OST
Ostschweizer
Fachhochschule

Wireless Communications

Prof. Dr. Heinz Mathis
Prof. Dr. Hans-Dieter Lang

2021/2022

”You see, wire telegraph is a kind of a very, very long cat. You pull his tail in New York and his head is meowing in Los Angeles. Do you understand this? And radio operates exactly the same way: you send signals here, they receive them there. The only difference is that there is no cat.”

Albert Einstein, when asked to describe radio

Preface

Historically, the term *Mobile Communications* was used to classify radio equipment that can be moved during operation. In recent times, the term has been adopted for equipment that is in use in the framework of a standardized wireless system network. For many years we have used the word in either meaning, particularly in the sense that the fundamentals in this course equally apply to standardized systems as well as to any proprietary RF link established. However, due to an increasing number of wireless communication systems that run on the same physical foundations, but not within the classical framework of a mobile system, the more generic expression *Wireless Communications* was introduced in 2015 for the lecture that had been taught for the previous 13 years at HSR.

This course is geared towards tuition in electrical engineering rather than computer science, so it does not cover aspects of higher-layer communications of a wireless network. Speaking in terms of layers, it concentrates on the lowest layer (physical layer). Rather than focussing on current standards, it provides the necessary prerequisites an electrical engineer should have when designing and building equipment as part of these technologies. Instead of detailed derivations and proofs of fundamental theory, more so-called back-of-the-envelope calculations are provided. So the first part is devoted entirely to the fundamentals in coding theory, RF design, antennas and propagation, and receiver architectures. The second part is structured into different applications. Rather than extensively covering wireless standards, introductory material is provided followed by selected topics that pose particular challenges to RF and DSP designers. The goal of the course is not to give the engineer a wide detailed knowledge of ever-changing standards but to equip him/her with the necessary basics to do work within any present or future wireless system.

Acknowledgment

Creating figures is the most time-consuming job, and sometimes not particularly taxing, especially when good figures have already been created. We would like to thank Prof. Dr. W. Bächtold from ETH Zurich, Prof. Dr. Simon Saunders from University of Surrey, U.K., Dr. Heinz Ochsner, Dr. Hanspeter Schmid, and Dr. Sigi Wyrsch for their generous allowance to use any figures from their respective lecture notes.

We would also like to thank my assistants, Marcel Kluser and Daniel Megnet, Mischa Sabathy, Patrick Fleischmann, Roman Gassmann, Nicola Ramagnano, Matthias Meienhofer, Michel Nyffenegger, and Nicolas Jost for preparing tables and figures in various chapters. Last but not least, Dieter Ehrismann has put many hours into proof-reading the lecture notes. Many thanks to him for his feedback.

Contents

I Fundamentals

II Applications

Part I

Fundamentals

1 Introduction

Wireless systems are ubiquitous these days. Whereas the transmission of electromagnetic waves from one point to another can have many purposes such as the transportation of energy, the initiation of a physical effect, the sending of information, the measurement of physical dimensions (e.g., RADAR, GPS) and many more, wireless is usually associated with information exchange. Although there is the verb *to wire*, established at the end of the 19th century to express the sending of information by means of a wired telegraph, no such equivalent exists for the wireless transmission process. Still, in the early days of radio broadcast, a receiver set was often referred to as a *wireless*.

1.1 History

Although Maxwell was one of the first who paved the way for wireless communications by stating the relationship between electric and magnetic fields, the practical breakthroughs in the fields are usually credited to Heinrich Hertz (1857–1894) and Guglielmo Marconi (1874–1937). Whereas Hertz did all his experiments in his lab, Marconi was the one who went out and tried to apply wireless radio over real useful distances.He was also the one who succeeded in the first transatlantic transmission. The problem was that all initial experiments were based on a spark transmitter, which generated a signal that was very hard to receive. Eventually, people started to transmit sinusoidal signals, which could be received by a variety of receivers, some of which will be reviewed later in Chapter 6. More details of the history of radio can be deferred from [37].

The history of wireless radio is embedded in a wider story of communication technology, a short overview of which shall be given in Table 1.1, which was partly taken from [85].

Year	**Accomplishment**
- 3000	People of Egypt develop hieroglyphs.
800	Arabians adopt numerical system from India.
1440	Johannes Gutenberg invents the printing press.
1752	Benjamin Franklin proves flashes to be electricity.
1827	Georg Simon Ohm states: $U = R \cdot I$.
1834	Carl F. Gauss and Ernst H. Weber build the electromagnetic telegraph.
1850	Gustav Robert Kirchhoff publishes the current laws.
1858	The first transatlantic cable is being laid and fails after only 26 days.
1864	James C. Maxwell predicts electromagnetic waves.
1876	Alexander Graham Bell develops and patents the first telephone.
1883	Thomas A. Edison discovers the flow of electrons in a vacuum.
1887	Heinrich Hertz confirms Maxwell's theory.
1894	Oliver Lodge demonstrates the wireless transmission over a distance of 140 meters.
1900	Guglielmo Marconi transmits the first wireless signal over the Atlantic ocean.
1905	Reginald Fessenden transmits voice and music using radio waves.
1906	Lee de Forest invents an amplifier based on the vacuum tube.
1920	The first radio programme is being broadcast in Pittsburgh (USA).
1920	J. R. Carson uses sampling in communication systems.
1923	Vladimir K. Zworkykin builds a television recording tube.
1926	J. L. Baird and C. F. Jenkins demonstrate television.
1927	Harold Black develops an amplifier with negative feedback.
1933	Edwin H. Armstrong invents frequency modulation (FM).
1935	The first practical radar is being developed by Robert A. Watson-Watt.
1936	England starts broadcasting television programmes.
1937	Alex Reeves develops pulse code modulation (PCM).
1947	Shockley, Brattain, and Bardeen invent the transistor.
1948	Claude E. Shannon publishes his work about information theory.
1950	Time multiplex is being used in telephony.
1950	Microwave transmission systems are being developed.
1953	Color television is introduced in the States.
1953	The first working transatlantic cable is being laid.
1957	The first satellite (Sputnik I) orbits the earth.
1958	A. L. Schawlow and C. H. Townes publish the basics of lasers.
1958	Jack Kilby builds the first integrated circuit (IC).
1961	Stereo-FM broadcasting starts.
1962	The first active satellite (Telstar I) transmits television signals.
1963	Dual-tone dialing is introduced.
1964	Error-correcting codes and adaptive equalization are being developed.
1965	The first commercial satellite (Early Bird) starts operation.
1968	Cable television (CATV) is being developed.
1971	The first microprocessor (4004) is developed by Intel.
1972	The first mobile telephone is demonstrated.
1976	The PC is being developed.
1980	Communications by fiber optics starts.
1980	The compact disc (CD) is developed jointly by Philips and Sony.
1990	A new era with digital signal processing on microprocessors and DSPs starts.
1992	Digital wireless telephony (GSM) is being launched.
2002	UMTS is starting operation. First handsets available.

Table 1.1 History of communication technology.

1.2 Disciplines

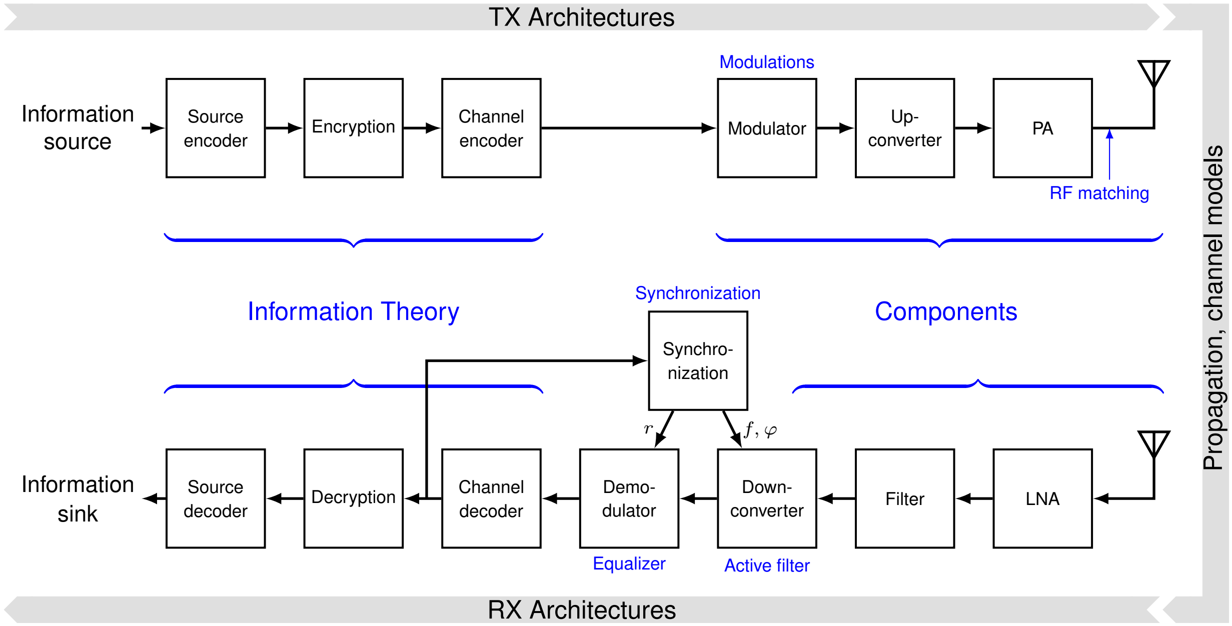


Figure 1.1 Disciplines involved in the design of a wireless communication system.

Every book on communication systems has a so-called *Figure 1*. It will always describe the chain from the information source to the information sink. We have made use of this figure here (although not number 1, but Fig. 1.1) to illustrate the multiplicity of disciplines involved in the design of such a system and thus in this lecture. The disciplines are the following:

- Information theory
- Modulation theory
- RF matching (RF design)
- RF components
- Transmitter and receiver architectures
- Antennas and propagation
- Synchronization and equalization

1.3 OSI model

Almost every network of a kind nowadays is structured into the layers of the OSI (Open System Interconnection) reference model. For reference, the seven layers are displayed in Table 1.2. Of special meaning to our lectures are the lower two layers. In wireless communications, the physical layer is often designated PHY. It corresponds to the radio front end and baseband signal processing. The lower half of the data link

Layer	Name
7	Application
6	Presentation
5	Session
4	Transport
3	Network
2	Data link
1	Physical

Table 1.2 OSI layers.

layer is called MAC (medium access channel) and defines how the medium is accessed. The protocol as to how error correction is handled, for example, is part of the MAC.

1.4 Distance range of systems

The range of commercial systems that are covered by wireless systems is displayed in Fig. 1.2 and varies from as low as 1 bit/s to almost 100 Mbit/s on the data rate, and from nearer than 1 m to as much as almost 10^5 km. Deep-space exploration, of course, employs wireless transmission over far longer distances. These are, however, not commercial services and thus excluded from the figure. The current data rates and distances are also shown in Fig. 1.2.

