recovery routines for visiting objects allowed us to be less concerned with the precise position of the robot, so that we didn't need to perform landmark triangulation as often. Since landmark triangulation is time consuming, requiring a movement and three images, the less often that it needs to be performed, the better.

It found and visited all ten objects in just, over nine minutes. This far out-paced the competition, none of whom could complete the task in the allotted twenty minutes. For the competition run it used two hard-coded view points from which it hoped to see all ten objects; it went immediately to these two points, storing objects that it found in its map. Should it not have found all ten objects at these first two view points, there were other view points to which it could have gone. Because it was able to find all ten objects at the first two viewpoints, its global object map was very accurate. It then proceeded to visit each object in turn, going to the nearest unvisited object first.

3.6. Integrating Real-Time AI Techniques in Intelligent Systems

There is a growing body of literature regarding the design of systems that apply artificial intelligence techniques in real-time task environments [Garvey and Lesser, 1994; Laffey, et al, 1988; Sttosnider and Paul, 1994]. The literature may be classified along two dimensions, the proposed technique and its level of application within a system. For example, proposed techniques include: approximate algorithms [Lesser, et ai, 1988], "anytime" algorithms [Dean and Boddy, 1988], and deliberative scheduling [Hayes-Roth, 1985; Boddy and Dean, 1993; Horvitz, et ai, 1989]. Techniques are applied at any of three different levels within a system: (a) the architecture underlying all of the system's operations; (b) reasoning methods the agent uses to solve particular problems; and (c) the control strategy by which the agent determines which of its possible computations to perform at each point in time. As discussed below, most approaches in the literature focus on points in this two-dimensional space. They apply

particular techniques at particular system levels to meet the real-time requirements of particular classes of applications. However, none of these approaches alone is sufficient to support the real-time requirements of intelligent systems that set and pursue multiple goals by perceiving, thinking, and acting in dynamic, complex, and uncertain environments.

Figure 3.1. presents an excerpt from a class hierarchy of "jobs" requiring intelligent systems. For example, an ICU (intensive care unit) monitoring agent must perceive a variety of patient variables, reason about the interpretation, diagnosis, prognosis, and treatment of the patient's condition, and act to guide the patient toward a satisfactory outcome.

At any point in time, systems may face opportunities to perform multiple instances of multiple activities. While it is engaged in particular instances of activities, the validity of its observations and inferences will decay and new information will become available

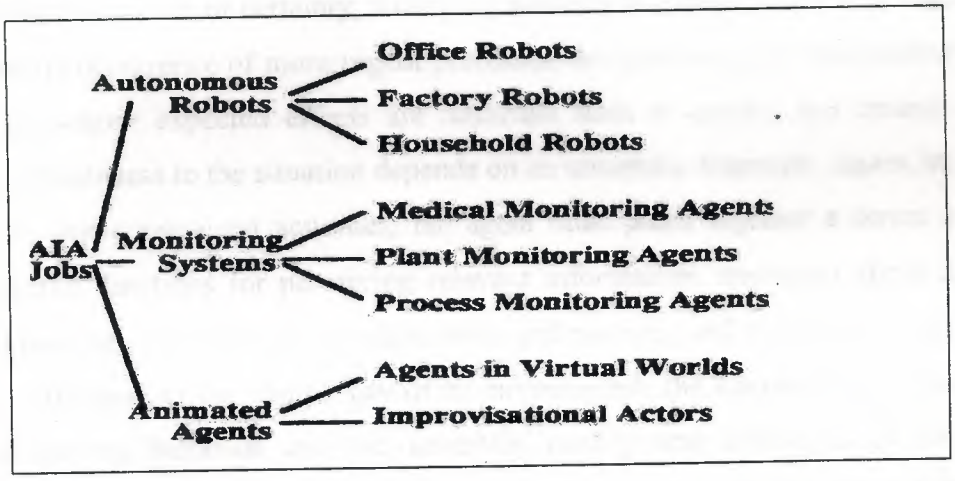


Figure 3.1. Excerpt from a class hierarchy jobs for intelligent systems

as exogenous phenomena evolve, smoothly or abruptly, in directions that are only partially predictable. Needless to say, there is no known algorithm for correctly monitoring ICU patients or performing any of the other jobs in Fig. 1. An agent must construct goal-seeking courses of behavior opportunistically out of component perception, cognition, and action behaviors that are triggered at run time and whose time requirements are variable and uncertain.

Given the dynamic, complex, and uncertain character of its environment, an intelligent system's job is pervaded by real time considerations. These include

occasional instances of "semi-hard" deadlines on the achievement of specific foreseeable goals. For example, an ICU monitoring agent should at least partially restore a patient's completely impeded oxygen flow within a few minutes. In addition to their intrinsic uncertainty, deadlines have context-specific uncertainty, since the agent's early actions and other exogenous factors may influence the environment (e.g., the patient's condition, or the behavior of a cooperating actor) and thereby increase or decrease the time available for action. The agent must remain sensitive to these changes even as it pieces together situation-triggered perception, cognition and action functions aimed at achieving its goal

In addition to goals with deadlines, an agent has more qualitative constraints on its behavior. For example, the time-constrained activities of an ICU monitoring agent include: tracking the time-varying course of a particular condition that may be measurable or inferable in different ways at different times for different costs and with different precision or certainty; balancing attention to current goals with vigilance to the possible occurrence of more urgent problems; or monitoring the execution of a plan of action whose expected effects are uncertain both in quality and timing and whose appropriateness to the situation depends on an uncertain diagnosis. Again, in performing these time-constrained activities, the agent must piece together a series of situation-triggered functions for perceiving relevant information, reasoning about the meaning and implications of perceived information and planning and executing effective actions.

Because of the complexity of its environment, the compositional character of its goal-seeking behavior and the uncertain quality and timeliness of its behavioral components, an agent's real-time effectiveness cannot be guaranteed by a single provably optimal technique within its architecture, reasoning methods, or control strategy. Instead, an agent's real-time effectiveness must emerge from effective interactions among multiple heuristic techniques at all three levels.

3.6.1. Real Time Techniques in The System Architecture

Figure 3.2. presents an overview of the proposed agent architecture. Basically it is blackboard architecture with a uniform mechanism for domain and control reasoning and controllable subordinate processes for perception and action. (See [Hayes-Roth. 1985; 1990; 1993] for additional discussion.) Real-time mechanisms are incorporated at several places in the architecture, as discussed below.

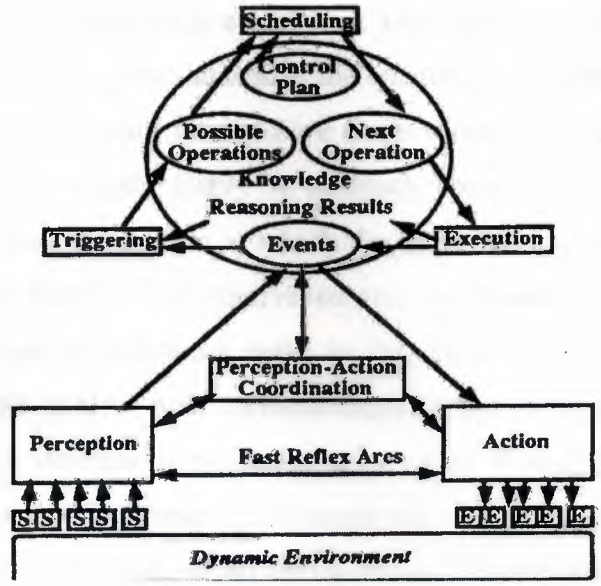


Figure 3.2. Overview of the system architecture

Loose coupling of Perception, Cognition, Action:

Most efforts to coordinate perception cognition and action couple them tightly, selecting and interleaving goal-directed sequences of perception, cognition, action functions [Simmons, et al. 1992; Georgeff and Lansky. 1987; Hendler and Agrawala. 1990]. This strategy works well for systems that work on one goal at a time can predict reliably when relevant perceptual information will be available and when external actions must be executed. However, this strategy is not appropriate for an agent that must: perceive its environment selectively, but continuously; reason about and act more or less promptly upon a selected subset of asynchronously detected situations; and coordinate different actions with different kinds of external events. To support these requirements, the proposed architecture provides dedicated computational resources for perception, cognition, and action, with asynchronous message passing among them.

Selective Perception:

In a complex task environment, resource-bounded agents cannot perceive all sensed data and cannot reason about all perceived information. Selective perception must achieve two objectives. Quantitatively, it should maximize the agent's awareness of its environment, while protecting its cognitive system from perceptual overload. Qualitatively, it should provide information that is relevant to the agent's current cognitive activities, while remaining vigilant for other potentially interesting

information. Both of these objectives carry real-time constraints: the agent's perception of and reasoning about interesting event must keep pace with the rate at which they occur. The proposed architecture addresses the quantitative objective by modulating the agent's global perceptual input rate relative to its dynamic cognitive processing load [Washington and Hayes-Roth. 1989]. A feedback control mechanism modulates the global input rate based on the rate at which the cognitive system retrieves perceptual data from its input buffer. A predictive control mechanism modulates it based on additions or deletions of reasoning tasks in the agent's cognitive control plan. As discussed below, the control plan contains descriptions of cognitive tasks the agent currently intends to perform-immediately or at some future time. The architecture addresses the qualitative objective by differentially distributing the global perceptual input rate among different kinds of perceptual inputs. In particular, it allocates a higher proportion of the global input rate to inputs that are relevant to the agent's current reasoning activities as indicated by messages from the cognitive system. The only exception to this policy occurs when critical values are perceived; they are sent immediately to me cognitive system regardless of relevance.

