Applied Informatics

Editors: Leonids Ribickis, Nadezhda Kunicina

Peteris Apse-Apsitis

Boris Assanovic

Jelena Caiko

Ilja Galkins

Dmitry Kovalenko

Elias Kyriakides

Nadezhda Kunicina

Viktar Liauchuk

Leonids Ribickis

Andrei Varuyeu

Anatolijs Zabasta

Anastasija Zhiravecka

Riga, 2018

This document has been prepared by the financial support of European Union. The authors from Riga Technical University and Belarusian State University are responsible for the content of this document. This publication reflects the views only of the authors, and it cannot be regarded as the European Union's official position.

The book is developed in a frame of the project “ERASMUS+ Capacity-building in the Field of Higher Education 2015 Call for Proposals EAC/A04/2014 561525-EPP-1-2015-1-LV-EPPKA2-CBHE-JP - ERASMUS+ CBHE.

The textbook is devised for students of applied physic and electrical engineering specialties. The textbook can be useful for students and professionals focusing on applied informatics issues. The book gives overview of current computer control of electrical technologies and IT elements, as well as explains their operating principles.

Key Action: KA2 - Cooperation for innovation and the exchange of good practices

Action: Capacity Building in Higher Education

Action Type: Joint Projects

Deliverable: 2.3. Five electronic courses e-Books by the declared directions

Co-funded by the
Erasmus+ Programme
of the European Union

Project Scientific Manager: Nadezhda Kunicina

Project Coordinator: Anatolijs Zabasta

Editors: Leonids Ribickis, Nadezhda Kunicina

Institution: Riga Technical University

ISBN - 978-9934-22-144-6 (pdf)

Contributors

Peteris Apse-Apsitis, associate professor, senior researcher. Head of Division of Industrial Electronic equipment of Institute of Industrial Electronics and Electrical Engineering, Riga Technical University, Faculty of Power and Electrical Engineering, 12/1 Azenes Str. - 510., Riga, LV 1048, Latvia, tel. +371 67089917, peteris.apse-apsitis@rtu.lv

Boris Assanovich, associate professor in the Department of Information Systems and Technologies, at YK State University of Grodno, Physics and Technical Faculty Coordinator of International Academic and Project Activity office: YKSUG Laboratory Campus, BLK-5, 107, Grodno 230009, Belarus, tel. +375 152431279, bas@grsu.by

Jelena Caiko, senior researcher of Institute of Industrial Electronics and Electrical Engineering, Riga Technical University, Faculty of Power and Electrical Engineering, 12/1 Azenes Str. - 503., Riga, LV 1048, Latvia, jelena.caiko@rtu.lv

Ilja Galkins, professor, senior researcher of Division of Industrial Electronic equipment, Institute of Industrial Electronics and Electrical Engineering, Riga Technical University, Faculty of Power and Electrical Engineering, 12/1 Azenes Str. - 502., Riga, LV 1048, Latvia, tel. +371 67089918, ilja.galkins@rtu.lv

Dmitry Kovalenko, dean of the Faculty of Physics and Information Technology, Francisk Skorina Gomel State University, 102 Sovetskaya Str., off. 2-9, Gomel, Belarus, 246019, tel. +375 232576557, dkov@gsu.by

Elias Kyriakides, associate professor in the Department of Electrical and Computer Engineering and the Associate Director for Research of the KIOS Research Center for Intelligent Systems and Networks, at the University of Cyprus. Office 1: Green Park 408, KIOS Research Center, Nicosia, Cyprus tel. +357 22892291, elias@ucy.ac.cy

Nadezhda Kunicina, professor, senior researcher of Division of Industrial Electronic equipment, Institute of Industrial Electronics and Electrical Engineering, Riga Technical University, Faculty of Power and Electrical Engineering, 12/1 Azenes Str. - 503., Riga, LV 1048, Latvia, tel. +371 67089052, nadezda.kunicina@rtu.lv

Viktar Liauchuk, associate professor. Head of the Department of Automated systems of information processing, Francisk Skorina Gomel State University, 102 Sovetskaya Str., off. 4-7, Gomel, Belarus, 246019, liauchuk@gmail.com

Leonids Ribickis, professor, rector of Riga Technical University, Head of Institute of Industrial Electronics and Electrical engineering, Faculty of Power and Electrical Engineering, Riga Technical University, Scientific Head of Electromechatronics Scientific laboratory, 1, Kalku Str. -217, Riga, LV 1658, Latvia, tel. +371 67089300, leonids.ribickis@rtu.lv

Andrei Varuyeu, associate professor of the Department of Automated systems of information processing, Francisk Skorina Gomel State University, 102 Sovetskaya Str., off. 4-7, Gomel, Belarus, 246019, +375 232578863, ang@gsu.by

Anatolijs Zabasta, senior researcher of Institute of Industrial Electronics and Electrical Engineering, Riga Technical University, Faculty of Power and Electrical Engineering, 12/1 Azenes Str. - 503., Riga, LV 1048, Latvia, tel. +371 67089051, anatolijs.zabashta@rtu.lv

Anastasija Zhiravecka, professor, senior researcher of Division of Industrial Electronic equipment, Institute of Industrial Electronics and Electrical Engineering, Riga Technical University, Faculty of Power and Electrical Engineering, 12/1 Azenes Str. - 509., Riga, LV 1048, Latvia, tel. +371 67089917, anastasija.zhiravecka@rtu.lv

Content

Introduction

The book “Applied Informatics” provides a technical background for students of applied physic and electrical engineering specialties. The book is addressed for all level of students and developers of electromechanical devices, automotive control schemes, motion control solutions, as well as a high level of automated infrastructure management technologies. The book is helpful to the development of modern intelligent systems and sensors network, using high-performance computing technology for signal processing, as well for the development of data collection devices, for development of models, methods and metrics that enable quantitative evaluation of electric power and telecommunications infrastructure critical impact and interdependences.

Modern industrial production and manufacturing systems have evolved in basically four generations. The first generation that enabled the industrial revolution dates back to around 1850 or so. The use of steam-powered machines enabled mass production of gods such as clothes, cars and many other products in the beginning of the 20th century (Delsing J., 2017). In the second generation, the use of efficient pneumatic systems became a widely adopted solution for mass-production. The combined use of pneumatic valves and sensors enabled automatic production systems to be used in industrial applications. The third generation systems evolved from pneumatic to electrical motors. The use of electricity as the energy source enabled even newer types of automatic control systems to be developed. Sensors and actuators were now connected to new types of monitoring and control systems like Distributed Control Systems, DCS and Supervisory Control and Data Acquisition, SCADA using technologies such as field buses. The hierarchical approach of device-level, DCS, and SCADA (known as ISA-95), soon became the de-facto architectural style for how industrial productions systems were designed and deployed. DCS and SCADA systems soon became networked, which enabled tight integration between control systems and Enterprise Resource Planning Systems (ERP) and Manufacturing execution system (MES). This is today the most widely used approach by the industry and has been so for at least the last 20-30 years. In the 90’ths the current state of the art architecture ISA-95 was established. Seemingly, the size of ISA-95 based automation systems was limited in respect of I/O points. This becomes a technology bottleneck in the view of the upcoming smart cities and smart energy grids.

The book is useful for students, who study industrial production lines and control tools developed in all generations. The book is also exploring the problems associated with modelling of an ongoing urban environment process in order to ensure a high level of automated infrastructure management, intelligent systems and sensors network technology.

The Chapter 1 of this book is devoted to communication networks with the clear focus on computer networks architecture, design, network standards and specific network elements. The goal of this chapter is to present the basics concepts of telecommunication systems, including OSI, network components, quality issues of network design, model with focus on digital and wireless, and the most important features of the propagation of telecommunication signals, as well as computer networks architecture and design, using standard and specific adaptive telecommunication network elements, and its application domains and deployments.

The Chapter 2 of this book is devoted to control theory, including evaluation of regulated system stability, qualitative parameters of system stability. The chapter addresses an issues of the automation control, in particular process of a technical object control without a human involvement. Additionally, the object should be able to perceive the control signals containing the information about the further object condition. In most of the cases these signals are generated on the basis of the information about the current condition of the object. Therefore, the flows of information and their relations are the basis of the automatic control process.

The Chapter 3 of this book is devoted to microcontrollers, in particular it explains, how design embedded systems with microcontrollers, the architectures of MCUs, main parameters, most popular MCUs presented on market, and peripheral devices of MCUs. The chapter describes specific of control systems, which contain actuators, sensors and microcontrollers, included in the devices. Case studies describe design and application of such embedded control systems.

The Chapter 4 of this book is devoted to general aspects of electrical engineering and automation, starting with principles, applications, and detailed explanation of main features of the physical background and elements of automation, like resonance phenomenon in AC electrical circuits, main principles of single-phase transformer, and basic realisation of electrical motor. The chapter describes measurements of active power and energy in AC circuits, measurements of power in three-phase electrical circuits, three-phase electrical motors, the means of automation in electrical systems as well as sensors and microprocessor applied in automation.

Chapter 1: Information and Communication Technologies

1.1. Communication network design and operational characteristics

1.1.1. Introduction

The Chapter 1 of this book is devoted to communication networks with the clear focus on computer networks architecture, design, network standards and specific network elements.

Before we can understand how to design a computer network, we should first agree on exactly what a computer network is. At one time, the term network meant the set of serial lines used to attach dumb terminals to mainframe computers. Other important networks include the voice telephone network and the cable TV network used to disseminate video signals. The main things these networks have in common are that they are specialized to handle one particular kind of data.

What distinguishes a computer network from these other types of networks? Probably the most important characteristic of a computer network is its generality. Computer networks are built primarily from general-purpose programmable hardware, and they are not optimized for a particular application like making phone calls or delivering television signals. Instead, they are able to carry many different types of data, and they support a wide, and ever growing, range of applications. Today's computer networks are increasingly taking over the functions previously performed by single-use networks. This chapter looks at some typical applications of computer networks and discusses the requirements that a network designer who wishes to support such applications must be aware of (Computer Networks 2012).

1.1.2. Telecommunications basics

The goal of this chapter is to present the basics concepts of telecommunication systems with focus on digital and wireless, and the most important features of the propagation of telecommunication signals.

Basic concepts include:

- Signal: Analog, Digital, Random
- Sampling
- Bandwidth
- Spectrum
- Noise
- Interference
- Channel Capacity
- BER
- Modulation

The purpose of any telecommunications system is to transfer information from the sender to the receiver by a means of a communication channel.

Let us talk a bit about what a signal actually is electronic signals specifically (as opposed to traffic signals, albums by the ultimate power-trio, or a general means for communication). The signals we are talking about are time-varying "quantities" which convey some sort of information. In electrical engineering the quantity that's time-varying is usually voltage (if not that, then usually current). So, when we talk about signals, just think of them as a voltage that is changing over time.

Signals are passed between devices in order to send and receive information, which might be video, audio, or some sort of encoded data. Usually the signals are transmitted through wires, but they could also pass through the air via radio frequency (RF) waves. Audio signals, for example might be transferred between your computer's audio card and speakers, while data signals might be passed through the air between a tablet and a Wi-Fi router (PSCES, 2012).

Analog signal

Because a signal varies over time, it is helpful to plot it on a graph where time is plotted on the horizontal, x-axis, and voltage on the vertical, y-axis. Looking at a graph of a signal is usually the easiest way to identify if it is analog or digital; a time-versus-voltage graph of an analog signal should be smooth and continuous**.**

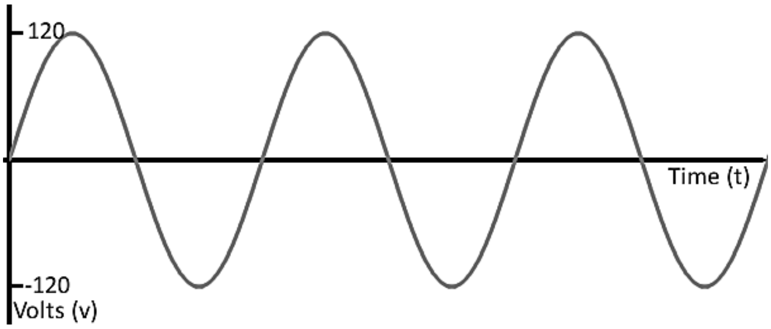


Figure 1.1. An analogue signal graph

While these signals may be limited to a range of maximum and minimum values, there are still an infinite number of possible values within that range.

For example, the analogue voltage coming out of your wall socket might be clamped between -120V and +120V, but, as you increase the resolution more and more, you discover an infinite number of values that the signal can actually be (like 64.4V, 64.42V, 64.424V, and infinite, increasingly precise values).

For analogue signals, these variations are directly proportional to some physical variable like sound, light, temperature, wind speed, etc.

Pure audio signals are also analogue. The signal that comes out of a microphone is full of analogue frequencies and harmonics, which combine to make beautiful music.

Example Analog Signals

Video and audio transmissions are often transferred or recorded using analog signals. The composite video coming out of an old RCA jack, for example, is a coded analog signal usually ranging between 0 and 1.073V. Tiny changes in the signal have a huge effect on the color or location of the video.

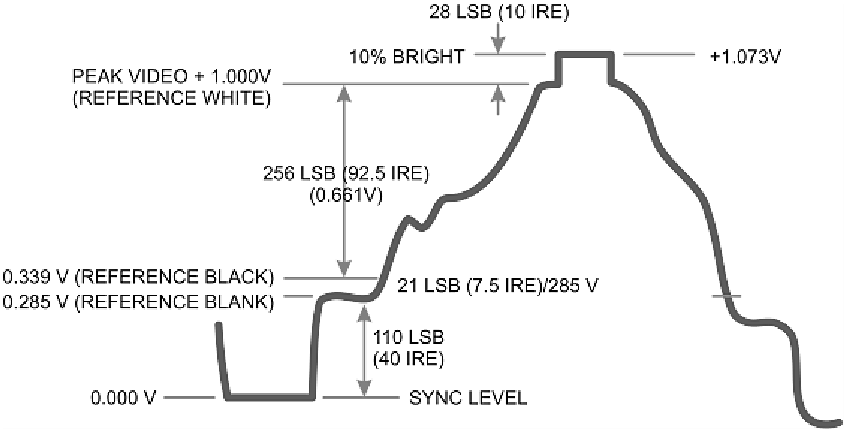


Figure 1.2. An analogue signal representing one line of composite video data

Characteristics of Analog Signal

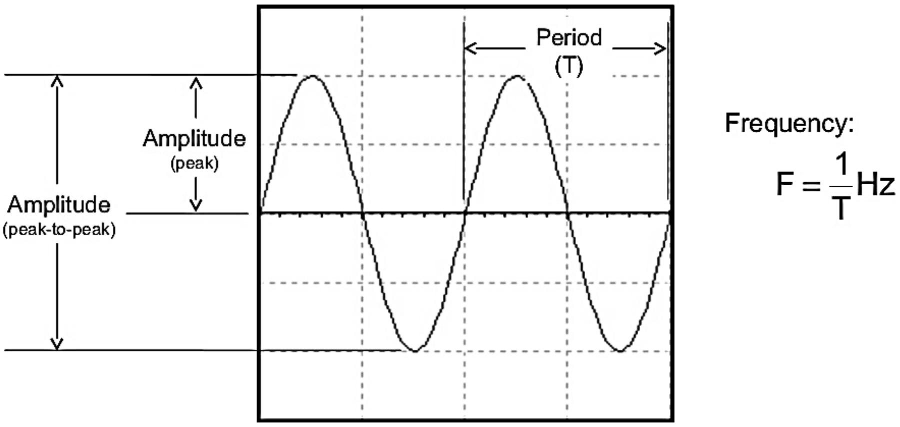


Figure 1.3. Characteristics of analogue signal

Information which is analog in its native form (audio and image) can vary continuously in terms of intensity (volume or brightness) and frequency (tone or color). Those variations in the native information stream are translated in an analog electrical network into variations in the amplitude and frequency of the carrier signal. In other words, the carrier signal is modulated (varied) in order to create an analog of the original information stream.

The electromagnetic sinusoidal (waveform) or sine wave can be varied in amplitude at a fixed frequency, using Amplitude Modulation (AM). Alternatively, the frequency of the sine wave can be varied at constant amplitude using Frequency Modulation (FM). Additionally, both frequency and amplitude can be modulated simultaneously.

- Analog signal can have infinite number of values and varies continuously with time.
- Analog signal is usually represented by sine wave.
- As shown in figure each cycle consists of a single arc above the time axis followed by a single arc below the time axis.
- Example of analogue signal is human voice. When we speak, we use air to transmit an analogue signal. Electrical signal from an audio tape, can also be in analogue form.

Amplitude

- Amplitude of a signal refers to the height of the signal.
- It is equal to the vertical distance from a given point on the waveform to the horizontal axis.
- The maximum amplitude of a sine wave is equal to the highest value it reaches on the vertical axis as shown in figure.
- Amplitude is measured in volts, amperes or watts depending on the type of signal. A volt is used for voltage, ampere for current and watts for power.

Period

- Period refers to the amount of time in which a signal completes one cycle.
- It is measured in seconds.
- Other units used to measure period are millisecond (10-3 sec.) microsecond (10-6 sec), nanosecond (10-9 sec) and picoseconds (10-12 sec).

Frequency

- It refers to the number of wave patterns completed in a given period of time.
- To be more precise, frequency refers to number of periods in one second or number of cycles per second.
- Frequency is measured in Hertz (Hz)
- Other units used to express frequency are kilohertz (103 Hz) Megahertz (106 Hz), gigahertz (109 Hz) and terahertz (1012 Hz).
- Frequency and period are the inverse of each other. Period is the inverse of frequency and frequency is the inverse of period.

Example: The power we use at home has a frequency of 60 Hz. The period of this sine wave can be determined as follows:

Solution: $T = \frac{1}{f} = \frac{1}{60} = 0.0166s = 0.0166 \times 10^3 ms = 16.6\, ms$

Example: The period of a signal is 100 ms. What is its frequency in kilohertz?

Solution: First we change 100 ms to seconds, and then we calculate the frequency from the period (1 Hz = 10^{-3} kHz).

$$100\,ms = 100 \times 10^{-3} s = 10^{-1} s$$

$$f = \frac{1}{T} = \frac{1}{10^{-1}} Hz = 10Hz = 10 \times 10^{-3} kHz = 10^{-2} kHz$$

Example: A sine wave is offset 1/6 cycle with respect to time 0. What is its phase in degrees and radians?

Solution: We know that 1 complete cycle is 360°. Therefore, 1/6 cycle is

$$\frac{1}{6} \times 360 = 60° = 60 \times \frac{2\pi}{360} rad = \frac{\pi}{3} rad = 1.046\,rad$$

Phase

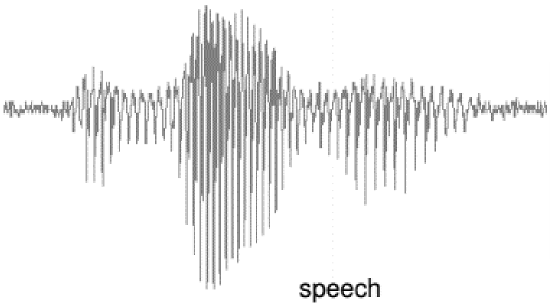
- Phase describes the position of the waveform relative to time zero.
- Phase describes the amount by which the waveform shifts forward or backward along the time axis.
- It indicates the status of first cycle.
- Phase is measured in degrees or radians.

A phase shift of 3600 indicates a shift of a complete period, a phase shift of 180° indicates a shift of half period and a phase shift of 90° indicates a shift of a quarter of a period as shown in fig. below.

Random signal

Random signal are the ones that are unpredictable and can be described only by statistical means. Many signals processed by computers can be considered as random.

Examples of random signal: speech, audio, video, digital communication, medical, biological, and economic signals.



speech

ECG

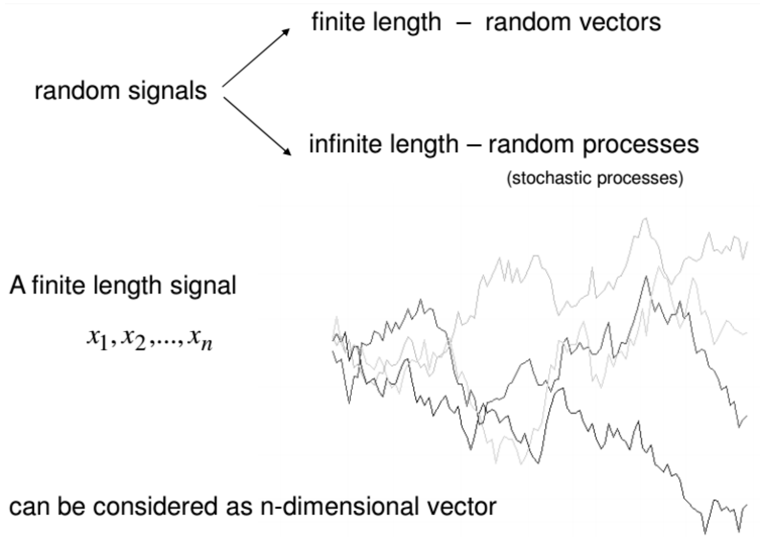

finite length – random vectors
random signals
infinite length – random processes
(stochastic processes)
A finite length signal
$x_1, x_2, \ldots, x_n$
can be considered as n-dimensional vector

Figure 1.4 Examples of random signals

Sampling

The information can also be transmitted by digital binary signals that will have only two values, a digital one and a digital zero.

Any analogue signal can be converted into a digital signal by appropriately sampling and then coding it.

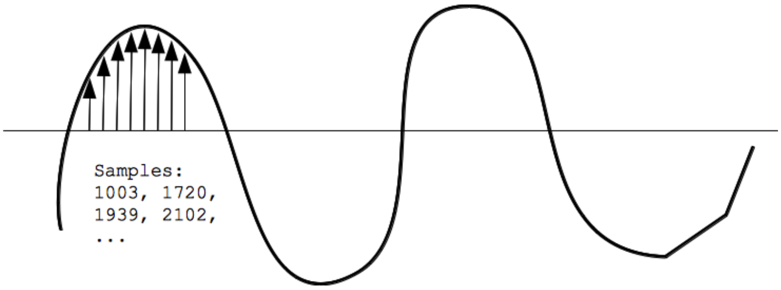


Figure 1.5. Sampling

The way an analog signal is brought into the computer is "digitization". The computer "samples" the signal very rapidly over time, noting the value (the height basically) of the curve each time, and recording that value as a number. CD quality audio samples the sound signal 44000 times per second, noting the "height" of the sound signal once for sample. Each sample is a whole number in the range -32768 .. 32767 (this is the range of number that can be stored in 2 bytes, 256 squared).

Audio CD digitization is very good, but not perfect. The signal might have a little wiggle that happens so fast, the 44000 samples/second are not fast enough to quite capture the wiggle perfectly. In reality, 44000 samples per second captures almost everything the human ear can hear anyway. There can also be a sort of rounding error -- the signal value might be in between 1452 and 1453, and the digitization has to pick just one value to represent it. As a practical matter, the 44000 samples per second is extremely good at capturing the level of detail that humans can hear.

So in essence, digitization translates a sound signal into just a series of numbers: 12000, 12002, 12006, 12007, 12010, 12005, 12006, ... and so on. Playing back the digital sound is just the reverse: a chip takes in the stream of numbers, say representing 44000 samples/second, and constructs an electrical signal that matches those numbers over time. In effect, this reconstructs the original sound signal from the numbers.

Example:

Sound needs to be converted into binary for computers to be able to process it. To do this, sound is captured - usually by a microphone - and then converted into a digital signal.

An analogue to digital converter will sample a sound wave at regular time intervals. For example, a sound wave like this can be sampled at each time sample point:

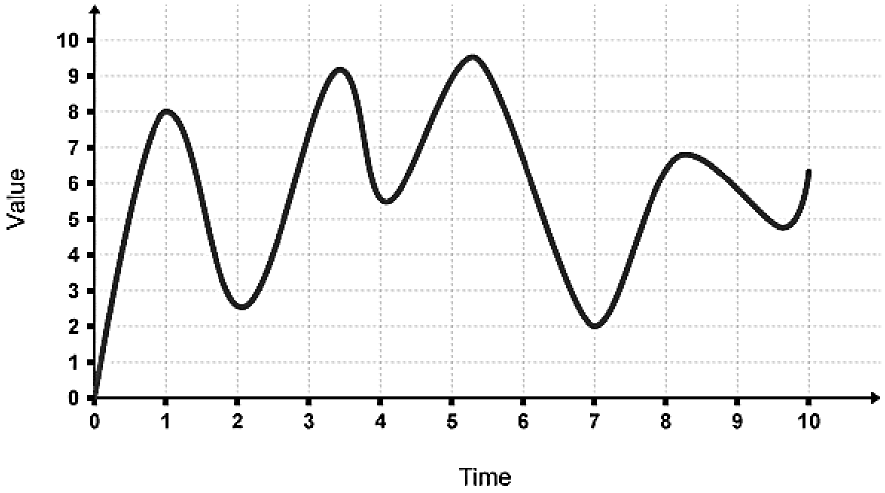


Figure 1.6. Sound representation

The samples can then be converted to binary. They will be recorded to the nearest whole number.

Time sample	1	2	3	4	5	6	7	8	9	10
Denary	8	3	7	6	9	7	2	6	6	6
Binary	1000	0011	0111	0110	1001	0111	0010	0100	0110	0110

If the time samples are then plotted back onto the same graph, it can be seen that the sound wave now looks different. This is because sampling does not take into account what the sound wave is doing in between each time sample.

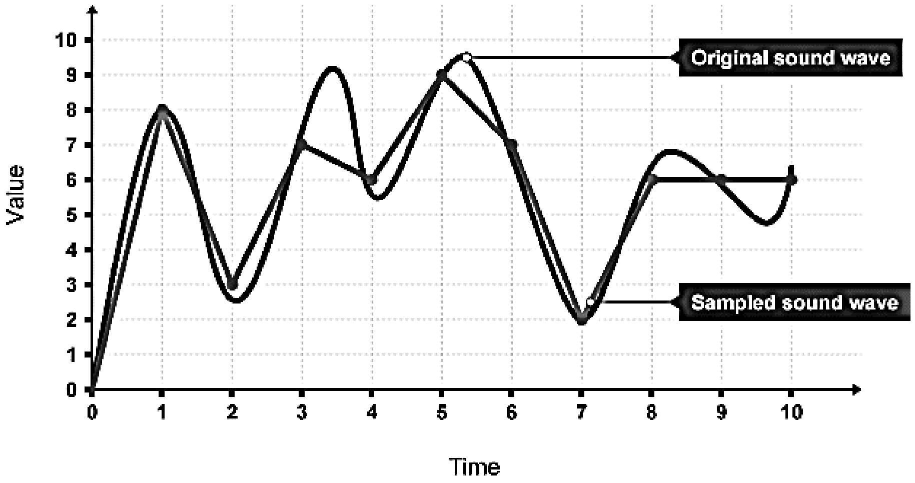


Figure 1.7. Sampled signal of sound

This means that the sound loses quality, as data has been lost between the time samples. The way to increase the quality and store the sound at a quality closer to the original is to have more time samples that are closer together. This way, more detail about the sound can be collected, so when it is converted to digital and back to analogue again it does not lose as much quality.

The frequency at which samples are taken is called the sample rate, and is measured in Hertz (Hz). 1 Hz is one sample per second. Most CD-quality audio is sampled at 44 100 or 48 000 KHz.

The sampling frequency must be at least twice the maximum frequency present in the signal in order to carry all the information contained therein.

The minimum sampling rate required in order to accurately reconstruct the analogue input is given by the Nyquist sampling rate, f_N, given by the formula:

$$f_N = 2f_{max}, \tag{1.1}$$

where f_{max} is the highest frequency component of the analogue input.

The Nyquist sampling rate (frequency) is the minimum. In practice, sampling rates are much higher, typically 2 times the Nyquist rate f_N. In other words, 4 times f_{max}.

Therefore, we say that the sampling frequency must be *more than* twice the value of the highest frequency component of the signal:

$$f_s \geq f_N \text{ , where } f_N \text{ is } 2f_{max} \tag{1.2}$$

If this rule is violated, a problem called aliasing results.

Quantizing

Quantization is the process of mapping the sampled analog voltages to discrete, binary levels. The number of bits determine the number of levels (1 bit is 2 levels, 2 bits is 4 levels, 2^n=#levels).

We determine the amplitudes associated with each sample point of an analog signal. We can then convert these amplitudes (real numbers) into binary numbers associated with their levels.

Quantizer is used that characterizes the length of the binary words they produce. An *N*-bit quantizer has 2^N levels and outputs binary numbers of length *N*.

So, you might see:

Telephones use 8-bit encoding → $2^8 = 256$ levels

CD audio uses 16-bit encoding → $2^{16} = 65{,}536$ levels.

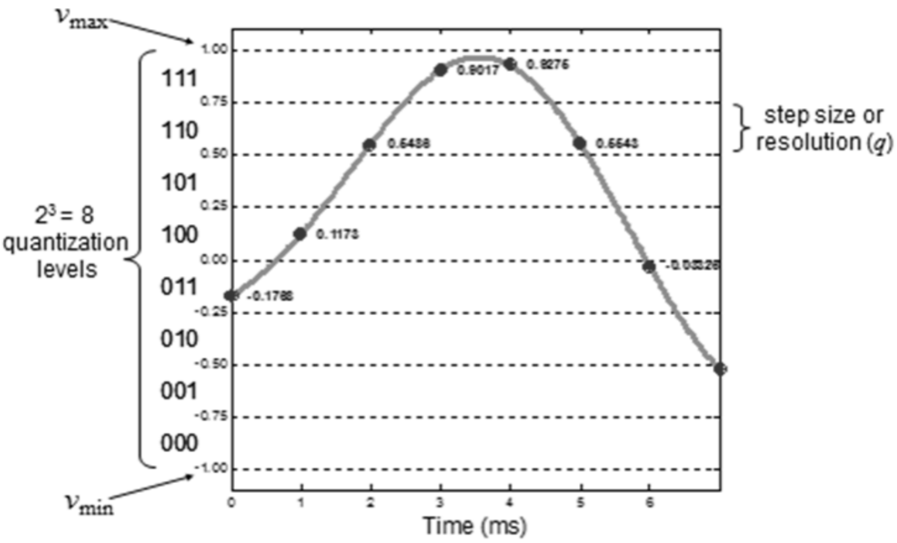


Figure 1.8. Quantized signal

Quantizers are limited to a specific voltage range. For the example above, we will assume that our analog input falls within a range of -1.0 to +1.0 volts.

The quantizer will partition this range into 2^N steps of size q , the quantizer step size, given by

$$q = \frac{v_{\max} - v_{\min}}{2^N} \quad . \tag{1.3}$$

So for the above example, step size will be

$$q = \frac{1-(-1)}{8} = .25V \, . \tag{1.4}$$

q, the step size, is formally called the resolution.

Each of the individual steps, called quantization intervals, is assigned a binary value from 0 to $2^N - 1$. So when you look at the above example, you can see our quantizer uses 3-bit encoding. This means the intervals will range from 0 to 7. So, 000 is assigned to the voltages from -0.75 to -1.0, 001 is assigned to the voltages from   -0.5 to -0.74999…, and so on.

If sampled point falls within that interval (or bin), it is assigned that binary value. For the first sample point at time 0, the voltage is -0.1768, which means that sample is assigned a binary value of 011.

The binary representation of the above signal is: 011 100 110 111 111 110 011 001

Digital signals

Digital signals are more robust and easier to store and transport, that is why nowadays digital signals prevail.

Digital signals must have a finite set of possible values. The number of values in the set can be anywhere between two and a-very-large-number-that's-not-infinity. Most commonly digital signals will be one of two values – like either 0V or 5V. Timing graphs of these signals look like square waves.

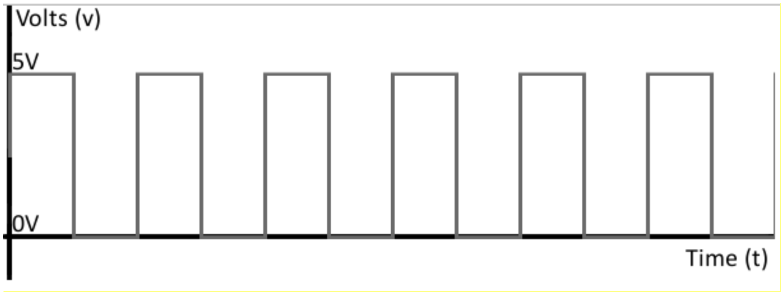


Figure 1.9. Timing graphs of a digital signal

A digital signal might be a discrete representation of an analogue waveform. Viewed from afar, the wave function below may seem smooth and analogue, but when you look closely there are tiny discrete **steps** as the signal tries to approximate values:

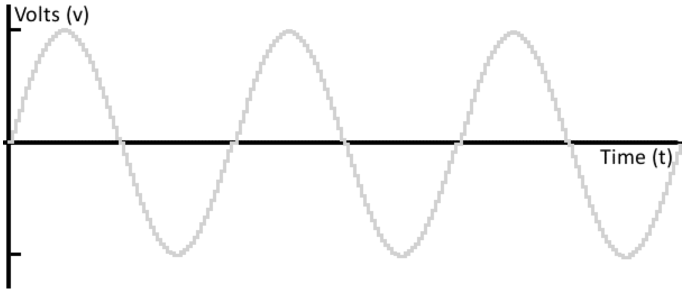
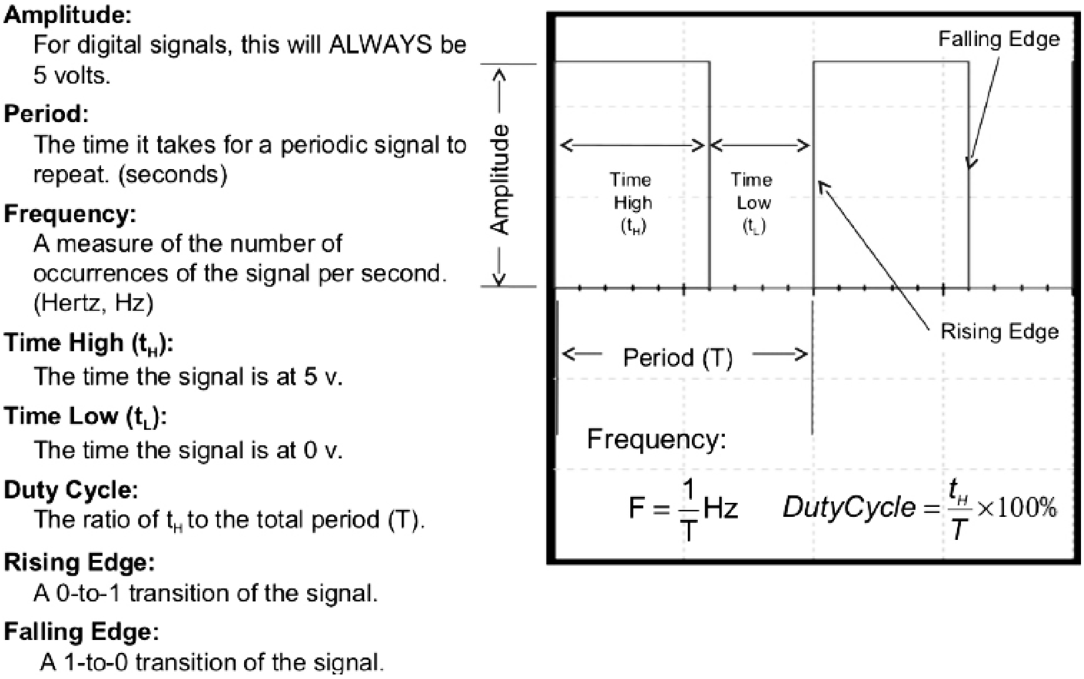


Figure 1.10. Digital signal as a discrete representation of an analogue waveform

That is the big difference between analog and digital waves. Analog waves are smooth and continuous, digital waves are stepping, square, and discrete.

Characteristics of a digital signal

Scientists and engineers often use a digitizer to capture analog data in the real world and convert it into digital signals for analysis. A digitizer is any device used to convert analog signals into digital signals. One of the most common digitizers is a cell phone, which converts a voice, an analog signal, into a digital signal to send to another phone. However, in test and measurement applications, a digitizer most often refers to an oscilloscope or a digital multimeter (DMM). This article focuses on oscilloscopes, but most topics are also applicable to other digitizers.



Amplitude:
For digital signals, this will ALWAYS be 5 volts.

Period:
The time it takes for a periodic signal to repeat. (seconds)

Frequency:
A measure of the number of occurrences of the signal per second. (Hertz, Hz)

Time High (t_H):
The time the signal is at 5 v.

Time Low (t_L):
The time the signal is at 0 v.

Duty Cycle:
The ratio of t_H to the total period (T).

Rising Edge:
A 0-to-1 transition of the signal.

Falling Edge:
A 1-to-0 transition of the signal.

$$F = \frac{1}{T} \text{Hz} \qquad DutyCycle = \frac{t_H}{T} \times 100\%$$

Figure 1.11. Characteristics of a digital signal

Regardless of the type, the digitizer is vital for the system to accurately reconstruct a waveform. To ensure you select the correct oscilloscope for your application, consider the bandwidth, sampling rate, and resolution of the oscilloscope.

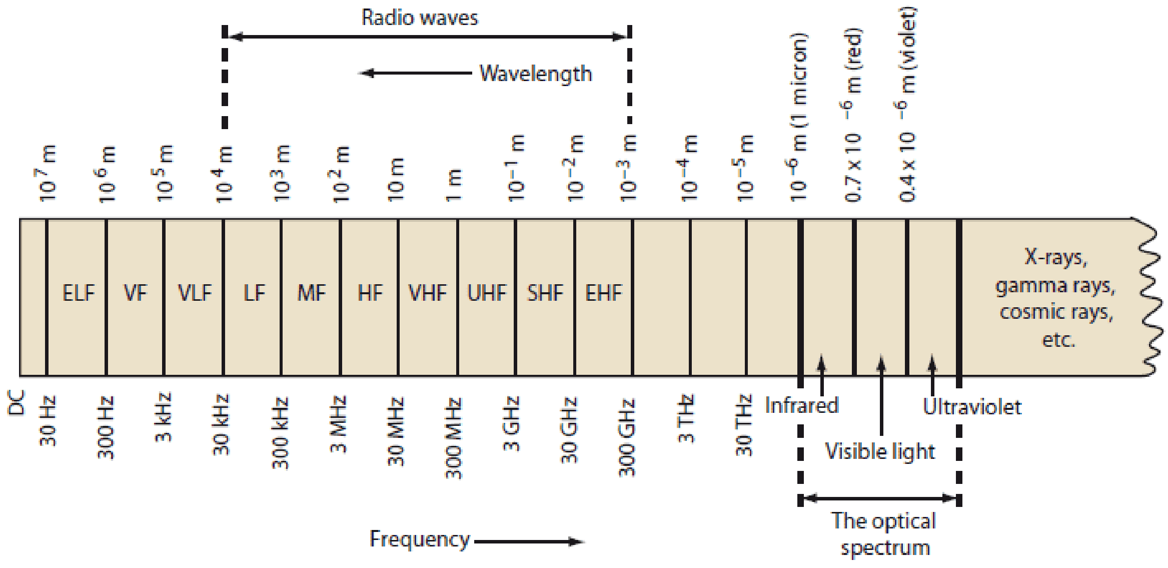
Electromagnetic Spectrum

A signal can be characterised by its behaviour over time or by its frequency components, which constitute its spectrum. Figure 1.11 gives you the big picture of the total spectrum from dc to light and beyond. Note both the frequency and wavelength measures. In the wireless world, we mainly use frequency. However, in the upper reaches of the spectrum, we use wavelength in meters as spectrum bands.

Figure 1.12 represents the entire electromagnetic spectrum. It goes all the way from very low frequency radio waves on the left, to very high frequency X-rays and gamma rays on the right.

In the middle, there is a very small region that represents visible light. In the scope of the entire electromagnetic spectrum, the range of frequencies that we can actually perceive with our eyes is very small. You can see on either side of visible light is infrared and ultraviolet.

But the area that we are interested in is the very narrow range of frequencies used by WiFi equipment. That is the very thin sliver at the low end of the microwave range.



1. The electromagnetic frequency spectrum ranges from dc to light. The lower radio frequencies are designated mainly by frequency. The optical ranges are referred to by wavelength.

Figure 1.12. Electromagnetic Spectrum

Electromagnetic (EM) communications with submarines are forced to use ***very low EM*** frequencies, because of the difficulties of propagation of higher frequency RF signals under water. Most of the other usages are concentrated on higher frequencies, because of the wider capacity available there (more channels and more data per channel). Examples are:

Shortwaves (international AM broadcast, maritime communications, radio amateurs HF bands, etc.): from 1 to 30 MHz

FM radio: from 88 to 108 MHz

TV broadcast: VHF channels in many bands from 40 to 250 MHz; UHF channels in many bands from 470 to 885 MHz (depending from the country)

VHF and UHF radio ham bands: around 140-150 and 440-450 MHz, together with many other users (services, security, police, etc...)

Mobile phones: 850, 900, 1800, 1900 and 2100 MHz for GSM and CDMA cellular networks;

GPS: 1227 and 1575 MHz

Wi-Fi: 2400-2485 MHz and 4915-5825 MHz (depending from the country). See http://en.wikipedia.org/wiki/List_of_WLAN_channels for details.

Radars: common bands for radars are: L band (1–2 GHz), S band (2–4 GHz), C band (4–8 GHz), X band (8–12 GHz) but others are also used.

Satellite TV: C-band (4–8 GHz) and Ku-band (12–18 GHz)

Microwave telecom links: for example in the United States, the band 38.6 - 40.0 GHz is used for licensed high-speed microwave data links, and the 60 GHz band can be used for unlicensed short range data links with data throughputs up to 2.5 Gbit/s. The 71-76, 81-86 and 92–95 GHz bands are also used for point-to-point high-bandwidth communication links.

The basic communication system is formed by a transmitter TX, a communication channel and a receiver RX.

The Transmitter injects a signal into the channel that delivers it to the receiver. The receiver must recover the information contained in the receiver signal despite the limitations introduced by the channel.

The channel can be a physical one, like a copper cable and an optical fibre, or simply air or even vacuum that transmits electromagnetic waves. Any channel is subject to some kind of electromagnetic "noise" and interference, will attenuate the signal and will change its shape (distortion).

Since it takes some time for the signal to traverse the channel, the received signal will have some latency with respect to the transmitted signal. This "latency" might change over time and contribute to "jitter" in the received signal.

The signal might also reach the receiver by means of different trajectories, and in this case, the different received versions will interact as a consequence of the "multipath". Multipath can completely obliterate a signal but it can also be used advantageously in some modern communications techniques. Although the effect of attenuation can easily be overcome with an amplifier, the amplifier will also enhance any noise introduced by the channel and inevitably introduce some extra noise of its own.

The wavelength (sometimes referred to as lambda, λ) is the distance measured from a point on one wave to the equivalent part of the next, for example from the top of one peak to the next. The frequency is the number of whole waves that pass a fixed point in a period of time.

Waves also have a property called amplitude. This is the distance from the centre of the wave to the extreme of one of its peaks, and can be thought of as the "height" of a water wave.

Unlike waves in water, electromagnetic waves require no medium to carry them through space. It may be said that the media that oscillates is the electromagnetic field.

The phase of a wave is the fraction of a cycle that the wave is offset from a reference point. It is a relative measurement that can be express in different ways (radians, cycles, degrees, percentage). Two waves that have the same frequency and different phases have a phase difference, and the waves are said to be out of phase with each other.

Wavelength and Frequency:

$$c = f * \lambda \tag{1.5}$$

c = speed (meters / second)

f = frequency (cycles per second, or Hz)

λ = wavelength (meters)

Example: If a wave on water travels at one meter per second, and it oscillates five times per second, then each wave will be twenty centimetres long:

1 meter/second = 5 cycles/second * λ

λ = 1 / 5 meters

λ = 0.2 meters = 20 cm

A wave has a certain speed, frequency, and wavelength. These are connected by a simple relation: Speed = Frequency * Wavelength

The wavelength (sometimes referred to as lambda, λ) is the distance measured from a point on one wave to the equivalent part of the next, for example from the top of one peak to the next. The frequency is the number of whole waves that pass a fixed point in a period of time. Speed is measured in meters/second, frequency is measured in cycles per second (or Hertz, abbreviated Hz), and wavelength is measured in meters.

Bandwidth

The bandwidth of a composite signal is the difference between the highest and the lowest frequencies contained in that signal.

Typically, the bandwidth of a filter is specified in terms of 3db (half-power) bandwidth -for a given transfer function, bandwidth spans the frequency range where the magnitude is greater than 3dB down from maximum gain. It can be easily measured by driving the circuit with a sinusoidal source and monitoring output level.

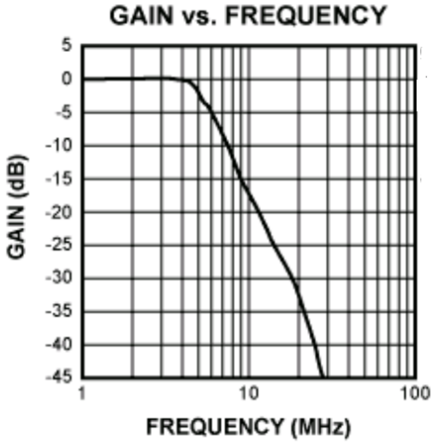


Figure 1.13. Gain versus frequency

Bandwidth is measured between the lower and upper frequency points where the signal amplitude falls to -3 dB below the passband frequency. This sounds complicated, but when you break it down it is actually relatively easy.

First, calculate -3 dB value.

$$-3\ dB = 20\log\frac{V_{out,pp}}{V_{in,pp}} \tag{1.6}$$

$V_{in,pp}$ is the peak-to-peak voltage of the input signal and $V_{out,pp}$ is the peak-to-peak voltage of the output signal. For example, if you input a 1 V sine wave, the output voltage can be calculated as $-3 = 20\log\frac{V_{out,pp}}{1}$ so $V_{out,pp} \approx 0.7\ V$.

Because the input signal is a sine wave, there are two frequencies at which the output signal hits this voltage; these are called the corner frequencies f_1 and f_2. These two frequencies go by many different names such as corner frequency, cut-off frequency, crossover frequency, half-power frequency, 3 dB frequency, and break frequency. However, all these terms refer to the same values. The center frequency, f_0, of the signal is the geometric mean of f_1 and f_2.

$$f_o = \sqrt{f_1 f_2} \tag{1.7}$$

Calculate the bandwidth (BW) by subtracting the two corner frequencies.

$$BW = f_2 - f_1 \tag{1.8}$$

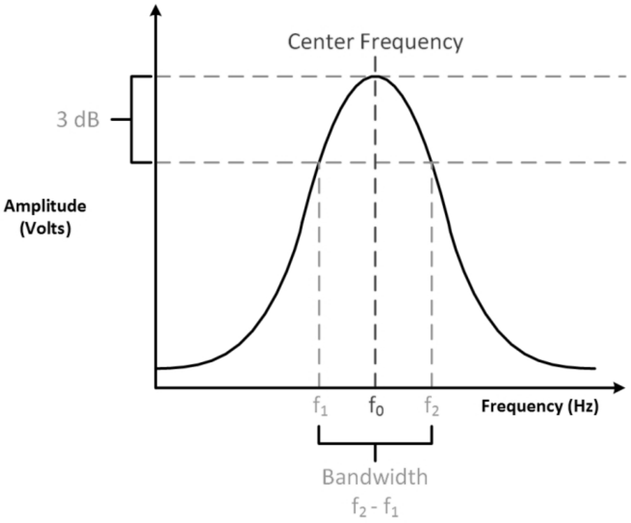
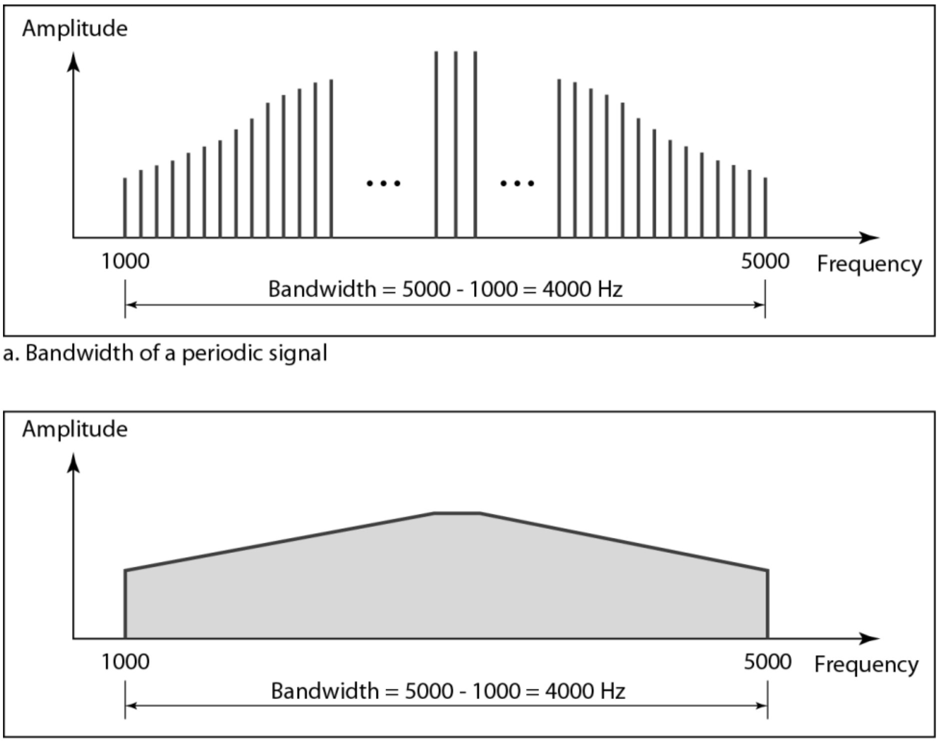


Figure 1.14. The bandwidth, the corner frequency, the center frequency, and the 3 dB point

Example:



a. Bandwidth of a periodic signal

b. Bandwidth of a nonperiodic signal

Figure 1.15. Bandwidth of a periodic and a non-periodic signals

Spectrum

Spectrum is:

- A signal is a function of time which can be represented by a series of sinusoidal functions or sinusoidal components.
- These sinusoidal components have different frequencies, different amplitudes, and different phases.
- Therefore, the plots of frequency versus amplitude and phase for the sinusoidal components which comprise the signal are called the Frequency Spectrum or Spectrum of the signal.

Example: If a periodic signal is decomposed into five sine waves with frequencies of 100, 300, 500, 700, and 900 Hz, what is its bandwidth? Draw the spectrum, assuming all components have a maximum amplitude of 10 V.

Solution: Let f_h be the highest frequency, f_l the lowest frequency, and B the bandwidth. Then $= f_h - f_l = 900 - 100 = 800Hz$. The spectrum has only five spikes, at 100, 300, 500, 700, and 900 Hz.

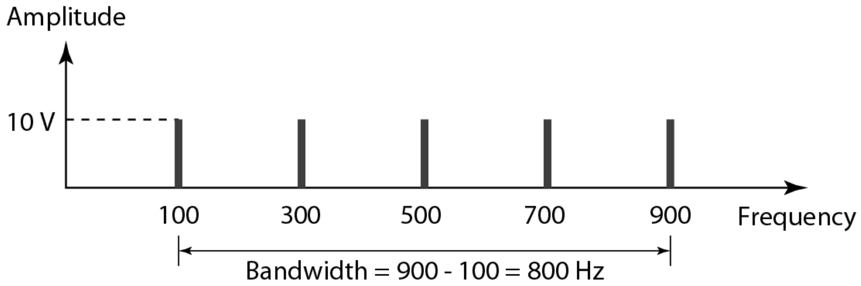


Figure 1.16. The spectrum of a periodic signal

Noise

Noise is a typical random signal, described by its mean power and frequency distribution.

Pure signal, e.g. put into one end of the phone

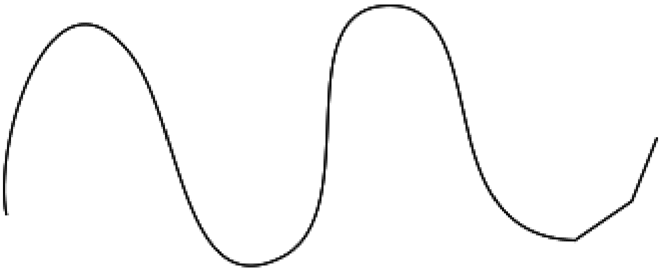


Figure 1.17. Pure signal

Signal + noise as it comes out of the other end of the phone

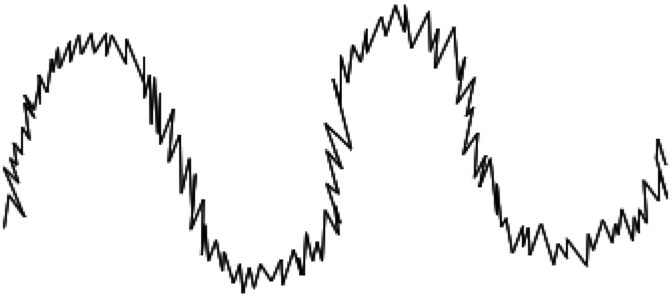


Figure 1.18. Signal + noise

What do "errors" look like in an analog system? The signal you care about is translated from sound to electricity and so on. With each translation step, little errors creep in. The microphone has a little stiffness, the wires don't carry the signal perfectly, and so on. The errors are called "noise" -- you can imagine the pure sound signal you wanted but it's been distorted by little up/down errors -- like fuzzy variations around the true signal.

Interference

Radio-frequency interference (RFI) when in the radio frequency spectrum, is a disturbance generated by an external source that affects an electrical circuit by electromagnetic induction, electrostatic coupling, or conduction.

In unlicensed bands (e.g., 802.11), there are lots of transmitters - 802.11 cards - 802.15.1 (Bluetooth) - 802.15.4 (ZigBee) - 2.4GHz phones - Microwave ovens

This interference can be stronger or weaker than the signal, and can prevent successful reception.

Sources of EMI

EMI sources, both natural and man-made, that compose the EME can be categorized into several primary categories. Some of these classifications of sources are listed below.

(1) Ambient EME that is composed of numerous sources of which the most significant are:

- Television transmissions both analogue and digital.
- Radio AM, FM, and Satellite.
- Solar Magnetic Storms which peak on a eleven year cycle.
- Lightning which occurs as a very high voltage and high current event.
- Utility power grid transmission lines which have high voltage, low current, and low frequency characteristics. In this category is also the new technology of Broadband over Power Lines (BPL) digital signals.
- Other ambient EME sources include airport port radar, telecom transmissions, electrostatic discharge (ESD), and white noise. Also in this category is the earth's magnetic field flux, which has a value of about 500 milligauss.
- Some other major product and system's emissions sources include switching mode power supplies, arc welders, motor bushes, and electrical contacts.

(2) High Powered Electromagnetic Pulse (HEMP) threats which are intended to disable electrical and electronic equipment. These sources are designed to be utilized by terrorist and military organizations. Currently existing HEMP devices include the following:

- Intentional Electromagnetic Interference (IEMI) source – a high powered pulse device utilized by combat, sabotage and terrorist organizations.
- High Altitude Nuclear Electromagnetic Pulse (HNEMP) – produced by the detonation of a nuclear device high above the earth's atmosphere.
- High Powered Microwave Weapon (HPM) – a device utilized by the military as a combat weapon.

- E-Bomb – a HEMP weapon employed by the military to disrupt an enemy’s intra structure that is delivered by an aircraft.
- EMP Cannon – a military tactical weapon.

(3) Power Quality degradation1 factors can effect the operation of equipment that is powered by a mains power source. These mains degradation factors include:

- Voltage surges, sages, dips, spikes, and high and low voltage.
- Brownouts and blackouts.
- Power line faults.
- Electrical Fast Transitions (EFT).
- Electrical noise superimposed on the mains power line.

These power quality degradation factors can occur simultaneously or independently, during any time interval.

(4) Railroad and Mass Transit Systems have some unique types of EMI source problems. These include:

- Propulsion system’s high voltage and high current operational mode emissions.
- Train signalling systems and their associated computer operating codes.
- Third rail shoes arcing broadband emissions.
- High voltage contact switching arcing broadband emissions.
- Train control system’s emissions.
- Track train control circuits.
- Right away emission sources.

(5) Medical equipment utilized in medical facilities has numerous EMI sources. Some of the more prominent of these are listed below:

- Life support equipment such as ventilators, cardiac defibrillators, infusion pumps, etc.
- Patient telemetry and assistance equipment which includes electrocardiographs and motorized wheelchairs.
- Electrical surgical units and their associated support equipment.
- Magnetic Resonance Imagine (MRIs) systems.
- X-ray units, both therapeutic and diagnostic.
- Gamma Beam Electron Accelerators and Therapeutic equipment.

Sources and their most significant effects

(1) Ambient (EME) – Can affect sensitive electronic equipment in the vicinity of the EMI sources. The closer the sensitive electronic equipment is to the EMI source, the higher the source’s radiated power level, and its in-band frequency the greater is the probability that the EMI will cause an interference problem.

In the case of the effects of ESD on sensitive electronic systems it can cause upsets, burn outs, and latch-ups in these units.

(2) High Powered Electromagnetic Pulse effects – High powered electromagnetic sources can totally destroy an electrical and electronic equipment’s function.

As an example, an HNEMP device detonation above the earth’s atmosphere of the United States can totally immobilize the whole of the continental United States’ infrastructure. IEMI, HPM, E-Bombs, and EMP Cannons can be utilized to disable electronic systems at specific locations.

(3) Power Quality distortions and transits that are present on the power main systems can affect the normal operation of the equipment that it supplies power. Transits such as power surges are capable of destroying interface electronic circuits. EFTs can cause electronic circuit upset conditions.

(4) Railroad and Mass Transit Systems have one primary source of EMI and that is the transit and railroad engine’s propulsion systems, which operates with high voltages, currents, and magnetic field levels. They have been known to affect other facilities that contain sensitive electrical equipment that are located near the railroad or mass transit systems right away. These propulsion systems have had EMI associated problems with other elements of their systems. Train control electronics can be affected by EMI sources such as third rail and other broadband frequency arcing sources if they are not adequate designed for EMC.

(5) Medical equipment and facilities sources include patient monitoring systems

Those are very susceptible to EMI interactions. The human body signals that they monitor are very weak. They are measured in unites of microvolts and micro-amps. Among other devices that are susceptible to EMI are hearing aids, wireless patient monitoring systems, magnetic resonance imaging systems, implantable cardiovascular devices, drug pumps, and portable diagnostic meters. As new technologies are developed and enter the marketplace at a fast pace the list will grow.

Signal to Interference-and-Noise Ratio (SINR)

Measured in dB:|S|/|(N+I)|

- S = -50dBm, N+I = -95dBm, SINR = 35dB
- S = -89dBm, N+I = -93dBm, SINR = 4dB

Channel capacity

The channel capacity determines the maximum transmission rate that a wireless channel can sustain with a negligible error probability in terms of bits per second per unit bandwidth. For the case of an ideal channel, where the only impairment in the wireless channel is the introduction of additive white Gaussian noise (AWGN), the channel capacity is given by Shannon's well-known formula.

$$C = \log_2 (1+\gamma) \text{ (bits/s/Hz)} \tag{1.9}$$

where γ is the ratio of the received signal power to the AWGN power, also known as the received signal-to-noise ratio (SNR). The consequence of Shannon's mathematical construct was the Shannon coding theorem and its converse. The Shannon coding theorem proves that there exists a code, which if utilized, allows to transmit data without errors at a rate r (bits/s/Hz) as long as $r < C$. While, the converse theorem showed that the error probability is always larger than zero if the transmission rate r is higher than the capacity C.

The channel capacity formula in (1.5) considers a simple scenario, by assuming an ideal AWGN channel. However, a realistic description of wireless propagation environment is far more complex. Mobile fading channels, particularly in urban environments, are generally classified as time-variant multipath fading channels, which can well be characterized with the help of proper statistical channel models.

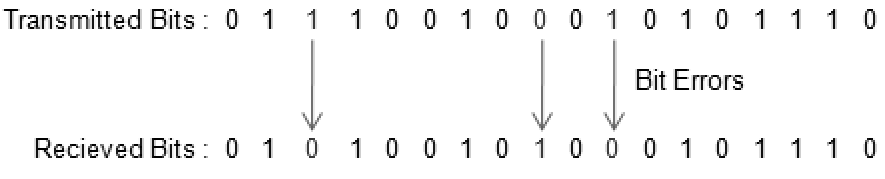
Bit Error Rates (BER)

The ultimate measure of quality in digital transmission is the *BER -Bit Error Rate* that corresponds to the fraction of erroneously decoded bits. Typical values of BER range between 10^{-3} and 10^{-9}.

On the surface, BER is a simple concept— its definition is simply:

BER = Errors/Total Number of Bits.

If the received bit is different from the transmitted bit, it is said that the data has bit error. As an example, if you have the transmitted bit stream and received bit stream as shown below, we can say this bit stream (data stream) has 3 bit errors. You see there are three locations where the transmitted bit and received bit is different.


Transmitted Bits : 0 1 1 1 0 0 1 0 0 0 1 0 1 0 1 1 1 0
Bit Errors
Recieved Bits : 0 1 0 1 0 0 1 0 1 0 0 0 1 0 1 1 1 0

It is the ratio of the number of bit errors and the total number of transmitted bits. It would be clear if you have a simple example as follows.

Example:

A million bits (10^6) was transmitted and 243 bits were found to be errors at the receiver. What is the bit error rate?

Solution: Bit Error Rate = (Number of Error Bits)/(Number of Transmitted Bits) = $243/10^6$ = 2.43×10^{-4}.

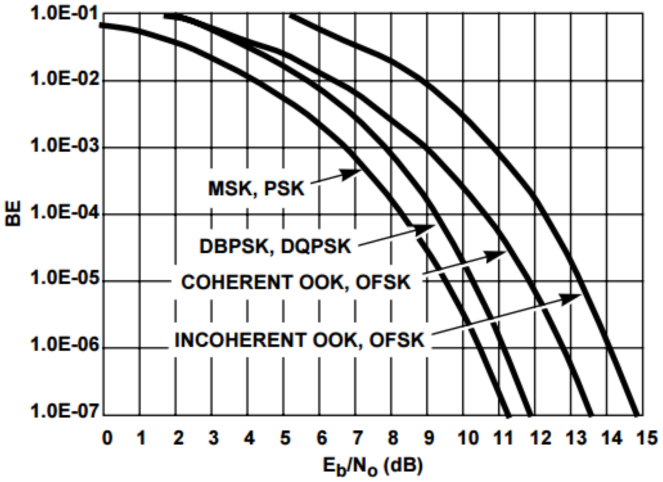


Figure 1.19. BER: Bit Error Rate

Modulation

Modulation is the process of superimposing the information contents of a modulating signal on a carrier signal (which is of high frequency) by varying the characteristic of carrier signal according to the modulating signal.

The robustness of the digital signal is also exemplified by the fact that it was chosen for the first trials of radio transmission. Marconi showed the feasibility of long distance transmission, but pretty soon realised that there was a need to share the medium among different users.

This was achieved by assigning different carrier frequencies which were modulated by each user's message. Modulation is a scheme to modify the amplitude, frequency or phase of the carrier according with the information one wants to transmit. The original information is retrieved at destination by the corresponding demodulation of the received signal.

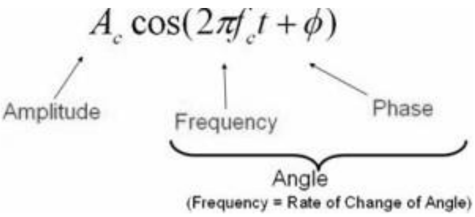
These modulation techniques are classified into two major types: analog and digital or pulse modulation. Prior to discussing further about the different types of modulation techniques, let us understand the importance of modulation.

The combination of different modulation schemes has resulted in a plethora of modulation techniques depending on which aspect one wants to optimise: robustness against noise, amount of information transmitted per second (capacity of the link in bits/second) or spectral efficiency (number of bits/s per Hertz).

For instance, BPSK -Binary Phase Shift Keying- is a very robust modulation technique but transmits only one bit per symbol, while 256 QAM -Quaternary Amplitude Modulation- will carry 8 bits per symbol, thus multiplying by a factor of eight the amount of information transmitted per second, but to correctly distinguish amongst the 256 symbols transmitted, the received signal must be very strong as compared with the noise (a very high S/N -Signal/Noise ratio- is required).

The modulation also allows us to choose which range of frequency we want to use for a given transmission. All frequencies are not created equal and the choice of the carrier frequency is determined by legal, commercial and technical constraints.

Analog Modulation


$A_c \cos(2\pi f_c t + \phi)$
Amplitude
Frequency
Phase
Angle
(Frequency = Rate of Change of Angle)

(1.10)

In this modulation, a continuously varying sine wave is used as a carrier wave that modulates the message signal or data signal. The Sinusoidal wave’s general function is shown in the figure below, in which, three parameters can be altered to get modulation – they are amplitude, frequency and phase, so the types of analog modulation are:

- Amplitude modulation (AM)
- Frequency modulation (FM)
- Phase modulation (PM)

In amplitude modulation, the amplitude of the carrier wave is varied in proportion to the message signal, and the other factors like frequency and phase remain constant. The modulated signal is shown in the below figure, and its spectrum consists of lower frequency band, upper frequency band and carrier frequency components. This type of modulation requires greater band width, more power. Filtering is very difficult in this modulation.

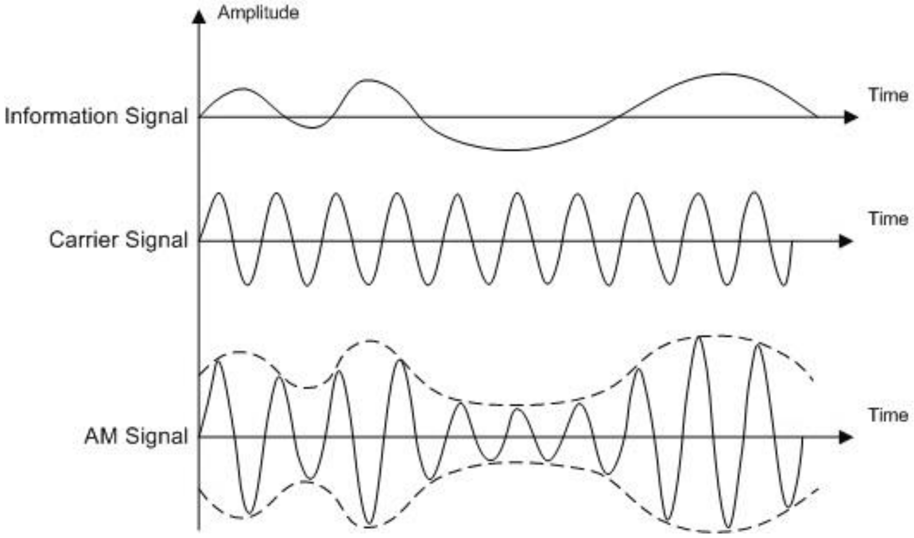


Figure 1.20. Amplitude modulation

Frequency modulation (FM) varies the frequency of the carrier in proportion to the message or data signal while maintaining other parameters constant. The advantage of FM over AM is the greater suppression of noise at the expense of bandwidth in FM. It is used in applications like radio, radar, telemetry seismic prospecting, and so on. The efficiency and bandwidths depend on modulation index and maximum modulating frequency.

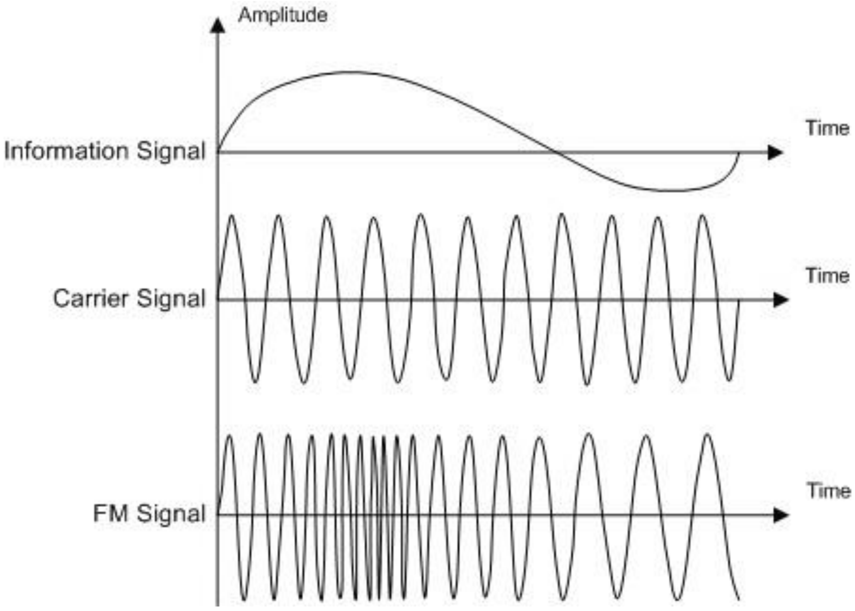


Figure 1.21. Frequency modulation

Phase modulation, the carrier phase is varied in accordance with the data signal. In this type of modulation, when the phase is changed it also affects the frequency, so this modulation also comes under frequency modulation.

Analog modulation (AM, FM and PM) is more sensitive to noise. If noise enters into a system, it persists and gets carried till the end receiver. Therefore, this drawback can be overcome by the digital modulation technique.

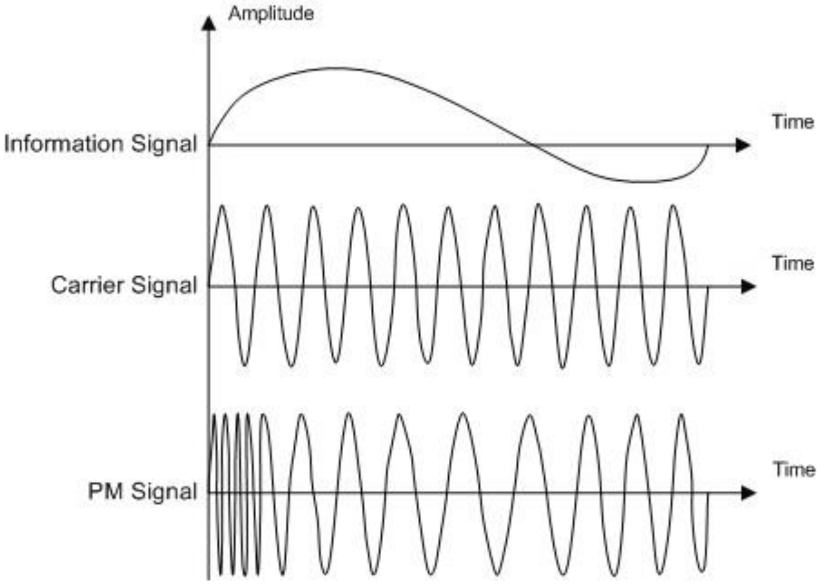


Figure 1.22. Phase modulation

Digital Modulation

For a better quality and efficient communication, digital modulation technique is employed. The main advantages of the digital modulation over analog modulation include permissible power, available bandwidth and high noise immunity. In digital modulation, a message signal is converted from analog to digital message, and then modulated by using a carrier wave.

The carrier wave is keyed or switched on and off to create pulses such that the signal is modulated. Similar to the analog, here the parameters like amplitude, frequency and phase variation of the carrier wave decides the type of digital modulation.

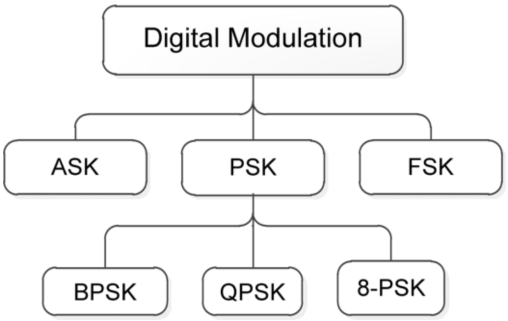


Figure 1.23. Types of Digital Modulation

Digital modulation is of several types depending on the type of signal and application used such as Amplitude Shift Keying, Frequency Shift Keying, Phase Shift Keying, Differential Phase Shift Keying, Quadrature Phase Shift Keying, Minimum Shift Keying, Gaussian Minimum Shift Keying, Orthogonal Frequency Division Multiplexing, etc., as shown in the figure.

Amplitude shift keying changes the amplitude of the carrier wave based on the base band signal or message signal, which is in digital format. It is used for low-band requirements and is sensitive to noise.

In frequency shift keying, the frequency of the carrier wave is varied for each symbol in the digital data. It needs larger bandwidths as shown in the figure. Similarly, the phase shift keying changes the phase of the carrier for each symbol and it is less sensitive to noise.

Multiplexing

Multiplexing is the sharing of a single communication channel among different users. The communication channel can be a copper wire, an optical fiber or the space between a transmitting and a receiving antenna.

Different users can be distinguished by means of different frequencies, time slots, codes or regions of space.

The process of making the most effective use of the available channel capacity is called Multiplexing. For efficiency, the channel capacity can be shared among a number of communicating stations just like a large water pipe can carry water to several separate houses at once. Most common use of multiplexing is in long-haul communication using coaxial cable, microwave and optical fibre.

The multiplexer is connected to the demultiplexer by a single data link. The multiplexer combines (multiplexes) data from these ‘n’ input lines and transmits them through the high capacity data link, which is being demultiplexed at the other end and is delivered to the appropriate output lines. Thus, Multiplexing can also be defined as a technique that allows simultaneous transmission of multiple signals across a single data link.

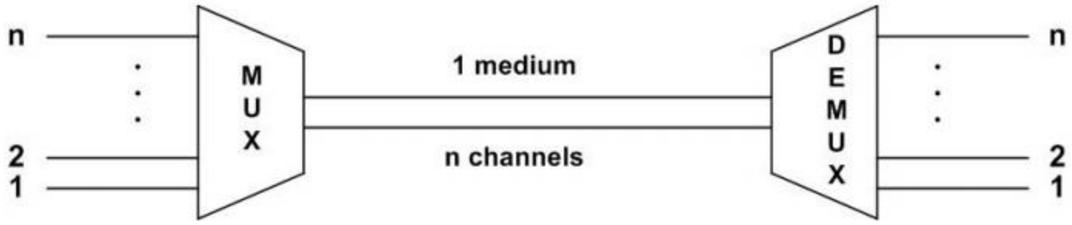


Figure 1.24. Basic concept of multiplexing

Multiplexing techniques

In FDMA (Frequency Division Multiple Access), each user has a different frequency band allocated. In frequency division multiplexing, the available bandwidth of a single physical medium is subdivided into several independent frequency channels. Independent message signals are translated into different frequency bands using modulation techniques, which are combined by a linear summing circuit in the multiplexer, to a composite signal. The resulting signal is then transmitted along the single channel by electromagnetic means as shown in Fig. 1.25.

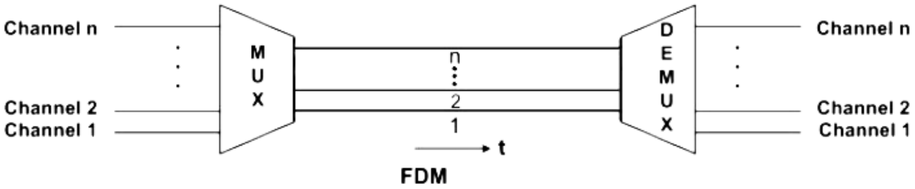


Figure 1.25. Basic concept of FDM

Basic approach is to divide the available bandwidth of a single physical medium into a number of smaller, independent frequency channels. Using modulation, independent message signals are translated into different frequency bands. All the modulated signals are combined in a linear summing circuit to form a composite signal for transmission.

In TDMA (Time Division Multiple Access), each user has d different time slot allocated, while the same frequency is shared among all the users of the service.

In frequency division multiplexing, all signals operate at the same time with different frequencies, but in Time-division multiplexing all signals operate with same frequency at different times. This is a base band transmission system, where an electronic commutator sequentially samples all data source and combines them to form a composite base band signal, which travels through the media and is being demultiplexed into appropriate independent message signals by the corresponding commutator at the receiving end. The incoming data from each source are briefly buffered. Each buffer is typically one bit or one character in length. The buffers are scanned sequentially to form a composite data stream. The scan operation is sufficiently rapid so that each buffer is emptied before more data can arrive. Composite data rate must be at least equal to the sum of the individual data rates. The composite signal can be transmitted directly or through a modem. The multiplexing operation is shown in Fig. 1.26.

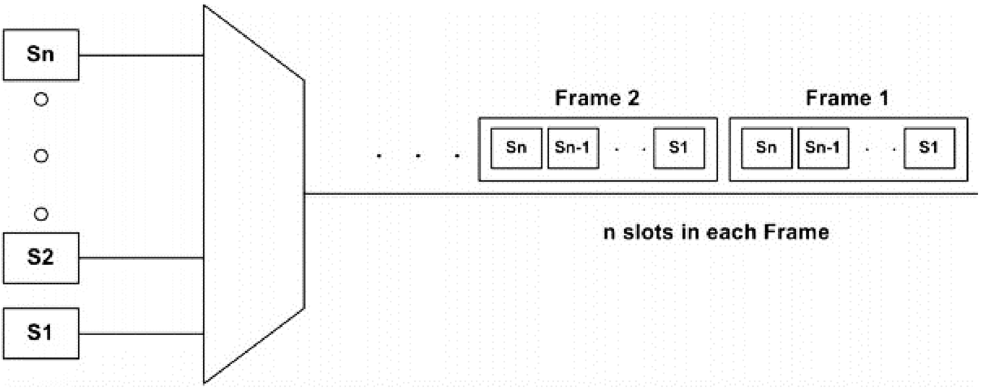


Figure 1.26. Time division multiplexing operation

As shown in the Fig 1.26. the composite signal has some dead space between the successive sampled pulses, which is essential to prevent inter channel cross talks. Along with the sampled pulses, one synchronizing pulse is sent in each cycle. These data pulses along with the control information form a frame. Each of these frames contain a cycle of time slots and in each frame, one or more slots are dedicated to each data source. The maximum bandwidth (data rate) of a TDM system should be at least equal to the same data rate of the sources. Synchronous TDM is called synchronous mainly because each time slot is preassigned to a fixed source. The time slots are transmitted irrespective of whether the sources have any data to send or not. Hence, for the sake of simplicity of implementation, channel capacity is wasted. Although fixed assignment is used TDM, devices can handle sources of different data rates. This is done by assigning fewer slots per cycle to the slower input devices than the faster devices. Both multiplexing and demultiplexing operation for synchronous TDM are shown in Fig. 1.27.

IN CDMA (Code Division Multiple Access), the users are distinguished by means of a special mathematical code, while sharing the same frequency and time slots. CDMA is also known as spread spectrum because it takes the digitized version of an analogue signal and spreads it out over a wider bandwidth at a lower power level. This method is called direct sequence spread spectrum (DSSS) as well (Fig. 1.28). The digitized and compressed voice signal in serial data form is spread by processing it in an XOR circuit along with a chipping signal at a much higher frequency. In the cdma IS-95 standard, a 1.2288-Mbit/s chipping signal spreads the digitized compressed voice at 13 kbits/s.

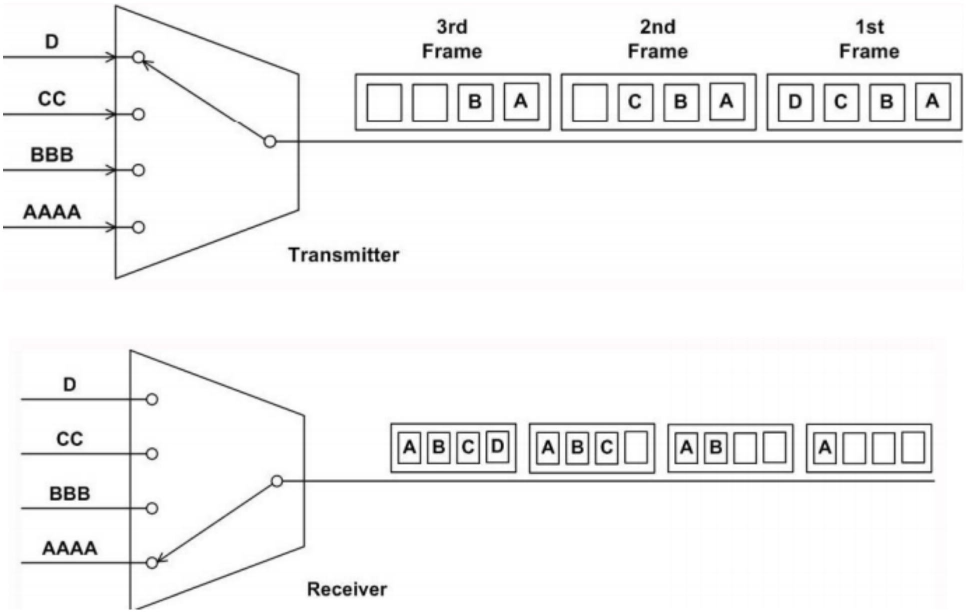


Figure 1.27. Multiplexing and demultiplexing in synchronous TDM

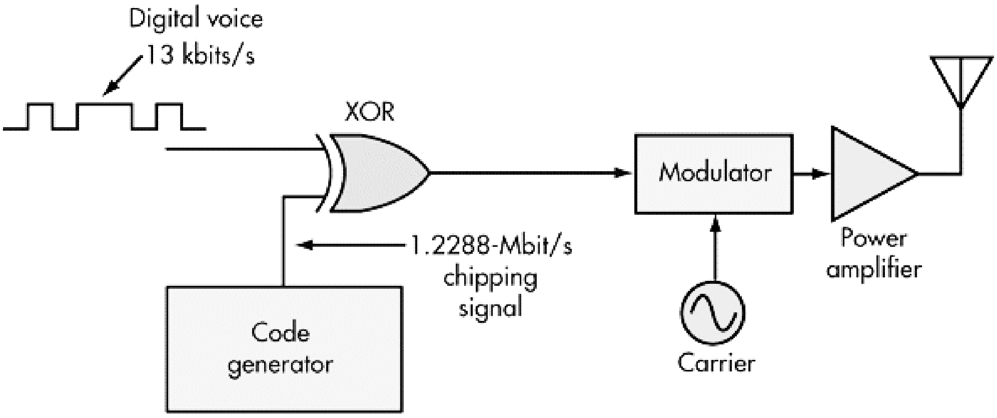


Figure 1.28. Spread spectrum is the technique of CDMA

The chipping signal is derived from a pseudorandom code generator that assigns a unique code to each user of the channel. This code spreads the voice signal over a bandwidth of 1.25 MHz. The resulting signal is at a low power level and appears more like noise. Many such signals can occupy the same channel simultaneously. For example, using 64 unique chipping codes allows up to 64 users to occupy the same 1.25-MHz channel at the same time. At the receiver, a correlating circuit finds and identifies a specific caller’s code and recovers it.