Name	Data rate	Distance
Bluetooth	1-3 Mbit/s	<100 m
Bluetooth LE	1 Mbit/s	≥10 m
DAB (via satellite)	3 Mbit/s	36’000 km
DCF77	1 bit/s	2000 km
DECT	2 Mbit/s	300 m
DVB-S	>70 Mbit/s	36’000 km
EDGE	384 kbit/s	20–35 km
GPS	50 bit/s	20’200 km
GPRS	171.2 kbit/s	20 km
GSM	9.6 kbit/s	20–35 km
HSCSD	57.6 kbit/s	20 km
RFID	10 kbit/s	1 m
UMTS	2 Mbit/s	20 km
LTE	≤300 Mbit/s	0.01-100 km
LTE-Advanced	≤1 Gbit/s	1-5 km
WLAN 802.11	1-2 Mbit/s	100 m
WLAN 802.11a	54 Mbit/s	100 m
WLAN 802.11b	11 Mbit/s	100 m
WLAN 802.11n	600 Mbit/s	70 m
WLAN 802.11ac	≤1.33 Gbit/s	70 m
WLAN 802.11ad	≤6.9 Gbit/s	10 m

Figure 1.2 Distances and data rates of popular wireless systems.

2 RF Basics

2.1 Introduction

RF circuit design is different from conventional circuit design in many ways. The direct translation of a schematic, element by element in its ideal assumption no longer works. A capacitor is not a simple capacitance, and therefore has to be modeled including parasitics as a result of its physical implementation such as connection leads. In a conventional, e.g. audio-related, design we are often expressing signals in voltage mode or in current mode. Between the two modes we know elements of transformation, e.g., voltage-controlled current amplifier. In RF design most elements work in power mode. There cannot be just voltage, or current for that matter. We need to provide the right impedance to a circuit, or else the power gets reflected and thus lost.

2.2 Parasitics

At high frequencies, components, transmission lines, and even simple things such as wires very often behave differently from what would be expected from an ideal assumption. The self-inductance of a wire is approximately given by [28]

$$L = 0.002\,l \left(\ln \frac{4l}{d} - 0.75 \right) \quad [\mu\text{H}], \tag{2.1}$$

where l and d are the length and diameter of the wire, respectively, both in cm. What at low frequencies might go under parasitic effects, may suddenly dominate when going to higher frequencies. Resistors, e.g., have to be modeled as a series of an inductance and a resistance in parallel to a capacitance, see Fig. 2.1. The equivalent circuits of a capacitor and an inductor look alike. They differ, of course, in the values of their respective elements. The equivalent circuit of transmission lines, cables, etc., will be discussed later. It is very important to take into account the parasitics of lumped elements when designing filters and the like. Fig. 2.2 shows what happens to a low-pass filter at high frequencies, when the parasitics of the lumped elements come into play. A low-pass filter may eventually end up as a high-pass filter, as (c) shows. Fig. 2.3 clearly shows the low-pass characteristics but also the band-pass and high-pass characteristics of the parasitic components.

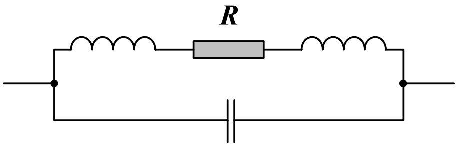
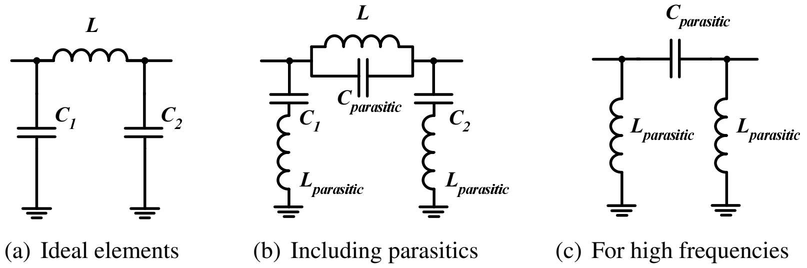


Figure 2.1 Equivalent circuit of a resistor.



(a) Ideal elements (b) Including parasitics (c) For high frequencies

Figure 2.2 Low-pass filter arrangements.

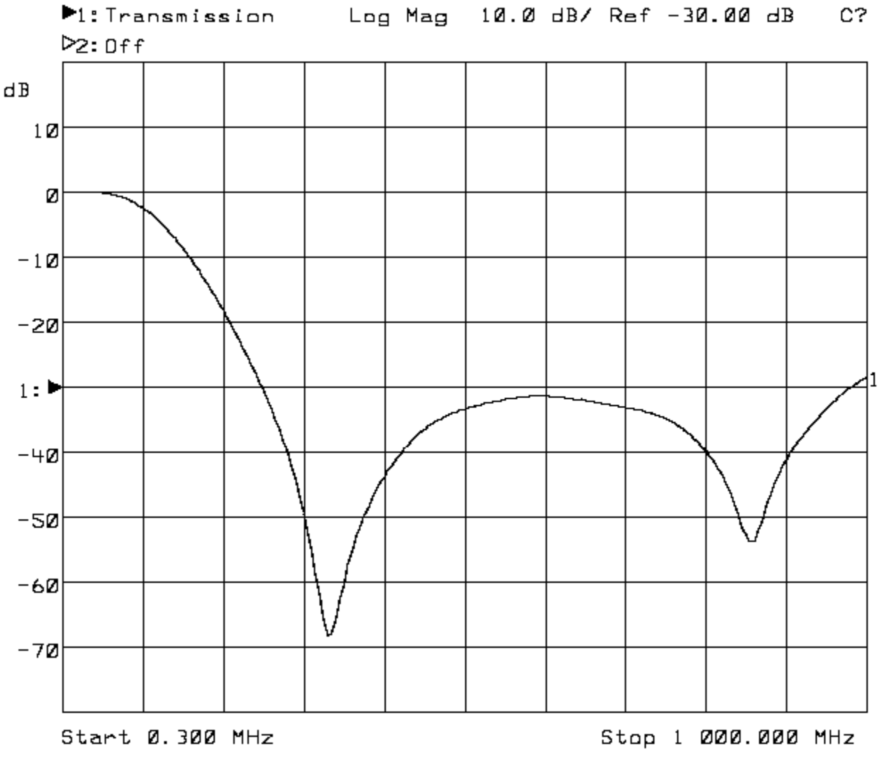


Figure 2.3 Low-pass filter transfer function.

2.3 Transmission Lines

2.3.1 Motivation

Although transmission lines may sound as being out of scope in a book on wireless communications, this is not the case. First, we always have to carry the signal to and from the antenna, hence, we need some kind of wires. Even if the antenna is mounted directly on the PCB, transmission line theory is heavily needed for the correct design of microstrip or coplanar 50-Ω matched transmission lines. Second, rather than using discrete components to match different components to each others' impedances, we can use an appropriate network of short transmission lines to do the matching, particularly at higher frequencies. This is more important for circuits working at 1 GHz and higher, where parasitics of discrete components start to take more dominant roles. Third, whole filter structures are possible using microstrip technology, a technique dependent on the comprehensive understanding of transmission line theory.

2.3.2 Frequency Ranges

Cables or transmission lines are not equally suited over the entire frequency range in technical use. Fig. 2.4 shows the frequency ranges and the most suitable corresponding means of transportation.

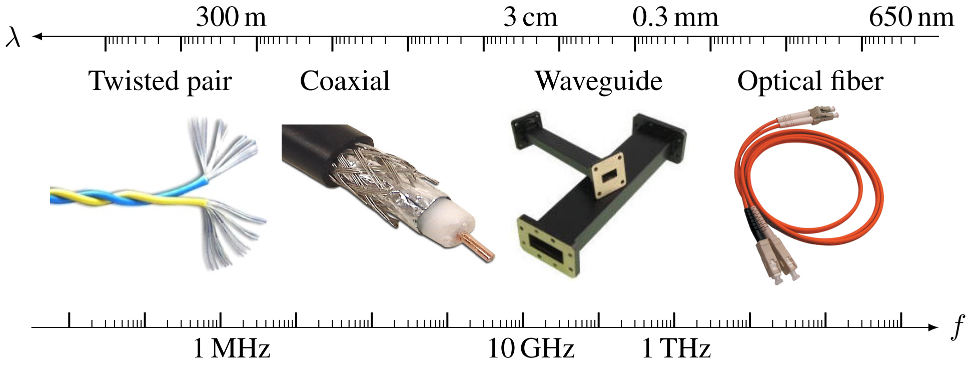


Figure 2.4 Frequency ranges and corresponding means of transportation.

2.3.3 Skin Effect

The phenomenon that conductivity is reduced to the outmost layer of a metal conductor at high frequencies is called *skin effect*. The theory behind the skin effect can be derived by the application of the Maxwell equations to the transmission lines. Essentially, the electromagnetic field forces the current toward the outer layers of a conductor. The current density is an exponential function of the location x (going from the outer skin towards the center of the conductor). In good approximation, see also Fig. 2.5, we find [3] for the absolute value of the current density

$$|J| = J_0 \, \mathrm{e}^{-x/\delta}. \tag{2.2}$$

With *skin depth*, we designate the difference in radius for which the current density has fallen to $1/\mathrm{e}$ or 37 % of the current density at the surface of the conductor. The skin depth is given by [38]

$$\delta = \sqrt{\frac{\lambda}{\pi\sigma\mu v}} = \sqrt{\frac{1}{\pi\sigma\mu f}}, \tag{2.3}$$

where λ is the wavelength, σ the specific conductivity, and μ the permeability of the material. v is the propagation velocity in the material at hand. For copper, for example, the skin depth is only 70 µm at a frequency of 1 MHz. The skin effect can have influence on the design of high-frequency equipment. Copper structures such as antennas are often silver-plated to reduce resistance. This might sound expensive, but considering the thickness needed, it is not.

Examples: Skin depths vs. frequency in some metals

Electric properties of some common metals:

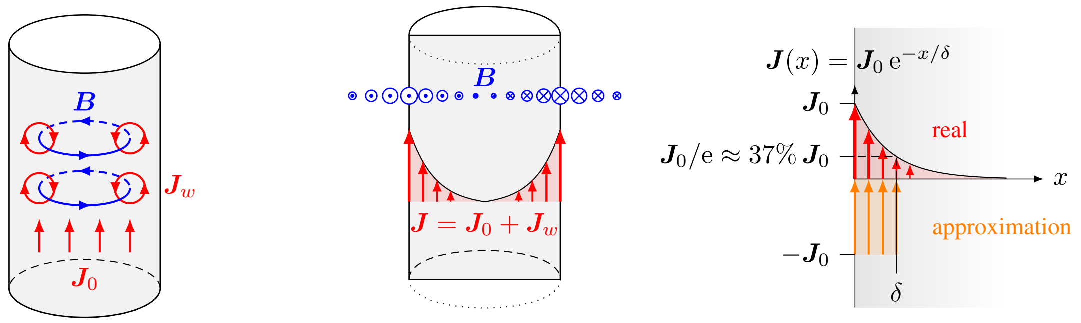
Metal	**Conductivity σ**	**Rel. Permeability μ_r**
Copper (Cu)	58.0 MS/m	$1 - 6.4 \times 10^{-6}$
Aluminium (Al)	36.6 MS/m	$1 + 2.2 \times 10^{-5}$
Iron (Fe)	9.6 MS/m	1000 (300...10 000)
Stainless Steel (X10CrNi18-8)	1.5 MS/m	1.002

Skin depths δ:

Frequency	**Copper**	**Aluminium**	**Iron**	**Stainless Steel**
50 Hz	9.4 mm	11.8 mm	0.7 mm	58.1 mm
1 kHz	2.1 mm	2.6 mm	160 µm	13.0 mm
10 kHz	0.7 mm	0.8 mm	51 µm	4.1 mm
100 kHz	209 µm	263 µm	16.2 µm	1.3 mm
1 MHz	66.1 µm	83.2 µm	5.1 µm	0.4 mm
10 MHz	21 µm	26.3 µm	1.6 µm	130 µm
100 MHz	6.6 µm	8.3 µm	0.5 µm	41.1 µm
1 GHz	2.1 µm	2.6 µm	162 nm	13 µm
10 GHz	0.7 µm	0.8 µm	51 nm	4.1 µm

Thus, as can be seen in the table:

- Iron is a bad conductor for high frequency signals, because of the very narrow skin depth.
- Even for power stations and power transfer the skin depth is of practical importance: a bus bar made of copper for AC at 50 Hz does not have to be thicker than $2\delta = 2 \cdot 10$ mm $=$ 20 mm, at least not for electrical reasons.