Interrupt Ability of Reasoning:

Like other blackboard and related architectures [Engelmore and Morgan. 1988; Georgeff and Lansky.1987], the proposed architecture provides a structure within which expertise is partitioned into discrete chunks that are independently triggered and selected for execution. In addition to permitting rapid context-switching when unanticipated time-sensitive events occur, the discrete triggering of chunks of expertise supports strategic approaches to meeting time constraints at both the reasoning and control levels

Anytime Knowledge Retrieval:

Several approaches have been taken to speeding up the cyclic knowledge retrieval process underlying all reasoning: optimizing algorithms for pattern matching [Forgy. 1982]; parallel pattern-matching algorithms [Gupta, 1985]; and restrictions on the expressiveness of patterns to be matched [Tambe and Rosenbloom. 1987]. Although improving the speed of knowledge retrieval helps an agent to meet real-time constraints on its behavior, it will not protect the agent from the effects of monotonic increases in the size of its knowledge base or sporadic changes in the rate at which important events

occur. To address these issues, an "anytime" knowledge retrieval mechanism is parameterized by the agent's dynamic focus of attention and time stress as expressed in its dynamic control plan (discussed below). The agent uses these parameters to heuristically order, prune, and interrupt its knowledge retrieval. This mechanism provides an additional speed-up in knowledge retrieval and. more importantly, a context-specific modulation of the speed-quality trade-off [Hayes-Roth and Collinot, 1993]. Thus an agent can "think quickly" by retrieving a small number of its potentially applicable operations before choosing the best one found so far for execution. At the other extreme, it can "think carefully" by retrieving all of its applicable operations before choosing the best available one for execution. The quality of its control plan determines the trade-off between the agent's speed of thought and the quality of the operations it executes. In particular, if the agent has a very specific control plan that is well designed to achieve its goal. It can afford to prune knowledge retrieval severely without compromising the quality of its behavior.

Temporal Representations:

Temporal representation permits an agent to reason about both temporal phenomena in its environment and the temporal properties of its own computations [Allen. 1983; Dean. 1985; McDermott, 1982; Sboham. 1989; Vere. 1981]. In the proposed architecture, an agent's mental Stale is organized as an interval-based timeline that supports reasoning with standard temporal relations (e.g.. It, Ie, eq, gt, ge). The timeline also distinguishes occurrences, expectations and intentions regarding conditions in the environment and permits. Instances of these entities may represent different types of phenomena (e.g.. actions, signs, inferences) and they may be linked to one another with semantic relations (e.g.. causes, predicts, explains, promotes). Thus, for example, the occurrence of a particular undesirable sign might justify an intended action that will cause an expected improvement in subsequently observed signs. Comparisons among corresponding intervals on these timelines inspire much of the agent's reasoning. For example, discrepancies between corresponding observations and expectations suggest either a perceptual error or an imperfect model of the world. Similarly, discrepancies between observations and intentions suggest either a perceptual error or an unsuccessful plan of action.

Dynamic Global Control of Reasoning:

The most distinctive feature of the proposed architecture is its explicit representation of control plans that can be changed by the agent at run time [Hayes-Roth. 1985]. An agent can use this data structure to record its reasoning and conclusions around the kinds of operations it intends to perform and to keep a history of its operations and their supporting control plans. The primary use of control plans is to determine which of the agent's situation-triggered operations it will perform at each point in time. The basic mechanism, which permits different implementations takes the current active control plans, uses them to evaluate current situation-triggered operations, and selects the operation with the highest evaluation to execute next. Additional control-related uses of the control data structure include providing parameter values to knowledge retrieval, selective perception, and action control processes. Other uses include explaining and justifying computational behavior and analyzing past behavior to improve future control policies. Of particular interest in this paper, control plans may reflect or enforce real-time constraints on behavior.

3.6.2. Real Time AI Techniques In The Agent's Reasoning Methods:

There are two general approaches to time sensitive reasoning in the literature. First, discretionary allocation of resources to "anytime" or "flexible" algorithms allows an agent to monotonically improve the quality of its results with increases in computation time [Dean and Boddy 1988; Horvitz 1988; Korf 1990]. Second, run-time choice among alternative "approximate" algorithms allows the agent to make different discrete time/quality trade-offs in the results of its reasoning [Bonissone and Halverson 1990; Decker, et al, 1990; Lesser, et ai, 1988]. Because of the variety of operating conditions they face, adaptive intelligent agents benefit from exploiting both of these approaches in each of their component tasks (e.g., pattern recognition, diagnosis, prediction, planning, explanation). This section illustrates the utility of having access to multiple approximate and flexible algorithms for one of these tasks: diagnosis. *PCT* (Parsimonious covering theory) [Peng and Reggia. 1990] is an associative method of diagnosis in which bipartite graphs connect signs to diseases. Given prior probabilities of specified diseases, causal strengths of links to specified signs, and truth information on the signs, PCT calculates relative likelihood scores and ranks any number of multiple-disorder hypotheses. As new data arrive, PCT updates its diagnostic

probabilities accordingly, PCT method offers wide coverage of the problem domain and can generate multiple-disorder hypotheses. Its weakness (in the present implementation) is its limitation to bipartite graphs and, therefore, its inability to compute probabilities for intermediate nodes representing pathophysiological states [pearl 1988]. From a real-time perspective, PCT offers weak "anytime" properties. It begins producing a diagnosis as soon as a fault sign is detected and refines its diagnosis (improves its estimates of probabilities on competing hypotheses) over time. At any point along the way, the agent can decide to interrupt diagnosis and act upon the current best hypothesis MFM (multi-level flow model) [Lind, 1990; Larsson, 1994] is a model-based method for diagnosis, prediction, and explanation. Its causal diagrams are organized around part-whole models of structure (anatomy) and behavior (physiology) and means-ends models of the functions achieved by those structures and behaviors (i.e., goals representing desired physiological dynamic conditions, faults representing pathophysiological conditions)). MFM qualitative advantage over the other diagnosis methods is its ability to explain it with an intuitive physical model. From a real-time perspective, it offers speed (MFM is the fastest of the four diagnosis methods) and weak "anytime" properties. Like PCT, MFM begins diagnosis as soon as a fault sign is detected and refines its diagnosis (improving its model of which structure or function components are causing the fault) as new data arrive. Again, at any point along the way, the agent can decide to act upon MFM's current diagnosis. ICE (Hewett and Hayes-Roth, 1990; 1994] is a different model-based method for diagnosis, prediction, and explanation. Starting with structure-function models similar to those of MFM, ICE more completely annotates those models width topological, geometric, pan-whole, causal, and other relations. It uses a set of similarly represented abstract models of generic systems that can be instantiated at a number of specific loci within different physical systems (e.g., flow systems, diffusion systems, delivery systems), along with causal models of their potential faults. By instantiating generic models within particular reasoning contexts, ICE can diagnose, predict, and explain a large number of specific faults from qualitative first principles. For example, ICE can use its generic flow model to reason about the flow of gases in the pulmonary system or in the ventilator, the flow of blood in the cardiac system, or the flow of nutrients in the digestive system. Its qualitative advantage over the other diagnosis methods is its ability to handle a much larger space of potential faults, including many that may not be represented explicitly in the agent's domain knowledge,

and its ability to explain its diagnoses and predictions with intuitive physical models of both the domain and abstract physical systems. From a real-time perspective, ICE offers little. It is by far the slowest of the four diagnosis methods and useful primarily when the agent has no hard time constraints or when the agent is dealing with an unfamiliar problem outside of its domain knowledge.

ReAct [Ash, Gold, Seiver, and Hayes-Roth, 1993;Ash and Hayes-Roth, 1993] is a decision-theoretic method for diagnosis and therapy planning. Unlike other decision-theoretic approaches, it organizes diagnostic knowledge in an action-based hierarchy of fault classes. Higher-order nodes represent classes of faults that: (a) manifest similar symptoms; and (b) are amenable to similar non-specific treatments. Leaf nodes in the hierarchy represent specific faults that are amenable to specific treatments. Given an observed abnormal sign, ReAct attempts to refine its current best hypothesis by performing tests that distinguish among its children. When a deadline occurs, ReAct performs the treatment action associated with its current best hypothesis. By definition, this action is expected to give positive value for any subordinate class of faults or specific fault. Thus, "ReAct provides stronger "anytime" properties than PCT or MFM. It begins diagnosis as soon as an abnormal sign is detected. It improves (increases the specificity of) its diagnosis monotonically over time as additional data become available. At any time along the way, the agent can decide to act upon ReAct's current best hypothesis. Note that ReAct offers a more substantive opportunity for action on deadline than either PCT or MFM because it precompiled the best applicable treatment actions with every node in the hierarchy. Finally, ReAct structures its fault hierarchies to guarantee the best possible anytime performance, given the set of specific fault, tests, and actions of interest. Within the proposed agent architecture, all reasoning methods are implemented as collections of situation-triggered operations and associated strategies for selecting and sequencing operations to achieve goals under varying contextual conditions. Thus, for example, when the agent detects a fault condition, any of the four diagnosis methods described above might be triggered and available for application to the problem. (In many cases, only a subset of the methods actually is triggered because they cover different parts of the fault space and have somewhat different information requirements.) At the same time, control planning operations are triggered and executed to establish the agent's intent to treat the detected problem and to impose context appropriate constraints on its approach. In particular, when real-time is

the critical factor, the control plan reflects this constraint and the agent bases its choice among triggered diagnosis methods on their relative speed and degree of any time properties. As discussed above, three of the four diagnosis methods offer only a subset of the conventional properties of "anytime" algorithms, but they offer additional advantages. Like conventional "anytime" algorithms, they can be interrupted at any point after they have begun working and, on average, they will deliver better results when allowed to work longer. In addition, ReAct provides a guaranteed optimal use of resources for a given set of operating conditions and run-time measurements of the cost and benefit of interrupting versus continuing the diagnosis. However, conventional "anytime" algorithms improve their results by computing longer, whereas these diagnosis methods improve their results by requiring more data-which happens to arrive over a period of time. Thus, an agent cannot improve its diagnosis simply by allocating more computing resources to the task. To improve its diagnosis, it must get more relevant data faster, which it may be able to do by deciding to perform other perception, reasoning, or action tasks. On the other hand, while an agent is waiting for relevant data during its diagnostic reasoning, it does not monopolize the computational resource. The agent can interleave other important reasoning activities without compromising the quality or timeliness of its diagnostic results.