The third generation (3G) cell-phone technology called wideband CDMA (WCDMA) uses a similar method with compressed voice and 3.84-Mbit/s chipping codes in a 5-MHz channel to allow multiple users to share the same band.

IN OFDM (Orthogonal Frequency Division Multiplexing)

OFDM is a transmission technique that has been around for years, but only recently became popular due to the development of digital signal processors (DSPs) that can handle its heavy digital processing requirements.

Orthogonal frequency division multiplexing (OFDM) is a modulation method that divides a channel into multiple narrow orthogonal bands that are spaced so they do not interfere with one another. Each band is divided into hundreds or even thousands of 15-kHz wide subcarriers.

The data to be transmitted is divided into many lower-speed bit streams and modulated onto the subcarriers. Time slots within each subchannel data stream are used to package the data to be transmitted (Fig. 1.29). This technique is very spectrally efficient, so it provides very high data rates. It also is less affected by multipath propagation effects.

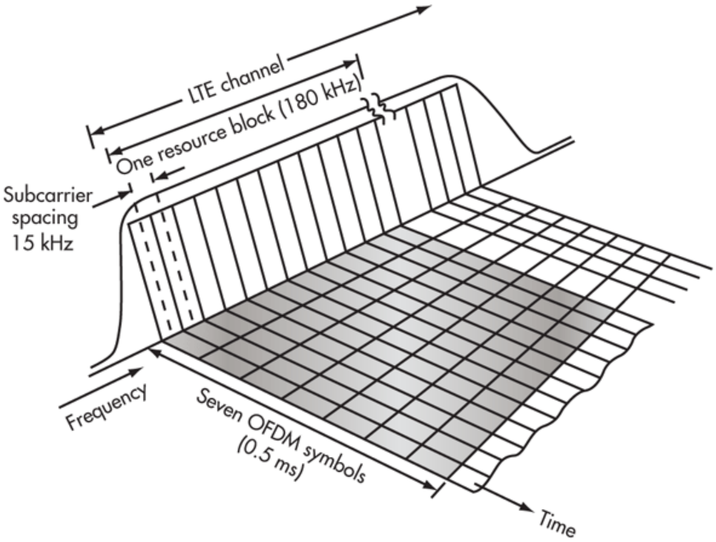

LTE channel
One resource block (180 kHz)
Subcarrier spacing 15 kHz
Frequency
Seven OFDM symbols (0.5 ms)
Time

Figure 1.29. OFDMA assigns a group of subcarriers to each user

To implement OFDMA, each user is assigned a group of sub channels and related time slots. The smallest group of sub channels assigned is 12 and called a resource block (RB). The system assigns the number of RBs to each user as needed.

Behaviour of radio waves

There are a few simple rules of thumb that can prove extremely useful when making first plans for a wireless network:

- The longer the wavelength, the further it goes
- The longer the wavelength, the better it travels through and around things
- The shorter the wavelength, the more data it can transport

Assuming equal power levels, waves with longer wavelengths tend to travel further than waves with shorter wavelengths. Lower frequency transmitters can reach greater distances than high frequency transmitters at the same power.

It is harder to visualize waves moving “through” solid objects, but this is the case with electromagnetic waves. Longer wavelength (and therefore lower frequency) waves tend to penetrate objects better than shorter wavelength (and therefore higher frequency) waves. For example, FM radio (88-108MHz) can travel through buildings and other obstacles easily, while shorter waves (such as GSM phones operating at 900MHz or 1800MHz) have a harder time penetrating buildings. This effect is partly due to the difference in power levels used for FM radio and GSM, but is also partly due to the shorter wavelength of GSM signals.

Absorption

When electromagnetic waves go through some material, they generally get weakened or dampened. Materials that absorb energy include:

- Metal. Electrons can move freely in metals, and are readily able to swing and thus absorb the energy of a passing wave.
- Water molecules jostle around in the presence of radio waves, thus absorbing some energy.
- Trees and wood absorb radio energy proportionally to the amount of water contained in them.
- Humans are mostly water: we absorb radio energy quite well!

Polarization

- Electromagnetic waves have *electrical* and *magnetic components* that oscillate perpendicular to each other and to the direction of the propagation.
- The *polarization* of the wave corresponds to the plane in which the electrical oscillations occur.

Another important quality of electromagnetic waves is polarization. Polarization describes the direction of the electrical field vector. If you imagine a vertically aligned dipole antenna (a straight piece of wire), electrons only move up and down, not sideways (because there is no room to move) and thus electrical fields only ever point up or down, vertically. The energy leaving the wire and traveling as a wave has a strict linear (and in this case, vertical) polarization. If we put the antenna flat on the ground, we would find horizontal linear polarization.

Most Wi-Fi antennas we work with are linearly polarized, but circularly polarized antennas are also sometimes used (for special purposes). The polarization of a transmitting and receiving antenna MUST MATCH for optimum communications.

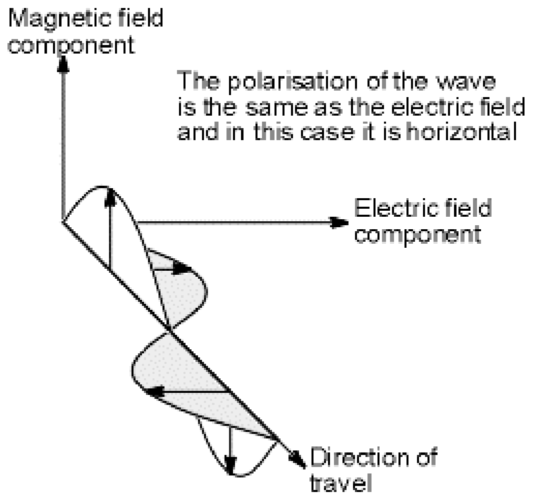

Magnetic field component
The polarisation of the wave is the same as the electric field and in this case it is horizontal
Electric field component
Direction of travel

Figure 1.30. The polarization of the wave

Conclusions

- The communication system must overcome the noise and interference to deliver the signal to the receiver.
- The capacity of the communication channel is proportional to the bandwidth and to the logarithm of the S/N ratio.
- Modulation is used to adapt the signal to the channel and to allow several signals to share the same channel.
- Higher order modulation schemes allows for a higher transmission rate, but require higher S/N ratio.
- The channel can be shared by several uses that occupy different frequencies, different time slots or different codes,
- Radio waves have a characteristic wavelength, frequency and amplitude, which affect the way they travel through space.
- Wi-Fi uses a tiny part of the electromagnetic spectrum. Lower frequencies travel further, but at the expense of throughput.
- Radio waves occupy a volume in space, the Fresnel zone, which should be unobstructed for optimum reception.

1.1.3. Wireless sensor networks in smart metering

Wireless network today refers to any kind of communication that can be implemented without wires, such as wireless energy transfer and wireless remote control. Automatic detection, prevention, and recovery from urban disasters are one of many potentials uses for wireless sensor networks. Despite their variety, all sensor networks have certain fundamental features in common. Sensors detect the world's physical nature, such as light intensity, temperature, sound, or proximity to objects. Similarly, actuators, affect the world in some way, such as toggling a switch, making a noise, or exerting a force. Sensors perform monitoring and remote diagnostics, and they usually use battery-powered transceivers.

A Wireless sensor network (WSN) consists of wireless sensor nodes or motes, which are devices equipped with a processor, a radio interface, an analogue-to-digital converter, sensors, memory, and a power supply. The processor provides the mote management functions and performs data processing. The sensors attached to the mote are capable of sensing temperature, humidity, light, etc. Due to bandwidth and power constraints, motes primarily support low data units with limited computational power and a limited sensing rate. Memory is used to store programs (instructions executed by the processor) and data (raw and processed sensor measurements). Motes are equipped with a low-rate (10–100 kbps) and short-range (less than 100 m) wireless radio, e.g., IEEE 802.15.4 radio to communicate among themselves. Since radio communication consumes most of the power, the radio must incorporate energy-efficient communication techniques. The power source commonly used is rechargeable batteries (P. Rawat, K. D. Singh, H. Chaouchi, J. M. Bonnin, 2013).

Since motes can be deployed in remote and hostile environments they must use little power and must employ built-in mechanisms to extend network lifetime. For example, motes may be equipped with effective power harvesting methods, such as solar cells, so they may be left unattended for years.

Sensor nodes can be deployed in an ad-hoc or a pre-planned manner. An ad-hoc deployment is good for large uncovered regions where a network of a very large number of nodes can be deployed and left unattended to perform monitoring and reporting functions. Network maintenance such as managing connectivity and detecting failures is difficult in such a WSN due to large number of nodes. On the other hand, preplanned deployment is good for limited coverage where fewer nodes are deployed at specific locations with the advantage of lower network maintenance and management cost.

WSN: challenges and requirements

The collaborative nature of WSNs brings several advantages over conventional wireless ad-hoc networks, including self-organization, rapid deployment, flexibility, and inherent intelligent-processing capability. However, the unique features of WSN present new challenges in hardware design, communication protocols, and application design. A WSN technology must address these challenges to realize the numerous envisioned applications. This requires modifying legacy protocols for conventional wireless ad-hoc networks or designing new effective communication protocols and algorithms (Rodrigues J.J., Neves P.A., 2010)

Table 1. Challenges vs. required mechanisms in WSN (P. Rawat, K. D. Singh, H. Chaouchi, J. M. Bonnin, 2013)

Challenges	**Required mechanisms**
Resource constraints	Efficient use of resources

Dynamic and extreme environment conditions	Adaptive network operation
Data redundancy	Data fusion and localized processing
Unreliable wireless communication	Reliability
No global identification (ID) for sensor nodes	Data-centric communication paradigm
Prone to node failures	Fault tolerance
Large scale deployment	Low-cost small-sized sensors with self-configuration and self-organization

Table 1 lists important challenges and corresponding required mechanisms to address them in WSN. Sensor nodes have resource constraints including limited energy, limited memory, and computational capacities. The limited energy supplies of the sensor nodes in the network impose lifetime constraints on the WSN. The problem of limited resources can be addressed by using them efficiently.

Dynamic network topologies and harsh environment conditions may cause sensor node failures and performance degradation. This requires WSN to support adaptive network operation including adaptive signal-processing algorithms and communication protocols to enable end-users to cope with dynamic wireless-channel conditions and varying connectivity.

The communication in WSN is unreliable due to error prone wireless medium with high bit error rates and variable-link capacity. Thus, a WSN should be reliable in order to function properly and depending on the application requirements, the sensed data should be reliably delivered to the sink node. WSNs are usually prone to unexpected node failures due to different reasons like nodes may run out of energy or might be damaged (in extreme environment conditions), or wireless communication between two nodes can be permanently interrupted. This requires WSNs to be robust to node failures. In WSN, fault tolerance can be improved through a high level of redundancy by deploying additional nodes than required if all nodes functioned properly.

Since WSNs may contain a large number of sensor nodes, the employed architectures and protocols must be able to scale to sizes of thousands or more. Moreover, a large scale deployment of WSN requires low-cost and small-sized sensor nodes. A WSN should be able to self-organize itself as the network topology may change due to reasons like node failure, mobility, and large scale deployments.

Types of WSNs

Presently many WSNs are deployed on land, underground and underwater. They face different challenges and constraints depending on their environment. We present five types of WSNs (Yick J, Mukherjee B, Ghosal D., 2008).

Terrestrial WSN consists in a large number (hundreds to thousands) of low-cost nodes deployed on land in a given area, usually in an ad-hoc manner (e.g., nodes dropped from an airplane). In terrestrial WSNs, sensor nodes must be able to effectively communicate data back to the base station in a dense environment. Since battery power is limited and usually non-rechargeable, terrestrial sensor nodes can be equipped with a secondary power source such as solar cells. Energy can be conserved with multi-hop optimal routing, short transmission range, in-network data aggregation, and using low duty-cycle operations. Common applications of terrestrial WSNs are environmental sensing and monitoring, industrial monitoring, and surface explorations.

Underground WSN consists of a number of sensor nodes deployed in caves or mines or underground to monitor underground conditions (Li M, Liu Y., 2007). In order to relay information from the underground sensor nodes to the base station, additional sink nodes are located above ground. They are more expensive than terrestrial WSNs as they require appropriate equipment to ensure reliable communication through soil, rocks, and water. Wireless communication is a challenge in such environment due to high attenuation and signal loss. Moreover, it is difficult to recharge or replace the battery of nodes buried underground making it important to design energy efficient communication protocol for prolonged lifetime. Underground WSNs are used in many applications such as agriculture monitoring, landscape management, underground monitoring of soil, water or mineral, and military border monitoring.

Underwater WSNs consists of sensors deployed underwater, for example, into the ocean environment. Such nodes being expensive, only a few nodes are deployed and autonomous underwater vehicles are used to explore or gather data from them. Underwater wireless communication uses acoustic waves that presents various challenges such as limited bandwidth, long propagation delay, high latency, and signal fading problems. These nodes must be able to self-configure and adapt to extreme conditions of ocean environment. Nodes are equipped with a limited battery which cannot be replaced or recharged requiring energy efficient underwater communication and networking techniques. Applications of underwater WSNs include pollution monitoring, under-sea surveillance and exploration, disaster prevention and monitoring, seismic monitoring, equipment monitoring, and underwater robotics.

Multimedia WSN consists of low-cost sensor nodes equipped with cameras and microphones, deployed in a pre-planned manner to guarantee coverage (Akyildiz I, Melodia T, Chowdhury K., 2007). Multimedia sensor devices are capable of storing, processing, and retrieving multimedia data such as video, audio, and images. They must cope with various challenges such as high bandwidth demand, high energy consumption, quality of service (QoS) provisioning, data processing, and compressing techniques, and cross-layer design. It is required to develop transmission techniques that support high bandwidth and low energy consumption in order to deliver multimedia content such as a video stream. Though QoS provisioning is difficult in multimedia WSNs due to variable link capacity and delay, a certain level of QoS must be achieved for reliable content delivery. Multimedia WSNs enhance the existing WSN applications such as tracking and monitoring.

Mobile WSN consists of mobile sensor nodes that can move around and interact with the physical environment. Mobile nodes can reposition and organize themselves in the network in addition to be able to sense, compute, and communicate.

A dynamic routing algorithm must, thus, be employed unlike fixed routing in static WSN. Mobile WSNs face various challenges such as deployment, mobility management, localization with mobility, navigation and control of mobile nodes, maintaining adequate sensing coverage, minimizing energy consumption in locomotion, maintaining network connectivity, and data distribution. Primary examples of mobile WSN applications are monitoring (environment, habitat, underwater), military surveillance, target tracking, search and rescue. A higher degree of coverage and connectivity can be achieved with mobile sensor nodes compared to static nodes.

1.1.4. WSN standards and technologies

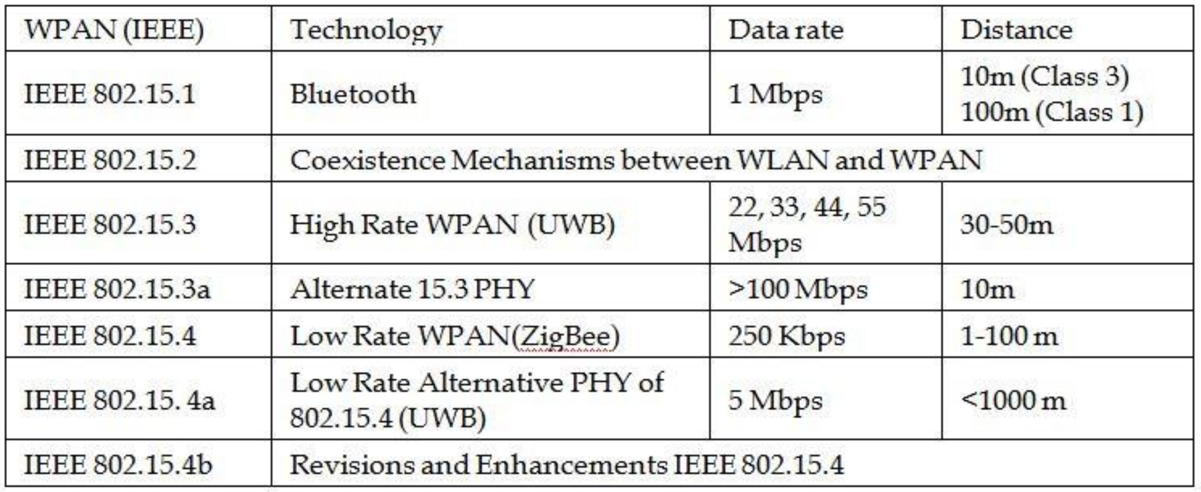
The process of standardization in the field of WSN has been very active in the last years. As compared to some well-known wireless communication standards such as IEEE 802.11 and IEEE 802.15, the standard IEEE 802.15.4 (IEEE 802.15.4-2006 standard) is specifically designed for low power, low data rate, and low-cost wireless sensor communication.

In comparison, Wi-Fi (IEEE 802.11) provides higher data throughput and range, but it consumes more energy resulting in a crucial disadvantage for WSNs. Bluetooth Low Energy (BLE) (Bluetooth, 2015) is considered as an attractive technology for WSN applications demanding higher data rates, but short range. Most of the WSN technologies operate in *the ISM band (Industrial, Scientific and Medical radio band*), which were internationally reserved for the use of RF (Radio Frequency) electromagnetic fields for industrial, scientific, and medical purposes other than communications.

The choice of technology to be used should be based on the target application as every WSN application has different requirements on the communication system. While some applications need a very low latency, others need a high secure connection or a long battery life. The development of new technologies is pushing WSN into new areas of application. While ISA100 (ISA-100.11a-2009) and Wireless HART (Kim A, Hekland F, Petersen S, Doyle P., 2008) technologies make WSNs a more viable possibility in traditional manufacturing environments, technologies like Bluetooth low energy, ZigBee green power, Wi-Fi direct and EnOcean will drive growth into areas such as medical devices, healthcare, automotive, energy efficient buildings, sports, and agriculture. In this section, we describe the potential WSN standards and technologies. In addition to the standard technologies, some commercial nonstandard technologies like ANT (ANT technology, 2014) are also considered.

A comparative study of emerging and existing radio technologies for WSNs is provided in Table 2.

Table 2. WSN standards and technologies



WPAN (IEEE)	Technology	Data rate	Distance
IEEE 802.15.1	Bluetooth	1 Mbps	10m (Class 3) 100m (Class 1)
IEEE 802.15.2	Coexistence Mechanisms between WLAN and WPAN		
IEEE 802.15.3	High Rate WPAN (UWB)	22, 33, 44, 55 Mbps	30-50m
IEEE 802.15.3a	Alternate 15.3 PHY	>100 Mbps	10m
IEEE 802.15.4	Low Rate WPAN(ZigBee)	250 Kbps	1-100 m
IEEE 802.15. 4a	Low Rate Alternative PHY of 802.15.4 (UWB)	5 Mbps	<1000 m
IEEE 802.15.4b	Revisions and Enhancements IEEE 802.15.4		

IEEE 802.15.4 standard

IEEE 802.15.4 (IEEE 802.15.4-2006 standard) is a standard defined by IEEE 802.15.4 Working Group for data communication devices operating in Low Rate Wireless Personal Area Networks (LR-WPANs). It provides low cost, short-range, low power, and low data-rate communication for sensor networks. It targets wireless sensor applications, which require short range communication to maximize battery life. The standard specifies the lowest two layers of the protocol stack; the physical (PHY) and medium access control (MAC) layers, based on the OSI model as shown in Fig. 4. The upper layers and interoperability sublayers of the protocol stack are separately defined by other architectures such as 6LoWPAN (Montenegro G, Kushalnagar N, Hui J, Culler D., 2007). Transmission of IPv6 packets over IEEE 802.15.4 networks. Internet proposed standard RFC 4944, ZigBee (ZigBee Alliance), ISA100.11a, and Wireless HART. Table 3 provides a technical comparison between the key IEEE 802.15.4-based WSN standards.

The IEEE 802.15.4 standard defines two types of network nodes: full-function device (FFD) and reduced-function device (RFD). RFDs are very basic nodes with little processing and memory resources. They can only act as end-systems in the network and communicate with FFDs, whereas FFDs are able to fully implement the standard. FFDs can act as coordinators (Personal Area Networks (PAN) or full network coordinators) and communicate with both FFDs and RFDs.

IEEE 802.15.4 supports two types of network topologies: star and peer-to-peer topology for communication between network devices as shown in Fig. 1.32. In the star topology, all the devices communicate with a central controller (FFD) while the peer-to-peer topology allows more complex network formations to be implemented, such as mesh networking topology.

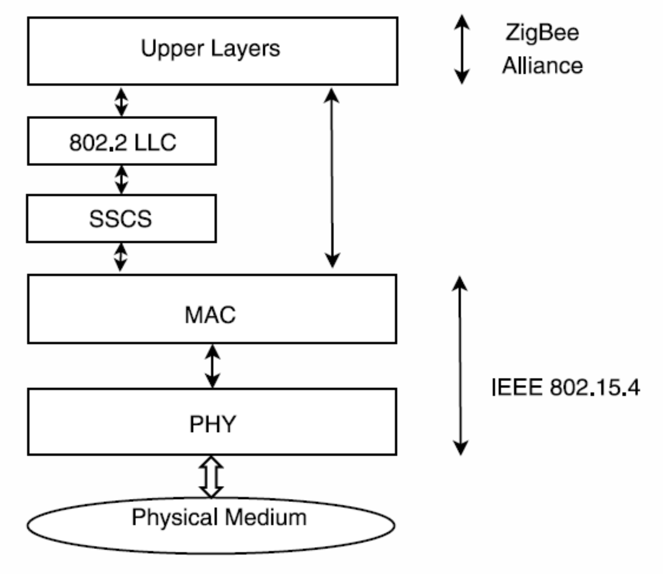


Figure 1.31. IEEE 802.15.4 protocol stack. Upper layers: ZigBee, 6LowPAN, etc.

Table 3. Key IEEE 802.15.4-based WSN standards

Standard	**Topology**	**Battery life (days)**	**Network nodes**	**Max Throughput**	**Range (m)**
ZigBee	Mesh	100–1000+	255	250 kbps	10–100
6LoWPAN	Mesh	100–365+	65536	250 kbps	1-100
Wireless HART	Mesh	760+	200	250 kbps	1-100
ISA100.11a	Mesh, Star	1000+		250 kbps	100

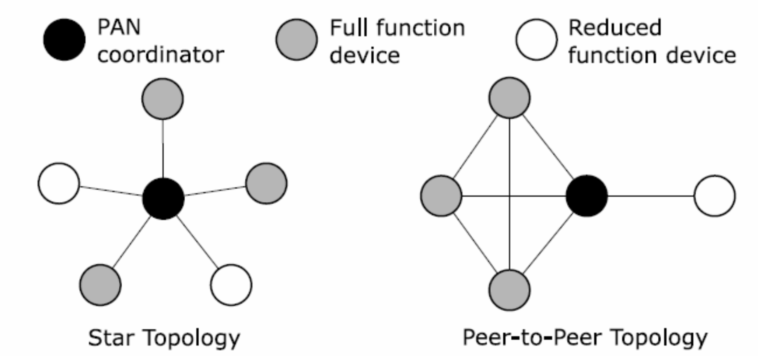


Figure 1.32. IEEE 802.15.4 - compliant network topologies: star and peer-to-peer topology

A peer-to-peer network can be ad-hoc, self-organizing, and self-healing. Star topology is preferred when coverage area is small and low latency is required by the WSN application, whereas peer-to-peer topology is suitable for a large coverage area where latency is not a critical issue.

The original 2003 version supports 868/915 MHz low bands with data rates of 20 and 40 kbps, and 2.4 GHz high bands with a rate of 250 kbps. The current version of the IEEE standard is 802.15.4-2006. It improves the maximum data rates of up to 100 and 250 kbps for the 868/915 MHz bands.

More information one can find in the chapter *1.4.5.Wireless technologies*.

6LoWPAN

The 6LoWPAN standard (RFC 4944) has been defined by IETF to adapt IPv6 communication on top of IEEE 802.15.4 networks. 6LoWPAN refers to IPv6 over.

Low power Wireless Personal Area Networks. It enables IPv6 packets communication over low power and low rate IEEE 802.15.4 links and assures interoperability with other IP devices. 6LowPAN devices can communicate directly with other IP enabled devices. IP for Smart Objects (IPSO) Alliance is promoting the use of 6LowPAN and embedded IP solutions in smart objects. 6LoWPAN provides an adaptation layer, new packet format, and address management to enable such devices to have all the benefits of IP communication and management. Since IPv6 packet sizes are much larger than the frame size of IEEE 802.15.4, the adaptation layer is introduced between MAC layer and the network layer to optimize IPv6 over IEEE 802.15.4. The adaptation layer provides mechanisms for IPv6 packet header compression, fragmentation and reassembly allowing IPv6 packets transmission over IEEE 802.15.4 links.

More information one can find in the chapter *1.4.5.Wireless technologies*.

Wireless HART

Wireless HART, released in 2007, is a wireless communications standard suitable for industrial applications such as process measurement and control applications. It adds wireless capabilities to the HART protocol while maintaining compatibility with existing HART devices. A Wireless HART network consists of wireless field devices, gateways, process automation controller, host applications, and network manager.

Field devices are connected to process or plant equipment and communicate with the host applications through gateways. The process automation controller serves as a single controller for continuous process. The network manager is responsible for configuring the network, scheduling communication between devices, managing routes, and monitoring network health. Wireless HART operates in the 2.4 GHz ISM band and to prevent interference from other applications, it uses frequency hopping with blacklisting of bad channels and has a high reliability in challenging environments.

The key features are its reliability, security, energy efficiency, compatibility with existing devices, and it enables mesh networking.

ISA100.11a

ISA100.11a standard is developed by the ISA100 standards committee which is a part of the International Society of Automation (ISA) organization. Its main application is in industrial automation and to meet the needs of industrial applications, it supports various network topologies, such as star and mesh networking. An ISA100.11a WSN consists of field devices, gateways, and hand-held devices. Field devices are responsible for gathering sensor data and some of them can also provide routing functionalities.

The gateways ensure connection between the WSN and the user application and also support interoperability with existing standards, such as Wireless HART by translating and tunnelling information between the networks. Handheld devices support device installation, configuration and maintenance. One of the key features of ISA100.11a is the low latency or fast response time of 100 ms. ISA100.11a uses only the 2.4 GHz ISM band with frequency hopping to increase reliability and prevent interference from other wireless networks.

IEEE 802.15.4a—ultra wideband

Ultra-Wideband (UWB) is a Radio Frequency (RF) communication technology in which the information is transmitted through a series of very short impulses emitted in periodic sequences (Porcino D, Hirt W., 2003). Ultra-wideband radio technology: potential and challenges ahead. IEEE Commun Mag 41(7):66–74]. A UWB signal can be defined as a signal with instantaneous spectral occupancy in excess of 500 MHz or a fractional bandwidth of more than 20 %. UWB has been a proposed technology for the IEEE 802.15.4a standard, which provides an alternative physical layer for low rate WPAN and is an amendment to IEEE 802.15.4. The advantages of UWB include its spectral efficiency, ability to transmit high data rates with low power, high precision ranging and location capability, and ability to cope with multipath environments. However, UWB is not suitable for communication over longer distances or measuring data from unsafe zone because of high peak energy of pulses. Impulse Radio-UWB (IR-UWB) that relies on ultrashort (nanosecond scale) waveforms is a promising UWB technique for WSN applications.

Bluetooth and Bluetooth low energy (BLE)

Bluetooth is a wireless technology for short-range and cheap devices intended to replace the cables in WPANs. It operates in the 2.45 GHz ISM band and uses frequency hopping to combat interference and fading. Bluetooth can cover a communication range of 10–100 m and allows data rate up to 3 Mbps. It was standardized as IEEE 802.15.1, but the standard is no longer maintained. Currently, Bluetooth is managed by the Bluetooth Special Interest Group, which adopted Bluetooth Core Specification Version 4.0 in 2010.

Bluetooth v4.0 (Bluetooth Specification v.4)] is the most recent version. It introduced Bluetooth Low Energy (BLE) technology that enables new low-cost Bluetooth Smart devices to operate for months or years on tiny, coin-cell batteries. Potential markets for BLEbased devices include healthcare, sports and fitness, security, and home entertainment.

BLE operates in the same 2.45 GHz ISM band as classic Bluetooth, but uses a different set of channels. Instead of Bluetooth's 1-MHz wide 79 channels, BLE has 2-MHz wide 40 channels. As compared to classic Bluetooth, BLE is intended to provide considerably reduced power consumption and lower cost, with enhanced communication range. BLE allows 1 Mbps data rates with 200 m range and has two implementation alternatives; single-mode and dual-mode. Single-mode BLE devices support only new BLE connections, whereas dual-mode devices support both classic Bluetooth as well as new BLE connections and have backward-compatibility.

Z-wave

The Z-Wave is a low powered RF-based wireless communications technology designed specifically for remote control applications in residential and light commercial environments. It was developed by Zensys and is currently, supported by ZWave Alliance (Z-Wave Alliance, 2016). Z-Wave's main advantage with respect to IEEE 802.15.4-based technologies is that it operates in sub-1 GHz band (around 900 MHz); unaffected to interference from Wi-Fi and other wireless technologies (Bluetooth, ZigBee, etc.) in the crowded 2.4-GHz range. The 868 MHz band used by Z-Wave in Europe is limited by European regulations to operate at or under 1 % duty cycle that can be sufficient for most of the control applications. Z-Wave technology supports mesh networking, operable data rates of 9.6 kbps and 40 kbps and maximum outdoor range of 30 m.

ANT technology

ANT (ANT technology) is a proprietary technology that features a wireless communication protocol stack for ultra-low power networking applications. It is designed to run using low cost, low power microcontrollers and transceivers operating in the 2.4 GHz ISM band.

ANT supports various topologies including peer-to-peer, star, tree, and other types of mesh networking in personal area networks (PAN) suited for sports, fitness, wellness, and home health applications. It is also suited for local area networks (LAN) in homes and industrial automation applications. ANT is energy-efficient and provides a data rate of 1 Mbps, which is much higher than that of IEEE 802.15.4 (250 kbps).

However, it lacks interoperability. This can be addressed by adding ANT+, an interoperability function to the base ANT protocol to make it interoperable. ANT+ enabled fitness monitoring devices, such as heart rate monitors, speed monitors, and weight scales can all work together to assemble and track performance metrics.

Wavenis technology

Wavenis is an ultra-low-power and long-range wireless technology developed by Coronis for WSN applications in which communication ability and device autonomy present conflicting requirements. It was originally developed as a proprietary technology, and is now promoted by the Wavenis Open Standard Alliance. Wavenis based devices are used in Telemetry, industrial automation, remote utility meter monitoring, home healthcare, access control and cold-chain monitoring. Its key features include reliability, power savings, network coexistence, and robustness against interferers.

Wavenis operates worldwide in the 868, 915, and 433 MHz ISM bands. Its data rates are programmable, from 4.8 kbps to 100 kbps. Most Wavenis applications communicate at 19.2 kbps.

EnOcean technology

EnOcean is an emerging WSN technology that is promoted by EnOcean Alliance. The EnOcean wireless standard (EnOcean, 2016)] is optimized for solutions with ultra-low power consumption and energy harvesting. The battery free EnOcean technology brings together wireless sensing and energy harvesting to enable energy harvester-powered WSNs. The goal of EnOcean's energy harvesting wireless sensor technology is to draw energy from the surroundings, for example, from motion, pressure, light or differences in temperature and convert that into energy that can be used electrically.

Thus, combining miniaturized energy harvesters and highly efficient wireless technology enable designing WSN that is supplied via energy harvesting. EnOcean provides wireless sensor solutions for buildings and industrial automation. It uses 868 MHz and 315 MHz and supports transmission range of up to 30 m indoor and 300 m outdoor. EnOcean products available in the market include battery-less self-powered wireless sensors and switches. Battery-less EnOcean modules with energy harvesting are available which reduce the life cycle cost as they are maintenance free.

1.1.5. Application domains and deployments

WSNs have been adopted in a large number of diverse application domains. It is envisioned that in future everyday objects will be embedded with sensors to make them smart. Smart objects can explore their environment, communicate with other smart objects, and interact with humans.

A taxonomy of WSN applications is shown in Fig. 1.33. In general, WSN applications can be of two types: monitoring and tracking. As shown in the taxonomy (Fig. 1.34), the leading application domains of WSNs include military and crime prevention, environment, health (Body Area Networks), industry and agriculture, and urbanization and infrastructure.

Military operations involving force protection with unattended ground sensors formed into intelligent networks around forward operating bases are receiving much attention. VigilNet (VigilNet, 2016) is an integrated sensor network system for energy-efficient surveillance missions. Another interesting example is networked mines called self-healing minefields that automatically rearrange themselves to ensure optimal coverage. Body Area Networks (BANs) integrated with soldier communication systems are also a key application, as vital health functions can be monitored when soldiers enter hazardous areas. In addition, the homeland security sector is showing great interest in WSNs for critical infrastructure monitoring (utilities, airports, etc.), border protection, incident detection, and crisis management. In the health sector, BANs for health applications are one of the emerging markets for WSNs.

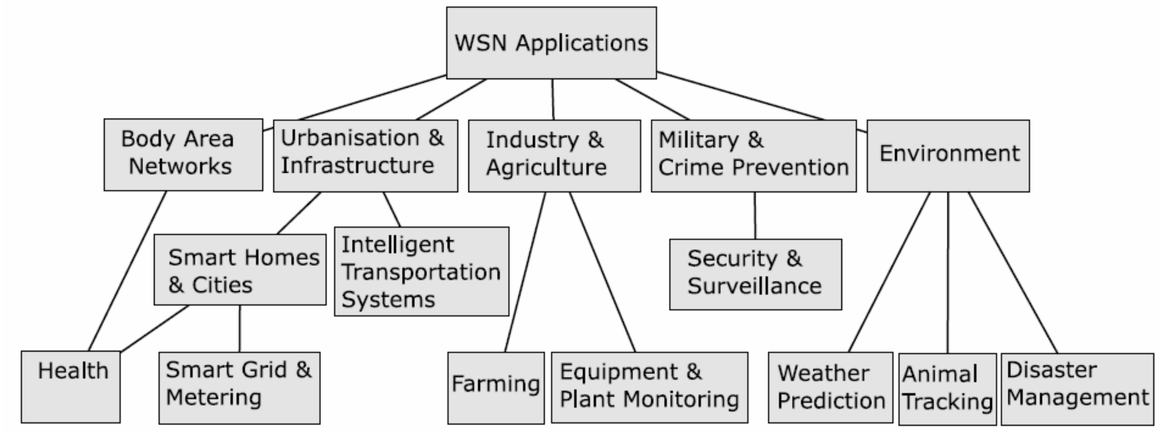


Figure 1.33. Taxonomy of WSN applications

Furthermore, in the civil sector, WSNs have generated a lot of interest from their smart infrastructure applications, such as smart grids, smart energy metering, smart transport and traffic management, smart roads, etc. Structural health monitoring enables detecting the health status of structures using a network of accelerometers and strain gages.

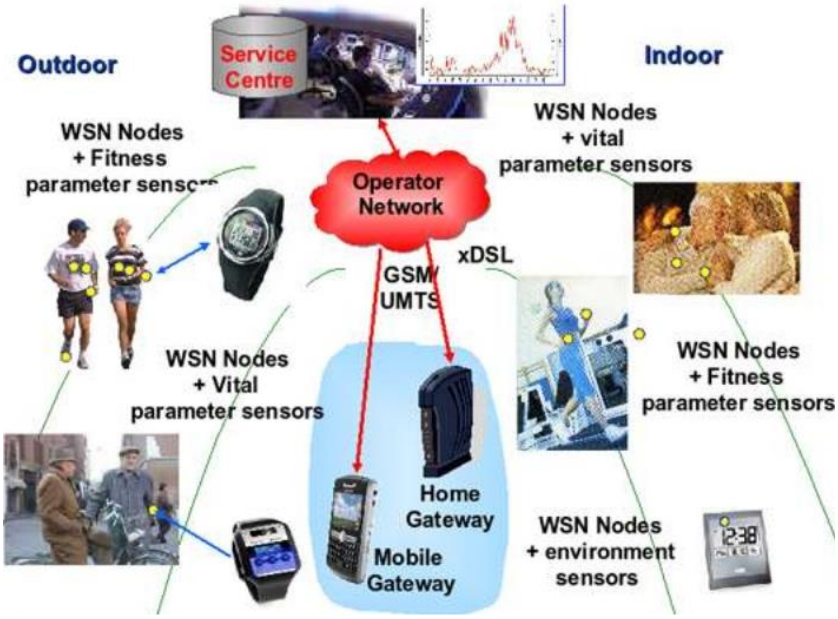

Outdoor
Service Centre
Indoor
WSN Nodes + Fitness parameter sensors
WSN Nodes + vital parameter sensors
Operator Network
xDSL
GSM/ UMTS
WSN Nodes + Vital parameter sensors
WSN Nodes + Fitness parameter sensors
Home Gateway
Mobile Gateway
WSN Nodes + environment sensors

Figure 1.34 Examples of WSN applications deployed in real environment

Examples of application of wireless sensor networks

The ***smart grid*** is an electrical grid that represents a next-generation electrical power system; it maximizes the use of information technology and communications, releases, and energy consumption. The sensors and LR-WPAN devices in terms of IEEE 802.15.4 add to power lines and are used in substations as well. Remote monitoring and managing the smart grid reliability can be maintained with the help of such automation. The combinations of IEEE 802.15.4 and its amendment IEEE 802.15.4g application scenario are shown in Figure 1.35.

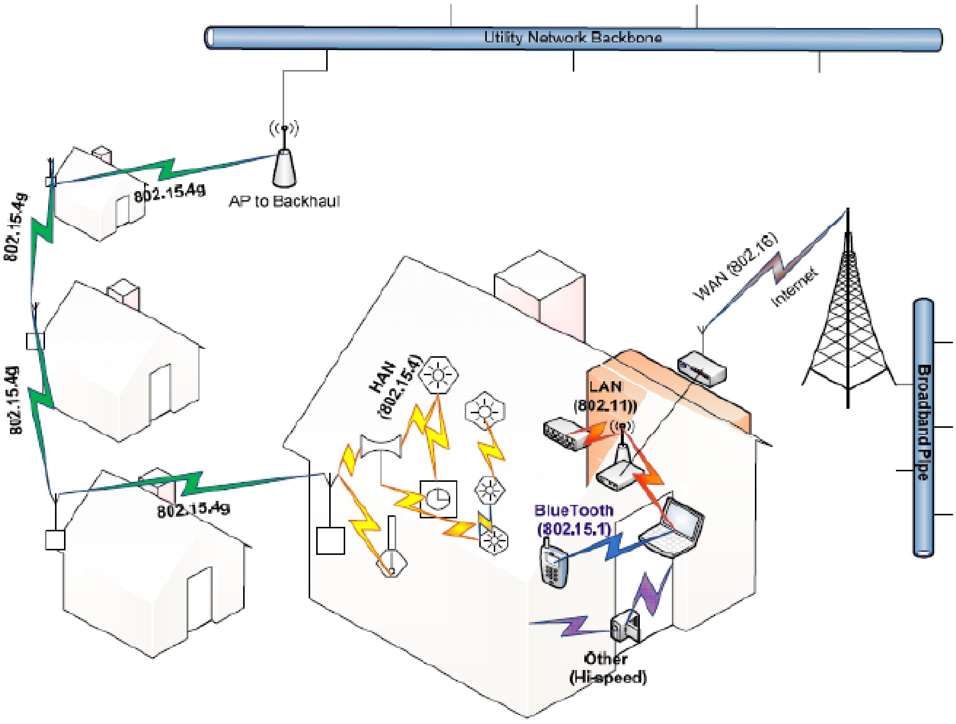


Figure 1.35. Combinations of IEEE 802.15.4 and IEEE 802.16 application scenario

IEEE 802.15.4 LR-WPAN convinces the requirements of wide range of applications. The current industrial market is using a lot LR-WPAN networks. With the use of wireless sensors and LR-WPAN devices together, information is collected, forwarded and analyzed. For example, wireless sensors and LR-WPAN devices can be placed at factories and industries to find out last-minute problems such as fires and risky chemical balances, and also be positioned in machinery to monitor and control regular operations. In short, the emphasis is monitoring in industrial applications where frequent updating and higher throughput is not the case. The longer battery powered devices are the key factor and they can get by on lower power consumption.

Automotive sensing minimizes the cost of wiring installation in vehicles and makes flexible connections. It has many applications, including tire pressure monitoring, smart badges and tags. For tire pressure monitoring, it is required to place four sensors on each tire.

These sensors forward the data or commands about the pressure changes every few miles to a central station, allowing drivers to take necessary steps based on the data. Once these sensors are battery powered, it is a completely wireless communications. There will be fewer bits of data and therefore, the life of battery-powered sensors can be longer. It is good if switching batteries are not necessary before the tires are replaced.

Home automation and networking is the fundamental and primary application of LRWPAN where they perfect fit compared to other IEEE standards or protocols. The star topology is preferable in this network due to presence of 50 to 100 devices. Pure home automation, consumer electronics, PC peripherals, health monitoring, toys and games are considered under this section.

In the home automation example, lighting monitoring, security system control, door and window locks, curtain controls, and HVAC (Heating, Ventilation and Air Conditioning) technology can be taken into account. Remote control, televisions, radios, VCRs, CDs, and DVDs come under the consumer electronics type. LR-WPAN makes an adaptable remote control to handle all of them. PC peripherals can use Bluetooth technology sometimes, but they typically use LR-WPAN devices such as wireless mice, keyboards, joysticks, PDAs with low-end inputs, and multimedia games. The requirements range for latency and throughput for PC peripherals are 15ms to 100ms and 10kbps to 115kbps, thus they are suitable for the LR-WPAN application.

Volcanic monitoring

A wireless sensor network of 16 sensor nodes equipped with seismoacoustic sensors was deployed on Volcan Reventador in northern Ecuador to monitor volcanic eruptions.The network collected seismic and acoustic data on volcanic activity over 3 km and transmitted the collected data through a multi-hop routing and over a long-distance radio link to a base station at the volcano observatory.

Fig. 1.36 illustrates the volcano monitoring sensor network architecture in Volcàn Reventador deployment (Harvard, 2012). The network observed 230 eruptions and other volcanic events over 3 weeks, generating useful data that enabled to evaluate the performance of large-scale sensor networks for collecting high-resolution volcanic data.

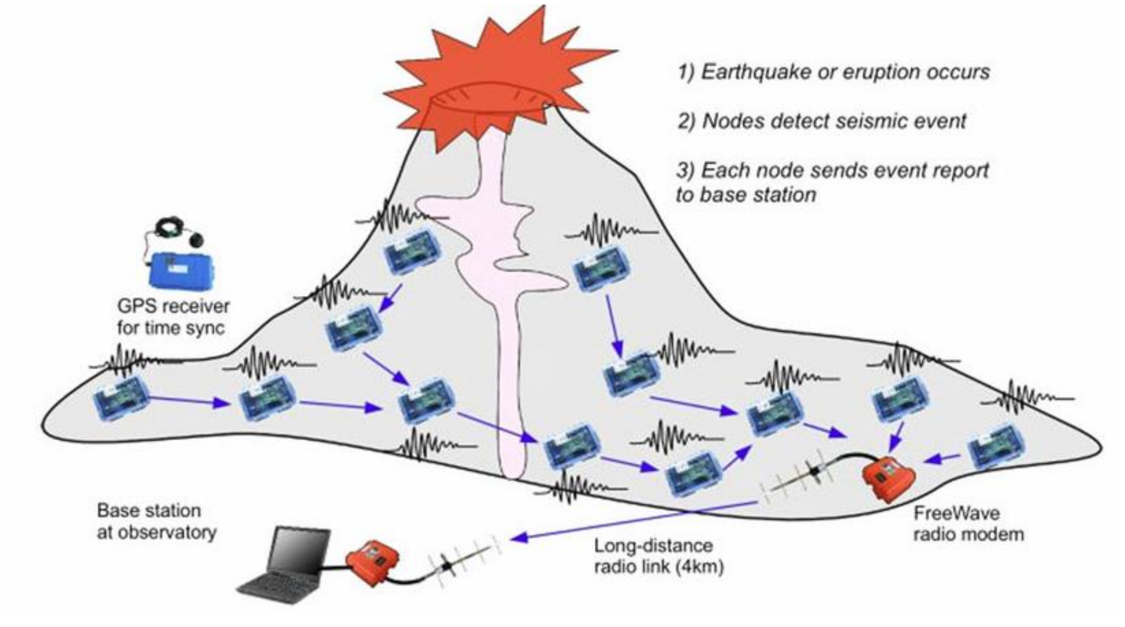

1) Earthquake or eruption occurs
2) Nodes detect seismic event
3) Each node sends event report to base station
GPS receiver for time sync
Base station at observatory
Long-distance radio link (4km)
FreeWave radio modem

Figure 1.36 Volcan Reventador deployment for volcano monitoring (Harvard, 2012)

1.1.6. Factors influencing sensor network design

A sensor network design is influenced by many factors, which include fault tolerance; scalability; production costs; operating environment; sensor network topology; hardware constraints; transmission media; and power consumption. These factors are addressed by many researchers as surveyed in this paper. However, none of these studies has a full integrated view of all factors that are driving the design of sensor networks and sensor nodes.

These factors are important because they serve as a guideline to design a protocol or an algorithm for sensor networks. In addition, these influencing factors can be used to compare different schemes (I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, 2002).

Fault tolerance

Some sensor nodes may fail or be blocked due to lack of power, have physical damage or environmental interference. The failure of sensor nodes should not affect the overall task of the sensor network. This is the reliability or fault tolerance issue. Fault tolerance is the ability to sustain sensor network functionalities without any interruption due to sensor node failures. The reliability $R_k(t)$ or fault tolerance of a sensor node is modelled in (G. Hoblos, M. Staroswiecki, A. Aitouche, 2000) using the Poisson distribution to capture the probability of not having a failure within the time interval (0; t):

$$R_k(t) = \exp(-\lambda_k t) \tag{1.11}$$

where λ_κ and t are the failure rate of sensor node k and the time period, respectively.

Note that protocols and algorithms may be designed to address the level of fault tolerance required by the sensor networks. If the environment where the sensor nodes are deployed has little interference, then the protocols can be more relaxed. For example, if sensor nodes are being deployed in a house to keep track of humidity and temperature levels, the fault tolerance requirement may be low since this kind of sensor networks is not easily damaged or interfered by environmental noise. On the other hand, if sensor nodes are being deployed in a battlefield for surveillance and detection, then the fault tolerance has to be high because the sensed data are critical and sensor nodes can be destroyed by hostile actions. As a result, the fault tolerance level depends on the application of the sensor networks, and the schemes must be developed with this in mind.

Scalability

The number of sensor nodes deployed in studying a phenomenon may be in the order of hundreds or thousands. Depending on the application, the number may reach an extreme value of millions. The new schemes must be able to work with this number of nodes. They must also utilize the high density nature of the sensor networks.

The density can range from few sensor nodes to few hundred sensor nodes in a region, which can be less than 10 m in diameter. The density can be calculated according to (N. Bulusu, D. Estrin, L. Girod, J. Heidemann, 2001) as

$$\mu(R) = (N\pi R^2)/A \tag{1.12}$$

where N is the number of scattered sensor nodes in region A; and R, the radio transmission range.

Basically, $\mu(R)$ gives the number of nodes within the transmission radius of each node in region A.

In addition, the number of nodes in a region can be used to indicate the node density. The node density depends on the application in which the sensor nodes are deployed. For machine diagnosis application, the node density is around 300 sensor nodes in a 5 x 5 m^2 region, and the density for the vehicle tracking application is around 10 sensor nodes per region (E. Shih et al, 2001). In general, the density can be as high as 20 sensor nodes/m^3 (E. Shih et al, 2001). A home may contain around two dozens of home appliances containing sensor nodes, but this number will grow if sensor nodes are embedded into furniture and other miscellaneous items. For habitat monitoring application, the number of sensor nodes ranges from 25 to 100 per region. The density will be extremely high when a person normally containing hundreds of sensor nodes, which are embedded in eye glasses, clothing, shoes, watch, jewellery, and human body, is sitting inside a stadium watching a basketball, football, or baseball game.

Production costs

Since the sensor networks consist of a large number of sensor nodes, the cost of a single node is very important to justify the overall cost of the networks. If the cost of the network is more expensive than deploying traditional sensors, then the sensor network is not cost-justified. As a result, the cost of each sensor node has to be kept low. The state-of-the-art technology allows a Bluetooth radio system to be less than 10 € (E. Shih et al, 2001).

The cost of a sensor node should be in range of euro cents in order for the sensor network to be feasible [70]. The cost of a Bluetooth radio, which is known to be a low-cost device, is even 10 times more expensive than the targeted price for a sensor node. Note that a sensor node also has some additional units such as sensing and processing units as described in Section 3.4. In addition, it may be equipped with a location finding system, mobilizer, or power generator depending on the applications of the sensor networks. As a result, the cost of a sensor node is a very challenging issue given the amount of functionalities.

Hardware constraints

A sensor node is made up of four basic components as shown in Fig. 1.37: a sensing unit, a processing unit, a transceiver unit and a power unit.

They may also have application dependent additional components such as a location finding system, a power generator and a mobilizer. Sensing units are usually composed of two subunits: sensors and analogue to digital converters (ADCs). The analogue signals produced by the sensors based on the observed phenomenon are converted to digital signals by the ADC, and then fed into the processing unit. The processing unit, which is generally associated with a small storage unit, manages the procedures that make the sensor node collaborate with the other nodes to carry out the assigned sensing tasks. A transceiver unit connects the node to the network. One of the most important components of a sensor node is the power unit. Power units may be supported by a power scavenging unit such as solar cells. There are also other subunits, which are application dependent.

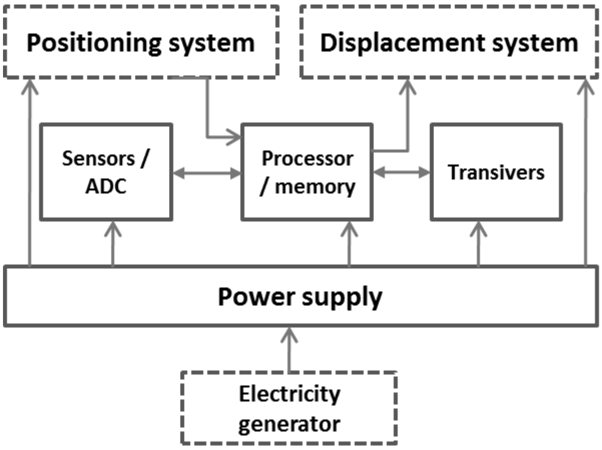


Figure 1.37. The components of a sensor node (adopted from I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, 2002)

All of these subunits may need to fit into a matchbox-sized module. The required size may be smaller than even a cubic centimetre, which is light enough to remain suspended in the air. Apart from the size, there are also some other stringent constraints for sensor nodes. These nodes must:

- consume extremely low power,
- operate in high volumetric densities,
- have low production cost and be dispensable,
- be autonomous and operate unattended,
- be adaptive to the environment.

Power consumption

The wireless sensor node, being a micro-electronic device, can only be equipped with a limited power source (<0.5 Ah, 1.2 V). In some application scenarios, replenishment of power resources might be impossible. Sensor node lifetime, therefore, shows a strong dependence on battery lifetime.

In a multi-hop ad hoc sensor network, each node plays the dual role of data originator and data router. The disfunctioning of few nodes can cause significant topological changes and might require re-routing of packets and re-organization of the network. Hence, power conservation and power management take on additional importance.

It is for these reasons that researchers are currently focusing on the design of power-aware protocols and algorithms for sensor networks.

Communication

Of the three domains, a sensor node expends maximum energy in data communication. This involves both data transmission and reception. It can be shown that for short-range communication with low radiation power (close to 0 dbm), transmission and reception energy costs are nearly the same.

Mixers, frequency synthesizers, voltage control oscillators, phase locked loops (PLL) and power amplifiers, all consume valuable power in the transceiver circuitry. It is important that in this computation we not only consider the active power but also the start-up power consumption in the transceiver circuitry. The start-up time, being of the order of hundreds of micro-seconds, makes the start-up power non-negligible. This high value for the start-up time can be attributed to the lock time of the PLL. As the transmission packet size is reduced, the start-up power consumption starts to dominate the active power consumption. As a result, it is inefficient in turning the transceiver ON and OFF, because a large amount of power is spent in turning the transceiver back ON each time.

In (E. Shih et al, 2001), the authors present a formulation for the radio power consumption (Pc) as

$$P_c = N_T[P_T(T_{on} + T_{st}) + P_{out}(T_{on})] + N_R[R_R(R_{on} + R_{st})] \tag{1.13}$$

$P_{T/R}$ - is the power consumed by the transmitter/receiver;

P_{out} - the output power of the transmitter;

T/R_{on} - the transmitter/receiver on time;

T/R_{st} - the transmitter/receiver start-up time;

$N_{T/R}$ - the number of times transmitter/receiver is switched on per unit time, which depends on the task and medium access control (MAC) scheme used.

Ton can further be rewritten as L/R, where L is the packet size and R, the data rate.

The power consumption in data processing (*Pp*) can be formulated as follows:

$$P_p = CV_{dd}{}^2 f \tag{1.14}$$

V_{dd} - the voltage swing;

C – the total switching capacitance;

f – the switching frequency.

1.1.7. A passive optical network (PON)

A passive optical network (PON) is a system that brings optical fibre cabling and signals all or most of the way to the end user. Depending on where the PON terminates, the system can be described as fibre-to-the-curb (FTTC), fibre-to-the-building (FTTB), or fibre-to-the-home (FTTH) (Optical Network, 2016).

A PON consists of an Optical Line Termination (OLT) at the communication company's office and a number of Optical Network Units (ONUs) near end users. Typically, up to 32 ONUs can be connected to an OLT. The passive simply describes the fact that optical transmission has no power requirements or active electronic parts once the signal is going through the network.

All PON systems have essentially the same theoretical capacity at the optical level. The limits on upstream and downstream bandwidth are set by the electrical overlay, the protocol used to allocate the capacity and manage the connection. The first PON systems that achieved significant commercial deployment had an electrical layer built on Asynchronous Transfer Mode (ATM, or "cell switching") and were called "APON." These are still being used today, although the term "broadband PON" or BPON is now applied. APON/BPON systems typically have downstream capacity of 155 Mbps or 622 Mbps, with the latter now the most common. Upstream transmission is in the form of cell bursts at 155 Mbps.

Multiple users of a PON could be allocated portions of this bandwidth. A PON could also serve as a trunk between a larger system, such as a CATV system, and a neighbourhood, building, or home Ethernet network on coaxial cable.

The successor to APON/BPON is GPON, which has a variety of speed options ranging from 622 Mbps symmetrical (the same upstream/downstream capacity) to 2.5 Gbps downstream and 1.25 Gbps upstream. GPON is also based on ATM transport. GPON is the type of PON most widely deployed in today's fibre-to-the-home (FTTH) networks in new installations and is generally considered suitable for consumer broadband services for the next five to 10 years. From GPON, the future could take two branches: 1) 10 GPON would increase the speed of a single electrical broadband feed to 10G; and 2) WDM-PON would use wavelength-division multiplexing (WDM) to split each signal into 32 branches.

A rival activity to GPON is Ethernet PON (EPON), which uses Ethernet packets instead of ATM cells. EPON should be cheaper to deploy, according to supporters, but it has not garnered the level of acceptance of GPON, so it is not clear how EPON will figure in the future of broadband access.

As an example of PON access solution for enterprises can serve a Campus access network, offered by Huawei (see Fig.1.38). The solution covers the following scenarios on the enterprise campus: enterprise office and production, residential community, video surveillance, video conferencing, wireless coverage, bulletins, and access control systems.

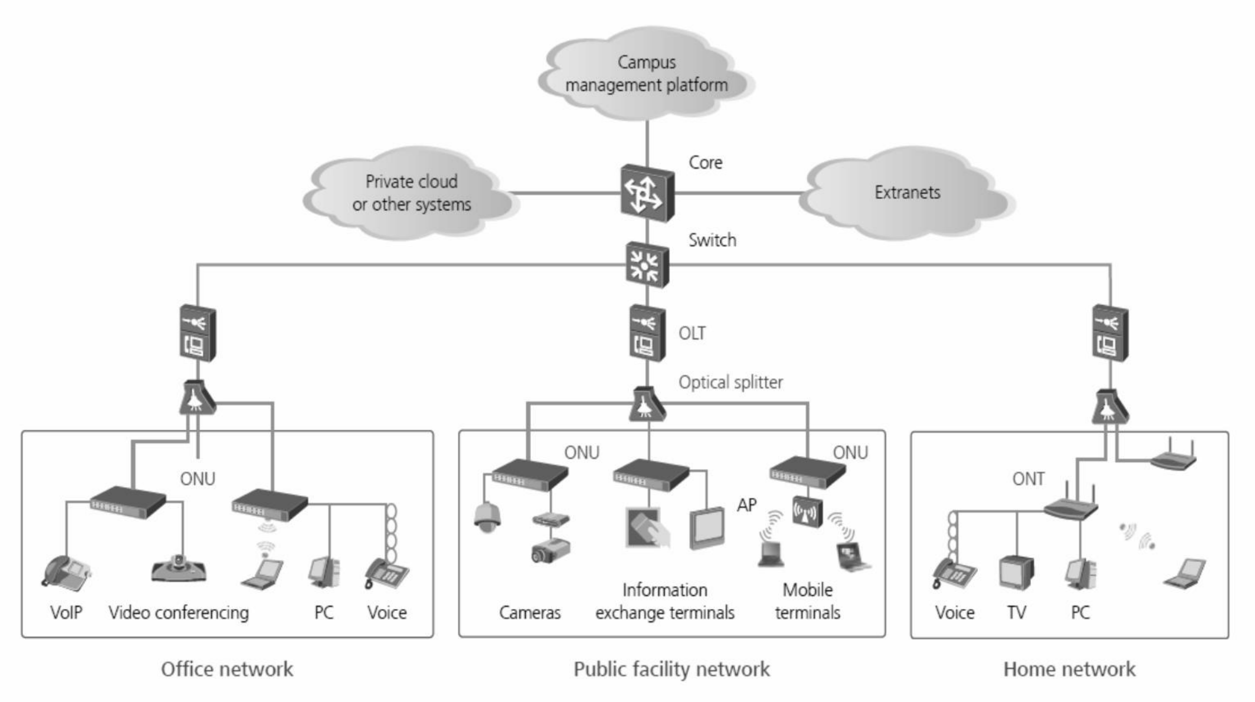


Figure 1.38. Campus access network

Campus access network scenarios can be divided into office network, public facility network, and home network. Office networks use FTTO-type ONUs to carry voice, broadband Internet access, video conferencing, and private line services. Public facility networks use indoor and outdoor ONUs to provide access control, surveillance, and Wi-Fi services. Typical home networks are those that use ONTs to carry triple play services.

1.2. Computer Networks a systems approach

1.2.1. Data to signal conversion in computer network

A *computer network* is used to send information from one point to another. This information needs to be converted to either a digital signal or an analog signal for transmission. Techniques for conversion digital and analog data to digital signal, commonly referred to as *encoding techniques* (Forouzan 2007).

There are two types of information transfer in the network - using signal modulation (passband transmission) or only with the signal conversion encoding (baseband transmission). The encoding includes three known techniques: *line coding, and block coding.*

Data rate and signal rate

Line coding is the process of converting digital data to digital signals. Line coding converts a sequence of bits to a digital signal. At the sender, digital data are encoded into a digital signal; at the receiver, the digital data are recreated by decoding the digital signal.

A data element is the smallest entity that can represent a piece of information: this is the bit. In digital data communication, a signal element carries data elements. A signal element is the shortest unit (timewise) of a digital signal. In other words, data elements are what we need to send; signal elements are what we can send. We define a *ratio r* which is the number of data elements carried by each signal element.

The *data rate* defines the number of data elements (bits) sent in one second. The unit is bits per second (*bps*). The *signal rate* is the number of signal elements sent in one second. The unit is the *baud*. The data rate is sometimes called the bit rate; the signal rate is sometimes called the *pulse rate*, or the *baud rate*. One goal in data communications is to increase the data rate while decreasing the signal rate. Increasing the data rate increases the speed of transmission; decreasing the signal rate decreases the bandwidth requirement. *Bandwidth* or radio bandwidth is a measure of the width of a range of frequencies used for signal transmission.

The relationship between data rate and signal rate depends on the value of *r* and can be expressed as

$$S = c*N*(1/r) \text{ baud}, \tag{1.15}$$

where *N* is the data rate in bit per second (bps); *c* is the case factor, which varies for each case; S is the number of signal elements; and *r* is the number of bits in one period of the signal.

Most digital signals we encounter in real life have an effective bandwidth with finite values. We can say that the baud rate, not the bit rate, determines the required bandwidth for a digital signal. The minimum bandwidth is defined, as

$$B_{min} = c*N*1/r. \tag{1.16}$$

We can solve for the maximum data rate if the bandwidth of the channel is given.

$$N_{\max} = 1/c*B*r. \tag{1.17}$$

No of signal levels

This refers to the number values allowed in a signal, known as signal levels, to represent data. A signal with L levels actually can carry $\log_2(L)$ bits per level. If each level corresponds to one signal element and we assume the average case (c = 1/2), then we have

$$N_{\max} = \frac{B \times r}{c} = 2B \times \log_2 L \,. \tag{1.18}$$

Hence several *Line Encoding Schemes* can be applied for data transmission with the use of different signal parameters.

1.2.2. Line Coding Schemes

The Line Coding Schemes can be classified as shown in Figure 1 (Forouzan B.A., 2007).

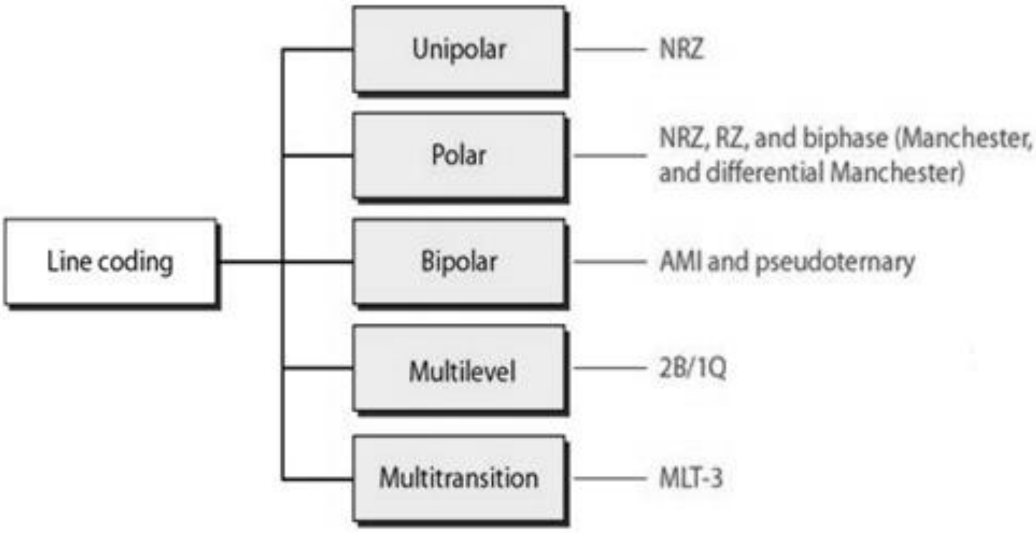


Figure 1.39. Line Coding Schemes

1.2.2.1. One-level Schemes

Unipolar Scheme

In a unipolar scheme, all the signal levels are on one side of the time axis, either above or below. Traditionally, a non-return-to- zero (NRZ) was designed for unipolar scheme. In this scheme a positive voltage defines bit 1 and a zero voltage defines bit 0. It is called NRZ because the signal does not return to zero at the middle of the bit. The following figure shows a unipolar NRZ scheme.

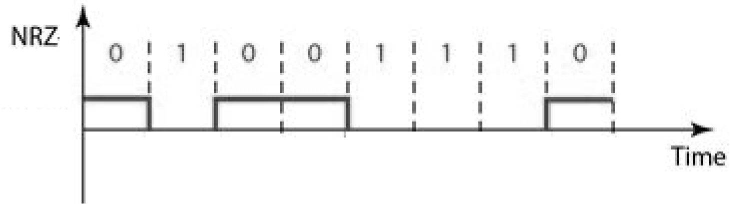


Figure 1.40. NRZ for Unipolar Scheme

Compared with its polar counterpart, the normalized power (power needed to send 1 bit per unit line resistance) is double that for polar NRZ. For this reason, this scheme is normally not used in data communications today.

In polar schemes, the voltages are on the both sides of the time axis. For example, the voltage level for 0 can be positive and the voltage level for 1 can be negative.

Non-Return-to-Zero (NRZ) for polar encoding. In polar NRZ encoding, we use two levels of voltage amplitude. We can have two versions of polar NRZ: NRZ-L and NRZ-I, as shown in Figure 1.41. The figure 1.41 also shows the value of *r*, the average baud rate, and the bandwidth.

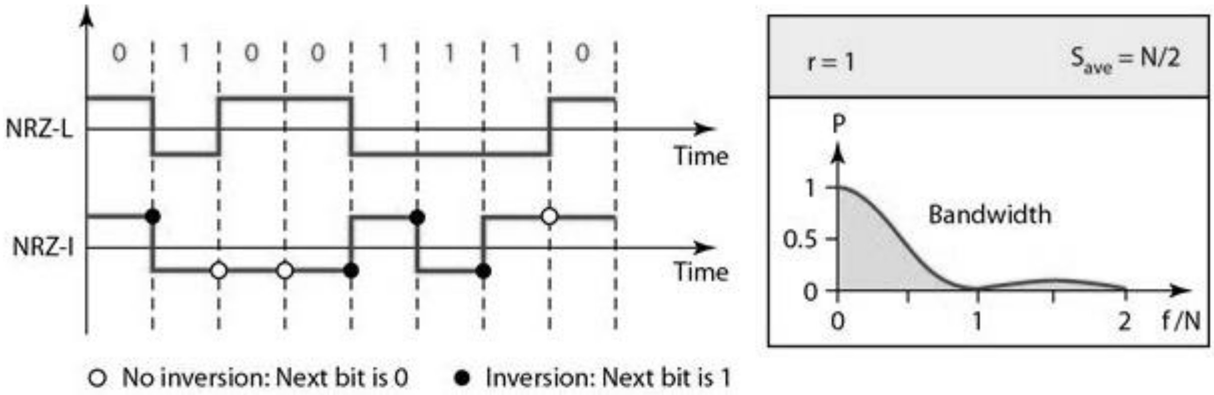


Figure 1.41. NRZ for Unipolar Scheme

In the first variation, NRZ-L (NRZ-Level), the level of the voltage determines the value of the bit. In the second variation, NRZ-I (NRZ-Invert), the change or lack of change in the level of the voltage determines the value of the bit. If there is no change, the bit is 0; if there is a change, the bit is 1.

Drawbacks. If there is a long sequence of 0s or 1s in NRZ-L, the average signal power becomes skewed. The receiver might have difficulty discerning the bit value. In NRZ-I this problem occurs only for a long sequence of 0s. If we eliminate the long sequence of 0s, we can avoid baseline wandering. The synchronization problem (sender and receiver clocks are not synchronized) also exists in both schemes NRZ-L and in NRZ-I schemes. This problem is more serious in NRZ-L than in NRZ-I. While a long sequence of as can cause a problem in both schemes, a long sequence of 1s affects only NRZ-L. Another problem with NRZ-L occurs when there is a sudden change of polarity in the system. For example, if twisted-pair cable is the medium, a change in the polarity of the wire results in all 0s interpreted as 1s and all 1s interpreted as 0s. NRZ-I does not have this problem. Both schemes have an average signal rate of N/2.

Return to Zero (RZ)

The main problem with NRZ encoding occurs when the sender and receiver clocks are not synchronized. The receiver does not know when one bit has ended and the next bit is starting. One solution is the return-to-zero (RZ) scheme, which uses three values: positive, negative, and zero. In RZ, the signal changes not between bits but during the bit. In the following figure, we see that the signal goes to 0 in the middle of each bit. It remains there until the beginning of the next bit (Figure 1.42).

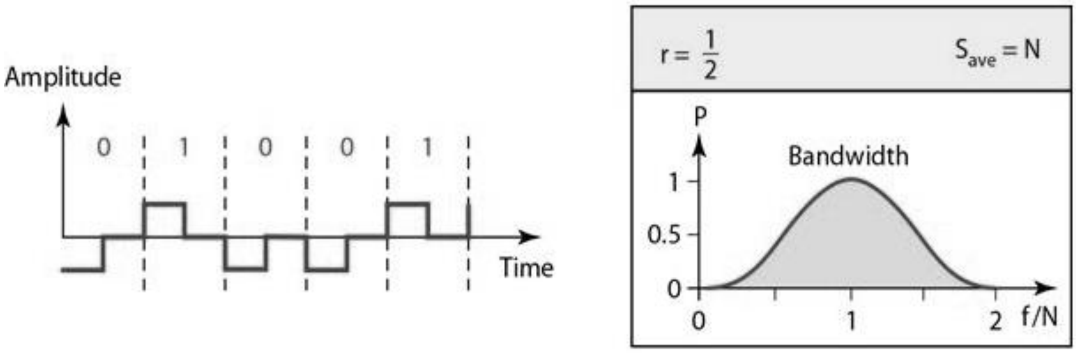


Figure 1.42. Return to Zero Scheme

Drawbacks. The main disadvantage of RZ encoding is that it requires two signal changes to encode a bit and therefore occupies greater bandwidth. Second: sudden change of polarity resulting in all 0s interpreted as 1s and all 1s interpreted as 0s, still exist here, but there is no DC component problem. Another problem is the complexity: RZ uses three levels of voltage, which is more complex to create and discern. As a result of all these deficiencies, the scheme is not used today.

Biphase Manchester and Differential Manchester

The idea of RZ (transition at the middle of the bit) and the idea of NRZ-L are combined into the Manchester scheme.

In Manchester encoding, the duration of the bit is divided into two halves. The voltage remains at one level during the first half and moves to the other level in the second half. The transition at the middle of the bit provides synchronization.

Differential Manchester, on the other hand, combines the ideas of RZ and NRZ-I. There is always a transition at the middle of the bit, but the bit values are determined at the beginning of the bit. If the next bit is 0, there is a transition; if the next bit is 1, there is none. The following figure shows both Manchester and differential Manchester encoding.

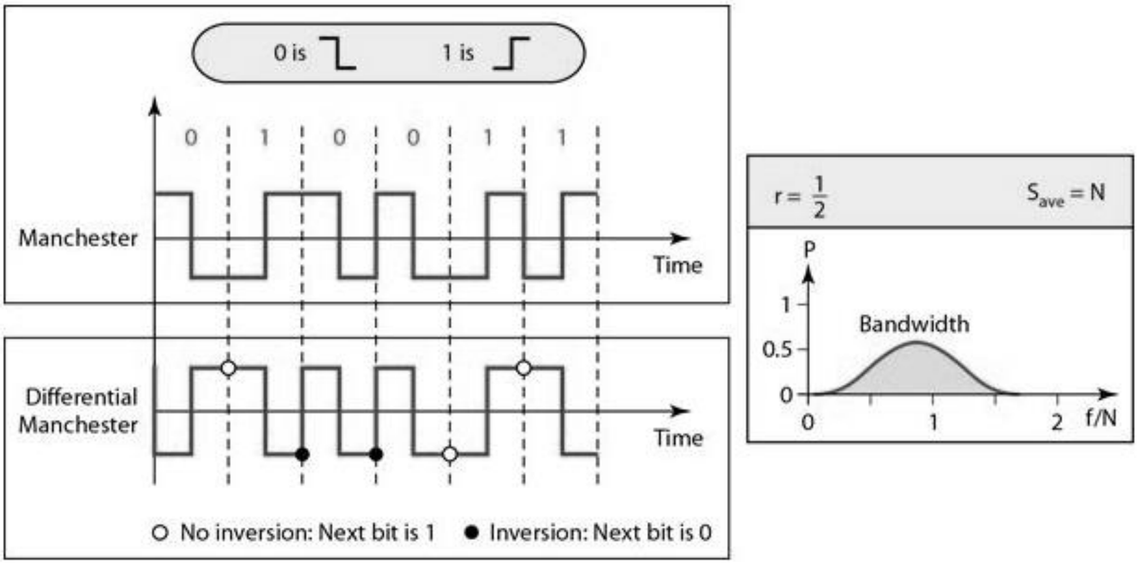


Figure 1.43. *Manchester* Schemes

The Manchester scheme overcomes several problems associated with NRZ-L, and differential Manchester overcomes several problems associated with NRZ-I. First, there is no baseline wandering. There is no DC component because each bit has a positive and negative voltage contribution. The only drawback is the signal rate. The signal rate for Manchester and differential Manchester is double that for NRZ. The reason is that there is always one transition at the middle of the bit and maybe one transition at the end of each bit.

1.2.2.2. Two-level Schemes

In bipolar encoding (sometimes called *Two-level* binary), there are three voltage levels, positive, negative, and zero. The voltage level for one data element is at zero, while the voltage level for the other element alternates between positive and negative.

AMI and Pseudoternary

A common bipolar encoding scheme is called bipolar *alternate mark inversion* (AMI). In alternate mark inversion, a neutral zero voltage represents binary 0. Binary 1s are represented by alternating positive and negative voltages. A variation of AMI encoding is called *Pseudoternary* in which the 1 bit is encoded as a zero voltage and the 0 bit is encoded as alternating positive and negative voltages.

The bipolar schemes were developed as an alternative to NRZ and presented in Figure 6. The bipolar scheme has the same signal rate as NRZ, but there is no DC component. The NRZ scheme has most of its energy concentrated near zero frequency, which makes it unsuitable for transmission over channels with poor performance around this frequency. The concentration of the energy in bipolar encoding is around frequency N/2.

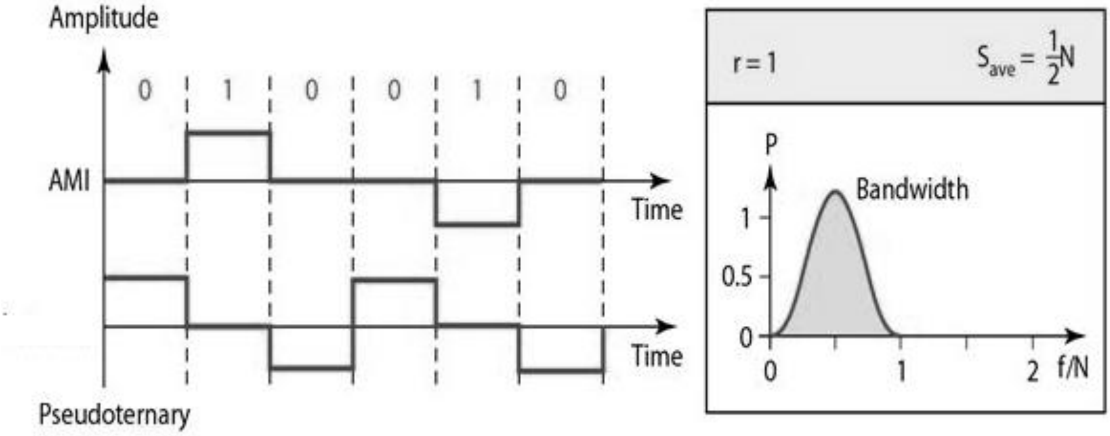


Figure 1.44. AMI and Pseudoternary Schemes

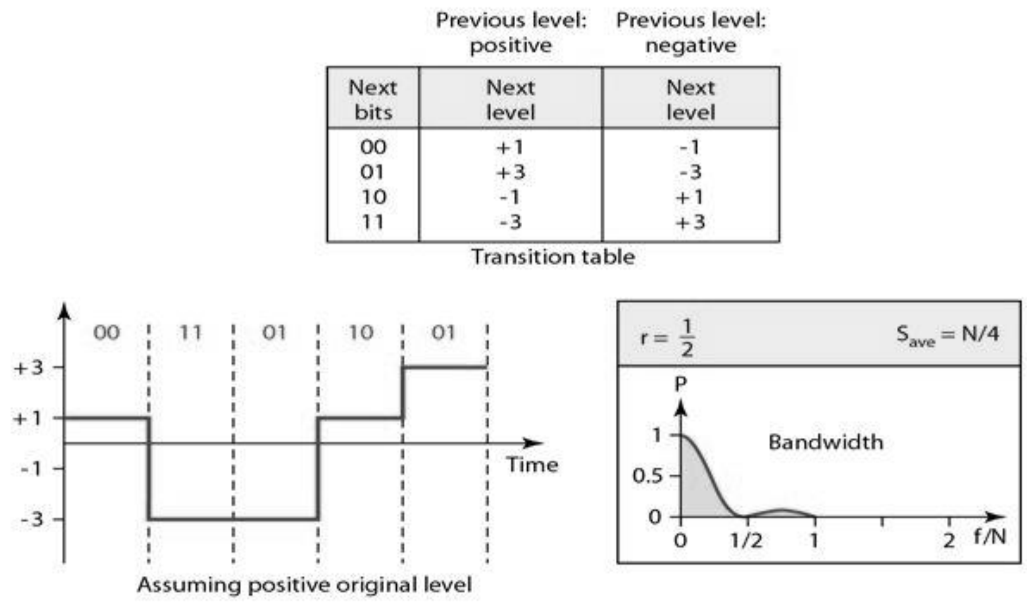
1.2.2.3. Multilevel Schemes

The desire to increase the data speed or decrease the required bandwidth has resulted in the creation of many schemes (Forouzan B.A., 2007). The goal is to increase the number of bits per baud by encoding a pattern of m data elements into a pattern of n signal elements. We only have two types of data elements (0s and 1s), which means that a group of *m* data elements can produce a combination of 2^m data patterns.

We can have different types of signal elements by allowing different signal levels. If we have *L* different levels, then we can produce L^n combinations of signal patterns. If $2^m<L^n$, data patterns occupy only a subset of signal patterns. The subset can be carefully designed to prevent baseline wandering, to provide synchronization, and to detect errors that occurred during data transmission. The code designers have classified these types of coding as *mBnL*, where *m* is the length of the binary pattern, *B* means binary data, *n* is the length of the signal pattern, and L is the number of levels in the signaling. A letter is often used in place of *L*: B(binary) for *L*=2, T (ternary) for *L* =3, and *Q* (quaternary) for *L* =4. Note that the first two letters define the data pattern, and the second two define the signal pattern.

The first *mBnL* scheme we discuss, two binary, one quaternary (2B1Q), uses data patterns of size 2 and encodes the 2-bit patterns as one signal element belonging to a four-level signal. In this type of encoding *m* =2, *n* =1, and *L* =4 (quaternary). The following figure shows an example of a 2B1Q signal.

The average signal rate of 2BlQ is *S* =*N*/4. This means that using 2B1Q, we can send data 2 times faster than by using NRZ-L. However, 2BlQ uses four different signal levels, which means the receiver has to discern four different thresholds.



	Previous level: positive	Previous level: negative
Next bits	Next level	Next level
00	+1	-1
01	+3	-3
10	-1	+1
11	-3	+3

Transition table

Figure 1.45. 2BIQ Scheme

Multiline Transmission

MLT-3. NRZ-I and differential Manchester are classified as differential encoding but use two transition rules to encode binary data (no inversion, inversion). If we have a signal with more than two levels, we can design a differential encoding scheme with more than two transition rules. MLT-3 is one of them. The multiline transmission, three level (MLT-3) scheme uses three levels (+V, 0 and -V) and three transition rules to move between the levels.