B
J_w
J_0
B
$J = J_0 + J_w$
$J(x) = J_0\,\mathrm{e}^{-x/\delta}$
J_0
$J_0/\mathrm{e} \approx 37\%\, J_0$
real
x
approximation
$-J_0$
δ

(a) Eddy currents due to the magnetic field inside the conductor

(b) Displaced current density

(c) The skin depth δ

Figure 2.5 The skin effect: principle, implication and approximation.

2.3.4 Characteristic Impedance

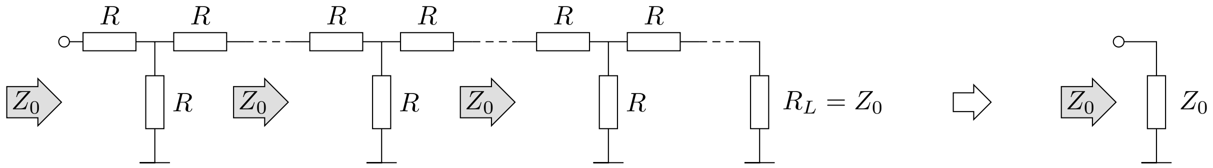


Figure 2.6 Illustration of characteristic impedance.

Imagine we have a cascade of T-networks[1] of resistors R as depicted in Fig. 2.6. Now let us design R such that if we load the output port by an impedance Z_0, we will see the exact same impedance Z_0 looking into the input port. A quick calculation will reveal the following: The input impedance can be computed by considering a circuit of resistor of value R, which is in series to another resistor, also of value R, which, in turn, is parallel to another series consisting of yet another R and the load, hence

$$\frac{(R+Z_0)R}{R+R+Z_0} + R = Z_0. \tag{2.4}$$

Solving this quadratic equation for R yields

$$R = \frac{Z_0}{\sqrt{3}}. \tag{2.5}$$

Now suppose we start cascading several such T-networks. The last one will be loaded by Z_0, making its input impedance Z_0, which, in turn, looks as the load of the one but last network and so on. Now, can you tell how many stages the whole network consists of by its input impedance? No, you cannot. Each network sees Z_0 at its output, transforms that to the same input impedance, which, in turn, lets the network to the left believe that it is loaded by Z_0. This is the concept behind the characteristic impedance. Z_0 is called the wave impedance or the characteristic impedance of the network. This concept is also valid for unequal resistances. Of course, only a fraction of the original power will reach the load attached to the last T-network, since our T-network is lossy. Does this mean by cascading networks of a certain impedance, we will incur insertion loss? Not necessarily. Similar ideas can be followed by using an L-network of pure reactances, for which the loss will be zero (at least in theory). For that purpose, see Fig. 2.7. Now we have

[1] The same theory might be shown using a Pi-network.

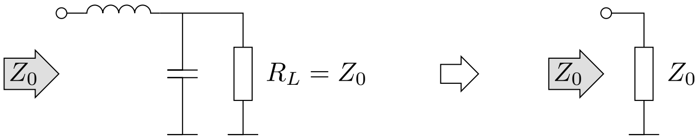


Figure 2.7 L-network of reactances.

an inductor in series with a capacitor which itself is parallel to the load that defines the input impedance.

$$\begin{aligned}\frac{Z_0\frac{1}{j\omega C}}{Z_0+\frac{1}{j\omega C}}+j\omega L &= \frac{Z_0}{1+j\omega C Z_0}+j\omega L\\ &= \frac{Z_0(1-j\omega C Z_0)}{1+\omega^2C^2Z_0^2}+j\omega L = Z_0.\end{aligned} \tag{2.6}$$

The right-hand side of Eq. (2.6) has no imaginary part. Thus, we set the imaginary part of the left-hand side to zero

$$CZ_0^2 = L(1+\omega^2C^2Z_0^2). \tag{2.7}$$

For the real parts of Eq. (2.6), we get

$$\frac{Z_0}{1+\omega^2C^2Z_0^2} = Z_0, \tag{2.8}$$

which constrains the operating frequency to $\omega^2C^2Z_0^2 \ll 1$. Under that condition, Eq. (2.7) simplifies to

$$\sqrt{\frac{L}{C}} = Z_0. \tag{2.9}$$

Using these considerations, we can model a transmission line by a cascade of such networks. For that purpose, we will cut the transmission line into infinitesimally small pieces and model these by networks similar to the one of Fig. 2.7. Doing so, we end up with something like Fig. 2.8, which might serve as the equivalent circuit of many different types of transmission lines. Since Fig. 2.8 represents the transmission line segment of an infinitely small length, we express all the parameters in terms of ’per unit’:

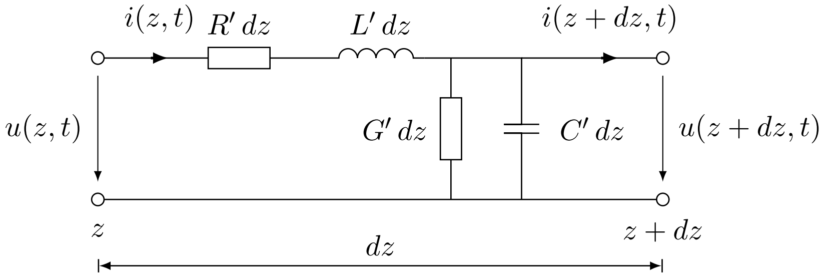


Figure 2.8 Circuit equivalent of a transmission line.

C' per unit capacitance [F/m], mostly given in [pF/m],

L' per unit inductance [H/m], mostly given in [mH/m],

R' per unit resistance [Ω/m],

G' per unit conductance [S/m].

By the Kirchhoff Laws, we can express the voltages and currents as

$$u(z,t) = iR'dz + \frac{\partial i}{\partial t}L'dz + u(z+dz,t), \tag{2.10a}$$

$$i(z,t) = uG'dz + \frac{\partial u}{\partial t}C'dz + i(z+dz,t). \tag{2.10b}$$

Eqs. (2.10a) and (2.10b) can be transformed into

$$\frac{u(z+dz,t)-u(z,t)}{dz} \triangleq \frac{\partial u}{\partial z} = -\left(R' + L'\frac{\partial}{\partial t}\right)i, \tag{2.11a}$$

$$\frac{i(z+dz,t)-i(z,t)}{dz} \triangleq \frac{\partial i}{\partial z} = -\left(G' + C'\frac{\partial}{\partial t}\right)u. \tag{2.11b}$$

Eqs. (2.11a) and (2.11b) build a system of linear partial differential equations of first order.

2.3.5 The Telegraph Equations

The telegraph equations describe the development of voltage and current along a transmission line. They result from partially differentiating Eq. (2.11a) with respect to z, using it in Eq. (2.11b) partially differentiated with respect to t, transforming some terms, and vice versa

$$\frac{\partial^2 u}{\partial z^2} = R'G'u + (R'C' + L'G')\frac{\partial u}{\partial t} + L'C'\frac{\partial^2 u}{\partial t^2}, \tag{2.12a}$$

$$\frac{\partial^2 i}{\partial z^2} = R'G'i + (R'C' + L'G')\frac{\partial i}{\partial t} + L'C'\frac{\partial^2 i}{\partial t^2}. \tag{2.12b}$$

Note that for lossless transmission lines, i.e., $R' = G' = 0$, Eqs. (2.12a) and (2.12b) can be simplified to the Telegraph equations for lossless transmission lines

$$\frac{\partial^2 u}{\partial z^2} = L'C'\frac{\partial^2 u}{\partial t^2}, \tag{2.13a}$$

$$\frac{\partial^2 i}{\partial z^2} = L'C'\frac{\partial^2 i}{\partial t^2}. \tag{2.13b}$$

If we now choose a frequency-dependent form for the voltages and currents along the line and in function of time, we can write

$$u(t,z) = \mathrm{Re}\left(U(z)\,\mathrm{e}^{j\omega t}\right), \tag{2.14a}$$

$$i(t,z) = \mathrm{Re}\left(I(z)\,\mathrm{e}^{j\omega t}\right), \tag{2.14b}$$

so that after settling we get

$$\frac{\partial^2 U}{\partial z^2} = (R' + j\omega L')(G' + j\omega C')U, \tag{2.15a}$$

$$\frac{\partial^2 I}{\partial z^2} = (R' + j\omega L')(G' + j\omega C')I. \tag{2.15b}$$

The solution to this system of differential equations is a sum of incident and reflected wave

$$U(z) = U_a + U_b = U_{a0}\,\mathrm{e}^{-\gamma z} + U_{b0}\,\mathrm{e}^{\gamma z}, \tag{2.16a}$$

$$I(z) = I_a - I_b = I_{a0}\,\mathrm{e}^{-\gamma z} - I_{b0}\,\mathrm{e}^{\gamma z}, \tag{2.16b}$$

where

$$\boxed{\gamma = \alpha + j\beta = \sqrt{(R' + j\omega L')(G' + j\omega C')}} \tag{2.17}$$

is the *wave propagation constant*[2] with the real part α denoting the attenuation and the imaginary part β the phase, respectively. Since the phase rotates by 2π for every wavelength, we have

$$\beta = \frac{2\pi}{\lambda}. \tag{2.18}$$

By choosing the negative sign in Eq. (2.16b), the two components become the incident and reflected current of a transmission line. Note that whereas U_a, U_b, I_a, and I_b, are the amplitudes of the incident and reflected voltage and current waves at the location z, the values for U_{a0}, U_{b0}, I_{a0}, and I_{b0}, represent the same for the location $z = 0$. By differentiating Eq. (2.16a) with respect to the location and using the impedance Z_0 as the factor between voltage and current

$$\begin{aligned}\frac{dU(z)}{dz} &= -\gamma U_{a0}\,\mathrm{e}^{-\gamma z} + \gamma U_{b0}\,\mathrm{e}^{\gamma z}\\ &= -\gamma Z_0 I_{a0}\,\mathrm{e}^{-\gamma z} + \gamma Z_0 I_{b0}\,\mathrm{e}^{\gamma z}\\ &= -\gamma Z_0 I(z).\end{aligned} \tag{2.19}$$

The phasor form of Eq. (2.11a) states

$$\frac{dU(z)}{dz} = -(R' + j\omega L')I(z). \tag{2.20}$$

A comparison of Eq. (2.19) and Eq. (2.20) reveals the characteristic or *wave impedance* of a line

$$\boxed{Z_0 = \frac{R' + j\omega L'}{\gamma} = \sqrt{\frac{R' + j\omega L'}{G' + j\omega C'}}.} \tag{2.21}$$

For lossless transmission lines Eq. (2.21) simplifies to

$$\boxed{Z_0 = \sqrt{\frac{L'}{C'}}.} \tag{2.22}$$

In this case, the attenuation α is zero and the phase is given by

$$\beta = \omega\sqrt{L'C'}. \tag{2.23}$$

The phase velocity is then given by

$$\boxed{v = \lambda \cdot f = \frac{2\pi}{\beta} \cdot f = \frac{\omega}{\beta} = \frac{1}{\sqrt{L'C'}} = \frac{1}{\sqrt{\mu\varepsilon}} = \frac{c_0}{\sqrt{\mu_r\varepsilon_r}}\ .} \tag{2.24}$$

[2] Propagation constant is in fact a misnomer, although well-accepted, since it is dependent on ω. Other name such as propagation coefficient are more appropriate.