3.6.3. Real-Time AI Techniques In The System's Control Strategy

AI Techniques for Task Control:

In conventional real-time systems [Liu, et al, 1991; Zhao, et al, 1987], an agent's performance of multiple tasks is controlled through deliberative scheduling of bounded-time computations. Following closely in this paradigm, some researchers introduce unbounded AI methods to improve the scheduling of bounded-time non-AI computations, the latter of which are guaranteed to meet hard deadlines [Hendler and Agrawala. 1990; Musliner, et al. 1993]. Note that this approach does not guarantee that the AI methods will run in time to have a useful effect on the schedule or that important higher-level goals will be achieved. This approach works best in domains where most of the agent's work is accomplished by schedulable non-AI methods and where opportunities for applying the AI methods are detected early enough for their results to be computed and applied effectively. It is less appropriate in domains where most of the agent's work requires a variety of uncertain AI methods and where the timeliness of those tasks influences the utility of its behavior. Other researchers use heuristic

techniques to produce a real-time schedule of the AI computations themselves. For example, the "design-to-time" approach [Garvey and Lesser, 1993] exploits the availability of alternative AI methods having different expected computation times to heuristically search for a feasible configuration of methods to achieve a goal directed schedule of AI tasks. When the requirements of desired tasks exceed the available resources, the agent can: (a) select faster methods and accept lower-quality results for some of the tasks within the current schedule; or (b) postpone some of the tasks. Similarly, the "deliberation scheduling" approach [Boddy and Dean, 1993; Horvitz and Rutledge, 1991; Russell and Zilberstein, 1991] uses decision-theoretic techniques to determine how much time to allocate to each of a goal-directed series of "anytime" methods. Here, when time requirements exceed resources, the agent can: (a) allocate less time and accept lower-quality results for some of the tasks within the current schedule; or (b) postpone some of the tasks. In both cases, the aim is to find the schedule of tasks, methods, and resources that achieve the best possible goal related results, given the available resources. As discussed below, both the design-to-time and deliberation scheduling techniques are useful within a more general framework for an agent's dynamic planning-control of its behavior.

Proposed Approach to Strategic Control:

In the proposed approach, which integrates and extends several of the ideas discussed above, the agent generates explicit plans for its own computational behavior, dynamically at run time. Basically, whenever the agent commits to a new goal, it generates (or retrieves and instantiates) a plan that specifies the sequence and timing of tasks it intends to perform in order to achieve the goal. Each of a goal's planned tasks inherits the goal's priority and is activated and deactivated when associated activation and deactivation conditions are met. In addition, planned tasks that have earliest start times and hard deadlines may be assigned positions on a real-time schedule. On each iteration of its control cycle, the agent identifies and executes me best currently triggered computation, where "best" is evaluated relative to currently active control plans, including any associated schedule constraints. This approach to control planning differs from the scheduling approaches discussed above in the following ways. First, the agent can construct and use control plans to describe its intended course of behavior at context-appropriate levels of abstraction [Hayes-Roth, 1993; Hayes-Roth, et al, 1993]. In particular, a control plan may describe the sequence of tasks the agent intends to

perform (e.g., confirm apparent fault. diagnose cause of apparent fault. plan action to correct cause of apparent fault), along with some constraints on their performance (e.g., reliable, fast, precise, complete), without necessarily specifying the particular methods (e.g., associative vs. model based diagnosis) to be used for the planned tasks. As it makes its way through the plan, the agent chooses situation-triggered computations that perform the planned tasks by whichever method best satisfies their constraints-including real time constraints as discussed below.

For example, here is a partial control plan for the cognitive behavior of an ICU patient monitoring agent that has detected a patient's suddenly high blood pressure:

(Treat (high BP, t)
    ({Importance = max} {urgency = high})
(Diagnose (BP, t),
    ({Importance = max} {speed = fast})
(Plan treatment (diagnosed fault),
    ({Importance = max} {speed = fast})
(Monitor (treatment plan),
    ({Importance = max} {vigilance = high})

The agent intends to treat the patient's observed high blood pressure at time t, which it assigns maximum importance and high urgency. Therefore, it instantiates a three-step control plan for treating important urgent problems. First, it will diagnose the blood pressure using its fastest applicable method. After completing this task, it will plan a treatment for the diagnosed fault again using its fastest applicable method. Finally, it will monitor execution of the treatment plan using its most vigilant method. As this example illustrates, the plan expresses only the intended tasks, their required parameter bindings, and discretionary constraints on the evaluation of alternative candidate methods. The specific diagnosis, treatment planning, and treatment monitoring behavior of the agent will depend on run-time circumstances. Abstract plans support real-time performance in two ways. First, they save time. The agent can generate abstract plans faster than specific plans because it has fewer decisions to make and easier constraints to satisfy. In addition, the agent doesn't have to replan, as often because abstract plans are more robust than specific plans. The agent can behave

consistently with its high-level goals in different run-time relations in which different combinations of task-relevant methods actually are triggered. Second, abstract plans promote the highest-quality feasible performance. When the agent's planned behavior runs more quickly than anticipated, it can select the most effective available methods for planned tasks given the actual time available, rather than committing early to less effective methods because of conservative predictions of available time. Of course, there are situations in which an agent needs to plan and control its course of behavior in detail. Other things being equal, as the specificity of the desired performance increases or the availability of resources decreases, the agent should plan its course of action in more detail. The control plan representation permits the agent to construct and follow plans at any desired level of abstraction. Second, the agent can construct and use a real time schedule for any context-appropriate subset of the tasks in its control plan [Lalanda and Hayes-Roth, 1994]. Note that inclusion of tasks in the real-time schedule reflects only their requirement to satisfy real-time constraints (e.g., hard deadlines), but is independent of their importance. Both important and unimportant tasks can be included or not included in the real-time schedule, at the agent's discretion. For example, the ICU monitoring agent discussed above might decide to assign its activities for treating the patient's high blood pressure to its real-time schedule. Thus, it might make a judgment about the deadline for alleviating the patient's high blood pressure, estimate the expected durations of its component tasks, and nm a scheduling algorithm to insert these activities at appropriate places in its current real-time schedule. The real-time schedule is a useful tool for the agent to use in determining when it will perform particular tasks. In addition to alerting the agent at run time to tasks with imminent deadlines, the schedule provides a global temporal perspective on the feasibility of performing planned tasks. Unlike the AI scheduling approaches discussed above, however, the real-time schedule is neither guaranteed nor strictly enforced. The inclusion of a task in the schedule does not guarantee that the task will meet its deadline or even that it will be performed. The schedule is only one of several considerations the agent makes in deciding which behaviors to execute at each point in time. Allowing some or many planned tasks to remain unscheduled saves scheduling time by reducing the complexity of the scheduling task and the number of times the agent must reschedule (i.e., whenever a new task must be scheduled or the current schedule fails). It also promotes the overall quality of its performance in terms of number and importance

of goals achieved. The agent can perform the most important tasks available by the highest-quality applicable methods, as determined at run-time, instead of unnecessarily pruning tasks from a carefully constructed schedule based on conservative predictions of time requirements. Third the agent can use context-appropriate methods for different planning and scheduling activities. [Johnson and Hayes-Roth, 1987] show how an agent can integrate the use of different planning methods. For example, it can use skeletal plan instantiation to generate a standard sequence of activities for achieving a known type of goal. It can use goal-directed reasoning to generate task sequences for achieving novel goals. It can use opportunistic reasoning to modify or augment an active Control plan, based on run-time events. Similarly, an agent can use different scheduling methods in different situations. For example, the agent can use deliberation scheduling [Baddy and Dean, 1993] when its situation presents: insufficient processing time, hard deadlines, and anytime reasoning methods. Alternatively, the agent can use design-to-time scheduling [Garvey and Lesser, 1993] when its situation presents: insufficient processing time, bard deadlines, and multiple reasoning methods with different time/quality trade-offs. A variation on the design-to-time algorithm [Lalanda and Hayes-Roth, 1994] schedules only tasks, not methods, and estimates time requirement for a task as the modal time requirement among the applicable methods. In addition, since some important tasks are not scheduled, the algorithm reserves some of the available time for them. Planning and scheduling its own behavior are reasoning tasks that an agent performs from time to time. Like other reasoning tasks, the agent decides when to perform them and which of several methods to apply. Thus, the agent's performance of these tasks benefits from all of the real-time techniques available for its other reasoning tasks. Finally, at each point in time, the agent evaluates situation-triggered possible operations against three criteria expressed in its control plan: the importance of the task the operation would performed, the quality of the expected result, and the urgency of the task. The importance of a last in the control plan is inherited directly from its goal. The quality of the expected result of applying an operation to a task is measured against constraints on the performance of the task (e.g., explainability, precision, completeness, certainty), which are expressed in the control plan and the degree to which the method represented by the operation meets those constraints. Some aspects of a method's quality may be stored with it as constant attributes; others may be computed as functions of contextual factors (e.g., certainty as a function of input data

properties). The urgency of a task is measured as an increasing function of the proximity of its deadline and as a decreasing function of the slack time in the schedule (the maximum delay in the start of a scheduled task that can be tolerated without causing subsequently scheduled tasks to miss their deadlines.). Other things being equal an operation's priority increases with its rating against each of these component criteria. By weighting them differentially, an agent can modulate its comparative emphasis on the timeliness, quality and importance of the goals it achieves.

3.6.4. Emergent Real-Time Properties in an Agent's Behavior

This section illustrates how techniques at different levels of the framework work together to support real-time properties in the agent's behavior. It presents an excerpt of the cognitive behavior of Guardian, an adaptive intelligent agent for monitoring ICU patients. Basically. Figure 3 shows Guardian using an explicit control plans to choose among situation-triggered cognitive operations and changing its control plan based on run-time information. The following discussion focuses on the interactions among real-time techniques in Guardian's architecture, reasoning, and control during three phases of this episode: phase 1, the initial monitoring; phase 2, the response to an urgent problem; and phase 3, the continued monitoring. During phase I, Guardian believes there are no immediate problems. On its timeline, both intentions and expectations specify normal values for all patient data variables. Guardian is minimally time stressed; its main concern is to do as much monitoring as possible without allowing its monitoring results to fall behind the rate of events in the world. Guardian makes two control plans one to monitor all patient data and one to update its control plan as necessary. Because it has only a single active reasoning task and because that task is to monitor all patient data. Guardian instructs its perception system to set a high global input data rate and to distribute it evenly among all perceived variables. As a result, the perception system sends a series of observations, each of which triggers a monitoring operation at the cognitive level. During the first few cycles, the feedback control mechanism monitors the rate at which perceptual inputs are retrieved from the cognitive input buffer for reasoning and adjusts the global input rate accordingly. After that, the global input rate is stable. Since there is only one active task. The global data rate has been set appropriately, and initially perceived values are normal as expected. Guardian has no trouble triggering and executing monitoring operations fast enough to keep up with