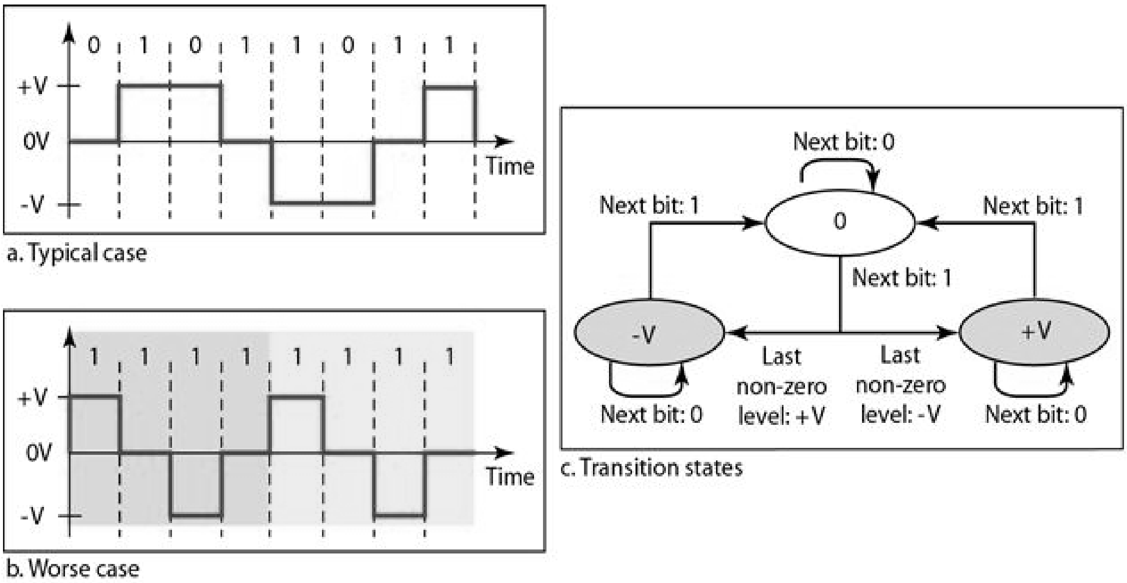
1. If the next bit is 0, there is no transition.

2. If the next bit is 1 and the current level is not 0, the next level is 0.

3. If the next bit is 1 and the current level is 0, the next level is the opposite of the last nonzero level.

The behavior ofMLT-3 can best be described by the state diagram shown in the following figure. The three voltage levels (-V, 0, and +V) are shown by three states (ovals). The transition from one state (level) to another is shown by the connecting lines. The following figure also shows two examples of an MLT-3 signal.

The signal rate is the same as that for NRZ-I, but with greater complexity (three levels and complex transition rules). It turns out that the shape of the signal in this scheme helps to reduce the required bandwidth. Let us look at the worst-case scenario, a sequence of Is. In this case, the signal element pattern +V0 – V0 is repeated every 4 bits.



a. Typical case

b. Worse case

c. Transition states

Figure 1.46. 2BIQ Scheme

Thus, a non-periodic signal has changed to a periodic signal with the period equal to 4 times the bit duration. This worst-case situation can be simulated as an analog signal with a frequency one-fourth of the bit rate. In other words, the signal rate for MLT-3 is one-fourth the bit rate. This makes MLT-3 a suitable choice when we need to send 100 Mbps on a copper wire that cannot support more than 32 MHz (frequencies above this level create electromagnetic emissions).

All discussed above schemes are widely used in data transmission over twisted pairs and copper lines.

The conversion of data elements to signals discussed above represents is the basis of the physical layer of the network communication. Now we want to discuss the properties of logical data structure in networks, which means that blocks of data (called frames), not bit streams, are exchanged between network nodes.

1.2.3. Data Link Control

Data link control deals with the design and procedures for communication between two adjacent nodes: node-to-node communication. Data link control functions include framing, and software implemented protocols that provide smooth and reliable transmission. The data link layer adds a header to the frame to define the addresses of the sender and receiver of the frame. If the rate at which the data are absorbed by the receiver is less than the rate at which data are produced in the sender, the data link layer imposes a flow control mechanism to avoid overwhelming the receiver. In general data transmission over any layer needs a *protocol* or a set of rules how to handle the transmitted data. Protocols are classified as character-oriented (byte-oriented) and bit-oriented protocols. Let us discuss several of them.

1.2.3.1. Byte-Oriented Protocols (BISYNC, PPP)

One of the known approaches to framing it has its roots in connecting terminals to mainframes—is to view each frame as a collection of bytes (characters) rather than a collection of bits. Such a byte-oriented approach is exemplified by older protocols such as the *Binary Synchronous Communication* (BISYNC) protocol developed by IBM in the late 1960s. The more recent and widely used Point-to-Point Protocol (PPP) provides another example of this approach.

Byte-Oriented Protocol BISYNC

Figure 1.47 illustrates the BISYNC protocol's frame format (Peterson L.L. and Davie B.S., 2012). BISYNC uses special characters known as sentinel characters to indicate where frames start and end.

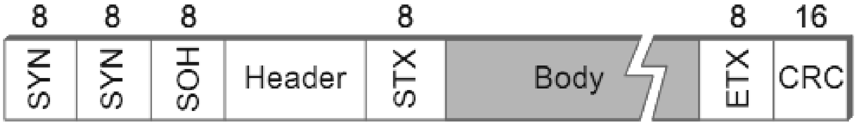


Figure 1.47. Byte-Oriented Protocol BISYNC Frame Format

The beginning of a frame is denoted by sending a special SYN (synchronization) character. The data portion of the frame is then contained between two more special characters: STX (start of text) and ETX (end of text). The SOH (start of header) field serves much the same purpose as the STX field. The problem with the sentinel approach, of course, is that the ETX character might appear in the data portion of the frame. BISYNC overcomes this problem by "escaping" the ETX character by preceding it with a DLE (data-link-escape) character whenever it appears in the body of a frame; the DLE character is also escaped (by preceding it with an extra DLE) in the frame body. This approach is often called character stuffing because extra characters are inserted in the data portion of the frame. Stuffing is necessary when there is a character with the same pattern as a special flag in a frame. The frame format also includes a field labeled CRC (cyclic redundancy check), which is used to detect transmission errors. Finally, the frame contains additional header fields that are used for the link- level reliable delivery algorithm.

Point-to-Point Protocol (PPP)

The more recent Point-to-Point Protocol (PPP), which is commonly used to carry Internet Protocol packets over various sorts of point-to-point links, is similar to BISYNC in that it also uses sentinels and character stuffing. The format for a PPP frame (Peterson L.L. and Davie B.S., 2012) is given in Figure 1.48.

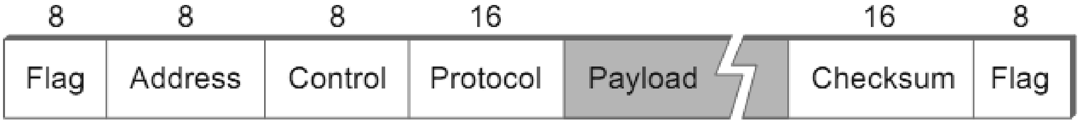


Figure 1.48. Point-to-Point Protocol Frame Format

The special start-of-text character, denoted as the Flag field in Figure 10, is 01111110.

The Address and Control fields usually contain default values and so are uninteresting. The Protocol field is used for demultiplexing; it identifies the high-level protocol such as IP or IPX (an IP-like protocol developed by Novell). The frame payload size can be negotiated, but it is 1500 bytes by default. The Checksum field is either 2 (by default) or 4 bytes long.

The PPP frame format is unusual in that several of the field sizes are negotiated rather than fixed. This negotiation is conducted by a protocol called the Link Control Protocol (LCP). PPP and LCP work in tandem: LCP sends control messages encapsulated in PPP frames—such messages are denoted by an LCP identifier in the PPP Protocol field—and then turns around and changes PPP's frame format based on the information contained in those control messages. LCP is also involved in establishing a link between two peers when both sides detect that communication over the link is possible (e.g., when each optical receiver detects an incoming signal from the fiber to which it connects).

1.2.3.2. Bit-Oriented Protocol (HDLC)

Unlike these byte-oriented protocols, a bit-oriented protocol is not concerned with byte boundaries—it simply views the frame as a collection of bits. These bits might come from some character set, such as ASCII; they might be pixel values in an image; or they could be instructions and operands from an executable file. The Synchronous Data Link Control (SDLC) protocol developed by IBMis an example of a bit-oriented protocol; SDLC was later standardized by the ISO as the High-Level Data Link Control (HDLC) protocol. In the following discussion, we use HDLC as an example (Peterson L.L. and Davie B.S., 2012); its frame format is given in Figure 1.49. HDLC denotes both the beginning and the end of a frame with the distinguished bit sequence 01111110. This sequence is also transmitted during any times that the link is idle so that the sender and receiver can keep their clocks synchronized. In this way, both protocols essentially use the sentinel approach. Because this sequence might appear anywhere in the body of the frame—in fact, the bits 01111110 might cross byte boundaries—bit-oriented protocols use the analog of the DLE character, a technique known as bit stuffing.

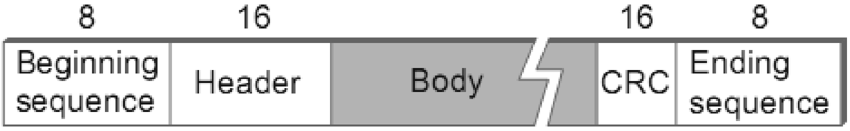


Figure 1.49. High-Level Data Link Control Frame Format

1.2.4. Error and Flow Control

Network is responsible for transmission of data from one device to another device. The end to end data transfer involves many steps, each subject to error. With the error control process, we can be confident that the transmitted and received data are identical. Data can be corrupted during transmission. Error is the process of detecting and correcting both the bit level and packet level errors. Now we consider the simple communication structure: sender, receiver, i.e. the parties that send and receive data, respectively and communication channel. Typical approach to make the reliable communication is to detect an error at the receiver side and notify sender that a message was corrupted. Then sender can retransmit a copy of the message.

1.2.4.1. Error Control

Error detection is the process of detecting the error during the transmission between sender and receiver. Two large classes of error correcting codes exist – convolutional codes and block codes (Sklar B., 2001). We discuss block error correcting codes that are mainly used in a network environment.

The main methods for error detection with the use of block codes include: 1. Parity checking. 2. Cyclic Redundancy Check (CRC). 3. Checksum.

Parity checking

Parity checking is based on linear operations with block codes and adds a single bit that indicates whether the number of one bits in the preceding data is even or odd. If a single bit is changed in transmission, the message will change parity and the error can be detected at this point.

Two-dimensional parity performs calculation of parity check bits for each bit position across each of the bytes contained in the frame. This results in an extra parity byte for the entire frame, in addition to a parity bit for each byte. Figure 1.50 illustrates how two-dimensional even parity works for an example frame containing 6 bytes of data (Peterson L.L. and Davie B.S., 2012). Notice that the third bit of the parity byte is 1 since there is an odd number of 1s in the third bit across the 6 bytes in the frame. It can be shown that two-dimensional parity catches all 1-, 2-, and 3-bit errors, and most 4-bit errors. In this case, we have added 14 bits of redundant information to a 42-bit message, and yet we have stronger protection against common errors. Parity checking is not very robust and often redundant bits more effectively are added in linear codes to control more than one errors in a data block. We define the basic their characteristics below.

Binary linear codes are defined by Hamming distance (minimum number of differing bits) between any binary codewords and a code rate (proportion of information in a block or frame).

Any linear code is completely defined by generator and parity-check matrix whose columns and rows are respectively linearly independent. The error correction capacity of linear error-correcting codes strictly depends on its *minimum distance* (the number of different elements between codewords). Errors in data transmission can be in the form of bit inversion, and bit insertion-deletion leading to the loss of synchronization. However, in the first and second cases an error correction in codeword corrupted by errors with the use of syndrome bounded distance decoding could be applied (Sklar B., 2001).

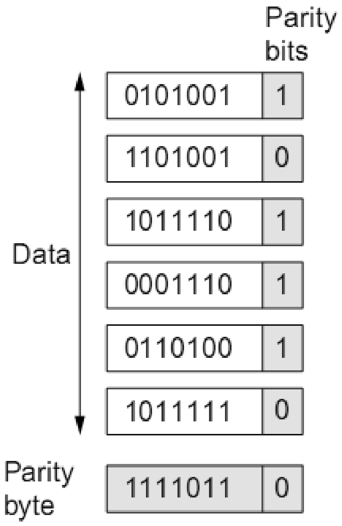


Figure 1.50. Two-dimensional parity checking

Any linear code C is completely defined by its generator matrix or parity-check matrix whose and columns are matrices respectively linearly independent.

Cyclic Redundancy Check (CRC)

Another class of error-correcting (detecting) codes - cyclic codes – is the largest class of linear block codes. They are usually described with the use of polynomials. Generally a binary cyclic code C is defined as a set of finite sequences (vectors) $\mathbf{u} = (u_0 ... u_{n-1})$ called codewords encoded with the use of corresponding message vectors $\mathbf{m} = (m_0 ... m_{k-1})$ from code symbols u_i, $m_i \in GF(2)$. Here, as before, n is a length of a codeword and k – is a length of its information part. A linear block code $C(n,k)$ is cyclic if, for any codeword, obtained by circular shift to the right of an element of its codeword, is also a codeword. Cyclic codes are usually represented as polynomials

of degree $n-1$ or

$$u(x) = u_0 + u_1 \cdot x + u_2 \cdot x^2 + ... + u_{n-1} \cdot x^{n-1} \tag{1.19}$$

From this form, we can define a cyclic code as a set of all polynomials of degree at most n-1, having as the common factor a fixed generator polynomial g(x).

Hence the addition of polynomials corresponds to the modulo 2 (mod2) addition of its coefficients corresponding to the powers of x, and the multiplication is performed in usual way. The only exception that resultant coefficients are obtained with the use of mod2 operation, and the subtraction is replaced by mod2 summation. For example, using a code determined by the generator polynomial $g(x) = 1 + x + x^3$, we can encode an arbitrary sequence, such as m=(0111), which corresponds to a polynomial $m(x)=x+ x^2 + x^3$. We multiply a polynomial $m(x)=x+ x^2 + x^3$ by x^{n-k} and get

$$m(x) \cdot x^{n-k} = m(x) \cdot x^3 = (x + x^2 + x^3) \cdot x^3 = x^4 + x^5 + x^6 . \tag{1.20}$$

Then we divide $m(x){\cdot}x^{n\text{-}k}$ on the generator polynomial $g(x)$ and obtain the remainder of the division $x^2{=}\rho(x)$. With the use of its addition we can write the expression

$\rho(x) + x^{n\text{-}k} \cdot m(x) = q(x){\cdot}\, g(x)$, which shows that polynomial $\rho(x){+}x^{n\text{-}k}{\cdot}m(x)$ is a multiple of g(x) and has a degree of n-1 or less. Consequently, the polynomial $\rho(x){+}x^{n\text{-}k}{\cdot}m(x)$ is a code polynomial corresponding to the encoded message sequence $m(x)$.

Thus, for this example we have $U(x) = 0 \cdot x^0 + 0 \cdot x^1 + 1{\cdot}x^2 + 0 \cdot x^3 + 1{\cdot}x^4 + 1{\cdot}x^5 + 1{\cdot}x^6$.

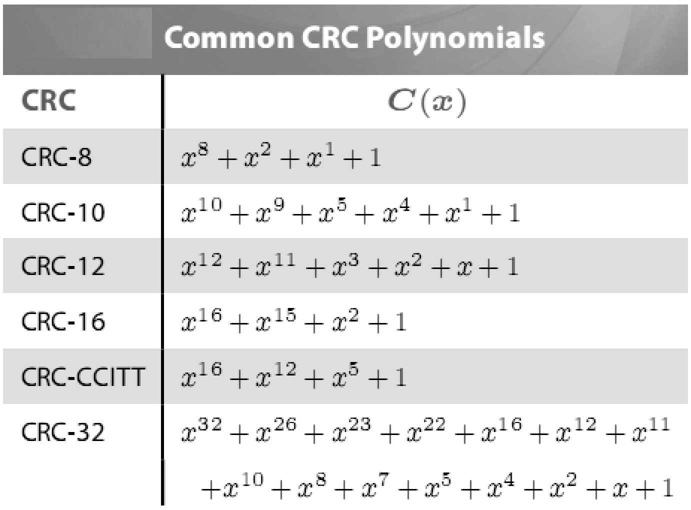
In this example, m = (0111) and the resulting codeword u = (0010111). It is seen that the multiplication carried out is the same as right shift of *m* to *n-k* positions, and the reminder of the division $\rho(x)$=001 is put to the lower positions of a codeword *u.* It is easy to see that a reminder is made from redundant bits of a codeword, can be used for error control, and is called a cyclic redundancy check (CRC). The algorithm for its calculation consists of the following steps:

1. Multiply m(x) by $x^{n\text{-}k}$; that is, add *n-k* zeros at the end of the message and get so-called zero-extended message p(x).
2. Divide p(x) by g(x) and find the remainder.
3. Add the remainder to p(x).

The analysis of encoding - decoding algorithms is convenient to carry out in a number of computing environments such as Matlab, where there are many built-in functions for them (Assanovich B.A. and Kiseleva N.N., 2006).

Finally, we note that the CRC algorithm, while seemingly complex, is easily implemented in hardware using a k-bit shift register and XOR gates. On an Ethernet network (Peterson L.L. and Davie B.S., 2012), for example, a frame carrying up to 12,000 bits (1500 bytes) of data requires only a 32-bit CRC code, or as it is commonly expressed, uses CRC-32. Such a code will catch the majority of errors.

Generally, 6 versions of generator polynomial g(x) are widely used in link-level protocols (shown in Figure 1.51). Ethernet uses CRC-32, while HDLC uses CRC-CCITT. ATM, uses CRC-8, CRC-10, and CRC-32. Remember that the highest order term of the polynomial is not present in the binary number representation, but implied by the algorithm itself.



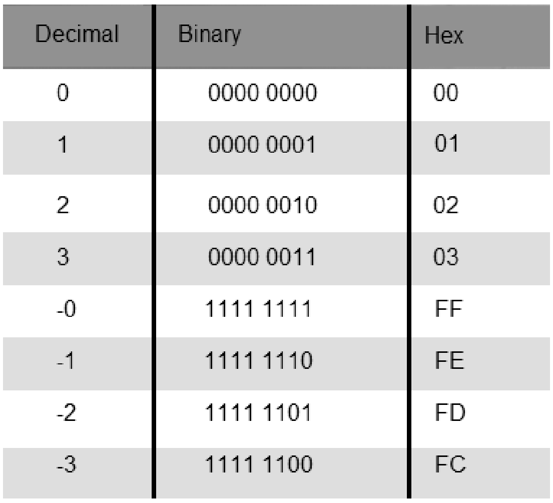
Common CRC Polynomials	
CRC	$C(x)$
CRC-8	$x^8+x^2+x^1+1$
CRC-10	$x^{10}+x^9+x^5+x^4+x^1+1$
CRC-12	$x^{12}+x^{11}+x^3+x^2+x+1$
CRC-16	$x^{16}+x^{15}+x^2+1$
CRC-CCITT	$x^{16}+x^{12}+x^5+1$
CRC-32	$x^{32}+x^{26}+x^{23}+x^{22}+x^{16}+x^{12}+x^{11}$ $+x^{10}+x^8+x^7+x^5+x^4+x^2+x+1$

Figure 1.51. Generator polynomials used in computer networks

Checksum

A second approach to error detection is the use of Internet checksum. Although it is not used at the link level, it nevertheless provides the same sort of functionality as CRCs and parity. The idea behind the Internet checksum is very simple—you add up all the “words” that are transmitted and then transmit the result of that sum. The result is the *checksum*. The receiver performs the same calculation on the received data and compares the result with the received checksum. If any transmitted data, including the checksum itself, is corrupted, then the results will not match, so the receiver knows that an error occurred.

The IP checksum is the 16 bit one's complement of the one's complement sum of all 16 bit words in the header. The 1's complement for fixed point integers (8-bit) is shown in Figure 14. It’s also necessary to remember that in 1's complement we have two zeroes 0 and -0.



Decimal	Binary	Hex
0	0000 0000	00
1	0000 0001	01
2	0000 0010	02
3	0000 0011	03
-0	1111 1111	FF
-1	1111 1110	FE
-2	1111 1101	FD
-3	1111 1100	FC

Figure 1.52. One's complement for fixed point integers

It is seen that we can perform: -3+5=2 or FC+05=01 01. By adding the carry (01) to the Least Significant Bit LSB (01) we can get the expected result: 01+01=02. So, the 1's complement sum is done by summing the numbers and adding the carry (or carries) to the result.

For example, suppose we send the packet FE 05 00, where 00 is the checksum field. Let's calculate and verify the Internet checksum FE + 05=01 03. This is the result of the normal (2's complement) addition. The 1's complement sum requires the addition of the carry to the 8-bit word (even though we will not get the same result) 03+01=04. So the 1's complement sum of FE+05 is 04. The 1's complement (showed with "tilde") of it or the Internet checksum will be ~04=FB and the packet FE 05 FB will be sent. Now, at the receiving end we add all the received bytes, including the checksum (using the 2's complement representation, which is also equivalent to taking the 1's compliment and then adding one) FE+05+FB=01 FE. The 1's complement sum is FE+01=FF=-0 which checks that the transmission was OK.

It is clear that the probability of an undetected error will be reduced with the increase of numbers representation dimension.

1.2.4.2. Flow Control

As we have seen above special error control codes like CRC are used to detect such errors. Even when error-detecting codes are applied some errors will be too severe to be corrected. As a result, some corrupt frames must be discarded. To deliver frames reliably the other mechanism named as "*flow control*" must be applied.

Flow control is an important design issue for the Data Link Layer that controls the flow of data in communication channel between sender and receiver. Sender sends data at higher rate and receives is too sluggish to support that data rate. To solve the above problem, *flow control* is introduced in Data Link Layer (DLL). It also works on several higher layers. The main concept of *flow control* is to introduce the efficiency in computer networks. The approaches of *flow control* realization include: 1) *feedback based flow control;* 2) *rate based flow control.*

In *feedback based flow control*, sender will not send the next data portion until it receives the feedback from receiver. The mechanisms of *feedback based flow* have defined the corresponding protocols: 1) *Stop-and-Wait Protocol;* 2) *Sliding Window Protocol.* The type-1 Protocol has 3 versions:

1. A One-Bit Sliding Window Protocol.
2. A Protocol Using Go Back N.
3. A Protocol Using Selective Repeat.

Stop-and-Wait Protocol

In this protocol there are the following assumptions: 1) It provides unidirectional flow of data from sender to receiver. 2) The communication channel is assumed to be error free.

In this protocol sender simply sends data and waits for the *acknowledgment* from receiver. In data networking, an *acknowledgement* is a signal passed between communicating processes or computers to signify acknowledgement, or receipt of response, as part of a communications protocols. That's why it is called *Stop-and-Wait Protocol.* This type of protocol is not so much efficient, but it is simplest way of *flow control*. A corresponding scheme working according to this protocol is shown in Fig.1.53.

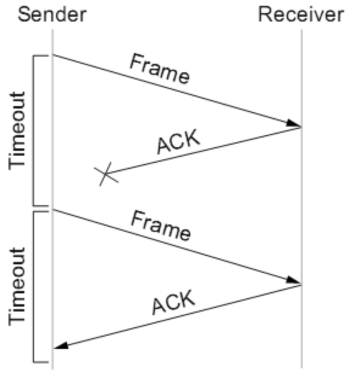


Figure 1.53. Stop-and-Wait Protocol

In this scheme we consider that a communication channel error free, but if *channel* has some errors then receiver is not able to get the correct data from sender so it will not possible for sender to send the next data portion (because it will not get the acknowledgement from receiver). As the result it will interrupt the process of communication. To solve this problem two concepts have been introduced.

1. *Timer*, if sender was not able to get acknowledgment in the particular time then, it sends the buffered data once again to receiver. When sender starts to send data, it starts timer.

2. *Sequence Number*, according to this concept sender sends data with the specific sequence number. Then after receiving data, receiver sends the acknowledgement with that sequence number. Sender is also expected the acknowledgment with the same sequence number.

The problem of this protocol is connected with fact that sender must wait for either positive acknowledgment from receiver or for *Timeout* to send the next frame to receiver. Sender is dependent on receiver. In this protocol only sender sends data and receiver must acknowledge it. The double bandwidth is used.

Sliding Window Protocol

In this protocol, sender and receiver both use buffer of the same size, so it is not necessary to wait for sender to send the second portion of data. It can send packets one after one without waiting for the receiver's acknowledgment, as it is shown in Fig. 1.54.

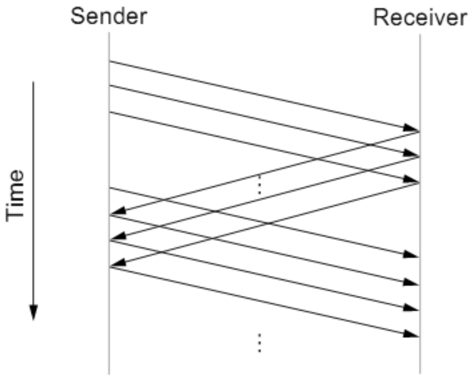


Figure 1.54. Sliding Window Protocol

This fact also solves the problem of increased bandwidth. I this scheme both use the channel for data transmission. Also receiver sends the acknowledgement with data which it wants to transfer to sender. There is no need for special bandwidth to be used for acknowledgment. The process is called piggybacking when both parties send acknowledge and data, i.e. both parties implement *flow control*.

A One-Bit Sliding Window Protocol

This protocol has a buffer size of one bit, so the only possibility for sender and receiver to send and receive packet is only 0 and 1. This protocol includes *Sequence, Acknowledgement,* and *Packet number*. It uses full duplex channel so there is two possibilities:

1. Sender first start sending data and receiver start sending data after it receive data.
2. Receiver and sender both start sending packets simultaneously.

First case is simple and works perfectly, but there will be an error in the second one. That error can be like duplication of the packet, without any transmission error.

A Protocol Using Go Back N

This protocol controls the pipelining process and resolves the problem of resending a corrupted data from the entire transmitted packet set. In this protocol, there are two possibilities at the receiver's side to have a large window size or just a window with size one. Suppose that sender wants to send packets with numbers from 1 to 10 with the use of window with size one at the receiver side (Figure 1.55a). Then if receiver has error in 2nd packet, sender will start retransmission of all packets starting from 0 number packet. Here we assume that sender has the Timeout interval that equals to 8. Timeout will occur after the 8 packets were sent, and only after this time sender will wait for the acknowledgment. The 2nd packet comes to receiver with an error, and all other 8 packets are discarded by it causing the loss of data.

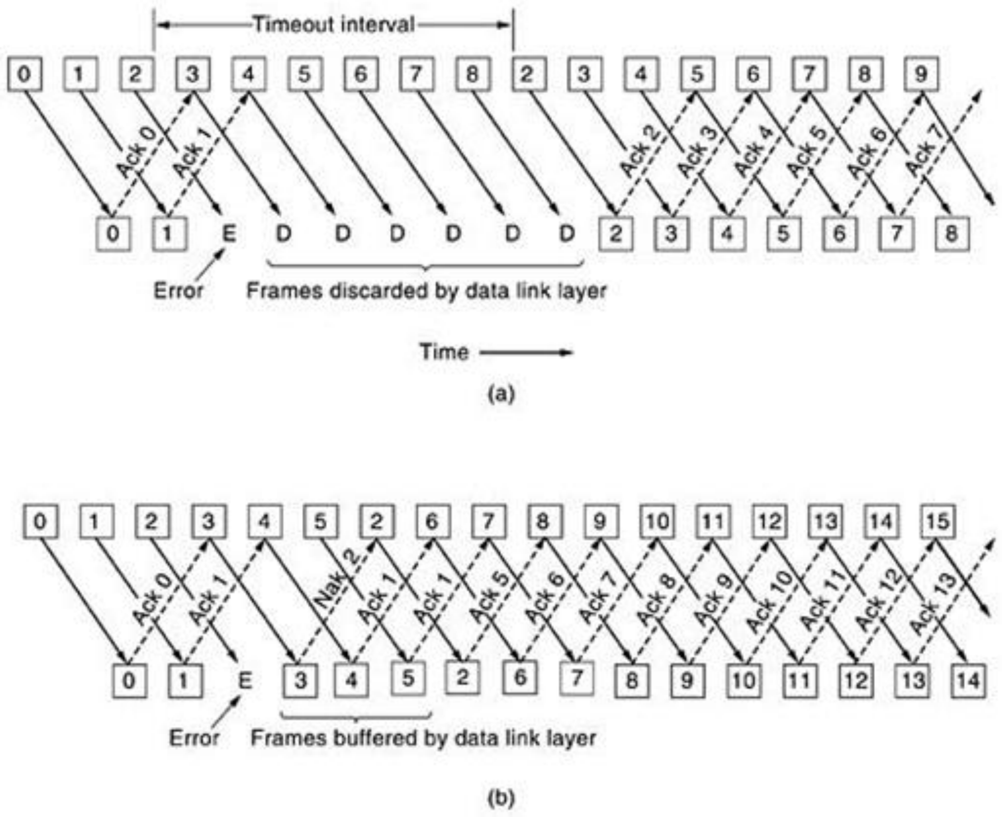


Figure 1.55. Go Back N Protocol (a- size one window; b- large size window)

Now suppose that in the case of large window the 2nd packet arrives with an error from sender. Receiver will accept the 3rd packet but it sends the negative acknowledgment (NAK) with value 2 to sender. Then it buffers the 3rd packet and sequentially buffers the 4th and 5th packet. When sender receives NAK for the 2nd packet it immediately resends it to receiver. After receiving the 2nd packet, receiver sends ACK for the 5th one signaling that it has received all packets up to 5th. So there is no need to resend 3d, 4th and 5th packet again, they are have buffered at the receiver side.

A Protocol Using Selective Repeat

The protocol mechanism “Go back N” is effective when channel quality is good, but when it is poor, it wastes time and bandwidth for frames retransmission. Hence, to provide the reliability, the Selective Repeat Protocol has been introduced. For this method, only specific damaged or lost frame is re-transmitted. In this protocol sender first sets it's window size with 0 and increases to some predefined maximum number. The size of the send window is much smaller. This means less efficiency in filling the pipe, but the fact that there are fewer duplicate frames can compensate for this. Receiver's window size is fixed and equal to the maximum size of sender's window. The receiver has a buffer reserved for the corresponding sequence of packets within its fixed window. Whenever a frame arrives, its sequence number is checked by algorithm to see if it corresponds to the number in the window. If it is true and if it hasn’t been already received, it is accepted and stored. This operation is repeated until it is transferred to the network layer. The difference of *Selective Repeat* protocol to *Go Back N Protocol* is that sender repeats the error frame (number 2 in Figure 16) without sending NAK by receiver, and after filling the buffer, sender retransmits frame that was get with error and receiver acknowledges it.

1.2.5. Random Access

We already have discussed data link control, a mechanism which provides a link with reliable communication. In the protocols we described, we assumed that there is an available dedicated link (or channel) between sender and receiver. This assumption mayor may not be true. Another user may want to use the same channel. That is why we usually consider the data link layer as two sublayers. The upper sublayer is responsible for data link control, and the lower sublayer is responsible for resolving random access to the shared media. The upper sublayer that is responsible for flow and error control is called the logical link control (LLC) layer; the lower sublayer that is mostly responsible for multiple-access resolution is called the media access control (MAC) layer.

The most widespread multiple access protocol is the contention based CSMA/CD protocol used in Ethernet networks. CSMA/CD means *Carrier Sense Multiple Access with Collision Detection*.

The Institute for Electrical and Electronic Engineers (IEEE) developed an Ethernet standard known as IEEE Standard 802.3. This standard also defines a set of rules for CSMA/CD. It uses a *carrier* sensing scheme in which a transmitting data station detects other signals while transmitting a frame, and stops transmitting that frame, transmits a jam signal, and then waits for a random time interval before trying to resend the frame. The *jam signal* or *jamming signal* is a signal that carries a 32-bit binary pattern sent by a data station to inform the other stations of the collision and that they must not transmit.

In other words, CSMA/CD is a set of rules determining how network devices respond when two devices attempt to use a data channel simultaneously (called a *collision*). Standard Ethernet networks use CSMA/CD. This standard enables devices to detect a collision. After detecting a collision, a device waits a random delay time and then attempts to re-transmit the message. If the device detects a collision again, it waits twice as long to try to re-transmit the message.

Ethernet frame structure

Each Ethernet frame is defined by the format (Peterson L.L. and Davie B.S., 2012), given in Figure 1.56.

The 64-bit preamble allows the receiver to synchronize with the signal; it is a sequence of alternating 0s and 1s. Both the source and destination hosts are identified with a 48-bit address. The packet type field serves as the demultiplexing key; it identifies to which of possibly many higher-level protocols this frame should be delivered. Each frame contains up to 1500 bytes of data. Minimally, a frame must contain at least 46 bytes of data, even if this means the host has to pad the frame before transmitting it. The reason for this minimum frame size is that the frame must be long enough to detect a collision. Finally, each frame includes a 32-bit CRC. The Ethernet protocol is a bit-oriented framing protocol.

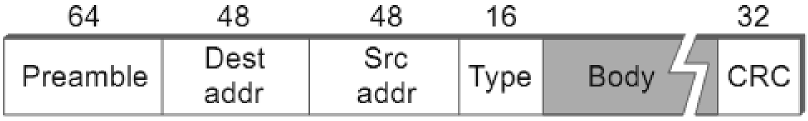


Figure 1.56. Ethernet frame structure

Every Ethernet frame must be at least 512 bits (64 bytes) long and 14 bytes of header plus 46 bytes of data plus 4 bytes of CRC transmitted over the line with length limited to only 2500 m. The reason is to avoid collision and a decrease a signal attenuation and frame round-trip delay. Finally, it is necessary to underline why Ethernets have been so successful. First, an Ethernet is extremely easy to administer and maintain especially when there were no switches in the original Ethernet. Second, it is very inexpensive requiring the network adaptor and cable on each host.

Nowadays there are a large number of Ethernet, depending on the type of physical layer used for communication.

Fast Ethernet standard

The *Fast Ethernet standard* (IEEE 802.3u) has been established for Ethernet networks that need higher transmission speeds. This standard raises the Ethernet speed limit from 10 Mbps to 100 Mbps with only minimal changes to the existing cable structure. There are three types of Fast Ethernet: 100BASE-TX for use with level 5 UTP cable; 100BASE-FX for use with fiber-optic cable; and 100BASE-T4 which utilizes an extra two wires for use with level 3 UTP cable. The 100BASE-TX standard has become the most popular due to its close compatibility with the 10BASE-T Ethernet standard.

Gigabit Ethernet

Gigabit Ethernet was developed to meet the need for faster communication networks with applications such as multimedia and Voice over IP (VoIP). Also known as “gigabit-Ethernet-over-copper” or 1000Base-T, GigE is a version of Ethernet that runs at speeds 10 times faster than 100Base-T. However the CSMA/CD method for gaining access to the physical medium is not employed. 10 Gigabit Ethernet uses multimode optical fiber up to 300 meters and single mode fiber up to 40 kilometers.

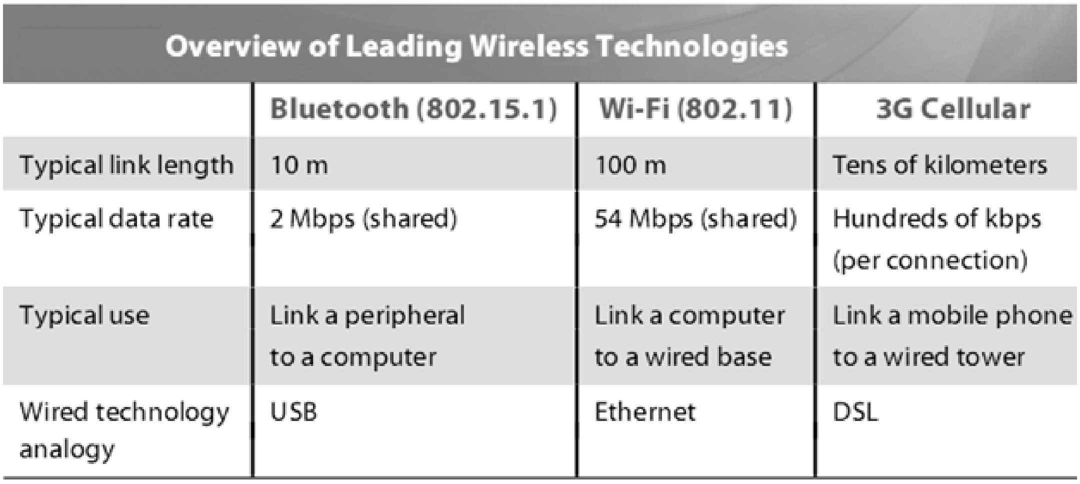
Recently, according to standardization process of IEEE P802.3bs Task Force, the new result was achieved that includes 200Gb/s over single-mode fiber and 400Gb/s over optical physical media. Optical and electrical signaling of 50Gb/s are also being developed to support both 200G and 400G transmission rates (Resources and analysis for electronics engineers, 2016).

1.2.6. Multiuser Access in Networks

As all networks are designed for multiuser communication, a special channel access method must be applied that allows several users to connect to the same multi-point transmission medium, to transmit over it and to share its capacity. A channel-access scheme is based on a *multiplexing* method that allows several data streams or signals to share the same communication channel.

Wireless networks

There is a baffling assortment of different wireless technologies, each of which makes different tradeoffs in various dimensions. One simple way to categorize the different technologies is by the data rates they provide and how far apart communicating nodes can be. In this section, we discuss three prominent wireless technologies: Wi-Fi (more formally known as 802.11), Bluetooth, and the third-generation or “3G” family of cellular wireless standards. Figure 1.57 gives an overview of these technologies and how they compare to each other. These wireless technologies are used for wireless networks of different type.



Overview of Leading Wireless Technologies

	Bluetooth (802.15.1)	Wi-Fi (802.11)	3G Cellular
Typical link length	10 m	100 m	Tens of kilometers
Typical data rate	2 Mbps (shared)	54 Mbps (shared)	Hundreds of kbps (per connection)
Typical use	Link a peripheral to a computer	Link a computer to a wired base	Link a mobile phone to a wired tower
Wired technology analogy	USB	Ethernet	DSL

Figure 1.57. The main wireless network technologies

Despite the widespread of wired data networks, it is often necessary to organize the wireless access to the Internet or to deploy the wireless local area networks. A wireless local area network (WLAN) is a *wireless computer network* that links two or more devices using a wireless distribution method (often *spread-spectrum* or OFDM radio) within a limited area.

Most modern WLANs are based on IEEE 802.11 standards. The IEEE 802.11 standard has two basic modes of operation: *infrastructure* and *ad hoc* mode. In *ad hoc* mode, mobile units transmit directly peer-to-peer. In infrastructure mode, mobile units communicate through an *access point* (AP) that serves as a bridge to other networks.

Infrastructure mode

In infrastructure mode, a base station acts as a *wireless access point* (AP) or *hub,* and nodes communicate through the hub. The hub usually, but not always, has a wired or fiber network connection, and may have permanent wireless connections to other nodes. Wireless access points are usually fixed, and provide service to their client nodes within range.

Peer-to-peer mode

An *ad hoc network* is a network where stations communicate only peer to peer (P2P). There is no base and no one gives permission to talk. In a Wi-Fi P2P group, the group owner operates as an access point and all other devices are clients. There are two main methods to establish a group owner in the Wi-Fi Direct group. In one approach, the user sets up a P2P group owner manually. IEEE 802.11 defines the physical layer (PHY) and MAC (Media Access Control) layers based on *CSMA/CA* (Carrier Sense Multiple Access with Collision Avoidance). *CSMA/CA* is a network *multiple access method* in which *carrier sensing* is used, but nodes attempt to avoid collisions by transmitting only when the channel is sensed to be "idle". When they do transmit, nodes transmit their packet data in its entirety. The 802.11 specification includes provisions designed to minimize collisions, because two mobile units may both be in range of a common access point, but out of range of each other.

It is particularly important for wireless networks, where the collision detection of the alternative CSMA/CD is unreliable due to the *hidden node problem.* In wireless networking, the *hidden node problem* or *hidden terminal problem* occurs when a node is visible from a wireless access point, but not from other nodes communicating with that AP (Peterson L.L. and Davie B.S., 2012) (see Figure 1.58). Each node is within communication range of the AP, but the nodes cannot communicate with each other, as they do not have a physical connection to each other. The problem is when nodes **A** and **C** start to send packets simultaneously to the access point **B**. Because the nodes **A** and **C** are out of range of each other and so cannot detect a collision while transmitting. Carrier sense multiple access with collision detection (CSMA/CD) does not work, and collisions occur, which then corrupt the data received by the access point.

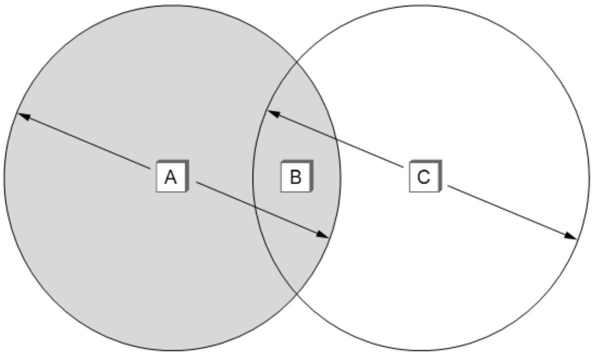


Figure 1.58. The hidden node problem

CSMA/CA is a protocol that also operates in the Data Link Layer (Layer 2**).** *Carrier Sense* in protocol means that prior to transmitting, a node first listens to the shared medium (such as listening for wireless signals in a wireless network) to determine whether another node is transmitting or not. *Collision Avoidance* is a mechanism for the node to stop transmitting when another node was heard, and wait for a period of time (usually random) before listening again for a free communications channel. It is particularly beneficial in a *wireless LAN* due to a common problem of multiple stations being able to see the Access Point, but not each other. This is due to differences in transmission power, and receive sensitivity, as well as distance, and location with respect to the AP. This will cause a station to not be able to 'hear' another station's broadcast. Devices utilizing 802.11 based standards can use the collision avoidance with RTS / CTS (Ready to Send-Clear to Send) handshake and *Point coordination function* (PCF), although they do not do so by default. By default they use a Carrier sensing mechanism called *exponential backoff'*, or Distributed coordination function (DCF) that relies upon a station attempting to 'listen' for another station's broadcast before sending. In a variety of computer networks the *binary exponential backoff* refers to an algorithm used to space out repeated *retransmissions* of the same block of data trying to chose a random number of slot times for retransmission after several collisions. However, because of the hidden terminal problem, just waiting for the absence of signals from other transmitters does not guarantee that a collision will not occur from the perspective of the receiver. CA, or PCF relies upon the AP (or receiver for Ad hoc networks) granting a station the exclusive right to transmit for a given period of time after requesting it (Request to Send / Clear to Send). This goes some way toward addressing the hidden terminal problem. The sender sends an RTS—a short packet—to the intended receiver, and if that packet is received successfully the receiver responds with another short packet, the CTS. Even though the RTS may not have been heard by a hidden terminal, the CTS probably will be. This effectively tells the nodes within range of the receiver that they should not send anything for a while—the amount of time of the intended transmission is included in the RTS and CTS packets. After that time plus a small interval has passed, the carrier can be assumed to be available again, and another node is free to try to send. The amount of time a given node delays is defined by an *exponential backoff algorithm.*

Frame 802.11 Format

Most of the 802.11 frame format (Peterson L.L. and Davie B.S., 2012), which is depicted in Figure 1.59. The first field of the frame having the main informative load, shown in more detail.

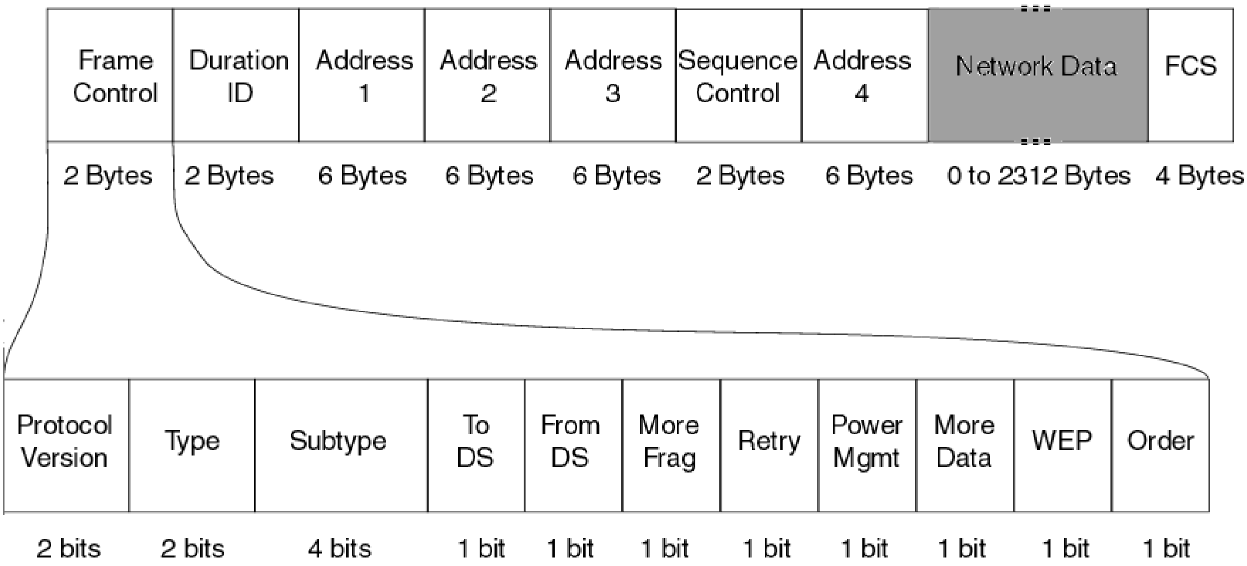


Figure 1.59. Frame 802.11 Format

The frame contains the source and destination node addresses, each of which is 48 bits long; up to 2312 bytes of data; and a 32-bit CRC. The Frame Control field contains three subfields of interest (not shown): a 6-bit Type field that indicates whether the frame carries data, is an RTS or CTS frame, or is being used by the scanning algorithm, and a pair of 1-bit field s—called ToDS and FromDS—that are described below. The peculiar thing about the 802.11 frame format is that it contains four, rather than two, addresses. How these addresses are interpreted depends on the settings of the ToDS and FromDS bits in the frame's Control field. This is to account for the possibility that the frame had to be forwarded across the distribution system, which would mean that the original sender is not necessarily the same as the most recent transmitting node. Similar reasoning applies to the destination address. In the simplest case, when one node is sending directly to another, both the DS bits are 0, Addr1 identifies the target node, and Addr2 identifies the source node. In the most complex case, both DS bits are set to 1, indicating that the message went from a wireless node onto the distribution system, and then from the distribution system to another wireless node. With both bits set, Addr1 defines the ultimate destination, Addr2 defines the immediate sender (the one that forwarded the frame from the distribution system to the ultimate destination), Addr3 identifies the intermediate destination (the one that accepted the frame from a wireless node and forwarded it across the distribution system), and Addr4 identifies the original source. The Frame Check Sequence (FCS) is the last four bytes in the standard 802.11 frame. Often referred to as the Cyclic Redundancy Check (CRC), it allows for integrity check of retrieved frames. As frames are about to be sent, the FCS is calculated and appended. When a station receives a frame, it can calculate the FCS of the frame and compare it to the one received. If they match, it is assumed that the frame was not distorted during transmission.

Bluetooth

Bluetooth fills the niche of very short range communication between peripheral devices (mobile phones, computers, etc.). It is a wireless technology standard for exchanging data over short distances in the Industrial, Scientific and Medical (ISM) band from 2.4 to 2.485 GHz) from fixed and mobile devices and building *personal area networks* (PANs). Bluetooth is managed by the *Bluetooth Special Interest Group* (SIG), which has more than 25,000 member companies in the areas of telecommunication, computing, networking, and consumer electronics (Resources and analysis for electronics engineers, 2016). The IEEE standardized Bluetooth as IEEE 802.15.1, but no longer maintains the standard.

Bluetooth operates at frequencies between 2402 and 2480 MHz, or 2400 and 2483.5 MHz including *guard bands* 2 MHz wide at the bottom end and 3.5 MHz wide at the top. Bluetooth uses a radio technology called *frequency-hopping spread spectrum.* Bluetooth divides transmitted data into packets and transmits each packet on one of 79 designated Bluetooth channels. Each channel has a bandwidth of 1 MHz. It usually performs 800 hops per second, with Adaptive Frequency-Hopping (AFH) enabled. Frequency hopping is a spread spectrum technique that involves transmitting the signal over a random sequence of frequencies; that is, first transmitting at one frequency, then a second, then a third, and so on. AFH (as used in Bluetooth) improves resistance to *radio frequency interference* by avoiding crowded frequencies in the hopping sequence. The sequence of frequencies is not truly random but is instead computed algorithmically by a pseudorandom number generator. The receiver uses the same algorithm as the sender and initializes it with the same seed; hence, it is able to hop frequencies in sync with the transmitter to correctly receive the frame.

The format originally chosen for Bluetooth in version 1 in basic rate (BR) mode has instantaneous data rate of 1 Mbit/s is possible. The term Enhanced Data Rate (EDR) achieves two data rates 2 and 3 Mbit/s. The combination of these (BR and EDR) modes in Bluetooth radio technology is classified as a "BR/EDR radio". The Bluetooth modulation schemes and the general format do not lend themselves to carrying higher data rates. For Bluetooth 3, the higher data rates are not achieved by changing the format of the Bluetooth modulation, but by working cooperatively with an IEEE 802.11g physical layer. In this way data rates of up to around 25 Mbps can be achieved.

Bluetooth is a packet-based protocol with a *master-slave* structure. One *master* may communicate with up to seven *slaves* in a piconet. All devices share the master's clock. Packet exchange is based on the basic clock, defined by the master. The devices can switch roles, by agreement, and the slave can become the master (for example, a headset initiating a connection to a phone necessarily begins as master—as initiator of the connection—but may subsequently operate as slave).

Bluetooth data transfer can be achieved using a variety of different data packet types and using different forms of links - asynchronous links and synchronous links. These different Bluetooth data file transfer formats provide flexibility, but they are invisible to the user who sees a connection being made and Bluetooth data being transferred.

The main types of links include: Asynchronous Connectionless communications Link (ACL); Synchronous Connection Orientated communications link (SCO).

The ACL Bluetooth link is used for carrying framed data - i.e. data submitted from an application to logical link control and adaptation protocol channel. The channel may support either unidirectional or bidirectional Bluetooth data transfer. There is a variety of different ACL formats that can be used - most of them incorporate forward error coding, FEC as well as header error correction to detect and correct errors that may occur in the radio link. We have already defined the main classes of FEC and algorithms for error control above. The ACL is enables data to be transferred via Bluetooth 1 at speeds up to the maximum rate of 732.2 kbps. When using Bluetooth 2 enhanced data rate, data rates of 2.1 Mbps may be achieved.

The SCO or Synchronous Connection Orientated communications link is used where data is to be streamed rather than transferred in a framed format. The idea of the SCO is to ensure that audio data can be streamed without suffering delays waiting for frames or packet slots to become available. A further form of link known as an eSCO or Extended SCO was introduced with version 1.2 of the Bluetooth standard with opportunity to send an acknowledgement and allow a limited number of re-transmissions if data is corrupted.

To use Bluetooth wireless technology, a device must be able to interpret certain Bluetooth profiles, which are definitions of possible applications. Two of them are fundamental to ensuring interoperability between BLE devices from different vendors:

Generic Access Profile (GAP). Covering the usage model of the lower-level radio protocols to define roles, procedures, and modes that allow devices to broadcast data, discover devices, establish connections, manage connections, and negotiate security levels, GAP is, in essence, the topmost control layer of BLE. This profile is mandatory for all Bluetooth Low Energy (BLE) devices, and all must comply with it.

Generic Attribute Profile (GATT). Dealing with data exchange in BLE, GATT defines a basic data model and procedures to allow devices to discover, read, write, and push data elements between them. It is, in essence, the topmost data layer of BLE.

The Bluetooth Core Specification provides for the connection of two or more piconets to form a scatternet, in which certain devices simultaneously play the master role in one piconet and the slave role in another as it is depicted for node 4 in piconets A and B in Figure 1.60.

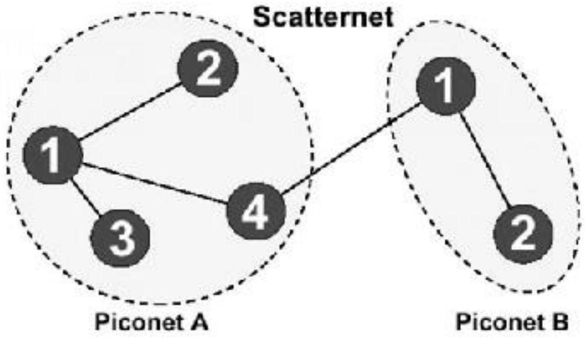


Figure 1.60. Scatternet structure

Bluetooth has been constantly evolving since it was conceived in 1994. The most recent update of Bluetooth, Bluetooth v4.0, is just beginning to gain traction in the consumer electronics industry, but some of the previous versions are still widely used.

The v1.x releases laid the groundwork for the protocols and specifications future versions would build upon. Bluetooth v1.2 was the latest and most stable 1.x version. These modules are rather limited compared to later versions. The 2.x versions of Bluetooth introduced *enhanced data rate* (EDR)**,** which increased the data rate potential up to 3 Mbps (closer to 2.1 Mbps in practice). Bluetooth v2.1, released in 2007, introduced *secure simple pairing* (SSP)**,** which overhauled the pairing process. It provides more information during the inquiry procedure to allow better filtering of devices before connection; and sniff subrating, which reduces the power consumption in low-power mode. Version 3.0 + HS of the Bluetooth Core Specification provides theoretical data transfer speeds of up to 24 Mbit/s, though not over the Bluetooth link itself.

The Bluetooth SIG completed the Bluetooth Core Specification version 4.0 (called Bluetooth Smart) and has been adopted as of 30 June 2010. It includes *Classic Bluetooth*, *Bluetooth high speed* and *Bluetooth low energy* protocols. Bluetooth high speed is based on Wi-Fi, and Classic Bluetooth consists of legacy Bluetooth protocols (Resources and analysis for electronics engineers, 2016).

Bluetooth low energy, previously known as Wibree, is a subset of Bluetooth v4.0 with an entirely new protocol stack for rapid build-up of simple links. In late 2011, new logos

"Bluetooth Smart Ready" for hosts and "Bluetooth Smart" for sensors have been introduced. The Bluetooth versions 4.1 and 4.2 introduced several technical features that improve consumer usability.

Bluetooth 5 was announced in June 2016 (Bluetooth, 2016). It will quadruple the range, double the speed, and provide an eight-fold increase in data broadcasting capacity of low energy Bluetooth connections, in addition to adding functionality for connectionless services like location-relevant information and navigation.

Bluetooth 5, projected for release in late 2016 to early 2017, will quadruple range and double speed of low energy connections while increasing the capacity of connectionless data broadcasts by 800 percent.

Cellular telephone technology

While cellular telephone technology had its beginnings around voice communication, data services based on cellular standards have become increasingly popular (thanks in part to the increasing capabilities of mobile phones or smartphones). One drawback compared to the technologies just described has tended to be the cost to users, due in part to cellular's use of licensed spectrum. The frequency bands that are used for cellular telephones (and now for cellular data) vary around the world. In Europe, for example, the main bands for cellular phones are at 900 MHz and 1800 MHz. In North America, 850-MHz and 1900-MHz bands are used. This global variation in spectrum usage creates problems for users who want to travel from one part of the world to another and has created a market for phones that can operate at multiple frequencies (e.g., a tri-band phone can operate at three of the four frequency bands mentioned above). That problem, however, pales in comparison to the proliferation of incompatible standards that have plagued the cellular communication business.

A cellular network or mobile network is a communication network where the last link is wireless. The network is distributed over land areas called cells, each served by at least one fixed-location *transceiver,* known as a *cell site* or *base station*. This base station provides the cell with the network coverage which can be used for transmission of voice, data and others.

Modern mobile phone networks use cells because radio frequencies are a limited, shared resource. Cell-sites and handsets change frequency under computer control and use low power transmitters so that the usually limited number of radio frequencies can be simultaneously used by many callers with less interference.

A cellular network is used by the *mobile phone operator* to achieve both coverage and capacity for their subscribers. Large geographic areas are split into smaller cells to avoid line-of-sight signal loss and to support a large number of active phones in that area. All of the cell sites are connected to *telephone exchanges* (or switches), which in turn connect to the *public telephone network*.

There are a number of different digital cellular technologies, including: *Global System for Mobile Communications* (GSM), *General Packet Radio Service* (GPRS), CDMA2000, Evolution-Data Optimized (EV-DO), Enhanced Data Rates for GSM Evolution (EDGE), Universal Mobile Telecommunications System (UMTS), Digital Enhanced Cordless Telecommunications DECT, etc.

The cellular mobile-radio network consists of the following structures:

- A network of radio base stations forming the *base station subsystem.*
- The *core circuit switched network* for handling voice calls and text.
- A *packet switched network* for handling mobile data.
- The *public switched telephone network* (PSTN) to connect subscribers to the wider telephony network.

GSM network is the most usable is wireless network. Despite the fact that the generation of cellular communication systems are quickly replaced the *third generation* (3G) systems are still widely used in practice. *3G* is the third generation of wireless mobile telecommunications technology. It is based on a set of standards used for mobile devices and mobile telecommunications services and networks that comply with the International Mobile Telecommunications-2000 (IMT-2000) specifications by *the International Telecommunication Union* (ITU). 3G finds application in wireless voice telephony, *mobile Internet* access, fixed wireless Internet access, video calls and mobile TV. 3G telecommunication networks support services that provide an information transfer rate of at least 200 kbit/s. Later 3G releases, often denoted *3.5G* and *3.75G*, also provide *mobile broadband* access of several Mbit/s to *smartphones* and *mobile modems* in laptop computers. This ensures it can be applied to wireless voice telephony, mobile Internet access, fixed wireless Internet access, video calls and mobile TV technologies. The following standards are typically branded 3G. The *Universal Mobile Telecommunications Service* (UMTS) system, first offered in 2001, standardized by 3GPP, used primarily in Europe and other regions. The cell phones are typically UMTS and GSM hybrids. Several radio interfaces are offered, sharing the same infrastructure: the original and most widespread radio interface is called *W-CDMA* (Wideband Code Division Multiple Access); the latest UMTS release *Evolved High Speed Packet Access* **(**HSPA+) that can provide peak data rates up to 56 Mbit/s in the downlink in theory (28 Mbit/s in existing services) and 22 Mbit/s in the uplink. Really it can achieve data rates of up to 42.2 Mbit/s. (HSPA is an amalgamation of several upgrades to the original W-CDMA standard and offers speeds of 14.4 Mbit/s down and 5.76 Mbit/s up). HSPA+ introduces antenna array technologies such as *beamforming* and *Multiple-input multiple-output communications* (MIMO). Beam forming focuses the transmitted power of an antenna in a beam towards the user's direction. MIMO uses multiple antennas at the sending and receiving side.

The above systems and radio interfaces are based on spread spectrum radio transmission technology. While the GSM EDGE standard ("2.9G"), DECT cordless phones also fulfill the IMT-2000 requirements and are approved as 3G standards by ITU, these are typically not branded 3G, and are based on completely different technologies. However, the following common standards comply with the IMT2000/3G standard: EDGE, that is a revision by the 3GPP organization to the older 2G GSM based transmission methods, utilizing the same switching nodes, base station sites and frequencies as GPRS, but new base station and cellphone

RF circuits. EDGE is still used extensively due to its ease of upgrade from existing 2G GSM infrastructure and cell-phones.

The Universal Mobile Telecommunications System, has been created and revised by the 3GPP. The family is a full revision from GSM in terms of encoding methods and hardware, although some GSM sites can be retrofitted to broadcast in the UMTS/W-CDMA format. W-CDMA is the most common deployment, commonly operated on the 2,100 MHz band. A few others use the 850, 900 and 1,900 MHz bands.

ITU has not provided a clear definition of the data rate that users can expect from 3G equipment or providers. Thus users sold 3G service may not be able to point to a standard and say that the rates it specifies are not being met. While stating in commentary that "it is expected that IMT-2000 will provide higher transmission rates: a minimum data rate of 2 Mbit/s for stationary or walking users, and 384 kbit/s in a moving vehicle".

The Long Term Evolution of the 3G services, eyes are now turning towards the next development, that of the truly 4G technology named IMT Advanced. The new technology being developed under the auspices of 3GPP to meet these requirements is often termed as fourth generation 4G or LTE Advanced. 4G is the fourth generation of wireless mobile telecommunications technology, succeeding 3G. A 4G system must provide capabilities defined by ITU in IMT Advanced. Potential and current applications include amended *mobile web access*, *IP telephony*, gaming services, *high-definition mobile TV*, *video conferencing*, *3D television*.

As opposed to earlier generations, a 4G system does not support traditional circuit-switched telephone service, but all - *Internet Protocol* (IP) based communication such as IP telephony. The spread spectrum radio technology used in 3G systems, is abandoned in all 4G candidate systems and replaced by *OFDMA multi-carrier* transmission and other *frequency-domain equalization* (FDE) schemes, making it possible to transfer very high bit rates despite extensive multi-path radio propagation (echoes). The peak bit rate is further improved by *smart antenna* arrays for MIMO communications.

Along with improved MIMO and OFDM there are a number of other techniques and technologies that will be employed.

- *Carrier Aggregation* (CA): As many operators do not have sufficient contiguous spectrum to provide the required bandwidths for the very high data rates, a scheme known as carrier aggregation has been developed. Using this technology operators are able to utilise multiple channels either in the same bands or different areas of the spectrum to provide the required bandwidth.

- *Coordinated Multipoint:* One of the key issues with many cellular systems is that of poor performance at the cell edges. Interference from adjacent cells along with poor signal quality lead to a reduction in data rates. For LTE-Advanced a scheme known as coordinated multipoint has been introduced.

- *LTE Relaying:* LTE relaying is a scheme that enables signals to be forwarded by remote stations from a main base station to improve coverage.

- *Device to Device,* (D2D): LTE D2D is a facility that has been requested by a number of users, in particular the emergency services. It enables fast swift access via direct communication. The example of D2D communication supported by the network is shown in figure below.

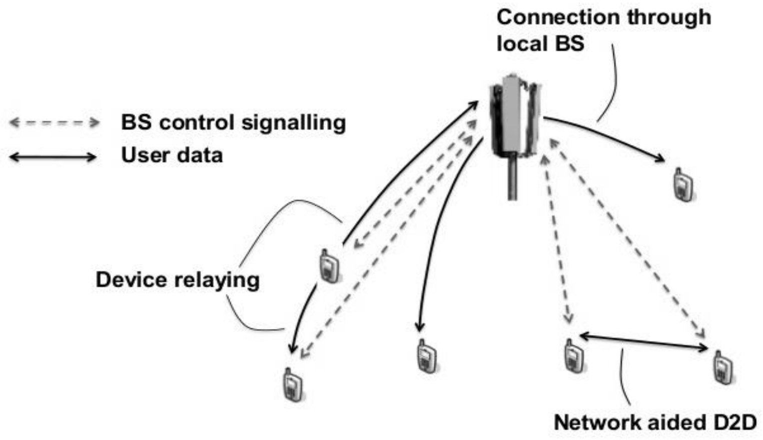


Figure 1.61. Network aided D2D communication

The 5th generation mobile networks or 5th generation wireless systems (5G) denotes the proposed next major phase of mobile telecommunications standards beyond the current 4G/IMT-Advanced standards. 5G planning includes *Internet connection* speeds faster than current 4G, and other improvements. Supposed that 5G will be based on Demand Attentive Network (DAN) philosophy and will uses very-wide-bandwidth radio channels (20 – 60 GHz). The main new technical features of 5G are presented below.

- *Multiple Access Schemes:* Again a variety of new access schemes are being investigated for 5G technology. Techniques including OFDMA, SCMA, NOMA, PDMA, MUSA and IDMA have all been mentioned.

- *Orthogonal frequency division multiple access* (OFDMA)*:* OFDMA has been widely used and very successful for 4G and could be used as a 5G multiple access scheme. However it does require the use of OFDM and requiring orthogonality between carriers and the use of a cyclic prefix has some drawbacks. As a result other multiple access schemes are being investigated.

- *Sparse Code Multiple Access* (SCMA)*:* SCMA is another idea being considered as a 5G multiple access scheme and it is effectively a combination of OFDMA and CDMA. Normally with OFDMA a carrier or carriers is allocated to a given user. However if each carrier has a spreading code added to it, then it would be able to transmit data to or from multiple users. This technique has been developed to use what are termed sparse code and in this way significant numbers of users can be added while maintaining the spectral efficiency levels.

- *Non-orthogonal multiple access* (NOMA): NOMA is one of the techniques being considered as a 5G multiple access scheme. NOMA superposes multiple users in the power domain, using cancellation techniques to remove the more powerful signal. NOMA could use orthogonal frequency division multiple access, OFDMA or the discrete Fourier transform, DFT-spread OFDM.

- *Modulation:* Whilst PSK and QAM have provided excellent performance in terms of spectral efficiency, resilience and capacity, the major drawback is that of a high peak to average power ratio (PAPR). Modulation schemes like Amplitude Phase Shift Keying (APSK) will provide advantages in several circumstances.

1.3. Computer networks architecture and design, using standard and specific adaptive telecommunication network elements

A computer network is the infrastructure that allows two or more computers (called hosts) to communicate with each other. The network achieves this by providing a set of rules for communication, called protocols, which should be observed by all participating hosts. The need for a protocol should be obvious: it allows different computers from different vendors and with different operating characteristics to ‘speak the same language’.

1.3.1. Network Components

Figure 1.62 shows an abstract view of a network and its hosts. The network is made up of two types of components: nodes and communication lines. The nodes typically handle the network protocols and provide switching capabilities. A node is usually itself a computer (general or special) which runs specific network software.

The communication lines may take many different shapes and forms, even in the same network. Examples include: copper wire cables, optical fiber, radio channels, and telephone lines.

A host is connected to the network by a separate communication line which connects it to one of the nodes. In most cases, more than one host may be connected to the same node. From a host’s point of view, the entire network may be viewed as a black box, to which many other hosts are connected. Each host has a unique address allocated to it by the network. For a host to communicate with another host, it needs to know the latter’s address. All communication between hosts passes through the nodes, which in turn determine how to route messages across the network, from one point to another.

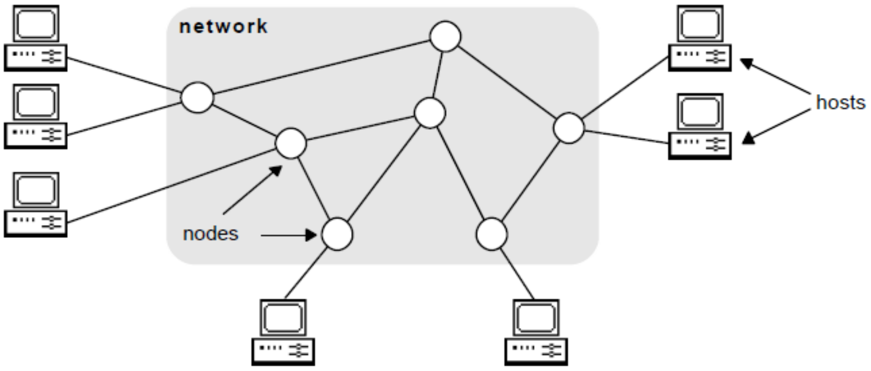


Figure 1.62. An abstract network

Throughout the rest of this book, there will be occasions when it is not necessary to distinguish between hosts and nodes. In such cases, we will use the term station to mean either.

Network Types

Networks may be divided into different types and categories according to four different criteria:

1. *Geographic spread* of nodes and hosts. When the physical distance between the hosts is within a few kilometres, the network is said to be a Local Area Network (LAN). LANs are typically used to connect a set of hosts within the same building (e.g., an office environment) or a set of closely-located buildings (e.g., a university campus). For larger distances, the network is said to be a Metropolitan Area Network (MAN) or a Wide Area Network (WAN).

MANs cover distances of up to a few hundred kilometres and are used for interconnecting hosts spread across a city. WANs are used to connect hosts spread across a country, a continent, or the globe. LANs, MANs, and WANs usually coexist: closely-located hosts are connected by LANs which can access hosts in other remote LANs via MANs and WANs, as illustrated in Figure 1.63.

2. *Access restrictions*. Most networks are for the private use of the organizations to which they belong; these are called private networks. Networks maintained by banks, insurance companies, airlines, hospitals, and most other businesses are of this nature. Public networks, on the other hand, are generally accessible to the average user, but may require registration and payment of connection fees. Internet is the most-widely known example of a public network. Technically, both private and public networks may be of LAN, MAN, or WAN type, although public networks, by their size and nature, tend to WANs.

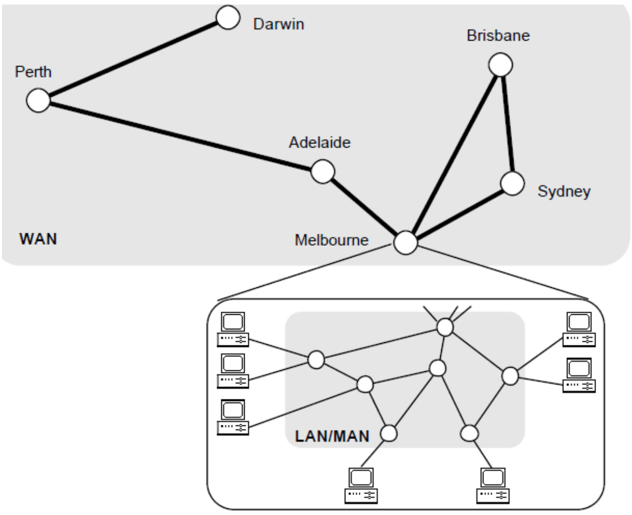


Figure 1.63. Example of a WAN between LANs

3. *Communication model* employed by the nodes. The communication between the nodes is either based on a point-to-point model or a broadcast model (see Figure 1.64). In the point-to-point model, a message follows a specific route across the network in order to get from one node to another. In the broadcast model, on the other hand, all nodes share the same communication medium and, as a result, a message transmitted by any node can be received by all other nodes. A part of the message (an address) indicates for which node the message is intended. All nodes look at this address and ignore the message if it does not match their own address.

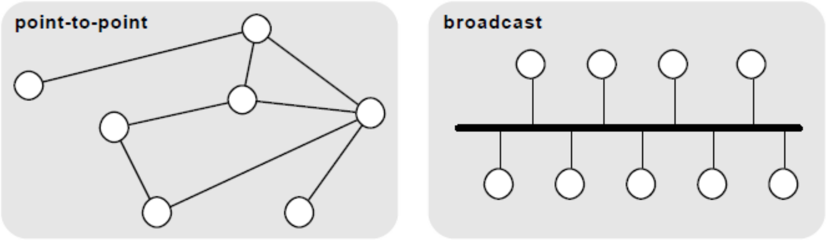


Figure 1.64. Communication models

4. *Switching model* employed by the nodes. In the point-to-point model, nodes either employ circuit switching or packet switching. Suppose that a host A wishes to communicate with another host B. In circuit switching, a dedicated communication path is allocated between A and B, via a set of intermediate nodes. The data is sent along the path as a continuous stream of bits. This path is maintained for the duration of communication between A and B, and is then released. In packet switching, data is divided into packets (chunks of specific length and characteristics) which are sent from A to B via intermediate nodes.

Each intermediate node temporarily stores the packet and waits for the receiving node to become available to receive it. Because data is sent in packets, it is not necessary to reserve a path across the network for the duration of communication between A and B. Different packets can be routed differently in order to spread the load between the nodes and improve performance.

However, this requires packets to carry additional addressing information.

1.3.2. The OSI Model

The International Standards Organization (ISO) has developed a reference model for network design called the Open Systems Interconnection (OSI). It proposes a seven-layer architecture for networks, as summarized by Figure 1.65. Each layer is characterized by a set of standard protocols which specify its behaviour.

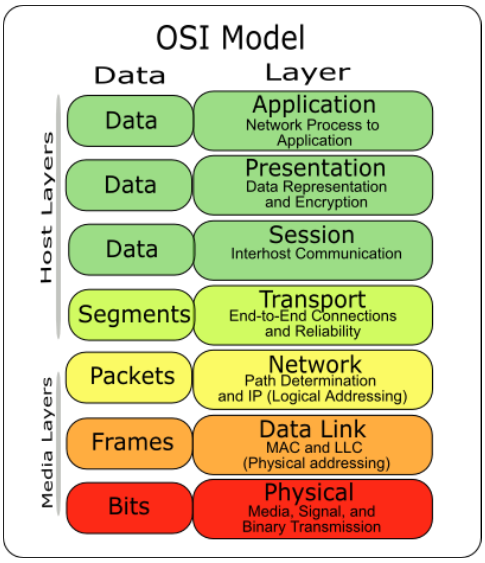

OSI Model
Data
Layer
Host Layers
Media Layers
Data
Application
Network Process to Application
Data
Presentation
Data Representation and Encryption
Data
Session
Interhost Communication
Segments
Transport
End-to-End Connections and Reliability
Packets
Network
Path Determination and IP (Logical Addressing)
Frames
Data Link
MAC and LLC (Physical addressing)
Bits
Physical
Media, Signal, and Binary Transmission

Figure 1.65. The OSI reference model

TCP/IP model

TCP/IP (Transmission Control Protocol / Internet Protocol) is a model which is similar to the OSI model, it contains 4 layers and a collection of protocols that make the network connectivity possible and now a day we are using TCP/IP on almost every single device that provides network connectivity.

The TCP/IP Layers:

- Layer 4: Application
- Layer 3: Transport
- Layer 2: Internet

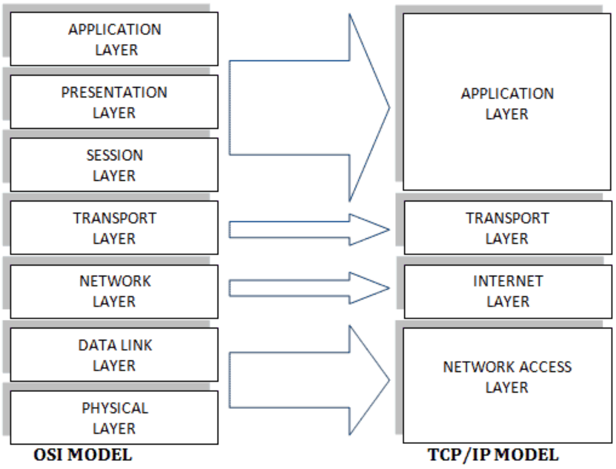


Figure 1.66. The OSI and TCP/IP models

These seven layers represent the protocol architecture for the communications component of a host. The nodes in a network implement only the lower three layers, as illustrated in Figure 1.66. The reason for this is that the upper four layers are irrelevant to the task of communication between the nodes.

In Figure 1.67, when host A sends a message to host B, the message moves down the successive layers of host A, from the application layer to the presentation layer, to the session layer, etc., until it reaches the physical layer. It is then transmitted across the communication line between host A and node X, and moves up the three layers of node X and down again. Then it is transmitted to node Y where it goes through the same procedure, and finally is transmitted to host B, where it moves up its seven layers, until it arrives at the application layer of host B.

Although actual communication takes place only at the physical layer, it is often useful to think of virtual communication between corresponding layers. For example, we can use an imaginary line of communication between the presentation layer on host A and the same layer on host B. This would be characterized by the presentation protocol.

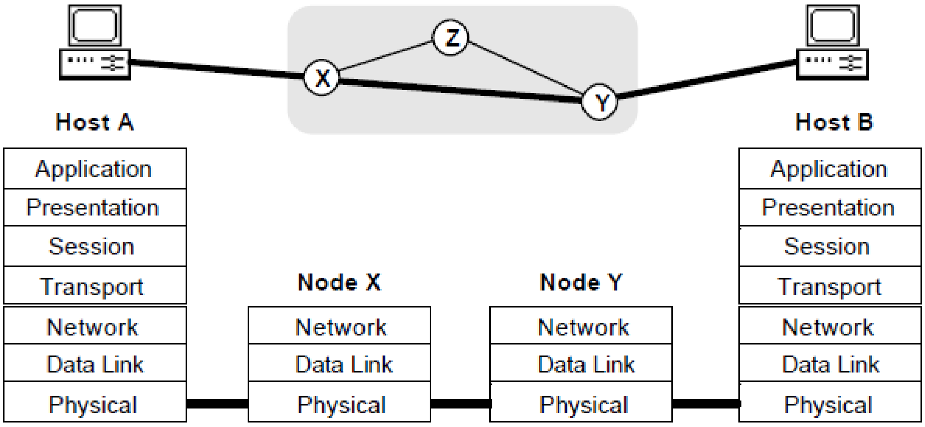


Figure 1.67. Nodes use only the lower 3 layers

The terms protocol and layer are often used interchangeably. This is harmless but not entirely accurate. Strictly speaking, protocol refers to the rules and conventions that the functions of a layer should conform to. Layer refers to a set of services and functions and their realization in hardware or software. A layer is therefore characterized by its protocol. A set of network layers is also commonly referred to as a protocol stack.

Each of the seven layers of the OSI model hides the implementation details of the lower layers from the upper layers. Well-defined protocols and interfaces for each of the layers make it possible for the layer to be designed and implemented in isolation from the other layers. Except for the physical layer, which is implemented in hardware, all other layers are implemented in software.1 For example, each of these layers may be implemented as a set of routines which communicate with the layer above and the layer below it via parameters passed in function calls. Alternatively, each layer may be implemented as a task (in a multi-tasking environment) which communicates with other tasks by message passing. Figure 1.68. illustrates the latter.

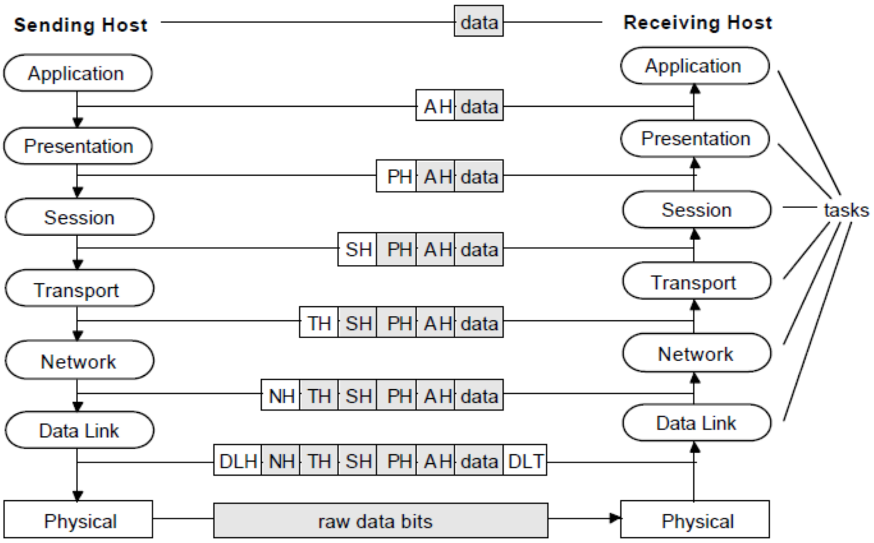


Figure 1.68. OSI layers as software tasks

For house-keeping purposes, each layer adds an additional piece of information to the message it is transmitting. The same layer removes the additional piece of information on the receiving end. The additional information appears in form of a header (e.g., TH = Transport Header). The data link layer adds a header as well as a trailer to its data.

Each of the seven layers of the OSI model is described below in more detail.

Subsequent chapters examine the layers in greater depth and discuss their main protocols. It should be pointed out that the OSI model is not the only model in use. It is, however, the benchmark for comparing other network architectures against.

The Physical Layer

The physical layer is concerned with the transmission of raw data bits over communication lines. Physical layer standards and protocols are concerned with issues such as the following:

- How a physical circuit is established between communicating devices.
- How the circuit is terminated when no longer needed.
- The physical form (e.g., voltages, frequencies, and timing) in which data bits (binary values 0 and 1) are represented.
- Whether transmission of data can take place in one or both directions over the same physical connection.
- Characteristics of the physical media that carry the signals (e.g., copper wire, optical fibre, radio waves).
- Characteristics of the connectors used for connecting the physical media.
- How data from a number of sources should be multiplexed before transmission and demultiplexed upon arrival, and the type of multiplexing technique to be used.
- The type of modulation to be used for transmitting digital data over analogue transmission lines.

The physical layer accounts for much of the tangible components of a network, including cables, satellites, earth stations, repeaters, multiplexers, concentrators, and modems. Physical layer protocols and standards are of mechanical, electrical, functional, and procedural nature.

The physical layer hides the above details from the higher layers. To the data link layer, it appears as a logical communication channel which can send a stream of bits from one point in the network to another (but not necessarily reliably).

The Data Link Layer

The data link layer is concerned with the reliable transfer of data over the communication channel provided by the physical layer. To do this, the data link layer breaks the data into data frames, transmits the frames sequentially over the channel, and checks for transmission errors by requiring the receiving end to send back acknowledgment frames. Data link protocols are concerned with the following issues:

- How to divide the data into frames.
- How to delimit frames by adding special bit patterns to the beginning and end of each frame. This allows the receiving end to detect where each frame begins and where it ends.
- Error detection. Some form of error check is included in the frame header. This is constructed by the transmitting end based on the contents of the frame, and checked for integrity by the receiving end. A change in the frame bits can be detected in this way.
- Error correction. When a frame arrives corrupted or is for any reason lost in the network, it is retransmitted. Lost acknowledgment frames may result in duplicate frames, which need to be detected and corrected as well.
- Flow control. In general, not all communication devices in a network operate at the same speed. Flow control provides a means of avoiding a slow receiver from being swamped by data from a fast transmitter.

The data link layer hides the above details from the higher layers. To the network layer, it appears as a reliable communication channel which can send and receive data packets as frames.

The Network Layer

The network layer is concerned with the routing of data across the network from one end to another. To do this, the network layer converts the data into packets and ensures that the packets are delivered to their final destination, where they can be converted back into the original data. Network layer protocols are concerned with the following issues:

- The interface between a host and the network.
- The interface between two hosts across the network.
- Routing of packets across the network, including the allocation of a route and handling of congestion.
- Correct ordering of packets to reflect the original order of data.
- Collection of statistical information (e.g., number of transmitted packets) for performance measurement and accounting purposes.
- Internetworking: communication between two or more networks.