2.3.6 Cable losses

In the following we consider lossy transmission lines. If the losses are moderate (otherwise the transmission line might not be very useful anyway), we can write Eq. (2.17) as

$$\begin{aligned}\gamma = \alpha + j\beta &= \sqrt{(R' + j\omega L')(G' + j\omega C')} \\ &= j\omega\sqrt{L'C'}\sqrt{\left(1 + \frac{R'}{j\omega L'}\right)\left(1 + \frac{G'}{j\omega C'}\right)} \\ &\approx j\omega\sqrt{L'C'}\sqrt{1 + \frac{R'}{j\omega L'} + \frac{G'}{j\omega C'}},\end{aligned} \tag{2.25}$$

where in the last step we have used the circumstance that the multiplication of two small terms may be neglected. Using a Taylor-series approximation of 1st order for $(1+x)^n \approx 1 + nx$ we can write

$$\begin{aligned}\gamma &\approx j\omega\sqrt{L'C'}\left(1 + \frac{R'}{2j\omega L'} + \frac{G'}{2j\omega C'}\right) \\ &= j\omega\sqrt{L'C'} + \frac{\sqrt{L'C'}R'}{2L'} + \frac{\sqrt{L'C'}G'}{2C'} \\ &= \underbrace{j\omega\sqrt{L'C'}}_{j\beta} + \underbrace{\frac{R'}{2Z_0} + \frac{Z_0 G'}{2}}_{\alpha}\end{aligned} \tag{2.26}$$

As can be seen above, the first part of α is contributed by the resistive loss along the transmission line, whereas the second part of α is due to dielectric losses.

2.3.7 Coaxial Cable

The transmission line type most often used is the coaxial cable, which is shown in Fig. 2.9. The term *coaxial* derives from the cable's design: the two conductors share a single axis. But as simple and straightforward the coaxial cable appears today, it dates back only as far as just a few years before Maxwell formed his equations. Some coaxial cables and their properties are listed in A.2.5. For a coaxial cable, capacitance and

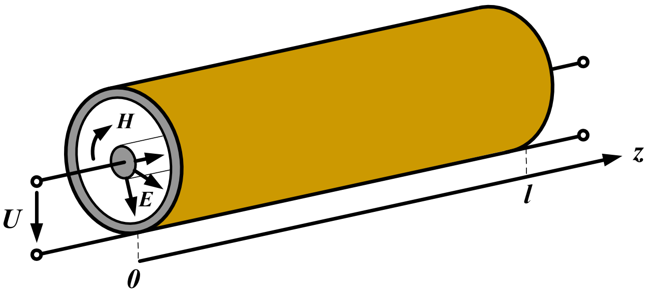


Figure 2.9 Coaxial cable with voltage, electric field, and magnetic field components. Note that the left end of the cable is denoted by $z = 0$, whereas the right end as $z = l$.

inductance are given as a function of the radii of the inner and outer conductor (r_1 and r_2, respectively) and the dielectric inbetween

$$\boxed{C' = \frac{2\pi\varepsilon_r\varepsilon_0}{\ln(r_2/r_1)},} \tag{2.27}$$

$$\boxed{L' = \frac{\mu_r\mu_0}{2\pi}\ln(r_2/r_1).} \tag{2.28}$$

By filling them into the definition of the characteristic impedance in lossless cases (2.22), the characteristic wave impedance of a coaxial cable can be found to be

$$\boxed{Z_{0\,\text{coax}} = \sqrt{\frac{\mu_r\mu_0}{\varepsilon_r\varepsilon_0}}\,\frac{\ln(r_2/r_1)}{2\pi} \approx \frac{60}{\sqrt{\varepsilon_r}}\ln\frac{r_2}{r_1}} \qquad (\mu_r \approx 1)\ . \tag{2.29}$$

Year	Event
1880	Oliver Heaviside patents the coaxial cable in England, (Patent No. 1407, 6. April 1880)
1884	Ernst Werner Siemens & Johann Georg Halske patent the coaxial cable in Germany (Patent No. 28.978, 27. March 1884)
1894	Nikola Tesla receives the patent for coaxial cable in the U.S. (U.S. Patent No. 0.514.167, "Electrical Conductor", New York, 6. February 1894)
1929	First modern coaxial cable patented by Lloyd Espenschied and Herman Affel of AT&T's Bell Telephone Laboratories (U.S. Patent 1.835.031)
1936	First transmission of TV pictures on coaxial cable, from the 1936 Summer Olympics in Berlin to Leipzig
1936	World's first underwater coaxial cable installed (Apollo Bay, near Melbourne, Australia, to Stanley, Tasmania, the 300 km, one broadcast channel and seven telephone channels)
1936	AT&T installs experimental coaxial telephone and television cable between New York and Philadelphia, with automatic booster stations every ten miles, completed in December, it can transmit 240 telephone calls simultaneously
1936	Coaxial cable laid by the General Post Office (now BT) between London and Birmingham, providing 40 telephone channels
1941	First commercial use in USA by AT&T, between Minneapolis, Minnesota and Stevens Point, Wisconsin. L1 system with capacity of one TV channel or 480 telephone circuits
1943	First submerged repeaters in the line between Anglesey and the Isle of Man, Irish Sea, linking the UK with Germany, the Netherlands, Belgium and Denmark
1956	First transatlantic coaxial cable laid, opened service on 25. September 1956, within the first 24 hours of service there were 588 calls from London to the US and 119 to Canada
1976	Robert Metcalfe invents the computer network over coxial cable as part of his dissertation, Ethernet (then called ALOHAnet) is born
1980	LAN is predominantly based on coax: *thicknet* (10BASE5) and *thinnet* or *cheapernet* (10BASE2), by about 1990 it was replaced by UTP (unshielded twisted pair, 10BASE-T)

Table 2.1 One century of coaxial cable history.

The Optimal Impedance Value for Coaxial Cables

Depending on the optimization criterium, we can find different values of wave impedance that show best performance. In the following, the optimization of the cable impedance is carried out for the three most important criteria:

- **Minimum loss (maximum efficiency)**: The loss of a transmission line is usually dominated by the skin effect. Using a model as shown in Fig. 2.10, we can easily see that the current through the load and the loss resistances (the skin resistances of the conductors) are the same. Thus, the power in each resistance is a linear function of the resistance. The relative loss is therefore

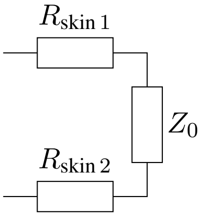


Figure 2.10 Cable loss modeled with resistances due to skin effects.

$$\text{loss} = \frac{R_{\text{skin1}} + R_{\text{skin2}}}{Z_0 + R_{\text{skin1}} + R_{\text{skin2}}} \approx \frac{R_{\text{skin1}} + R_{\text{skin2}}}{Z_0}, \qquad R_{\text{skin}} \ll Z_0. \tag{2.30}$$

where R_{skin1} and R_{skin2} are due to the skin effect of inner and outer conductor of the coaxial line. Since the resistance due to the skin effect is inversely proportional to the radii of the conductors, we note that the loss is proportional to

$$\text{loss} \propto \frac{\frac{1}{r_1} + \frac{1}{r_2}}{Z_0}, \tag{2.31}$$

where r_1 and r_2 are the radius of the inner and the outer conductor, respectively. The wave impedance of coaxial cables can be obtained by (2.29). Now, the loss is proportional to

$$\text{loss} \propto \frac{\frac{1}{r_1} + \frac{1}{r_2}}{\ln \frac{r_2}{r_1}} = \frac{1}{r_2} \frac{1 + \frac{r_2}{r_1}}{\ln \frac{r_2}{r_1}}. \tag{2.32}$$

As Z_0 is a function of only the ratio $\frac{r_2}{r_1}$ and not of the absolute value of r_2, we minimize the loss by differentiating the second factor of Eq. (2.32) with respect to $r = r_2/r_1$. Hence, for the outer radius held constant, we can minimize the following function

$$\alpha(r) = \frac{1+r}{\ln r}. \tag{2.33}$$

By setting the first derivative of $\alpha(r)$ to zero we get

$$\ln r = 1 + \frac{1}{r}, \tag{2.34}$$

the solution of which is $r \approx 3.59$. The corresponding wave impedance for minimum loss is thus

$$Z_0 = \frac{60}{\sqrt{\varepsilon_r}} \ln 3.59 = \frac{76.7}{\sqrt{\varepsilon_r}} \qquad [\Omega]. \tag{2.35}$$

For air ($\varepsilon_r = 1$), the minimum-loss value is thus $Z_0 = 76.7\,\Omega$, for polyethylene (PE), with $\varepsilon_r = 2.25$, it is $Z_0 = 51.1\,\Omega$. Hence, the two values of $Z_0 = 75, \Omega$ and $Z_0 = 50\,\Omega$ are very popular choices. For Teflon (PTFE), with $\varepsilon_r = 2.1$, we get $Z_0 = 52.9\,\Omega$.