important events in its environment As it executes monitoring operations for different variables (e.g.. blood pressure and heart rate) on successive cycles, Guardian records its interpretations of perceived values and trends as occurrences on its timeline. During phase 2, Guardian is in a very different situation. One of its monitoring operations has revealed a low value for the patient's blood pressure. Because this contradicts the normal value for blood pressure recorded in the corresponding interval of its intention and expectation timelines, Guardian identifies the low blood pressure as a serious and time-critical problem. Now Guardian is time stressed; it needs to resolve the blood pressure problem quickly to head off more serious consequences. Guardian makes a new control plan to respond (diagnose, treat, and monitor) quickly to the low blood pressure. Because it has planned an additional reasoning task, it has reduced resources for handling perceptual inputs; it instructs the perception system to reduce the global input data rate. It also instructs the perception system to allocate a greater proportion of its global data late to data that are relevant to the new control plan and to distribute the remaining pan of the global data rate among non-relevant variables. Similarly, the importance of the new plan and its time sensitivity cause the knowledge retrieval mechanism to focus on instantiating reasoning operations that meet the plan's requirements: fast operations for diagnosing, treating, and monitoring the patient's blood pressure problem. For example, Guardian chooses diagnosis method D1 because it is faster than method D2. On each cycle, as soon as Guardian instantiates an operation that matches its plan, the knowledge retrieval mechanism is interrupted so that the matching operation can be executed immediately. As with its monitoring operations during phase I, Guardian records the results of its reasoning about the patient's low blood pressure (diagnostic hypotheses and conclusions. planned treatments, expected results of treatments, etc.) as occurrences, expectations, and intentions assigned to appropriate intervals on its timeline. The plan for responding to the patient's low blood pressure continues to dominate Guardian's performance until the problem is resolved. Other kinds of reasoning operations (e.g.. monitoring of nonrelevant patient data) are triggered and executed only when no plan-relevant operations are available. During phase 3, Guardian has resolved the blood pressure problem, terminated the associated plan, and resumed monitoring all patient intentions and expectations of normal values for all patient data variables. It instructs its perception system to raise the global input data late back to its previous high level and to distribute it evenly among all perceived variables.

After a few cycles of feedback control, the global input late is stable. Again, under these parameters, Guardian has no trouble triggering and executing monitoring operations fast enough to keep up with important.

3.7. Neural Networks For Robot Control

The use of neural networks for robot control tasks constitutes an example of a very powerful approach to nonlinear process control: the observations or experience-based learning of the controllers, as opposed to the model-based design of classical controllers. In this context, learning in the neural networks can efficiently replace the detailed mathematical modeling of some complex control processes, as has been shown in the literature in the last few years (see e.g. Narendra, 1990; Cembrano et al., 1992; Hunt et, al., 1993). Robot control is one of the fields where extensive research is being performed along this line, especially oriented towards achieving greater degrees of autonomy in robot operation. The aim of this paper is to present a brief review of the state of the art in the application of neural networks to the solution of control problems in robotics. One of the most important applications of neural networks in control is the identification of complex nonlinear systems, because of their ability to approximate broad classes of nonlinear functions. In the domain of robotics, the most relevant system identification problems are the kinematics and dynamics. In its direct and inverse versions, the kinematics problems are concerned with the mapping between the joint coordinates of the robot and the resulting end-effector position and orientation in the physical space. Similarly, the direct and inverse dynamics problems deal with the relationship between joint motor commands and the resulting response of the end-effector. Neural net, works have also been shown to perform efficiently as controllers of nonlinear processes, alone or in conjunction with other type of controllers. Due to the adaptivity properties of the neural learning models, they are especially useful for adaptive control problems. One of the best examples of such problems in robotics is the dynamic control, concerned with the on-line generation of the appropriate control commands at the robot joints so as to achieve a desired trajectory of the end effector. This is a complex nonlinear control problem, where unknown and variable parameters appear. Another broad area for the use of neural networks in control is sensor motor control and, in particular, image-based control. The use of visual control in robotics has become a major focus of research, largely due to the advent of computers with sufficient

speed to allow including image processing in the control loop. Classical techniques for image processing and image-based positioning remain, however, severely limited by the requirements of very specific and precise knowledge of the environment" the objects to be viewed, the camera and the robot characteristics. The flexibility and the learning capability of neural networks can be efficiently exploited for tackling the problem of visual positioning, even in an unstructured or unknown environment. The following sections describe the authors' research on the use of neural networks in the solution of t inverse kinematics, dynamic control and visual positioning problems. In each section, a brief summary of the state of the art, including classical techniques, is provided, as well as the description of the adopted neural solutions and their results.

Overview of Applicable Technology:

The knowledge of inverse kinematics, which relates the end-effector coordinates with the required coordinates in the robot joints, is essential in any robot control application. Most commercial robots provide explicit mathematical models of their inverse kinematics, which are implemented in their execution controllers. However, robots frequently undergo mechanical changes during operation causing miscalibrations with respect to the original kinematical model. It is reasonable to expect that a robot arm with a certain degree of autonomy to be capable of relearning or recalibrating its inverse kinematical relationship. Most of the neural approaches to inverse kinematics have involved supervised learning schemes, applied to solving the non-redundant problem. The two learning rules most widely used have been LMS and backpropagation. Among these works, we find Miller et al. (1987), Kawato et al. (1987), Guez and Selinsky (1988) and Goldberg and Pearlmutter 1988). The most significant result in these papers is the demonstration of the ability of neural networks to learn this complex mapping. Other researchers, like Ahmad and Guez (1990), have studied the combination of backpropagation networks and a conventional method, or the use of specialized neural structures (Kozakiewicz et al., 1992) in order to improve the overall accuracy and convergence. The problems of redundancy and avoidance of singular configurations, which usually pose serious difficulties for conventional methods, has been treated in the literature, for example by Tanaka and Shimizu (1991) and DeMers and KreutzDelgado (1993). The general approach in these works is to optimize a performance criterion that includes manipulability and singularity avoidance terms, through the use of the Jacobian of the inverse transformation. A different

approach to inverse kinematics is to deal with the problem from a more behavioral point of view: relating images of the end-effector position and orientation to the required coordinates in the joints. Consequently, in this approach, it is not required to know the actual Cartesian coordinates of the end-effector, but the goal is to relate sensory information directly to joint coordinates. This approach is, in fact, very appropriate for visuomotor control and is sometimes referred to as the hand-eye coordination problem.

Self-organizing Topologic Maps for Learning Hand-eye Coordination:

The approach by Martinetz et al.(1990) and Ritter et al. (1992), with its special form of Kohonen topologic maps, is based on unsupervised learning. The target position of the end-effector is defined as a spot registered by two cameras looking at the workspace from two different vantage points. Neurons are arranged in a 3D lattice to match the dimensionality of physical space. The learning process makes this lattice converge to a discrete representation of the workspace. Each neuron i has an associated four-dimensional vector W_i representing the retinal coordinates of a point of the workspace. The response of the network to a given input u is the vector of joint angles θ_k and the 3x4 Jacobian matrix A_k associated with the winning neuron k. The joint angles produced for this particular input are then obtained with the expression:

$$\theta(\mathbf{u}) = \theta_k + \mathbf{A}_k(\mathbf{u} - \mathbf{w}_k).$$

A learning cycle consists of the following four steps:

1. First, the classical Kohonen rule is applied to the weights:

$$\mathbf{w}_i^{new} = \mathbf{w}_i^{old} + c\, h_k(i)\, (\mathbf{u}(t) - \mathbf{w}_k(t)),$$

Where c is the learning rate and $h_k(.)$ is a Gaussian function centered at k used to modulate the adaptation steps as a function of the distance to the winning neuron.

2. By applying 9(u) to the real robot, the end-effector moves to position u' in camera coordinates. The difference between the desired position u and the attained one u' constitutes an error signal that permits applying an error-correction rule, in this case the LMS rule:

$$\theta^* = \theta_k + \Delta\theta = \theta_k + \mathbf{A}_k(\mathbf{u} - \mathbf{u}').$$

3. By applying the correction increment Ak(Uu') to the joints of the real robot, a refined position u" in camera coordinates is obtained. Now, the LMS rule can be applied to the Jacobian matrix by using Δu = (u -u") as the error signal:

$$\mathbf{A}^* = \mathbf{A}_k + (\Delta\theta - \mathbf{A}_k\Delta\mathbf{u})\frac{\Delta\mathbf{u}^T}{\|\Delta\mathbf{u}\|^2}.$$

4.Finally, the Kohonen rule is applied to the joint angles:

$$\theta_i^{new} = \theta_i^{old} + c'\ h'_k(i)\ (\theta_i^* - \theta_k(t)),$$

And the Jacobian matrix:

$$\mathbf{A}_i^{new} = \mathbf{A}_i^{old} + c'\ h'_k(i)\ (\mathbf{A}_i^* - \mathbf{A}_k(t)),$$

Where again c' is the learning rate and h'$_k$ (.) is a Gaussian function centered at k used to modulate the adaptation steps as a function of the distance to the winning neuron. The previously cited works showed this method to converge and to self-organize into a reasonable representation of the workspace in a limited number of iterations.

With appropriate modifications, the algorithm described above has been used by the authors' group for the purpose of inverse kinematics recalibration in a simulated space robot. Firstly, a 3-degreeof-freedom case was solved and the system is currently being extended to 6 degrees of freedom with success. The implementation in the real robot is now in its early stages, and is scheduled to be concluded at the end of 1994. A number of decalibration tests have been performed, ranging from slight variations in parameters of the kinematic model to structural changes in the links or the joints. In all cases, the system has been able to reconstruct its kinematic mapping in a limited number of learning iterations. For the worst-case problems, involving a complete re-learning of the inverse kinematics transformation, this has been achieved in 3000 learning cycles, to an accuracy of less than 1mm in position. The accuracy of the

inverse kinematics correction is dependent on the discretization of the space chosen for the 3-D Kohonen map, but, in principle, the method should converge to any desired accuracy, if the appropriate number of Kohonen cells is provided.