The network layer hides the above details from the higher layers. To the transport layer, it appears as a uniform data transfer service, regardless of the location of the communicating devices and how they are connected.

The Transport Layer

The aim of the transport layer is to isolate the upper three layers from the network, so that any changes to the network equipment technology will be confined to the lower three layers (i.e., at the node level). Transport layer protocols are concerned with the following issues:

- Establishment and termination of host-to-host connections.
- Efficient and cost-effective delivery of data across the network from one host to another.
- Multiplexing of data, if necessary, to improve use of network bandwidth, and demultiplexing at the other end.
- Splitting of data across multiple network connections, if necessary, to improve throughput, and recombining at the other end.
- Flow control between hosts.
- Addressing of messages to their corresponding connections. The address information appears as a part of the message header.
- Type of service to be provided to the session layer (e.g., error-free versus errorprone connections, whether messages should be delivered in the order received or not).

The transport layer hides the above details from the higher layers. To the session layer, it appears as a customized data transfer service between two hosts, isolating the underlying network technology from it.

The Session Layer

The session layer provides a structured means for data exchange between user processes on communicating hosts. Session layer protocols are concerned with the following issues:

- Negotiating the establishment of a connection (a session) between user processes on communicating hosts, and its subsequent termination. This includes the setting of various communication parameters for the session (e.g., synchronization and control).
- Correct ordering of messages when this function is not performed by the transport layer.
- Recovery from interrupted transport connections, if necessary.
- Grouping of messages into a larger message, if necessary, so that the larger message becomes available at the destination only when its constituent messages have all been delivered successfully.

The session layer hides the above details from the higher layers. To the presentation layer, it appears as an organized communication service between user processes.

The Presentation Layer

The presentation layer provides a mutually-agreeable binary representation of the application data communicated between two user processes. Since there are many ways of encoding application data (e.g., integers, text) into binary data, agreement on a common representation is necessary. Presentation layer protocols are concerned with issues such as the following:

- Abstract representation of application data.

- Binary representation of application data.
- Conversion between the binary representation of application data and a common format for transmission between peer applications.
- Data compression to better utilize network bandwidth.
- Data encryption as a security measure.

The presentation layer hides the above details from the higher layers. To the application layer, it appears as a universal communication service between user processes, regardless of their system-specific idiosyncrasies, allowing them to converse in a common syntax.

The Application Layer

The application layer is concerned with the semantics of data, i.e., what the data means to applications. The application layer provides standards for supporting a variety of application-independent services. Examples include:

- Virtual terminal standards to allow applications to communicate with different types of terminals in a device-independent manner.
- Message handling system standards used for electronic mail.
- File transfer, access, and management standards for exchanging files or parts thereof between different systems.
- Transaction processing standards to allow different companies with different systems to access each other’s on-line databases (e.g., in banking and airline reservation).
- On-line directory standards for storing details of individuals, organizations, and network components.
- Standards for exchanging formatted documents.

Application layer standards have paved the way for open software systems, in which data can be communicated between incompatible base systems (i.e., different hardware and software architectures) without loss of meaning or usefulness.

1.3.3. Protocol Notations

OSI network protocols are specified in a variety of notations. This section describes two popular notations, sequence diagrams and state transition diagrams, which are extensively used in standards and the literature. Both rely on the notion of a service primitive which is described first.

Service Primitives

A service primitive is an abstract representation of the interaction between a service provider and a service user. Service primitives are concerned with what interactions take place rather than how such interactions are implemented. Service primitives may be of one of the following four types:

- Request Primitive. This is issued by a service user to the service provider to request the invocation of a procedure.
- Indication Primitive. This is issued by the service provider to a peer service user (usually in response to a request primitive) to indicate that a procedure has been requested.
- Response Primitive. This is issued by a peer service user to the service provider (usually in response to an indication primitive) to indicate that the requested procedure has been invoked.
- Confirm Primitive. This is issued by the service provider to a service user to indicate that an earlier request for the invocation of a procedure has been completed.

An actual service primitive consists of a command and, if appropriate, a set of associated parameters. A simple convention is used for naming primitives: a primitive name consists of the first letter of the layer to which it belongs, followed by its command name, followed by its type. For example, a request type primitive at the network layer for initiating a connection is named ‘N-CONNECT request’.

Sequence Diagrams

A sequence diagram defines a service protocol by specifying the permissible sequence of service primitives that may be exchanged between service users and service providers. Service users and service providers are represented by vertical bars. Service primitives are represented by directed lines between the bars. For clarity, primitive parameters are not included.

Figure 1.69. shows a simplified example of requesting a connection at the network layer. According to the diagram, a service user can request, from the service provider, a connection to a peer service user. The service provider in turn issues a connection indication to the peer service user. The peer service user responds to the service provider which, in turn, confirms the cycle with the original service user.

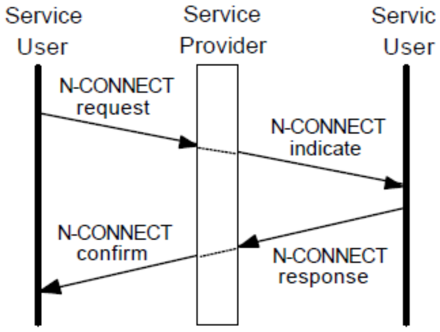


Figure 1.69. A simple sequence diagram

State Transition Diagrams

A state transition diagram describes the various execution states a station can assume and how service primitives cause it to transit from one state to another. States are represented by circles or boxes, and are labelled with a meaningful name that describes the state. A state transition is represented by a directed line from one state to another, and is labeled with the service primitive that triggers the transition.

Figure 1.70 shows an example which describes (in a simplified form) the states of a station at the network layer. According to the diagram, assuming that a station is in the idle state, if it issues a connection request to another station, it enters the attempting to connect state where it waits for a connection to be confirmed, in which case it moves to the connected state, or disconnected, in which case it returns to the idle state. A similar scenario applies to an incoming connection which starts with the station receiving a connection indication. Note that the N-DISCONNECT primitives can be either of request or confirmation type.

It is worth noting the complementary nature of sequence diagrams and state transition diagrams. The former specifies a service protocol from an outside observer's point of view, while the latter describes the same protocol from a station's point of view. The two notations, combined, provide a complete picture of how a protocol operates.

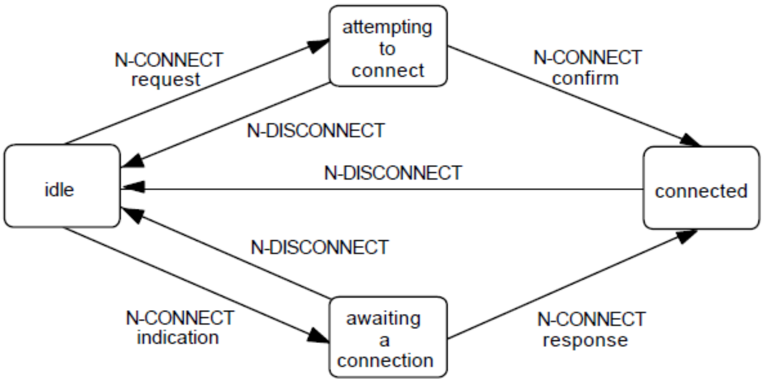


Figure 1.70. A simple state transition diagram

Standards

The importance of standards in the field of communication cannot be overstressed.

Standards enable equipment from different vendors and with different operating characteristics to become components of the same network. Standards also enable different networks in different geographical locations (e.g., different countries and continents) to be real cost savings: the same end-user device can be used for access to a variety of networks and services.

Standards are developed by national and international organizations established for this exact purpose. During the course of this book we will discuss a number of important standards developed by various organizations, including the following:

- The International Standards Organization (ISO) has already been mentioned. This is a voluntary organization with representations from national standards organizations of member countries (e.g., ANSI), major vendors, and end-users. ISO is active in many area of science and technology, including information technology. ISO standards are published as ISO serial-no (e.g., ISO 8632).

- The Consultative Committee for International Telegraph and Telephone (CCITT) is a standards organization devoted to data and telecommunication, with representations from governments, major vendors, telecommunication carriers, and the scientific community. CCITT standards are published as Recommendation L.serial-no, where L is a letter of the alphabet (e.g., I.440). These are revised and republished every four years. CCITT standards are very influential in the field of telecommunications and are adhered to by most vendors and carriers.
- The Institute of Electrical and Electronic Engineers (IEEE) is a US standards organization with members throughout the world. IEEE is active in many electric and electronic-related areas. The IEEE standards for local area networks are widely adopted and will be discussed in Chapter 9. IEEE standards are published as IEEE serial-no (e.g., IEEE 908).
- The Electronic Industries Association (EIA) is a US trade association best known for its EIA-232 standard, which will be discussed in the next chapter.
- The European Computer Manufacturers Association (ECMA) is a standards organization involved in the area of computer engineering and related technologies. ECMA directly cooperates with ISO and CCITT.

In addition to these organizations, and because of their global market influence, large vendors occasionally succeed in establishing their products as de facto standards. We will also look at a few standards of this nature later in the book.

1.3.4. Local Area Networks

Local Area Networks (LANs) have become an important part of most computer installations. Personal computers have been the main driving force behind the LAN proliferation. As personal computers became more widely used in office environments, so it became desirable to interconnect them to achieve two aims: to enable them to exchange information (e.g., e-mail), and to enable them to share scarce and expensive resources (e.g., printers). LANs have been so successful in realizing these aims that their cost is well justified even when there are only a handful of participating computers.

Current LANs are used for interconnecting almost any type of computing devices imaginable, including mainframes, workstations, personal computers, file servers, and numerous types of peripheral devices. Many LANs are further connected to other LANs or WANs via bridges and gateways, hence increasing the reach of their users.

In this chapter we will first look at some basic LAN concepts, and then discuss a number of widely-adopted LAN standards. As before, our aim will be to concentrate on general principles and protocols of importance rather than to get involved in the details of vendor-specific products.

Basic Concepts

A LAN consists of four general types of components:

- User station. This provides the user with access to the LAN. The most common example is a personal computer. The user station runs special network software (usually in form of a driver) for accessing the LAN.
- LAN protocol stack. This implements the LAN protocol layers. This usually takes the form of a hardware card inside the user station, containing a microprocessor and firmware which implements the non-physical protocols.
- Physical Interface Unit. This directly interfaces the user station-based LAN hardware to the LAN physical medium. The exact form of the PIU is highly dependent on the LAN physical medium. Coaxial cable connectors and cable TV taps are common examples.
- Physical Medium. This provides a physical path for signals to travel between stations. Coaxial cable, optical fiber, and infra-red light are examples.

The LAN protocol stack will be the main focus of our attention in this chapter.

Topologies and Access Protocols

There are two general categories of LAN topologies: bus and ring (see Figure 1.71).

The bus topology uses a broadcast technique, hence only one station at a time can send messages and all other station listen to the message. A listening station examines the recipient address of the message and if it matches its own address, copies the message; otherwise, it ignores the message.

The ring topology uses a closed, point-to-point-connected loop of stations. Data flows in one direction only, from one station to the next. As with the bus topology, transmission is restricted to one user at a time. When a station gains control and sends a message, the message is sent to the next station in the ring. Each receiving station in the ring examines the recipient address of the message and if it matches its own address, copies the message. The message is passed around the ring until it reaches the originator which removes the message by not sending it to the next station.

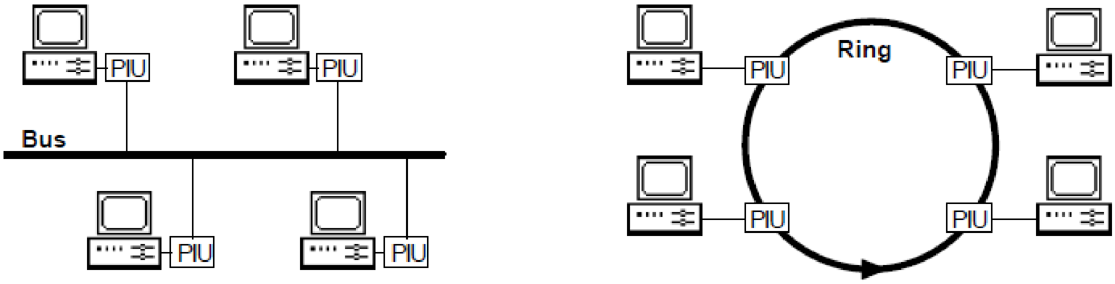


Figure 1.71. LAN topologies

Given that access to the bus or ring is restricted to one station at a time, some form of arbitration is needed to ensure equitable access by all stations. Arbitration is imposed by access protocols. A number of such protocols have been devised:

- ***Carrier Sense***. This protocol is applicable to a bus topology. Before a station can transmit, it listens to the channel to see if any other station is already transmitting. If the station finds the channel idle, it attempt to transmit; otherwise, it waits for the channel to become idle. Because of an unavoidable delay in a station's transmission to reach other stations, it is possible that two or more stations find the channel idle and simultaneously attempt to transmit. This is called a collision. Two schemes exist for handling collisions:

- ✓ ***Collision Detection***. In this scheme a transmitting station is required to also listen to the channel, so that it can detect a collision by observing discrepancies in the transmission voltage levels. Upon detecting a collision, it suspends transmission and re-attempts after a random period of time. Use of a random wait period reduces the chance of the collision recurring.
- ✓ ***Collision Free***. This scheme avoids collisions occurring in the first place. Each station has a predetermined time slot assigned to it which indicates when it can transmit without a collision occurring. The distribution of time slots between stations also makes it possible to assign priorities.

- Token Ring. This protocol is applicable to a ring topology. Channel access is regulated by a special message, called a token, which is passed around the ring from one station to the next. The state of the ring is encoded in the token (i.e., idle or busy). Each station wishing to transmit needs to get hold of the idle token first. When a station gets hold of the idle token, it marks it as busy, appends to it the message it wishes to transmit, and sends the whole thing to the next station.
- The message goes round the ring until it reaches the intended recipient which copies the message and passes it on. When the message returns to the originator, it detaches the message, marks the token as idle and passes it on. To ensure fair access, the token should go round the ring, unused, at least once before it can be used by the same station again.
- Token Bus. This protocol is applicable to a bus topology but makes it behave as a ring. Each station on the bus has two other stations designated as its logical predecessor and its logical successor, in a way that results in a logical ring arrangement (see Figure 1.72.). A special message is provided which plays the role of a token. Each station receives the token from its predecessor, readdresses it to its successor, and retransmits it on the bus. The rest of the protocol is as in a token ring.

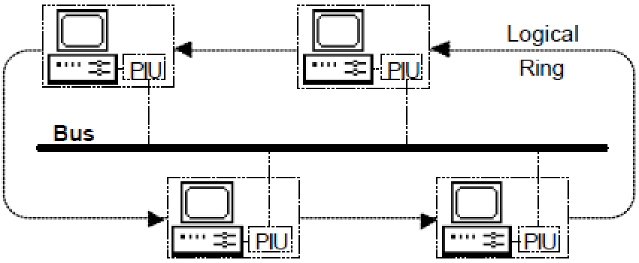


Figure 1.72. Token bus arrangement

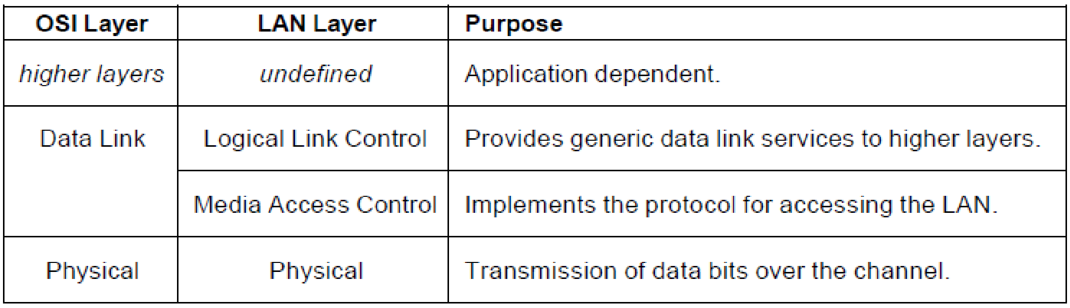
Architecture

Figure 1.73. depicts the LAN protocol layers in relation to the OSI model. The role of the physical layer is the same as in the OSI model. It includes the connectors used for connecting the PIU to the LAN and the signalling circuitry provided by the PIU.

(The next section describes the transmission methods employed by this layer.)

The OSI data link layer is broken into two sublayers. The Media Access Control (MAC) layer is responsible for implementing a specific LAN access protocol, like the ones described earlier. This layer is therefore highly dependent on the type of the LAN. Its aim is to hide hardware and access protocol dependencies from the next layer. As we will see shortly, a number of MAC standards have been devised, one for each popular type of access protocol.

The Logical Link Control (LLC) layer provides data link services independent of the specific MAC protocol involved. LLC is a subset of HDLC and is largely compatible with the data link layer of OSI-compatible WANs. LLC is only concerned with providing Link Service Access Points (LSAPs). All other normal data link functions (i.e., link management, frame management, and error handling) are handled by the MAC layer.



OSI Layer	LAN Layer	Purpose
higher layers	*undefined*	Application dependent.
Data Link	Logical Link Control	Provides generic data link services to higher layers.
	Media Access Control	Implements the protocol for accessing the LAN.
Physical	Physical	Transmission of data bits over the channel.

Figure 1.73. LAN protocol architecture

LANs are not provided with a network layer (or any other higher layer) because such a layer would be largely redundant. Because the stations are directly connected, there is no need for switching or routing. In effect, the service provided by the LLC is equivalent to the OSI network layer service.

Transmission

LAN transmission techniques are divided into two categories: baseband and broadband. In the baseband technique, the digital signal from a transmitting device is directly introduced into the transmission medium (possibly after some conditioning).

In the broadband technique, a modem is used to transform the digital signal from a transmitting device into a high frequency analogue signal. This signal is typically frequency multiplexed to provide multiple FDM channels over the same transmission medium.

Baseband is a simple and inexpensive digital technique. By comparison, broadband has additional costs: each device requires its own modem; also, because transmission is possible in one direction only, two channels typically need to be provided, one for either direction. Broadband, however, has the advantages of offering a higher channel capacity which can be used for multiplexing data from a variety of sources (e.g., video, voice, fax), not just digital data. It is also capable of covering longer distances, typically tens of kilometres compared to up to a kilometre for baseband.

1.4. Specialized telecommunications networks, the use for infrastructure control

1.4.1. State-of-the-art meter reading

Advanced Metering Infrastructure (AMI) refers to systems that measure, collect and analyse energy and water usage, from devices such as electricity meters, gas meters, and water meters, respectively, through various communication media on request or on a pre-defined schedule. AMI may also include end user/customer associated systems and meter data management (MDM) and it allows collection and distribution of information to customers, suppliers, utility companies and service providers. In an AMI, the data collected can be captured, stored, and forwarded to a central computer. This can include events and alarms such as tamper, leak detection, low battery, or reverse flow.

AMI can be differentiated from traditional Automated Meter Reading (AMR) and Automatic Meter Management (AMM) in that the latter may be considered subsystems of an AMI, in the way communications takes place, and in the complexity of networks and protocols. Moreover, AMI aims at an open system with a connection to the HAN.

AMR is a technology which automatically collects data from metering devices like water, gas, heat, electricity and transfers these data to a central database for analysis and billing purposes. Many AMR devices can also do data logging. The logged data can be used for water or energy use profiling, time of use billing, demand forecasting, demand side management (DSM), rate of flow recording, leak detection, flow monitoring, etc.

AMM allows control or influence energy consumption of customers through automatic reading of meters within fine time intervals in order to realise price elastic energy or water consumption or demand response.

A HAN is a network contained within a user's premises that interconnects home IT and entertainment devices and their peripherals as well as home security systems, and "smart" appliances such as lighting, heating, cooling, etc.

1.4.2. General concept of an AMI

Figure 1.74. displays the generic concept and architecture of an Advanced Metering

Infrastructure that may be subdivided into three segments:

- The local network segment
- The access network segment
- The back-haul network segment

The local network connects AMI-enabled meters belonging to the same entity (home, building, facility) as well as end-user applications (HAN) to a node acting as a local data collector and gateway between access and local network.

The access network comprises the networks between house gateway and a hub/data concentrator or the data management centre ic case there is no data concentrator, whilst the term back-haul network is used to designate the final segment between Hub/data concentrator and the data & management centre for utility services and customer-related services.

In some cases there may be no dedicated hub/data concentrator e.g. if public access networks are used.

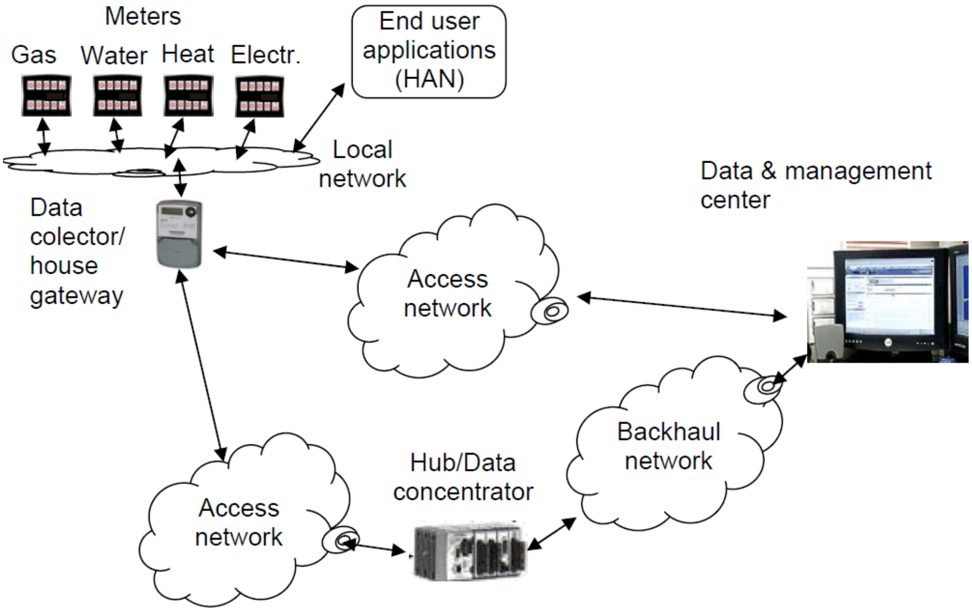


Figure 1.74. Generic concept and architecture of an Advanced Metering Infrastructure

To collect metering data from AMI-enabled meters belonging to the same entity (facility, building) and to enable energy-related end-user applications (HAN) a local network is typically installed using wired or wireless communications technologies.

1.4.3. Wire-line technologies

Data over PSTN

Data transmission over the Public Switched Telephone Network (PSTN) was a pioneer solution in order to have internet access in 1990's. With advances in voice-band modem technologies in the recent years, transmission speed has been steadily increasing virtually reaching Shannon limit of the bandwidth limited end-to-end analogue phone line channel (0.3 – 3.1 kHz) with the V.34 standard. The V.34bis standard achieves a maximum throughput of 33.4 kbps over an entirely analogue channel. It is based on trellis coded higher order modulation and sophisticated equalization techniques.

With the digitalisation of the PSTN behind local exchange (end office) providing a 64 kbps digital channel per phone call (the basic DS0 circuit), the use of even faster modem standards have become possible to bridge the remaining analogue 'last mile' segment only.

The V.92 standard e.g. provides 56 kbps downstream and approximately 33 kbps upstream.

The use of the PSTN is principally an interesting option for AMR and related applications, since today in Europe every house is connected to the PSTN via at least one twisted pair.

Data rates achievable with PSTN compatible modem standards (V.34 and V.92) would be adequate for these purposes. Apart from the modem equipment in the house gateway and possible installation costs to connect the house gateway to the POTS, there are no CAPEX.

However, since there is typically only one customer phone line available, dial-in charges would be on customers account. Moreover, the use of the analogue phone line by both customer and AMI may lead to conflicts particularly, if the PSTN is frequently accessed.

The use of an ISDN technology providing at least 2 independent channels at even higher speed may be an alternative to circumvent this problem. However, ISDN is not available in every household.

xDSL

ADSL (Asymmetric Digital Subscriber Line) is a data communications technology that enables much faster data transmission over copper telephone lines than conventional voice band modems can provide. It does this by utilising frequencies well above the voice band (> 25 kHz). The band from 25.875 kHz to 138 kHz is used for the upstream, whilst 138 kHz -1104 kHz is used for the downstream.

Offering ADSL services to end users requires a so-called DSL Access Multiplexer (DSLAM) in the local exchange/end office. Data carried by the ADSL is typically routed over the telecom operators’ data networks and the Internet.

The latest ITU G.992.1 Annex ATH standard (ADSL over POTS/twisted pair copper) supports up to 12 Mbps down-stream and 1.3 Mbps upstream. In real channel conditions, achievable data rates are usually significantly lower depending on line length and cross-talk interference levels. ADSL can generally only be used over line length typically less than 4km assuming a minimum service offer of a few hundred kbps.

Today a high percentage of households in Europe use ADSL for Internet access over POTS lines. If correctly installed, ADSL does not conflict with the use of the PSTN.

In cities, where telephone lines are typically shorter, the VDSL technology using frequencies far into the MHz range (<20 MHz for VDSL and <30 MHz for VDSL2) are presently deployed.

These ADSL evolution standards can support data rates up to 100 Mbps under optimum conditions enabling high quality IP-based TV services typically requiring several Mbps.

Today in Western Europe, a high percentage of PSTN local loops are equipped with ADSL. Its use at flat-rate tariff makes this technology an interesting candidate in an AMI concept. However, since ADSL cannot support point-to-multi-point connections, bandwidth cannot be shared with a second ADSL modem that is part of the AMI house gateway. An ADSL terminal equipment common to both AMI and customer Internet access may principally be possible. However, its practical implementation would require introduction of new business concepts where infrastructure is shared among energy and telecom service providers.

FTTB, FTTH

Fibre to the premises (FTTP) is a form of fibre-optic communication delivery in which an optical fibre is run directly onto the customers' premises. This contrasts with other fibre-optic communication delivery strategies such as fibre to the node (FTTN), fibre to the curb (FTTC), or hybrid fibre-coaxial (HFC), all of which depend upon more traditional methods such as copper wires or coaxial cable for "last mile" delivery.

Fibre to the premises can be further categorized according to where the optical fibre ends:

- FTTH (fibre to the home) is a form of fibre optic communication delivery in which the optical signal reaches the end user's living or office space.
- FTTB (fibre to the building) is a form of fibre optic communication delivery in which the optical signal reaches the private property enclosing the home or business of the subscriber or set of subscribers, but where the optical fibre terminates before reaching the home living space or business office space, with the path extended from that point up to the user's space over a physical medium other than optical fibre.

Different companies like telecom or electricity have started to invest in this new technology.

M-Bus

M-Bus “Meter bus” is a European standard, used mainly for one-way or two-way data exchange with gas, water and heat meters. It can also be used with various sensors and actuators. M-bus also supports remote powering of communication devices via the data wires (power over data).

It is standardised by CEN TC 294, "Communication systems for meters and remote reading of meters" in the EN 13757 series. TC 294 covers data exchange with all utility meters except electricity meters, which are covered by IEC / CENELEC TC 13.

EN 13757-1 is a general standard for meter data exchange, covering several physical media, protocols and the COSEM application data model. The parts EN 13757-2 ... EN.

13757-6 specify the M-Bus protocol layers. M-Bus uses an OSI three-layer (collapsed) protocol architecture, consisting of:

- The wired or the wireless physical layer;
- The data link layer based on IEC 60870-5-1 and IEC 60870-5-2;
- The (M-Bus) dedicated application layer.

M-Bus can alternatively be used with DLMS/COSEM: the COSEM Application layer specified in IEC 62056-53, the COSEM objects specified in IEC 62056-62 and the OBIS identification system specified in IEC 62056-61 (electricity) and in EN 13757-1 (other than electricity).

On the other hand, devices using DLMS/COSEM – e.g. electricity meters or house gateways

– may be set up as M-Bus masters, exchanging data with devices using M-Bus.

M-bus supports the following physical interfaces:

- Twisted pair local bus with base band signalling, according to EN 13757-2;

- Wireless in the unlicensed 868 – 980 MHz SRD (Short Range Device) band, according to EN 13757-4. It is suitable for in-house data exchange, up to 15 m. The action radius can be extended by using the relaying methods specified in EN 13757-5;
- Local bus, according to EN 13757-6.

M-Bus is a protocol optimised from the point of view of the meter, allowing simple and low-cost implementations and long battery life.

1.4.4. Powerline communications (PLC)

The use of existing powerlines (MV and LV) for communications in a future open AMI providing energy-related end-user applications is obvious, since most devices and appliances of interest are already connected to the power grid.

Today, many proprietary technologies and industrial (open) standards exist for communications over powerline (PLC) and systems have already been deployed in large scale. Normally, a distinction is made between

- Narrowband PLC systems operating in the so-called CENELEC band essentially below 148.5 kHz based on a commonly accepted normative basis. (for utility-related services CENELEC Band A from 3 to 95 kHz applies).
- Broadband PLC systems typically operating between 2 – 30 MHz on an interim normative basis not yet commonly accepted.

In the U.S., there exist commonly accepted standards enabling PLC solutions for MV and LV networks at frequencies up to about 450 KHz.

Apart from EMC concerns and regulatory issues not yet resolved, PLC operating in the MHz (HF) range has advantages over systems running at low frequencies in the CENELEC band. These are:

- Coupling and surge/transient protection becomes simpler
- Higher access impedance
- Signal propagation is lesser affected by changing network loading

An important issue of PLC-based AMR in general is the intermeshing of the power distribution network which may pose problems in certain cases, e.g. when parts of the LV grid are disconnected and reconnected to another feeding transformer or other similar scenarios. A PLC based AMI should be capable of handling such situations.

Open narrowband PLC technology standards

IEC 61334-5-1 S-FSK

IEC 61334-5-1, specifying the physical and the MAC layers, is the only PLC standard (DLC in IEC terminology) supported by IEC. It uses Spread Frequency Shift Keying modulation (SFSK).

The IEC 61334 series provide the specification of all OSI layers necessary for efficient PLC communication. In particular, the application models of IEC 62056-61/62 (DLMS/COSEM) with the IEC 52056-53 application layer can be connected via the IEC 61334-32 link layer to IEC 61334-5-1. The physical and MAC layers are supported by chips from several manufacturers. Smart metering systems based on IEC 61334-5-1 are offered by several meter/system providers. The IEC 61334-5-1 standard is part of the Dutch NTA DSMR specification.

PLC systems based on the Spread Frequency Shift Keying standard achieve reliable communication by combining frequency diversity, a robust repeating scheme, extensive error control with a lean protocol stack. However, supported data rates may be too low with respect to requirements of a future AMI.

IEC 61334-5-2

IEC 61334-5-2 describes an FSK based Physical Layer for PLC. The PHY layer relies on a binary differential FSK modulation. Its main characteristics are:

- Half Duplex operation
- Synchronous transmission
- Bit rate of 600 bit/s or 1200 bit/s

IEC61334-5-4

IEC61334-5-4 describes an OFDM-based PHY layer for communication via power lines. The basic modulation scheme is DPSK. Carrier spacing is 4.5 kHz leading to a brut data rate per carrier of 4.5 kbps. To increase the robustness with respect to channel impairments, a rate ½-convolutional code is used.

CENELEC EN50090 (KNX – PL)

The Konnex protocol is a communication protocol suite for home automation approved as

International Standard (ISO/IEC 14543-3), European Standard (CENELEC EN 50090 and

EN 13321-1) and Chinese Standard (GB/Z 20965). Konnex supports communications via several physical media, among others: power line, twisted pairs, RF, and IP tunnelling.

Regarding power line two profiles are defined:

- PL110 profile is an S-FSK scheme based on the EIB PL110 physical layer. Center Frequency is 110 kHz (CENELEC band B) with maximum signalling speed of 1.2 kbps.
- PL132 profile is an FSK scheme based on EHS physical Layer. Center frequency is 132.5 kHz (CENELEC band C) with maximum signalling speed of 2.4 kbps.

1.4.5. Wireless technologies

Drive-by

Mobile or "Drive-by" meter reading (ref. Figure 1.75) is a concept where a reading or interrogating device is installed in a vehicle. The meter reader drives the vehicle while the interrogating device automatically collects the meter readings. Often for drive-by meter reading the reading equipment includes navigational and mapping features supported by GPS and mapping software. With mobile meter reading, the reader does not normally have to read the meters in any particular route order, but just drives the service area until all meters are read.

In principle, meter readings/data can be directly forwarded to the billing centre via public cellular networks or private utility networks if available.

This concept has been widely used in less densely populated areas where walk-by isn’t an economical solution. Drive-by is quite popular in US but there is only limited use in EU.

Components often consist of a vehicular roof-top mounted antenna, a transceiver, a laptop and a software, RF receiver/transceiver, and external vehicle antennas.

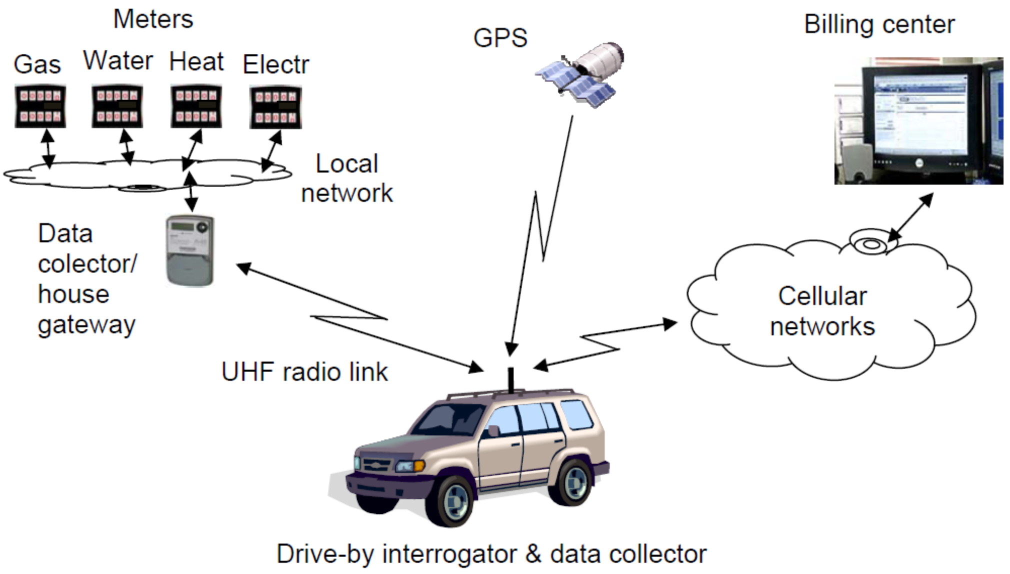


Figure 1.75. Drive-by concept

For drive-by meter reading the following wireless technologies/standards are commonly employed:

- ZigBee based on IEEE 802.15.4U MAC & PHY operating in the unlicensed 2.4 GHz
- Proprietary or industrial radio standards (e.g. ZigBit 900) operating in the UHF ISM band (433 MHz or 868 MHz in Europe, 915 MHz in North America) or frequency hopping-based long range UHF RFID (Wavenis).

Transmit power may be in the order of 10 mW.

To wirelessly forward meter readings/data to a central computer (billing centre) the following communications technologies/standards are commonly used:

- U2G/2.5G GSM/GPRS/EDGE
- U3G UMTS
- For navigation and positioning UGPS

Wireless communications solutions in fixed networks are always attractive, since they do not require new wiring. This is particularly true if a wireless network infrastructure that already exists could be used e.g. public mobile network, municipal WiFi or WiMAx networks, etc. On the other hand, solid buildings with concrete walls, together with the installation of meters predominantly in basements- which is typical for European countries - are obstacles for radio-based metering solutions. In these cases, additional equipment e.g. house gateway/repeaters installed at more favourable places become necessary. Therefore, wireless metering solutions so far have been mainly successful in countries like the USA, where houses typically are built less solid and without basements.

Open standard wireless technologies

IEEE 802.15.4 (WPAN)

PHY and MAC

IEEE 802.15.4 is a standard which specifies the physical layer and media access control (MAC) layer for low-rate wireless personal area networks (LR-WPANs). It is maintained by the IEEE 802.15 working group.

This standard intends to offer the fundamental lower network layers of a type of wireless personal area network (WPAN), which focuses on low-cost, low-speed ubiquitous communication between devices (in contrast with other, more end user-oriented approaches, such as Wi-Fi). The emphasis is on very low cost communication of nearby devices with little to no underlying infrastructure, intending to exploit this to lower power consumption even more.

The basic framework conceives a 10 to 75 meter communications area with a transfer rate of 250 kbps. Trade-offs are possible to favour more radically embedded devices with even lower power requirements, through the definition of not one, but several physical layers. Lower transfer rates of 20, 40 and 100 kbps are defined in the standard as well.

The IEEE 802.15.4 standard operates on one of three possible unlicensed frequency bands:

- 868-868.8 MHz: Europe, allows one communication channel
- 902-928 MHz: North America, up to thirty channels
- 2400-2483.5 MHz: worldwide use, up to sixteen channels

Important features of IEEE 802.15.4 include real-time suitability by reservation of guaranteed time slots, collision avoidance through CSMA/CA and integrated support for secure communications. Devices also include power management functions.

Networks can be built as either peer-to-peer (point-to-point) or star networks:

- Peer-to-peer networks can form arbitrary patterns of connections, and their extension is only limited by the distance between each pair of nodes. They are meant to serve as the basis for ad hoc networks capable of performing self-management and organization
- A more structured star pattern is also supported, where a network coordinator is assigned and plays the role of the central node.

Regarding secure communications, the Medium Access Control (MAC) sublayer offers facilities that can be harnessed by upper layers to achieve the desired level of security.

Higher-layer processes may specify keys to perform symmetric cryptography to protect the payload and restrict it to a group of devices or just a point-to-point link; these groups of devices can be specified in access control lists.

Finally, indicate that the IEEE 802.15.4 is the basis for such upper layer specifications as ZigBee, 6loWPAN, Wireless HART. These specifications attempt to offer a complete networking solution by developing the upper layers that are not covered by the IEEE 802.15.4 standard.

ZigBee

ZigBee standards-based protocols that provide the network infrastructure required for wireless sensor network applications. ZigBee is the communications protocol (network and applications layer) that is based on IEE 802.15.4 MAC and PHY layers.

ZigBee mesh networks are ideal for some metering applications because of their inherent redundancy, self-configuring and self-healing capabilities. The interoperable nature of

ZigBee means that different applications can work together. This feature is however not unique to ZigBee.

The ZigBee specification provides a low-cost, low-power, wireless mesh networking technology. The low cost allows this technology to be widely deployed in wireless control and monitoring applications, the low power-usage allows longer life with smaller batteries, and the mesh networking provides high reliability and larger range.

ZigBee is a specification for a suite of high level communication protocols using small, low power digital radios based on the IEEE 802.15.4 standard for wireless personal area networks (WPANs). ZigBee devices are required to conform to the IEEE 802.15.4, which specifies the lower protocol layers—the physical layer (PHY), and the medium access control (MAC) portion of the data link layer (DLL).

The ZigBee application layer consists of the Application Support sub-layer (APS), the ZigBee Device Object (ZDO) and the manufacturer-defined application objects.

The responsibilities of the APS sub-layer include maintaining tables for binding, which is the ability to match two devices together based on their services and their needs, and forwarding messages between bound devices. Another responsibility of the APS sub-layer is discovery, which is the ability to determine which other devices are operating in the personal operating space of a device.

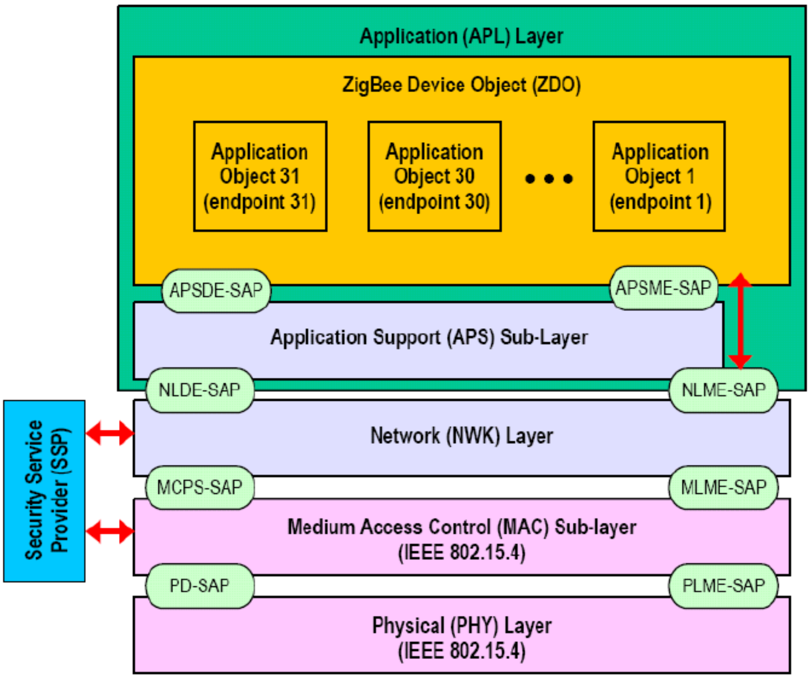

Application (APL) Layer
ZigBee Device Object (ZDO)
Application Object 31 (endpoint 31)
Application Object 30 (endpoint 30)
• • •
Application Object 1 (endpoint 1)
APSDE-SAP
APSME-SAP
Application Support (APS) Sub-Layer
NLDE-SAP
NLME-SAP
Security Service Provider (SSP)
Network (NWK) Layer
MCPS-SAP
MLME-SAP
Medium Access Control (MAC) Sub-layer (IEEE 802.15.4)
PD-SAP
PLME-SAP
Physical (PHY) Layer (IEEE 802.15.4)

Figure 1.76. ZigBee communication stack

The responsibilities of the ZDO include defining the role of the device within the network (e.g., ZigBee coordinator or end device), initiating and/or responding to binding requests and establishing a secure relationship between network devices. The manufacturer-defined application objects implement the actual applications according to the ZigBee-defined application descriptions

Depending on the implemented ZDO and manufacturer-defined applications, several application profiles can be defined.

6loWPAN

6loWPAN is an acronym of IPv6 over Low power Wireless Personal Area Networks. 6LoWPAN is the standard from the Internet Engineer Task Force IETF published in 2007, which optimises IPv6 for use with low-power, low-bandwidth communication technologies such as the IEEE 802.15.4.

The 6loWPAN group aimed at defining header compression mechanisms that allow IPv6 packets to be sent to and received from over IEEE 802.15-based networks. The base specification developed by the 6lowpan IETF group is the RFC 4944.

Whereas IEEE802.15.4 devices are intentionally constrained to reduce costs, IPv6 requires a considerable higher bandwidth. Therefore, header compression mechanisms standardized in RFC4944 can be used to provide header compression of IPv6 packets over such networks. Furthermore, RFC 4944 proposes an adaptation layer to allow the transmission of IPv6 datagrams over IEEE 802.15.4 networks. Moreover, since the compression is completely stateless, it means that it creates no binding state between the compressors any neighbour in compact form at all times.

6loWPAN brings the advantages already described in the IEEE 802.15.4 standard: offer the fundamental lower network layers of a type of wireless personal area network (WPAN), which focuses on low-cost, low-speed ubiquitous communication between devices; and the

IP protocol advantages which has proven itself a long-lived, stable, and highly scalable communication technology that supports both a wide range of applications, devices, and underlying communication technologies.

Finally, it shall be noted that the ZigBee specification and the 6lowpan RFC are not competing technologies but on the contrary, they can be well complementary as it can be shown in the diagram below:

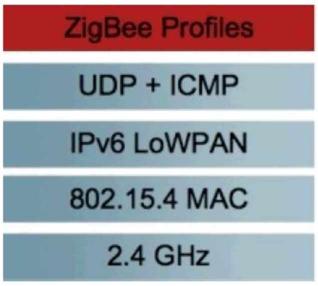


Figure 1.77. ZigBee over IPv6

Recently, the IETF has launched collaboration with the ZigBee alliance in order to approve an IETF specification for using ZigBee profiles over UDP/IP.

IEEE 802.11 (WLAN/WiFi)

In the last ten years, broadband wireless technologies based on the IEEE 802.11 standard have found broad acceptance worldwide for wireless local area networking (WLAN). These

WLAN technologies are designed to operate either in the 2.4 GHz or 5 GHz ISM frequency bands.

WLAN supports services similar to those offered by wired LANs (e.g. Ethernet) and can be used to build either stationary or mobile computer networks. Current WLAN systems provide data rates of typically 54 Mbit/s, with vendor-specific extensions reaching up to 108 Mbit/s.

Due to channel access and signalling overhead, roughly 50% of the physical layer data rates are available as actual user throughput.

Today, the IEEE 802.11 standard family is the most established. Its variants operate in the ISM-band, at either 2.4 GHz providing up to 11 Mbit/s, or in the 5 GHz range offering up to 54 Mbit/s.

WLAN equipment that is truly tested and certified by the Wireless Internet Compatibility Alliance (WICA) in accordance to the IEEE 802.11 family of wireless LAN standards is also known as WiFi (Wireless Fidelity).

Public WLANs (PWLAN) (WiFi hot spots) are deployed in many cities and areas as the most important commercial use of the WLAN technology today.

The use of WLANs for AMR has been frequently reported. In the US, several cities (e.g. Santa Clara, CA) already use WLANs to automatically collect readings from electricity, gas, and water meters. Since indoor coverage and particularly coverage in basements at frequencies above 2 GHz is relatively poor, WLAN-based AMR solutions require either meters mounted above ground level and outside wall of houses or an in-house wired or wireless local network with a WLAN equipped house gateway/data collector mounted at position well covered by an 802.11-based access network.

In cities or areas where there is already a public Wi-Fi network (e.g. a Metropolitan Area Network), the WLAN-based access solution may result in a more attractive business case, also taking into consideration that community-wide WiFi networks will become prevalent in future in many areas.

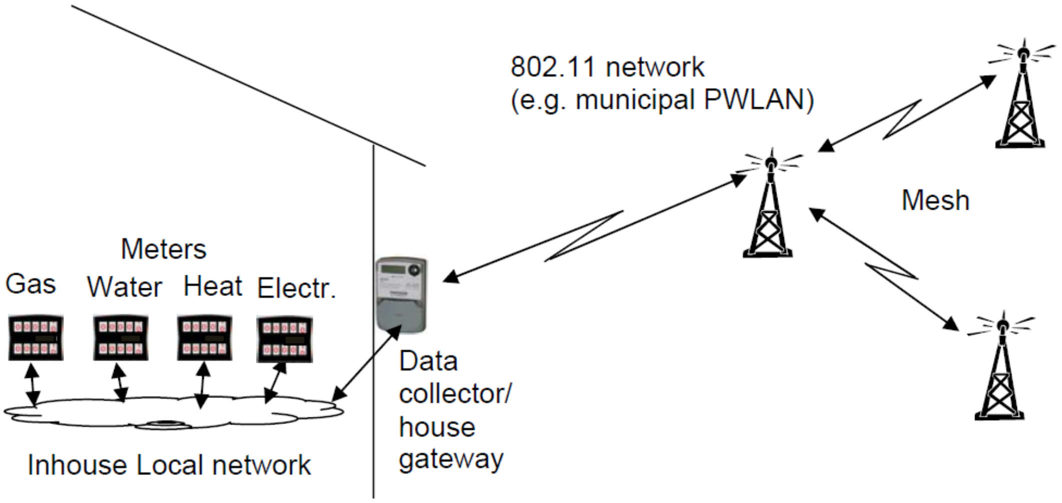


Figure 1.78. AMR using WLAN access network

Summary about wireless technologies

The LR-WPAN standard proves that the IEEE 802.15.4 standard is the most appropriate standard for such uses among other wireless connectivity standards, like IEEE 802.11 and Bluetooth. In similar applications, IEEE 802.11 Wireless Local Area Networks (WLANs) provide wireless connectivity but it is less suitable for resembling low cost applications and not cost-effective. Hence, the use of IEEE 802.11WLAN is overabundance technology. The second option can be Bluetooth technology; it is more suitable and replaces IEEE 802.11 WLAN in this low cost application. However, Bluetooth’s high complexity flaw is negative. These standards will require battery replacement so frequently, in few months, and battery charging in a day or so, they are almost impractical and unsuitable in temperature sensors or other LR-WPAN applications.

LR-WPAN device can operate on more than one frequency band, employing spreading parameters and data parameters. In summary, a general comparison of WLANs, WPANs and LR-WPANs is presented in Table 3:

Table 3. COMPARISON OF LR-WPAN WITH OTHER WIRELESS TECHNOLOGIES

Parameters	**WLAN** **(802.11)**	**WPAN** **(Bluetooth)**	**LR-WPAN** **(IEEE 802.15.4)**
Range	~100 m	~ 10-100 m	~ 10 m
Data throughput	~ 2-11 Mb/s	1 Mb/s	< 0.25 Mb/s
Power consumption	Medium	Low	Ultra-low
Size	Larger	Smaller	Smallest
Cost/complexity	> 6	1	0.2

The high level properties of LR-WPAN are summarized in Table 4.

Table 4. SUMMARY OF PROPERTIES OF LR-WPAN STANDARDIZATION (BITTNER M. ET AAL, 2009)

Property	**Range**

Raw data rate	868 MHz: 20kbps;
	915 Mhz: 40kbps;
	2.4 GHz : 250 kbps

Range Normally	10 cm to 10 m
Battery life	Application dependent , optimized for long battery life
Latency	10 to 15 ms
Channel Access	CSMA-CA and slotted CAMA-CA
Location Awareness	Optional
Nodes per network	Capable up to 65,534
Addressing Short	16-bit and 64-bit IEEE
Topology	Star and peer-to-peer (leads to mesh) topology
Complexity	Lower compared to current IEEE standard
Types of traffic	Data centric asynchronous communication
Temperature	Industrial temperature range -40 °C to +85 °C
Desired frequency band and channels	International band -868 MHz: 1 channel International band -915 MHz: 10 channels Unlicensed band-2.4 GHz: 16 channels

1.4.6. Open data/application models' standards

The data/application model describes the basic principles on which Objects are built. (An Object is a collection of attributes and methods.)It also gives a short overview on how interfaces are used for communication purposes. Data collection systems and metering equipment from different vendors, following these specifications, can exchange data in an interoperable way. However, a data/application model is not necessarily object based. ANSI C.12.19 e.g. uses tables.

Proprietary data/application models

Proprietary data/application models are typically created and maintained by equipment manufacturers. It is up to the manufacturer to make the models publically available or to make them available only to selected partners and to protect them by IP rights. Because there is no "public organization" guaranteeing proper maintenance of these models, typically the specifications are not stable and may vary with equipment type changes.

Interoperability with proprietary data/application models may be achieved on central system level by providing different "drivers" to handle equipment from different manufacturers (or even different types of the same manufacturer). This solution is not very cost effective but guarantees a recurring business for the central system providers.

Open data/application models standards: IEC 62056-61/62 COSEM

The application models cover classical AMR and smart metering applications for Electricity Gas, heat and Water. Besides the application models the standards also contain a unified object identification system (OBIS). The standards are permanently maintained by the DLMS-User Association considering the latest market needs and technology developments.

The COSEM standards define the applications independently of the supporting communication media and protocols. The use of the COSEM standards on PSTN or GSM IP-based communication channels as well as for PLC communication is specified in the IEC 62056 series.

The DLMS/COSEM offers three step approach that provides "orthogonality" such as separation of a model and protocols:

- Data model, to view the meter functionality at its interface(s) for COSEM objects and OBIS Object Identification System
- Messaging method to communicate with the model and to represent data as a series of bytes (APDUs)
- Transportation method to carry the messages over the media between the meter and remote parties

Object modelling assumes that any real-world things can be described by some attributes, when the attribute' name identifies the data, hence each attribute has a meaning, a data type and a value range. Methods in its turn allow performing operations on attributes, furthermore, attributes and methods constitute an object (Fig.1.79). Using the object means one can read or write the attributes and to invoke the methods.

The COSEM server model can be structured into three hierarchical levels: Physical device, Logical device and Accessible COSEM objects.

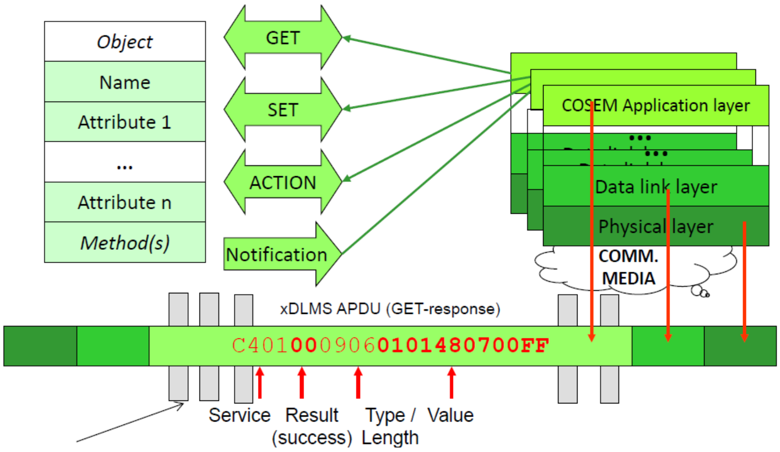


Figure 1.79. Building the messages services to access the objects and protocols to transport the information (Kmethy, Fuchs, Varjú, Roelofsen, 2015)

1.5. Solution of problems caused by Layer 1 Redundancy of communication channels

1.5.1. Reliability of systems with redundancy of communication channels

Unlike wireless communication media, the use of a cable when transporting network traffic leads to a conflict between two characteristics of networked environments: the reliability of data transmission and the speed of network traffic.

The expectation is that the network is always available to users who rely on it. This requires a network architecture that is built to be fault tolerant. A fault-tolerant network is one that limits the effect of a failure, so that the fewest number of devices are affected by it. It is also built in a way that enables quick recovery when such a failure occurs. Fault-tolerant networks depend on multiple paths between the source and destination of a message. If one path fails, the messages can be instantly sent over a different link. Having multiple paths to a destination is known as redundancy, as shown in Figure 1.80. [1]

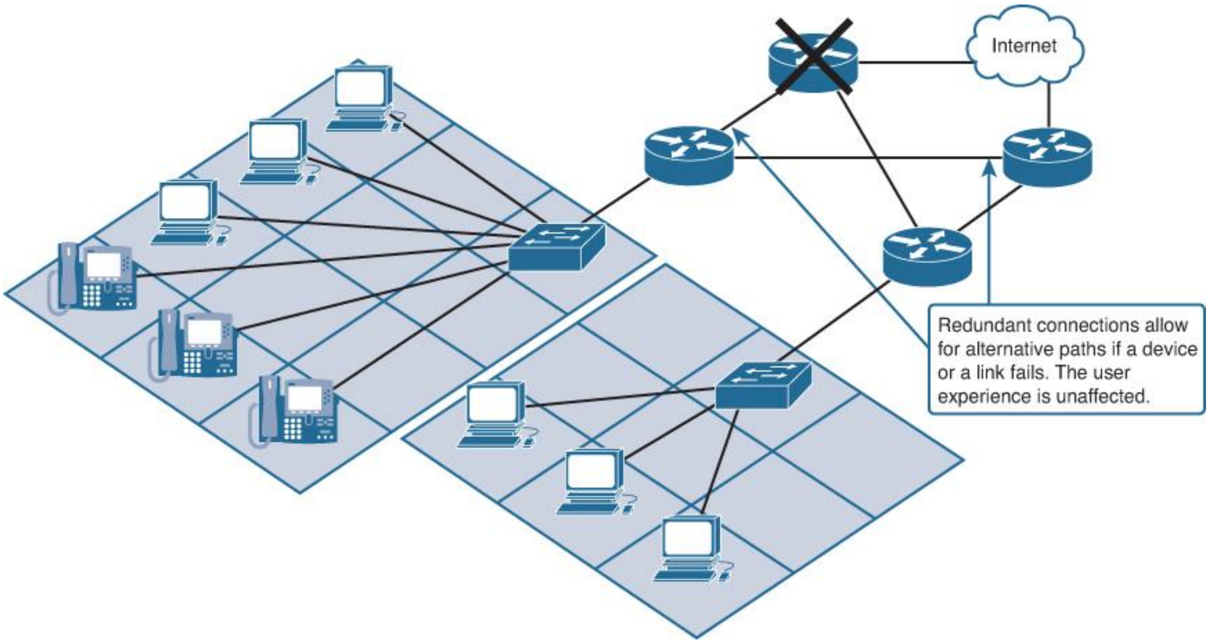

Internet
Redundant connections allow for alternative paths if a device or a link fails. The user experience is unaffected.

Figure 1.80. Fault Tolerance

To understand the need for redundancy, we can look at how early telephone systems worked. When a person made a call using a traditional telephone set, the call first went through a setup process. This process identified the telephone switching locations between the person making the call (the source) and the phone set receiving the call (the destination). A temporary path, or circuit, was created for the duration of the telephone call. If any link or device in the circuit failed, the call was dropped. To reconnect, a new call had to be made, with a new circuit. This connection process is referred to as a circuit-switched process and is illustrated in Figure 1.81.

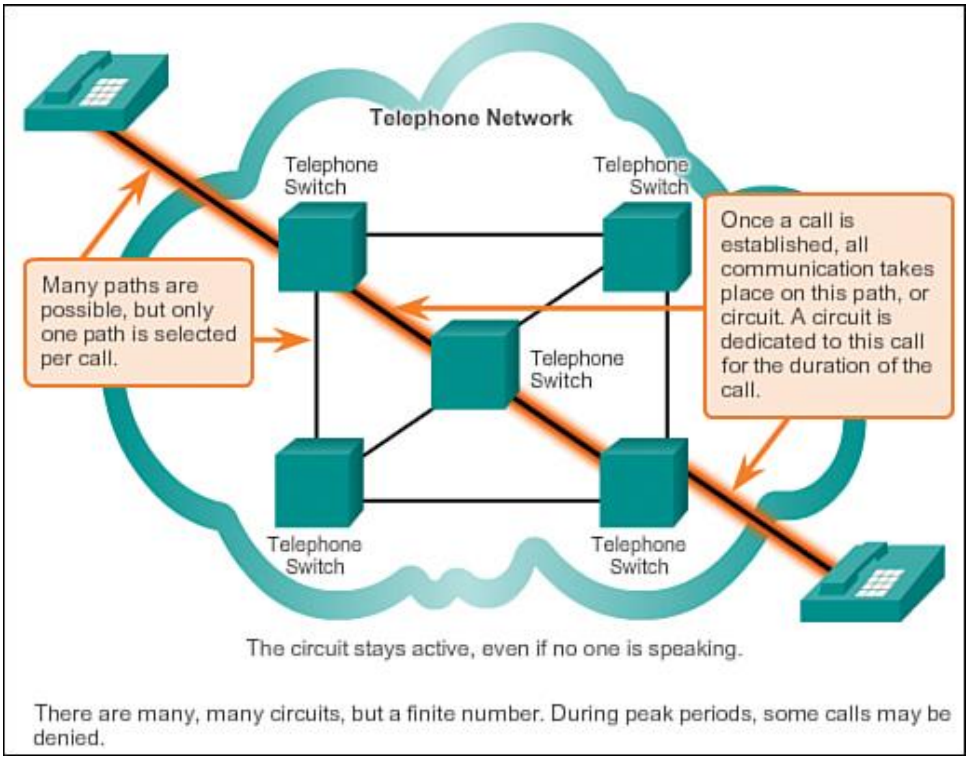

Telephone Network
Telephone Switch
Telephone Switch
Telephone Switch
Telephone Switch
Telephone Switch
Many paths are possible, but only one path is selected per call.
Once a call is established, all communication takes place on this path, or circuit. A circuit is dedicated to this call for the duration of the call.
The circuit stays active, even if no one is speaking.
There are many, many circuits, but a finite number. During peak periods, some calls may be denied.

Figure 1.81. Fault Tolerance

Many circuit-switched networks give priority to existing circuit connections at the expense of new circuit requests. After a circuit is established, even if no communication is occurring between the persons on either end of the call, the circuit remains connected and resources are used until one of the parties disconnects the call. Because there are only so many circuits that can be created, it is possible to get a message that all circuits are busy and a call cannot be placed. The cost to create many alternative paths with enough capacity to support a large number of simultaneous circuits, and the technologies necessary to dynamically re-create dropped circuits in the event of a failure, are why circuit-switched technology was not optimal for the network.

1.5.2. Increase in the number of copies of transmitted data

The role of switched networks has evolved dramatically in the last two decades. It was not long ago that flat Layer 2 switched networks were the norm. Flat Layer 2 switched networks relied on the Ethernet and the widespread use of hub repeaters to propagate LAN traffic throughout an organization. As shown in Figure 1, networks have fundamentally changed to switched LANs in a hierarchical network. A switched LAN allows more flexibility, traffic management, and additional features:

- Quality of service.
- Additional security.
- Support for wireless networking and connectivity.
- Support for new technologies, such as IP telephony and mobility services.

Figure 1.82 shows the hierarchical design used in the borderless switched network.

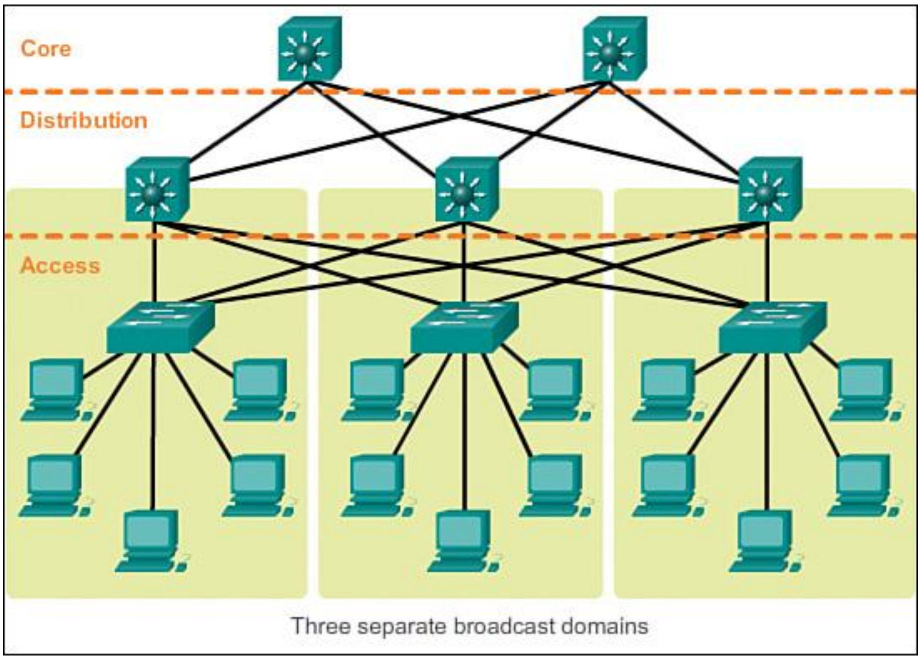


Figure 1.82. Hierarchical Network

It can be seen that in switched networks (Layer 2 ISO/OSI Model), switches can have redundant links Layer1 to each other. This redundancy is good because it minimizes downtime, but it may result in broadcasts continuously circling the network, which is called a broadcast storm. [2]

A collection of interconnected switches forms a single broadcast domain. Only a network layer device, such as a router, can divide a Layer 2 broadcast domain. Routers are used to segment broadcast domains, but will also segment a collision domain.

When a device sends a Layer 2 broadcast, the destination MAC address in the frame is set to all binary ones.

The Layer 2 broadcast domain is referred to as the MAC broadcast domain. The MAC broadcast domain consists of all devices on the LAN that receive broadcast frames from a host.

When a switch receives a broadcast frame, it forwards the frame out each of its ports, except the ingress port where the broadcast frame was received. Each device connected to the switch receives a copy of the broadcast frame and processes it. Broadcasts are sometimes necessary for initially locating other devices and network services, but they also reduce network efficiency. Network bandwidth is used to propagate the broadcast traffic. Too many broadcasts and a heavy traffic load on a network can result in congestion, which slows down network performance.

When two switches are connected together, the broadcast domain is increased. In this case, the broadcast frame is sent to all the connected ports on the switch, then it returns through the "loop" and is sent again, as shown in Figure 1.83.

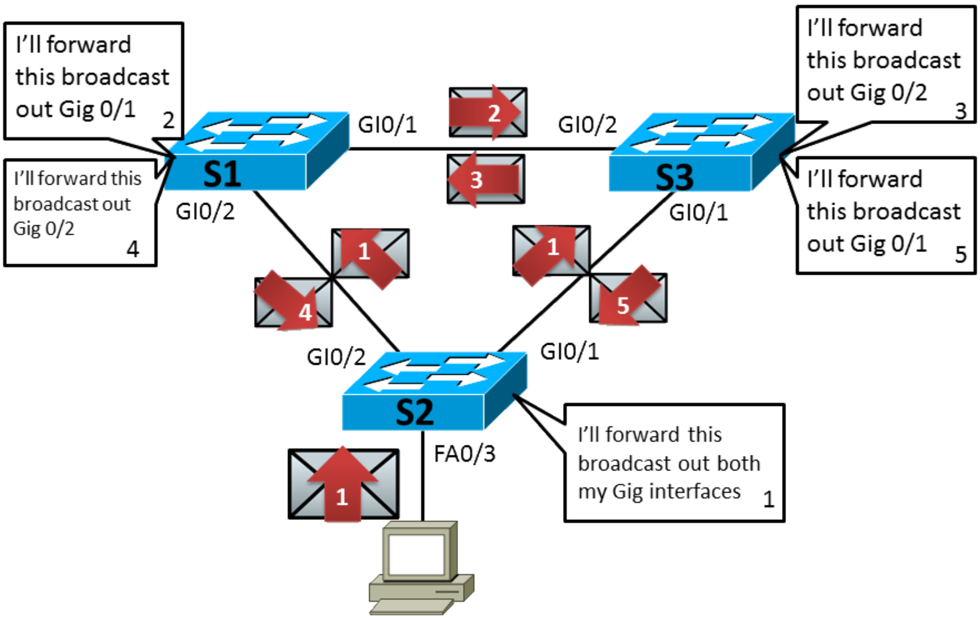

I'll forward this broadcast out Gig 0/1 2
I'll forward this broadcast out Gig 0/2 4
I'll forward this broadcast out Gig 0/2 3
I'll forward this broadcast out Gig 0/1 5
I'll forward this broadcast out both my Gig interfaces 1
S1
S2
S3
GI0/1
GI0/2
GI0/2
GI0/1
GI0/2
GI0/1
FA0/3
1
2
3
4
5

Figure 1.83. Broadcast storm

Ethernet frames do not have a time to live (TTL) attribute. As a result, if there is no mechanism enabled to block continued propagation of these frames on a switched network, they continue to propagate between switches endlessly, or until a link is disrupted and breaks the loop.

Broadcast frames are forwarded out all switch ports, except the original ingress port. This ensures that all devices in a broadcast domain are able to receive the frame. If there is more than one path for the frame to be forwarded out of, an endless loop can result. When a loop occurs, it is possible for the MAC address table on a switch to constantly change with the updates from the broadcast frames, which results in MAC database instability.

1.5.3. Spanning Tree Algorithm

To control the redundancy of communication channels, the spanning tree protocol (STP) family are used: STP (1998/2004), PVST +, Rapid PVST +, MSTP (Multiple STP), SPB (Shortest Path Bridging). [3]

STP ensures that there is only one logical path between all destinations on the network by intentionally blocking redundant paths that could cause a loop. A port is considered blocked when user data is prevented from entering or leaving that port. This does not include bridge protocol data unit (BPDU) frames that are used by STP to prevent loops. Blocking the redundant paths is critical to preventing loops on the network. The physical paths still exist to provide redundancy, but these paths are disabled to prevent the loops from occurring. If the path is ever needed to compensate for a network cable or switch failure, STP recalculates the paths and unblocks the necessary ports to allow the redundant path to become active.

To control the redundancy of communication channels, the spanning tree protocol family STP, PVST+, Rapid PVST+, MSTP (MultipleSTP), SPB (Shortest Path Bridging).

IEEE 802.1D STP and RSTP use the Spanning Tree Algorithm (STA) to determine which switch ports on a network must be put in blocking state to prevent loops from occurring. The STA designates a single switch as the root bridge and uses it as the reference point for all path calculations. In the figure 1.84, the root bridge (switch S1) is chosen through an election process. All switches that are participating in STP exchange BPDU frames to determine which switch has the lowest bridge ID (BID) on the network. The switch with the lowest BID automatically becomes the root bridge for the STA calculations.

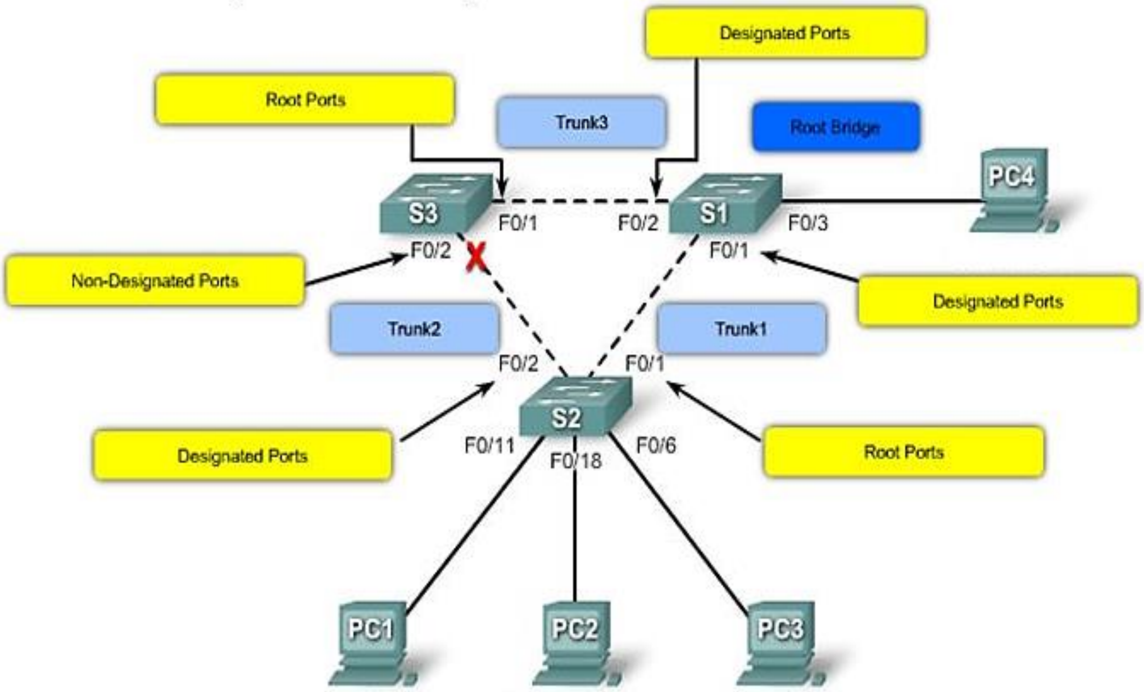


Figure 1.84. Port roles in Spanning Tree Algorithm

The BID is made up of a priority value, an extended system ID, and the MAC address of the switch. The bridge priority value is automatically assigned but can be modified. The extended system ID is used to specify a VLAN ID or a multiple spanning tree protocol (MSTP) instance ID. The MAC address field initially contains the MAC address of the sending switch.

All switches in the broadcast domain participate in the election process. After a switch boots, it begins to send out BPDU frames every two seconds. These BPDUs contain the switch BID and the root ID.

The switch with the lowest BID will become the root bridge. At first, all switches declare themselves as the root bridge. Eventually, the switches exchange BPDUs, and agree on one root bridge.

As the switches forward their BPDU frames, adjacent switches in the broadcast domain read the root ID information from the BPDU frames. If the root ID from a BPDU received is lower than the root ID on the receiving switch, then the receiving switch updates its root ID, identifying the adjacent switch as the root bridge. However, it may not be an adjacent switch. It could be any other switch in the broadcast domain. The switch then forwards new BPDU frames with the lower root ID to the other adjacent switches. Eventually, the switch with the lowest BID ends up being identified as the root bridge for the spanning tree instance.

There is a root bridge elected for each spanning tree instance. It is possible to have multiple distinct root bridges for different sets of VLANs. If all ports on all switches are members of VLAN 1, then there is only one spanning tree instance. The extended system ID includes the VLAN ID, and plays a role in how spanning tree instances are determined.

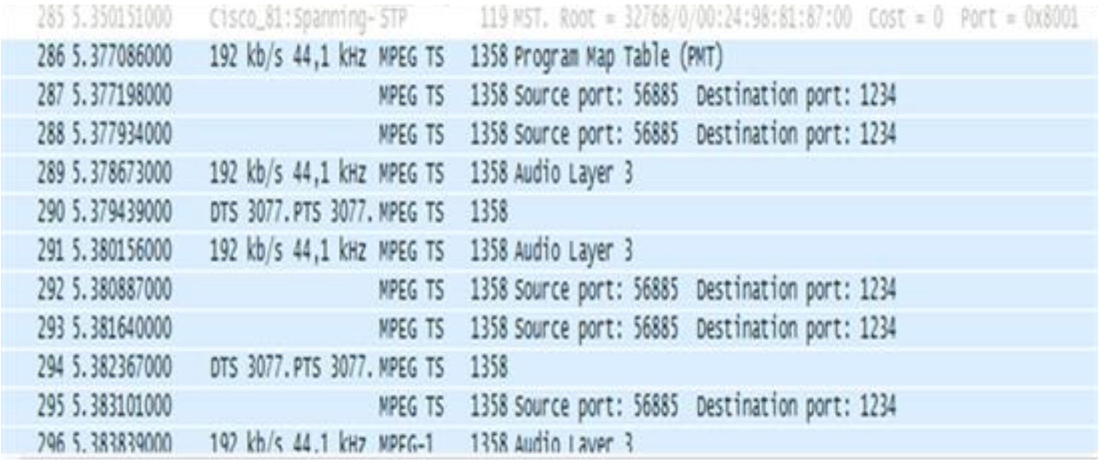
The BID consists of a configurable bridge priority number and a MAC address. Bridge priority is a value between 0 and 65,535. The default is 32,768. If two or more switches have the same priority, the switch with the lowest MAC address will become the root bridge.

An additional factor to consider when working with STP protocols is the device preparation time. It is increased to collect preliminary information about the network topology, the distribution of roles between the device ports and the assignment of their modes. This delay can affect the performance of a number of network services (for example, DHCP distribution).

Unmanaged devices do not offer the network service personnel to change the internal algorithm of operation, therefore it is recommended to use intelligent switching devices with console control or through remote connection mode to realize the redundancy of the cable structure.

Let's build on the allocated equipment stand the redundant topology of the network, similar to that shown in Figure 1.84. In the experiment, switches from several manufacturers participate in monovendor communication schemes and multivendor combinations.

Data exchange between two clients is monitored in a data transfer medium that is free from other traffic. As a test load, we used streaming video packages. Traffic analysis and time slots were carried out using the Wireshark network statistics collection package, in Figure 1.85 you can see an example of tracking packets.



```
285 5.350151000   Cisco_81:Spanning-STP      119 MST. Root = 32768/0/00:24:98:81:87:00  Cost = 0  Port = 0x8001
286 5.377086000   192 kb/s 44,1 kHz MPEG TS  1358 Program Map Table (PMT)
287 5.377198000                     MPEG TS  1358 Source port: 56885  Destination port: 1234
288 5.377934000                     MPEG TS  1358 Source port: 56885  Destination port: 1234
289 5.378673000   192 kb/s 44,1 kHz MPEG TS  1358 Audio Layer 3
290 5.379439000   DTS 3077.PTS 3077.MPEG TS  1358
291 5.380156000   192 kb/s 44,1 kHz MPEG TS  1358 Audio Layer 3
292 5.380887000                     MPEG TS  1358 Source port: 56885  Destination port: 1234
293 5.381640000                     MPEG TS  1358 Source port: 56885  Destination port: 1234
294 5.382367000   DTS 3077.PTS 3077.MPEG TS  1358
295 5.383101000                     MPEG TS  1358 Source port: 56885  Destination port: 1234
296 5.383839000   192 kb/s 44,1 kHz MPEG-1   1358 Audio Layer 3
```

Figure 1.85. Example of statistics gathered by Wireshark

At carrying out of check of system response to rupture of the shortest path between the two devices when transferring data from Windows to Linux and the opposite situation when working with protocols such as PVST+, Rapid PVST+, and MSTP were measured delay that occurred at this crossing de alternative channel after failure of the primary and the transition to primary after a failure alternative for each of these protocols, as shown in Figure 1.86.

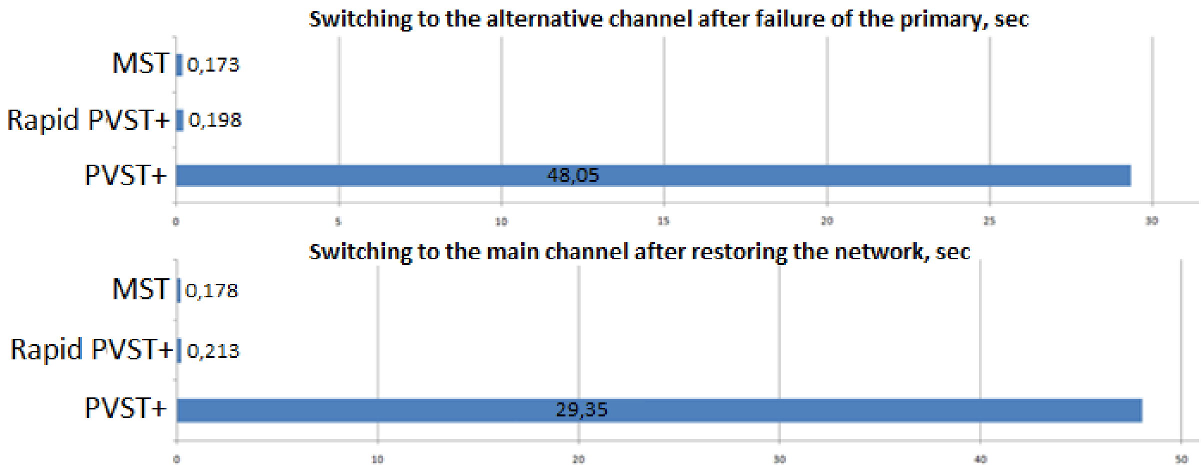


Figure 1.86. The delay of the communication channel in the protocols PVST+, Rapid PVST+, MSTP when connection is restored

PVST+ is based on the old implementation of STP, he gives the worst result in time delays. Protocols Rapid PVST+ and MSTP is based on a new protocol version is RSTP give much better results, many times superior to RSTP. Given the fact that Rapid PVST+ is a proprietary protocol and requires considerable resources from the switch, MSTP is a standardized protocol which can divide VLAN groups to reduce resource consumption, its use in a real network is more preferable option.

When building a switched network by using the redundancy of communication channels should use the new protocols of Spanning tree, as the old protocols do not provide the proper speed of transition to an alternative channel. You should also pay attention to the configure switches, since with incomplete adjustment, even in new protocols of Spanning tree speed the transition to alternative channel will be reduced significantly and will approach the speed of legacy protocols, which is unacceptable in current networks, where the slightest disruption in the network will have serious consequences.

Consider the Rapid PVST + protocol. Within this protocol, an independent instance of Rapid STP operates in each VLAN. In a properly configured network, RSTP can achieve convergence much faster, sometimes in just a few hundred milliseconds. The RSTP protocol again determines the types of ports and their states. If the port is configured as an alternative or backup, it can immediately go to the forwarding state, without waiting for the network to converge. Also an important parameter is portfast, which is indicated on the port connected to the end device. This option allows the port not to pass training and immediately go to the forwarding mode. Figures 1.87, 1.88 show a comparison of the convergence time for various switch settings and for different directions of the streaming video transmission.