- **Maximum withstand voltage:** The *withstand voltage*, also called *electric strength* is an important parameter which describes what voltage the cable can stand without resulting in a breakdown of the insulating material. The electric field inside a coaxial cable is a function of the distance x from the center of the cable
$$E = \frac{U}{x \ln \frac{r_2}{r_1}}, \tag{2.36}$$
where U is the voltage applied. The maximum field strength occurs at the inner conductor, $x = r_1 = r_2/r$, and can be found again by differentiating Eq. (2.36) with respect to $r = r_2/r_1$. Doing so, we can find that the minimum electric field strength is obtained at $\ln r = 1$. The corresponding wave impedance is thus
$$Z_0 = \frac{60}{\sqrt{\varepsilon_r}} \qquad [\Omega]. \tag{2.37}$$

- **Maximum power transfer:** For a given outer radius r_2 and the highest allowed field strength (in order not to damage the cable), the power that is transferred over the cable can be written as
$$P = \frac{U^2}{Z_0} = \frac{E^2(r_1 \ln r)^2}{Z_0} = \frac{E^2(r_1 \ln r)^2}{\frac{60}{\sqrt{\varepsilon_r}} \ln r} = \frac{E^2 \left(\frac{r_2}{r}\right)^2 \ln r}{\frac{60}{\sqrt{\varepsilon_r}}}. \tag{2.38}$$
where U is the voltage applied. Apart from the constant r_2, the power is proportional to
$$\alpha(r) = \frac{\ln r}{r^2}. \tag{2.39}$$
By setting the first derivative of $\alpha(r)$ to zero, we get $\ln r = 1/2$, the solution of which is $r = \mathrm{e}^{1/2}$. The corresponding wave impedance for maximum power transfer is
$$Z_0 = \frac{30}{\sqrt{\varepsilon_r}} \qquad [\Omega]. \tag{2.40}$$

Additionally, $50\,\Omega$ is often chosen as a compromise to get the least reflection from connecting a cable to a $\lambda/2$ dipole (feed-point impedance $Z_0 = 73\,\Omega$) or a $\lambda/4$ monopole (feed-point impedance $Z_0 = 40\,\Omega$). Table 2.2 shows an overview of all optimum impedances considered, Fig. 2.11 shows the simulation results leading to the optimum impedances, and Fig. 2.12 compares common coaxial cables and their properties. Remark: The commonly appearing abbreviation "RG" stands for *radio guide* and was originally a unit indicator for bulk RF cable in the U.S. military, which first standardized such cables and their specifications.

Criterion	**Air**	**Polyethylene (PE)**
Minimum loss	$76.7\,\Omega$	$51.1\,\Omega$
Minimal electric field strength	$60\,\Omega$	$40\,\Omega$
Maximum power transfer	$30\,\Omega$	$20\,\Omega$
Best match to $\lambda/2$ dipole	$73\,\Omega$	
Best match to $\lambda/4$ monopole	$40\,\Omega$	

Table 2.2 Optimum coaxial cable impedances.

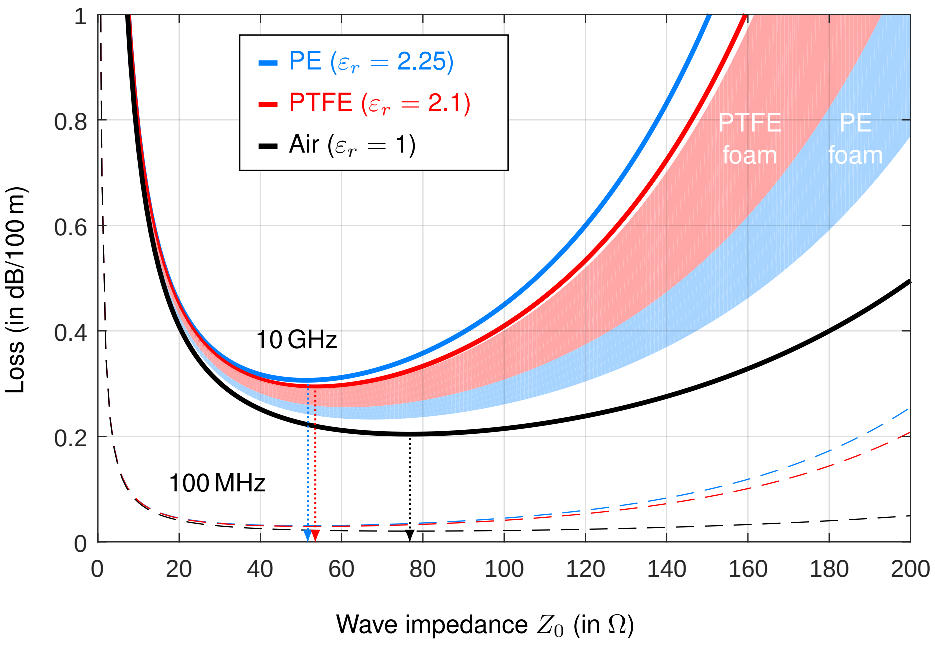


Figure 2.11 Conduction loss of coaxial cables vs. their wave impedances for various dielectrics at 10 GHz (solid lines) and 100 MHz (dashed lines).

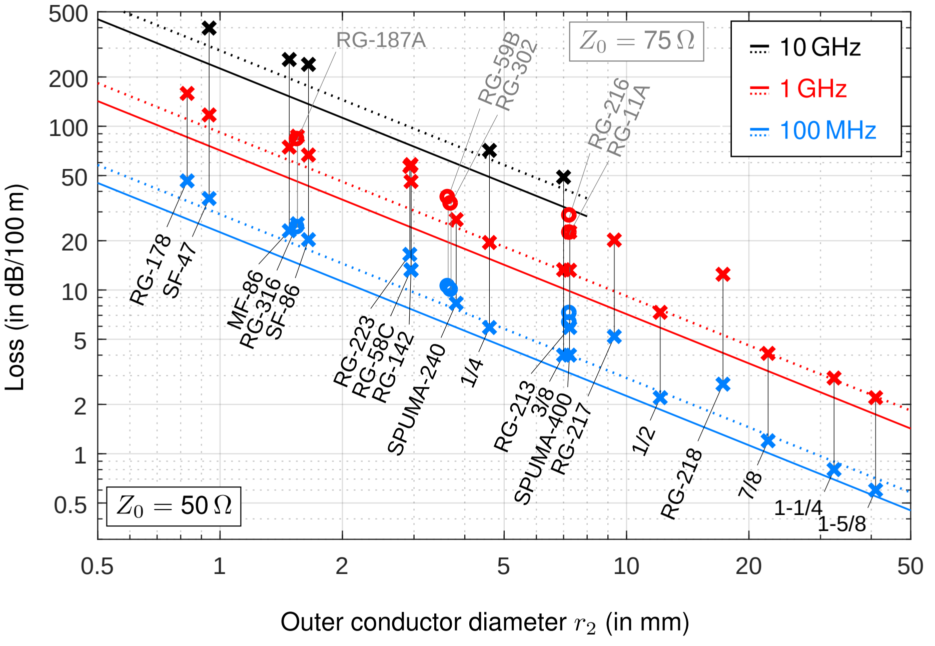


Figure 2.12 Loss (in dB/100 m) of various coaxial cables (wave impedance $Z_0 = 50\,\Omega$ with cross markers and black labels, $Z_0 = 75\,\Omega$ with circle markers and gray labels) at 100 MHz, 1 GHz, and 10 GHz. The solid and dashed lines are the theoretical losses for air ($\varepsilon_r = 1$) and dielectric foam ($\varepsilon_r = 2$) fillings.

2.3.8 Waveguides

Another means of transfering RF power along is the so called waveguide, which usually consists of a rectangular section. It is basically shaped as a long metallic box. The section dimensions need to be large enough to allow at least one mode to propagate. Roughly speaking, the larger of the two section dimensions need to be half a wavelength or larger. A waveguide is displayed in Fig. 2.13. In a waveguide, only electric and magnetic fields are present. The corresponding voltages and currents do not exist. Waveguides have applications in pulse radar due to their inherent power transfer capabilities.

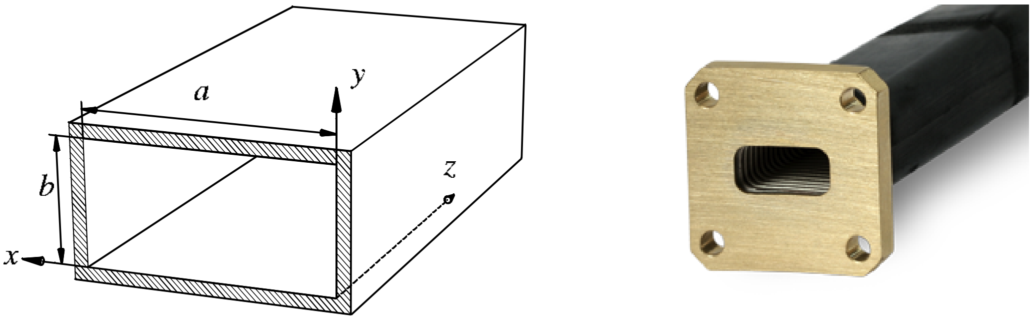
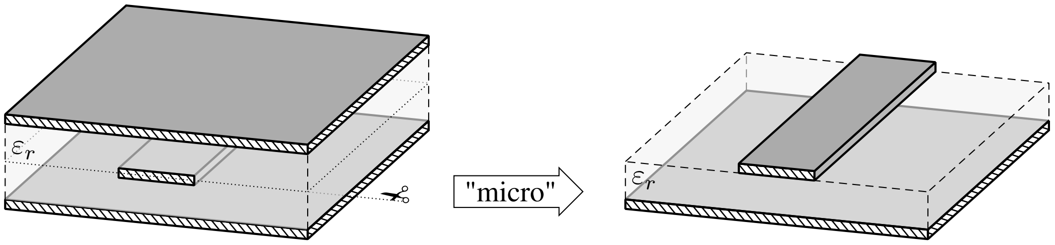


Figure 2.13 Waveguide. Sources: [3] and `safthenika.com`.

2.3.9 PCB Transmission Lines

Microstrip Transmission Lines

Microstrip transmission lines, illustrated in Fig. 2.14, are frequently used when a transmission line needs to be place on a PCB. This transmission-line type was introduced in 1952 as “novel approach to microwave transmission and components” [27]. While being more difficult to analyze, having dispersive (frequency-dependent) properties and not shielding as well as striplines, the ease of both series- and shunt-connecting active and passive components to the lines, as well as the simpler manufacturing process (and cost) quickly made microstrip lines the first choice for many applications.



(a) Stripline (b) Microstrip line

Figure 2.14 From stripline to microstrip.

Whereas for AF (audio frequency) applications a PCB can be designed in straightforward manner, once the schematic is born, thorough planning of the trace routes is necessary when the line lengths start to become comparable to the wavelength of the signal of interest.