Controlling Inverse Dynamics With Neural Networks:

The inverse dynamics problem consists of the control of the dynamic trajectory of the arm end-effector, through the motor commands at its joints, so that it follows a desired reference path, with a high degree of accuracy and precision. In state-space-representation terms, the state variables of this problem are the position, velocity and acceleration of the end effector and its control variables are the motor commands (currents or voltages) at each joint. The efficient solution of the dynamics control problem through conventional control schemes world require a deep knowledge of the system behavior, translated into a very accurate nonlinear mathematical model (see for example Zomaya and Nabhan, 1993). However, in the applications described in this paper, as in many others, this type of model is not available. Furthermore, the complexity of the model makes real-time implementations computationally intensive and the addition of payloads to the system may greatly affect the overall dynamic behavior. A feasible approach in conventional adaptive control is the determination of an approximate piecewise-linear model of the nonlinear plant, and the synthesis of appropriate linear controllers the different operating conditions corresponding to each one of the linear models. A common procedure in this approach is the use of indirect adaptive control, where an adaptive identifier of the plant updates the parameters of the plant model online and the parameters of the linear controller are then computed with the estimated plant values. Although this approach has been successful in several applications, the stability conditions for convergence of the adaptive schemes are based on a relatively slow time variation of the plant parameters with respect to the control adaptation times (see for example Astrom 1991 and Sastry and Bodson, 1991). The use of linear models to approximate nonlinear processes leads, in general, to models with time-varying parameters, where the parameter variations in time may be too fast to guarantee convergence of the adaptive control algorithms. For these reasons, the neural identification and control is very relevant to the problem of the robot dynamics. On the one hand, the ability to learn a nonlinear behavior through appropriate examples of inputs and outputs may overcome the modeling difficulties. On the other, it is expected that the use of neural networks will produce more efficient nonlinear approximations of

the plant model, which do not require such a fast adaptation of parameters as linear approximations, thus providing more adequate conditions for stability of the closed-loop systems. A general approach to adaptive control, proposed by Narendra (1990), consists of an indirect model reference adaptive scheme with a series-parallel structure. Similarly, several concepts related to neural adaptive control are treated in Cembrano and Wells (1992) and Hunt et al. (1993). A number of more refined versions of backpropagation have been proposed in the literature for adaptive control and proved in a variety of applications e.g.. Tzirkel-Hancock et al. (1992), Hoskins et al.(1993), Sbarbaro-Hofer et al. (1993), Loke and Cembrano (1994). A different type of neural network, CMAC (Cerebellar Model Articulation Controller), originally developed by Albus (1975), has also been used in similar neural adaptive control schemes. It has been demonstrated by Miller et al. (1990), (1992), (1994) for real-time control of industrial robots. A comparison between an adaptive dynamic control scheme using back-propagation and a CMAC-based control is provided by Ananthraman and Garg (1993), for a 2-dof SCARA robot. The CMAC scheme is shown to have considerably better model-tracking capabilities in this simplified problem, as well as a much faster convergence of the learning algorithm.

CMAC For Dynamic Control of A Robot Arm:

The CMAC network, based on the cerebellar model for neuromuscular control, is basically a nonlinear table look-up technique that maps each N-dimensional input state-space vector to a corresponding output vector of the same or different dimension. Each input vector activates exactly c input neurons with overlapping receptive fields, where c is a variable parameter representing the extent of generalization within the state space. The potentially very large virtual state space is mapped to a smaller physical weight table using a fixed random hashing function. A supervised training method, resembling the WidrowHoff rule, is used to adjust the CMAC memory values based on these observations. The learned information is used to predict the command signals required to produce desired changes in the sensor outputs. Miller et al. have studied a neural-based learning control system for the dynamic control of robot manipulators with multiple feedback sensors and multiple command variables. The system has been studied for the control of two and five-jointed robot arms without vision, and a five-jointed robot arm with vision. In this scheme, a neural network is used, in place of an explicit system model, to adaptively learn an approximate dynamic model of the

controlled robot in appropriate regions of the system state space. The CMAC neural control module adaptively learns the unknown nonlinear mapping between the sensor outputs and the system command variables from on-line observations of each during system operation. In these experiments, the CMAC neural controller is implemented in a closed-loop control system, where it is used to predict the actuator torques required to make the robot follow a desired trajectory. These torques are a function of the current joint positions, velocities and accelerations, and are used as feedforward terms in parallel with a fixed-gain linear feedback controller. The control signal input to the robot is the sum of the terms from the CMAC module and the feedback controller. At first, the network memory values are initialized to zero, and the feedback controller provides initial movements upon which the CMAC controller may learn and improve. The network is trained by adapting the CMAC memory values after each control cycle, based on observations of the control inputs and obtained sensor outputs. The difference between the observed sensor output and that computed by the CMAC module is used to calculate the memory adjustment for the current control cycle using a form of the WidrowHoff LMS training rule. After several control iterations, the network outputs approach the correct values, the error input to the fixed-gain controller decreases to zero (and hence its output also), and the CMAC module effectively takes over full control of the system. The concepts of CMAC have been applied to the control of the OHR in simulation. The simulator of the robot dynamics is a variable structure state-space model of the telescopic harvesting arm of the OHR, which takes into account variations of inertia moments and friction coefficients with the elongation. In order to generate acceptable training data, a PID controller is added to the robot model, since it would otherwise produce unstable responses. The implementation in this project differs from those of the aforementioned works, especially in the complexity of the dynamic behavior of the robot (variable structure), which poses serious difficulties to the learning process. Firstly, a discrete-time state-space representation of the dynamic process was preferred to the original input output scheme and the output error signals were used as feedback signals. Therefore, the inputs to the CMAC network include information on the current state and next desired state, as well as tapped delays of both. Additionally, the original control scheme is not readily applicable in this case, since the linear controller used with the OHR requires at least a derivative component. While in the original scheme the proportional controller could be made to operate in parallel with the

CMAC controller in on-line learning, this is not the case when the linear controller contains derivative and integrator terms. Therefore, in the adopted CMAC controller scheme, shown in figure 3.3. the outputs of the linear and the CMAC controllers are not additive. Rather, the linear controller is initially used for a first phase of training the CMAC controller during this phase; the CMAC module is not connected to the robot. When both controllers provide control signals, which are equal, to a certain degree of accuracy, the CMAC controller is switched on, and a period of on-line training starts, where the neural controller can improve its performance with respect to the PID controller results. The linear controller is a backup system that takes over in case of malfunction of CMAC.

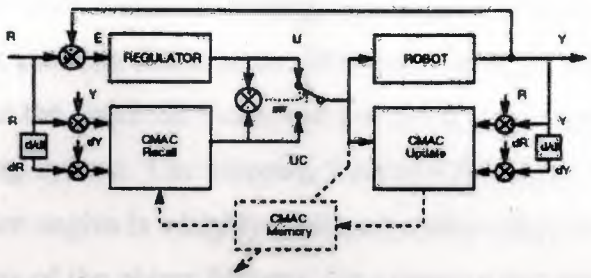


Figure 3.3. Scheme CMAC control of robot dynamics

The results of this implementation show that. CMAC can efficiently learn the inverse dynamics of this complex robot model and that it can improve the results of applying a classical PID controller. Moreover, it is important to take into account that the PID controller was specially designed for the mathematical model of the simulator, so that the PID control results are the best case situation, while poorer results are to be expected in real conditions, where deviations from the model will undoubtedly appear. Conversely, the CMAC learning is not model dependent, so that the real robot conditions should not pose additional difficulties for control.

Learning Visual Positioning:

Many authors have tackled the problem of visual positioning of robot manipulators, for applications ranging from inspection, grasping and assembly of parts to docking and navigation of autonomous guided vehicles. Most of the existing works have tended to rely on simple geometrical features extracted from images, such as

points, lines or circles, together with projection transformations to analytically derive the mapping from 2D image space to the robot Cartesian coordinates (Abidi, 1990; Mandel, 1987; Kabuka, 1987; Chaumett.e, 1991; Espiau, 1992). Although existing approaches based on analytical methods have produced useful results, they all depend on simple geometric features that are assumed to be always visible and extractable in the camera image. In addition, the physical relationship between all object features must be known in order to derive the projection transformations. Furthermore since more than one cue is typically used, matching must be performed to find the correspondence between each feature in the observed image and a feature in the reference image or physical scene. Finally, in order to derive the complete transformation from image features to robot coordinates, the precise relationship between the camera coordinates and those of the end effector must be known, requiring calibration of the camera as well as knowledge of its intrinsic parameters. Obviously, most of these assumptions place strong restrictions on the observed scene, and the operating conditions and robustness to noise of the resulting system. The mapping between 2D image feature deviations and robot position or joint angles is a highly nonlinear relationship that depends on the type and relative positions of the object features, the camera-robot relationship, the intrinsic camera parameters, and the robot kinematics. A neural network may be used to learn the entire transformation implicitly based on training examples, thus avoiding explicit computation of all intermediate transformations. A first work in this area was that of Hashimoto (1992) who used backpropagation networks to learn the mapping between the image deviations of 4 projected points of a viewed object with respect to a "reference" or desired image, and the corresponding joint angles of a robot with a camera mounted on its end effector. Hashimoto's work showed the capability of a neural network to perform visual positioning with a degree of accuracy similar to that obtained using analytical techniques. However, many possibilities by which neural networks can be exploited to produce more general-purpose visual positioning systems remain open for research. For example, training neural networks to map from more global image descriptors to robot commands could make visual servoing methods much less sensitive to the presence and extractability of a few simple geometric features in the observed scene. Similarly, the need for explicit feature matching may be avoided and robustness to obstructed features, noisy images, and changing reference positions may be increased. The approach of this work was to develop more flexible and robust visual

positioning methods based on neural networks. In all the experiments performed, the overall objective was to train a neural network to represent the mapping between the variations, with respect to a desired "reference" image, of several prespecified image features or descriptors as seen from a camera mounted robot end effector, and the 3D Cartesian position and orientation of the end effector relative to the reference pose. Unlike in earlier neural approaches (Hashimoto, 1992), inverse kinematics of the robot was not included in the mapping. In this way aspects related to the visual mapping problem could be studied independently from the robot kinematics and the additional benefit was achieved that the learned transformation is completely independent from the chosen reference position, and may therefore be used to correctly position the robot relative to the object regardless of their initial locations. Two sets of experiments were performed, in which a neural network was used to learn the mapping between a chosen set of extracted image features and the 6D movements required to approximate the camera to a prespecified reference position (a more detailed description of this application is given in Venaille et al., 1994). In initial experiments, four point features were used, thus allowing comparison of results with similar existing works based on analytic and neural methods. The differences between the x, y coordinates of the four points in the reference and observed images were used as inputs to the network. In the second set of experiments, the features used consisted of 32 Fourier descriptors used to encode the shape of the extracted silhouette of an observed object on a uniform background. As before, the differences between the descriptors in the reference and observed images comprised the network inputs. In both cases, training sets were constructed by moving the robot-mounted camera to 1000 random positions in the vicinity of the reference pose, and the extracted feature deviations for each image, along with the applied 6D movement were used as training examples. Backpropagation net works were trained using these data sets and tested In a closed-loop visual positioning system to test their ability to guide the camera back to the reference position from any given initial position and orientation within a limited range. In both sets of experiments, the neural visual positioning systems were capable of converging on the reference position to an accuracy of less than half a pixel in an average of 2 to 10 movements, depending on the range of initial displacements used. Even when the camera was displaced to three times the range of movements used to generate the training set, the networks generalized quite well and were able to achieve the same final positioning

accuracy in just a few more approach movements. Thus, this work demonstrated the capability of using neural networks to accurately position a robot manipulator based on image information even when complex image features are used for which it is not possible to derive explicit transformation relationships or perform feature matching.