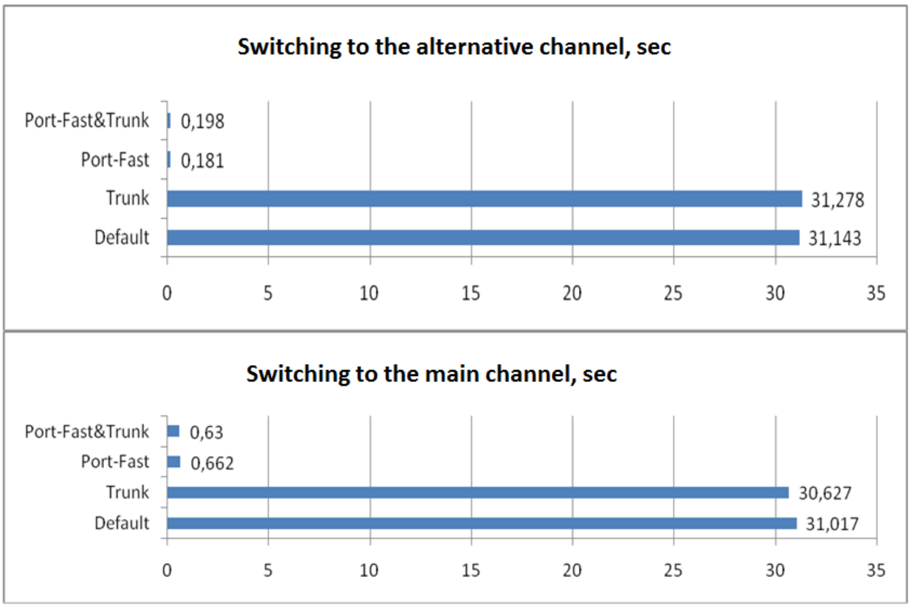


Figure 1.87. Transmission data frames from Windows10 to Linux Ubuntu 16.4

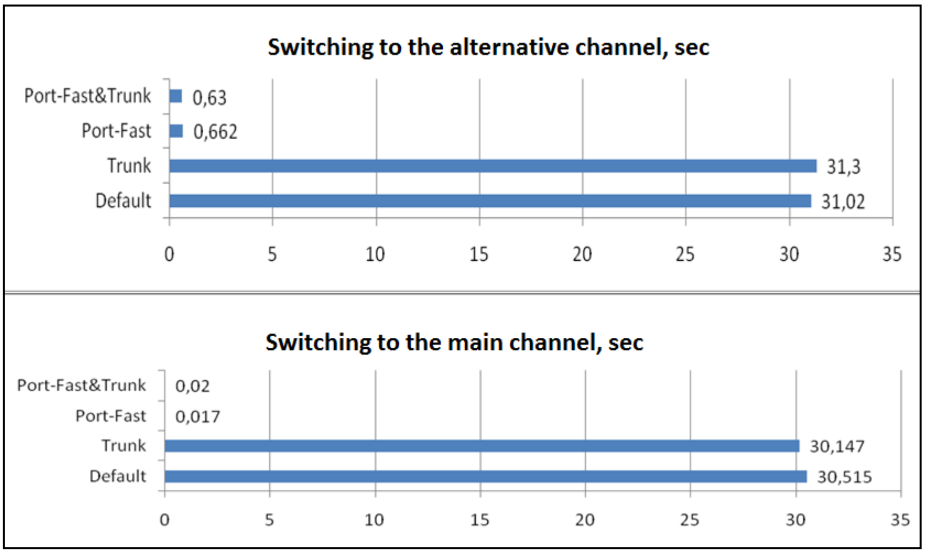


Figure 1.88. Transmission data frames from Linux Ubuntu 16.4 to Windows10

The results of the experiments showed that the time passing to the alternate channel and back depends on:

- the selected version of the protocol for negotiating the state of redundant ports between switches;
- combinations of devices from different vendors and versions of their embedded software;
- the direction of transmission of data frames and operational systems that carry out the reception and transmission of packages.

1.5.4. Aggregation of communication channels

The increase in the transmission rate of the data of the communication channel of the network is most often achieved by switching to a communication standard with high speed characteristics. To implement such a transition, new network devices are purchased, or the exchangeable module of operating network equipment changes to a new one, which supports higher speed characteristics. In addition, it is necessary to interfere with the cable network structure, if it does not meet the requirements of the implemented standard. This procedure requires a temporary shutdown of the computing process in the network and an additional period of adaptation of the network in new conditions (network convergence and traffic rebalancing).[3]

Another popular way to improve the data transfer speed is to aggregate two or more physical links between the network devices in order to summarize their throughput. The EtherChannel, Multi-Link Trunking (MLT), DMLT, SMLT, DSMLT, R-SMLT, and similar protocols aggregate their function to the STP family protocols, so blocking the communication channels of the group does not occur, as shown in Figure 1.89.

For example, EtherChannel reduces part of the binary pattern that the addresses in the frame form to a numerical value that selects one of the links in the channel in order to distribute frames across the links in a channel. EtherChannel frame distribution uses a specialized hashing algorithm. The algorithm is deterministic; if you use the same addresses and session information, you always hash to the same port in the channel. This method prevents out-of-order packet delivery.

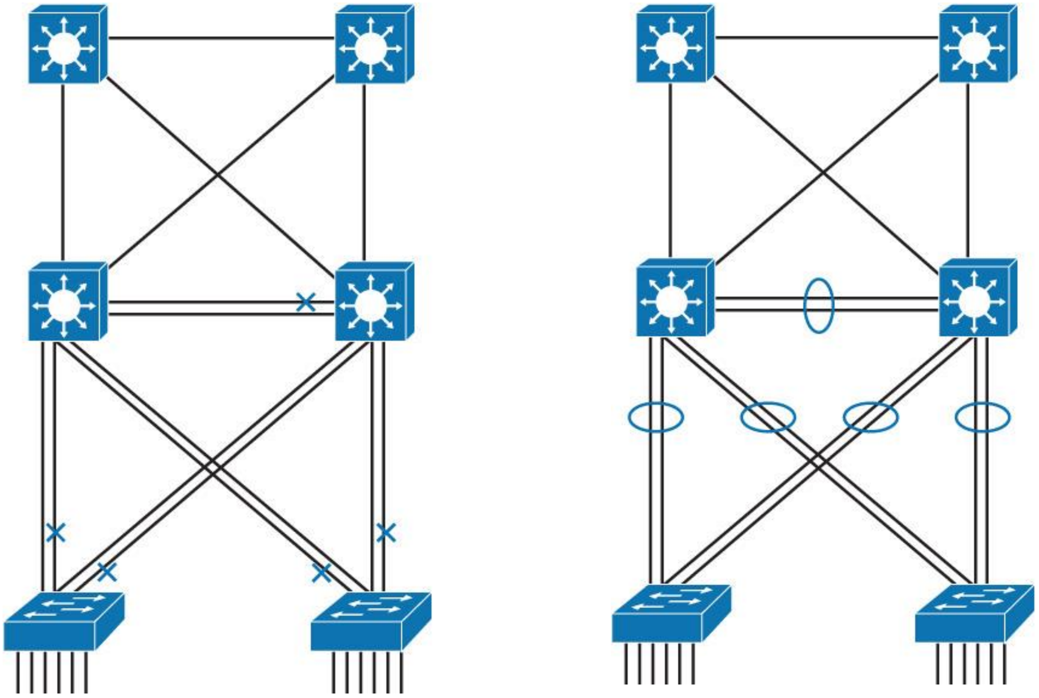


Figure 1.89. EtherChannel integration in the network

Aggregation of communication channels has a higher reliability than "acceleration" of the channel to higher speeds. The group can include backup channels, which will automatically activate if any of the main channels fail.

Correct balancing of traffic between the channels of the group allows to provide an optimal route for moving data frames to the destination and minimizing the loss of frames in case of failure.

1.5.5. Aggregation of communication devices

Distribution of channels of the group of aggregated communication channels can be performed both on a single device and in a group of devices. In this case, the task of increasing the speed of the communication channel is solved with simultaneous increase in the level of reliability of the network, since the channel is allocated for simultaneous support by two or more devices. [3]

The following combinations of communication channel combinations are proposed:

- LAG (Link aggregation) – the port group includes the ports of only one switching device;
- MC-LAG (Multi-Chassis Link Aggregation Group) the port group includes the ports of several switching devices located on one side of the connection, while the switching devices on the opposite side are connected by a strict sequence of ports;
- High Availability MCLAG – the port group includes the ports of several switching devices located on one side of the connection, the switching devices on the opposite side being connected by a cross-section of the ports between the various devices.

An example of combining aggregation modes is shown in the figure 1.90.

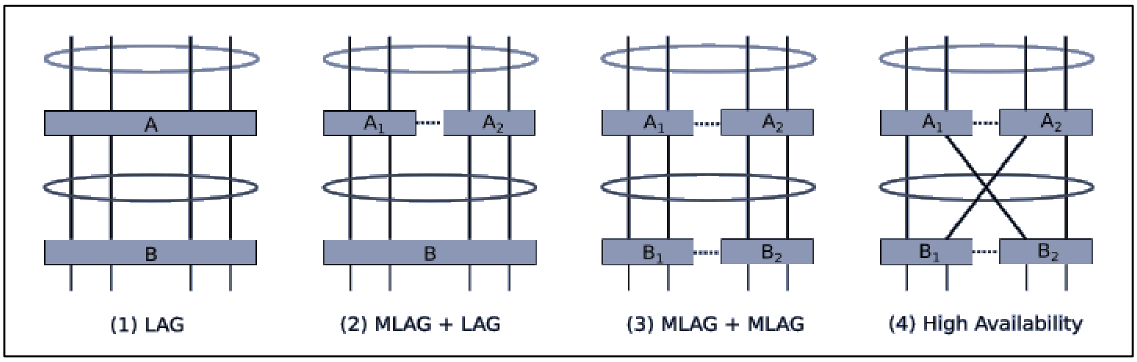


Figure 1.90. Port roles in Spanning Tree Algorithm

References:

[1] http://www.ciscopress.com/articles/article.asp?p=2158215&seqNum=5

Exploring the Modern Computer Network: Types, Functions, and Hardware

By Cisco Networking Academy.

Sample Chapter is provided courtesy of Cisco Press.Date: Dec 19, 2013.

[2] http://www.ciscopress.com/articles/article.asp?p=24101

Cisco Network Topologies and LAN Design

By Anthony Bruno, Jacqueline Kim.

Sample Chapter is provided courtesy of Cisco Press.Date: Nov 16, 2001.

[3] Воруев, А.В. Выбор протоколов дублирования и агрегирования каналов связи в гетерогенной сети передачи данных // Воруев А.В. / IV Международная научная конференция Проблемы взаимодействия излучения с веществом. (Гомель, 9-11 ноября 2016 года): в 2 ч. Ч. 2. – Гомель : ГГУ, 2016. – с.209-213

Chapter 2:Control Theory

2.1. Basics of Control Theory

The term „automatic control” means the process of a technical object control without a human involvement. Additionally, the object should be able to perceive the control signals containing the information about the further object condition. In most of the cases these signals are generated on the basis of the information about the current condition of the object. Therefore, the flows of information and their relations are basic in the processes of automatic control.

Control theory is an interdisciplinary branch of engineering and mathematics that deals with the behaviour of dynamic systems with inputs, and how their behaviour is modified by feedback. The usual objective of control theory is to control a system, often called the plant, so its output follows desired control signal, called the reference, which may be a fixed or changing.

Although a major application of control theory is in control of engineering systems dealing with the design of process control systems for industry, other applications range far beyond this. As the general theory of feedback systems, control theory is useful wherever feedback occurs. A few examples are in physiology, electronics, climate modelling, machine design, ecosystems, navigation, neural networks, production theory. The control theory is an engineering science, which describes control processes in technical applications. It is a subpart of the automation control as well as the control technique. A technical control process is a selective influence for example of physical or chemical values in a technical system. Even if there is a disturbance in the system, the so named control variables should be tried to stay constant (constant value regulation) or to influence it in a way to keep it in a default temporal change (follow-up control). One familiar application for example is the cruise control to keep the speed in a car constant.

Control systems may be thought of as having four functions: measure, compare, compute and correct. These four functions are completed by five elements: detector, transducer, transmitter, controller and final control element. The measuring functions completed by detector, transducer and transmitter. The compare and compute functions are completed within the controller, which may be implemented electronically by proportional, PI, PID controller, hysteretic control or programmable logic controller.

To focus on the main topic, this subtopic will only be slightly introduced. In automation technologies the regulation is, beside the control, also important. For regulation the input- and output-values can be analogue or digital. Dependent on these two cases, it will be differentiated among analogue and binary regulations. Since the analogue regulations predominantly are used in the control theory, the binary regulations are primarily used in control technology.

Analogue regulation

In contrast to a binary or a digital signal an analogue signal may have a keen run and theoretically it can take unlimited values in the dynamic area.

Binary regulation

The term of control technology is seldom based on analogue regulations but on the treatment of control devices with many binary in- and outputs. The binary regulations with two-valued signals are differing in combinative and sequential behaviour. Combinative behaviour means logical binary sign-combination.

Advantages of regulations

- Known and unknown not measurable disturbances were reduced or eliminated - Approximated models for the analysis of a control loop are sufficient

Disadvantages of regulations

- The control loop can suffer from undesired change of parameters like from deterioration - Exact and fast measurements of the control variable can be expensive

Advantages of continuous controls

- If an interference happen, the process can manually be operated - No instable behaviour or excessive amplitude of the control variable are possible

Disadvantages of continuous controls

- Only known and measurable interferences can be compensated through appropriate action - There´s no feedback if the command variable u(t) are reached through the control factor y(t)

2.2. Concept of Transfer Function

Analysis of the quality of the systems automated control is based on the investigation and evaluation of their static and dynamic characteristics.

The static characteristics reflect the correlation between input and output in steady-state condition. The systems with one input and one output are described with the only static characteristic:

$$X_{st}=Ay_{st}, \tag{2.1}$$

where A is an amplification factor. For linear systems A=const, for non-linear A=f(y).

The systems with more inputs are described with a set of static functions. If a system has n inputs, then for an i-th input its reaction for an j-th output can be:

$$X_{st}=A_{i,j}y_{i\ st}, \tag{2.2}$$

where $y_{i\ st}$ =var.

The dynamic characteristics reflect the transient processes of the system as a reaction for different types of inputs. The dynamic characteristics can be formed on the basis of transfer function of the system.

Transfer function is a ratio of Laplace operation forms of the output and input values at zero initial conditions. For one input system transfer function is determined:

$$F(s)=\frac{\bar{x}}{\bar{y}}\ , \tag{2.3}$$

where $\bar{x} = X(s), \bar{y} = Y(s)$ are the Laplace operation forms of input and output values at zero initial conditions.

With more inputs the transfer function of type (2.3) is obtained for each input assuming that rest of the inputs are zero.

The dynamic properties of the systems are also estimated with the help of the plot of frequency response. This property characterises the response of the element onto an input harmonic disturbance like the following:

$$y(t)=A_{in}\ (\omega)\ e^{j(\omega t+\phi_{in}(\omega))}, \tag{2.4}$$

with the changing of frequency ω from zero to ∞. The frequency response is usually applied for stability investigation as well as for forming of transient processes of the automated control systems. Each typical dynamic controller can be represented with its own frequency response reflecting its dynamic properties. Frequency response is represented as a dependence of phase and relative value of output signal amplitude on the frequency of input disturbance oscillations. The amplitude value is a ratio of output and input signals (voltages) amplitudes. For example, if the input of an aperiodic controller is supplied with sine-form voltage (fig.2.1) with frequency of ω_1 and amplitude U_{in} then at the output we will get a sine-form voltage of the same frequency but amplitude U_{out} shifted in phase for angle ϕ_1.

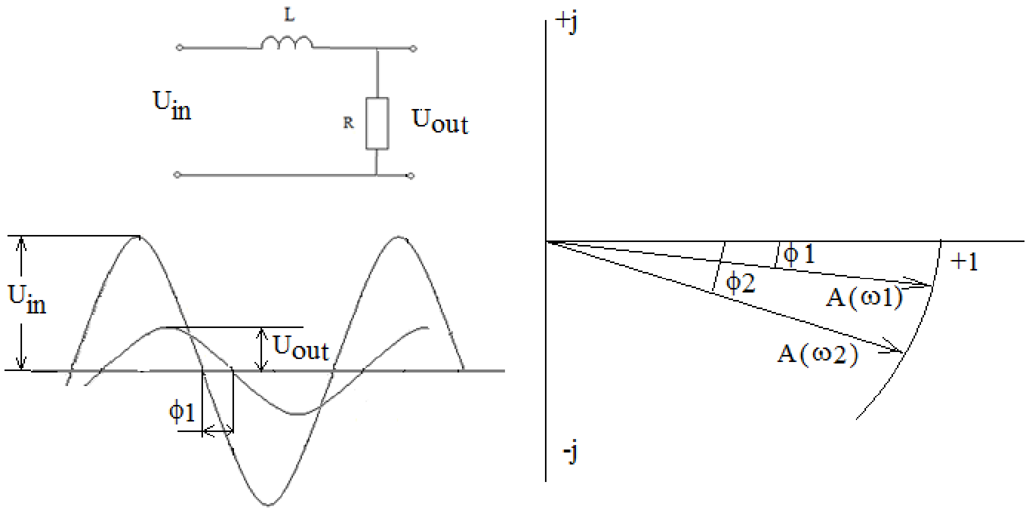


Figure 2.1. To the notion of graphic representation of frequency response

Parametric plot of frequency response (Nyquist plot). Frequency and phase responses of control systems in open and close condition can be built on the basis of frequency response of the dynamic controllers the system consists of. Fig. 2.2 presents a single loop system the transfer function of which is

$$\Phi(s) = \frac{K(s)}{1+W(s)}, \tag{2.5}$$

where $W(s) = K(s)K_f(s)$

is a transfer function of the system open in point B. Replacing operator s for jω in transfer function $W(s) = K(s)K_f(s)$ the frequency response of the open-loop system is

$$W(j\omega) = K_1(j\omega)K_2(j\omega)\dots K_n(j\omega)K_f(j\omega) =$$

$$A_1(\omega)e^{j\varphi_1(\omega)}A_2(\omega)e^{j\varphi_2(\omega)}\dots A_n(\omega)e^{j\varphi_n(\omega)}A_f(\omega)e^{j\varphi_f(\omega)} = A_p(\omega)e^{j\varphi_p(\omega)}, \tag{2.6}$$

where $A_p(\omega) = A_1(\omega)A_2(\omega)\dots A_n(\omega)A_f(\omega)$;

$$\varphi_p(\omega) = \varphi_1(\omega)\varphi_2(\omega)\dots\varphi_n(\omega)\varphi_f(\omega)$$

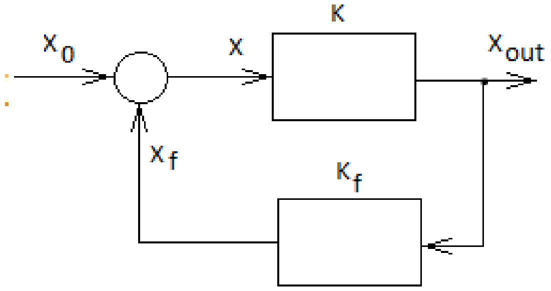


Figure 2.2. Example of a single loop system

Parametric plot of frequency response (Nyquist plot) is the ratio of complex expressions

$$\Phi(j\omega) = \frac{\dot{x}(t)}{\dot{y}(t)} \;, \tag{2.7}$$

where $\dot{x}(t) = A_{out}(\omega)e^{j(\omega t+\varphi_{out}(\omega))}$ is a harmonic output signal that traditionally written as

$$\Phi(j\omega) = A(\omega)e^{j\varphi(\omega)}. \tag{2.8}$$

A(ω) is an amplitude response:

$$A(\omega) = \frac{A_{out}(\omega)}{A_{in}(\omega)} = \frac{|\dot{x}(t)|}{|\dot{y}(t)|}, \tag{2.9}$$

ϕ(ω) – phase response:

$$\phi(\omega) = \phi_{out}(\omega) - \phi_{in}(\omega)\,. \tag{2.10}$$

Parametric plot of frequency response is a complex variable and can be illustrated as

$$\Phi(j\omega) = P(\omega) + jQ(\omega), \tag{2.11}$$

where P(ω) is a real frequency response, Q(ω) – an imaginary frequency response of the system.

For the investigation of dynamic properties such as stability of the systems the control theory applies Bode plot (graph of frequency response). They are widely used for determination of structure and parameters of the controllers forming an expected transient process of the system.

Applying logarithm for PAFC (2.4)

$$ln\Phi(j\omega) = lnA(\omega) + j\varphi(\omega)\,. \tag{2.12}$$

Expressions lnA(ω) and ϕ(ω) are the logarithm amplitude and phase response correspondingly. These properties are represented in graphic way. For estimation of the variables a logarithmic unit decibel (dB) is used. Values L and A are correlated as

$$L = 20lgA \text{ [dB]} \tag{2.13}$$

The investigation of transient processes of the systems is based on the application of differential or integral equations called dynamic equations forming a mathematical model of the control processes only. The mathematical model is formed by means of structural schemes consisting of dynamic elements. The dynamic elements are described with linearized differential equations. A full system of such differential equations of dynamic elements represents a mathematical model of a full physical system and serves to get its total differential equation.

2.3. Types of Controllers and Their Description

The character of transient process in a system of automatic control (SAC) depends on the dynamic properties of the elements it consists of. In accordance with the area of SAC application these elements can differ in purposes, construction, operation principles, etc. These elements can be machines, apparatuses, devices of different action – mechanical, electrical, hydraulic, pneumatic, etc. But independently on their real technical properties all these elements can be sorted into a limited number of elements with same dynamic properties and called typical dynamic elements.

Each dynamic element can transform physical values into other in one direction only – from input to output. The static characteristic of a element is a ratio of its output value to input value in steady-state condition. The dynamic elements can be linear and non-linear; static characteristics of the linear elements are represented as linear functions $x_{out}=f(x_{in})$ analytically or graphically, but non-linear elements – mostly with the help of graphical representation. Equation of a static characteristic of a linear element is a linear function (Fig.2.3)

$$x_{out} = x_0 + kx_{in} \ , \qquad (2.14)$$

where x_0 is an initial level of the output value with $x_{in}=0$; $k = \Delta x_{out}/\Delta x_{in}$ – is called a static amplification factor (transfer factor) of the element. When x=0 $= x_{out}/x_{in}$.

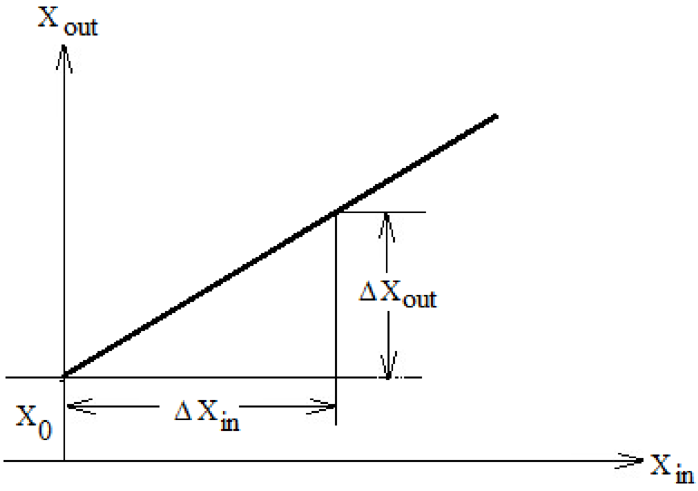


Figure 2.3. Static characteristic of a linear element

The dynamic properties of elements can be determined based on differential equation of the element behaviour in transient mode. A solution of differential equation allows obtaining transfer characteristic of dynamic element representing an output value dependence on time with a particular changing of input value in time. Differential equations are solved at zero initial conditions of the output value. Besides the output characteristic the dynamic properties of elements are also represented by other means, e.g. frequency, pulse characteristics, etc. (see § 2.2).

According to the character of transient processes, the typical dynamic elements are divided into proportional, aperiodic, oscillating, derivative, integral and delay elements.

1. Proportional controller.

Proportional controller with transfer function

$$W(p) = K_0\,, \tag{2.15}$$

multiplies an input signal without influencing its form, i.e. k(ω)=K_0=const and ϕ(ω)=0. If K_0 has no measurement unit then it is an amplification factor. If measurement unit occurs then it is a transfer factor. Proportional controller usually goes in combination with other elements.

In accordance with the transfer function:

$$K(j\omega) = k\,, \tag{2.16}$$

in complex system is determined with a point on a real axis located at k distance from initial zero coordinate (fig. 2.4, a). Equations of real and imaginary components of frequency response (fig.2.4, b) are

$$P(\omega) = k; Q(\omega) = 0 \tag{2.17}$$

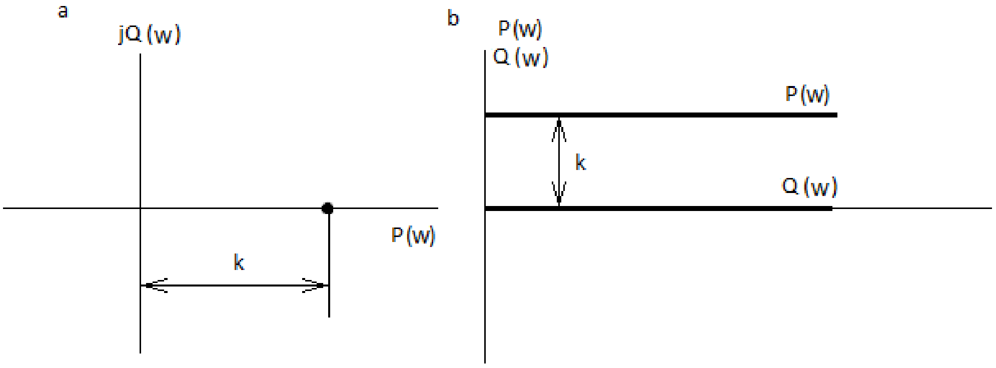


Figure 2.4. Parametric plot of frequency response (a), real and imaginary (b) variables plot

Taking a logarithm of (2.16) results in description of Bode plot.

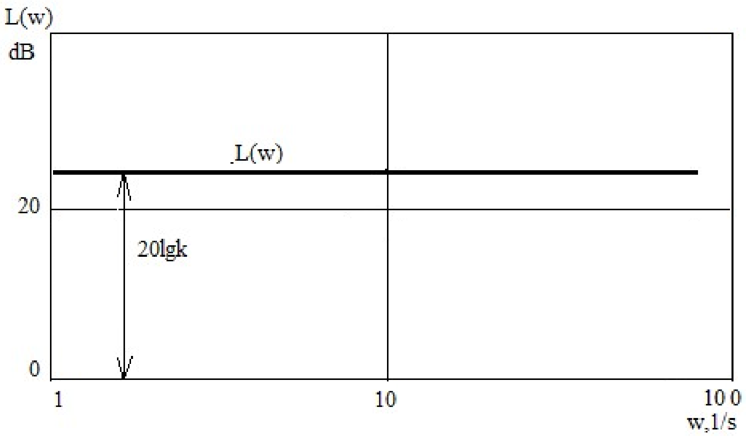


Figure 2.5. Bode plot of proportional controller

2. Aperiodic controller.

An aperiodic (or the first order inertia) controller has an output value changing in exponential dependence. This controller is described with a first-order differential equation:

$$T\frac{dx_{out}}{dt} + x_{out} = kx_{in}\,, \tag{2.18}$$

the Laplace transformation of which with zero initial conditions id represented as

$$(Tp+1)x_{out} = kx_{in}\ , \tag{2.19}$$

where T is a time constant of the controller.

Therefore, the transfer function of the controller is

$$W(p) = 1/(1+Tp). \tag{2.20}$$

Aperiodic controller can be associated with RL and RC circuits, magnetic amplifiers, DC generators, etc.

The following expression is used to get AFC of the aperiodic controller

$$K(j\omega) = \frac{k}{\sqrt{\omega^2T^2+1}}e^{-jarctg\omega T}\ , \tag{2.21}$$

where $\frac{k}{\sqrt{\omega^2T^2+1}}$ - is a modulus, arctg ωT – an argument of vector K(jω). Therefore, AFC is a circle with radius k/2 with a centre in point O located at k/2 distance from the axis initial point. With ω changed from 0 to ∞ vector K(jω) turns for an angle ϕ=-π/2. Real and imaginary frequency responses

$$P(\omega) = \frac{k}{\omega^2T^2+1}; Q(\omega) = -\frac{k\omega T}{\omega^2T^2+1} \tag{2.22}$$

are demonstrated in Fig.2.6, b.

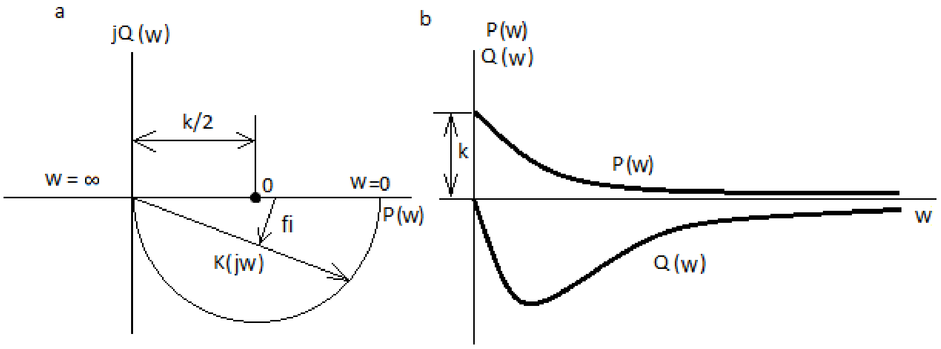


Figure 2.6. Parametric plot of frequency response, real and imaginary variables of aperiodic controller

Taking logarithm of (2.15) provides the following:

$$L(\omega) = 20lgk - 20lg\sqrt{\omega^2T^2+1};\ \varphi(\omega) = -arctg\omega T. \tag{2.23}$$

Considering the second component of the expression results in the following conclusions: while $\omega \ll 1/T$ and $\omega^2T^2 \ll 1$ the value $k(\omega)\approx 1$ $W(s) = K(s)K_f(s)$ and $k(\omega)[dB]\approx 0$. But if $\omega \gg 1/T$ and $\omega^2T^2 \gg 1$ the value $k(\omega)\approx(\omega^2T^2)^{-1/2}=1/\omega T$. Therefore, the bandwidth of the aperiodic controller is not limited at the left side, but starting with frequency $\omega_s=1/T$, $k(\omega)[dB]$ declines with -20dB/dec, but the phase bias is close to $-\pi/2$. Frequency ω_s is called junction frequency. Fig.2.5. Frequency ω_{st} at which LAC crosses the abscise is called frequency of cutoff.

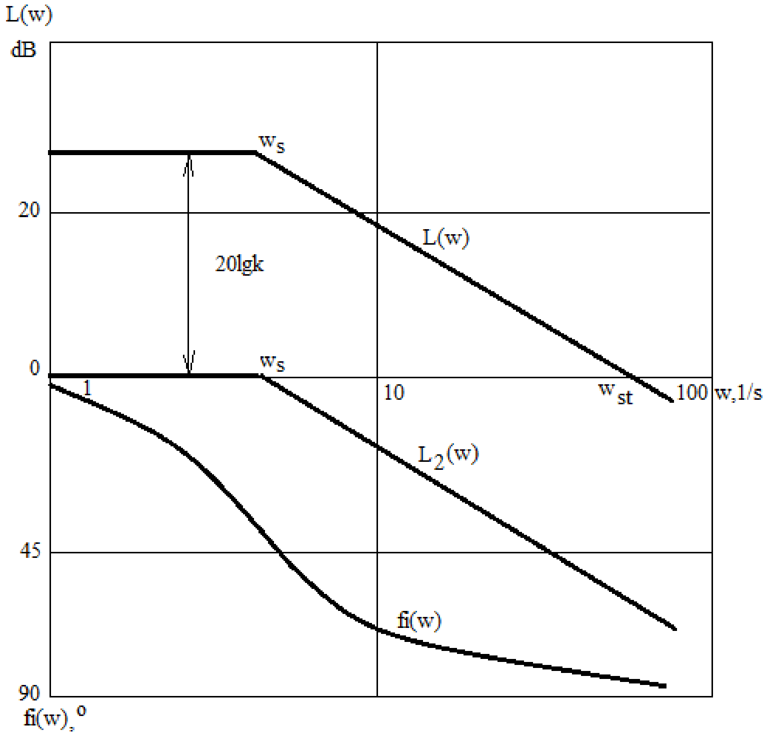


Figure 2.7. Bode plot of aperiodic controller

3. Derivative controller

The output value of derivative controller is proportional to the speed of input parameter changing

$$x_{out} = k\frac{dx_{in}}{dt}. \tag{2.24}$$

The transfer function of such elements is p

$$W(p) = Tp. \tag{2.25}$$

However, no ideal derivative controller exists in practice – all of them contain some inertia. Therefore a differential equation of such controller is

$$T\frac{dx_{out}}{dt} + x_{out} = kT\frac{dx_{in}}{st}, \tag{2.26}$$

and the transfer function therefore

$$W(p) = \frac{kTp}{Tp+1}\ , \tag{2.27}$$

where T is a time constant of the controller, k – amplification factor. This controller can be considered as a series connection of an ideal derivative controller with transfer function (2.29) and aperiodic controller with transfer function (2.24) with equal time constants.

Derivative controllers are used for correction (or regulation) of transfer process. The examples could be stabilising transformers, capacitor circuits, bridge type schemes, etc.

Equation of APC of the derivative controller is

$$K(j\omega) = \frac{jk\omega T}{j\omega T+1}. \tag{2.28}$$

Real and imaginary components are calculated as

$$P(\omega) = \frac{k\omega^2T^2}{\omega^2T^2+1}; Q(\omega) = \frac{k\omega T}{\omega^2T^2+1} \tag{2.29}$$

and the correspondent responses are in figure 2.8

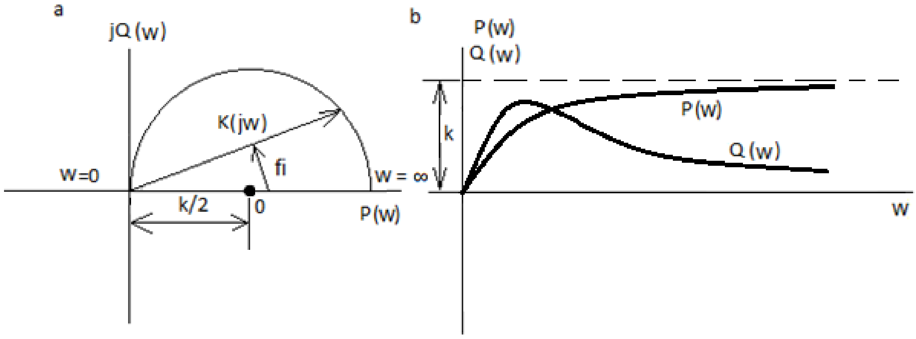


Figure 2.8. Parametric plot of frequency response, real and imaginary variables of derivative controller

The logarithm of the frequency response we can find:

$$L(\omega) = 20lgk + 20lg\omega T - 2 - lg\sqrt{\omega^2T^2 + 1};\ \varphi(\omega) = arctg\frac{1}{\omega T}. \tag{2.30}$$

Therefore the Bode plot of the derivative controller consists of three components: a straight line parallel to abscise, a line with an inclination of 20 dB/dec and crossing a point of an abscise, correspondent to a junction frequency, and the third with two asymptotes with a common point, one of which lies along abscise and the other with a negative declination -20 dB/dec.

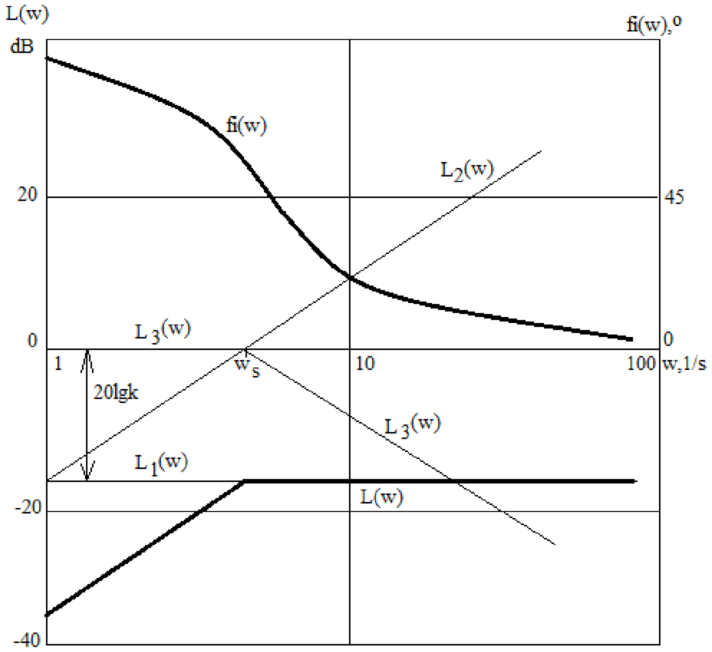


Figure 2.9. Bode plot of derivative controller

4. Integrating controller

Integrating controller is a controller the output value of which is proportional to the time integral of the input value:

$$x_{out} = k \int x_{in} dt, \tag{2.31}$$

where k is a transfer factor of the controller equal to the speed of the output value changing according to the input. The example of the integrating controller could be an R-C circuit or an armature of DC motor (with constant excitation flux) the input value of which is input voltage but the output is the angle of the shaft turning ϕ.

The transfer function of the integrating controller is

$$W(p) = \frac{k}{p}. \tag{2.32}$$

For the integrating controller

$$K(j\omega) = \frac{k}{j\omega} = \frac{k}{\omega} e^{-j\frac{\pi}{2}}, \tag{2.33}$$

that is illustrated with a straight line coinciding with an axis of negative imaginary numbers (fig.2.10). At any frequency the output oscillations lack behind the input for angle ϕ=-π/2. The equations of real and imaginary frequency responses (fig.2.10) are

$$P(\omega) = 0; Q(\omega) = -\frac{k}{\omega}. \tag{2.34}$$

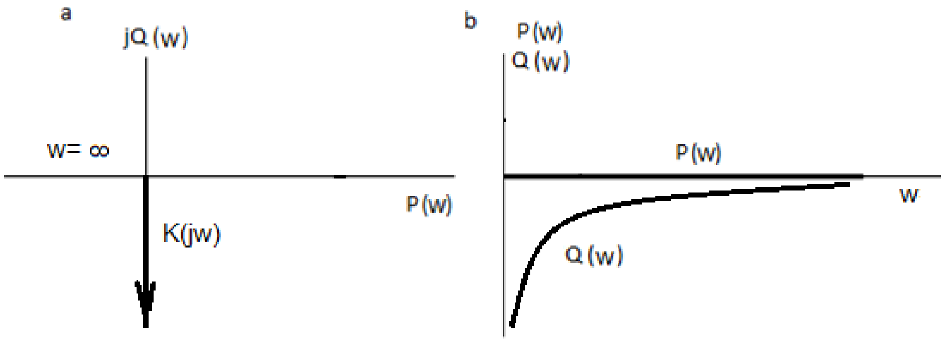


Figure 2.10. Parametric plot of frequency response of integrating controller

LAFC is formed in accordance with

$$L(\omega) = 20lgk - 20lg\omega, \tag{2.35}$$

LFC in its turn:

$$\varphi(\omega) = -\frac{\pi}{2}. \tag{2.36}$$

They are illustrated in fig.2.11.

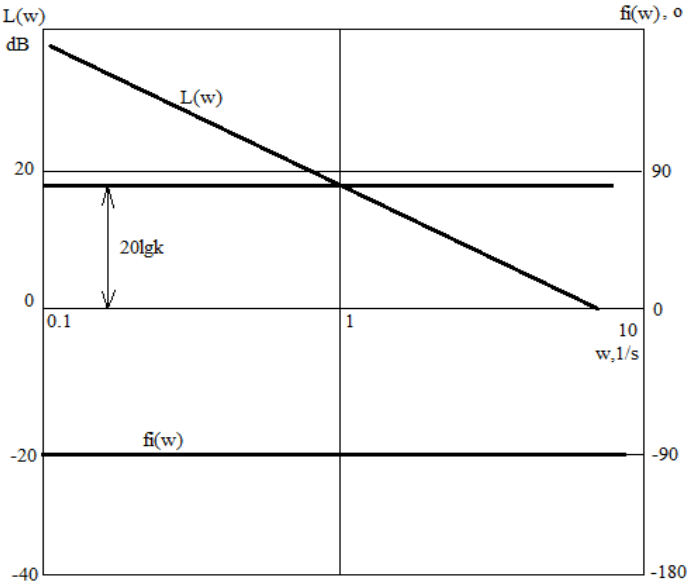


Figure 2.11. Bode plot of integrating controller

5. <u>Oscillating controller (second order aperiodic or inertia controller)</u>

This is a controller the output value of which at the supply of a step impact aims to a steady-state value making damped oscillations. The transient process of this controller is described as

$$T_1 T_2 \frac{d^2 x_{out}}{dt^2} + T_1 \frac{dx_{out}}{dt} + x_{out} = kx_{in} \,, \tag{2.37}$$

where T1 and T2 are the time constant of the system; k – is an amplification factor of the controller.

The transient process is determined with the roots of the equation depending on the level of the time constants and can be aperiodic or oscillating. This type of controller can be illustrated with electric circuits consisting of inductance, capacity and resistance, electro-mechanic element like motor able to accumulate kinetic energy in its armature and electro-magnetic energy in the armature circuit, mechanic elements with mass, elasticity and viscous friction, etc.

Transfer function of this controller is

$$W(p) = \frac{k}{T_1 T_2 p^2 + T_1 p + 1} \quad . \tag{2.38}$$

The equation of APC from the transfer function is

$$K(j\omega) = \frac{k}{-T_1 T_2 \omega^2 + T_1 \omega + 1}. \tag{2.39}$$

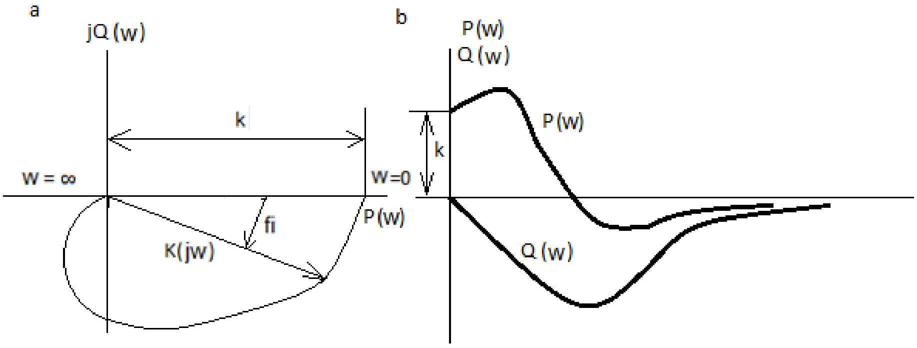


Figure 2.12. Parametric plot of frequency response of oscillating controller

APC is located in two quadrants. As frequency ω is changed from 0 to ∞ K(jω) turns for an angle ϕ=-π (fig.2.12). Real and imaginary components are calculated according to

$$P(\omega) = \frac{k(1 - T_1 T_2 \omega^2)}{(1 - T_1 T_2 \omega^2)^2 + T_1^2 \omega^2}; Q(\omega) = \frac{k T_1 \omega}{(1 - T_1 T_2 \omega^2)^2 + T_1^2 \omega^2}. \tag{2.40}$$

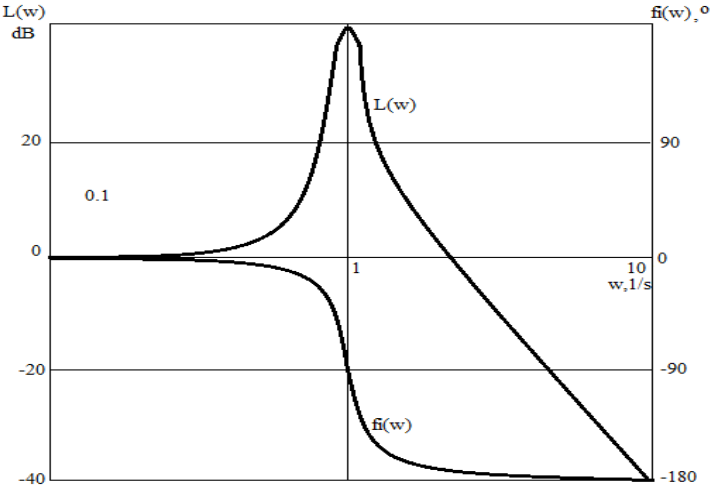


Figure 2.13. Bode plot of oscillating controller

2.4. Evaluation of Regulated System Stability

Stability does not mean static. A stable system is not necessarily steady. Dynamically systems can be stable such as steady systems. Some stable systems oscillate in perpetuity without ever stopping at a fixed position, such as climate and ecological systems. Dynamically stable systems have internal dampening mechanisms that return the system to balance. Dynamically stable systems are self-correcting even if there are disturbances or imperfections. There are forces that counteract independent forces.

Unstable systems can exceed performance levels of stable systems but lose abruptly stability. To hold an unstable system as a stable system they require an external energy source. Dynamically unstable systems cannot self-correct and will fail if disturbed. A steady unstable system has an instability in its own. It is unstable in the beginning and cannot reach a stable situation without an external force.

Stability of a system is its independent ability to reach steady-state condition after supply a disturbance to any of its outputs. Analysis of the system for stable operation and providing of it are the basic tasks of control theory.

In the frequency and phase response functions most part of the elements contain factor e^{-vt}, where -v is a real part of the root. It means that as v>0 the parametric plot of frequency response is damping, it is infinitively close to zero at $t \to 0$. Therefore, with v>0 the elements are statically stable as with any disturbance they return to the steady-state condition as soon as the disturbance is completed.

With v<0 the parametric plot of frequency response is infinitively increasing at $t \to \infty$, as $e^{|v|t}$ is faster increasing than any other function. Therefore, with v<0 the response of an element will infinitively increase and will not return to its steady-state until it is compensated or its properties are changed. The element is unstable (an example could be a bicycle).

A special case takes place with v=0. If it is an integrating element then its output signal will neither infinitively increase with an input disturbance nor get the steady-state. A step function with an amplitude opposite proportional to its time constant is formed at the output of the element. The second special case is with the second order aperiodic element. With v=0 its frequency response is of sine form. Ideally it could be considered as stable, however such element is complex for realisation – a dissipation of the energy of sine-from oscillations, that is close to the situation with v>0. In mechanical systems the energy is dissipated in friction, but in electric and electronic systems – for the losses in active resistances (heat) and electromagnetic waves radiation.

Therefore the basic condition of the stability of a linear system is an absence of roots with a positive real component in frequency response denominator. If all roots have a negative real component, then all the coefficients of this polynomial are with same sign (or their component is zero).

Therefore: the stability of a linear system requires the same signs of all coefficients of frequency response denominator. But this requirement is enough for the first and second order elements only. For other systems and elements this is necessary but not enough. The system is unstable if it contains at least one unstable element in its structure.

While investigating the stability of a system the following tasks could be solved:

1. Analysis of the system stability at given parameters;
2. Opportunity of some parameters changing without disturbing of the system stability;
3. Analysis of the structure of the system and determining of the parameters resulting in its stability.

2.4.1. Rauss-Gurvits stability criterion

Let us assume that the equation of the transfer function denominator is:

$$a_0p^n + a_1p^{n-1} + a_2p^{n-2} + \cdots + a_{n-1}p + a_n = 0. \qquad (2.41)$$

Factor a_0 is assumed to be positive and all the coefficients are real values.

The Rauss method gives an opportunity to determine the occurrence of roots with positive real component analysing all the coefficients from a_0 to a_n. Firstly the coefficients (including zero) are grouped into two lines. Then with crossed multiplication of the first column onto the each further the n-1 lines are formed, etc.

Rauss-Gurvits criteria states that: equation (2.41) does not contain the roots with positive real component in the only case when all the components of the first column are positive. With this condition fulfilment all components in the lines are also positive.

Disadvantages of the criteria application:

1. At the experimental evaluation of the system the determining of all the coefficients of the equation is not always possible with necessary accuracy.
2. The method based on the criteria becomes too complicated while analysing of the high-order systems.
3. The criteria does not state clearly the achievement of the system stability.

2.4.2. Nyquist criterion of stability

Nyquist criterion allows to judge about the stability of a close system on the basis of its frequency response in open condition. A complex transfer factor of the close system can be obtained as:

$$W_C(j\omega) = \frac{W(j\omega)}{[1-\beta(j\omega)W(j\omega)]} \qquad (2.42)$$

Assume that $W_L(j\omega)=\beta(j\omega)W(j\omega)$. It is obvious that parameter $W_C(j\omega)$ is infinitively high, i.e. the system losses its stability, when the product $W_L(j\omega)=1$. The Nyquist plot of the open system $W_L(j\omega)$ and component 1- $W_L(j\omega)$ can be represented by means of hodograph of vector in a complex coordinate system. For the feedback to be negative the signal in the feedback loop should be inverted, i.e. change the sign for an opposite. It corresponds to a phase shift for +/-π. If due to additional shifts the total phase shift is equal to 0 or achieves +/-2π then the feedback becomes positive. Therefore for stability of the close system the modulus of Nyquist plot vector of the open system $k(\omega)=k_\beta(\omega)k_D(\omega)$ should be less than 1 at the frequencies at which the phase shift in the loop is 0 or 2π-fold. The value of k(ω) at some of these frequencies higher than 1 cannot state that the system is unstable, but it is rather unstable. The frequency of resonance oscillations lies in the range where the frequency response is close to 1; and the closer is the system to different critical conditions the slower are the damping resonance oscillations.

All these conditions are stated in the Nyquist criterion:

Close system is stable if its open system is stable and the Nyquist plot does not encircle point (-1; j0). The example of the hodograph is in fig. 2.14 – this is the graph of a stable system.

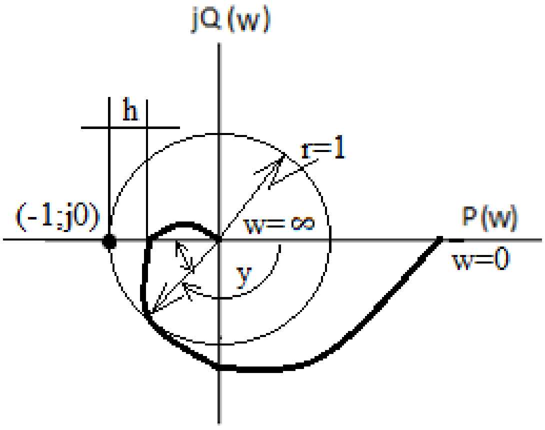


Figure 2.14. To the Nyquist criterion of system’s stability

2.5. Qualitative Parameters of System Stability

The quality of the system's control and dynamic errors is estimated by means of different criteria. The most important of them are two: the time of regulation t_p (time of transient process) and overshoot ΔX parameters (Fig.2.15).

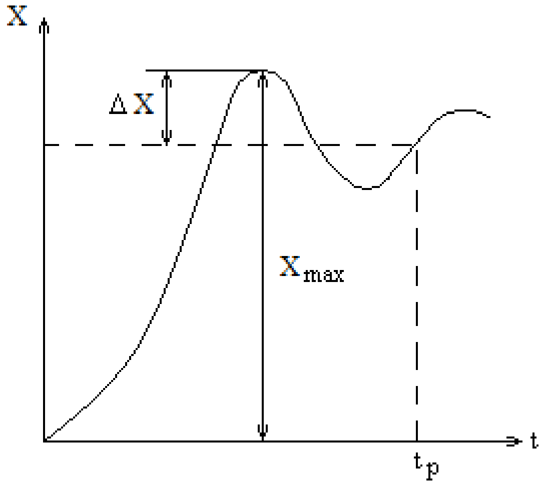


Figure 2.15. Parameters for the system's stability estimation

1. Time of regulation (or time of transient process) is the time during which the response of the system achieves steady-state condition. In fact this time could be determined at the moment when declination of the parameter under consideration is becoming long-term less than 5% of the ideal steady-state.
2. The overshoot is a result of the inertia of any component of the system. Therefore at the initial moment the difference of h(∞)-h(t) will always be maximum.

The maximum deviation of the regulated parameter from its steady-state value h(∞) with the same sign is called overshoot estimated in percent.

If the time of regulation and overshoot are given then the necessary transient process and its quality is also given.

Selection of the control law and controllers tuning controllers

Approach to the automated control systems based on the calculations and computer simulation only is rarely possible and successful as the properties of the plants are almost always known not to its full extent as well as the factors and coefficients of the transfer function of the plant.

Therefore, usually first of all a type of the controller is selected taking into account the information about the plant and general requirements to the regulation system. In the very simplest case it can be a proportional (P) controller with the transfer function:

$$W_R(p) = K_R \tag{2.43}$$

But P-controllers are applied only in the cases of low requirements to the control accuracy. Application of integrating controller (I) in parallel with the proportional with the total transfer function:

$$W_R(p) = K_R + {}^{1}/_{T_I p} \tag{2.44}$$

where T_I is the time constant of the controller. This is so called PI controller type. The application of integrating controller gives an opportunity to exclude static error of the system. PI approach to the controller is the most often used in the industrial systems.

The third component –differential – is introduced into the control in two cases:

1. If the large sudden changes are possible in the system – e.g. sharp changes of the load, etc.
2. For the increasing of stability reserve in respect to phase.

The transfer function of the controller is:

$$W_R(p) = K_R + {}^{1}/_{T_I p} + T_D p \tag{2.45}$$

where T_D is the time constant. This controller is called PID. The inertia of the control is described by means of PID controller. Differentiating component reduces the inertia of the control as the variations of the regulated components are discovered before they take place (or finished). Therefore for the different types of controllers output variable r(t) and the error signal ε(t) at its input are described by the following expressions:

for P-controller $r(t) = K_R \varepsilon(t)$ (2.46),

for PI-controller $r(t) = K_R \left[\varepsilon(t) + \left({}^{1}/_{T_I}\right) \int_0^t \varepsilon(t) dt\right]$ (2.47),

for PID-controller $r(t) = K_R \left[\varepsilon(t) + \left({}^{1}/_{T_I}\right) \int_0^t \varepsilon(t) dt + T_D \varepsilon'(t)\right]$ (2.48).

The process of the selection of the parameters and structure of the controllers first includes the analysis of stability. Then a criterion of the quality of the process is selected, e.g. being described mathematically by means of some equation with the given restrictions of time of control, overshoot, etc., and the parameters are tuned giving the closest correspondence to the selected criterion.

For the controllers tuning two methods could be basically applied. The first method is well suited for most of the processes as soon as the second is good to be applied for slow processes. Both methods take into account the requirements to the stability and give enough accurate results as they count the dynamic characteristics of both plant and controller.

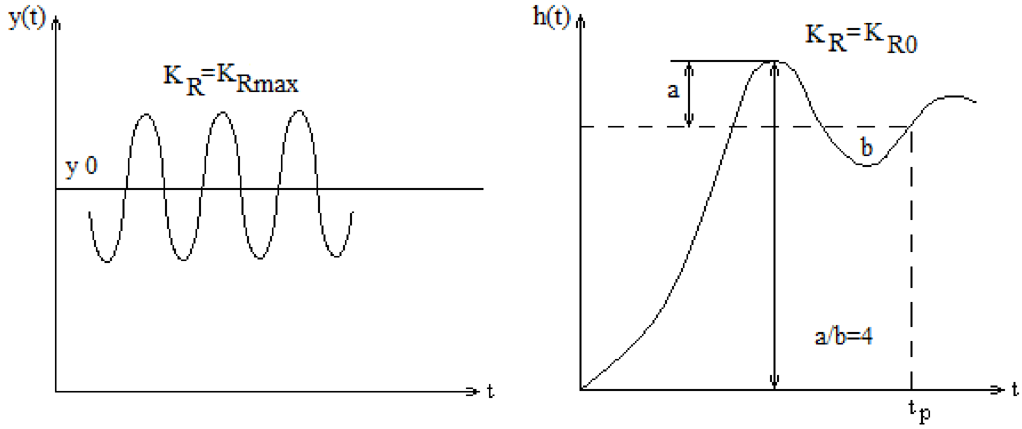


Figure 2.16. Qualitative parameters of transient process

In the method of a closed loop the parameters of the P, I and D elements are fixed. Then the factor of proportional amplification KR is increased until the plant gets the steady oscillations, fig. 2.16,a. This value of KR is the maximum amplification KRmax. If the duration of the maximum cycle of the oscillation is λ then the recommended parameters are the following:

P-controller: $K_R = 0.5K_{Rmax};$ (2.49)

PI-controller: $K_R = 0.45K_{Rmax};$ (2.50)

$T_I = {}^{\lambda}/_{1.2};$ (2.51)

PID-controller: $K_R = 0.6K_{Rmax};$ (2.52)

$T_I = {}^{\lambda}/_{2};$ (2.53)

$T_D = {}^{\lambda}/_{8}.$ (2.54)

The mode of the steady oscillations is not always allowable. Therefore for the specific slow process the method of a given damping is applied. The controllers are tuned also in accordance with the condition of system's stability. But such value KR = KR0 is selected that the amplitude of oscillations is four times damped with each cycle. In this method the system does not reach the boundaries of stability. The recommended parameters of controllers are:

P-controller: $K_R = K_{R0};$ (2.55)

PI-controller: $K_R = 0.9K_{R0};$ (2.56)

$T_I = \lambda\ ;$ (2.57)

PID-controller: $K_R = 1.2K_{R0};$ (2.58)

$T_I = \lambda\ ;$ (2.59)

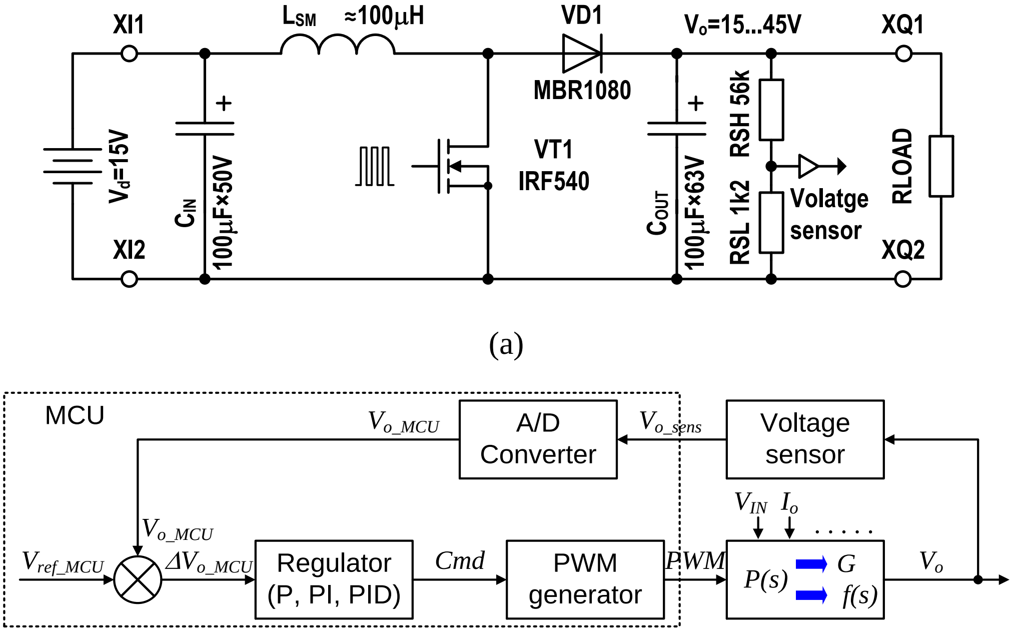
$T_D = {}^{\lambda}/_{4}\ .$ (2.60)

Chapter 3:Microcontrollers

3.1. Embedded systems with microcontrollers

Modern electrical devices are getting more and more new functionality and performance. For example, even such a simple device like electric kettle undergone significant changes during recent years. Nowadays electric kettle is not just a vessel with heating element, but a device capable of keeping constant temperature or change it with the correspondence with required pattern. Obviously, such devices contain not only actuators (executive elements), but also sensors and a control system, included in the device. Such control systems, nowadays mostly microprocessor, therefore regarded as embedded control systems.

Let's discuss an example of the embedded system – a control unit for automatic voltage stabilization. If the voltage is stabilized means of a step-up converter, Figure 3.1(a), the control system must be capable of producing PWM signal to the transistor of the converter, measuring the voltage produced by the converter and making necessary calculations for implementation of regulator, as well as of entering the reference signal form external devices of from a human-machine interface, Figure 3.1(b). Therefore, the corresponding embedded control system must incorporate a calculation unit (processor or CPU), analog-to-digital converter (ADC), generator of PWM signal (timer and compare unit), general purpose digital inputs/outputs, communication module, as well as some memory for storing the program and data. This is a quite typical configuration of the embedded control system: CPU + memory + peripheral devices. When implemented in an IC it is called microcontroller unit (or MCU).



(a)

(b)

Figure 3.1. Boost converter (a) and its representation in regulation system and main parts of its control system (b)

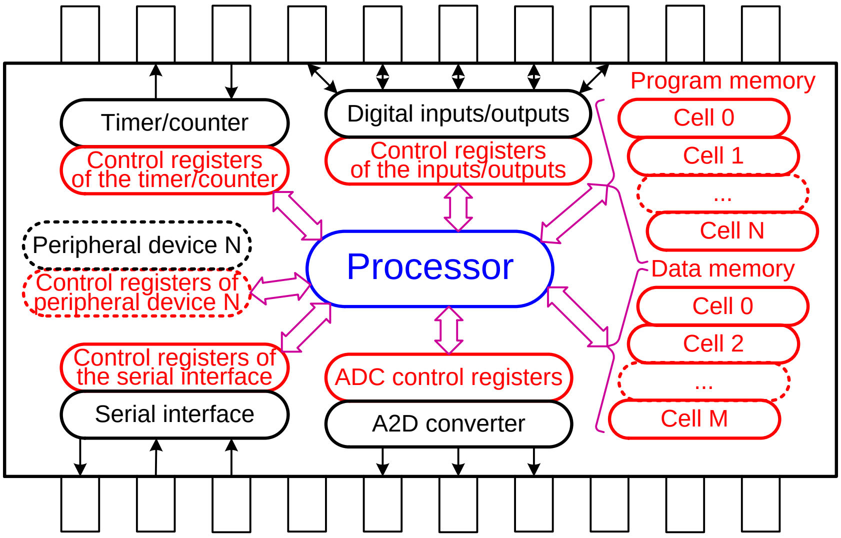


Figure 3.2. Microcontroller a combination of CPU, memory and peripheral devices

All microcontroller peripherals (including the analogue) connected to the processor through an intermediate - control registers. The control registers are special memory cells whose individual bits of the control switches play a role. So, in general, it can be said that the microcontroller on the processor are connected to a number of memory cells - general or special (peripheral control registers) - Figure 3.2. Such a memory cell array can be characterized by the number of bits of data cells and cell numbers (memory).

3.2. Architectures of MCUs, their main parameters, most popular MCUs presented on market

It is clear that each memory cell cannot be added to the processor with only a dedicated group of managers in the total number of connections would be too great. One way to avoid such compounds - is the parallel connection of cells. Parallel connections to line equivalent memory cell data input / output connected together and connected to the processor data input / output. This creates a management team that determines recordable and readable information. These leaders are called data bus. It is obvious that the girder is such signals (hereinafter - D), the data bit is a memory cell and the processor. Saying, for example, "8-bit Microcontrollers" - is meant that the microcontroller data bus is 8 signals the processor may process 8-bit numbers and memory consists of 8-bit memory cells.

Parallel to the event processor at a time can be connected to only one of the memory cells. The other memory cell does not work then - otherwise make connections uncertainty. Primitives can be assumed that only one memory cell is fed feeding, but the rest remain excluded. Such cells or facilities continue to call it "activated" or "active" (in English literature using the word "enabled"). To identify cells that are activated require additional signal group called an address bus. Information addresses girder down the processor thus activating certain memory cell. If the address girder is A signals (and given that each address signal can be 0 or 1), it may 2A code versions and can be addressed 2A memory cell (that is - "may indicate one of 2A cells" or "you can choose one of the memory cells 2A"). In other words we can say that 2A - is the maximum amount of memory units corresponding to the memory cell and the size of the data bus D (the most common 8-bit bytes or 16-bit words). For example, if A = 10 and D = 8, the maximum amount of memory in the system is 210 = 1kB = 1024B (1024 bytes or 1kB). Sometimes evaluating the microprocessor memory total volume is also used for bit units. Can be calculated that if the address is A girder signals and can be addressed 2A cells, and each cell size is D bits, the total memory is M = 2A x D bits. In this example, at A = 10 and D = 8, meet at M = 8192b = 8KB (8192 bytes or 8 kilobits).

To determine what kind of data exchange processes take place, as well as to align storage unit and processor operation requires some more signals. For example, it is necessary recording signal W to initiate the recording triggers the memory cell, the read signal R that determine the reading of data. The signals W and R are often combined in a single operation selection signal. Such individual signals together often referred to as the control bus. More Control Bus may be synchronization signals, the signals that regulate the exchange of data between the processor and devices with a significantly slower speed, etc.

Figure above shows an example of a microprocessor system with six 8-bit memory cells. Thus, the memory and the processor is 8-bit, 8 dataset signals. In turn, addressed to 6 cells need at least 3 address signals, representing the address bus. In addition, the system is a signal that defines the data exchange process.

Each cell is realized here as a parallel register (REG0 REG7 ...) by recording the inputs and activation input. Such a registry entry must be made at, but reading - off and. If both, it is a 1, the register stored just previously recorded data. Register activation signal form descrambler DC, the inputs (A, B and C) are connected to the address bus, the outputs (Q0 ... Q7) - to register activation inputs. Each of the descrambler outputs activated (give 0) at certain combinations of addresses girder. Then the registry is available for reading or recording.

Pay attention to the outputs Q3 and Q4 which are not connected. Therefore, formal addresses 011b (binary number 011) and 100b no real memory cell and does not meet the actual recording or reading with these addresses is not possible. This situation is quite common. For example, below is shown MSP430 microcontroller memory topology with significant areas, which do not meet any physical memory.

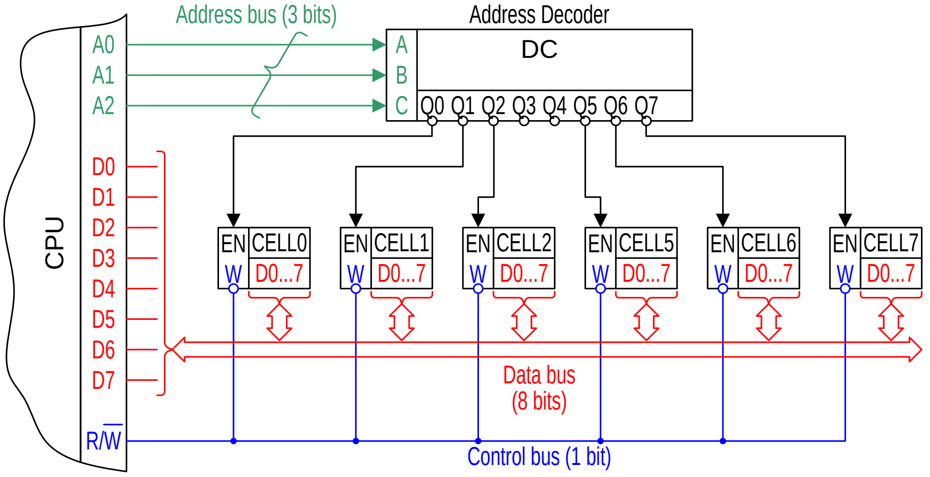


Figure 3.3. Connection of several memory cells to CPU – meaning of busses

As has been noted in microcontroller memory is not homogeneous. This means that the storage part (sometimes says "segments") and even individual cells differ in design technology, and the role of the microprocessor system.

Technologically memory may be permanent, operational, and the hardware attached to the memory cells. Constant-volatile memory does not lose data after power removal, but the recording time is relatively long, so the memory is difficult to use operational requirements. Memory is a faster, but its information after power-off disappears. In turn, the hardware attracts the memory cell content depends on the state of the equipment.

From a functional point of view, can be divided into program memory (the content processor interprets the command codes), data memory (the content of which is processed the processor) and the peripheral control registers. The control registers are the same with hardware-related memory cells. In fact, it is not memory as such, but important hardware parameters of the digital image. It can also be interpreted as a permanent overwriting memory. At these "memory" processor turns when it is necessary to change something in the machine's operation.

Memory is the most important segment of memory. The processor turns to program memory each time it reads a new command. This means that the execution of the program during the program memory is read regularly. At the same time, the need to change the program memory occurs only when you need to change the program itself, it is - configuring microcontrollers. While configuring a very fast memory recording (compared to the data recording / reading microprocessor) is not very up to date. It should be noted that after the power supply feeding the memory already contain a program for the microcontroller can begin to work. Therefore, the program memory is usually carried out volatile performance.

Data memory is read both regular and regularly entered. Read / write frequency depends on the tasks for which the microcontroller and may vary (can also be comparable with the processor speed). Similarly, the need to retain data after power disappearances can be (different rate of conversion tables, etc. In the case), but can also be a (temporary data only). Consequently, the data memory is often both the permanent and the operational part.

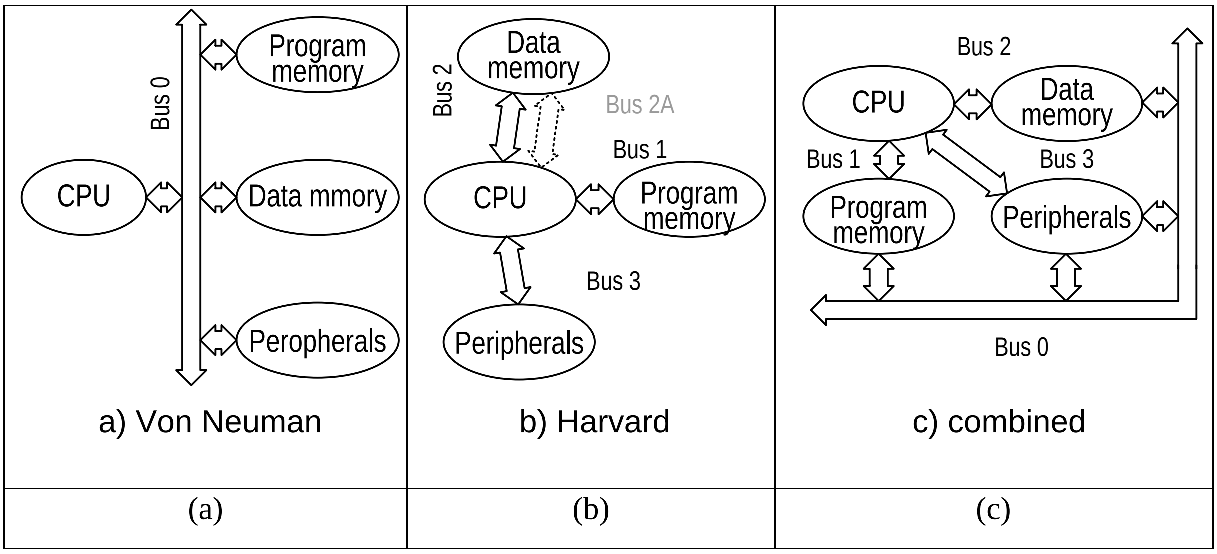
There are various ways to connect to all these memory blocks to the processor. If the program memory, data memory and register control for connection to the processor used the same truss kit,

Figure 3. **4**(a), then such a microcontroller architecture is the name "Von Neumann" architecture. All memory segments in this case are in the same address space, but differ in their addresses. All memory blocks can be modified with the same teams that facilitates the absorption of microcontroller and programming in assembler language. At the same time, the processor is not possible to simultaneously apply to both program memory and data at and for the plant, which makes the team and the performance of the program longer. Such architectural functional differences between the program and data memory gets pretty fuzzy. On the one hand the program code may be recorded in the operational data memory and run from there, but on the other hand the data can be found (since the configuration) Permanent program memory and can be successfully from there taken (only ensure that the processor does not comply with these data words as commands) . As the "Von Neumann" architecture representatives can mention company "Texas Instruments" MSP430xXXX microcontrollers.

Alternative architecture option requires that every important memory block is connected with its busbar assembly,

Figure 3. **4**(b), and there are multiple address spaces. This means that programs the memory cell with address 0, data of the memory cell with address 0 and can also control register of address 0. This microcontroller architecture called "Harvard" architecture. With this architecture processor can turn off all the memory blocks at the same time, which allows to organize the so-called conveyors (when the processor simultaneously reads one command executes each other, take the data for the third and entered fourth command execution results while working with peripherals) thus speeding up the program execution. At the same time, access to the various memory blocks are different, should be applied in a variety of commands, and programmatic access to the program memory is often not at all (it is modified only during configuration). Another noted that the use of "Harvard" architecture microcontroller manufacturers often combine data memory with the control registers and program memory is connected separately. Sometimes, in order to simultaneously provide both data reading and recording both the data memory is connected by two truss kits. As the "Harvard" architectural examples can be mentioned the company "Microchip" PIC1XxXXX microcontrollers.

Finally, it should be noted that there are also mixed architecture, which provides a common bus for all the memories and even special access to certain memory blocks separately, Figure 3. **4**(c). This allows you to achieve the advantages of both architectures in a single device. It has built the company "Atmel" AVR microcontrollers.



(a)	(b)	(c)

Figure 3. 4. Generalized functional diagrams of computer architectures:
(a) von Neumann, (b) Harvard, (c) combined

3.3. Rules of successful MCU handling

Practical implementation of an embedded control system in an MCU requires making several practical steps. Let’s discuss them taking an MSP430 MCU as an example.

(1) Microcontroller must be connected to a **power supply**. Although this point seems trivial, it must not be overlooked. Practical experience has shown that most of the hardware problems such or otherwise associated with a missing or inappropriate feeding. It includes a number of features.

(1-a) **Voltage value.** Modern microcontrollers usually only one supply voltage, which can be charged with a fairly wide range. However, a variety of microcontroller blocks supply voltage size may vary. For example, "flash" memory voltage at which the memory can be recorded, are often lower. MSP430 microcontrollers supply voltage (by the technical passport data) is from 1.8V to 4.1V, but "flash" memory recording voltage is 2.7V and higher.

(1-b) **Power connection.** In the event that the system is only a digital chip with a single supply voltage, this voltage connection of each chip (including microcontroller) provides for two contacts. If there are multiple power supply voltages, the land is generally common, but each voltage provides specific contact.

If the microprocessor system is not only digital, but also analog equipment, a common power supply, Figure 3. **5**(a) is not desirable. Digital equipment pulsed current causes serious disruption to the power supply voltage, which has a negative impact analog equipment performance - reduces the accuracy of measurement, leads to self-oscillation and so on. Therefore, in mixed systems used to distribute power Figure 3. **5**(b). Compression plate then there are certain paths analog equipment feeding / ground and digital equipment feeding / ground. These paths are merged only in the vicinity of the power supply to pulse current digital and analog current of mutual influence would be less. Best quality solution is to create a special analogue equipment power supply, and analog and digital power ground connect.

MSP430x1xx microcontroller digital supply voltage contact file is called VCC (DVCC or if there is an analogue power supply) and ground contact - CSF (DVSS). Analog power supply connected to the AVCC and AVSS. MSP430F123 (2) microcontroller with SO28 Power supply contacts are: 2 - 4 and VCC - VSS. Between supply voltage and ground contacts, as close to these contacts, connected to a small capacitor with low internal resistance (typically ceramic with a value of 10nF ... 1mkF Figure 3.6(a).

Power Consumption. Modern microcontroller self-consumption is very small - from a few mkA to mA and can provide any power supply. But even such a power supply current to be carefully assessed if Microcontrollers feeding on an autonomous energy source: batteries, battery or supercapacitor.

In calculating the total consumption of microcontroller systems to be reckoned with microcontroller general contact currents flowing in the same microcontroller and affects the power losses and heat mode. First is subject to the following conditions: (i) If the microcontroller acts as a contact output, output logic 1 (which corresponds to the output voltage close to Vcc) and is connected to a load, the resulting "comes out" the power circuit Figure 3.6(b). This circuit closes through the power supply, microcontroller power contact, the upper transistor digital outputs, output contacts and load. Closed-circuit current is mainly determined by the load impedance; (ii) If the microcontroller contact acts as an output, the output logic 0 (which corresponds to the output voltage close to 0V) and is connected to the load, then it formed "flowing into" power circuit, which closes through the power supply load, the output contacts, digital output bottom transistor and microcontroller earth contact Figure 3.6(c).

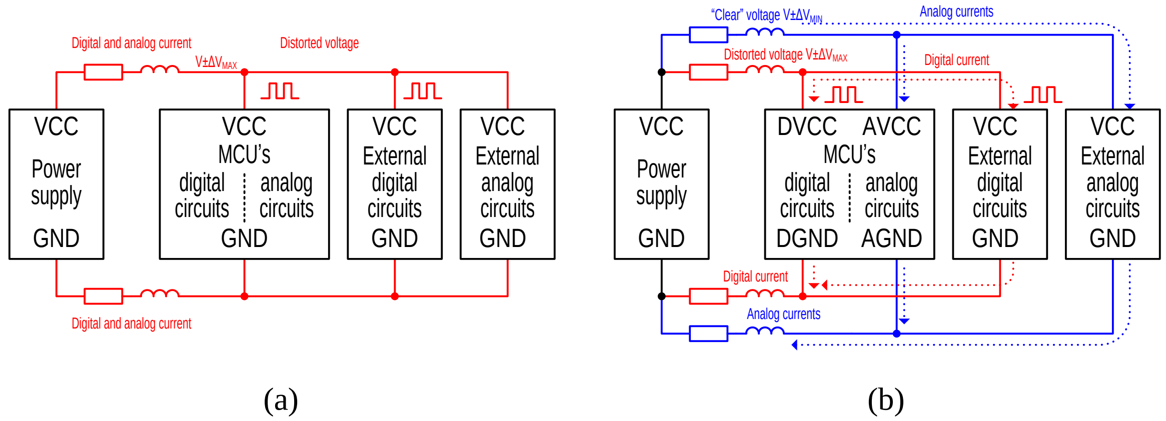


Figure 3. 5. Mixed (a) and separate (b) power supply for digital and analog circuits

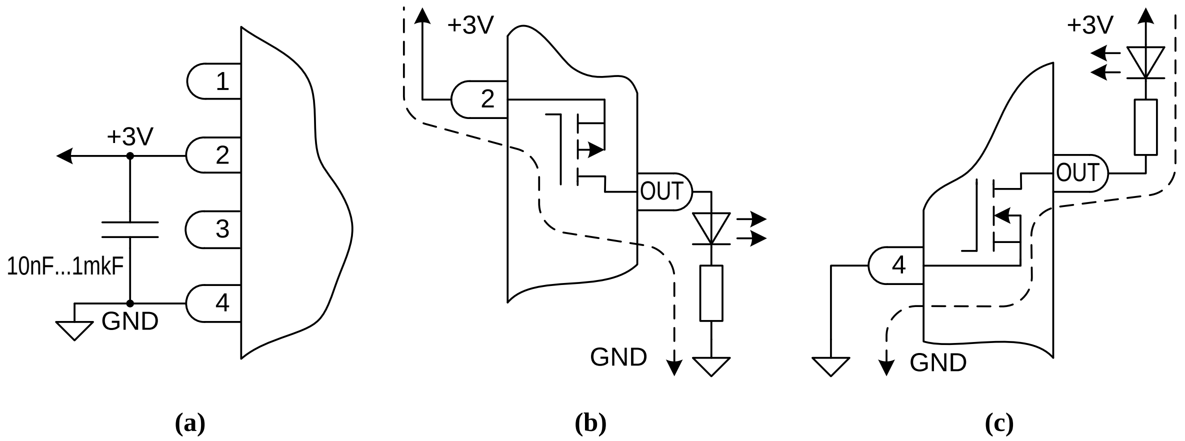


Figure 3.6. Power consumption in MSP430F1232 MCU's: (a) through power terminal; (b) outgoing currents; (c) incoming currents

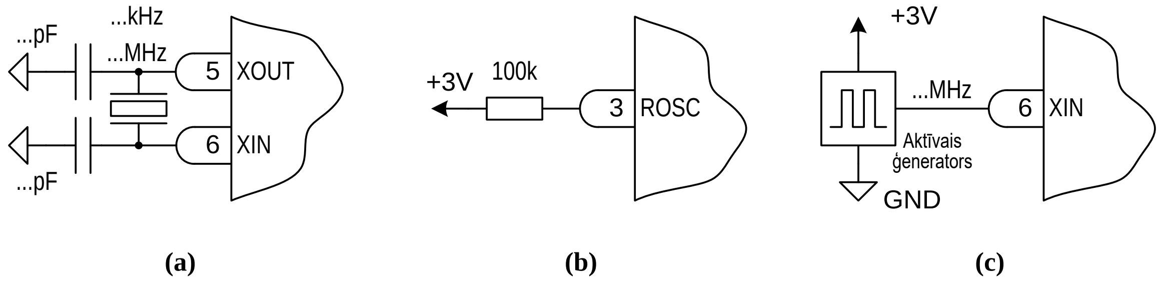


Figure 3.7. External elements of the basic clock system of MSP430F1232: (a) crystal; (b) resistor; (c) active clock source

(2) Operations in MCU are synchronous, therefore, **clock signals** must be provided.

This means that the microcontroller's internal circuit switching and internal operations are only changing a specific signal, called the synchronization signal. Modern microcontrollers normally has an internal clock synchronization signal generator, in addition to which is connected only some frequency referencing elements.

The most accurate way of posing frequencies are those of quartz or ceramic resonator assistance (Figure 3.7-a). Resonator connected to the microcontroller special contacts in parallel by connecting them to the supply of land with small (a few pF) capacitors.

Clock frequency may also be charged with the help of the RC chain. Usually outside the microcontroller is only one element, and it is more resistor, such as MSP microcontroller (Figure 3.7-b).

Finally microcontroller may be hopped with a signal from an external generator (Figure 3.7-c).

MSP430 microcontrollers have advanced clock system. First, it has a number of clock signals, which can vary greatly in terms of frequency, which can be quite free to choose a separate microcontroller part from hunting. Second, the clock frequency can be used for setting all the abovementioned elements (also simultaneously). Third, the MSP430 clock system is erasable - that is, its configuration can be quickly adjusted for the current needs. Additionally, this microcontroller has a generator with internal frequency elements and programmable frequency setup. Immediately after the power-up feeding MSP430F123(2) are synchronized directly from the generator to the default frequency of 750kHz, which can then be reprogrammed.

(3) Correct **reset** circuitry.

Sometimes microcontroller needs to run again - that is, to put the microcontroller to execute your program from the beginning. This process is called, MCU’s return or reset. So that would be no need to disable and re-connect the power, microcontrollers are usually a special return entrance. As long as this input is active (usually a logical 0 - that is connecting the return contact at digital terrestrial) Microcontrollers remain in the returned position and the program does not comply. Once the active signal is lost, microcontroller starts to execute the program again.