For microstrip lines, some of the field vectors penetrate the medium above the PCB, usually air, see Fig. 2.15. This changes the effective dielectric permittivity[3] somewhat. A coarse approximation of the effective dielectric permittivity is given by [68]

$$\boxed{\varepsilon_{r\,\mathrm{eff}} \approx \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\frac{1}{\sqrt{1 + 12h/w}}\,,} \tag{2.41}$$

where h and w are the height of the PCB and the width of the microstrip line, respectively. A plot of the effective permittivity as given by Eq. (2.41) is shown in Fig. 2.16(a). The wider the microstrip lines, the

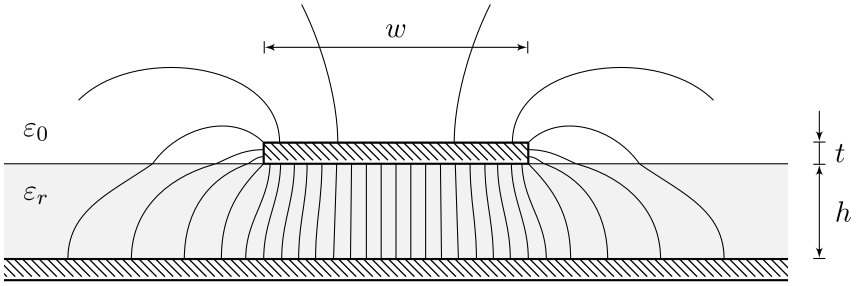


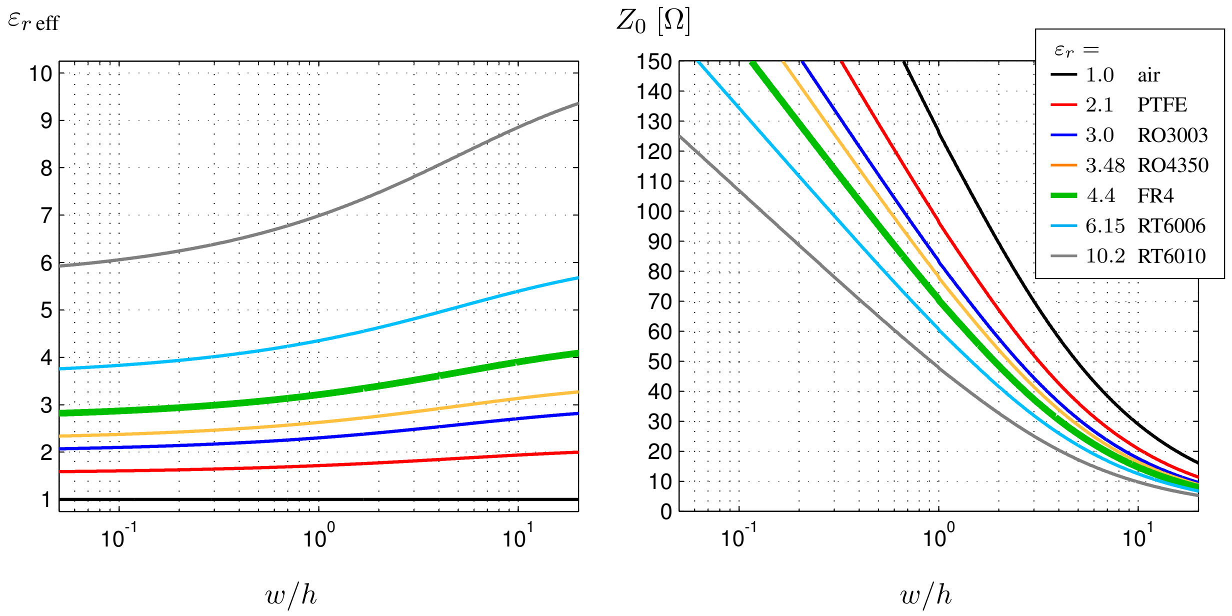
Figure 2.15 Electric field in microstrip lines.

[3] Note that the former expression ’dielectric constant’ is called ’dielectric permittivity’ these days, since this physical measure is frequency dependent (and thus not constant).

smaller the difference between effective and relative permittivity is, since most of the field is contained in the PCB. The wave impedance of microstrip lines can be approximated by [30]

$$Z_0 = \begin{cases} \dfrac{60}{\sqrt{\varepsilon_{r\,\mathrm{eff}}}} \ln\left(8\dfrac{h}{w} + 0.25\dfrac{w}{h}\right) , & w/h \leq 1 \\[2ex] \dfrac{120\pi/\sqrt{\varepsilon_{\mathrm{eff}}}}{w/h + 1.393 + 0.667\ln(w/h + 1.444)} . & w/h \geq 1 \end{cases} \qquad [\Omega] \qquad (2.42)$$

A plot of the wave impedance of microstrip lines as given by Eqs. (2.42) is shown in Fig. 2.16(b). For microstrip lines with additional ground planes running along the transmission lines, the impedance stays roughly constant as long as the gap between the line and the ground is at least twice the thickness of the substrate.



(a) Effective permittivity $\varepsilon_{r\,\mathrm{eff}}$

(b) Wave impedance Z_0

Figure 2.16 Characteristics of microstrip lines as a function of their width-to-height ratio w/h.

Example: Effective dielectric permittivity and line impedances

Effective dielectric permittivity $\varepsilon_{\mathrm{eff}}$, propagation velocity and microstrip line width w for air and common PCB materials with height $h = 1.6$ mm, for 50 Ω wave impedance.

PCB material	ε_r	$\varepsilon_{\mathrm{eff}}$	w[mm]	v/c
Air	1.0	1.0	7.86	100%
PTFE	2.1	1.82	5.08	74%
RO3003	3.0	2.45	4.01	64%
RO4350	3.48	2.77	3.62	60%
FR4	4.4	3.38	3.04	54%
RT6006	6.15	4.49	2.33	47%
RT6010	10.2	6.93	1.46	38%

Note that whereas for low frequencies, loss in the dielectric is usually not discussed, for higher frequencies the loss becomes a serious threat. Often the loss is considered in the imaginary part of ε_r, which introduces a loss admittance G' in parallel to the C' in the equivalent circuit of an infinitesimally small transmission line. The loss factor, also known as the loss tangent or dissipation factor (or loss factor) is then given as

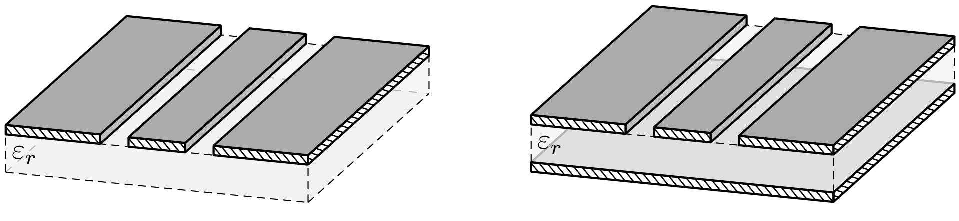
$$\tan\delta = -\frac{\operatorname{Im}\varepsilon_r}{\operatorname{Re}\varepsilon_r}. \tag{2.43}$$

The conductance of the material can then be computed from the capacitance of the dielectric using

$$G' = \omega C' \tan\delta. \tag{2.44}$$

Coplanar Waveguide

Cheng P. Wen invented the coplanar waveguide[4], CPW, while working at RCA's Sarnoff Laboratories. His 1969 IEEE paper titled *"Coplanar waveguide: a surface strip transmission line suitable for nonreciprocal gyromagnetic device applications"* is largely taken proof enough that the original concept belongs to him. Nowadays, the coplanar waveguide is an ubiquitous transmission line, especially in low- and medium-power microwave electronics. Generally, when we are talking about coplanar waveguides on PCB we actually mean the *grounded* CPG (GCPW, also called coplanar waveguide with ground [81] or conductor-backed CPW [70]), whereas the original CPW is also called *conventional* coplanar waveguide [70]. Both cross sections are depicted in Fig. 2.17. As for the microstrip line, the lower ground plane provides mechanical strength, which makes it particularly well-suited for applications on thin and fragile wafers (e.g. GaAs), and because of the connection to the top layer, where the components are located, it also provides good head sink capability (e.g. for medium-power amplifiers).



(a) CPW (conventional)

(b) GCPW (grounded)

Figure 2.17 Conventional and Grounded Coplanar Waveguide.

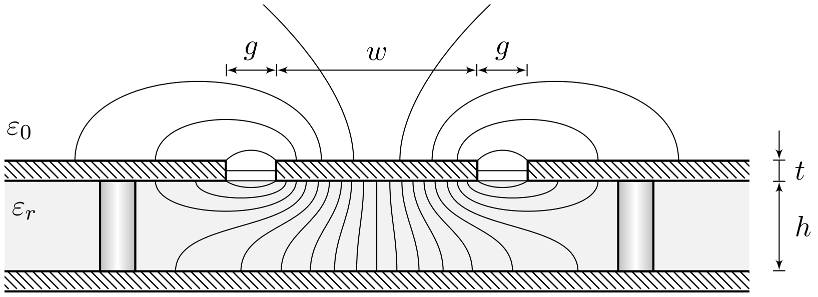


Figure 2.18 Electric field in grounded coplanar waveguide lines.

[4] Some people presume the name might have come after the abbreviation, which also stands for Cheng P. Wen...

The overall working principle is similar to the one of the microstrip line: again, a part of the field is carried inside the dielectric (generally more than for the microstrip line) and a part propagates in the surrounding air. Important are the fringing fields on the edges toward the coplanar ground areas, which let the fields concentrate more in the area near the strip.

As for the microstrip line, analytical results of the wave impedance (or the effective dieletric permittivity) cannot be found[5]. Instead of using complicated numerical methods or even general field simulators (which of course achieve the best results, but with huge computational expense) "simple" approximative models have been built. The characteristic wave impedance of grounded coplanar waveguides is most commonly modelled by

$$Z_0 = \frac{60\pi}{\sqrt{\varepsilon_{r\,\mathrm{eff}}}} \, \frac{1}{\dfrac{K(k_1)}{K(k_1')} + \dfrac{K(k_2)}{K(k_2')}} \,, \tag{2.45}$$

with $k_1 = a/b$ or $k_1 = w/(w+2g)$, and $k_n' = \sqrt{1-k_n^2}$, as well as

$$k_2 = \frac{\tanh \dfrac{w\,\pi}{4h}}{\tanh \dfrac{(w+2g)\,\pi}{4h}} \tag{2.46}$$

and the effective relative permittivity

$$\varepsilon_{r\,\mathrm{eff}} = \frac{1 + \varepsilon_r \, \dfrac{K(k_1')}{K(k_1)} \, \dfrac{K(k_2)}{K(k_2')}}{1 + \dfrac{K(k_1')}{K(k_1)} \, \dfrac{K(k_2)}{K(k_2')}} \,. \tag{2.47}$$

Both the model for the effective relative permittivity $\varepsilon_{r\,\mathrm{eff}}$ as well as the one for the wave impedance Z_0 use the so-called *complete elliptic integral of the first kind* given by

$$K(k) = \int_0^{\pi/2} \frac{1}{\sqrt{1-k^2\,\sin^2\theta}}\, d\theta \,, \tag{2.48}$$

which can be found in [10] or in Matlab using the command `ellipke`[6]. The model is known to estimate somewhat too high impedances, owing to the fact that it does not include effects due to the finite thickness of the strip and the coplanar ground areas. Cleary, those effects are even more important than for the microstrip line and can be accounted for by widening the strip and the slot widths a bit and using the model provided before. As correction for both widths