CONCLUSION

In this thesis the basic concepts of the radio communication, interfaces between computer and modem and intelligent control methods are trying to be covered. As a result, my aim in this project is to design a kind of mobile robot, which can go from base point to destination point. Robot structure is simple and robust based obstacle avoidance and navigation feedback. A compass sensor can be also added to the electronic of the robot. This circuit takes the data from the beginning of the movement of the robot and sends them to the computer in wireless transmission. A 56 Kbaud RF modem is taken for this purpose. And in the computer side there must be a program running to receive the data from the robot via this modem and sends the commands to the robot again via this modem. On the computer side the path of the robot, edges of the corridor that it goes through, obstacles and temperature informations can be monitored. Well, this system can be thought as a basic concept for industrial manufacturing controlled via wireless data communication and intelligent control systems. It can be imagined there can be several robots like this one I described above in a factory and a main control unit that controls these terminals simultaneously. It can be seen such as these techniques with RF cells and a control unit in high industries.

REFERENCES

1. Agre, Philip E., and Chapman, David. 1990 What are plans for? Robotics and Autonomous Systems.

2. Ballard, Dana H. 1991. Animate Vision. Artificial Intelligence

3. Betke, M., and Gurvits, L. Mobile Robot Localisation Using Landmarks. Technical Report, SCR-94-TR-474, Siemens Corporote Research, Princeton, N.J.

4. Borenstein, Johann, and Koren, Yoram 1992. Noise Rejection for Ultrasonic Sensors in Mobile Robot Applications. In Proceedings of the IEEE International Conference on Robotics and Automation. 1727-1732. Los Alamitos, Calif:IEEE Computer Society Press.

5. Borenstein, Johann and Koren, Yoram 1989. Real Time Obstacle Avoidance For Fast Mobile Robots. IEEE Transactions on Systems Man and Cybernetics 19(5):1179-1186

6. Garne, F. Bryan. Telecommunications. PRIMER: Data, Voice and Video Communications, International Editions, 1999

7. Mamedov, Fakhreddin Telecommunications. Nicosia 200

8. Parsons J.D., Jardine D., Gardiner J.G. Mobile Communication Systems Blackie(Glasgow), Halsted (New York),1989.

9. Rhee Man Young . Cellular Mobile Communications and Network Security. Prentice Hall International Editions. 1999

10. Warren Hioki. Telecommunication. 2nd edition. Prentice Hall International Editions,1995.

Online References

Data Communication, http://www.us-epanorama.net

Mobile computing, http://www.yahoo.com/Computers/Mobile_Computing

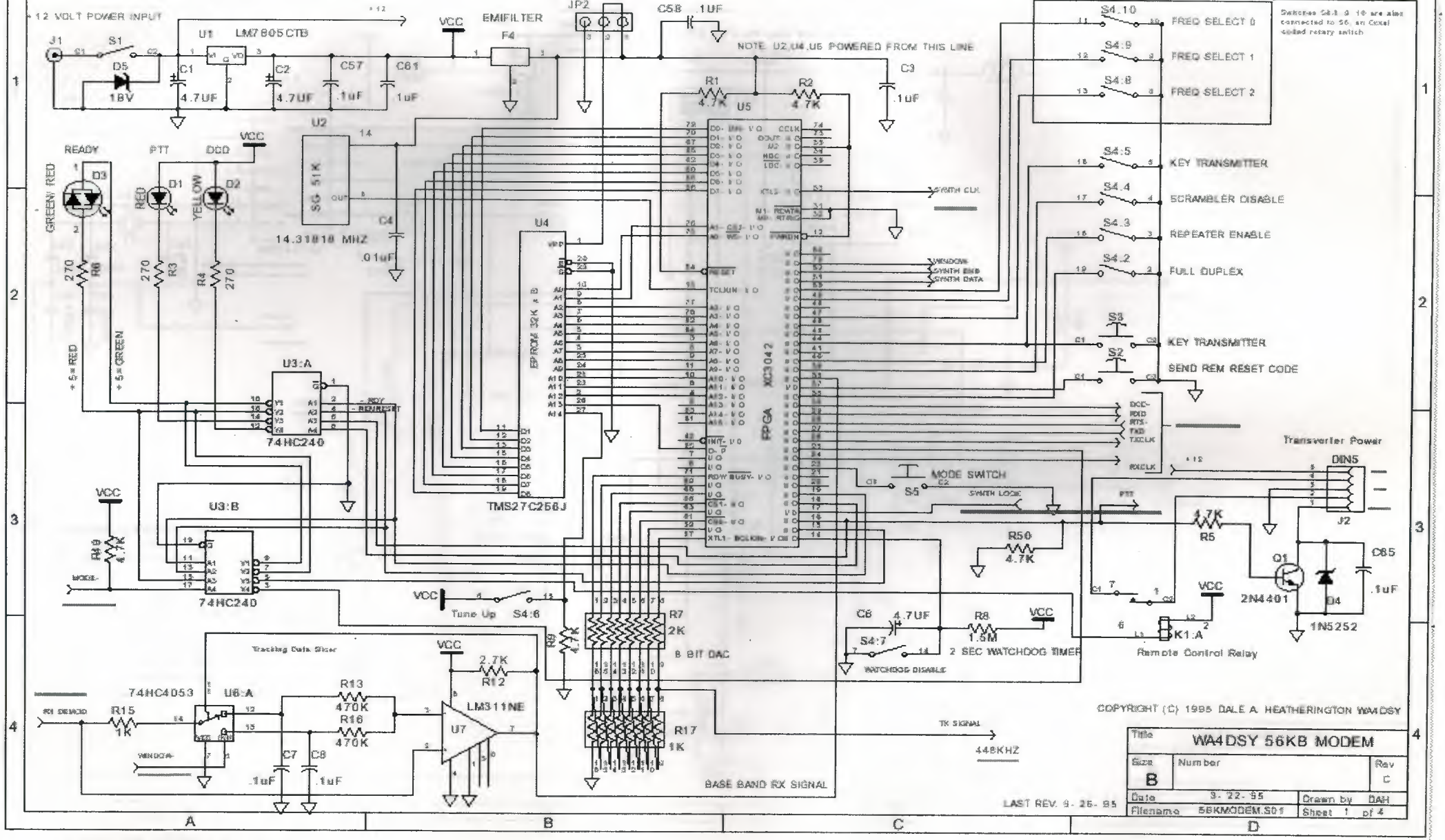
Optical Navigation, http://www.semiconductor.agilent.com

RF Modem Design, http://www.wa4dsy.net

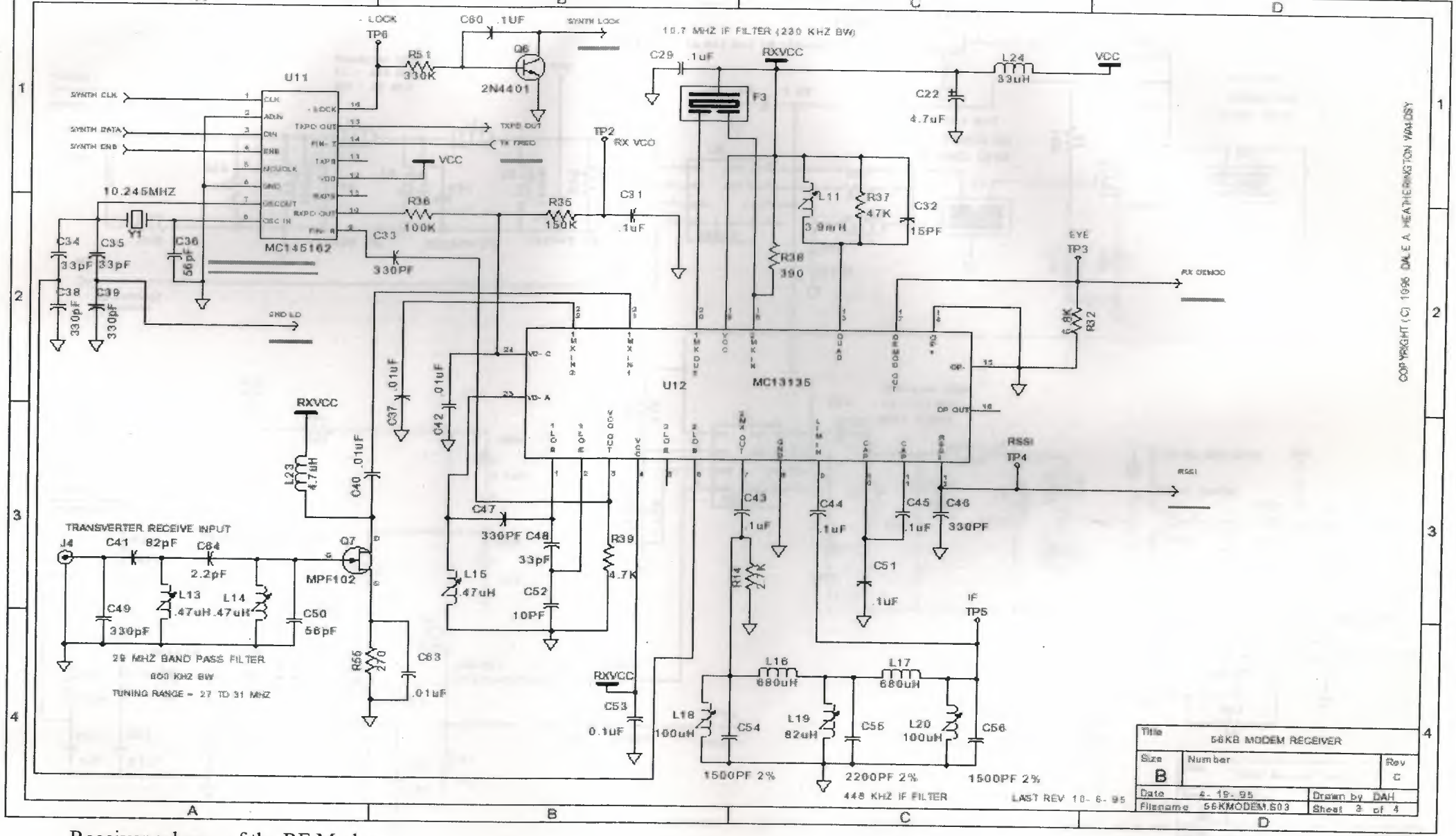
Schematics and PCBs, http://www.rlocman.com.ru/en