Reset pin should not be left unconnected because then contact’s electrostatic charge can be interpreted as 1 or 0 (for CMOS ICs) and can reboot microcontroller accidentally. If the return value of the active signal is "0", then return input can be simply connected to the supply voltage (Figure 8-a and b). Then the microcontroller will start only once - immediately after the power feeding.

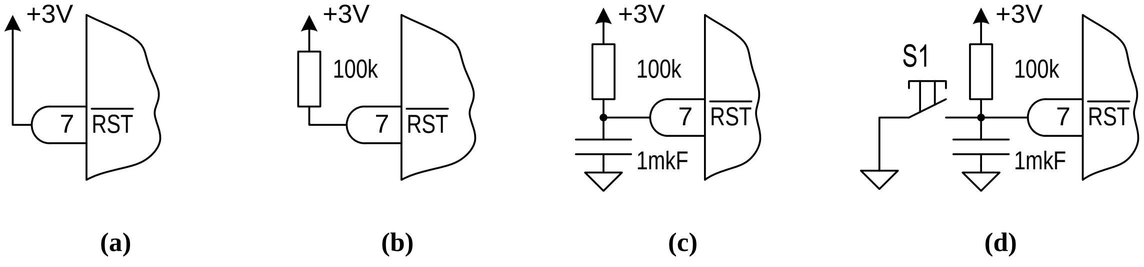


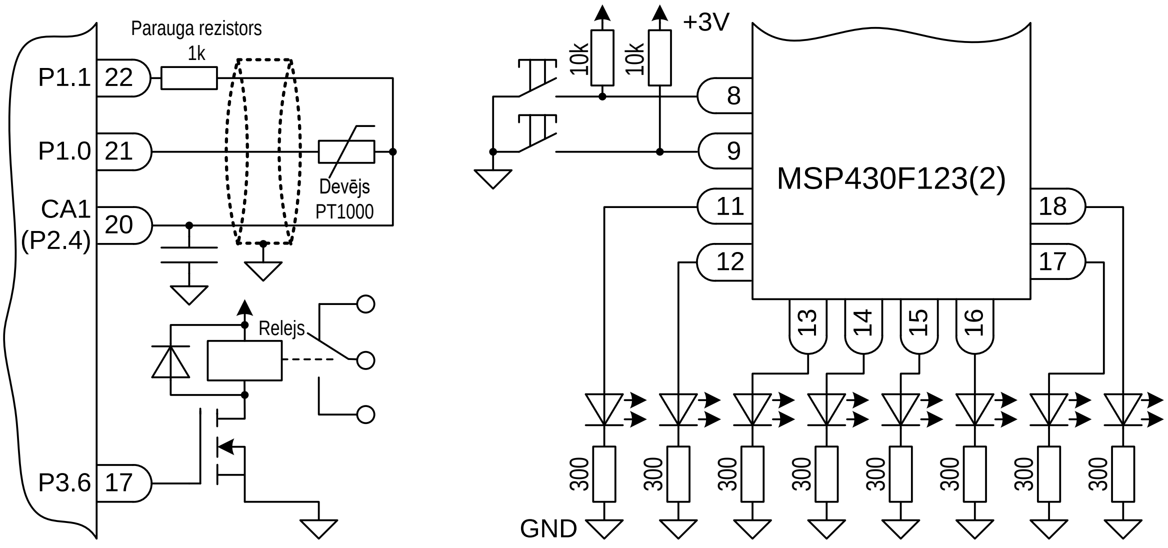
Figure 3.8. Typical configurations of return (rest) circuitry of MCP430

Some microcontroller processor begins to operate at lower supply voltage than the other blocks. If the supply voltage is increasing slowly, it can lead to mistakes. Microcontroller to prevent the return of power after feeding time return entrance connected to the RC chains (Figure 3.8-c). The processor starts only after the power supply reaches all the blocks on the supply voltage level.

Finally, if there is a need to reboot manually microcontrollers, then return entrance connected to the ground via a push-button (Figure 3.8-d), which pushes rebooting case.

(4) MCUs inputs and outputs

While the microcontroller is complete microprocessor unit, its performance can be observed only when it is connected to any external devices. This external device number and content is determined by the nature of the task. For example, if the microcontroller is realized in temperature control, it is connected to the temperature sensor and an actuator, such as a relay (Figure 3.9-a). Microcontroller learning time is often used primitive configuration with mirdzdiodēm and Popper (Figure 3.9-b).



a) Temperature regulator;

b) Training configuration;

Figure 3.9. Examples of GPIOs connections

(5) Last but not least - according to the task **program** has to be composed and located in MCU.

Microcontrollers is a unit with a microprocessor. In turn, the microprocessor is based on a certain agenda. Therefore, the correct and proper exercise program compilation is microcontroller equipment design process, an integral part. Moreover, microcontroller programming and debugging is more labor-intensive part of the design process. Thus, it is mostly turn to microcontroller programming.

The main problem is that the processor does not understand anything other than binary numbers (in their physical representation). In addition, the individual bit values it detects on the basis of the voltage (or current) level. Until recently, the code memo recorded with special equipment - programmers asking the address codes and data in binary form by mechanical switchgear and pressing the button "write". It was a very laborious process. An example of machine code for MSP430F123(2) connected to LEDs and pushbuttons as in Figure 9-b that provides LED blinking at approximately 2 Hz is given in Example 1.

Code example 1. Machine code of MSP430F123(2)

```
0x40B2, 0x5A80, 0x0120      ; Stop watchdog
0x43D2, 0x001A              ; P3.0 - output
0x403F, 0xC350              ; Setup counter
0x831F                      ; Decrement
0x23FE                      ; Repeat untill 0
0xE3F2, 0x0019              ; Toggle P3.0
0x3FF9                      ; Repeat Always
```

Even presented in hexadecimal system this machine code is not easy to understand. This is why MCUs programs are not usually written directly in the machine code, but rather in assembly code, which instructions are abbreviations of the corresponding MCU operations and have direct correspondence to the machine codes.

Code example 2. Assembly code (program) of MSP430F123(2).

```
main         mov    #WDTPW+WDTHOLD, &WDTCTL      ;Stop watchdog
             mov.b  #1,        &P3DIR           ;P3.0 - output
Rpt_forever    mov    #50000,      R15           ;Setup counter
Rpt_250ms      dec    R15                        ;Decrement
            jnz    Rpt_250ms                    ;Repeat until zero
            inv.b  &P3OUT                       ;Toggle P3.0
            jmp    Rpt_forever                  ;Repeat Always
```

Assembly code is simpler, but not easy to use when complex algorithms and sophisticated data processing is necessary. In such cases writing of the initial program in some algorithmic language or object oriented language etc. saves program development time. The above-discussed LED-blinking task solved in C is presented in Example 3. It must be noted also that this program for other microcontrollers will look very similar.

Code example 3. C code for the discussed example.

```
void main(void)
{
 WDTCTL = WDTPW + WDTHOLD;          // Stop watchdog timer
 P3DIR |= 0x01;                     // Set P3.0 to output direction
 for (;;)
 { unsigned int i;
   i = 50000;do i--;while (i != 0);    // Software Delay
   P3OUT ^= 0xFF; }                  // Toggle P3.0 using exclusive-OR
}
```

In the same time between the programs in higher level programming language and output codes is no direct correlation, as determined by the code compiler and compiler optimization specifics. Consequently, the final code operation cannot always be predicted. For example, the assembler program shown in "direct translation" of the C language (Example 3) also causes LEDs flicker plate (Figure 1 9-b), but the flicker frequency will be lower - about 1Hz. Program code is also not the same as the compiled from assembler (Examples 1 and 2).

3.4. Operation of CPU

The main element of any microcontroller is its central processing unit (CPU) because it does the most of the data processing tasks. Its performance depends on the number and complexity of the operations that can be executed by CPU, as well as on their execution time. This, in turn, depends on the instruction set, available data access (addressing) modes and internal “quick” memory of the CPU. Let us to discuss this details looking at MSP430.

MSP430 CPU’s registers

CPU of MSP430 microcontrollers contains 16 bit registers: 12 are general purpose memory cells, and 4 - are special purpose registers. However, from the point of view of data flow these 16 registers are equal – they may serve as data source or data destination. In assembler program registers are identified by their name (or even a few names). The most important special registers are program counter (PC or R0) and CPU status register (SR or R2).

Program counter is a pointer (its content is interpreted as an address) to a memory cell from where a new instruction word of data word will be taken. During the execution of an instruction it is incremented by 2 or 4 or 6 so, that at the end of the execution of this instruction it contains the address of the next one. Upon reset, PC takes address form the main vector cell (0xFFFE).

Status register is not interpreted as a number or address, but as a set of bits with particular meaning. There are two groups of bits. Bits SCG1, SCG0, OSCOFF, CPUOFF and GIE are control bits, but V, N, Z and C – flags indicating some result.

15 9	8	7	6	5	4	3	2	1	0
Rezerved	V	SCG1	SCG0	OSCOFF	CPUOFF	GIE	N	Z	C

rw-0

Figure 3.10. CPU Status Register of MSP430

Processor status register bits are as follows:

bit 0 - is called the flag C. flag. This flag will install when an arithmetic operation results in a carry-over from the oldest significant bit (8 or 16). If the gear does not, then this bit shall be deleted. Similarly, C bits changing the subtraction operations, but to the fact that the withdrawal takes place as counting with the addition of code, the bit significance inverted (that is 1 if the drawing is not). Gear C flag is changed to shear operations, so that all the shear occurs through this flag;

bit 1 - zero flag Z, which will install if arithmetic or logic operation result is zero. It can also occur in counting the event of a gear. For example, if count the bytes 10010 = 15610 = 011,001,002 and 100,111,002 then formed result = 25610 1000000002nd The result is a 9-bit and installing gear flag C (which is actually the result of 9 bits), but the formal result of all 8 bits are 0 why also installed zero flag Z.

bit 2 - is a sign flag N and set, if an arithmetic operation result is a negative number (in addition to the code - see. terminological dictionary). This flag is a formal installation. The flag shall also be installed, if the figure is processed without the use of additional code, but the result is the oldest bit first

8 bits - is overflow flag V. This sets if one arithmetic operation result cannot be represented by the same format. This flag installation complex formula and the definition can be explained as follows:

1) when addition of two positive numbers formally produces negative result:

$100_{10} + 50_{10} = 01100100_2 + 00110010_2 = 10010110_2$ this binary number is 150_{10} of no 2's complement is used or -106_{10} if it is assumed that 2's complement encoding is used for negative number.

2) when addition of two negative numbers formally provide positive result :

$(-80_{10}) + (-60_{10}) = 10110000_2 + 11000100_2 =$(C = 1) 011101002=$116_{10}$

Similarly, it will rise when: 3) subtracting a negative number from a positive number produces negative result; 4) if the subtracting a positive number from a negative number produces positive result.

bit 3 - is the global interrupt enable bit GIE. If it is set, the CPU is allowed to execute interrupt subroutines.

4 bits – is the processor disabling bit CPUOff (no flag). If the control bit is set, then the processor is stopped and the program is no longer executed. Microcontroller mode when processor is still called "sleep" mode. "Sleep" mode ends after interrupt calls or after reset.

bits 5 ... 7 - is the most important bits of the clock system OSCOFF, SCG0 and SCG1;

The remaining bits (9 ... 15) are reserved for special purposes of following MSP430 generations. MSP430x1xx microcontroller can use them for general purposes.

MSP430 command system

There are two types of MSP430 instructions. The first group includes the instructions with unique machine code - let's call them the basic commands. The second group includes basic commands with special operands – one can call it emulated. MSP430 command system. Has 27 basic commands and 24 - emulated (together the 51). The instructions can be either 8-bit (they handled certain bytes of memory, or processor register latest byte) or 16-bit (the memory is processed byte over - the words, but the processor registers are processed in full). For the attention of 16-bit data word applied to the latest byte address (pairs) and separately (with the suffix [.W]) indicate that the word is being addressed. At the same time, processor can address individual bytes separately. The address can be any and it is applied suffix [.b].

Program commands actually one, two or three 16-bit words in memory codes programs. The first word is the word operations encoding the same command. Depending on how we formed the operation name and looks like the same basic command it is assigned to another three groups (Figure 3.11) teams with two operands (Format I) with a operand (Format II) and the short transition commands (Format III).

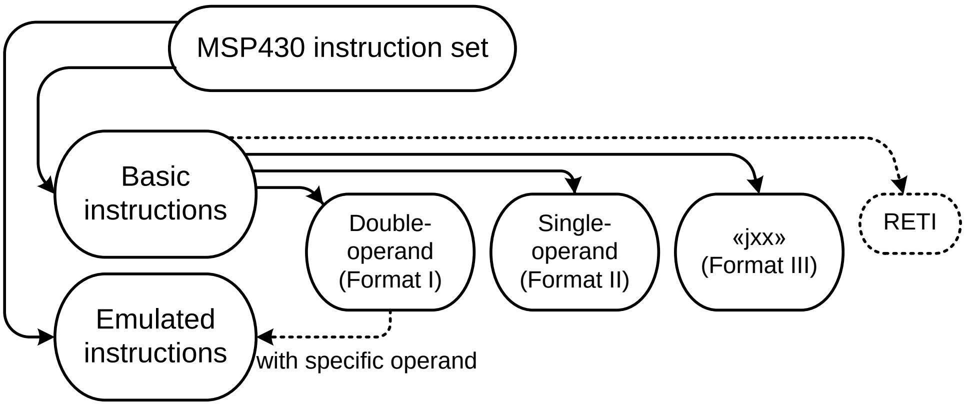


Figure 3.11.

Double-operand instructions.

The usage of the double-operand instructions is following:

cmnd[.b] src, dst

where:

cmnd mnemonics

[.b] extension: if it is omitted or **«.w»**, the instruction processes 16-bit numbers from the range 0…65535 (or -32767…32767); it can use even addresses, but automatic increments (if required) are always by 2; 16-bit instructions deal with entire registers of CPU or with couples of bytes from memory;

If the extension is **«.b»**, the instruction is 8-bitu: processes numbers from the range 0…255 (or -127…127), addresses can be odd, but automatic increments are by 1; 8-bit instructions deals with separate bytes from memory or with least significant byte from a CPU register (the most significant byte then become 0);

src 1st operand – data source; All addressing modes are applicable;

dst 2nd operand – data source (for instruction **«mov»** – formal) and destination; Only several addressing modes are applicable.

The double operand instructions produce 1, 2 or 3 word (2, 4 or 6 byte) machine code, where the first word is instruction word, but the others – is data word, or address od address offset. The instruction word is synthesized as

15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
C3	**C2**	**C1**	**C0**	**S3**	**S2**	**S1**	**S0**	**Ad**	**B/W**	**As1**	**As0**	**D3**	**D2**	**D1**	**D0**

where:

C3…C0 – operation code;

S3…S0 – source (src) register code: 0 – PC, 1 – SP, 2 – SR and CG1, 3 – CG2, 4...15 – general purpose registers R4…R15;

D3…D0 – destination (dst) register code: 0 – PC, 1 – SP, 2 – SR and CG1, 3 – CG2, 4...15 – general purpose registers R4…R15;

As1, As0 – source addressing code (00 – register, 01 – indexed, 10 – indirect, 11 – indirect auto-increment);

Ad – destination addressing code (0 – register, 1 – indexed);

B/W – 8-bit (1) or 16-bit (0) operation.

The double-operand instructions are listed in Table I.

Table I. Double operand instructions and their features

Instruction	C3...C0	Operation	Flags			
			V	N	Z	C
MOV src,dst	0100	dst=src	-	-	-	-
ADD src,dst	0101	dst=src+dst	+	+	+	+
ADDC src,dst	0110	dst=src+dst+C	+	+	+	+
SUB src,dst	1000	dst=dst-src	+	+	+	+
SUBC src,dst	0111	dst=dst-src-C	+	+	+	+
CMP src,dst	1001	dst-src	+	+	+	+
DADD src,dst	1010	dst=BCD(src+dst+C)	+	+	+	+
BIT src,dst	1011	src AND dst	0	+	+	+
BIC src,dst	1100	dst=NOT(src) AND dst	-	-	-	-
BIS src,dst	1101	dst=src OR dst	-	-	-	-
XOR src,dst	1110	dst=src XOR dst	+	+	+	+
AND src,dst	1111	dst=src AND dst	0	+	+	+

Single-operand instructions.

The real core instructions with one operand is quite a bit. They should not be confused with the emulated commands – they produce different machine code and their usage is different. In addition, single operand instructions have only “src” operand, but the emulated commands with one operand – can only be "dst". One operand command usage is as follows:

cmnd[.b] src

where:

cmnd	mnemonics;
[.b]	extension;
src	operand that can utilize all addressing modes.

The single operand instructions produce 1 or 2 word (2 or 4 byte) machine code, where the first word is instruction word, but the second – is data word, or address od address offset. The instruction word is synthesized as

15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
C8	C7	C6	C5	C4	C3	C2	C1	C0	B/W	As1	As0	S3	S2	S1	S0

where:

C8…C0 – operation code;

S3…S0 – source and destination (src/dst) register code: 0 – PC, 1 – SP, 2 – SR and CG1, 3 – CG2, 4...15 – general purpose registers R4…R15;

As1, As0 – source/destination addressing code (00 – register, 01 – indexed, 10 – indirect, 11 – indirect auto-increment);

B/W – 8-bit (1) or 16-bit (0) operation.

Single operand instructions are given in Table II.

Table II. Single operand instructions and their features

Instruction	C8...C0	Operation	Falgs			
			V	N	Z	C
RRC src	000100000	C->MSB MSB->...->LSB;LSB->C	+	+	+	+
SWPB src	000100001	8MSB<->8LSB	–	–	–	–
RRA src	000100010	MSB->MSB MSB->...->LSB;MSB->C	0	+	+	+
SXT src	000100011	8MSB=7th bit	0	+	+	+
PUSH src	000100100	SP=SP-2 MEM(SP)=src	–	–	–	–
CALL dst	000100101	SP=SP-2 MEM(SP)=PC+2 PC=dst	–	–	–	–

Instructions of short jumps.

The third instruction format regards short jumps. These instructions do not have any data operand. In assembly code they are followed by a label or address offset or address, but this argument is fully in the instruction word. So the usage of short jumps is following:

jxx Label

where:

jxx „xx" is condition at which the jump occur, for example, „**jnz**";

Label label or address, which location is closer than 512 words forward or backward away form the location of the jump instruction.

Short jumps forms the following machine code:

15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
0	0	1	Cn2	Cn1	Cn0	A9	A8	A7	A6	A5	A4	A3	A2	A1	A0

where:

001 – identifier of all jump instructions;

Cn2…Cn0 – code of condition;

A9...A0 – 10-bit signed jump offset –511…+512 with respect to the upcoming PC value.

Short jump instructions are given in Table III.

Table III. Short jump instructions and their features

Instruction	Cn2...Cn0	Operation
JNE/JNZ lb	000	Jump, if Z=0
JEQ/JZ lb	001	Jump, if Z=1
JNC lb	010	Jump, if C=0
JC lb	011	Jump, if C=1
JN lb	100	Jump, if N=0
JGE lb	101	Jump, if grater or equal (after CMP) – i.e. jump if V=N
JL lb	110	Jump, if less (after CMP) – i.e. jump if V≠N
JMP lb	111	Jump unconditionally

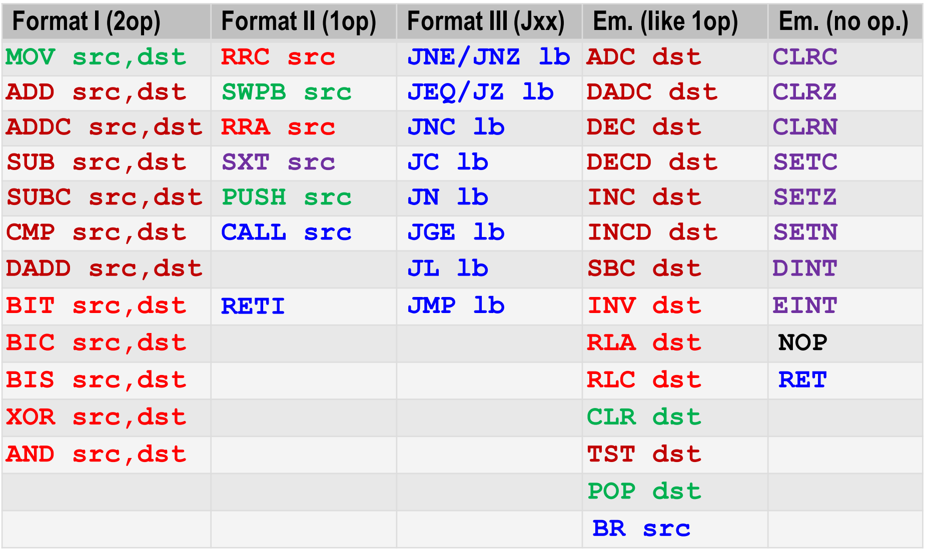
Emulated instructions.

In fact are double-operand instructions with a special 1st or/and 2nd operand. For example, instruction «inc dst» in fact is «add #1,dst». In s/w development environment emulated instructions are converted into core instructions and compiled automatically as double operand instructions. So the emulated instructions can be regarded as an alternative symbolic representation of double operand instructions. Emulated instructions are listed in Table IV.

Table IV. Emulated instructions and their features

Instruction	= core-instruction	Operation	Flags affected			
			V	N	Z	C
ADC dst	ADDC #0,dst	dst=dst+C	+	+	+	+
DADC dst	DADD #0,dst	dst=src+dst	+	+	+	+
DEC dst	SUB #1,dst	dst=src+dst+C	+	+	+	+
DECD dst	SUB #2,dst	dst=dst-src	+	+	+	+
INC dst	ADD #1,dst	dst=dst-src-C	+	+	+	+
INCD dst	ADD #2,dst	dst-src	+	+	+	+
SBC dst	SUBC #0,dst	dst=dst-C	+	+	+	+
INV dst	XOR #(-1),dst	dst=0FFFFh XOR dst	+	+	+	+
RLA dst	ADD dst,dst	dst=dst+dst	+	+	+	+
RLC dst	ADDC dst,dst	dst=dst+dst+C	+	+	+	+
CLR dst	MOV #0,dst	dst=0	-	-	-	-
NOP	MOV #0,#0	r3=0	-	-	-	-

Table V. Summary on MSP430 instruction set (green – data copy, brown – arithmetic, red – logic, blue – branch/jump, as well as violet/black – special instructions.)



Format I (2op)	Format II (1op)	Format III (Jxx)	Em. (like 1op)	Em. (no op.)
MOV src,dst	RRC src	JNE/JNZ lb	ADC dst	CLRC
ADD src,dst	SWPB src	JEQ/JZ lb	DADC dst	CLRZ
ADDC src,dst	RRA src	JNC lb	DEC dst	CLRN
SUB src,dst	SXT src	JC lb	DECD dst	SETC
SUBC src,dst	PUSH src	JN lb	INC dst	SETZ
CMP src,dst	CALL src	JGE lb	INCD dst	SETN
DADD src,dst		JL lb	SBC dst	DINT
BIT src,dst	RETI	JMP lb	INV dst	EINT
BIC src,dst			RLA dst	NOP
BIS src,dst			RLC dst	RET
XOR src,dst			CLR dst	
AND src,dst			TST dst	
			POP dst	
			BR src	

Data operands

The most of the instructions (except short jumps) have data operands. Therefore, it assumes that some data have to be delivered to CPU for processing. The mechanism that defines how the data are delivered is called addressing mode. When the addressing modes are discussed it is necessary to emphasize (1) how they are used in assembly instruction and which are the limitations of this use, (2) where data are located, (3) how the addressing affect the machine code including instruction word and extra words, (4) as well as how addressing affect the execution time of the instructions (Table VI). A – a

Table VI. Addressing modes and their features

Addressing	**Assembly usage**	**src**	**dst**	**Ad**	**As**	**Extra word**	**Extra clock SRC**	**Extra clock DST**
Immediate (direct)	#Data	+	–	11	–	Data	1 (to read data)	–
Register	Rx	+	+	00	0	–	–	–
Symbolic	Address	+	+	01	1	Address offset	2 (to read offset and data)	3 (to read offset and read/write data)
Absolute	&Address	+	+	01	1	Address	2 (to read address and data)	3 (to read address and read/write data)
Indirect	@Rx	+	–	10	–	–	1 (to read data)	–
Indirect auto-increment	@Rx+	+	–	11	–	–	1 (to read data)	–

Indexed	Offset(Rx)	+	+	01	1	Address offset	2 (to read offset and data)	2 (to read offset and read/write data)

Arithmetical conditions in MSP430 microcontroller

Program branching a key element of the condition. Microcontrollers, including the MSP430, the conditions are related to the condition of transition commands. It can be distinguished arithmetical conditions (transition after arithmetic operations) and logic conditions (transition bit after analysis).

Any arithmetic operation affects the processor status register SR flags (zero «Z», gear "C", marks the «N» and overflow «V»). Flags "Z", "C" directly analyzed by commands "JZ", "jnz", "JC" and "JNC". Direct way - with a team of "jn" - may also be analyzed flag «N». In turn, the flags "N" and "V" together allow a relatively simple comparison of numbers format with a sign (in addition to the code).

Comparison, form a special group of arithmetic rules, which are typical of commands combinations. In such command blocks first team is "cmp", which is essentially custodial commands (although without result conservation) and affects all flags. Following the command "cmp" it is located in one or more short transition commands in making the transition, conditional execution case (Table VII and Table VIII). If the condition is not met, then it runs the command, which is located immediately behind the transition commands. It should be noted that the transition condition acts on the "cmp" command against the second operand first. For example, the condition "R15≥100" exercises with the team «jge» with pre-command "cmp", the second operand is R15, but first - number 100 (Table VII-c).

The conditions a) ... c) and e) is implemented with a short transition command. Condition "≤" (d) cannot be implemented with a single command. Therefore, this is carried out by two commands - «jeq" and "jl", which works on "or" principle. Also, the condition ">" (f) cannot be implemented with a single command. However, its implementation is more complex - is located at the beginning of the transition team "jeq $ + 4" that bypass the other team "jge» uniformity case. Therefore «jge» can only be triggered under the condition ">". Assembler notation "$ + 4" means the address which is offset from the current 4 bytes forward (with a minus - can also back). In this case it is the team's office, which is located immediately behind the team "jge AtYES» unfulfilled conditions or commands.

Command «jeq" and "JNE" migrating uniformity and inequalities cases. Given that the command "cmp" is actually a team of deprivation, these cases account for, respectively, installed and uninstalled zero flag «Z». From these considerations, we can see that the command «jeq" the essence of the command «jz» Alternative appearance (do not mix with the emulated commands), but the command "jne" meets «jnz».

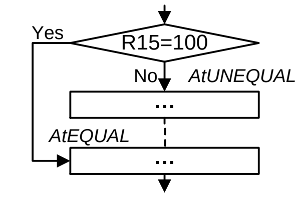
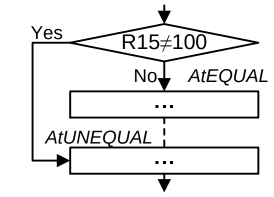
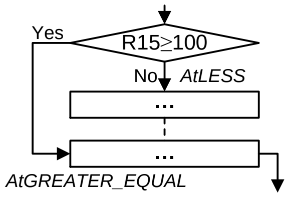
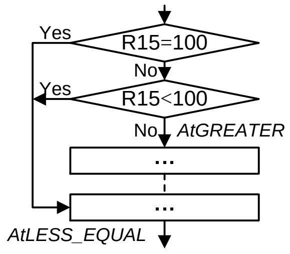
Table VII refers to cases where the numbers may have to be treated with a negative part (2’s complement) the format is called "Signed integer" - integers with a sign in the range -32,767 ... + 32,767). If the format is "unsigned integer" - integers without signs in the range 0 .. + 65535 then have to use a slightly different command set (Table VIII), where the "jl" ("jump IF less") commands are used instead «jl» ("jump IF lower ") and "jge" command ("jump IF greater or equal") - instead "jhs" ("IF jump higher or the same ').

The same could be said about the 8-bit numbers, which can be analyzed with an 8-bit command «cmp.b». Table VII Imight be useful format with sign (-127 ... 127), but the format without the sign (0 ... 255) should be used in a second table (Table 3 2), provided that the "cmp" is aizveidots with "cmp. b. "

Of course 16-bit numbers 0 ... + 32767 (as well as 8-bit 0 ... + 127) can be used for both command sets in this range of numbers the picture is the same in both formats.

Carefully analyzing the "JHS" and "JLO" command may be noticed that these two commands has been associated with the "C" flag. If it is 1, then the transition team performs "JHS", but if 0 - then "JLO". This suggests that these two commands are, respectively, alternative symbolic images of "JC" and "JNC" commands.

Table VII. Arithmetical conditions for unsigned data

	Condition	Assembly code
a)	„Is R15=100?”  	... cmp #100,R15 jeq AtEQUAL AtUNEQUAL ...
b)	„Is R15≠100?”  	... cmp #100,R15 jne AtUNEQUAL AtEQUAL ...
c)	„Is R15≥100”  Yes R15≥100 No AtLESS ... ... AtGREATER_EQUAL	... cmp #100,R15 jge AtGREATER_EQUAL AtLESS ...
d)	„Is R15≤100?”  Yes R15=100 No Yes R15<100 No AtGREATER ... ... AtLESS_EQUAL	... cmp #100,R15 jeq AtLESS_EQUAL jl AtLESS_EQUAL AtGREATER ...

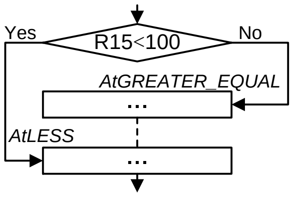
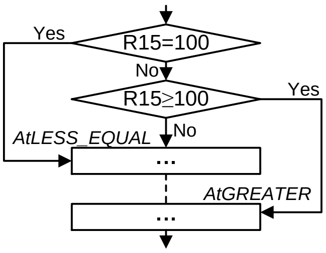
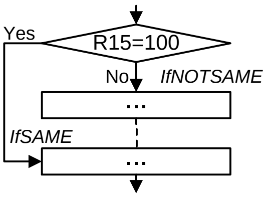
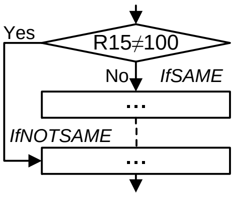
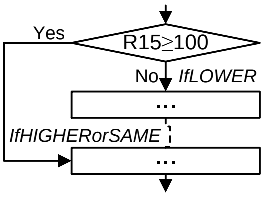
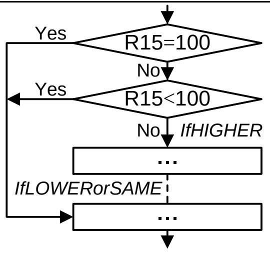
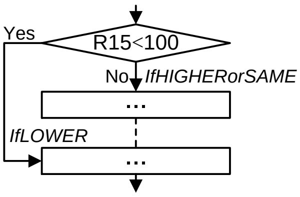
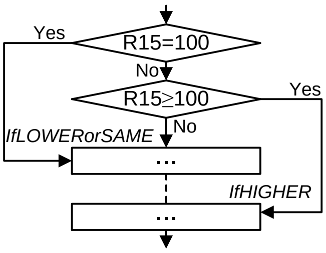
e)	„Is R15<100?” 	... cmp #100,R15 jl AtLESS AtGREATER_EQUAL ...
f)	„Is R15>100?”  Yes R15=100 No R15≥100 Yes AtLESS_EQUAL No … AtGREATER …	... cmp #100,R15 jeq $+4 jge AtGREATER AtLESS_EQUAL ...

Table VIII. Arithmetical conditions for signed data.

	Condition	Assembly code
a)	„Is R15=100?” 	... cmp #100,R15 jeq IfSAME IfNOTSAME ...
b)	„Is R15≠100?” 	... cmp #100,R15 jne IfNOTSAME IfSAME ...
c)	„Is R15≥100?” 	... cmp #100,R15 jhs IfHIGHERorSAME IfLOWER ...
d)	„Is R15≤100?”	... cmp #100,R15 jeq IfLOWERorSAME jlo IfLOWERorSAME

	 Yes R15=100 No Yes R15<100 No IfHIGHER ... IfLOWERorSAME ...	**IfHIGHER ...**
e)	„Is R15<100?”  Yes R15<100 No IfHIGHERorSAME ... IfLOWER ...	**...** **cmp #100,R15** **jlo IfLOWER** **IfHIGHERorSAME ...**
f)	„Is R15>100?”  Yes R15=100 No R15≥100 Yes No IfLOWERorSAME ... IfHIGHER ...	**...** **cmp #100,R15** **jeq $+4** **jhs IfHIGHER** **IfLOWERorSAME ...**

Arithmetical conditions in MSP430 microcontroller

Logic conditions split the program, depending on the individual bit or bit group status. Microcontroller peripheral control register bits play a key role in either the signal of a hardware event or a setting the control mode. Therefore, logic conditions microcontroller assembler programming is also of general interest. For example, the condition "is flag WDTIFG set" can also be interpreted as "is watchdog counting cycle complete" either as "did watchdog overflow occur".

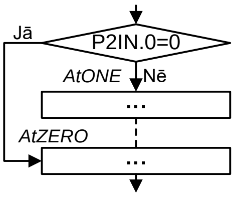
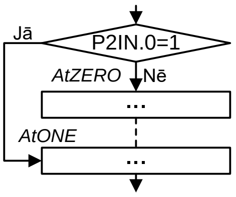
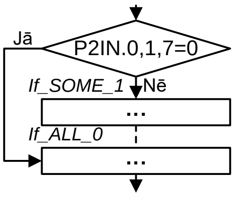
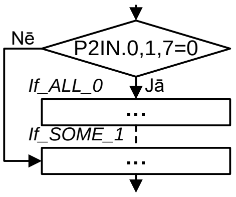
Individual bit analysis is provided for the command "bit" with a test mask (the first argument put directly - or through direct addressing), where under test Bit to 1 and the remaining bits are 0 (assembler teams «bit» it usually asked in binary form). Such command set zero flag Z, if the tested bit is 0, or delete this flag, if the first bit in fact, it can be said that the zero flag Z "bit" command execution becomes equal to the inverted bit analyzed. If the condition is "or bit is zero," then with the team «JZ» can make the transition to the program site, which is fulfilled in the case of zero (Table IX-a). Either the contrary - if the condition is "BITS is the number one", then the command «jnz» The transition to the 1 treatment (Table IX-b). Of course, these two conditions are mutually inverse logic and after the transition team can deploy a team that meets the unfulfilled conditions (or section opposite condition).

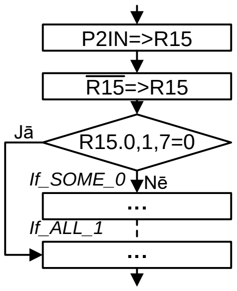
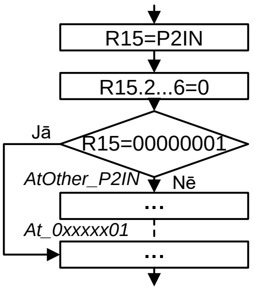
With the command "bit" bit groups may also be analyzed. The condition will work similarly, but it should be remembered that the zero flag Z will be 1 only if all analyzed bits are equal to zeros. Therefore, if the command "bit" is a multi-bit mask, regardless of the type of the transition command ("jz" or "jnz"), the condition is a "test if all bits are 0" (Table IX-c).

MSP430 command system does not have special commands that might be tested a number of bits, or all of them are equally singlehanded. Therefore, the following conditions shall be implemented by a number of command. For example, the test data can invert and then tested on the zeros. Generally, in order to leave the same data without any changes to be tested with a copy of a processor register. Then condition includes copying, inversion, and only their own testing (Table IX-d).

Often you need to test whether certain bits are fixed values. For example, the condition could be "or P2IN.7 = 0 and = 0 and P2IN.1 P2IN.0 = 1". If the remaining bit values are known, such testing could be done with the command "cmp". However, other bit of attention is usually not defined. To avoid this uncertainty, the test data processor copies the registry, delete the unwanted bits (if installed) and the results of tests for compliance with the specified value. In this case, after copying the bits 2 ... 6 erased. If bit 7 is 0, 1 bit is 0 and the latest bit is 1, the formed test code "00000001" (Table IX-e).

Table IX. Bit test conditions

	Condition	**Assembly code**
a)	„Is P2IN.0=0?” 	... bit.b #00000001b,&P2IN jz AtZERO AtONE ...
b)	„Is P2IN.0=0?” 	... bit.b #00000001b,&P2IN jnz AtONE AtZERO ...
c)	„iIs P2IN.0=P2IN.1=P2IN.7=0?”  Jā P2IN.0,1,7=0 If_SOME_1 Nē ... If_ALL_0 ...	... bit.b #10000011b,&P2IN jz If_ALL_0 If_SOME_1 ...
	 Nē P2IN.0,1,7=0 If_ALL_0 Jā ... If_SOME_1 ...	... bit.b #10000011b,&P2IN jnz If_SOME_1 If_ALL_0 ...

d)	„Is P2IN.0=P2IN.1=P2IN.7=1?”  P2IN=>R15 $\overline{R15}$=>R15 Jā R15.0,1,7=0 If_SOME_0 Nē ... If_ALL_1 ...	... mov.b &P2IN,R15 inv.b R15 bit.b #10000011b,R15 jz If_ALL_0 If_SOME_1 ...
e)	„Is P2IN.7,1=0 and P2IN.0=1?”  R15=P2IN R15.2...6=0 Jā R15=00000001 AtOther_P2IN Nē ... At_0xxxxx01 ...	... mov.b &P2IN,R15 bic.b #01111100b,R15 cmp.b #00000001b,R15 jz At_0xxxxx01 AtOtherP2IN ...

Finally, mention should be the case when it is necessary to analyze all test-bit combinations. Continuing the previous example, we can say that three bits 23 = 8 combinations and testing should be repeated 8 times. However, copying and deleting that little bit can not repeat, as a result of repetitive commands only "cmp" and "JZ".

Code example 4. Testing of all possible combinations of P2IN.0, P2IN.0 and P2IN.7.

```
...
mov.b   &P2IN,        R15          ;Copy data for testing
bic.b   #01111100b,    R15          ;Clear bits 2...6
cmp.b   #00000000b,    R15           ;Test for 0xxxxx00
jz      At_0xxxxx00
cmp.b   #00000001b,R15              ;Test for 0xxxxx01
jz      At_0xxxxx01
cmp.b   #00000010b,R15              ;Test for 0xxxxx10
jz      At_0xxxxx10
cmp.b   #00000011b,R15              ;Test for 0xxxxx11
jz      At_0xxxxx11
cmp.b   #10000000b,R15              ;Test for 1xxxxx00
jz      At_1xxxxx00
cmp.b   #10000001b,R15              ;Test for 1xxxxx01
```

```
        jz     At_1xxxxx01
        cmp.b   #10000010b,R15              ;Test for 1xxxxx10
        jz     At_1xxxxx10
At_1xxxxx11    ...                          ;Process code 1xxxxx11
```

If tested bits are taking neighboring positions it is easy to implement so-called calculated branch. This technique is particularly advantageous if the test bit and the number of combinations is large. For example, if, depending on the register R4 8, 9, 10 and 11-bit values, one of 16 programs has to be chosen, it can be done as a calculated branch. Then: 1) data are copied; 2) bits 8 ... 11 are aligned with positions 1 ... 4 bits (in this case it is more effective to apply the team «swpb» and «RLA» seven «RRA» team in place); 3) the resulting code 0,2,4 ... 32 is added to the program counter PC thus making the transient to a "JMP" command, which further provides a transient to the corresponding program. If these blocks are small and of the same length, then "jmp"-s are replaced by the program blocks but the calculated shifts are equal to their lengths.

Code example 5. Calculated branch.

```
        ...
        mov    R4,          R15         ;Copy data
        swpb   R15                      ;Exchange bytes in R15
        rla    R15                      ;Move bits to 1...4
        bic    #1111111111100001b,R15        ;Clear bits except 1..4
        add    R15,         PC          ;Add BITS*2 to PC
        jmp    At_X0XX
        jmp    At_X1XX
        jmp    At_X2XX
        jmp    At_X3XX
        jmp    At_X4XX
        ...
        jmp    At_XFXX
At_X0XX        ...                      ;Process code 0000
```

Cycles in assembly code

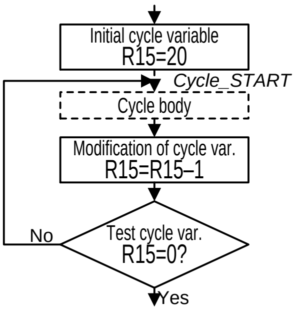
Consequently, the assembler language often does not include any special cycle command. This is why cycles are also implemented as a number of instructions. Functional meaning of the cycle instructions are: verifying the condition of cycle continuation (typically a bit condition) and transient to the beginning of the cycle. If the cycle has a limited number of repetitions, the key role in the formation of the cycle plays cycle variable which is defined outside the cycle and before it. During the execution of the cycle it is modified and checked for compliance with cycle condition. Cycle execution completes when this condition is met. So the cycle algorithm contains the following blocks: 1) defining a cycle variable; 2) changing the cycle variable within the cycle; 3) cycle condition test.

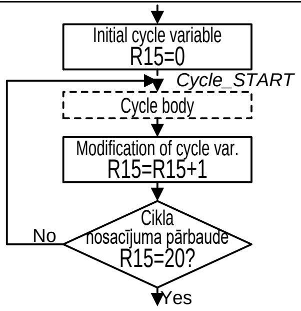
In MSP430 assembler language, the cycle variables are usually defined in the processor registers (compilers of higher-level languages also do that quite often). Like any other operand processing with register addressing, the cycle counting is faster and takes up less space in program memory. This advantage becomes decisive taking into account that the cycle variables are handled regularly and often.

Theoretically cycle variable may be modified at the beginning and at the end. However, if there are no additional conditions, it is more advantageous if the cycle variable in each cycle is reduced by 1 until zero is obtained (Table X-a). Then its initial value is the number of repetitions. Explicit comparison with 0 is not necessary due to the fact that after the decrement instruction, if the outcome is zero, zero flag Z is automatically set and it is possible to immediately make the transient to the beginning of the cycle with the command «jnz».

Increment instructions also set flags. For example, 255+1 (if an 8-bit operation) forms a zero result and sets zero flag Z. However, the cycle variable in the beginning of the cycle is not equal to the number of repetitions of the cycle, but equal to 256 – "number of repetitions", which reduces the readability of the program. For this reason, the ascending cycle variable starts usually from zero. Then the cycle ends when the loop variable becomes equal to the number of repetitions (Table X-b). In order to verify this condition an extra "cmp" instruction has to be added to the cycle.

Table X. Assembly cycles with cycle variable

	Type	**Cycle implementation**
a)	Descending cycle variable  Initial cycle variable R15=20 Cycle_START Cycle body Modification of cycle var. R15=R15–1 No Test cycle var. R15=0? Yes	**...** **mov #100,R15** **Cycle_START ... ;cycle's body** **dec R15** **jz Cycle_START** **...**
b)	Ascending cycle variable	**...** **mov #0,R15** **Cycle_START ... ; cycle's body** **inc R15** **cmp #100,R15**

	 Initial cycle variable R15=0 Cycle_START Cycle body Modification of cycle var. R15=R15+1 Cikla nosacījuma pārbaude R15=20? No Yes	**jnz Cycle_START** **...**

Cycle variables are often used for other purposes. For example, a cycle variable can also be useful in the case if the cycle processes an array. Typically, the initial array is located outside the processor and array elements are accessed by means of the indexed addressing. Then the applied index is equal to the array start address, but the content of the register is the address offset from the beginning. It is proportional to the length of the array elements and to the element number. For example, copying a 32-bit number from the registers R10 and R11 into the 10th element of 32-bit (4-byte) numeric array can be done by the instructions:

Code example 6. Defining a 32-bit element of and array (>10 elements).

```
Write_10th    mov    #10*4,        R15          ;Prepare address shift
          mov    R10,          0x200(R15)     ;Copy LS byte
          mov    R11,          0x200+2(R15)   ;Copy MS byte
```

Here the array start address is 0x200 and the array element numbering starts from 0. Cycle often handles all array elements, by bringing to each array element in turn. In this example, assuming that the array has 20 elements, the offsets of their addresses are 0, 4, 8, 76 ... (or the same in reverse order). Then (Table XI) cycle variable changes from 20 to 1 (0 is already the illegal value) or 0 to 19 (in this case the value which is not processed is 20).

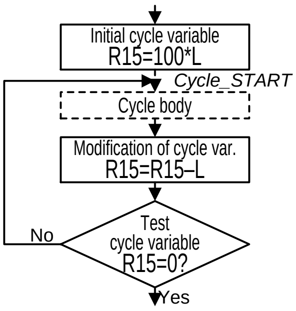
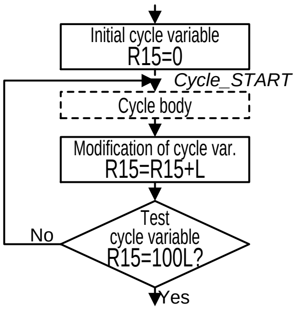
In this case, a well-formed cycle would use a variable that corresponds to the address offset. However the cycle variable should be reduced / increased by L = 4 (the length of the array elements). If the array must be processed in descending order, the cycle of changing from the starting value 20 × 4 = 80, but the end-condition - zero cycle variable (Table XI-a). If the array numbers is processed in the ascending order, the initial cycle variable is 0, but the cycle ends, if it is 20 × 4 = 80 (Table XI-b).

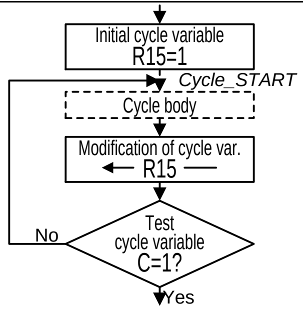
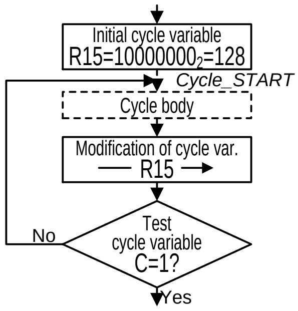
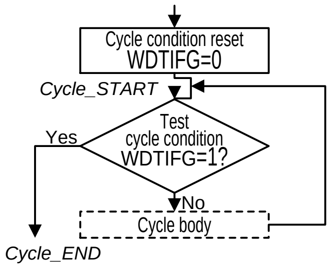
Another case where the loop variable can be used in a special way, a cyclical analyzing / setting / clearing of a bit in a register or in a memory cell (such as a digital port control registers). Consequently, the number of bits to be analyzed is usually small (usually - 8 or 16), the elementary cycle variable varies in a small range - from 1 to 8 or 1 to 16. However, the cycle still needs changing bit mask. For example, if the number of bits is 8 and the beginning of analyzes 0 bit, the cycle variable is 0, the bit mask is 00000001b=1. In the second sweep the variable is 1, the mask 00000010b=2, in the third – 2 and 00000100b=4 and so on. It can be noticed, that the mask is 2 in power "loop variable". In such cases, much more efficient to create a cycle variable that is a bit mask. If bits are analyzed from the least significant to the most significant, then the starting value is 00000001b = 1, it is cyclically shifted to the left through

C and in the end C = 1, which is the condition for cycle ending (Table XI-c). If, by contrast, the bits are analyzed from the most significant to the least significant, then the starting value is 10000000b = 128, it is cyclically shifted to the right through carry flag C, but the end-criterion is again C = 1 (Table XI-d). 16-bit version is similar, but bit masks and cycle variable maximum value is 1000000000000000 = 2^{15} = 32768.

Finally, the cycles without cycle variable, but with cycle ending condition should be mentioned. Such cycles are often used if the ending criterion is a hardware event, which sets the dedicated flag-bit. It is assumed that before the first cycle repetition this criterion is not truth, but over the cycle, these criterion is checked. Then there is no need to define and control the cycle variables and the cycle consists only of its body (useful part) and cycle ending condition (Table XI-e).

Table XI. Cycles with double use of the cycle variable

	Type	**Cycle implementation**
a)	Descending address offset  Initial cycle variable R15=100*L Cycle_START Cycle body Modification of cycle var. R15=R15–L Test cycle variable R15=0? No Yes	@ L=4 **...** **mov #80,R15** **Cycle_START ... ; cycle's body** **sub #4,R15** **jz Cycle_START** **...**
b)	Ascending address offset  Initial cycle variable R15=0 Cycle_START Cycle body Modification of cycle var. R15=R15+L Test cycle variable R15=100L? No Yes	@ L=4 **...** **mov #0,R15** **Cycle_START ... ; cycle's body** **add #4,R15** **cmp #80,R15** **jnz Cycle_START** **...**
c)	Bit shifted to the left	**...** **mov.b #00000001b,R15** **Cycle_START ... ; cycle's body** **clrc** **rlc.b R15**

	 Initial cycle variable R15=1 Cycle_START Cycle body Modification of cycle var. R15 No Test cycle variable C=1? Yes	jnc Cycle_START ...
d)	Bit shifted to the right  Initial cycle variable $R15=10000000_2=128$ Cycle_START Cycle body Modification of cycle var. R15 No Test cycle variable C=1? Yes	... mov.b #10000000b,R15 Cycle_START ... ; cycle's body clrc rrc.b R15 jnc Cycle_START ...
e)	Cycle without variable  Cycle condition reset WDTIFG=0 Cycle_START Yes Test cycle condition WDTIFG=1? No Cycle body Cycle_END	... bic.b #WDTIFG,&IFG1 Cycle_START bit.b #WDTIFG,&IFG1 jnz Cycle_END ... ; cycle's body jmp Cycle_START Cycle_END ...

Assembly arithmetic

It can be noticed that the MSP430 assembler arithmetic command group is rather limited. It is actually just addition and subtraction commands. Also, the data formats longer than 16 bits cannot be processed by a solo instruction. Therefore, the longer data formats, as well as multiplication, division or more complex operations are carried out as program subroutines. Assembler arithmetic is a broad theme, which out of the scope of this material. So this chapter deals with only the most frequently applied operations with data: with multiplication and division of whole constant integral constant.

Longer formats

Addition and subtraction for the data words, longer than 16 bits are carried out through carry flag C starting from the list significant word. Processing of the least significant parts of the longer data units can be done with the same instructions as 8/16 bit data words – "add" and "sub". These instructions can modify carry flag C. This is why further parts have to take C into account – instructions "addc" and "subc" are suitable for these most significant parts.

Let's continue Code example 6 and perform the addition of 10th 32-bit element of the array located at 0x200 to the 10th 32-bit element of another array located at 0x300. This can be done in steps by 8 (Code example 7) or by 16 bits (Code example 8). It is obvious, that the firs example has longer execution time and bigger machine code. So, 16 access is preferable, except for the cases where remaining groups of bits has length <8, for example, when processing 24 bit data words.

Code example 7. 32-bit addition of array elements (8-bit access).

```
Add_32      mov    #10*4,        R15          ;Prepare address shift
            add.b   0x200(R15),    0x300(R15)     ;Add 8 LS bits
            addc.b  0x200+1(R15),  0x300+1(R15)   ;Add bits 8...15
            addc.b  0x200+2(R15),  0x300+2(R15)   ;Add bits 16...31
            addc.b  0x200+3(R15),  0x300+3(R15)   ;Add 8 MS bits
```

Code example 8. 32-bit addition of array elements (16-bit access).

```
Add_32      mov    #10*4,        R15          ;Prepare address shift
            add     0x200(R15),    0x300(R15)     ;Add 16 LS bits
            addc    0x200+2(R15),  0x300+2(R15)   ;Add 16 MS bits
```

Multiplication by means of addition

Most primitive multiplication by the constant K can be made just by adding the multiplier to a sum K times. The corresponding program actually only defines the initial zero-sum cycle and organize a number of repetitions K as well as adding the first multiplier to the sum. For example, multiplication of R10 by 105 could be done by the following program:

Code example 9. Multiplication by means of addition.

```
            mov    #105,        R11          ;(2) R11=K=105
            mov    #0,          R4           ;(1) R4=S=0
Multiply105    add    R10,         R4           ;(1) R4=S=S+X
            dec    R11                       ;(1) Repeat 105 times
            jnz    Multiply105               ;(2)
```

It is not difficult to see that this type of multiplication is quite time-consuming. In addition, the higher is the constant the longer is execution time. For example, the execution time of Code example 9 can be calculated as T_{105}=(2+1)+105(1+1+2)=3+105*4=423 clock cycles. In the same time, the multiplication by 840 would require T_{840}=3+840(1+1+2)=3+840*4=3363 clock cycles.

Multiplication by means of additions and shifts.

Much faster multiplication with the constant K can be executed as a multiplication by the sum of powers of two. Such multiplication assumes that M-bit binary K is represented by the formula

$K= K_M\times 2^M+ K_{M-1}\times 2^{M-1}+ ... + K_2\times 2^2+ K_1\times 2^1+ K_0\times 2^0$,

where $K_M \ldots K_0$ are bits of the constant. Then

$X\times K= X\times (K_M\times 2^M+K_{M-1}\times 2^{M-1}+ ... +K_2\times 2^2+K_1\times 2^1+K_0\times 2^0)$

and

$X\times K= X\times K_M\times 2^M+X\times K_{M-1}\times 2^{M-1}+ ... +X\times K_2\times 2^2+X\times K_1\times 2^1+X\times K_0\times 2^0$.

Since the bits are 0 or 1, the sum actually consists only of a multiplication of the bit and a power of two. In addition, if bit is 1, the sum of the product involved, but if it is 0, then not present. In this case, this result can be expressed as follows:

$X\times 105=X\times 1101001_2=X\times (1\times 2^6+1\times 2^5+0\times 2^4+1\times 2^3+0\times 2^2+0\times 2^1+1\times 2^0)=64\times X+32\times X+8\times X+X$.

In turn, multiplication by two un power N is easy - it means the addition of the multiplier to itself N times, or shifted by n positions to the left (arithmetical shift “rla”). In fact, it shows out that simple multiplication routine consists of multiple additions of the multiplicand to itself (multiplication by 2) or to the result. The program, which multiplies R10 by 105 looks like this:

Code example 10. Multiplication by 105 by means of additions and shifts.

```
MUL105_Fast   mov   #0,       R4       ;(1) R4=S=0
        add   R10,      R4       ;(1) R4=0+X=X
        add   R10,      R10      ;(1) R10=2X
        add   R10,      R10      ;(1) R10=4X
        add   R10,      R10      ;(1) R10=8X
        add   R10,      R4       ;(1) R4=X+8X=9X
        add   R10,      R10      ;(1) R10=16X
        add   R10,      R10      ;(1) R10=32X
        add   R10,      R4       ;(1) R4=9X+32X=41X
        add   R10,      R10      ;(1) R10=64X
        add   R10,      R4       ;(1) R4=41X+64X=105X
```

The program takes 11 clock cycles. In the same time, the multiplication by the other constant (840) requires completely different formula:

$X\times 840=X\times 11010010002=X\times (1\times 26+1\times 25+0\times 24+1\times 23+0\times 22+0\times 21+1\times 20)=512\times X+256\times X+64\times X+8\times X$,

which corresponds to the program:

Code example 11. Multiplication by 840 by means of additions and shifts.

```
MUL840_Fast   mov   #0,        R4        ;(1) R4=S=0
          add   R10,       R10       ;(1) R10=2X
          add   R10,       R10       ;(1) R10=4X
          add   R10,       R10       ;(1) R10=8X
          add   R10,       R4       ;(1) R4=0+8X=8X
          add   R10,       R10       ;(1) R10=16X
          add   R10,       R10       ;(1) R10=32X
          add   R10,       R10       ;(1) R10=64X
          add   R10,       R4       ;(1) R4=8X+64X=72X
          add   R10,       R10       ;(1) R10=128X
          add   R10,       R10       ;(1) R10=256X
          add   R10,       R4       ;(1) R4=72X+256X=328X
          add   R10,       R10       ;(1) R10=512X
          add   R10,       R4       ;(1) R4=328X+512X=840X
```

the execution time is 14 clock cycles, which is only 3 clock periods more than before. The longest case of K M-bit integer with 1 in all bits, would require time is 2M (multiplication by 2 takes 1 cycle and adding to the sum or result – 1 more, 1 clock additionally - the defining of initial sum, and minus 1 because 0 in power 2 is the initial multiplicand).

Dividing by means of subtraction.

Primitive dividing can be realized similar to multiplication by addition. This time is the divider D subtracted cyclically from the X, each time increasing the result by 1 until the current balance X * is greater than separating D. As before, the beginning of the program is to define a zero result. This cycle is wholesome of comparison, which is located at the beginning of the cycle as the sections may be less than the divider (the result is 0). If the balance is not necessary, then the comparison can be carried out by «sub»:

Code example 12. Dividing by means of subtraction.

```
          mov   #105,      R11       ;(2) R11=D=105
          mov   #0,        R4        ;(1) R4=Res=0
Divide105   sub   R11,       R10       ;(1) R10=Rem-D=???
          jl    Next                 ;(2) Exit if X<D
          inc   R4                   ;(1) Res=Res+1
          jmp   Divide105            ;(2)
Next        ...
```

However, the program can also be an additional goal - to find the reminder after division. Then, the program shall include a comparison, but subtraction is executed only if the current balance X * is greater than the divisor D. If the dividend is register R10, the program is as follows:

Code example 13. Dividing by means of subtraction and reminder finding.

```
            mov    #105,        R11         ;(2) R11=D=105
            mov    #0,          R4          ;(1) R4=Res=0
Divide105      cmp    R11,        R10         ;(1) R10-R11=Rem-D=???
            jl     Next                     ;(2) Exit if X<D
            sub    R11,         R10         ;(1) Rem=Rem-D
            inc    R4                       ;(1) Res=Res+1
            jmp    Divide105                ;(2)
Next        ...
```

Similarly to the multiplication with addition, subtraction with division fulfilled a long time. Also here - the higher the reult, the longer execution time. For example, if the dividend is 10500_{10} (which gives a result of 100), then given during program execution:

T_{105}=(2+1)+100(1+2+1+1+2)=3+100*7=703 clock cycles,

but if the divider is 525 (5 times higher) it is

T_{105}=(2+1)+500(1+2+1+1+2)=3+500*7=3503 clock cycles.

Multiplication by a fractional constant and dividing by a constant as multiplication

Any division of a whole number can be represented as multiplication with fractions. For example, division by 8 is a multiplication by 0.125. Similarly, multiplication by an integer division of the basic operations can be performed using the binary number value formula (only this time - fractional part):

$D= D_{-1}\times2^{-1}+ D_{-2}\times2^{-2}+ ... + D_M\times2^M$.

Then the dividing is

$X/D^*=X\times D=X\times(D_{-1}\times2^{-1}+D_{-2}\times2^{-2}+ ... +D_M\times2^M)$

Here it should be noted that converting fractions in binary form, often obtained an infinite number of decimal places. Then obtaining fractions of multiplier D* must take as many bits as providing accurate answer. One can use a simple rule that 10 binary characters corresponding to 3 decimal. For example, dividing by 5 obtains a multiplier 0210, a binary image with an accuracy of 0.001_{10} is 0.0011001100_2. When you open a parenthesis in the previous terms, we obtain:

$X/D^*=X\times D_{-1}\times2^{-1}+X\times D_{-2}\times2^{-2}+ ... + X\times D_M\times2^{-M}$.

Just as multiplication by an integer, this sum consists of shifted X multiplied by 0 or 1 – a partial result. This time, the divisor is negative, which actually means dividing. Divide by 2N is as easy as multiply. It is a shift of N bits, but this time - to the right. Multiplication by a fraction has a fundamental feature – fractional part of partial results, which should not be ignored (that often happens when drawing up the assembler program). For example, 63 multiplied $0.875_{10}=0.111_2$ gives 55.125≈55. When multiplying if the partial multiplications of are not ignored, it produce exact result (Figure 3.12.-a). However, if you only count the integer partial multiplications, the result is 53 (Figure 3.12.-b). Error is the greater, the more bits a multiplier D. For example, if 63 is multiplied by $0.9375_{10} = 0.1111_2$, the exact result is 59.0625≈59 (Figure 3.12.-c), but inaccurate - 56 (Figure 3.12.-d). Error (3) is greater than the above (2). It follows to an important conclusion - multiplication by a fractional number requires that the fraction parts of the partial result has to be stored, which makes the program more complex.

```
     1 1 1 1 1
   1 1 1 1 1 1 1
   .................
   0 1 1 1 1 1.1 0 0 = 3 1.5 0 0
+  0 0 1 1 1 1.1 1 0 = 1 5.7 5 0
+  0 0 0 1 1 1.1 1 1 =   7.8 7 5
   -----------------------------
   1 1 0 1 1 1.0 0 1 = 5 5.1 2 5
```

a) 63×0.875 (exact)

```
     1 1 1
   1 1 1 1 1
   ...........
   0 1 1 1 1 1 = 3 1
+  0 0 1 1 1 1 = 1 5
+  0 0 0 1 1 1 =   7
   -----------------
   1 1 0 1 0 1 = 5 3
```

b) 63×0.875 (partial fractional parts ignored)

```
       1 1 1 1
     1 1 1 1 1 1
   1 1 1 1 1 1 1 1
   .......................
   0 1 1 1 1 1.1 0 0 0 = 3 1.5 0 0 0
+  0 0 1 1 1 1.1 1 0 0 = 1 5.7 5 0 0
+  0 0 0 1 1 1.1 1 1 0 =   7.8 7 5 0
   0 0 0 0 1 1.1 1 1 1 =   3.9 3 7 5
   ---------------------------------
   1 1 1 0 1 1.0 0 0 1 = 5 9.0 6 2 5
```

c) 63×0.9375 (exact)

```
       1 1
     1 1 1 1
   1 1 1 1 1
   ...........
   0 1 1 1 1 1 = 3 1
+  0 0 1 1 1 1 = 1 5
+  0 0 0 1 1 1 =   7
   0 0 0 0 1 1 =   3
   -----------------
   1 1 1 0 0 0 = 5 6
```

d) 63×0.9375 (partial fractional parts ignored)

Figure 3.12. Multiplying by a fractional number

As an example, let’s look at the division by 5 (or multiplication by $0.2_{10} = 0.00110011_2$). Fraction parts of partial results will be kept in the register R11, but the fraction part of result – in register R5. Integer part of partial result and common dividend are in register R10, but the integer part or result – in register R4. Then the program will look like this:

Code example 14. Exact dividing by a constant.

```
DIV5      clr    R4              ;Res=R4.R5=0
          clr    R5
          rra    R10             ;R10.R11=X/2
          rrc    R11             ;
          rra    R10             ;R10.R11=X/4
          rrc    R11             ;
          rra    R10                 ;R10.R11=X/8
          rrc    R11                 ;
          add    R11,      R5          ;Res=Res+X/8
          addc   R10,      R4          ;
          rra    R10                 ;R10.R11=X/16
          rrc    R11                 ;
          add    R11,      R5          ;Res=Res+X/16
          addc   R10,      R4          ;

          rra    R10                 ;R10.R11=X/32
          rrc    R11                 ;
          rra    R10                 ;R10.R11=X/64
          rrc    R11                 ;
          rra    R10                 ;R10.R11=X/128
          rrc    R11                 ;
          add    R11,      R5          ;Res=Res+X/128
          addc   R10,      R4          ;
          rra    R10                 ;R10.R11=X/256
          rrc    R11                 ;
          add    R11,      R5          ;Res=Res+X/256
          addc   R10,      R4          ;
```

As can be seen, the program will not be short, but its execution time does not depend on the value. The increased length is due to the need to process the fractions parts. They are, as previously found a significant impact on the result. However, there are some roads as to avoid them. The first - to use a special format. For example, if all the numbers are from the range of 0 ... 255 and requires numerical accuracy to 3 decimal places, the decimal point can be located between the 16-bit word 7 and 8 bits (this is a simple example of a fixed format comma). Then, for example, it looks like a 63_{10} dividend 0011111100000000_2 but multiplier-divider 0.2_{10} – 0.000000000110011_2. Then all operations take place within one 16-bit word within (the MSP430 processor register within):

Code example 15. Exact dividing by a constant (enhanced).

```
DIV5      clr    R4                    ;Res=R4=0
          rra    R10                 ;R10=X/2
          rra    R10                 ;R10=X/4
```

```
        rra   R10                ;R10=X/8
        add   R10,      R4       ;R4=Res=Res+X/8
        rra   R10                ;R10=X/16
        add   R10,      R4       ;R4=Res=Res+X/16
        rra   R10                ;R10=X/32
        rra   R10                ;R10=X/64
        rra   R10                ;R10=X/128
        add   R10,      R4       ;R4=Res=Res+X/128
        rra   R10                ;R10=X/256
        add   R10,      R4       ;R4=Res=Res+X/256
```

Actual result is 8 MS bits of R4. It can be moved to 8 LS bits by means of "swpb"

```
FIXPOINT2INT   bic   #0x00FF,    R4      ;R4.7...0=0
        swpb   R4                ;R4.15...8<->R4.7...0
```

Correct rounding requires the analysis of bit 7 and adding it to the result

```
FIXPOINT2INT   swpb   R4                 ;R4.15...8<->R4.7...0
        bit   #1000000000000000b,R4      ;Test bit 15 (fromer 7)
        jz    $+4
        inc   R4                 ;Increment @ XXX.1b
        bic   #0xFF00,     R4    ;R4.8...15=0
```

Another way to avoid the calculation of fraction part of the divider increase multiplier-divider by 2^N (N times be moved to the left), and then the results must divided by this number (be moved to the right). In this case, multiplication is by an integer, but dividing by 2^N simply shift to the right. For example, multiplication by $0.2_{10} \approx 0.00110011_2$ can meet as multiplication by $110011_2 = 51_{10}$ and division is by $2^8 = 100000000_2 = 256_{10}$. Ecaxtly it is equal to 0.1992187510, which is a 3-digit accuracy corresponds to the desired ratio 0.2. This multiplication corresponds to the program:

Code example 16. Exact dividing by a constant (enhanced 2).

```
DIV5      clr   R4               ;Res=R4=0
        add   R10,      R4       ;R4=Res=0=Res+X
        rla   R10                ;R10=2X
        add   R10,      R4       ;R4=Res=X+2X=3X
        rla   R10                ;R10=4X
        rla   R10                ;R10=8X
        rla   R10                ;R10=16X
```

```
add    R10,        R4          ;R4=Res=3X+16X=19X
rla    R10                     ;R10=32X
add    R10,        R4          ;R4=Res=19+32X=51X
rra    R4                      ;R4=51X/2
rra    R4                      ;R4=51X/4
rra    R4                      ;R4=51X/8
rra    R4                      ;R4=51X/16
rra    R4                      ;R4=51X/32
rra    R4                      ;R4=51X/64
rra    R4                      ;R4=51X/128
rra    R4                      ;R4=51X/256=0.199X
jnc    $+4                     ;Rounding:
inc    R4                      ;Res=51X/256+1 at XXX.1b
```

As can be seen, the MSP430 microcontroller processor lets you multiply and divide by the constant. However, it should be noted that the program's multiplication and division are not as effective as multiplication by hardware means, for example, with a help of a hardware multiplier installed in some MSP.

3.5. Peripheral devices of MCUs

In a well-designed MCU-based embedded system the most of the work done by the MCU typically is done by its peripheral devices of the MCU. The program executed by MCU's processor tunes the peripherals devices and transfers data between them. The number and functionality of the peripheral devices depends on the particular model of the used MCU. However, it is possible to split the peripheral devices of any MCU into two parts: basic peripheral devices which are necessary for elementary operation of the MCU and advanced peripheral devices which simplifies the development of particular technical solution. Basic peripheral devices are digital general purpose inputs and outputs – knowing their operation principles allows to build a complete MCU based control system. In the case of MSP430 basic peripheral devices are watchdog timer and clock system – their programming is necessary for successful and optimal start of MSP430.

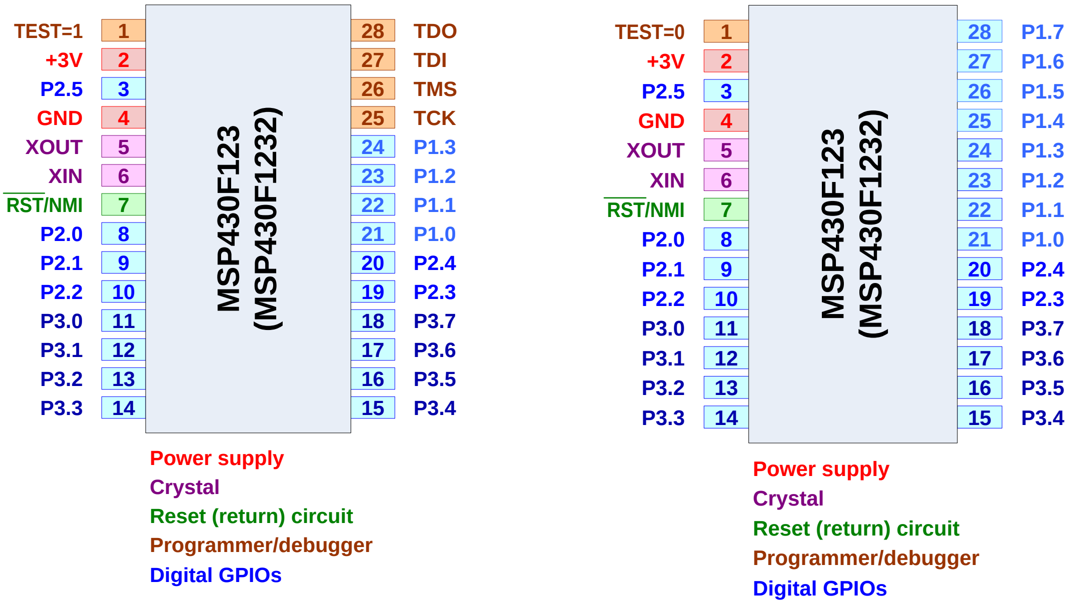
General purpose inputs and outputs (GPIOs)

MSP430 pins

Microcontroller chips have many contacts and is mostly of a general digital input / output. For example, MSP430F123 microcontrollers (2) produced in the SO and TSSOP packages with 28 pins. Of these, two pins are designed for power supply (2 - + 3V and 4 - land), another 2 (5 and 6) - to connect to the microcontroller quartz resonator and one (7) - an external signal RST return connection. In addition, the 4 contacts (25 ... 28th) are required to connect the debugging device (typically a computer with a special unit - programmers). In cases where the chip contacts are few in number, spend 4 contacts only for debugging purpose, it is too irrational. Therefore, the contacts can also serve as a universal digital input / output. However, to select the contact mode, you need another contact TEST (1 foot). If the contact connected to the supply voltage, then 25 feet ... 28 serves debugging purposes (Figure 4 1-a), but if the test is connected to the ground, then 25 ... 28 digital input / output (Figure 12-b).

In general MSP430F123 (2) microcontrollers have two configurations. The first (Figure 12-a) is effective when running the debug interface. Then microcontroller 10 contacts with special functions and 18 digital inputs / outputs. The second configuration of the (Figure 12-b) debugging does not occur. Then microcontroller 6 of outstanding contacts and 22 digital inputs /outputs.

Microcontroller digital inputs / outputs are usually grouped together (most often - by eight). Such a contact group are referred to as input / output ports. Each port generally controlled by special control register a group at each subsequent contact port control register group the bits. MSP430F123 (2) microcontroller has three ports: 6-bit port P2, 8-bit port P3 and the port P1, the debug configuration combines four, but the operating mode - eight digital inputs / outputs. Porto contacts are positioned so as to ensure the compatibility of the contact 28 and 20 pin microcontrollers (with no port P3). Therefore, all port P3 contacts are placed in the same housing end having four of each contact row (Figure 3.13).

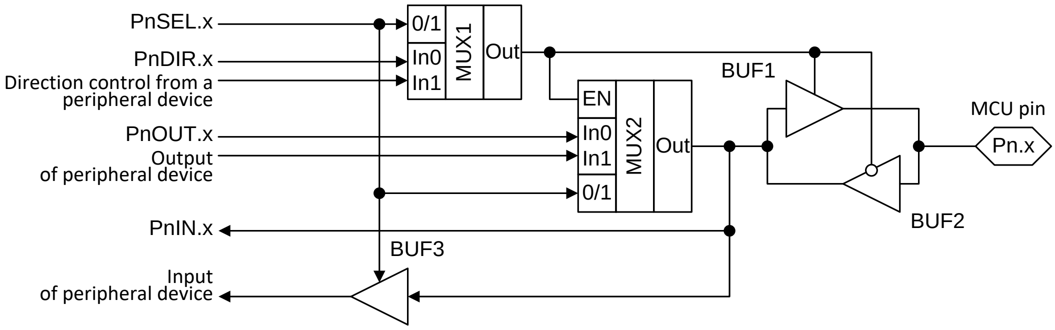


a) debugging configuration b) application configuration

Figure 3.13. Contacts of MCP430F123(2)

Pin circuitry and GPIOs control registers of MSP430

Port operations and schematic structure defined by two features: 1) port contacts can act as both the input and the output; 2) The same contacts can operate not only as a universal I / O, but also as a peripheral equipment input / output (what exactly - is usually found in a particular chip in its description). The result of each contact is connected to the circuit consisting of three multiplexers which are controlled by port control register bit (Figure 3.14).



PnSEL.x
PnDIR.x
Direction control from a peripheral device
PnOUT.x
Output of peripheral device
PnIN.x
Input of peripheral device
0/1
In0
In1
MUX1
Out
EN
In0
In1
0/1
MUX2
Out
BUF1
BUF2
BUF3
MCU pin
Pn.x

Figure 3.14. MSP430 pin circuitry and control bits

In this scheme, the crucial role played bit that defines port «n» contact «X» function - general digital input / output peripherals or input / output. It is a management registry PnSEL bit "x". If it is 0, the contact is guided from port circuit where 1 - then the peripherals. After power feeding PnSEL register bits are zero and contacts are connected to the port schemes. Schematically, the control bit PnSEL.x selects which multiplexer MUX1 and MUX2 input will be connected to the exit. This means that of the conversion scheme selection control signal. PnSEL.x bit with the value 1 activates the repeater BUF3 that feeds the signal from the contact to the peripheral equipment.

If the contact «X» are of general validity (at PnSEL.x = 0), then contact the direction determined by the Management registry PnDIR bit "x". The value "0" corresponds to the entrance, but "1" - output. If the contact "owns" peripherals (at PnSEL.x = 1), then the direction of the contact determine its peripherals. The actual direction selection signal form the contact circuit (Figure 4 2) multiplexer MUX1 via bit PnSEL.x.

The actual direction selection signal with a value of 1 activates the output multiplexer MUX2, which determines how the information is removed - from the port or peripherals. Multiplexer for input selection signal is the same control bits PnSEL.x but the switched input connected to the output peripherals and another contact management schemes bit - PnOUT.x. It follows that if the contact works as an overall output (at PnSEL.x PnDIR.x = 0 and = 1), then the voltage level of this contact (or the output information) determines the bit "x" in the register PnOUT. If, on the contrary, acts as a contact output peripherals (at PnSEL.x = 1), then the output data is determined peripherals.

Registry PnIN bits reflect the voltage level (or peak value) of the contacts, regardless of how they work (as input or output) and to which are connected (to the port or to peripherals). If the contact works as an entrance to it and the level of information entered addressing external equipment, but if it works as output register PnIN reflects the output information. If the contact is connected to the peripheral equipment that removes some information, then using this bit, you can check whether the output is adequate information that can be used for debugging purposes.

Programming of MSP430 GPIOs

Port management instructions are mostly individual bit and bit groups modifying and testing instructions. Port control registers PnSEL, PnDIR and PnOUT usually modified but register PnIN tested. It should be remembered that each port are grouped into 8 contacts, port control registers are 8-bit and they are treated with 8-bit commands.

If you need to fix the connection or direction, or acquiring the port information for all contacts (for some - so few - otherwise), then the control registers (respectively PnSEL, PnDIR and PnOUT) is modified by using the command "mov". For example, if the port P3 0 ... 3 contacts need to set the output information, the contacts 4 ... 7 - to input the command applies:

mov.b #00001111b,P3DIR

If you need to adjust only some contacts port, use the command "bis" (bit set) or "BIC" (bit clear). For example, a command that port P3 contacts 0, 1, 4 and 5 to fit the output, while the rest remain without changes is as follows:

bis.b #00110011b,&P3DIR

By contrast, commands that tunes port P3 contacts 1, 3, 5 and 7 to do the information input, but the rest of the port does not change their contacts functions, are as follows:

and.b #01010101b,&P3DIR or **bic.b #10101010b,&P3DIR**

Sometimes arises the need to invert the current bit position. This is done with the inversion command "inv" (all bits), or "xor" (individual bits). For example, commands that inverts the entire port P3 escaping information are as follows:

inv.b &P3OUT or **xor.b #11111111b,P3OUT**

If only 0th bit has to be toggled the second command will look like

xor.b #00000001b,&P3OUT

Port control registers PnIN not intended for recording. This attempt to modify the registry does not change the actual contents of the registers, but the microcontroller current in recording time is greater. So, some PnIN register bits are usually tested by an instruction "bit". Naturally, the command "bit" is used together with the commands "jz" and "jnz» of branching programs. Below an example is a program that after port P2 contact 0 reading: at 1 – inverts the output port P3 information, but at 0 - output port P3 binary code "00001111":

Code example 17. GPIO testing example.

```
         bit.b   #00000001b,     &P2IN         ;Test P2.0

         jz     AtZero

AtOne        inv.b   P3OUT                     ;Toggle P3, at 1

         ...

AtZero       mov.b   #00001111b,     &P3OUT        ;P3=00001111=0Fh at 0

         ...
```

If it becomes necessary to analyse several PnIN registry bits, then it is reasonable to use calculated branches – PnIN contents at the beginning is read, unnecessary bits deleted and the remaining bits are then analyzed as a solid piece of data. For example, if at the port P2 contacts 0 ... 2 are connected to 3 switches, they can define the microprocessor controller operating mode. In this case, in order not to analyse each port pin, the branch address can be easily calculated:

Code example 18. Calculated branch after GPIOs reading.

```
         mov.b   &P2IN,       R15          ;Read P2

         bic    #0xFFF8,      R15          ;Clear 13 MS bits

         add    R15,         R15          ;(0...7)X2=(0...14)

         add    R15,         PC           ;Branch to ...

         jmp    Mode0                     ;... mode 0

         jmp    Mode1                     ;... mode 1

         jmp    Mode2                     ;... mode 2

         jmp    Mode3                     ;... mode 3

         jmp    Mode4                     ;... mode 4

         jmp    Mode5                     ;... mode 5

         jmp    Mode6                     ;... mode 6

         jmp    Mode7                     ;... mode 7
```

```
Mode0       ...                         ;Mode0 Continued
        ...
```

Here, the at the beginning all port contacts are read with the command "mov" to register R15. 13 MS bits of this register are then deleted with the command "bic". As a result, the register contains the code from 0 to 7. The first command «add» multiply this code to 2, while the second adds calculated offset (0,2,4 ... 14) to the program counter PC, resulting in a branch to one of the instructions "JMP", which selects the appropriate mode program branch.

Watchdog timer

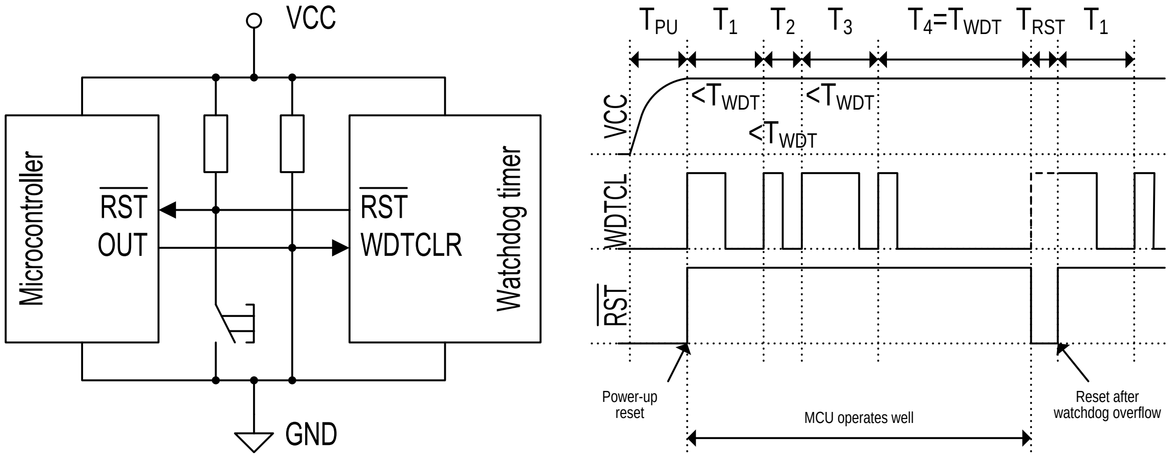
Function of watchdog

Watchdog timer a facility that performs microcontroller return in the case of the incorrect action. But the question arises, what is the microcontroller incorrect operation? In this case, the find operation that does not have the same Watchdog timer regular deletion (or return). In other words - if Watchdog timer too long is not affected, it starts to affect microcontroller operation. Of course, such a principle largely guaranteed return when microcontroller program execution for some reason stops or repeats an endless loop. However, the mentioned principle will not prevent the program error that does not stop the same microcontroller and its program.

Watchdog timer often sold as an independent external device that is associated with the microcontroller return input (RST), and a microcontroller with a digital output (Figure 3.15. - a). After power feeding and stabilization (during TPU) Watchdog timer begin to deduct define the time interval TWDT. If the lapse of time before this microcontroller has time to change the digital output (usually - a certain way), then the time interval TWDT counting begins anew. If the digital output to use and not Watchdog timer TWDT deduct the time, then it generates a specific length TRST microcontroller return pulse. Show a example (Figure 3.15.-b) Microcontrollers 3 times suggests time TWDT recount, providing Watchdog timer extinguishing signal WDTCL growth. However, the fourth time it is not able to do it and after a while TWDT Watchdog timer microcontroller generates a return pulse TRST.