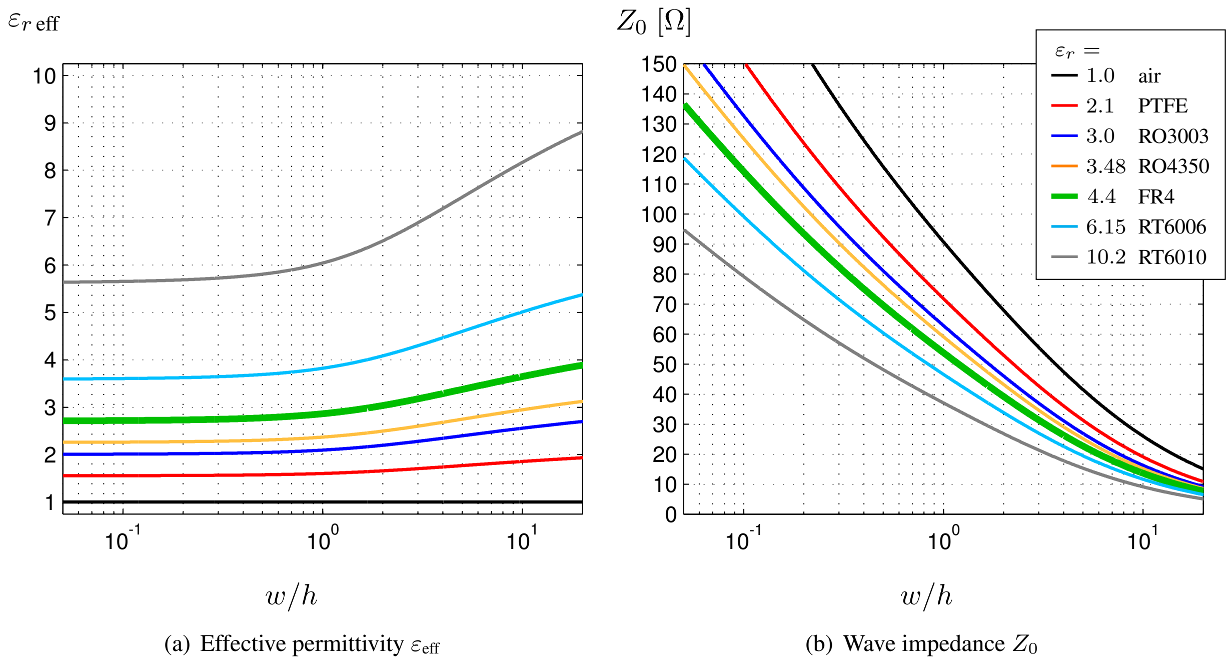
$$\Delta g = \Delta w = 1.25\,t\,\frac{1+\ln(2h/t)}{\pi} \tag{2.49}$$

is given in [70], after which however somewhat too low impedances are received. The effective relative permittivity and wave impedances of grounded CPW are plotted in the graphs in Fig. 2.19.

The vias connecting the coplanar ground areas with the lower ground layer are not generally necessary (at sufficiently high frequencies), since the large capacitance between the layers provides enough coupling

[5] Only for the conventional CPW with a dielectric of ε_r equal to the surrounding dielectric, analytical solutions have been derived.

[6] Actually, `ellipke` is equivalent to $K(m) = \int_0^{\pi/2}(1 - m\,\sin^2\theta)^{-2}\,d\theta$ where $m = k^2$. Thus, the command should be called according to `ellipke(k^2)`.



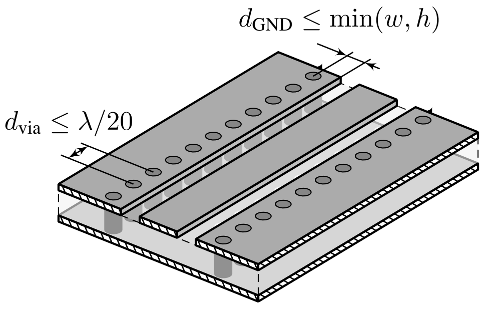
(a) Effective permittivity $\varepsilon_{\mathrm{eff}}$

(b) Wave impedance Z_0

Figure 2.19 Characteristics of CPW lines as a function of their width-to-height ratio w/h with constant gap width $g = 0.25\,\mathrm{mm}$.

for the waveguide to work. However, besides accounting for better heat transfer, the vias have two more purposes: suppressing the spurious differential slot mode as well as the cavity waveguide mode. These enhancements make coplanar waveguide better suited for both higher and lower frequencies, thus increasing its broadband capabilities. As commonly recommended in design rules and application notes, there should generally be "as many vias as possible" and they should be distributed all over the ground areas. As illustrated in Fig. 2.20, the vias spread along the line should not be separated by more than $\lambda/20$ and not more than about one strip width or one thickness of the dielectric (PCB), whichever is smaller, away from the ground edge. Placing the vias even closer to the strip can be favourable, but then they have greater influence on the impedance, making the calculations by the model provided above less reliable. Also, the vias should not be too small in diameter, since that would account for larger inductance and diminish their purpose.

Generally, narrow strips (and slots) are to be preferred, since they allow more ground plane and therefore better isolation and lower radiation loss. However, very narrow strips account for increased conduction loss, since evidently narrower strips show increased resistance. This also applies for the slots: narrow



$d_{\mathrm{GND}} \leq \min(w, h)$
$d_{\mathrm{via}} \leq \lambda/20$

Figure 2.20 Via placement and separation for GCPW.

slots account for higher (fringing) electric fields between the strip and the ground, which due to the finite conductance of the dielectric causes increased dielectric loss. Thus, in practice all those effects have to be considered well and a suitable compromise has to be found.

Microstrip vs. Coplanar Waveguide in Applications

One point that is always under consideration, especially for broadband and pulse application is frequency dispersion, i.e., the frequency-dependent behavior of a transmission line. For example, some frequencies can be attenuated (or even annihilated), some might be delayed whereas others will not suffer at all, leading to distortion and making that transmission channel unsuited for broadband applications.

Frequency dispersion occurs for all non-TEM[7] transmission lines, such as waveguides (both the cavity waveguide as well as coplanar waveguides), microstrip lines, etc. Generally said, dispersion occurs, wherever scattering of a wave (e.g. in a waveguide) or an (electrically) asymmetrical structure of the transmission line is present. For example, coax and stripline are both symmetrical structures (and do not involve wave scattering) and in fact neither of them is dispersive, both are so-called *pure TEM media*.

Fig. 2.21 shows the frequency dependence of the effective permittivity for both microstrip and coplanar waveguides. Clearly, both are frequency dependent – but which is worse? Since dispersion is a measure of the frequency dependence, larger fluctuations in narrower frequency ranges are worse than a slow steady progression. The effective dielectric permittivity of microstrip is somewhat flatter with frequency, especially both at low and high frequencies. CPW on the other hand is only flat for low frequencies (where the strip and the slots seem very narrow compared to the wavelength) and dispersion effects increase towards higher frequencies. Therefore, CPW is considered to be more dispersive than microstrip and should only be used in low-frequency regions, meaning the wavelength should be large compared to all geometric extents (e.g. the strip and slot widths as well as the substrate thickness h). In Table 2.3 some other properties of microstrip and coplanar waveguides are compared.

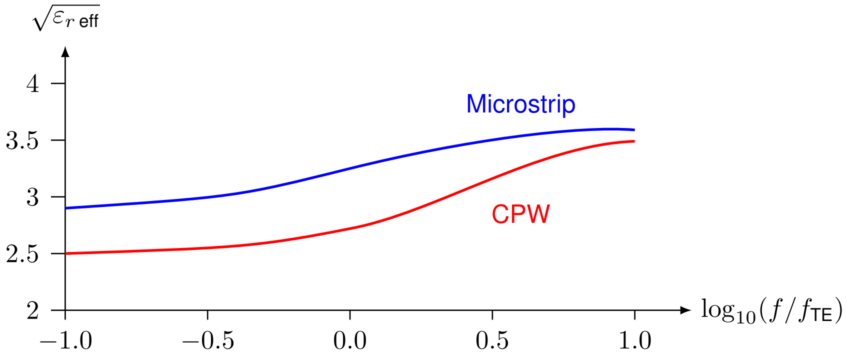


Figure 2.21 Comparison of effective permittivity vs. frequency.

[7] TEM stands for Transverse Electro-Magnetic mode, meaning the electric and magnetic fields stand exactly orthogonal to each other. This is the best-case scenario, but rarely the case. Other modes are TE, TM and hybrid modes.

Property	**Microstrip**	**(Grounded) Coplanar Waveguide**
Loss*	Highly dependent on the substrate material, but microstrip lines can be made very low-loss. Radiation loss can be a problem, especially for open-connected line ends.	Usually, due to the smaller distances between the signal trace and the ground areas, GCPW is more lossy than microstrip for the same application and materials.
Isolation	Fields between adjacent traces can couple significantly, corresponding to low isolation. Thus, microstrip is well-suited for coupled-line applications and unqualified for applications with multiple traces in close proximity.	Radiation loss, cross-talk and coupling is lower for GCPW, making it better suited for highly sensitive and miniature applications than microstrip. GCPW is less suited for couplers.
Bandwidth, Dispersion	Microstrip is less dispersive than CPW and, therefore, somewhat better suited for applications using broader frequency ranges, e.g. pulse applications.	CPW can provide very high frequency responses (up to 100 GHz), but special methods of enhancement have to be applied.
Simplicity of design	Microstrip lines are very easy to design and the models are quite reliable for a large variety of materials.	CPW has more degrees of freedom for a given impedance, i.e. can be made smaller for the same application. On the other hand, models are known to be less reliable.
Insertion of components	Serial components, stubs and open lines (e.g. filter structures) can easily be incorporated. Shunt (parallel) components require additional ground areas on the same layer.	CPW is well-suited for lumped circuits as it can employ both serial and parallel elements. On the other hand, it is less suited for stubs and filter applications.
Area of application	Microstrip is more common in areas where the lines are shorter and low loss crucial.	CPW is especially common in integrated circuits, (M)MIC.
Availability	Microstrip has been around for quite some time and, thus, is very well known and available in all microwave CAD programs.	CPW is newer and still less common in (especially less elaborate and expensive) microwave CAD programs.

*) In integrated devices both microsrip and CPW are built on suspended substrates, i.e. "in the air", to further minimize loss.

Table 2.3 Comparison of some properties of microstrip and coplanar waveguide lines.

2.4 Impedance Matching

It is well known that the power transfer from a DC source is maximal if the load resistance is equal to the internal resistance of the source. At AC (see Fig. 2.22), this is the case if the load impedance Z_L is the complex conjugate of the source impedance Z_S.

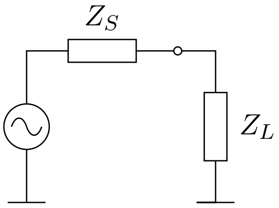


Figure 2.22 Load and source impedance in a circuit.

Fig. 2.23 shows that the reactive parts cancel in this case and only the real parts of the source and load impedance face each other. The place of the matching circuit between the source and the load is illustrated in Fig. 2.24.

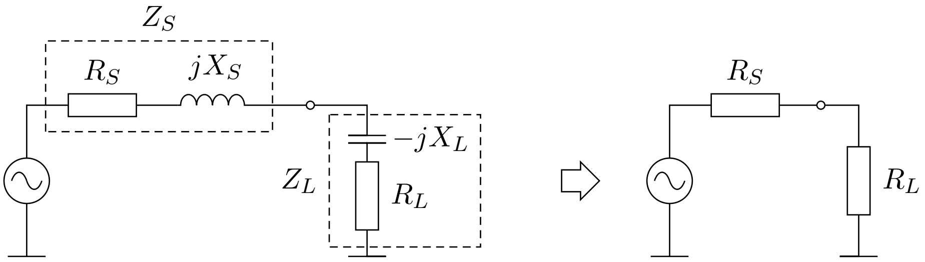


Figure 2.23 Source with an impedance Z_S that is the complex conjugate of the load impedance Z_L.