Temperature Sensor, http://www.national.com

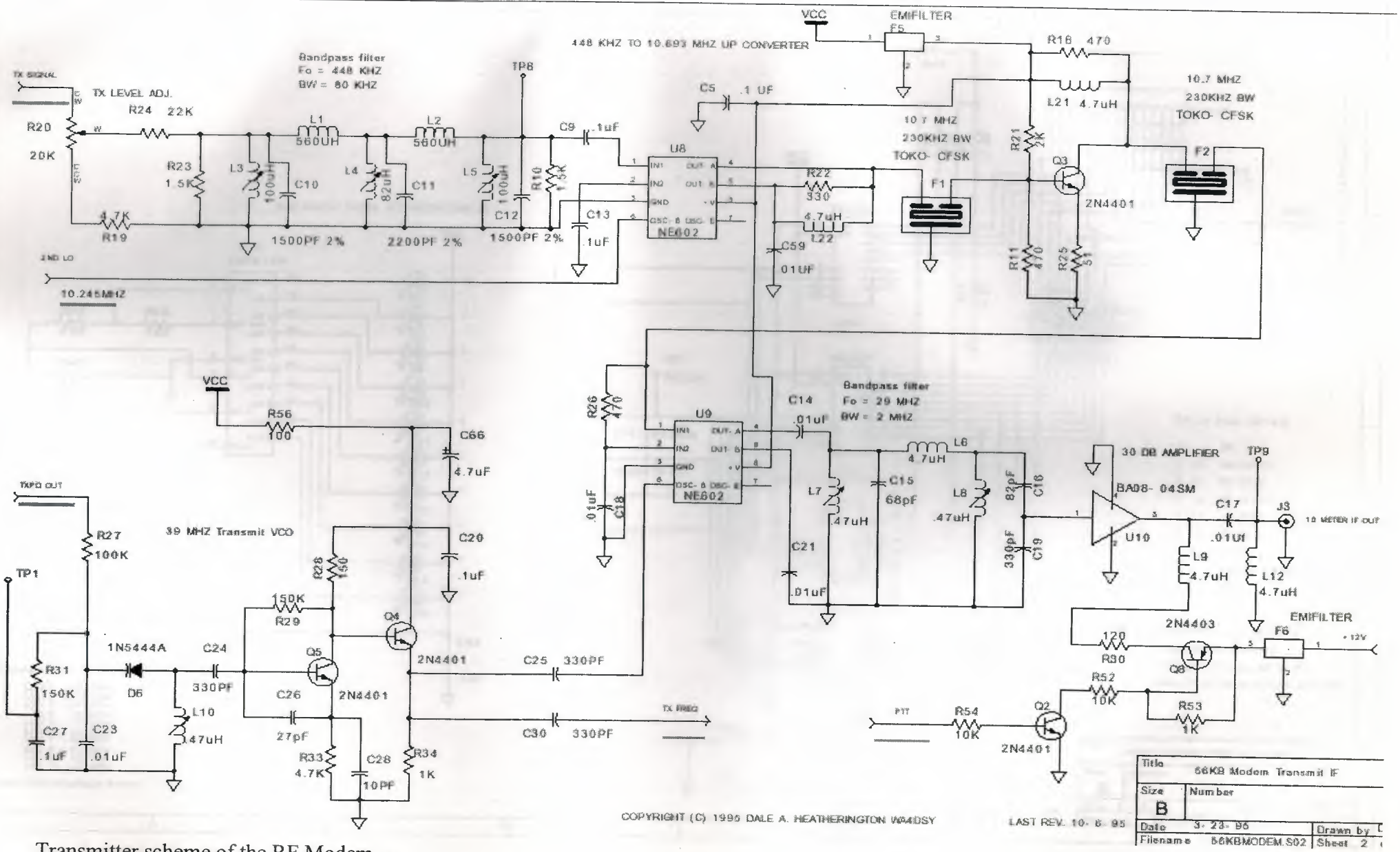
Wireless news group: comp.std.wireless



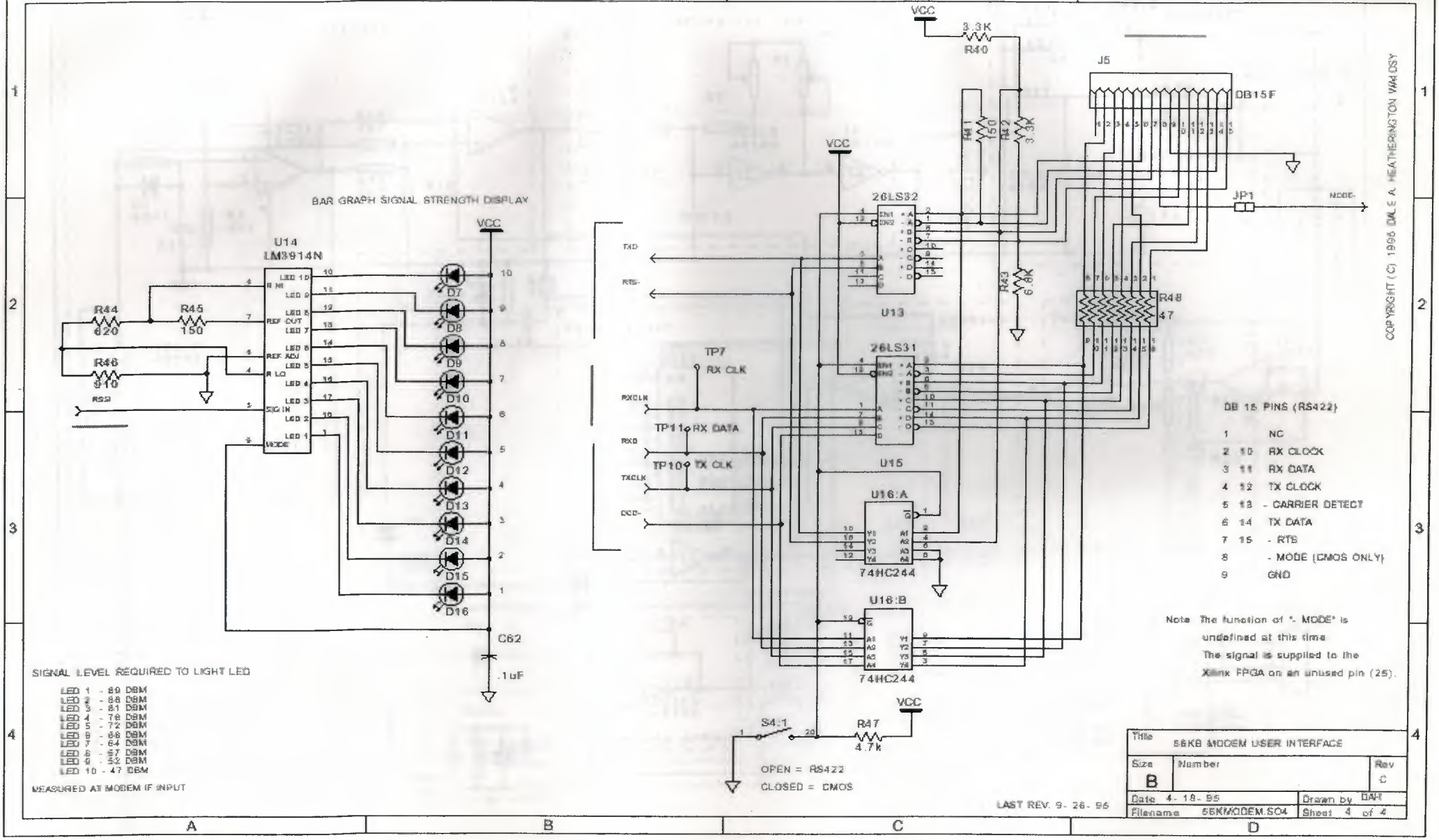
Main scheme of RF Modem



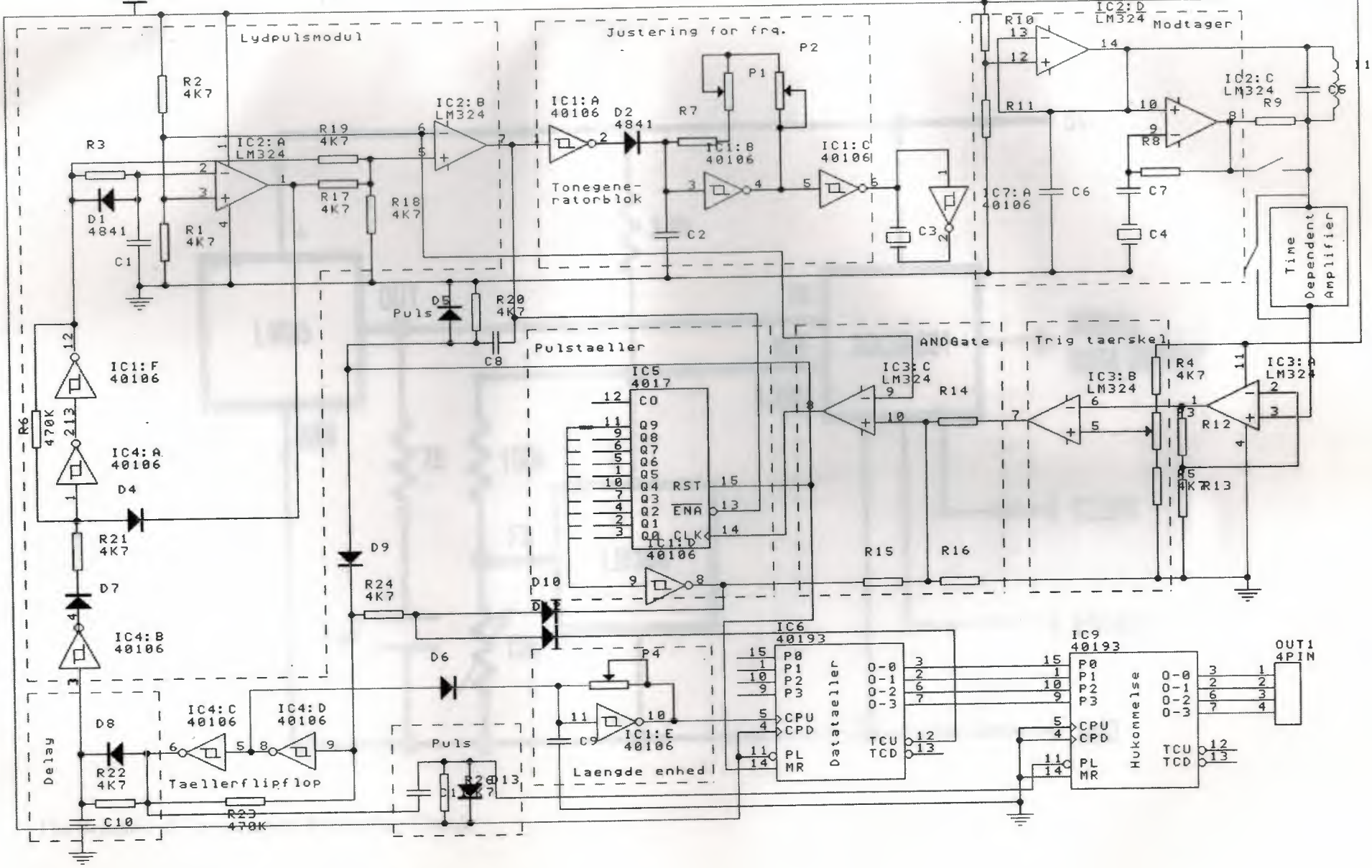
Receiver scheme of the RF Modem