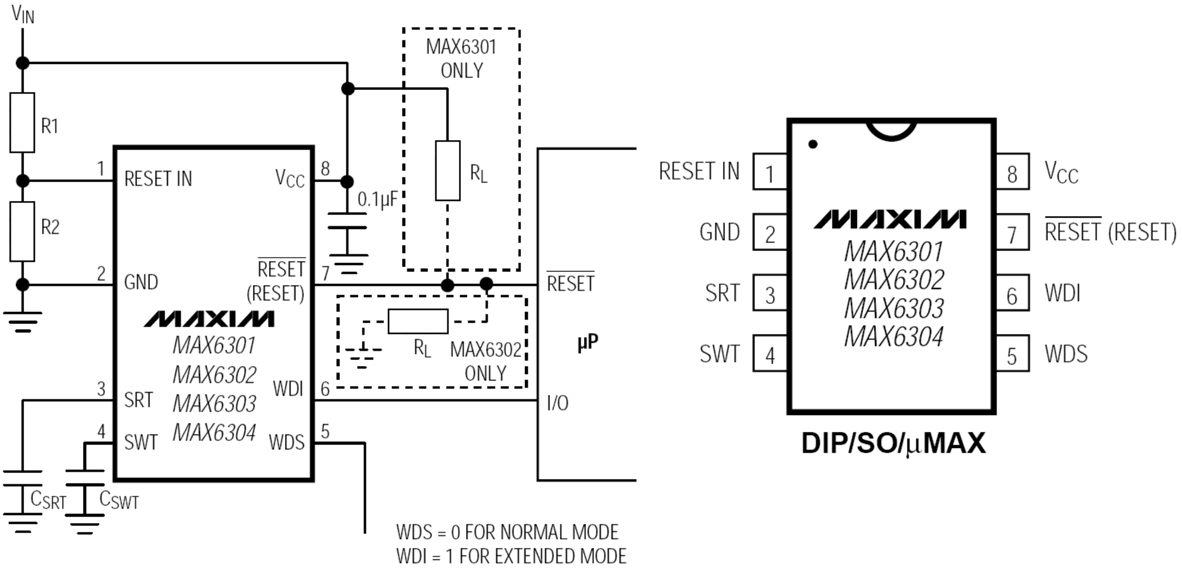
Watchdog chips often have extra features. For example, the return of a decrease in supply voltage or a special entrance to the push-button connection.

Watchdog chips as an example may be mentioned MAX6301, MAX6302, MAX6303 and MAX6304 chip "Maxim-Dallas' branded product (Figure 3.16.). As can be seen, in general, the connections comply with the above mentioned. Chips are also input mode for setting, as well as specific output voltage return for asking. The same contact can be used to return the push-button connection.



a) connections b) operation

Figure 3.15. Watchdog function



a) connections b) pinout

Figure 3.16. Example of watchdog IC.

Features of MSP430 watchdog timer

Microcontroller manufacturers often watchdog timer built into the microcontroller. For example, the MSP430 microcontroller watchdog timer is implemented as a peripheral device with programmatic access.

MSP430 microcontroller built watchdog timer is a special clock pulse counter that the event generates Overflow microcontroller return signal. Therefore, if watchdog timer is started, it is periodically erased before it rolls over to Microcontrollers reset. These characteristics are used to increase the stability of the software. In certain areas of the program are placed in watchdog timer deletion commands. If the program is working properly, then watchdog timer deletion commands are triggered and overwhelm. If the program works incorrectly, then delete the team will not be triggered, watchdog timer rolls over, but the microcontroller is reset. MSP430 watchdog timer has the following features:

- Watchdog timer based on the counter, the contents of which can not read or write directly, but can be deleted; After microcontroller return (and power feeding) it is activated;
- The program can select 4 thresholds (numbers);
- The program can choose 2 counting signals;
- There are two modes of operation:
    - microcontroller return mode in which watchdog timer after tripping threshold being reached (overfilling) forms return signal;
    - universal timer / counter mode when overflow case simply set a specific bit WDTIFG;
- watchdog timer management mainly takes place through the 16-bit control register, the recording is protected by a special code (password).

watchdog timer (Figure 3.17) The main components are: a 16-bit counter WDTCNT counting signal (or watchdog timer clock signal) Multiplexer MUX1 and meter output multiplexer MUX2. watchdog timer scheme installed special watchdog timer WDTIFG flag, as well as affect the microcontroller return signal generating circuit. watchdog timer control scheme watchdog timer management registry WDTCTL bits.

If at watchdog timer meter is connected to a pulse signal, it will increase. But counted pulse value can not be read (or write) and programmable use. The only chance to directly influence the content of the counter is to delete the modal logic 1 to the redemption entrance R which, in turn, is controlled by the control registry WDTCTL bit WDTCNTCL and the installed by program. It should be noted that bit WDTCNTCL remain installed on a single microprocessor clock period, which determines its short-term impact and provide their own watchdog timer effect (otherwise watchdog timer are removed continuously).

Multiplexer MUX1 is connected to the meter watchdog timer one of the microcontroller clock signals SMCLK or ACLK. The Multiplexer input selection provides control bit WDTSSEL. Immediately after the return is 0, which is determined by the meter input signal SMCLK with a default frequency of 750kHz (with corresponding pulse counting period 1.33mks). In turn, the signal ACLK after the return has not been activated (as it can not be activated if the microcontroller is connected to the quartz resonator), and this signal can not be readily used watchdog timer from hunting (hence the signal after the return is not used). Multiplexer MUX1 is the activation input EN, which is controlled by the control bit WDTHOLD. If WDTHOLD = 1, then the multiplexer MUX1 is blocked counter WDTCNT does not count but watchdog timer generally your return function is failing.

The second multiplexer MUX2 selects one of the meter outputs Q15, Q13, Q9 or Q6 and further added to the return / break signal generation scheme. Multiplexer control signals WDTIS0 and WDTIS1 after the return is 0, which determines watchdog timer overflow threshold 215 = 32,768th If this is achieved, then the flag is set and WDTIFG (return mode) is generated microcontroller return signal. Return signal activation and installation watchdog timer

WDTIFG bit counter Overflow case is designed especially for watchdog timer circuit elements - the pulse generator and a trigger.

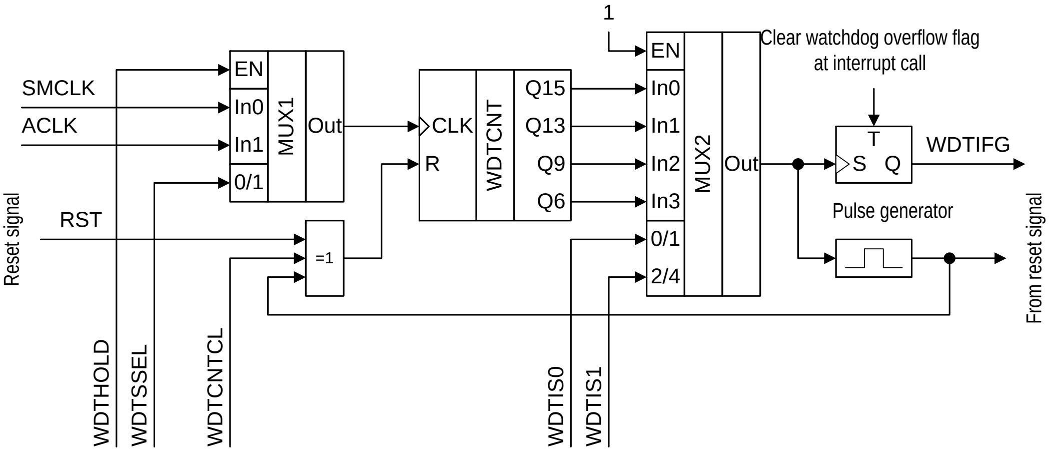


Figure 3.17. Functional diagram of MSP430 watchdog timer

Multiplexers MUX1 and MUX2 together determine watchdog timer overfilling time. Immediately after returning microcontroller or power feeding when the default counting signal is SMCLK with a default frequency of 750kHz, but the default threshold of 32768, counting period constitutes 32768 × 1.33mks = 44ms.

Another important part is watchdog timer control register WDTCTL. It is a 16-bit data and is composed of parts (the latest byte) and password parts (senior byte). Password part is connected to a comparator that compares the recorded 16-bit word senior 8 bits password 01011010b (5Ah). If the name of the oldest byte password coincides with the latest byte is entered in the register of the data part. Recording with an incorrect password is equivalent meter overfilling. Watchdog timer recording password programs are usually written as WDTPW that actually 0x5A00 - password entered with a zero byte oldest part of the latest (or simply multiplied by 256). It should be remembered that the password can only be recorded. Password-bit reading actually read test number 69h (or full form 6900h).

Control register of MSP430 watchdog timer

The structure of MSP430 watchdog timer control register is shown in Figure 3.18.

15 ... 8	7	6	5	4	3	2	1 0
WDTPW	WDTHOLD	WDTNMIES	WDTNMI	WDTTMSEL	WDTCNTCL	WDTTSSSEL	WDTISx
r-69h w-5Ah	rw-0	rw-0	rw-0	rw-0	r-0(w)	rw-0	rw-00

Figure 3.18. Control bits of MSP430 watchdog timer

The meaning of these bits is following:

WDTPW – Read as 0x69, Must be written as 0x5A – i.e. must be written by “mov.w”, Otherwise WDT behaves as at overflow;

WDTTMSEL – 0 selects watchdog mode, 1 selects interval timer mode;

WDTIS… – Select threshold (00 select 2^{15}, 01 – 2^{13}, 10 – 2^{9}, 11 – 2^{6});

WDTSSEL – Selects clock (WDTSSEL=0 selects SMCLK, 1 – ACLK);

WDTCNTCL – 1 clears the counter (reset automatically);

WDTHOLD – 1 stops counting (disables the counter).

Table XII. Overflow periods of MSP430 watchdog counter (at f_{SMCLK}=750kHz, f_{ACLK}=32768Hz)

WDTSSEL	**WDTIS1**	**WDTIS0**	Counting threshold	Counting (clock) signal	Overflow period [ms]
0	0	0	2^{15}=32768	SMCLK	$t_{SMCLK}\times2^{15}$=44ms
0	0	1	2^{13}=8192	SMCLK	$t_{SMCLK}\times2^{13}$=11ms
0	1	0	2^{9}=512	SMCLK	$t_{SMCLK}\times2^{9}$=683μs
0	1	1	2^{6}=64	SMCLK	$t_{SMCLK}\times2^{6}$=85μs
1	0	0	2^{15}=32768	ACLK	$t_{SMCLK}\times2^{15}$=1s
1	0	1	2^{13}=8192	ACLK	$t_{SMCLK}\times2^{13}$=0.25s
1	1	0	2^{9}=512	ACLK	$t_{SMCLK}\times2^{9}$=16ms
1	1	1	2^{6}=64	ACLK	$t_{SMCLK}\times2^{6}$=1.9ms

Watchdog flag indicating its overflow is located in a generic flag register IFG1 (Figure 3.19).

7	6	5	4	3	2	1	0
			NMIIFG				WDTIFG
			rw-0				rw-0

Figure 3.19. Control bits of MSP430 watchdog timer

Programming of MSP430 watchdog timer

Once programmed in a timer mode with instructions like (for 2^6 and ACLK)

```
mov #WDTPW+WDTIS1+WDTIS0+WDTSSEL,&WDTCTL
```

the watchdog counter can be stopped by

```
mov #WDTPW+WDTHOLD,&WDTCTL
```

or periodically cleared by

```
mov #WDTPW+WDTCNTCL,&WDTCTL
```

More complicated Code example 19 of an LED blinking (attached to P3.0) shows how MSP430 watchdog can be used to count time intervals.

Code example 19. Use of MSP430 watchdog counter to count time.

```
main        mov    #WDTPW+WDTTMSEL,&WDTCTL       ;Start WDT as timer
            mov.b  #00000001b,    &P3DIR        ;P3.0 - output
;-------------------------------------------------------------------------
Rpt_forever mov    #5,          R15          ;Setup WDT cycle counter
Rpt_040ms   bic.b  #WDTIFG,      &IFG1        ;Clear WDTIFG
Rpt_New     bit.b  #WDTIFG,      &IFG1        ;Test WDTIFG
            jz     Rpt_New                    ;Repeat until WDTIFG=1
;-------------------------------------------------------------------------
            dec    R15                        ;Decrement WDT counter
            jnz    Rpt_040ms                  ;Repeat untill 0
;-------------------------------------------------------------------------
            inv.b  &P3OUT                     ;Toggle P3.x
            jmp    Rpt_forever                ;Repeat forever
```

Basic clock system of MSP430

The structure of MSP430 clock system is given in Figure 3.20. It supports 3 main clock signals (MCLK for CPU, SMCLK for peripherals and ACLK also for peripherals), 4 clock sources (digitally controlled generator DCOCLK, very low power generator VLOCLK 10kHz, 2 generators with crystal – LFXT1CLK and XT2CLK). The clock system is flexible, programmable in-system and on-the-go. By default DCO generator is tuned for f_{DCO}=750kHz, but f_{MCLK}= f_{DCO}= and f_{SMCLK}= f_{DCO}. The default frequency can be easy modified by means of clock control registers DCOCTL and BCSCTL1 (3 LS bits).

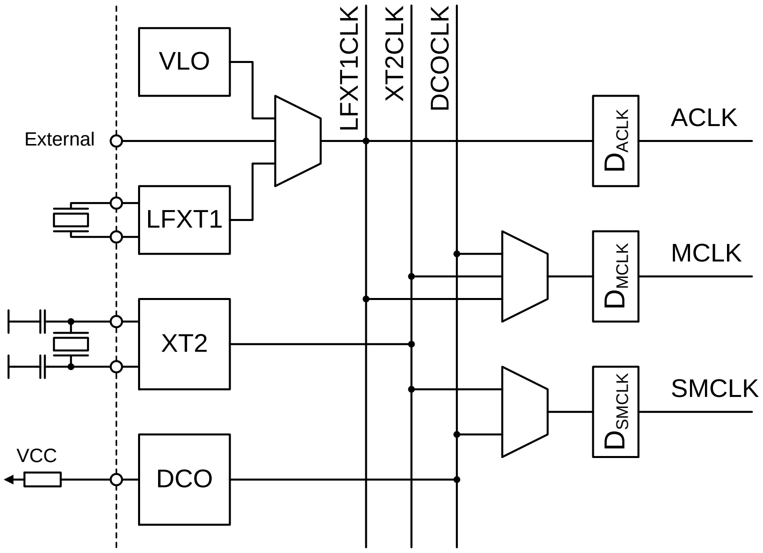


Figure 3.20. MSP430 clock system

Chapter 4:Electrical engineering and automation

4.1. Principles and applications

4.1.1. Resonance phenomenon in AC electrical circuits

Resonance of Voltage

In a complex scheme with R, L, C connected in series the situation of equality of X_C and X_L could occur, that is

$$\omega L = {}^{1}/_{\omega C} \tag{4.1}$$

or

$$\omega = \omega_0 = \sqrt{{}^{1}/_{LC}}. \tag{4.2}$$

This situation could be possible in two cases :

1) when L and C are unchangeable but the input frequency varies (fig.4.1);
2) with the invariable input frequency but when meanings of L and C are varied.

Independently on a method of situation of realisation, when vectors U_L and U_C are opposite in direction and equal to each other (fig.4.1), total reactive voltage appears equal to zero and the current in the circuit is limited only with the resistor:

$$I_R = \frac{U_1}{R}. \tag{4.3}$$

Undoubtedly, the current reaches the maximum possible value for this circuit (fig.4.1).

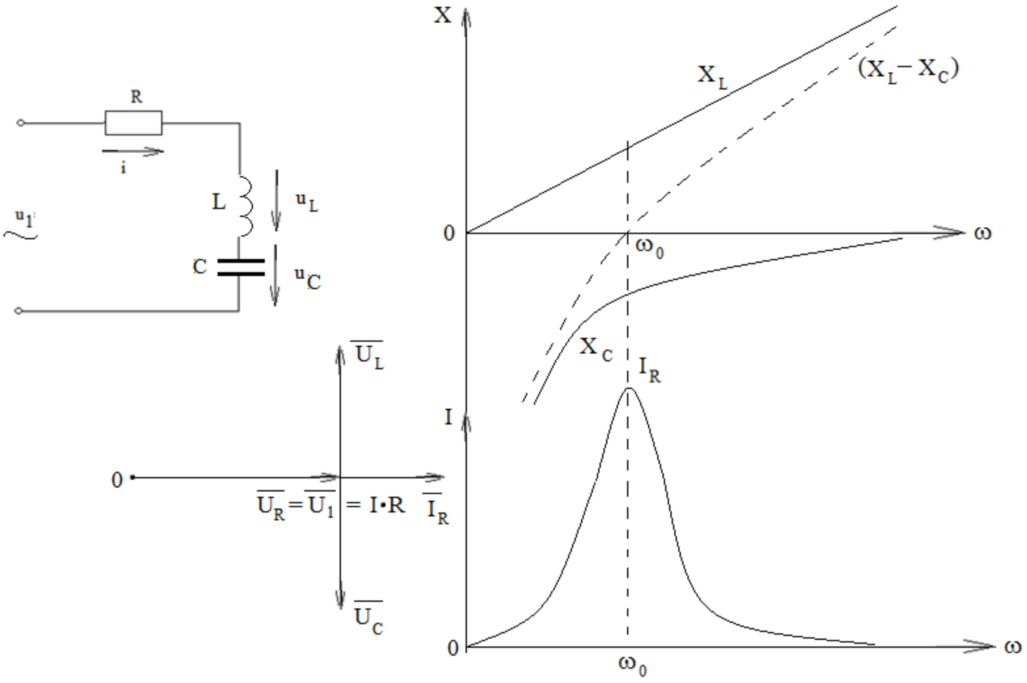


Figure 4.1. Diagrams for voltage resonance

At the resonance the current is in phase to the input voltage and cosφ=1. The total power of the circuit is

$$S = P = I^2 \cdot R\,. \tag{4.4}$$

Nevertheless, the voltages across the reactive elements could reach very high values

$$U_L = U_C = \frac{U_1 \cdot \omega_0 \cdot L}{R} = \frac{U_1}{R \cdot \omega_0 \cdot C} \quad . \tag{4.5}$$

For example, if $\frac{(\omega_0 \cdot L)}{R} = 10$, then the voltage across the reactor is ten times of the input voltage. The ratio of $(\omega_0 \cdot L)/R$ or $1/(R \cdot \omega_0 \cdot C)$ characterises the resonant effect.

Resonance of Currents

Resonance of currents could appear in a parallel circuit of reactor L and capacitor C (fig.4.2).

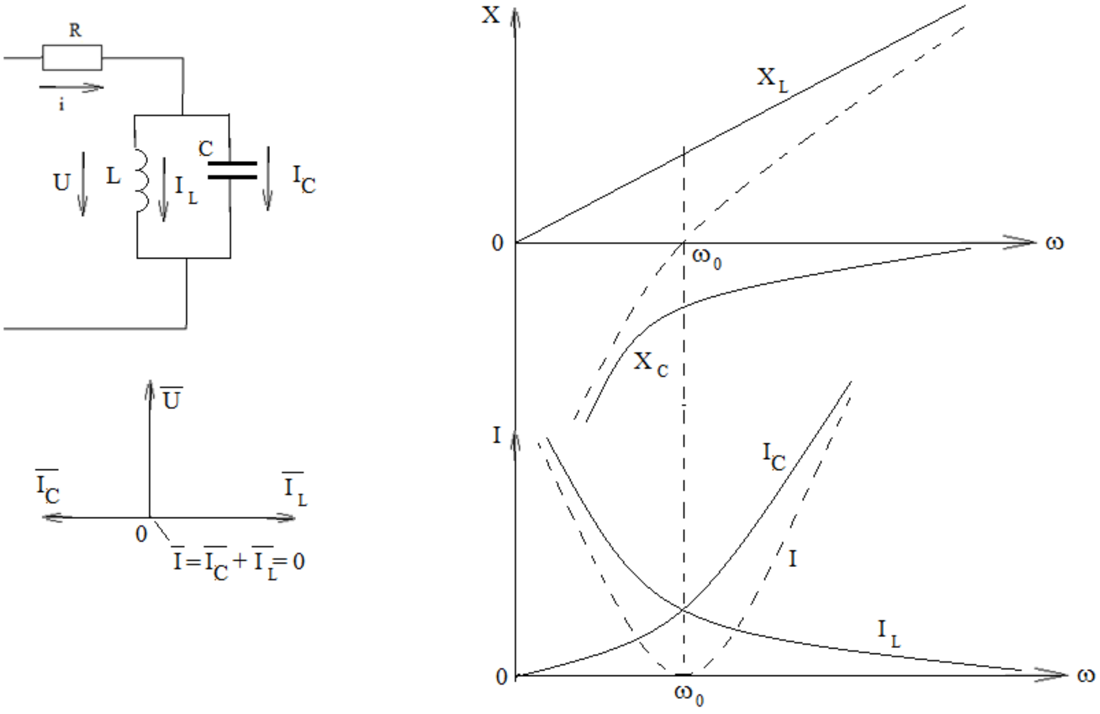


Figure 4.2. Diagrams for resonance of currents

The total current of the circuit I is defined as a vector sum of the current of capacitor I_C and the reactor I_L, but they are opposite directed to the total input voltage U (fig.4.2). With the equality of both currents, when their impedances are the same, currents $I_L=I_C>0$, but their summative current is equal to zero. The total impedance in this case is infinitely large. Like for the case of resonance of voltage

$$X_L = \omega \cdot L = X_C = {}^1/_{\omega C}$$

or

$$\omega = \omega_0 = \sqrt{{}^1/_{LC}} \tag{4.6}.$$

4.1.2. Single-Phase Transformer

Transformer changes the sine-form magnetic flux created by current of one winding into sine-form alternating voltage of other winding or windings. To realise this transformation a common magnetic core should be turned with two windings with w_1 and w_2 number of turns (fig.4.3, a).

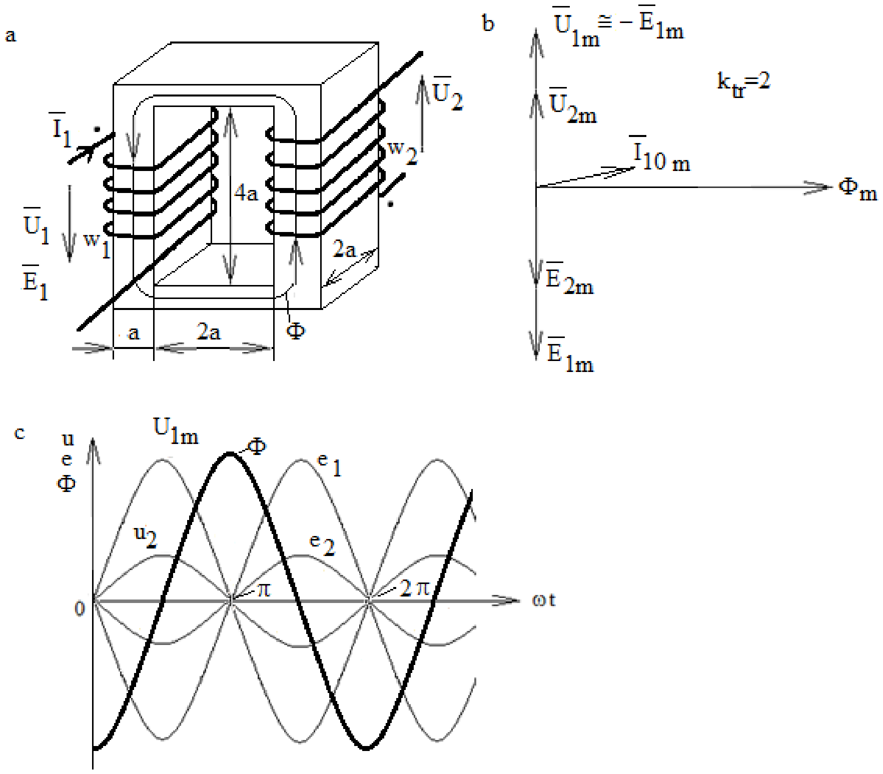


Figure 4.3. Single-phase transformer and its diagrams

If both windings of transformer are turned in the same direction (in fig.4.3,a the windings in respect to the initial clamps marked with the point are turned clockwise) then positive directions of voltage and current of the windings could be shown in the same way. Starting clamp of each winding is nominated with a point and the points note on the situation when currents in both windings are directed in the same way and magnetic fluxes in both windings then also are coinciding in direction.

If the directions of EMF and voltage are the same then

$$u_1 = -e_1 = w_1 \frac{d\Phi}{dt} = U_1 \sin\omega t. \tag{4.7}$$

Therefore, the changes of the magnetic flux are calculated as

$$\Phi = \frac{U_{1m}}{w_1} \int \sin\omega t dt = -\frac{U_{1m}}{w_1 \cdot \omega} \cos\omega t = -\Phi_m \cos\omega t. \tag{4.8}$$

As we can see (fig.4.3,b) the vector of magnetic flux lags behind the vector U_{1m} by 90^0 in phase. When the sine-form magnetic flux goes through the core and crosses the loop of winding w_2, EMF is induced in the latter

$$e_2 = -w_2 \frac{d\Phi}{dt} = -w_2 \cdot \omega \cdot \Phi_m \cdot \sin\omega t = -E_{2m}\sin\omega t, \qquad (4.9)$$

and voltage u_2 is equal to $u_2 = -e_2 = U_{2m}\sin\omega t.$ (4.10)

As we can see from the expressions, the vectors of the voltages of the primary and secondary windings are in phase, but those of both EMF are in counter-phase to the vectors of voltage (fig.4.3, b). The maximum value of the flux is

$$\Phi_m = \frac{U_{1m}}{w_1 \cdot \omega}. \qquad (4.11)$$

Substituting this flux expression into that of EMF e_2 and voltage u_2 we get

$$U_{2m} = w_2 \cdot \omega \cdot \frac{U_{1m}}{w_1 \cdot \omega} = \frac{w_2}{w_1} U_{1m}. \qquad (4.12)$$

The relationship of the two windings values is the following

$$\frac{U_1}{U_2} = \frac{w_1}{w_2} = k_{tr} \ , \qquad (4.13)$$

where k_{tr} is a factor of transformation.

The following could be defined from the equation of magnetic flux: $\sqrt{2} \cdot U_1 = w_1 \cdot \omega \cdot \Phi_m$.

Taking into account that $\omega = 2\pi \cdot f, \Phi_m = B_m \cdot s$, the following could be obtained:

$$U_1 = 4.44 \cdot w_1 \cdot f \cdot B_m \cdot s. \qquad (4.14)$$

The constructive parameters of the transformer are derived from (4.11): the number of turns of the primary winding w_1 and cross-section area of the iron core s.

The operation mode shown in fig.4.3 is called a mode of open-circuit. In this case I_2=0 and E_2=U_2 , there is a low current I_{10} in the primary winding. If the active resistance of winding w_1 is equal to zero, vector I_{10} is in phase with the vector of the flux. But some losses of power appear in the resistance R_1 therefore I_{1m} lags behind U_{1m} in phase by angle less than 90^0 (fig.4.3, b).

Having connected an active inductive load to the secondary winding, EMF e_2 induces current

$$I_2 = -\frac{E_{2m}}{Z} \sin(\omega t - \varphi), \qquad (4.15)$$

where Z – is the total impedance of the secondary circuit of the transformer, φ is the phase shift between vectors I_2 and E_2 (fig.4.4).

At these circumstances the current I_1 is increasing because MMF of the primary winding should compensate the influence of the current I_{10} and I_2:

$$\overline{I_1} \cdot w_1 = \overline{I_{10}} \cdot w_1 + (-\overline{I_2} \cdot w_2). \qquad (4.16)$$

The following should be taken into account as with a rated load I_{10} is 5% only of I_{1R}, and therefore

$$I_1 \cdot w_1 = I_2 \cdot w_2$$

or

$$\frac{I_1}{I_2} = \frac{w_2}{w_2} = \frac{1}{k_{tr}}. \tag{4.17}$$

Therefore with a load

$$U_1 \cdot I_1 \approx U_2 \cdot I_2, \tag{4.18}$$

i.e. the apparent powers of the primary and secondary windings are approximately equal to each other, that means the efficiency factor of the transformer is close to 1.

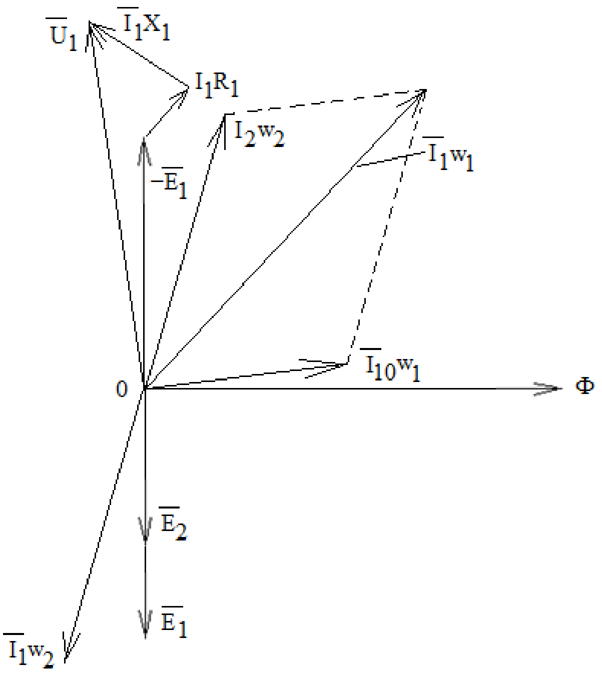


Figure 4.4. Phase diagram for unloaded single-phase transformer

If the current I2 is corrected so that $I_2' \cdot w_1 = I_2 \cdot w_2$, where I_2' is the current of the load reduced to that of the primary winding, then the balance equation of the currents should be represented in one substitution scheme (fig.4.5). The related parameters are marked with upper index “ ’ ”.

The following equations are valid for the substitution scheme (fig.4.5):

$$\begin{gathered} \overline{U_1} = -\overline{E_1} + \overline{I_1} \cdot R_1 + \overline{I_1} \cdot X_1 ; \\ \overline{-E_2'} = \overline{U_2'} + \left(\overline{-I_2'} \cdot R_2'\right) + \left(\overline{-I_2'} \cdot X_2'\right). \end{gathered} \tag{4.19}$$

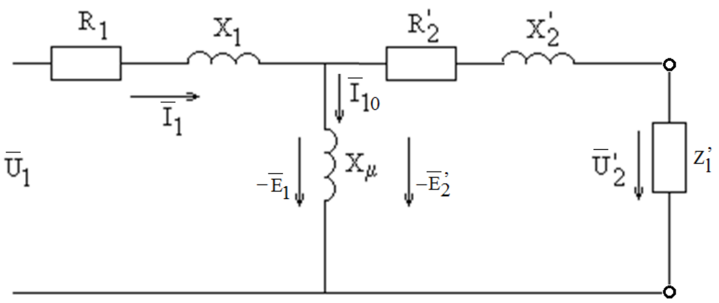


Figure 4.5. Substitution scheme of a single-phase transformer

The impedance of the secondary winding is defined as

$$R'_2 = R_2 \cdot k_{tr}^2 ;$$

$$X'_2 = X_2 \cdot k_{tr}^2. \qquad (4.20)$$

Using the expression $U_1 = 4.44 \cdot w_1 \cdot f_1 \cdot B_m \cdot s$ as well as the geometric parameters of the core in fig.4.3, the main parameters of the transformer are calculated as

$$S = 2 \cdot a^2,$$

$$w_1 = \frac{8 \cdot a^2 \cdot 0.5 \cdot 0.35 \cdot j}{I_1},$$

$$a = \sqrt[4]{\frac{U_1 \cdot I_1}{4.44 \cdot 8 \cdot j \cdot f \cdot B_m}}, \qquad (4.21)$$

where the meaning of B_m is taken equal to 1.1…1.2 T, and $j \approx 2 \cdot 10^6 \ \mathrm{A/m^2}$.

4.1.3. Basic realisation of electrical motor

The operation of any electric motor is based on the electromagnetic induction phenomenon - interaction of a conductor with electric current and magnetic field. Three conceptions of the motor realisation are possible:

1 – when the current in the conductor influences the flux induced by a permanent magnet (or electroamgnet) or opposite transformation;

2 – when the current of the conductor influences a secondary magnetic flux induced by the current itself;

3 – when the magnetic flux induced by the current influences a ferromagnetic body (so called reluctance principle).

The realisation of the first conception is shown in fig.4.6. Two separate contact rings a and b are set on a rotating axis (a shaft) of the motor. These rings are supplied with the current through unmovable coal brushes. The rings are connected with the frame of conductors which has two edges 1-2 and 3-4 located in parallel with the shaft normally to the magnetic lines.

In the position in fig.4.6,a the direction of the current in the edge 1-2, in accordance with the law of left hand provides the movement of this edge to the left. The current of the other edge 3-

4 moves it to the right, i.e. the motive force F counter clockwise movement appears in this case. The force is maximum in its value when the frame is in vertical position

$$F_m = I \cdot l \cdot B \cdot w, \qquad N, \tag{4.22}$$

where I is the current of the frame with w number of turns, l – length of the active edges of the frame, B – is the flux density, T. The product of the force and radius of the frame is called an electromagnetic torque

$$M_{em} = F \cdot R \quad , \qquad Nm. \tag{4.23}$$

To continue the counter clockwise movement of the frame the current of it should change its direction to the opposite when the both edges pass 90^0 from the vertical position (fig.4.6, b). It means that the contact rings a-b should be supplied with the voltage of the opposite polarity than that in case “a”, thus the rings should be supplied with an alternating voltage.

The angular velocity of the frame is in accordance with the frequency of the voltage supplying the frame:

$$\omega = 2\pi \cdot f. \tag{4.24}$$

While turning the frame the following power is induced in the motor

$$P_{em} = M_{em} \cdot \omega \, , \qquad W \tag{4.25}$$

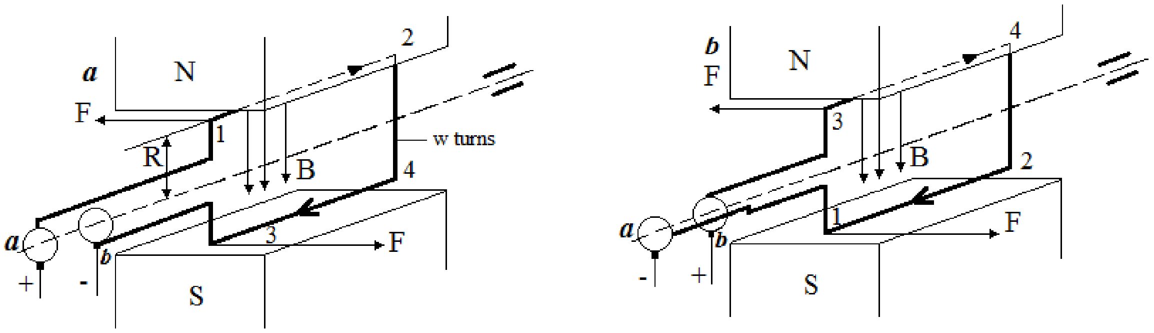


Figure 4.6. Motor realisation in accordance with the first conception

Power losses of the rotor should be subtracted from this power P_{em} , thus we can get the useful (active) power P. The supplying power of the source is

$$P_1 = U_1 \cdot I_1 \cdot \cos\varphi, \tag{4.26}$$

which is larger than P_{em} and P. The ratio of the useful (active) power to the supply one is an efficiency factor

$$\eta = {}^{P}/_{P_1}. \tag{4.27}$$

Therefore, the current supplied from the source is

$$I_1 = \frac{P}{U_1 \cdot \cos\varphi \cdot \eta} \quad . \tag{4.28}$$

The construction of the motor under consideration has only one pair of poles. The motor with four poles (two pairs of poles) is also possible. These poles are located in the order N-S-N-S.

Thus two periods of the supplying voltage will take place during the period of one full turning of the frame and in general case

$$\omega = \frac{2\pi \cdot f}{p}, \qquad (4.29)$$

where p – is the number of pairs of poles.

The number of revolutions per minute is calculated as:

$$n = \frac{60 \cdot \omega}{2\pi} = \frac{60 \cdot f}{p}. \qquad \text{RMP} \qquad (4.30)$$

With the aim to change the direction of the frame's rotation the polarity of the current at the positions a and b should be opposite. In practice it means we must have a sensor which detects the upper position of the edge 1-2 then provides supplying it with the positive or negative pole, depending on the required direction of rotation. This concept is applied in motors with commutation (fig.4.7) when the source of direct current could be connected to the upper position of the edge 1-2 with the help of the contactor S1 (to rotate the shaft in the counter clockwise direction) or S2 (in the opposite direction).

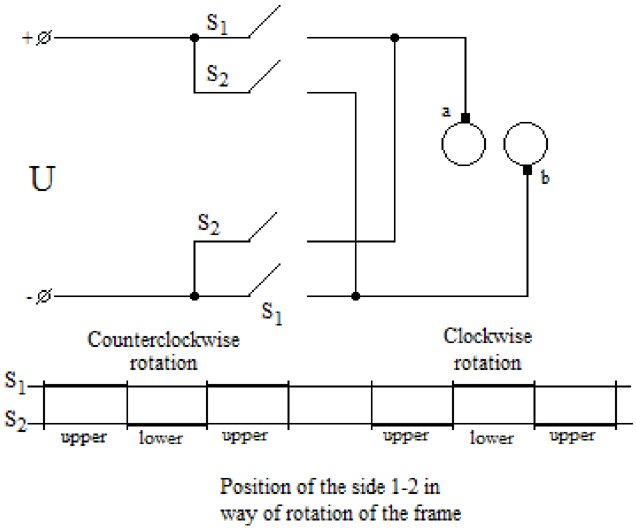


Figure 4.7. Control of the brushless DC motor with commutation

In practice this type of motors in the case of supplying with an alternating current or commutated direct current is realised with a magnet, fixed on the rotor, and stator with windings. Synchronous AC motor or so called brushless DC motor are obtained in this way. Synchronous AC motor is working without position sensors but brushless DC is operating with a commutation controlled by position sensor. When a synchronous conception is realised some problems at the initial rotation (starting) exist. These problems are connected with AC magnetic flux direction changing twice in one cycle at unmovable magnets of rotor. For a successful starting of the synchronous motor an auxiliary initial acceleration of the rotor must be realised.

Nevertheless in the case of direct current the commutation could be obtained dividing the contact ring so that the edge 1-2 is connected to the side "a" of the ring (fig.4.8), but the side 3-4 – to the side "b" of the ring. But the direct current in its turn is supplied to the rings through two unmovable brushes.

In such way the DC motors were developed using the divided contact rings with a pair of brushes which represent a sensor of position as well as a commutation of direct current. Independently on its construction the rotating frame crosses the magnetic lines and, therefore, EMF is induced in the windings with the maximum value

$$E_m = 2 \cdot w \cdot B \cdot l \cdot V \quad , \tag{4.31}$$

where V is a tangential velocity

$$V = \omega \cdot R = \frac{2\pi \cdot f \cdot R}{p} . \tag{4.32}$$

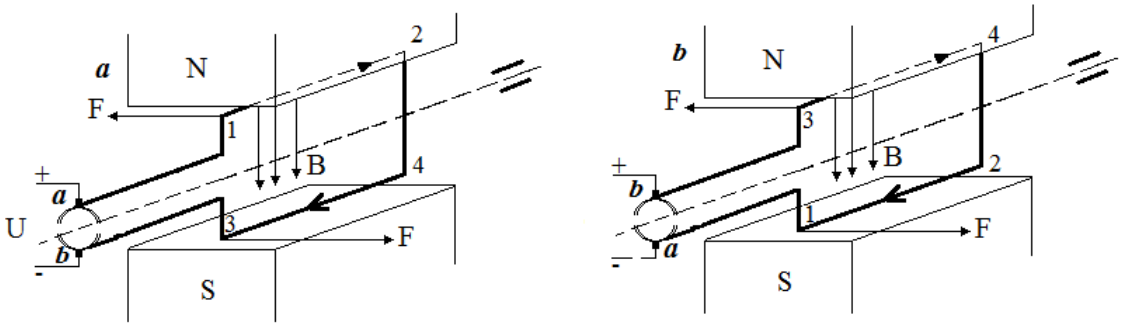


Figure 4.8. Realisation of the DC electric motor

The Kirchhoff equation for the frame is the following:

$$\overline{U} = \overline{E} + \overline{I} \cdot R + \overline{I} \cdot X \quad . \tag{4.33}$$

In the case of commutation of direct current X=0 and

$$U = E + I \cdot R \quad , \tag{4.34}$$

where R is a resistance of a rotating frame. In DC motors

$$E = c \cdot \Phi \cdot n \quad , \tag{4.35}$$

where c is a constant depending on the number of pair of poles, number of turns and a type of realisation, Φ is magnetic flux of the poles.

Therefore, the rotation velocity of the direct current motor is

$$n = \frac{U - I \cdot R}{c\Phi} . \quad \text{RPM} \tag{4.36}$$

To realise the second concept of the motors the easiest way is to introduce two normally located to each other stator frames (fig.4.9). The rotor here is a metallic solid cage without insulated winding.

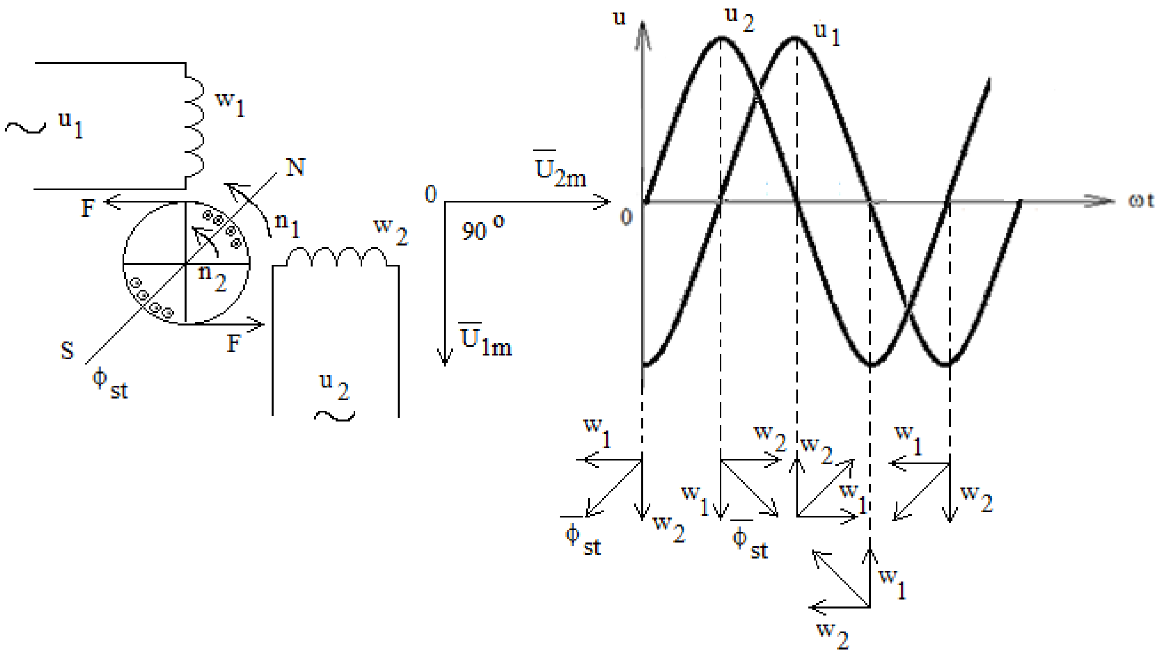


Figure 4.9. Realisation of AC induction motor

If the windings w_1 and w_2 are supplied from two separate single-phase sources of alternating current with voltages u_1 and u_2 with 90^0 shift angle between their vectors, then with zero angle the winding w_2 creates the vector of magnetic flux of the stator turned to the right. Otherwise, when the voltage u_2 of the winding w_2 is positive the downward vector Φ of the flux is created. When u_2 becomes negative vector Φ turns to the left, etc., thus the rotating magnetic field of the stator is induced, crossing the rotor. At the initial position an EMF is induced in the rotor in accordance with the right hand rule, this EMF induces current in the direction shown in fig.4.7.

Otherwise, in accordance with the left hand rule the result of the current of the rotor and the rotating magnetic field of the stator is a tangential force F, which induces a torque rotating in the same direction with the stator magnetic field. Therefore the rotating velocity of the magnetic field of the stator is

$$n_1 = \frac{60 \cdot f}{p} \ , \qquad (4.37)$$

where f is the frequency of supplying voltage, p – the number of the poles pairs of the stator.

If the velocities n_1 and n_2 are equal to each other the crossing of the rotor by the rotating magnetic field will not take place and the torque will be equal to zero. The relative difference of

$$s = \frac{n_1 - n_2}{n_1} \ . \qquad (4.38)$$

This conception of the motor is called the double-phase induction motor of double-phase asynchronous short-cage motor.

The direction of stator rotation could be changed with the changing places of the both of leads of one winding relatively to the terminals of the supply source (e.g. winding w_2). Thus if vector

u_2 lags behind u_1 by 90^0 it means the magnetic field of the stator rotates in the opposite direction to that defined in fig.4.9.

Electric power of the motor is

$$P_1 = 2 \cdot U \cdot I \cdot \cos\varphi \,, \tag{4.39}$$

where U, I cosφ are the RMS values of current and voltage and the power factor of each winding.

The current of the winding lags behind by 90^0 and, if one of the windings is supplied directly but the second through a phase supplying unit, the total current is

$$I_k = I\sqrt{2} \,, \tag{4.40}$$

where I is RMS value of the winding current.

The third conception is based on the attraction of ferromagnetic body placed in a magnetic field to the state of minimum magnetic reluctance. The rotor in this case is a kind of cross-form made of some type of iron (fig.4.10), but the stator is carried out as a system of three pair of poles. If voltage is supplied to the pole A the rotor remains unmovable. If pole A is disconnected and pole B is connected to the supply the rotor turns for 30^0 clockwise. Further disconnecting B and connecting C the rotor turns for 30^0 more, etc., i.e. the order of the poles A, C, B, A, C….. in its turn gives the opposite direction. This kind of motor is called reluctance motors, i.e. magnetically controlled modulator.

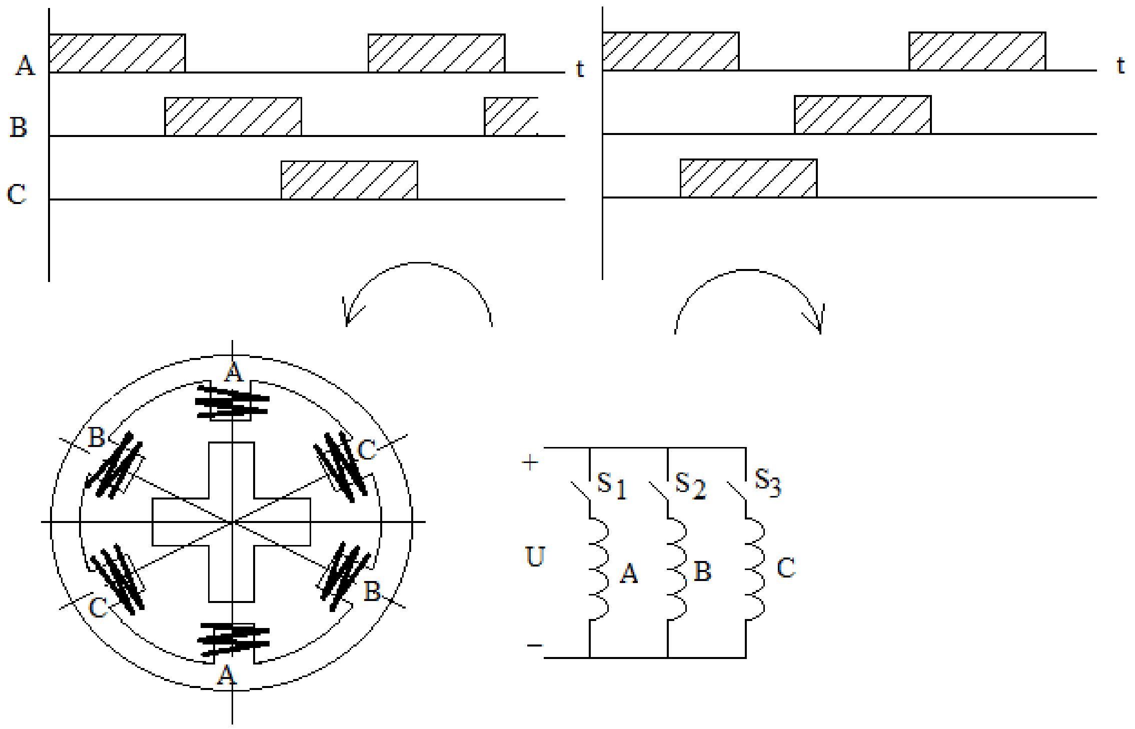


Figure 4.10. Realisation of the reluctance motor

4.1.4. Measurements of active power and energy in AC circuits

Active power in AC circuit can be measured by wattmeter containing two coils. The voltage coil of wattmeter has a large resistance and the current of the coil is coinciding in phase with AC voltage:

$$\mathrm{i_u = \frac{U_m}{R} sin\omega t} \ . \qquad (4.41)$$

In its turn a magnetic field created by the current of the series coil is:

$$\mathrm{B = k_l \cdot I_m \cdot sin(\omega t - \varphi)} \ , \qquad (4.42)$$

where φ is shifting angle between voltage and current of the estimated AC circuit.

A connected with pointer movable coil of wattmeter is turned by the force in accordance with the left hand rule:

$$\mathrm{F = i_u \cdot B \cdot l = l\frac{U_m}{R} \cdot k_l \cdot I_m \cdot sin\omega t \cdot sin(\omega t - \varphi) = k_l \cdot U \cdot I \cdot [cos\varphi - cos(2\omega t - \varphi)]}.(4.43)$$

But the inertial voltage coil of the wattmeter can not move with the fast variations of the double AC frequency component of the force. Because of this reason the power measured with wattmeter is proportional to the first component of the force:

$$\mathrm{F = k_l \cdot U \cdot I \cdot cos\varphi} \ , \qquad (4.44)$$

i.e. wattmeter measures active power

$$\mathrm{P = U \cdot I \cdot cos\varphi} \ . \qquad (4.45)$$

Measurement device for the monitoring of energy can be made on the base of electrical motor action principle. Horizontal metallic disc1 with vertical rotation axis (fig.4.11) is placed in a gap of vertical core 2 of the coil of AC voltage. Created by the supply voltage u the current in the winding of this coil in its turn creates magnetic flux Φ_u of sine form which induces EMF and a corresponding current in the metallic disc.

The disc will be unmovable unless the current of the measured circuit passes though another coil 3 upon the horse-shoe magnetic core. The situation with a disconnected current exists because coil 2 can not create a rotational magnetic field but pulsating only, i.e. disc can not get a rotational torque.

If the current of the circuit is passing through coil 3 then the created by current magnetic flux Φ_1 interacts with the current of the disc I_2 and in accordance with the left hand rule a rotational torque appears. Disc starts rotation and the number of revolutions per time unit denominates the amount of measured energy. Similar to the case of wattmeter the rotation speed of the disc is proportional to the active power of the measured circuit:

$$\mathrm{n \equiv I_2 \cdot \Phi_1 \cdot cos\varphi \equiv U \cdot I \cdot cos\varphi} \ . \qquad (4.46)$$

If n is measured as a number of revolutions within the time duration t, then the total number N of the revolutions within a certain time interval t_n corresponds to the correspondent energy:

$$\mathrm{N \equiv A \equiv U \cdot I \cdot cos\varphi = P \cdot t_n} \text{ , Ws=J.} \qquad (4.47)$$

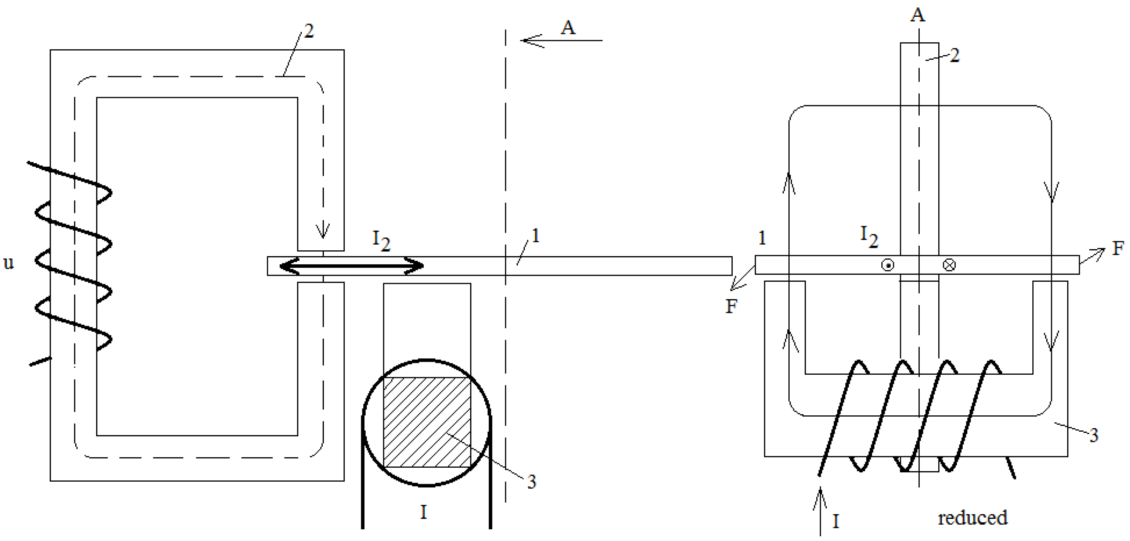


Figure 4.11. Realization of the energy monitoring device

4.1.5. Measurement of power in three-phase electrical circuits

To measure the active power a wattmeter should be used in each phase of a three-phase system (fig.4.12) where the coils of current are connected serially according to the line current in each line of the circuit, but the voltage coils are connected to phase voltages.

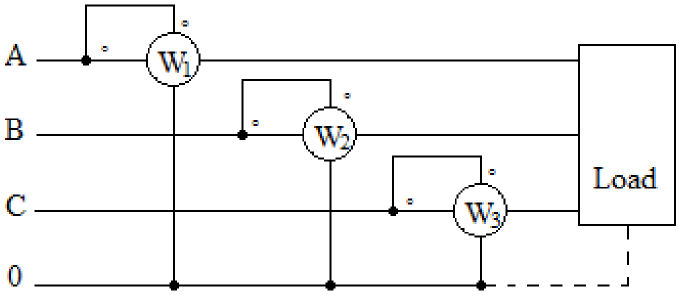


Figure 4.12. Electrical scheme for power measurement in three-phase system

Thus the measured power is

$$P_3 = I_A \cdot U_A \cdot \cos\varphi_A + I_B \cdot U_B \cdot \cos\varphi_B + I_C \cdot U_C \cdot \cos\varphi_C = P_A + P_B + P_C \quad . \qquad (4.48)$$

If the load is fully symmetric then only one wattmeter could be used and the total active power is obtained by multiplication its indication by three.

If there is no zero wire (three-phase Y-type connection of the load and Δ-type connection as well) the total active power could be measured with two wattmeters where the coils of current are connected to two phase circuits (fig.4.13), but the voltage coils are connected between their corresponding phases and a free one. Then the total instantaneous power is

$$p_3 = p_A + p_B + p_C = u_A \cdot i_A + u_B \cdot i_B + u_C \cdot i_C \quad . \qquad (4.49)$$

Taking into account that $i_A + i_B + i_C = 0$, if there is no zero wire, total power is

$$p_3 = u_A \cdot i_A + u_B \cdot i_B - u_C \cdot i_A - u_C \cdot i_B \quad (4.50)$$

or

$$p_3 = i_A(u_A - u_C) + i_B(u_B - u_C) = i_A \cdot u_{AC} + i_B \cdot u_{BC} \; . \quad (4.51)$$

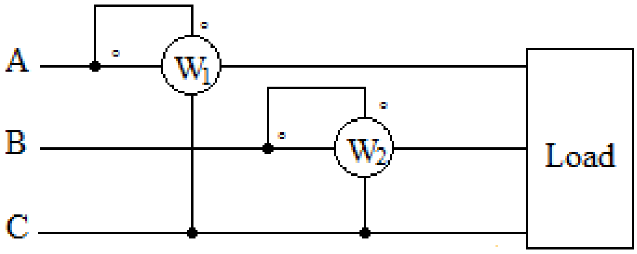


Figure 4.13. Power measurement in three-phase system applying two wattmeters

Correspondingly, an average power which is shown by wattmeter

$$P_3 = I_A \cdot U_{AC} \cdot \cos(I_A U_{AC}) + I_B \cdot U_{BC} \cdot \cos(I_B U_{BC}) = P_1 + P_2 \; , \quad (4.52)$$

i.e. a total active power in three-phase system is equal to the sum of indications of the both wattmeters.

To measure reactive power in the case of symmetric load the only one wattmeter could be used, the coil of current of which is connected to any phase of the circuit, but the voltage coil is connected to other two phases (fig.4.14). The power is the following:

$$P_1 = I_A \cdot U_{BC} \cdot \cos(I_A U_{BC}) = I_A \cdot U_{BC} \cdot \cos(90^0 - \varphi) = I_A \cdot U_{BC} \cdot \cos\varphi. \quad (4.53)$$

A total three-phase reactive power

$$Q = \sqrt{3} \cdot I \cdot U_f \cdot \sin\varphi = 3P_1 \quad , \text{var} \quad (4.54)$$

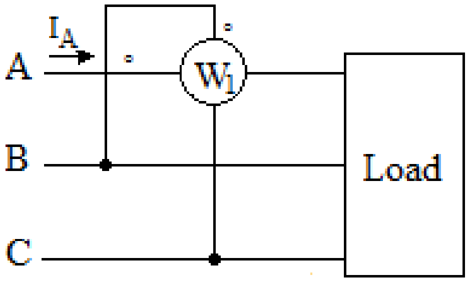


Figure 4.14. Electrical scheme for measurement of reactive power

In the case if symmetric load the reactive power could be measured with two wattmeters which are connected as shown in fig.4.13. Indication of the first wattmeter

$$P_1 = I_A \cdot U_{AC} \cdot \cos(\varphi - 30^0) \; ,$$

but

$$P_2 = I_B \cdot U_{BC} \cdot \cos(\varphi + 30^0) \; , \quad (4.55)$$

where φ is a shift angle between corresponding phases of current and voltage.

4.1.6. Three-phase electrical motor

The principle of rotating magnetic field is applied in a three-phase electric motor. The motor could be operated according to synchronous as well as asynchronous principle. Figure 4.15 demonstrates it better. Three windings shifted in space for 120^0 are supplied from a three-phase AC voltage system u_A, u_B, u_C with an angular speed of rotating voltage vectors ω. When $0<\omega t<60^0$ voltages u_A and u_C are positive, but u_B negative. It means that the current in the conductors of the 1st and 3rd windings are directed from the reader, but the 2nd – to the reader. Thus a clockwise inclining magnetic flux appears around the beginning of the 1st and 3rd windings and the of the 2nd winding. A magnetic flux inclining in an opposite direction appears around the ends of the 3rd and 1st windings and the beginning of the 2nd one. For this case the direction of the total vector of the magnetic flux N-S is shown in fig. 4.15.b.

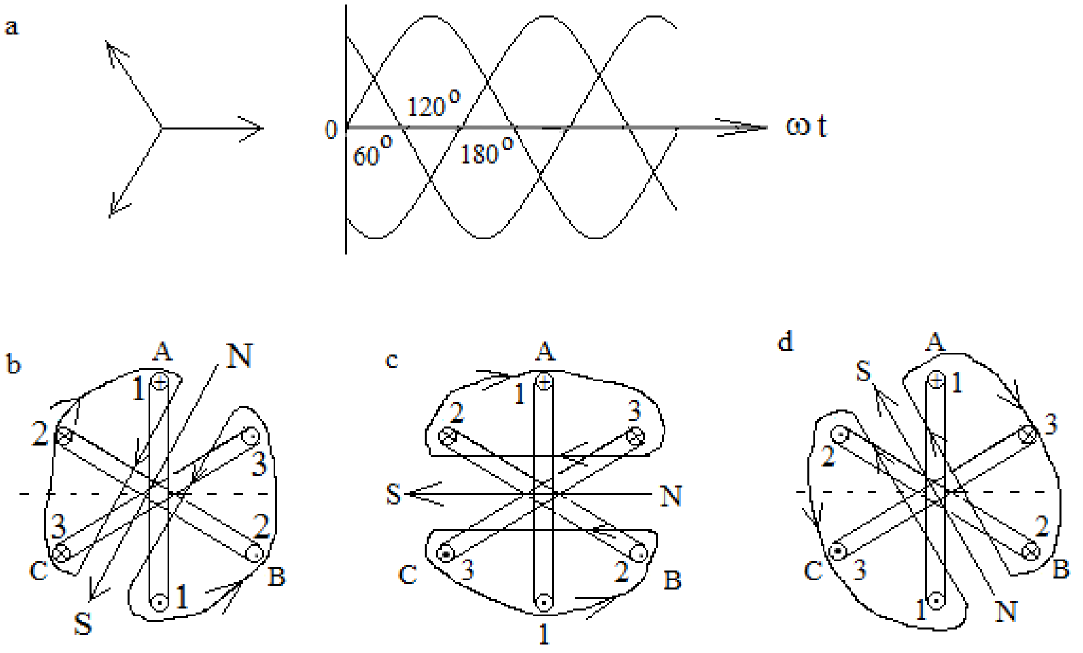


Figure 4.15. To the explanation of rotating magnetic field in three-phase system

The situation is changed when $60^0<\omega t<120^0$, then u_A is positive, u_B and u_C are negative. It means that windings 1 and 2 are in previous situation, but the direction of the current in winding 3 is changed to opposite. Thus the flux inclining in clockwise direction is induced around the ends of the 2nd and 3rd windings and the beginning of the 1st winding. The flux with the opposite direction in its turn is around the beginnings of the 2ns and 3rd and the end of the 2st windings. The total vector of magnetic flux now id directed horizontally to the left, i.e. the vector has turned for 60^0 clockwise (fig.4.15.c).

When $120^0 < \omega t < 180^0$, then u_A, u_B are positive and u_C is negative. Then the further clockwise turn of the vector of the total magnetic flux for 60^0 indicates the new directions of the currents in windings (fig.4.15.d).

During the whole period of time the changes of the voltage have 6 such positions and the total vector of magnetic flux makes the full revolution in clockwise direction. With the aim to change the direction of the vector rotation voltage u_A should be supplied to the 1st winding, u_C – to the 2nd, and u_B – to the 3rd one, i.e. the windings of 2 phases should be replaced (fig.4.16).

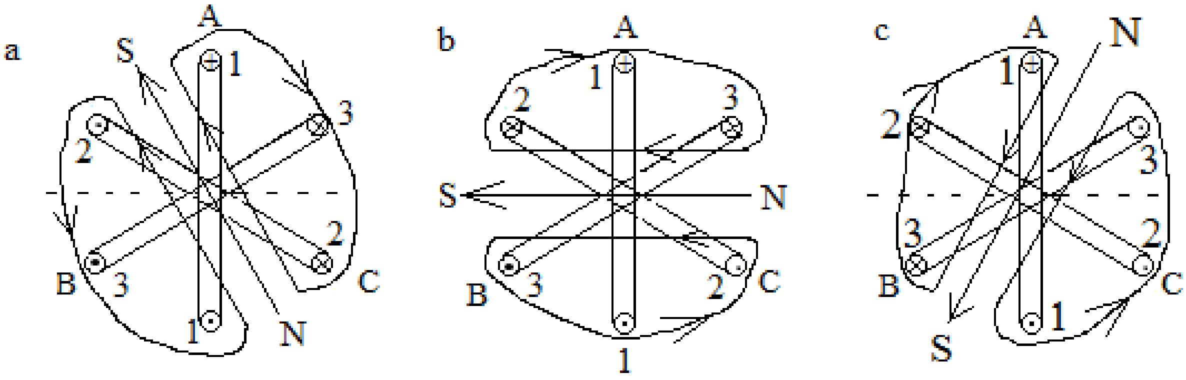


Figure 4.16. Rotating magnetic field at reverse of the connection of supply wires

The figure demonstrates that the total vector of magnetic flux in this case rotates in a clockwise direction.

If to place these three windings on the stator but to realise the rotor as a permanent magnet of electromagnet we get a synchronous motor (fig.4.17). Fig.4.17,a shows the situation when the total vector of magnetic flux is inclined to the up-right position and the rotor turns for the angle of 30^0 clockwise, i.e. the rotor follows the total vector's direction. Like in this situation the total vector of the stator winding rotates in a clockwise direction then the rotor turns in the same direction.

Nevertheless the problem with a synchronous motor could appear with the connecting to supply voltage. Of the rotor has a high level of inertia it will not start its rotation immediately and in a half-period the total vector of the flux makes the rotor to turn into the opposite direction (fig.4.17.b).

The situation mentioned above will be repeated with further rotation, therefore the rotor cannot start is rotation. Thus to provide the normal operation the rotor should be moved with some additional starting winding or an auxiliary engine at the very initial point of operation, that complicates the own construction of the motor and limits the areas of its application.

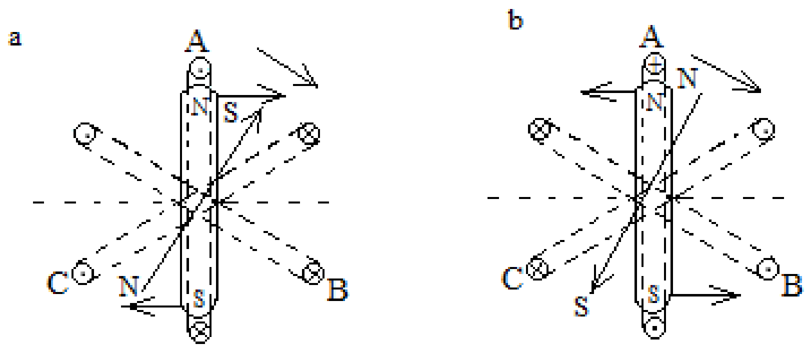


Figure 4.17. Starting position of synchronous motor with high level of inertia

The second possibility is to realise a three-phase asynchronous electric motor (or induction motor) where the rotating magnetic field of stator induces a current in rotor winding with no insulated conductor bars and interacts with the magnetic field of this current (fig.4.18). In the figure the vector of rotating magnetic field of the stator windings rotates clockwise with the velocity n_0. According to the right hand rule EMF and current shown in the figure are induced in the massive rotor bars. According to the left hand rule the mechanical forces directed correspondingly to the direction of stator's magnetic flux rotation are induced in the system. These forces create an electro-magnetic torque M_{em} which rotates the shaft of the motor.

To keep the magnetic field of the rotor it is necessary to keep its intersection with the rotating magnetic field of the stator. This intersection takes place if the rotation velocity of the rotor n_1 is less than that of the stator rotating field n_0

$$n_0 = \frac{60 \cdot f_1}{p}, \tag{4.56}$$

where p is the number of poles pairs of the stator winding; f_1 is the frequency of supply voltage of the stator winding.

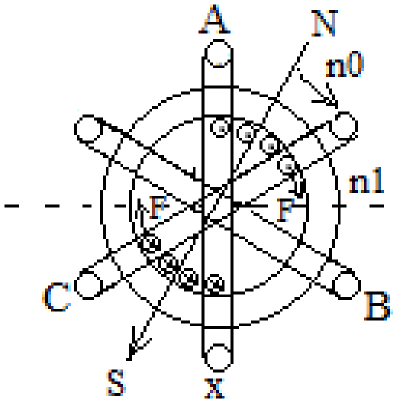


Figure 4.18. Representation of the interaction of rotor and stator windings of asynchronous motor

This difference of the velocities is characterised with a parameter called slip that is a very important for this motor operation

$$s = \frac{n_0 - n_1}{n_0} . \tag{4.57}$$

Multiplying an electric magnetic torque M_{em} of the asynchronous motor by angular velocity of the stator magnetic field ω_0 we can get electro-magnetic power

$$P_{em} = M_{em} \cdot \omega_0 = M_{em} \frac{2\pi \cdot n_0}{60} , \tag{4.58}$$

where n_0 is measured in rev/min. The rated power of the motor P_N is the power produced on the motor shaft and it is less than P_{em}. But the following power supplied from the network $P_1 = \sqrt{3} \cdot U_1 \cdot I_1 \cdot \cos\varphi$ is higher than P_{em}. The ratio between the P_N and P_1 is an efficiency factor. Then the rated current of the stator could be defined as

$$I_{1N} = \frac{P_N}{\sqrt{3} \cdot U_1 \cdot \cos\varphi \cdot \eta} , \tag{4.59}$$

where the power factor cosφ and efficiency factor η for an induction motor of average size (about 20 kW) are about 0.85…0.9. Under the normal circumstances of the motor operation the frequency of the rotor current is not high, if the rated slip s is about 0.03; and the losses of power in the rotor are also not high $\Delta P_{rot} = P_{em} \cdot s$.

4.1.7. Three-phase transformer

A three-phase transformer transforms three-phase voltage of a some particular level into another voltage with either higher or lower level. This kind of transformation is realised on the base of three bars core with equal cross-section area (fig.4.19). Correspondent primary and secondary phase winding are placed on each of these bars. The both windings in the figure have Y-connection.

If the load and supply voltage are symmetric then all rms values of magnetic fluxes of each bar are equal to each other

$$\overline{\Phi_A} + \overline{\Phi_B} + \overline{\Phi_C} = 0 \quad . \tag{4.60}$$

The rms values of the currents in primary windings are also equal

$$\overline{I_{1A}} + \overline{I_{1B}} + \overline{I_{1C}} = 0 \quad , \tag{4.61}$$

The same is in the secondary winding

$$\overline{I_{2A}} + \overline{I_{2B}} + \overline{I_{2C}} = 0 \quad . \tag{4.62}$$

The primary and secondary windings could have also Δ-connection (fig.4.20). Because of that reason the secondary winding in figure fig.4.20,a has no zero wire. In fig.4.20,b the primary is Δ-connection, but the secondary Y-connection. In fig.4.20,c both windings have connection of Δ-type.

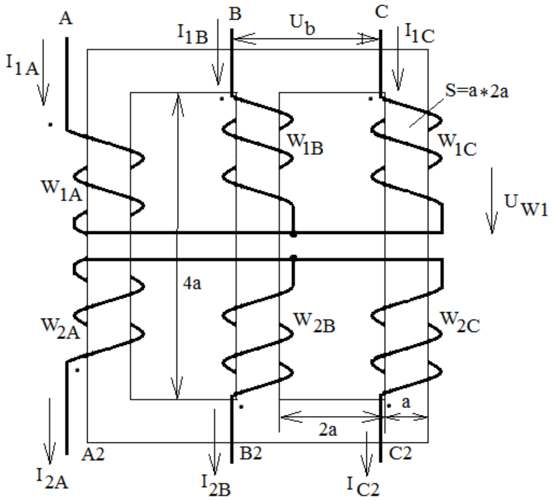
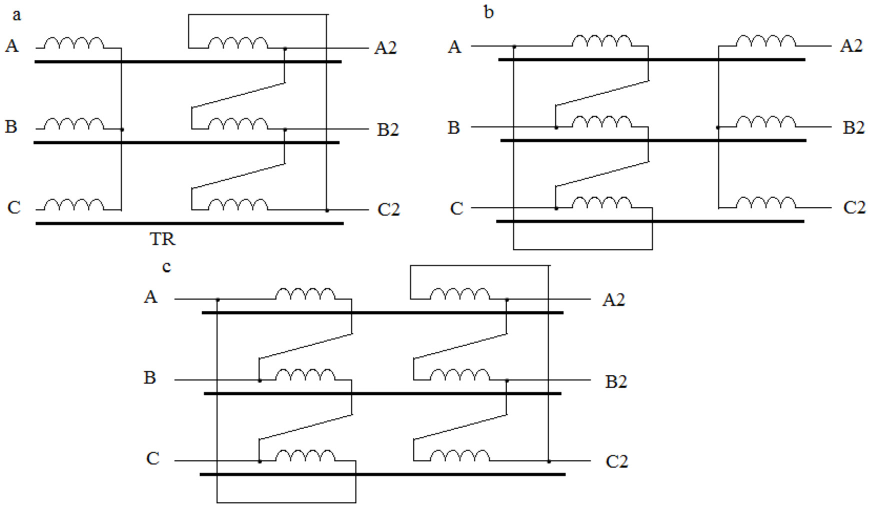


Figure 4.19. Magnetic scheme of three-phase transformer's realisation

If primary windings are of Y-connection the windings of each phase should be calculated according to the phase-to-zero voltage $U_{1ph} = U_1/\sqrt{3}$, where U_1 is phase-to-phase voltage:

$$U_{w1} = U_{1ph} = 4.44 \cdot f \cdot w_1 \cdot B_m \cdot S \quad , \tag{4.63}$$

where f is frequency, B_m is an extreme flux density (about 1.2 T), w_1 is a number of turns for the primary winding, S – cross-section area of the bar.



0

Figure 4.20. Three schemes for ways of connection of three-phase transformer's windings

If primary windings have Δ-connection then the winding of each phase should be calculated in accordance to the phase-to-phase voltage:

$$U_{w1} = U_1 = 4.44 \cdot f \cdot w_1 \cdot B_m \cdot S \,. \tag{4.64}$$

Independently on the way of connection each of the windows of the core has four windings – two primary and two secondary:

$$K_{wf} \cdot S_w = 2 \cdot w_1 \cdot S_{w1} + 2 \cdot w_2 \cdot S_{w2} \,, \tag{4.65}$$

where K_{wf} is a factor characterising a fulfilling of window with the conductor wires (this factor is approximately close to 0.3), S_w is an area of the open space of the window (according to the proportions given in fig.4.19 equal to $8a^2$), w_1 and w_2 are the number of turns of the primary and secondary windings correspondingly, S_{w1} and S_{w2} are cross-section areas for the conductor wires of the windings:

$$S_{w1} = \frac{I_{pR}}{j};$$

$$S_{w2} = \frac{I_{sR}}{j}, \tag{4.66}$$

where I_{pR} and I_{sR} are the rated currents of the primary and secondary windings ($I_{pR} = I_{sR}/K_{TR}$), j is the density of the current in the wires (about $2 \cdot 10^6\ A/m^2$).

Taking into account geometrical parameters shown in figure 4.19 and a ratio of the voltages of the primary and secondary windings $\frac{U_{w1}}{U_{w2}} = K_{TR}$, we can get

$$w_1 = \frac{U_{w1}}{4.44 \cdot f \cdot B_m \cdot 2a^2};$$

$$w_2 = \frac{U_{w1}}{K_{TR} \cdot 4.44 \cdot f \cdot B_m \cdot 2a^2} \tag{4.67}$$

$$\text{but} \quad S_{w1} = \frac{I_{pR}}{j};\ S_{w2} = \frac{I_{pR} \cdot K_{TR}}{j}. \tag{4.68}$$

Taking into consideration that all these relations we can obtain the expression for calculations of parameters a:

$$a = \sqrt[4]{\frac{U_{w1} \cdot I_{pR}}{4.44 \cdot f \cdot B_m \cdot j \cdot 4 \cdot K_{TR}}} \quad . \tag{4.69}$$

The optimal construction can be evaluated by the surface losses of the coil (W/m^2). If the construction is optimal these losses are in the range from 1000 to 1200 W/m^2 providing normal operation of the transformer coils at the temperature of the conductors close to the allowed. These losses can be calculated using an equivalent resistance for the both windings of one bar (fig.4.21).

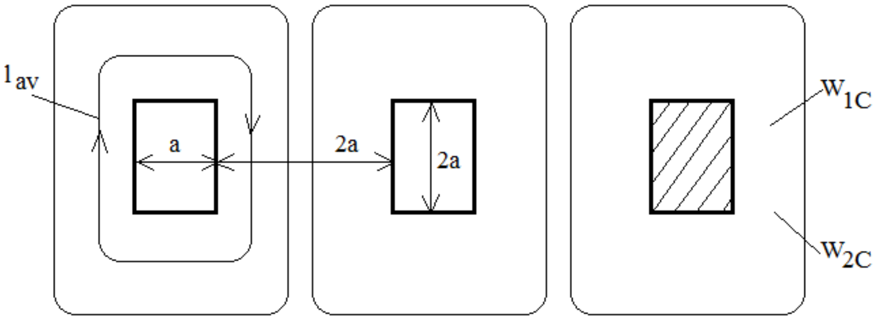


Figure 4.21. Location of the phase winding upon the bars of magnetic core

Reducing the secondary winding to the primary we get the following meaning of the equivalent resistance for a coil

$$R_{sp} = \rho \cdot 2 \cdot w_1 \cdot \frac{l_{av}}{S_{w1}} \qquad , \tag{4.70}$$

where ρ is $0.02 \cdot 10^{-6}\ \Omega \cdot$ m , but l_{av} is an average length of one turn of the coil.

Losses of the equivalent coil are

$$\Delta P_{sp} = I_{pR}^2 \cdot R_{sp} = I_{pR}^2 \cdot \rho \cdot 2 \cdot w_1 \cdot \frac{10a}{S_{w1}} . \tag{4.71}$$

Then the cooling surface of the coil is

$$S_{cool} = 14a \cdot 4a = 56a^2 , \tag{4.72}$$

where 14a is a perimeter of the surface (fig.4.21) but 4a is a height of the coil.

4.2. Means of automation in electrical systems

4.2.1. Parameters and characteristics of elements

Generators

Generators are needed in many applications. Except from the public supply of electricity they are also needed in many situations where independent supply is needed. The generators are used in the public supply networks in which a number of high-power generator sets may operate in parallel, private or independent generators which may run in parallel with the public supply or isolated from it and portable supplies where no alternative supply is available. The private or independent generators may be used to reduce the maximum demand of electricity by a user, to protect the supply to critical circuits such as hospitals or water supplies, for temporary supply needed by the construction industry, or in cases of breakdown.

The cylindrical rotor and the salient pole are the two main types of generators. They are both synchronous machines meaning that the rotor turns in exact synchronism with the rotating magnetic field in the stator. The largest generators used in major power stations are usually cylindrical rotor generators. They operate at high speeds and are usually directly coupled to a steam or gas turbine. The salient pole generator is more commonly used in smaller and medium power ranges.

Except from these two types there are also other types like the induction generators and inductor alternators but they are less commonly used. Induction generators have a simple form of rotor construction making it much cheaper to manufacture and much more reliable. The machine has characteristics which suit wind turbines very well, and they also provide a low-cost alternative for small portable generators. Inductor alternators are usually used for specialized applications requiring high frequency.

The basic operation of all these generator types can be explained using two simple rules, the first for magnetic circuits and the second for the voltage induced in a conductor when subjected to a varying magnetic field.

The flux Φ in a magnetic circuit which has a reluctance R_m is the result of a magnetomotive force (mmf) F_m, which itself is the result of a current I flowing in a coil of N turns.

$$\Phi = F_m/R_m \quad \text{and} \quad F_m = I\,N \tag{4.73}$$

For the salient-pole generator DC current is supplied to the rotor coils through brushes and sliprings. The product of the rotor or field current I and the coil turns N results in F_m and this acts on the reluctance of the magnetic circuit to produce a magnetic flux. As the rotor turns, the flux pattern created by the F_m turns with it. When a magnetic flux Φ passes through a magnetic circuit with a cross section A, the resulting flux density B is given by

$$B = \Phi/A \tag{4.74}$$

As the rotor turns, its magnetic flux crosses a stator with a single coil and axial length l with a velocity v. Electromotive force (emf) V is generated, where

$$V = B\,v\,l \tag{4.75}$$

The direction of the voltage is given by Fleming's right-hand rule. As the magnetic field rotates, the flux density at the stator coil changes. When the pole face is next to the coil, the air gap flux density B is at its highest, and B falls to zero when the pole is 90° away from the coil. The induced emf or voltage V therefore varies with time in the same pattern as the flux density varies around the rotor periphery. The waveform is repeated for each revolution of the rotor; if the rotor speed is 3000 rpm then the voltage will pass through 50 cycles/second (50 Hz). This is the way in which the frequency of the electricity supply from the generator is established.

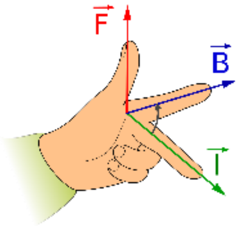


Figure 4.22. Fleming's right-hand rule

The general rule relating the synchronous speed n_s (rpm), number of poles p and the generated frequency f (Hz) is given by

$$f = n_s\, p/120 \tag{4.76}$$

A simple voltage output can be delivered to the load with a pair of wires as a single phase supply. If more coils are added to the stator and if these are equally spaced, then a three-phase output can be generated. The positive voltage peaks occur equally spaced, one-third of a cycle apart from each other. The three coils either supply three separate loads for three electric heating elements, or more usually they are arranged in either star or delta arrangement in a conventional three-phase circuit.