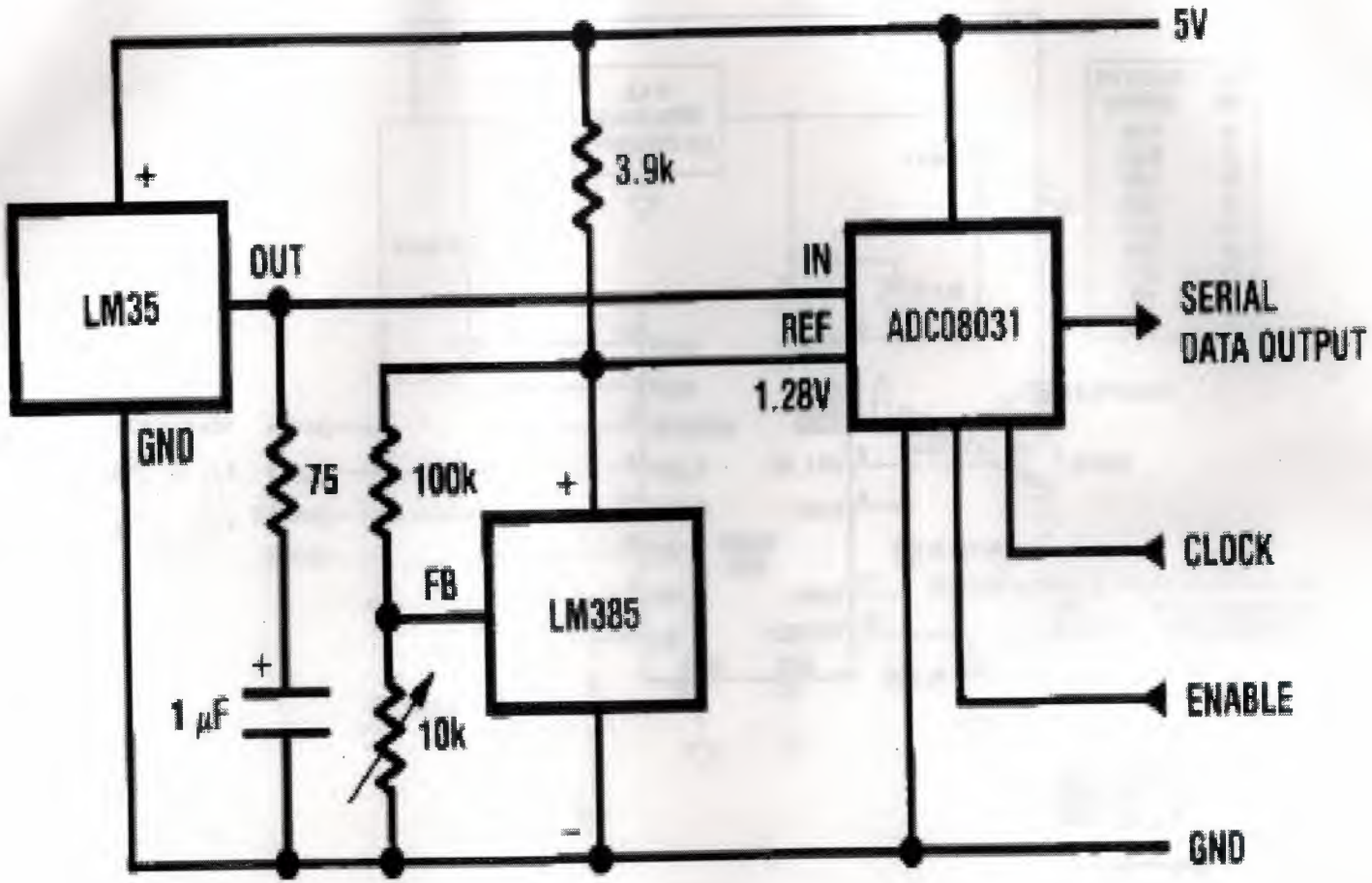
Transmitter scheme of the RF Modem



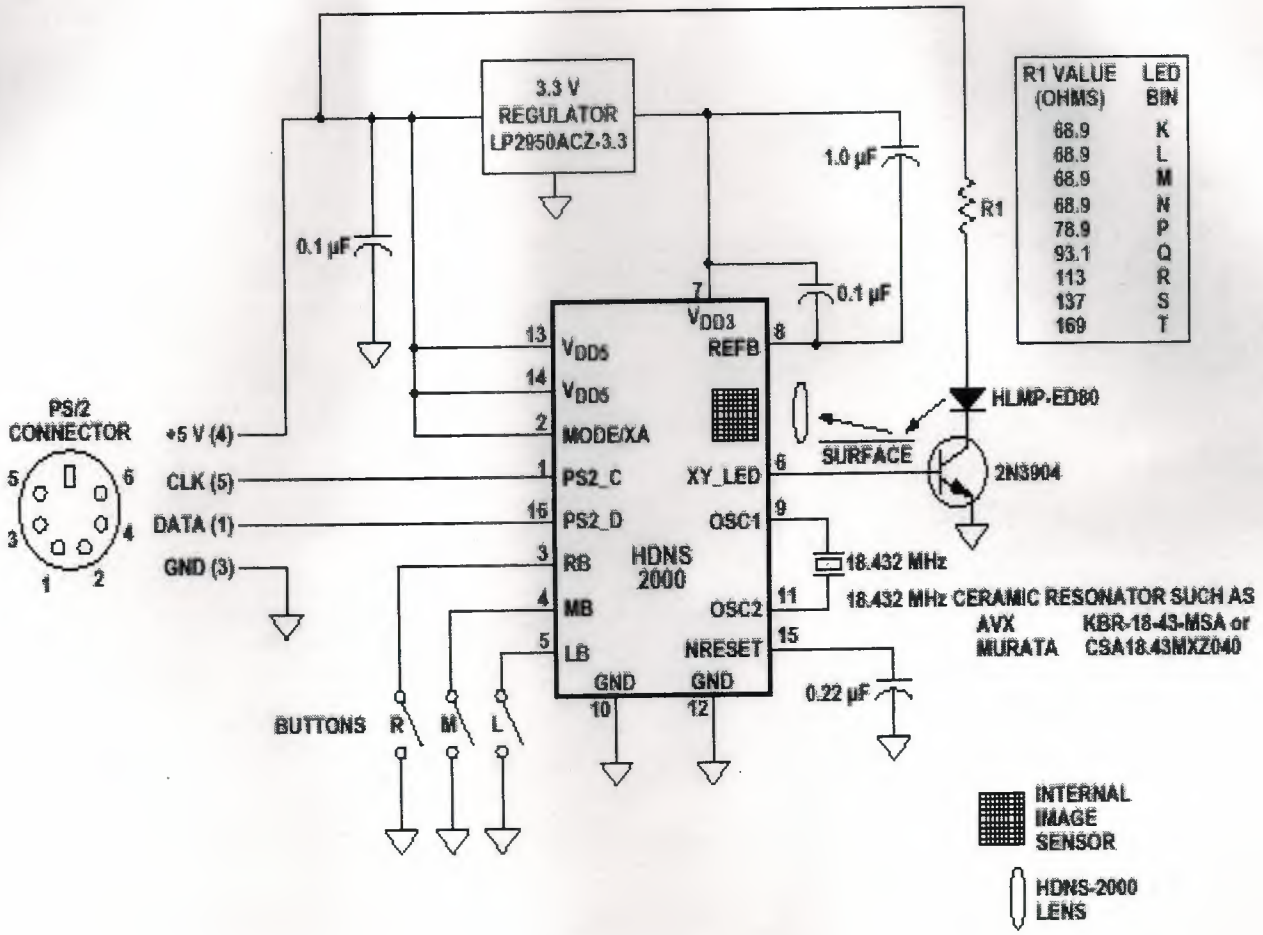
User Interface Control of the RF Modem



The Scheme of Distance Measuring Circuit



The Scheme of Temperature Measuring Circuit



Scheme of Navigation Feedback