Loads

Loads are any devices in which power is consumed. From the circuit perspective, a load is defined by its impedance, which comprises a resistance and a reactance. The impedance of an individual device may be fixed, as in the case of a simple light bulb, or it may vary, for example, if an appliance has several operating settings. In the larger context of power systems, loads are usually modeled in an aggregated way: rather than considering an individual appliance. Load may refer to an entire household, a city block, or all the customers within a certain region. In the language of electric utilities, the term load therefore has attributes beyond impedance that relate to aggregate behavior, such as the timing of demand. From a physical perspective, we would think of loads in terms of the electrical characteristics of individual devices. If we consider a load as being defined by its impedance, there are theoretically three types of loads: purely resistive loads, inductive loads, and capacitive loads. Resistive loads are those consisting basically of a heated conductor. Inductive loads are the most common and include all types of motors, fluorescent lights, and transformers like those used in power supplies for lower voltage appliances. Capacitors, by contrast, do not lend themselves for doing mechanical or other practical work outside electrical circuitry. We know of capacitive loads as standard components of electronic circuits, but not on the macroscale among utility customers.

Loads also differ in the type of electric power they can use (AC or DC). The simplest type of load is a purely resistive load, that is, one without capacitive or inductive reactance. The power factor of such a load is 1.0 and dissipates power according to the relationship $P = I^2R$. What people care about is how much power, in watts, a device consumes. In general, resistive loads are the simplest to operate and the most tolerant of variations in power quality, meaning variations in the voltage level above or below the nominal V, or departures of AC frequency from the nominal 50 or 60 Hz. Reducing the voltage for appliances other than simple resistors may damage them, or they may not work at all.

Electric motors represent an important fraction of residential, commercial, and industrial loads. Motor loads comprise fans, pumps of all kinds including refrigerators and air conditioners, basically anything electric that moves. A motor is essentially the same thing as a generator operated backwards. Electrical and mechanical energy are converted into one another by means of a magnetic field that interacts with both the rotating part of the machine and the electrons inside the conductor windings. Aside from differences in size and power, there are three distinct types of motors that correspond to the three main types of generators: induction, synchronous, and DC. In each case, the motor is similar to its generator counterpart. Induction motors are the least expensive and by far the most common. Besides motor type, another important distinction is between single and three phase motors. Three-phase motors operate more smoothly and much more efficiently than single-phase motors, though they are also more expensive. They are commonly used for large industrial and commercial applications where high performance, including high horsepower output and high efficiency, is essential, and where three-phase utility service is standard. Unlike resistive loads, electric motors are sensitive to power quality, including voltage, frequency, harmonic content and, in the case of three-phase machines, phase imbalance. One of the key problems that tend to afflict motors is unequal and excessive heating of the windings. Because of these sensitivities, it is common for owners of expensive and sophisticated motor systems to install their own protective or conditioning equipment, so as to guarantee power quality beyond the standard provided by the local utility.

Consumer electronics are powered by low-voltage DC. They may be operated either by batteries or through a power supply that delivers lower-voltage DC by way of a step-down transformer and rectifier. While individual electronic loads tend not to be large power consumers, their proliferation in number is turning them into a significant load category. The power consumption of any electronic appliance can be gauged by how warm it gets during operation. This type of heat is very much like the waste heat resulting from mechanical friction, which can be reduced by clever design but never completely eliminated. Since the main job of an electronic circuit component is not to deliver power but to relay the information whether a particular circuit is “on” or “off,” this job can be performed with a very small current. To the power circuit, mixed electronic and power appliances appear as resistive or inductive loads, depending on the feature that dominates energy consumption. From the power system perspective, they differ from the plain variety of resistive and inductive loads mainly in their sensitivity to power quality.

From the point of view of the power grid, individual customers and their appliances are small, numerous, and hardly discernible as distinct loads. Therefore, frequently we deal with aggregate load, that is, the combined effect of many customers both in terms of the magnitude and timing of electric demand. While consumers typically think of their electricity usage in terms of a quantity of energy (in kilowatt-hours) consumed over the course of a billing period, the quantity of interest to system operators and planners is the power (in kilowatts or megawatts, measuring the instantaneous rate of energy flow) demanded at any given time. The term demand thus refers to a physical quantity of power, not energy. Serving that instantaneous demand under diverse circumstances is the central challenge in designing and operating power systems, and the one that calls for the majority of investment and effort. An entire discipline is dedicated to predicting what the demand might be at a given future time.

Transmission Lines

Transmission lines are an important component of an electric power system. They enable transmission of electricity from the power plants to the consumers by carrying electric power from one end of the line to the other. A line has four distributed electrical parameters affecting the way it transmits electric power from sending to receiving end: series resistance, series inductance, shunt capacitance, and shunt conductance. The distributed resistance and inductance form the series impedance of the line, while the capacitance and the conductance present between conductors or conductor to neutral form the shunt admittance of the line. The value of these parameters depends on the cable material characteristics and on the electric and magnetic fields along and around the conductors. Therefore, the geometrical configuration of the lines also plays an important role in the determination of these parameters.

The line resistance is defined by

$$R = \frac{P_{loss}}{|I|^2}\Omega, \tag{4.77}$$

where

R: Effective resistance of a conductor

P_{loss}: Power loss in a conductor (W)

I: rms current in the conductor (A)

This definition is most applicable at steady-state. If the effective resistance is equal to the dc resistance of the conductor:

$$R_o = \frac{\rho l}{A}\Omega \tag{4.78}$$

where

Ro: DC Resistance of the Conductor

ρ: Resistivity

l: Length

A: Cross-Sectional Area

For direct current, the current distribution throughout a conductor is uniform but with alternating current as the frequency increases, the non-uniformity increases and skin effect must be considered.

In order to study the line as part of the power system, it is important to be able to determine how line parameters influence the flow of power through the system. The line behavior in terms of wave propagation is of interest. Specifically, voltage and current relationships along the lines are needed.

A simplification of this model is more commonly utilized for studies on power system behavior. In the distributed parameter model, the line parameters are uniformly distributed throughout the length of the line. The four line are quantified in per unit length. The model is obtained as a summation of differential sections (or segments). A differential section of the line model is made of two components: a series impedance z and a shunt admittance y, quantified respectively by

$z = r + jwl$, Series impedance per unit length

$y = g + jwc$ y, Shunt admittance per unit length to neutral

where

ω: Angular Frequency (rad/sec)

r: Series Resistance (Ω per unit length)

l: Series Inductance (H per unit length)

g: Shunt Conductance (S per unit length)

c: Shunt Capacitance (F per unit length)

A differential section of the distributed line model is shown in Fig. 4.23 for a section of length dx. Its series impedance and shunt admittance are then z dx and y dx, respectively.

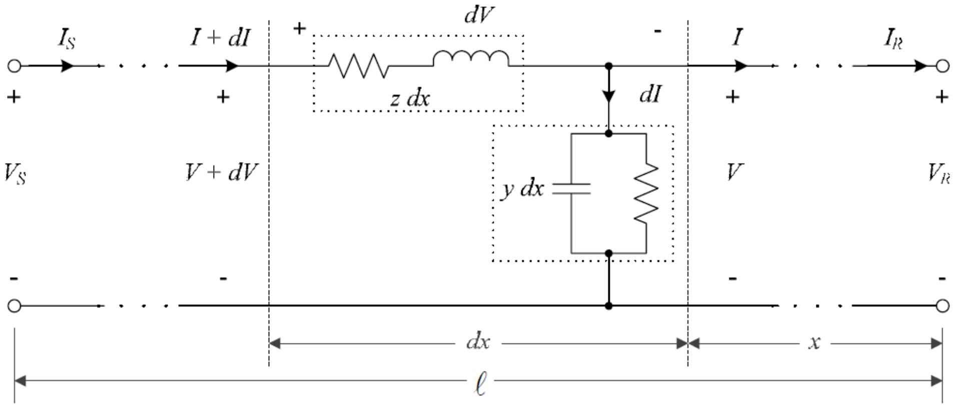


Figure 4.23. Representation of the distributed parameter model

The voltage and current equations in terms of x, point along the line are given by:

$$V = V_R \cosh(\gamma x) + Z_C I_R \sinh(\gamma x) \tag{4.79}$$

$$I = I_R \cosh(\gamma x) + \frac{V_R}{Z_C} \sinh(\gamma x) \tag{4.80}$$

where

$\gamma = \sqrt{yz}$ is the propagation constant

$Z_c = \sqrt{\frac{z}{y}}$ is the characteristic impedance of the line

As mathematically expressed by the above equations, with the distributed line model the steady-state voltages and currents can be determined at any point along the line. However, the relationship between the terminal voltages and the terminal currents are often sufficient in power system studies:

$$V_s = V_R \cosh(\gamma l) + Z_C I_R \sinh(\gamma l)$$

$$I_s = I_R \cosh(\gamma l) + \frac{V_R}{Z_C} \sinh(\gamma l) \tag{4.81}$$

These equations can be re-written in matrix form as follows:

$$\begin{bmatrix} V_s \\ I_s \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} V_R \\ I_R \end{bmatrix} \tag{4.82}$$

where

$$A = \cosh(\gamma l) \qquad B = Z_c \sinh(\gamma l)$$

$$C = \frac{1}{Z_c} \sinh(\gamma l) \qquad D = \cosh(\gamma l)$$

A lumped parameter model of the line is then created to be an equivalent circuit with the same transmission parameters, A, B, C, and D, as the distributed line model. The relationships of terminal voltages and currents are therefore maintained, but information on voltage and current propagation along the line is lost.

Common examples of lumped parameter models are the pi (π) and the gamma (Γ) models shown in the figures. The lumped-equivalent circuits are obtained by selecting the model components Z' and Y' so that the terminal behavior of the distributed line model is preserved with the use of passive elements.

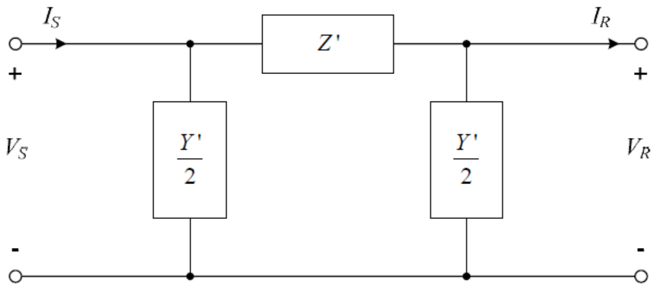


Figure 4.24. π-Equivalent model

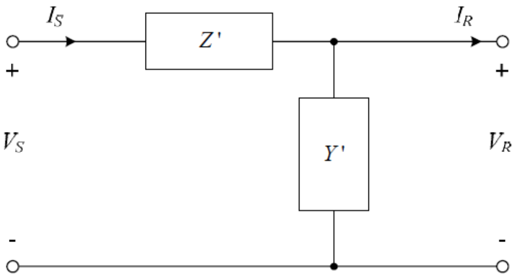


Figure 4.25. Γ-equivalent model

The voltage and current relationships for these equivalent circuits are simply obtained by applying Kirchoff's voltage law (KVL) and current law (KCL). The appropriate Z' and Y' are determined by:

$$Z' = Z_c \sinh(\gamma l)$$

$$\frac{Y\prime}{2} = \frac{1}{Z_c}\tanh(\frac{\gamma \mathrm{l}}{2}) \tag{4.83}$$

At the fundamental frequency, the lumped model is in fact proven to have accurate terminal behavior with respect to the distributed parameter model.

Transformers

A transformer consists of two windings connected by a magnetic core. One winding is connected to a power supply and the other to a load. The circuit containing the load may operate at a voltage which differs widely from the supply voltage, and the supply voltage is modified through the transformer to match the load voltage. In a practical transformer, there may be more than two windings as well as the magnetic core.

With no load current flowing, the transformer can be represented by two windings on a common core. The input and output voltages and currents in a transformer are related by the number of turns in these two windings, which are usually called the primary and secondary windings as shown below.

The magnetic flux density in the core is determined by the voltage per turn and represents a key relationship between the frequency, the number of turns in a winding and the size of the core.

$$\frac{V_1}{N_1} = 4.44\, f\, B_m\, A \tag{4.84}$$

In the no-load case, a small current I_0 flows to supply the mmf which drives the magnetic flux around the transformer core; this current lags the primary voltage by almost 90°. This I_0 is limited in magnitude by the effective resistance (Rc) and reactance (Xc) of the magnetizing circuit.

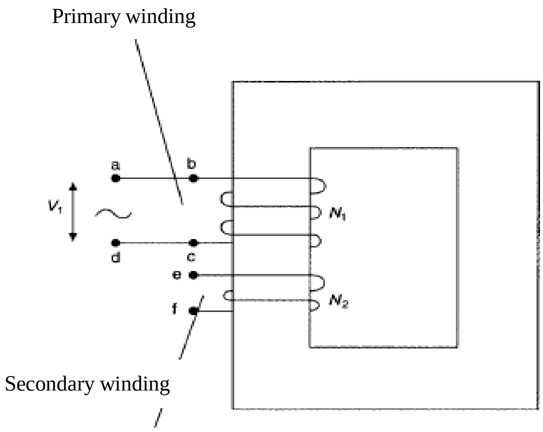


Figure 4.26. Simple transformer circuit

$$\frac{V_1}{V_2} = \frac{N_1}{N_2}; \quad \frac{I_1}{I_2} = \frac{N_2}{N_1} \tag{4.85}$$

When the transformer is loaded, there is an internal voltage drop due to the current flowing through each winding. The voltage drops due to the primary and secondary winding resistances (R1 and R2) are in phase with the winding voltage, and the voltage drops due to the primary and secondary winding leakage reactances (X1 and X2) lag the winding voltage by 90°. The leakage reactances represent those parts of the transformer flux which do not link both the windings; they exist due to the flow of opposing currents in each winding and they are affected strongly by the winding geometry.

The current flow and voltage drops within the windings can be calculated using the equivalent circuit. This circuit is valid for frequencies up to 2 kHz.

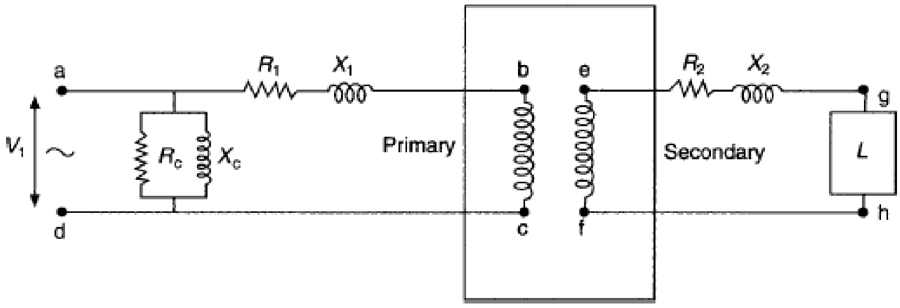


Figure 4.27. Equivalent circuit of transformer

The decrease in output voltage when a transformer is on load is known as regulation. The output voltage is less than the open-circuit voltage because of voltage drops within the winding when load current flows through the resistive and reactive components. The resistive drops are usually much smaller than the reactive voltage drops, especially in large transformers, so the impedance Z of the transformer is predominantly reactive. Regulation is usually expressed as a percentage value relating the vector addition of the internal voltage drops to the applied voltage.

Resistance and reactance values can be referred from secondary to primary windings (and vice versa) using the square of the turns ratio.

$$R = R_1 + R_2 \left(\frac{N_1}{N_2}\right)^2$$

$$X = X_1 + X_2 \left(\frac{N_1}{N_2}\right)^2$$

$$L' = L \left(\frac{N_1}{N_2}\right)^2 \tag{4.86}$$

The impedance of the transformer is given by: $Z = (R^2 + X^2)^{1/2}$ (4.87).

Values for R1 and R2 can be established by measuring the resistance of the windings. The value of X is determined by calculation or by derivation from the total impedance Z, which can be measured with one winding of the transformer short circuited. This Z is given by $Z = \frac{V}{I}$,

where V is the voltage necessary to circulate the full-load current I in the windings under short circuit conditions. When V is expressed as a percentage of rated voltage, this gives Z as a percentage value referred to rated power.

When the transformer is energized, but without a load applied, the no-load power loss is due to the magnetic characteristics of the core material used and the flow of eddy currents in the core laminations. The loss due to magnetizing current flowing in the winding is small and can be ignored. When the transformer is loaded, the no-load loss is combined with a larger component of loss due to the flow of load current through the winding resistance. Additional losses on load are due to eddy currents flowing within the conductors and to circulating currents which flow in metallic structural parts of the transformer, stray losses. The total losses consist of the no-load loss (or iron loss) which is constant with voltage and the load loss (or copper loss) which is proportional to the square of load current.

Capacitors

Power capacitors provide several benefits to power systems. Among these include power factor correction, system voltage support, increased system capacity, reduction of power system losses, reactive power support, and power oscillation damping.

Power Factor Correction: In general, the efficiency of power generation, transmission, and distribution equipment is improved when it is operated near unity power factor. The least expensive way to achieve near unity power factor is with the application of capacitors. Capacitors provide a static source of leading reactive current and can be installed close to the load. Thus, the maximum efficiency may be realized by reducing the magnetizing (lagging) current requirements throughout the system.

System Voltage Support: Power systems are predominately inductive in nature and during peak load conditions or during system contingencies there can be a significant voltage drop between the voltage source and the load. Application of capacitors to a power system results in a voltage increase back to the voltage source, and also past the application point of the capacitors in a radial system. The actual percentage increase of the system voltage is dependent upon the inductive reactance of the system at the point of application of the capacitors.

Increased System Capacity: The application of shunt or series capacitors can affect the power system capacity. Application of shunt capacitors reduces the inductive reactive current on the power system, and thus reduces the system kVA loading. This can have the effect of increasing system to serve additional load. Series capacitors are typically used to increase the power carrying capability of transmission lines. Series capacitors insert a voltage in series with the transmission line that is opposite in polarity to the voltage drop across the line, which decreases the apparent reactance and increases the power transfer capability of the line.

Power System Loss Reduction: The installation of capacitors can reduce the current flow in a power system. Since losses are proportional to the square of the current, a reduction in current will lead to reduced system losses.

Reactive Power Support: Capacitors can help support steady-state stability limits and reactive power requirements at generators.

Power Oscillation Damping: Controlled series capacitors can provide an active damping for power oscillations that many large power systems experience. They can also provide support after significant disturbances to the power system and allow the system to remain in synchronous operation.

The most common application of power capacitors is shunt connected capacitors and they are either energized continuously or switched on and off during load cycles. A fixed series capacitor bank is an assembly of different components inserted in series with a power line to reduce the apparent reactance, increase the power transfer capability and improve system stability.

Reactors

A reactor is a coil with a large number of turns. They are added to the system for protection specially to limit the short circuit currents that may cause damage to the equipment. Reactors may be used for arc suppression, to filter out harmonics, in series with low reactance autotransformers or induction regulators, to protect from high voltage waves, surges and lighting and to control starting current of motors. Reactors have similar uses like capacitors but have contrary effect to the system.

Protection Relays

Systems incorporating protection relays can disconnect high currents in high-voltage circuits which are beyond the scope of fuse systems. In general, relays operate in the event of a fault by closing a set of contacts or by triggering a thyristor. This results in the closure of a trip-coil circuit in the circuit breaker which then disconnects the fault. The presence of the fault is detected by current transformers, voltage transformers or bimetal strips.

Electromechanical and solid-state relays are both widely used, but the latter are becoming more widespread. All components of the protection system including current transformers, relays and circuit breakers must have the correct current, voltage and frequency rating, and the I^2t let-through and interrupting capacity of the entire system depends upon the circuit breaker. A protection relay must have a minimum operating current which is greater than the rated current of the protected circuit, and other properties of the relay must be chosen correctly in order for it to operate the circuit breaker as required by the application.

Protection levels, time delays and other characteristics of both electromechanical and solid-state relays can be changed on site. In general, electromechanical relays can be adjusted continuously and solid-state relays are adjusted in steps. The adjustment of operating times is used to provide time-grading and co-ordination. An alternative system uses protection relays which operate only when a fault is in a clearly defined zone. To achieve this, a relay compares quantities at the boundaries of a zone. Protection based on this principle can be quicker than with time-graded systems because no time delays are required.

Protection relays may be ‘all-or-nothing’ types, such as overcurrent tripping relays, or they may be measuring types which compare one quantity with another. An example of the latter is in synchronization, when connecting together two sources of power.

A protection relay may be classified according to its function. Common applications are undervoltage and overvoltage detection, overcurrent detection, overfrequency and underfrequency detection. These functions may be combined in a single relay. For example, a relay may be used in power stations and provides overvoltage and undervoltage, overfrequency and underfrequency protection, and it rapidly disconnects the generator in the case of a failure in the connected power system.

Another application in power systems is the protection of transmission lines by distance relays. Current and voltage inputs to a distance relay allow detection of a fault within a predetermined distance from the relay and within a defined zone. The fault impedance is measured and if it is less than a particular value, then the fault is within a particular distance.

Another form of protection for lengths of conductor is the pilot wire system. Current transformers are placed at each end of a conductor and they are connected by pilot wires. Relays determine whether the currents at the two ends of the conductor are the same, and they operate if there is an excessive difference.

Other applications involving relays include the protection of motor starters against overload, checking phase balance, the protection of generators from loss of field and the supervision of electrical conditions in circuits.

Switchgear

Switchgears are used to connect and disconnect electric power supplies and systems. It is a general term which covers the switching device and its combination with associated control, measuring, protective and regulating equipment, together with accessories, enclosures and supporting structures. Switchgears are applied in electrical circuits and systems from low voltage right up to transmission networks. The main classes of equipment are disconnectors, or isolators, switches, fuse switch combinations, circuit breakers and earthing switches.

A *disconnector* is a mechanical switching device which in the open position provides a safe working gap in the electrical system which withstands normal working system voltage and any overvoltages which may occur. It is able to open or close a circuit if a negligible current is switched, or if no significant change occurs in the voltage between the terminals of the poles.

A *switch* is a mechanical device which is able to make, carry and interrupt current occurring under normal conditions in a system, and to close a circuit safely, even if a fault is present.

A *fuse* and a *switch* can be used in combination with ratings chosen so that the fuse operates at currents in excess of the rated interrupting or breaking capacity of the switch.

A *circuit breaker* is a mechanical switching device which is not only able to make, carry and interrupt currents occurring in the system under normal conditions, but also to carry for a specified time and to make and interrupt currents arising in the system under defined abnormal conditions, such as short circuits. It experiences the most onerous of all the switching duties and is a key device in many switching and protection systems.

An *earthing switch* is a mechanical device for the earthing and short-circuiting of circuits. It is able to withstand currents for a specified time under abnormal conditions, but it is not required to carry normal service current.

4.2.2. Converters of voltage, current and frequency

Power converters are electronic circuits associated to the conversion, control, and conditioning of electric power. The power range can be from milliwatts, mobile phone, for example, to megawatts, in electric power transmission systems. Reliability of the power converters become a key industrial focus. Electronic devices and control circuit must be highly robust in order to achieve a high useful life. A special accent must be set on the total efficiency of the power electronic circuits. Firstly, because of the economic and environmental value of wasted power and, secondly, because of the cost of energy dissipated that it can generate.

Even a small improvement in converter power efficiency translates to improved profitability of the investment in the electronic market.

Among all electronic converters, the most common technology is switched-mode power converters (SMPC). They convert the voltage input to another voltage signal, by storing the input energy temporarily and then releasing that energy to the output at a different voltage. This switched-mode conversion has a particular interest due to the fact that it can switch at high frequency in a very efficient way. Power is controlled (even modified) by controlling the timing that the electronic switches are “on” and “off”.

A much greater emphasis is required on achieving high-power efficiency in low-power level electronic technology, since few low-power circuits can tolerate a power efficiency less than 85%. Converters are used in these circuits in order to change the supply voltage in the blocks of the System on Chips (SoCs) according to performance requirements, for power efficiency reasons. Research has been focused on developing electronic circuits that can be employed as switches. e.g. approximating ideal closed or open switches, as the Vdd-hopping converter.

Converters classification

Power converters control the flow of power between two systems by changing the character of electrical energy: from direct current to alternating current, from one voltage level to another voltage, or in some other way.

Here, some important way to classify the power converters are described. The aim of this section is not to make a rigorous converter classification, either to make a state of the art, because it is not the purpose of this thesis. It is only desired to understand some properties of these kind of circuits.

The most common classification of power conversion systems is based on the waveform of the input and output signals, in the case whether they are alternating current (AC) or direct current (DC), thus:

- **DC to DC**.
- **DC to AC**. Inverter.
- **AC to DC**. Rectifier.
- **AC to AC**. Transformer.

At the same time, the devices within converters can be switched in different ways. If the devices switch at the line frequency (normally, 50*Hz* or 60*HZ*), they can be *line frequency converters* (naturally commutated converters) or *high-frequency switching* (forced-commutated converters).

Depending on the character of the input source, they may be *voltage-source converters* or *current-source converters*. Moreover, converters may be of low, medium or high voltage and/or current level. Another sort of classification may be performed according to the size of the output signal obtained from the input signal; if the converter accomplishes a lower output signal it is well known as *step-down*, and if it obtains a larger signal, it is known as *step-up*.

DC-DC

i) DC-DC converter

DC-DC converters are electronic circuits that change the DC operating voltage or current. They have recently aroused the interest in the current market due to its wide range of applicability. Normally, they are designed in order to transfer power from the input to the output in one direction. However, in the case of switches topologies, the power moving may be also bidirectional, being very useful to develop new converter topologies for other applications, as can be an inverter topology.

They have a particular interest in low-power circuits, as cellular phone and personal computers (PCs). This sort of technology are composed of many sub-circuits that require an own voltage level from an external supply (higher, lower or even negative) or battery. DC-DC converters have a special role in these kind of systems, since they can be employed to change the voltage from a partial lowered battery voltage thereby. This is based on the Dynamic Voltage Scaling technic (DVS). The main idea of DVS is to vary the supply voltage in order to consume a minimal amount of energy. This fact improves the power efficiency and saves space in spite of using multiple batteries to accomplish the same voltage level.

ii) Voltage regulator

A voltage regulator is designed to automatically maintain a constant voltage level. A voltage regulator may be a simple "feed-forward" design or may include negative feedback control loops. It may use an electromechanical mechanism, or electronic components. Depending on the design, it may be used to regulate one or more AC or DC voltages.

Electronic voltage regulators are found in devices such as computer power supplies where they stabilize the DC voltages used by the processor and other elements. In automobile alternators and central power station generator plants, voltage regulators control the output of the plant. In an electric power distribution system, voltage regulators may be installed at a substation or along distribution lines so that all customers receive steady voltage independent of how much power is drawn from the line.

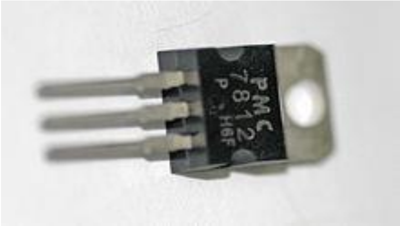


Figure 4.28. An integrated circuit voltage regulator in a TO-220 style package.

iii) Linear regulator

A linear regulator is a system used to maintain a steady voltage. The resistance of the regulator varies in accordance with the load resulting in a constant output voltage. The regulating device is made to act like a variable resistor, continuously adjusting a voltage divider network to maintain a constant output voltage and continually dissipating the difference between the input and regulated voltages as waste heat. By contrast, a *switching regulator* uses an active device that switches on and off to maintain an average value of output. Because the regulated voltage of a linear regulator must always be lower than input voltage, efficiency is limited and the input voltage must be high enough to always allow the active device to drop some voltage.

Linear regulators may place the regulating device in parallel with the load (shunt regulator) or may place the regulating device between the source and the regulated load (a series regulator). Simple linear regulators may only contain a Zener diode and a series resistor; more complicated regulators include separate stages of voltage reference, error amplifier and power pass element. Because a linear voltage regulator is a common element of many devices, integrated circuit regulators are very common. Linear regulators may also be made up of assemblies of discrete solid-state or vacuum tube components.

DC-AC

i) DC-AC converter

DC-AC converters, or commonly named inverters, can obtain a certain amplitude and frequency of the AC voltage and/or current without using normally an intermediate DC stage. This electrical device is a power electronic oscillator. An electronic oscillator is just an electric circuit that produces a repetitive signal, as a sine-wave output signal. Generally, they are SMPCs.

These kind of circuits require an efficient control for the switches devices that, in many occasions, can be quite complex due to system structure. Therefore, to design a suitable control law currently is a subject of much research.

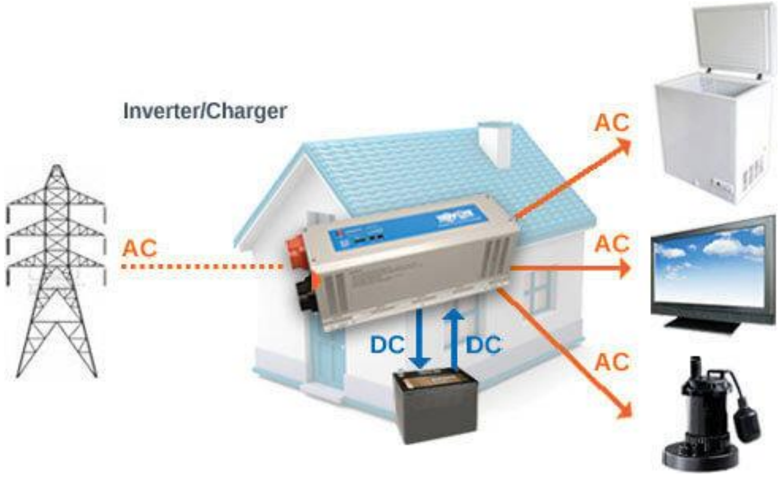


Figure 4.29. DC-AC converter/ inverter

AC-DC

i) AC-DC converter

The process that converts AC to DC is known as rectification, hence, these converters are also called rectifier. Among others applications, they are used in power supplies and detector of radio signals.

The rectification can be half-wave or full-wave. In the first case (half-wave rectification), only one half of the input waveform can be employed to reach the desired output. Therefore, only this half AC wave (positive or negative) is converted. The efficiency will depend of the kind of application. It is clear, that it is not useful for power transfer. The full-wave rectification can convert the whole of the input waveform to achieve the constant output signal. It becomes more efficient.

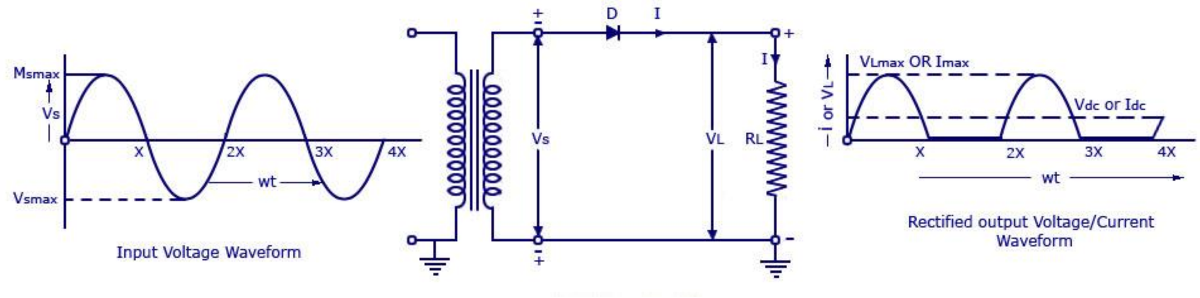


Figure 4.30. Half-wave rectifier

AC-AC converter

i) AC-AC converter

AC-AC converters are employed to transform an AC input signal to another AC output signal with an arbitrary amplitude. Likewise, depending on the converter complexity, the frequency can be changed as well. The efficiency of these kind of systems depends on the type of circuit employed. It is clear that a higher power density and reliability will be obtained with a conversion in one single stage.

ii) Transformer

A transformer is an electrical device that transfers electrical energy between two or more circuits through electromagnetic induction. Electromagnetic induction produces an electromotive force within a conductor which is exposed to time varying magnetic fields. Transformers are used to increase or decrease the alternating voltages in electric power applications.

A varying current in the transformer's primary winding creates a varying magnetic flux in the transformer core and a varying field impinging on the transformer's secondary winding. This varying magnetic field at the secondary winding induces a varying electromotive force (EMF) or voltage in the secondary winding due to electromagnetic induction. Making use of Faraday's Law (discovered in 1831) in conjunction with high magnetic permeability core properties, transformers can be designed to efficiently change AC voltages from one voltage level to another within power networks.

Since the invention of the first constant potential transformer in 1885, transformers have become essential for the transmission, distribution, and utilization of alternating current electrical energy. A wide range of transformer designs is encountered in electronic and electric power applications. Transformers range in size from RF transformers less than a cubic centimeter in volume to units interconnecting the power grid weighing hundreds of tons.

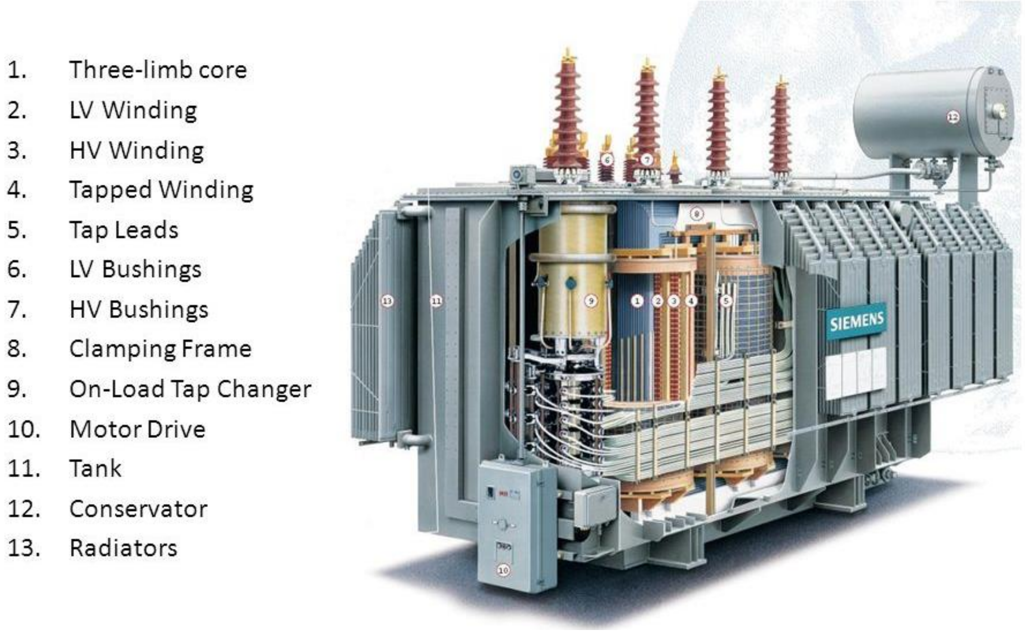


Figure 4.31. Transformer main parts

iii) Autotransformer

Autotransformer (Auto-Transformer) is an electrical transformer in which there is one winding, a portion of which is common to both the primary and the secondary circuits. An autotransformer uses common winding and offer no interference or disturbance isolation. The current in the high-voltage circuit flows through the series and common winding. The current in the low-voltage circuit flows through the common winding and adds vectorially to the current in the high-voltage circuit to give the common winding current.

The principle difference between an autotransformer and an isolation transformer is the separation of the secondary windings. Because the auto-transformer uses a single coil winding for both the primary input and the secondary output, any electrical noise, voltage spikes, sags or any other undesirable condition will pass through unchecked. Equipment susceptible to damage by poor line conditions will not be protected. And noise and harmonics generated by components on the secondary side will be allowed to transmit onto the main supply line. Because the auto-transformer can transmit line disturbances directly, local building codes may prohibit their use in certain areas. Auto-Transformers also should not be used in closed delta connections as they will introduce into the circuit a phase shift which causes higher power use.

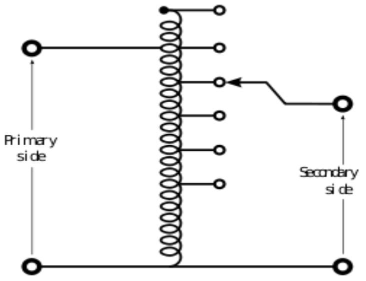


Figure 4.32. Autotransformer

iv) Cycloconverter

A cycloconverter (CCV) or a cycloinverter converts a constant voltage, constant frequency AC waveform to another AC waveform of a lower frequency by synthesizing the output waveform from segments of the AC supply without an intermediate DC link. There are two main types of CCVs, circulating current type or blocking mode type, most commercial high power products being of the blocking mode type.

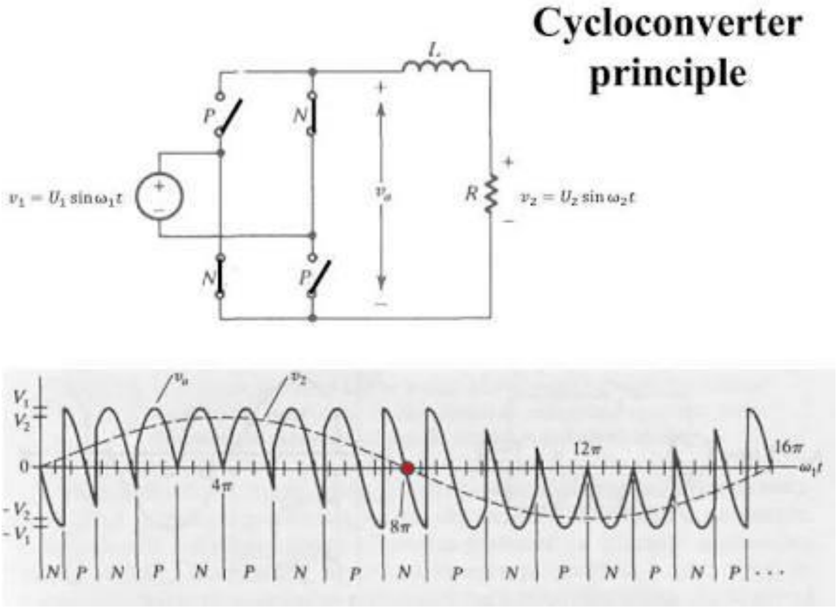


Figure 4.33. Cycloconverter principle

4.2.3. Sensors

Current Transformer (CT)

Current Transformer (CT), is one of the main types of instrument transformers. Instrument transformers are high accuracy electrical devices which are widely utilized for changing currents and voltages in power systems to values more suited to the range of conventional measuring devices. More specifically, CT is a current transducer that provides a current signal directly proportional in magnitude and phase to the current flowing in the primary circuit. Thus it reduces high voltage currents to a much lower value and provides a convenient way of safely monitoring the actual electrical current flowing in an AC transmission line using a standard ammeter.

Current transformers are often constructed by passing a single primary turn (either an insulated cable or an uninsulated bus bar) through a well-insulated toroidal core wrapped with many turns of wire as shown in Fig. 4.34. This affords easy implementation on high voltage bushings of grid transformers and other devices by installing the secondary turn core inside high-voltage bushing insulators and using the pass-through conductor as a single turn primary. Due to this type of arrangement, the current transformer is often referred as a “series transformer” as the primary winding, which never has more than a very few turns, is in series with the current carrying conductor supplying a load. An accurate current transformer requires a close coupling between the primary and secondary winding to ensure that the secondary current is proportional to the primary current over a wide current range. The current in the secondary winding is the current in the primary winding (assuming a single turn primary) divided by the number of turns of the secondary. For low primary currents, typically below 100 A, multiturn primary windings consisting of two or more turns may be used in order to achieve sufficient ampere-turns output to operate the secondary connected equipment.

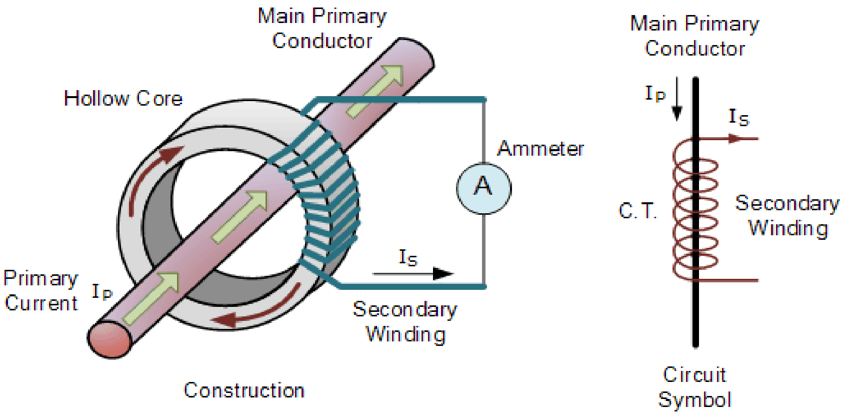


Figure 4.34. Operation of current transformer

Furthermore, it is a common practice to utilize current transformers and ammeters together as a matched pair in which the design of the current transformer is such as to provide a maximum secondary current corresponding to a full-scale deflection on the ammeter. In most current transformers an approximate inverse turns ratio exists between the two currents in the primary and secondary windings. Due to this reason, the calibration of the CT is commonly for a specific type of ammeter.

Generally, CTs are separated into two types: bushing and wound. The core of a bushing transformer is annular, while the secondary winding is insulated from the core and also there is no primary winding. In the case of a wound transformer the primary winding consists of several turns that encircle the core and in addition both, primary and secondary windings are insulated from each other and from the core. Bushing transformers have lower accuracy than the wound ones, but they are less expensive. Because of this favorable low-cost they are very often used for performing protection functions. Similarly, because of their great accuracy with low currents, wound transformers are usually applied in metering and similar applications. Another benefit of bushing transformers is their convenient placement in the bushings of power transformers and circuit breakers. This means that they take up no appreciable space in the substation. The core of bushing transformers encompasses the conductor carrying the primary current. Because of such a design, the core presents relatively large path for the establishment of electromagnetic (EM) field, necessary for the conversion of current. This is the primary reason for their lower accuracy, when compared with wound transformers. However, bushing transformers are also built with increased cross-sectional area of iron in the core. The advantage of this is the higher accuracy in scaling of fault currents that are of large multiples of nominal current, when compared to wound transformers. High accuracy for high fault currents is desirable in protective relaying. Therefore, combined with their low cost, the bushing transformers represent the best choice for protective applications.

Voltage Transformer (VT):

Like CT, Voltage Transformer (VT) is also a type of instrument transformer. It is also called as potential transformer (PT) and it's a parallel connected type of instrument transformer. VT gives an accurate representation in magnitude and phase of the voltage of the primary winding. At transmission voltages VTs are always single phase connected, whereas at distribution voltages they may have either a three-phase or a single-phase connection. Furthermore, at distribution voltages the primary winding is always star connected with its neutral point generally insulated and ungrounded. The secondary winding is usually connected in a star arrangement to provide a standard phase-to-phase secondary voltage.

VTs are separated in two types: electromagnetic voltage transformers (EVTs) and capacitor voltage transformers (CVTs). The EVT is very similar to a conventional power transformer, where the main difference is that the EVT is connected to a small and constant load. Designs of EVTs usually consider an earthed electrical shield between the high voltage (HV) and the secondary winding, for protection. More specifically, in the event of a HV breakdown, fault currents will flow to the earth via the shield rather than through the secondary winding.

The main disadvantage of the EVTs for usage at transmission voltages are their high cost and thus a more inherently reliable and less costly VT was developed, known as CVT. This concept uses one or more HV capacitor assemblies, each of which is enclosed within its own porcelain housing. The capacitors are mounted on top of each other to form a series assembly of HV capacitors. The complete assembly is usually connected on an earthed tank which consists of a capacitor and an electromagnetic (matched) transformer. The bottom capacitor forms the lower leg of a capacitor divider assembly and a voltage signal (typically 12-25 kV) is taken from the interface of the HV and LV capacitors. This signal is then fed via a reactor to a transformer which has a secondary winding giving an output of 63.5 V at the system rated voltage. CVT has two main designs: 1) the coupling-capacitor device and 2) bushing device. The first design consists of a series of capacitors (arranged in a stack), where the secondary of the transformer is taken from the last capacitor in series (called auxiliary capacitor). The second design uses capacitance bushings to produce secondary voltage at the output. Lastly, CVTs are tuned devices which means that they are highly depended on their tuned frequency and thus their accuracy and output will fall considerably when considering other frequencies.

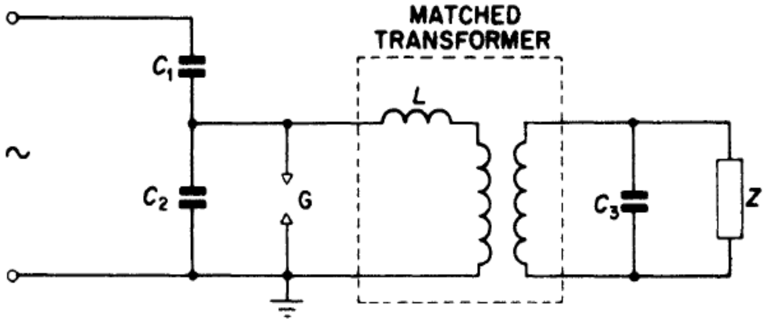


Figure 4.35. Simplified example of CVT configuration

Phasor Measurement Unit (PMU)

The phase angles of voltage phasors of the power system buses have always been of special interest to power system engineers. It is well-known that active (real) power flow in a power line is very nearly proportional to the sine of the angle difference between voltages at the two terminals of the line. As the power flow, and more specifically the flow of real power, is vital for the planning and operation of the power system, measuring angle differences across transmission has been of concern for many years.

The modern era of phasor measurement technology has its roots on the research conducted on computer relaying of transmission lines. The first prototypes of the modern "phasor measurement units" (PMUs) using the Global Positioning System (GPS) were built at Virginia Tech in early 1980s. In the last decade, synchronized phasor measurements have become the measurement technique of choice for the most of the Transmission System Operators (TSOs) worldwide. The main reason is that PMUs provide positive sequence voltage and current measurements synchronized to within a microsecond. This has been made possible by exploiting the GPS and the sampled data processing techniques developed for computer relaying applications. In addition to positive sequences voltages and currents, these systems can also measure local frequency and rate of change of frequency, and may be customized to measure harmonics, negative and zero sequence quantities, as well as individual phase voltages and currents. At present there are about 24 commercial manufacturers of phasor measurement units (PMUs), and industry standards developed in the Power System Relaying Committee of IEEE has made possible the interoperability of units from different manufacturers.

The increasing penetration of PMUs and their wide scale deployment in the transmission system is mainly due to the recent spate of spectacular blackouts on power systems throughout the world. Positive sequence measurements provide the most direct access to the state of the power system at any given instant. Many applications of these measurements have been discussed in the technical literature, especially in the area of Wide Area Monitoring and Control (WAMC) systems, and no doubt many more applications will be developed in coming years.

One of the most important features of the PMU technology compared to the conventional measurements, is that its measurements are time-stamped with high precision at the source. This has as a result for the data transmission speed to be no longer a critical parameter in making use of this data. However, the compensation of communication delays between the PMUs and the WAMC system is slowly becoming a challenging task, since they can affect the performance of the latter. All PMU measurements with the same time-stamp are used to present the state of the power system at the instant defined by the time-stamp. PMU data could arrive at a central location at different times depending upon the propagation delays of the communication channel in use and thus time-tags associated with the phasor data provide an indexing tool which helps create a coherent picture of the power system out of such data.

Fig.15 is based upon the configuration of the first PMUs built at Virginia Tech. The analog inputs represent the currents and voltages obtained from the secondary windings of the CTs and VTs. All three phase currents and voltages are used so that positive-sequence measurements can be obtained. The current and voltage signals are converted to voltages by utilizing appropriate shunts or instrument transformers so that they are matched with the requirements of the analog-to-digital converters. The sampling rate chosen for the sampling process dictates the frequency response of the anti-aliasing filters. These filters are commonly analog-type filters with a cut-off frequency less than half the sampling frequency in order to satisfy the Nyquist criterion, regarding the aliasing phenomenon.

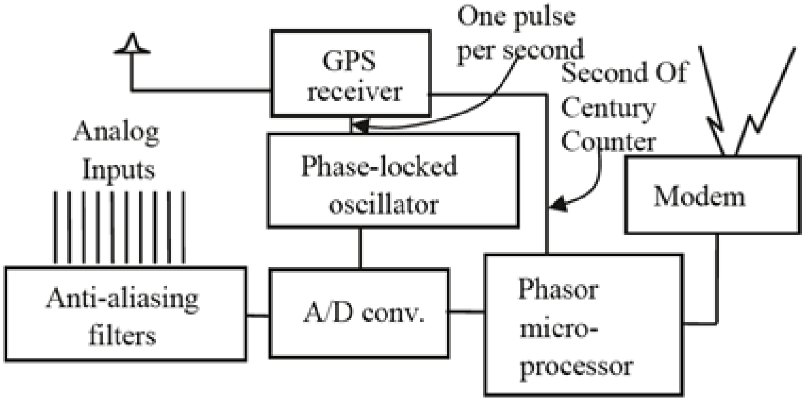


Figure 4.36. Components compromising a PMU

The sampling clock is phase-locked with the GPS clock pulse. Furthermore, sampling rates have been going up steadily over the years - starting with a rate of 12 samples per cycle of the nominal power frequency in the first PMUs and going up to 96 or 128 samples per cycle in more modern devices, as faster analog-to-digital converters and microprocessors have become available. Even higher sampling rates are expected in the future leading to more accurate phasor estimates, since higher sampling rates do lead to improved estimation accuracy. The microprocessor calculates positive-sequence estimates of all the current and voltage signals using various techniques. Other estimates of interest are frequency and rate of change of frequency measured locally, and these are also included in the output of the PMU.

The time-stamp is created by utilizing two of the signals derived from the GPS receiver. The time-stamp identifies the identity of the “universal time coordinated" (UTC) second and the instant defining the boundary of one of the power frequency periods as defined in the IEEE standard. Finally, the principal output of the PMU is the time-stamped measurement to be transferred over the communication medium through suitable modems to a higher level in the measurement system hierarchy.

PMUs represent the first level of the measurement system hierarchy. More specifically, the PMUs are installed in power system substations, where their selection depends upon the use to be made of the measurements they provide. In most applications, the phasor data are needed at locations remote from the PMUs and thus an architecture involving PMUs, communication links, and data concentrators must exist in order to realize the full benefit of the PMU measurement system. The measurements provided by the PMUs are stored in local data storage devices, which can be accessed from remote locations for post-mortem or diagnostic purposes. Due to the fact that the local storage capacity is limited, the stored data belonging to an interesting power system event can be flagged for permanent storage so that they are not overwritten when the local storage capacity is exhausted. Furthermore, the phasor data are also available for real time applications in a steady stream as soon as the measurements are made.

There may well be some local application tasks which require PMU data, in which case it can be made available locally for such tasks.

However, the main use of the real-time data is at a higher level where data from several PMUs are available. The next level of the measurement system hierarchy is commonly known as "phasor data concentrators" (PDCs). Typical functions of the PDCs are to gather and forward data from several PMUs, reject bad data, align the time-stamps, and create a coherent picture from simultaneously recorded data of the power system. There are local storage facilities in the PDCs, as well as application functions (e.g. WAMC applications) which require the PMU data available at the PDC. This can be made available by the PDCs to the applications in real time. Obviously, the communication and processing delays at the PDCs will create an additional latency in the real-time data, but all practical experience shows that this is not unachievable.

Smart Meter

The growing increase in the purchase of electric appliances has as a result for the energy demand in the households to rise. Furthermore, the inefficient use of these appliances causes a waste of energy. A quick way to achieve reduction of the residential energy demand is by simply informing the consumers in real-time time about their consumption and energy costs. By doing this, the consumers can reduce their consumption during times where the demand (and thus the energy cost) is high and move the time of energy use to off-peak times such as nighttime and weekends. By adding this feature, the consumers are becoming active players in the energy market, who can monitor and audit the actual consumed amounts as well as the values saved, as new habits are established. In the future these consumers will be able to trade energy and emissions, in more flexible ways as "prosumers". In order to have an efficient energy demand management or demand side management (DSM) whereby the energy use of different types of consumers can be adapted, the implementation of smart meters is strongly promoted by political as well as economic organizations.

A smart meter is an advanced energy meter that measures the energy consumption of a consumer and furthermore it can monitor the daily consumption pattern and provide it (along with other data) to the utility company. The latter can be very beneficial for the utility companies, which based on the daily consumption reports they can predict more accurate demand forecasts and thus they can manage the demand and the amount of the energy supply more efficiently. Another important feature is that a smart meter can communicate with other devices in order to detect potential imbalances in advance and respond quickly. Furthermore, the term smart meter often refers to an electricity meter, but it also may mean a device measuring natural gas or water consumption.

Other capabilities that this device provides are: monitor and control the household's appliances according to the desirable maximum load demand; provide the energy bill along with beneficial dynamic tariff schemes to the customer; introduce direct load control in order to decrease energy demand whenever is needed and avoid the construction of additional plants; efficient control and monitoring of the power system; and support decentralization of the grid.

DSM can help reduce peak demand and energy consumption while still allowing for the same level of comfort within the household. Key in this context could be the so-called smart appliances. These appliances are designed to work within smart grids. However, the smart meter is a necessary device for the implementation of these applications in the house. Refrigerators, freezers, washing machines, clothes dryers and dishwashers are amongst the most energy consuming appliances used in households. Smart technology can help reducing their energy use. An example of the application of smart technology is the possibility to partly or completely switch off an appliance during its run-time without any noticeable consequences for the consumer.

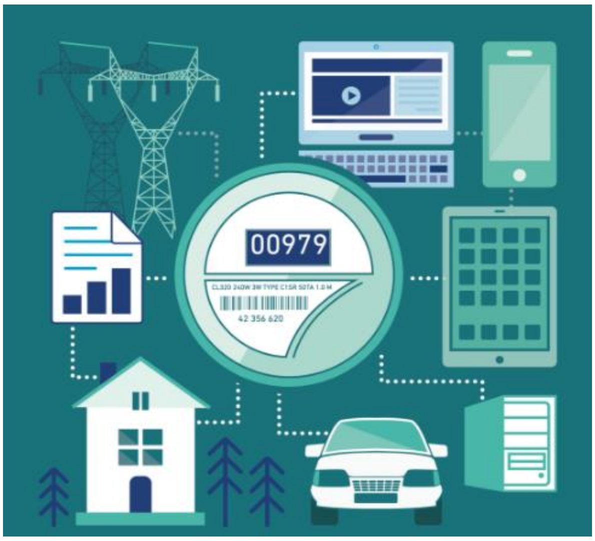


Figure 4.37. Smart meter connection and communication capabilities

The main technological problem in deploying and wide-spread the smart meter technology is communication. Each meter must be able to reliably and securely communicate the information collected to some central location. Fig. 16 presents an example of the communication capabilities of the smart meter. Considering the varying environments and locations where meters are found, that problem can be daunting. Among the solutions proposed are: the use of cell and pager networks, satellite, licensed radio, combination licensed and unlicensed radio, and power line communication. In addition, not only the medium used for communication purposes, but also the selection of the type of network used, is critical as well. Up until now the network types are separated into fixed wireless, mesh networks or a combination of the two. There are several other potential network configurations possible, including the use of Wi-Fi and other internet related networks. To date no one solution seems to be optimal for all applications. Rural utilities have very different communication problems from urban utilities or utilities located in difficult locations such as mountainous regions or areas ill-served by wireless and internet companies. In addition to communication with the head-end network, smart meters may need to be part of a Home Area Network. Technologies for this network will vary from country to country but generally include Power line communication and ZigBee.

Frequency Measurement

One of the most important variables of the power system is its frequency, which actually represents the balance between generation and demand. The earliest frequency measurement for power frequency voltages was performed by mechanical devices which employed mechanical resonators tuned to a range of frequencies around the nominal power frequency. An example of a frequency meter from mid-1950s is shown in Fig. 4.38 (a). Another frequency measuring instrument of about the same period is a resonance-type device. This device utilizes tuned resonant circuits at different frequencies, which are energized by the secondary voltage obtained from a voltage transformer, and thus the circuit which is in resonance provides the frequency measurement (Fig. 4.38 (b)). Typical resolution of these meters was of the order of 0.25 Hz.

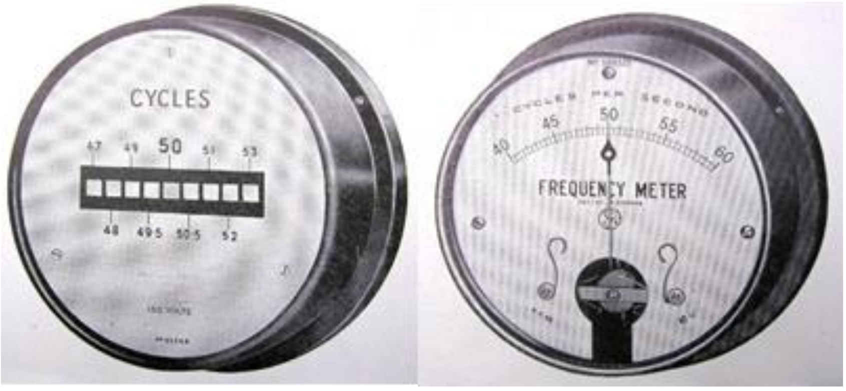


Figure 4.38. (a) 50 Hz mechanical resonance-type frequency meter (b) 50 Hz electrical resonance-type frequency meter

The next advance in frequency measurement came with the introduction of precise time measurement techniques. More specifically, by measuring the time interval between consecutive zero crossings of the voltage waveform, the frequency of the voltage could be determined. Clearly the accuracy of such a measurement is directly depended upon the precision of time measurement, as well as on the detection accuracy of the waveform's zero-crossings. Especially the zero-crossing detection can be challenging as it is affected by the presence of noise in the measurement, varying harmonic frequencies and levels, and the performance of the zero-crossing detector circuits.

Nowadays, new technologies have been developed to provide more accurate frequency measurements which can eliminate many of the aforementioned error sources:

- Frequency counter: A frequency counter is a digital instrument that can measure and display the frequency of any periodic waveform. The signal waveform whose frequency is to be measured is converted into trigger pulses and applied continuously to one of the terminals of an AND gate. In addition, a pulse of duration of 1 sec is applied to the other terminal of the gate, and thus the number of pulses counted at the output terminal indicate the frequency of the signal. The accuracy of a frequency counter is strongly dependent on the stability of its timebase. A timebase is very sensitive like the hands of a watch, and can be changed by movement, interference, or even drift due to age, meaning it might not "tick" correctly. This has as a result for a frequency reading, when referenced to the timebase, seem higher or lower than the actual value. A solution for higher accuracy measurements, is to utilize an external frequency reference tied to a very high stability oscillator such as a GPS disciplined rubidium oscillator.

- Frequency Disturbance Recorder: Typically, only static frequency measurements are widely available in the power system. This is due to the fact that most of frequency measurement devices assume a single system frequency, and thus they use long periods of data averaging in order to achieve good estimation accuracy. This is not a problem when the system is in its steady state. However, the power system is almost never in steady state and thus most of the frequency data (and also the most valuable data) are obtained during small or large disturbances, when the system frequency is time varying, and when frequencies could be very different in various areas of the system. Frequency monitoring network (FNET) is a wide area power system frequency measurement system, which deploys a type of PMU, known as frequency disturbance recorders (FDR) in order to measure the power system frequency accurately. More specifically, FDR is a GPS-synchronized single-phase PMU that is installed at ordinary 120 V outlets. Because the voltages involved are much lower than those of a typical three-phase PMU, the device is relatively inexpensive, it has high accuracy and it is simple to install. Just like a PMU, each FDR measures a voltage signal and determines the phase angle, amplitude, and frequency of the waveform. Precise timing across a wide area is made possible through the use of the GPS, which as it was mentioned previously, it allows the synchronization of the measurements. The FNET system is used to monitor the changing frequency in continuous time and in different locations. Dynamic measurement accuracy is critical and the frequency estimation algorithms developed for the FNET system have virtually zero algorithm error in the 52-70Hz range.

4.2.4. Microprocessor means of automation

Automatic Voltage Regulator

The automatic voltage regulator is used to regulate the voltage. It takes the fluctuate voltage and changes them into a constant voltage. The fluctuation in the voltage mainly occurs due to the variation in load on the supply system. The variation in voltage damages the equipment of the power system.

The variation in the voltage can be controlled by installing the voltage control equipment at several places likes near the transformers, generator, feeders, etc., The voltage regulator is provided in more than one point in the power system for controlling the voltage variations.

In DC supply system the voltage can be controlled by using over compound generators in case of feeders of equal length, but in the case of feeders of different lengths the voltage at the end of each feeder is kept constant using feeder booster. In AC system the voltage can be controlled by using the various methods likes booster transformers, induction regulators, shunt condensers, etc.

The automatic voltage regulator works on the principle of detection of errors. The output voltage of an AC generator obtained through a potential transformer and then it is rectified, filtered and compared with a reference. The difference between the actual voltage and the reference voltage is known as the error voltage**.** This error voltage is amplified by an amplifier and then supplied to the main exciter or pilot exciter.

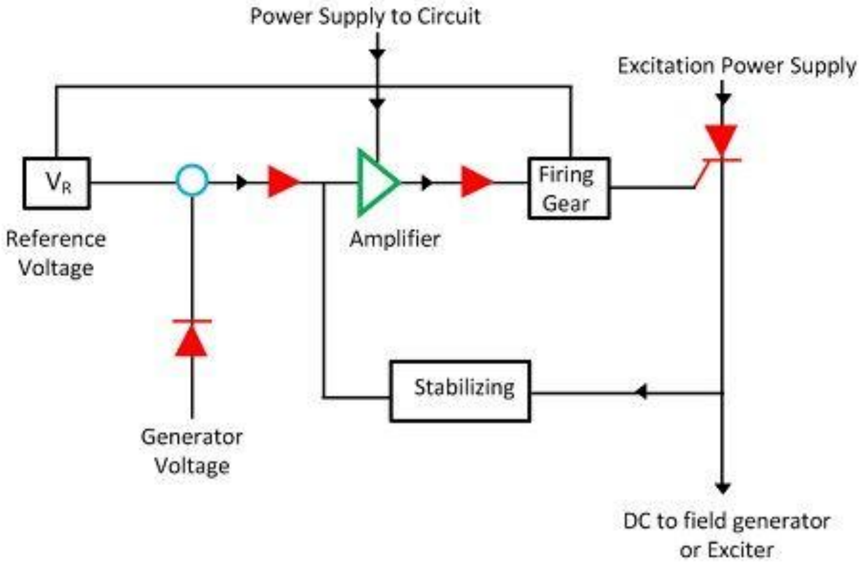


Figure 4.39. The automatic voltage regulator

Thus, the amplified error signals control the excitation of the main or pilot exciter through a buck or a boost action (i.e. controls the fluctuation of the voltage). Exciter output control leads to the controls of the main alternator terminal voltage.

The main functions of an AVR are as follows.

1. It controls the voltage of the system and has the operation of the machine nearer to the steady state stability.

2. It divides the reactive load between the alternators operating in parallel.

3. The automatic voltage regulators reduce the overvoltages which occur because of the sudden loss of load on the system.

4. It increases the excitation of the system under fault conditions so that the maximum synchronising power exists at the time of clearance of the fault.

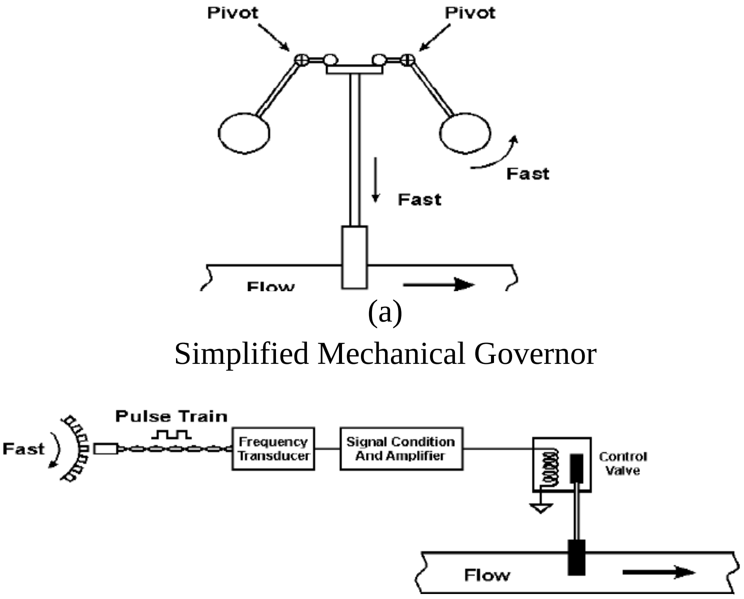
When there is a sudden change in load in the alternator, there should be a change in the excitation system to provide the same voltage under the new load condition. This can be done by the help of the automatic voltage regulator. The automatic voltage regulator equipment operates in the exciter field and changes the exciter output voltage, and the field current. During the violent fluctuation, the ARV does not give a quick response.

For getting the quick response, the quick acting voltage regulators based on the overshooting the mark principle are used. In overshoot mark principle, when the load increase the excitation of the system also increase. Before the voltage increase to the value corresponding to the increased excitation, the regulator reduces the excitation of the proper value.

Speed governors

Speed governors vary prime mover output(torque) automatically for changes in system speed (frequency). The speed sensing device is usually a flyball assembly for mechanical-hydraulic governors and a frequency transducer for electro-hydraulic governors. The output of the speed sensor passes through signal conditioning and amplification (provided by a combination of mechanical-hydraulic elements, electronic circuits, and/or software) and operates a control mechanism to adjust the prime mover output (torque) until the system frequency change is arrested. The governor action arrests the drop in frequency, but does not return the frequency to the pre-upset value (approximately 60 Hz) on large interconnected systems. Returning the frequency to 60 Hz is the job of the AGC (Automatic Generation Control) system. The rate and magnitude of the governor response to a speed change can be tuned for the characteristics of the generator that the governor controls and the power system to which it is connected.

Simplified schematics of mechanical and electronic speed governing systems are shown in Fig. 4.40. If a decrease in system frequency occurs, due to a loss of generation or an increase in load, the shaft speed of each connected synchronous generator will also decrease. This speed decrease is transmitted to a mechanical governor flyball assembly by means of a shaft-mounted PMG (permanent magnet generator) and a ballhead motor, and to an electro-hydraulic governor frequency transducer by a toothed wheel or the generator potential transformers. As the flyballs spin more slowly, they move in causing the valve in Fig. 4.40 (a) to move up and allow more flow (fuel, steam, water, etc.) to the prime mover. In the same manner, the frequency decrease sensed by a frequency transducer will be amplified and used to open the valve in Fig. 4.40(b). Thus, the output power (torque) of the controlled prime mover will increase and help arrest the frequency drop.



(a)

Simplified Mechanical Governor

(b)

Simplified Electro Hydraulic Governor

Figure 4.40. Simplified governor schematics

Flexible Alternating Current Transmission System (FACTS)

Power System Stability is one of the most important topics at transmission level. Power systems are being pushed closer to their stability and thermal limits and in addition they are subjected to various types of disturbances that induce oscillations and which may even lead to instability. In order to protect the transmission system, engineers came up with developing power electronic based devices that maximize the AC power transmission. These devices are called Flexible AC Transmission Systems (FACTS).

FACTS were introduced in transmission lines with the main goal of operating them up to their thermal limits. Without FACTS in the transmission system, the lines still transmit lots of power but with a great percentage of reactive power. Reactive power is the main issue that does not allow active power flow to increase. Thus, less active power can flow across the line since a great amount of reactive power is also flowing and increasing the line's temperature. When lines are overloaded, FACTS can modify their apparent impedance in order to allow these lines to pass more power. This can give more time for the system operator to control generators and loads to come back to normal condition. In addition, FACTS can also improve considerably the stability of the system, voltage control and power oscillation damping. To do all these, they are composed by reactive elements in order to modify the line reactance and thus compensate for the excess of reactive power that does not allow the active power to increase.

When referring to FACTS devices, it is important to explain the terms of 'dynamic' and 'static'. The term 'dynamic' is used to express the fast controllability of FACTS-devices provided by the power electronics. This is one of the main differentiation factors from the conventional devices. The term 'static' means that the devices have no moving parts like mechanical switches to perform the dynamic controllability. Therefore, most of the FACTS devices can equally be static and dynamic.

Furthermore by utilizing FACTS in the transmission system, its operation improves with minimal infrastructure investment, lower environmental impact, and less implementation time compared to the construction of new transmission lines, offering utilities and industry the ability to: (i) dynamically control the power flows on specific transmission and distribution routes, (ii) allow secure loading of transmission and distribution lines to their full thermal capacity, and (iii) improve power quality.

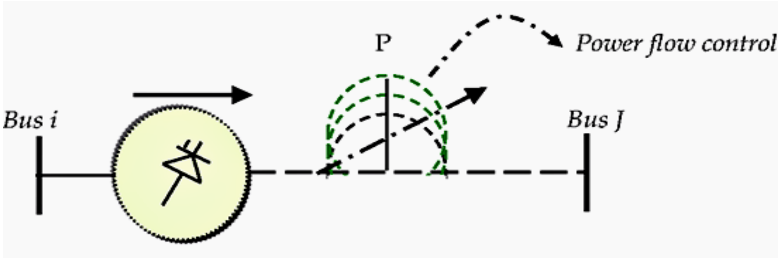


Figure 4.41. Series FACTS controller

FACTS controllers can be divided into three categories: series controllers, shunt controllers and combined series-shunt controllers. In series controllers, the line impedance is modified, which means that the net impedance is decreased and thereby increasing the transmittable active power. More specifically, the series controller is a variable impedance, such as capacitor, reactor or a power electronics based variable voltage source. A series controller generates reactive power that in a self-regulating manner balances a fraction of the line's transfer reactance. The result is that the line is electrically shortened, which improves angular stability, voltage stability and power sharing between parallel lines. In principle, all series controllers inject a voltage in series with the line. As long as the injected voltage is in phase quadrature with the line current, then the series controller can only supply or consume variable reactive power. Typical types of series controllers are the static synchronous series capacitor (SSC), thyristor controlled series capacitor (TCSC), thyristor controlled series reactor (TCSR), thyristor switched series capacitor (TSSC) and thyristor switched series reactor (TSSR).

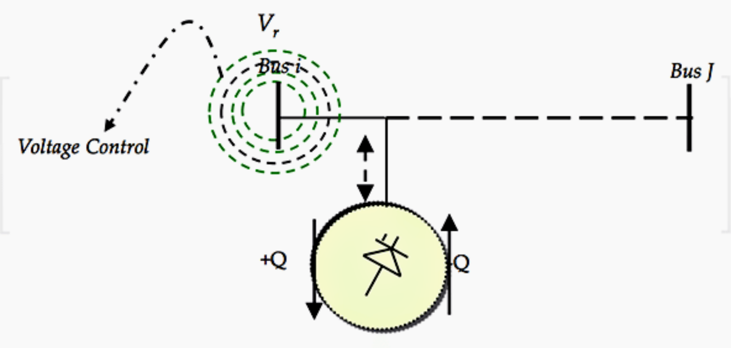


Figure 4.42. Shunt FACTS controller

As in the case of the series controller, the shunt controller may be a variable impedance (reactor or capacitor), variable source, or a combination of these. For shunt controllers, reactive current is injected into the line in order to regulate the voltage at the point of connection. Even a variable shunt impedance connected to the line voltage causes a variable current flow and hence represents injection of shunt current into the line. As long as the injected phase current is in quadrature with the line voltage, the shunt controller only supplies or consumes variable reactive power. Any other phase relationship results to the injection of real power as well.

Furthermore, this category includes the most known and commonly used FACTS, the Static synchronous compensator (STATCOM) and the Static VAR compensator (SVC).

In the case of the combined series-shunt controller, a combination of shunt and series controllers is considered, which are controlled in a coordinated manner in the case of a multiline transmission system. In principle, combined series-shunt controllers inject current to the system with the shunt part of the controller and voltage in series in the line with the series part of the controller. However, when the shunt and series controllers are unified, they can also ex

change real power between the lines via the common DC link. As expected their combination offer several advantages over the other FACTS, such as better voltage regulation, more flexibility and controllability, better damping of the system's oscillations, more secure loading of the transmission lines close to their thermal limits, etc.

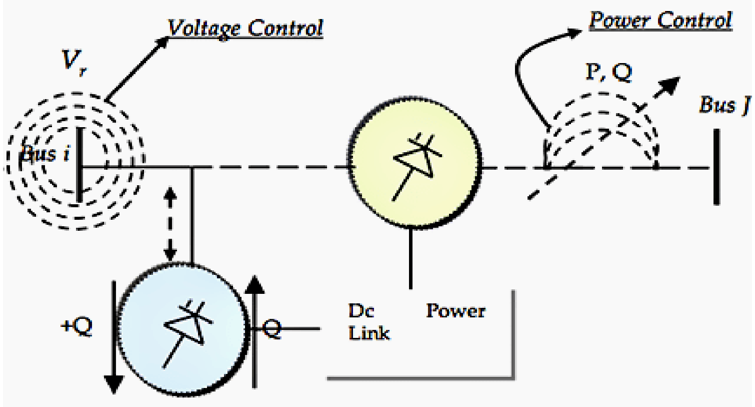


Figure 4.43. Combined series-shunt controller

Following a more analytic description of the most known shunt (SVC, STATCOM) and series (TCSC) FACTS is presented:

- SVC: As mentioned, the SVC is a FACTS fully dedicated to voltage support. More specifically, it is an electrical device that provides fast-acting reactive power on high voltage electric transmission networks. SVCs operate the electronic soft switching of their own shunt reactors and/or capacitors in order to achieve a fast and continuous reactive power variation. Compared to the typical capacitor banks, which are commonly used for reactive power compensation, the SVC offers a near-instantaneous response, it has a higher capacity, it is faster and more reliable but on the other hand it is much more expensive. In addition, they are also used in HVDC converter stations, where the fast control of voltage and reactive power flow are required. The main benefits of equipping the system with an SVC are the increased stability limits and the more controllable voltage profile. This results in a more robust system with enhanced capacity, flexibility and predictability. The SVC main control features can be resumed as voltage control, reactive power control, damping of power oscillations and unbalance control. The compensation along the electrical lines requires a midpoint dynamic shunt. With a dynamic compensator at the midpoint, the symmetrical line behavior is achieved. The midpoint voltage will vary with the load, and an adjustable midpoint susceptance is a way to maintain constant voltage magnitude, therefore, the advantage of utilizing SVCs is evident. In principle the SVC consists of Thyristor Switched Capacitors (TSC) and Thyristor Switched or Controlled Reactors (TSR / TCR).

- STATCOM: It is an SVC with voltage source converter, which has characteristics similar to the synchronous condenser. Its capacitive of inductive output current can be controlled independent of the ac system voltage. As an electronic device, it has no inertia and many advantages such as the better dynamics, lower investment cost, and lower operating and maintenance costs. It can be considered as a voltage source behind a reactance, which provides reactive power generation as well as absorption purely by means of electronic processing of voltage and current waveforms in a voltage source converter. This means that capacitor banks and shunt reactors are not needed for generation and absorption of reactive power. The advantage of a STATCOM is that the reactive power provision is independent from the actual voltage on the connection point. A STATCOM is consisted by thyristors with turn-off capability like GTO or to-day IGCT or with more and more IGBTs. The next step in

STATCOM development is the combination with energy storages on the DC-side. The performance for power quality and balanced network operation can be improved much more with the combination of active and reactive power.

- TCSC: It is an electrical device that works as a controllable voltage source. It is used on long lines to shorten them indirectly by changing the impedance of the line. Moreover, it provides damping of low frequency electromechanical oscillations. TCSC addresses two specific dynamical problems in transmission systems. Firstly, it increases damping when large electrical systems are interconnected, by providing strong damping torque on inter-area electromechanical oscillations. Secondly it can overcome the problem of Sub-Synchronous Resonance (SSR), a phenomenon that involves an interaction between large thermal generating units and series compensated transmission systems. The main features of the TCSC are: reduction of the line voltage drop, limitation of the load dependent voltage drops, influence the load flow in parallel transmission lines by controlling the current, increase the power transfer capability, reduction of the transmission angle and increase the system stability. From a principal technology point of view, the TCSC resembles the conventional series capacitor. All the power equipment is located on an isolated steel platform, including the thyristor valve that is used to control the behavior of the main capacitor bank. Likewise, the control and protection is located on ground potential together with other auxiliary systems. There are two main principles supporting TCSC technology. First, the TCSC provides electromechanical damping between large electrical systems by modulating the reactance of one or more specific interconnecting power lines, offering that way a variable capacitive reactance. Second, the TCSC can change its apparent impedance for subsynchronous frequencies in such a way that a potential SSR is avoided. The TCSC achieves both objectives by using control algorithms that work concurrently. The controls will function on the thyristor circuit (this in parallel to the main capacitor bank) such that controlled charges are added to the main capacitor, making it a variable capacitor at fundamental frequency but a "virtual inductor" at subsynchronous frequencies.

References

1. Akyildiz I, Melodia T, Chowdhury K., 2007. A survey on wireless multimedia sensor networks. Computer Networks 51(4):921–960
2. Akyildiz I.F., W. Su, Y. Sankarasubramaniam, E. Cayirci, 2002. Wireless sensor networks: a survey, Computer Networks 38 (2002) 393–422
3. ANT technology. http://www.thisisant.com/technology
4. Assanovich B.A., Kiseleva N.N., 2006. Projecting in the environment MatLAB. YKSUG. Grodno.
5. Bittner M., Widmer H., Pajot A., Alberdi G., Hohl H., Kmethy G., 2009. “Open Public Extended Network metering”, Project Funded by the European Commission under the 7th Framework Program.
6. Bonaventure O., 2011. Computer Networking: Principles, Protocols and Practice, Release 0.25, October 30, 2011.
7. Bluetooth. Bluetooth core specification v4.0. Specification/adopted-specifications
8. Bulusu N, Estrin D, L. Girod, J. Heidemann, 2001. Scalable coordination for wireless sensor networks: self-configuring localization systems, International Symposium on Communication Theory and Applications (ISCTA 2001), Ambleside, UK, July 2001.
9. Computer Networks 2012. A Systems Approach. Copyright © 2012 Elsevier, Inc. DOI: 10.1016/B978-0-12-385059-1.00001-6.
10. Delsing J., 2017. IoT based Automation - made possible by Arrowhead Framework, p.401, (in press).
11. EnOcean. [Accessed in November 2016, http://www.enocean.com/en/enocean-wireless-standard/]
12. Forouzan B. A., 2007. Data Communications and Networking 4th ed. McGraw-Hill.
13. Harvard, 2012. Sensor Networks Lab. Volcano monitoring. Available online: http://fiji.eecs.harvard.edu/Volcano.
14. Hekmat C., 2005. Communication Networks, PragSoft Corporation, www.pragsoft.com, p.198.
15. Hoblos G, Staroswiecki M, A. Aitouche, 2000. Optimal design of fault tolerant sensor networks, IEEE International Conference on Control Applications, Anchorage, AK, September 2000, pp. 467–472.
16. IEEE 802.15.4-2006 standard for information technology part 15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low rate wireless personal area networks (LRWPANs).

17. Kim A, Hekland F, Petersen S, Doyle P., 2008. When HART goes wireless: understanding and implementing the WirelessHART standard. In: IEEE international conference on emerging technologies and factory automation, pp 899–907.

18. Kmethy G., P. Fuchs, V. Varjú, B. Roelofsen, 2015. "IEC 62056 DLMS/COSEM seminar", DLMS Seminar EUW November 2015 Vienna.

19. Li M, Liu Y., 2007. Underground structure monitoring with wireless sensor networks. In: Proceedings of the 6th international conference on information processing in sensor networks. ACM, New York, p 78.

20. Montenegro G, Kushalnagar N, Hui J, Culler D., 2007. Transmission of IPv6 packets over IEEE 802.15.4 networks. Internet proposed standard RFC 4944

21. Optical Network Design and Transport 101, [accessed in December 2016, http://searchtelecom.techtarget.com/definition/passive-optical-network]

22. Peterson L.L., Davie B. S., 2012.Computer Networks: A Systems Approach**.** 5th ed. Morgan Kaufmann Ser.

23. PSCES, 2012. Fundamentals of telecommunications, PSCES Project, The Abdus Salam International Centre of Theoretical Physics, [Accessed in November 2016, http://wtkit.org/]

24. Porcino D, Hirt W., 2003. Ultra-wideband radio technology: potential and challenges ahead. IEEE Commun Mag 41(7):66–74

25. Rawat P., Singh K. D., H. Chaouchi, J. M. Bonnin, 2013. Wireless sensor networks: a survey on recent developments and potential synergies Springer Science+Business Media New York 2013.

26. Resources and analysis for electronics engineers. [accessed in December 2016, http://www.radio-electronics.com]

27. Rodrigues J.J., Neves P.A., 2010. A survey on IP-based wireless sensor network solutions. Int J Commun Syst 23(8):963–981.

28. Roger L. Freeman, 2004. Telecommunication System Engineering Fourth Edition, John Wiley & Sons, Inc., Hoboken, New Jersey, pp.1-1008.

29. Ribickis L.,Kuņicina N.,. Čaiko J., Agafonovs J., 2011. Industriālo datortīklu pamati, Riga Technical University, Riga, 2011, p.64.

30. Shih E, S. Cho, N. Ickes, R. Min, A. Sinha, A. Wang, A. Chandrakasan, 2001. Physical layer driven protocol and algorithm design for energy-efficient wireless sensor networks, Proceedings of ACM MobiCom'01, Rome, Italy, July 2001, pp. 272–286.

31. Sklar B., 2001. Digital Communications: Fundamentals and Applications. 2nd ed. Prentice-Hall.

32. VigilNet. [Accessed in November 2016, http://www.cs.virginia.edu/wsn/vigilnet/]

33. Wireless systems for industrial automation: process control and related applications. ISA-100.11a-2009

34. Yick J, Mukherjee B, Ghosal D., 2008. Wireless sensor network survey. Computer Networks 52(12):2292–2330

35. Z-Wave Alliance. [Accessed online December 2016: http://www.z-wavealliance.com/technology/]