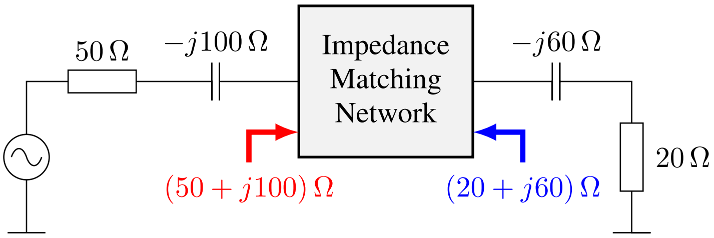


Figure 2.24 Location of matching network to make load look like the complex conjugate of the source impedance.

For wireless communication systems, impedance matching is of high importance. In receivers, where the sensitivity is important, no or little signal power (which is already very small) should be lost. In transmitters, on the other hand, as much of the produced power as possible shall be transferred to the antenna. This is paramount for efficient use of batteries, but also ensures that the power amplifier is not harmed by reflected power. In summary, matching is important due to the following reasons:

- power loss (decrease of sensitivity or efficiency)
- power reflection (heating up amplifier)
- reflection (additional phase)

2.4.1 Parameters Characterizing Reflection and Transfer

Before we talk about the Smith chart, the most practical and intuitive way to visualize network behaviour and to design matching networks, we have to introduce some parameters that are important to RF design in general. There are a number of parameters that characterize the reflection and transfer functions of a two-port. These parameters are usually collected in matrices. Among the most known matrices are

- the $\boldsymbol{Z}$ matrix (impedance),
- the $\boldsymbol{Y}$ matrix (admittance),
- the $\boldsymbol{A}$ matrix (ABCD parameters), and
- the $\boldsymbol{H}$ matrix (hybrid).

They all uniquely describe a linear, time-invariant network. Their determination, however, involves short-circuit and open-circuit measurements, which lead to problems at high frequencies. Besides, transmission lines exist for which voltage and current are meaningless, e.g., waveguides (dt. *Hohlleiter*) and optical fibres (dt. *Glasfasern*).

S-Parameters

In RF and microwave design, so-called *scattering parameters*, collected conveniently in the matrix $\boldsymbol{S}$, are often used. These parameters are determined by loading the ports with a reference impedance (usually 50 Ω) and measuring incident and reflected wave amplitudes (or powers) and phases. The scattering parameters are then used to describe the wave parts that are transferred and reflected at each port.

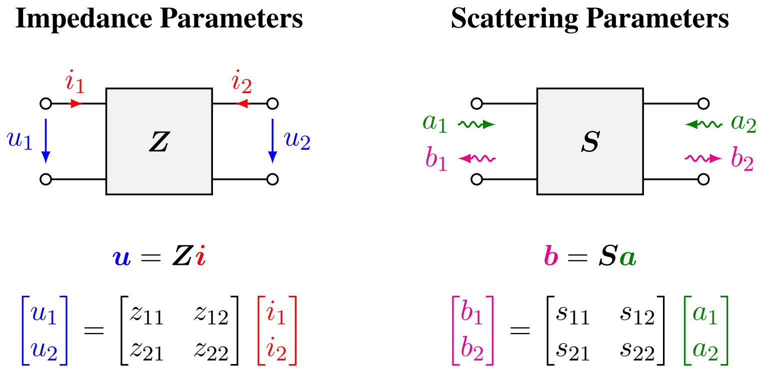

Impedance Parameters
Scattering Parameters
i_1
i_2
u_1
u_2
$\boldsymbol{Z}$
a_1
b_1
a_2
b_2
$\boldsymbol{S}$
$\boldsymbol{u} = \boldsymbol{Z}\boldsymbol{i}$
$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}$
$\boldsymbol{b} = \boldsymbol{S}\boldsymbol{a}$
$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$

Figure 2.25 Two-port parameters: The impedance parameter formulation (left) well-known from circuit theory is replaced by the scattering parameter formulation (right), for RF and microwave engineering. Instead of voltages and currents, these consider incident and reflected waves.

Fig. 2.25 shows a comparison of two-ports, once in terms of impedance parameters (left) and once in terms of scattering parameters (right). As indicated, a_1, b_1 designate the incoming and outgoing waves of port 1 (e.g. the input port), and a_2, b_2 the incoming and outgoing waves of port 2 (e.g. the output port), respectively.

The four scattering parameters are defined as follows:

$$\text{input reflection coefficient:} \qquad s_{11} = \left.\frac{b_1}{a_1}\right|_{a_2=0}, \tag{2.50}$$

$$\text{forward gain:} \qquad s_{21} = \left.\frac{b_2}{a_1}\right|_{a_2=0}, \tag{2.51}$$

$$\text{backward gain:} \qquad s_{12} = \left.\frac{b_1}{a_2}\right|_{a_1=0}, \tag{2.52}$$

$$\text{output reflection coefficient:} \qquad s_{22} = \left.\frac{b_2}{a_2}\right|_{a_1=0}. \tag{2.53}$$

If voltages and currents are meaningful for the two-port under consideration, the reflected waves are usually put in relation to them using

$$a = \frac{U_a}{\sqrt{R_w}} = I_a \cdot \sqrt{R_w}, \tag{2.54}$$

$$b = \frac{U_b}{\sqrt{R_w}} = -I_b \cdot \sqrt{R_w}, \tag{2.55}$$

where U_a, U_b, I_a, and I_b are the incoming effective voltage, the outcoming voltage, the incoming current, and the outcoming current, respectively, and R_w is a (real) reference impedance. Thus, the incoming and reflected waves carry the average powers, $P_a = |a|^2$ and $P_b = |b|^2$, respectively. The scattering matrix, for two-ports it is given by

$$\boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}, \tag{2.56}$$

consists of all the scattering parameters, i.e. from each port to all other ports and itself. It is used to describe the relationship between all incoming and outcoming waves:

$$\boldsymbol{b} = \boldsymbol{S}\boldsymbol{a}. \tag{2.57}$$

The two-port example is mathematically formulated in detail on the left of Fig. 2.25 and illustrated in Fig. 2.26, where white arrows indicate transmitted wave paths and black arrows correspond to reflected paths.

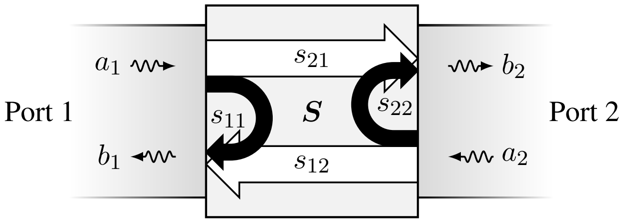


Figure 2.26 Illustration of the physical meanings of the reflection (black arrows) and transmission (white arrows) parameters of a two-port in the scattering matrix.

Reflection

The goal is usually to transfer as much power as possible and to reduce the part that is reflected. As we have already seen, the reflected part heavily depends on the quality of matching, i.e., the better a load impedance matches a source impedance (or the complex conjugate in the complex case), the smaller the reflections are.

The reflection coefficient Γ is the ratio of the reflected and the incoming wave at an interface:

$$\Gamma = \frac{b}{a} \, . \tag{2.58}$$

For a simple one-port, the input reflection coefficient Γ_{in} is simply s_{11} of that one-port, which is equivalent to the load reflection coefficient $\Gamma_L = (R_L - Z_0)/(R_L + Z_0)$, as illustrated in Fig. 2.27. The reflected power follows as

$$|\Gamma_{\text{in}}|^2 = \frac{P_{\text{ref}}}{P_{\text{in}}} = \left.\frac{P_b}{P_a}\right|_{\text{in}} \, . \tag{2.59}$$

Hence, in this example, the power dissipated in the load R_L is $P_L = (1 - |\Gamma_{\text{in}}|^2)P_{\text{in}}$, while the reflected power is $P_{\text{ref}} = |\Gamma_{\text{in}}|^2 \, P_{\text{in}}$.

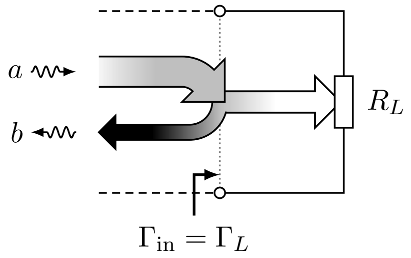


Figure 2.27 Illustration of the reflection (black arrow) and transmission (white arrow) at a (potentially mismatched) load R_L. As this is a one-port scenario (i.e. there is only one interface), the input reflection coefficient Γ_{in} is equal to the load reflection coefficient Γ_L.

The input reflection of a two-port connected to a certain load (again given by the reflection factor Γ_L) is given by

$$\boxed{\Gamma_{\text{in}} = s_{11} + \frac{s_{12}\, s_{21}\, \Gamma_L}{1 - s_{22}\, \Gamma_L}} \tag{2.60}$$

and illustrated in Fig. 2.28. As can be seen, only if the output load of the two-port is matched to the system impedance, in other words, $\Gamma_{\text{L}} = 0$, the input reflection is simply $\Gamma_{\text{in}} = s_{11}$. In general, for reflection factors $\Gamma_{\text{L}} \neq 0$, this is only true if either $s_{21} = 0$ (no transmission, usually not reasonable) or $s_{12} = 0$ (perfect isolation).

Return Loss

The term “return loss” (dt. *Reflexionsdämpfung, Rückflussdämpfung*) is somewhat of a misnomer, and therefore often confusing. Return loss, measured in dB, expresses the forwarded power divided by the reflected power. The goal is to make it as large as possible. A return loss of infinite dB represents a very well matched case, whereas 0 dB means that all power is reflected. The return loss of a network or a component is the

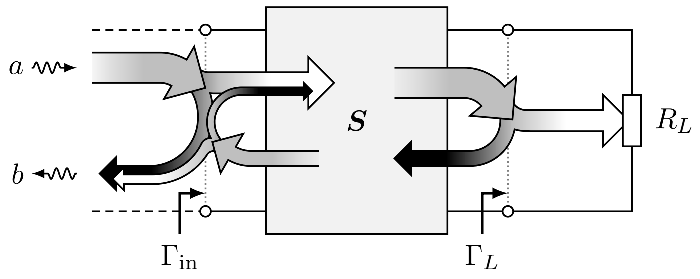


Figure 2.28 Illustration of the reflections (black arrows) and transmissions (white arrows) at one side of a two-port $\boldsymbol{S}$ when the other is connected to a load R_L. The input reflection coefficient Γ_{in} and the load reflection coefficient Γ_L are defined at two different interfaces; they are connected via the S-parameters of the two-port, according to (2.60).

amount of power that is 'lost' to the load and does not return as a reflection. Most often, the goal is to maximize the return loss. Mathematically the return loss is given as

$$\text{RL} = -10\log_{10}|\Gamma_{\text{in}}|^2 = -20\log_{10}|\Gamma_{\text{in}}|\,. \tag{2.61}$$

VSWR

A slightly different view of the reflection relations is given by the use of the voltage standing wave ratio or short VSWR. The VSWR expresses the ratio of maximum and minimum voltage of a standing wave pattern, as shown in Fig. 2.29:

$$\text{VSWR} = \frac{u_{\max}}{u_{\min}}\,. \tag{2.62}$$

As opposed to the return loss, the VSWR is infinite for total reflection and one for matched impedances.

Fig. 2.30 shows the different quality ranges of matching and the corresponding return loss and VSWR figures. Up to a return loss of 0 to 10 dB, impedances are considered not matched or not aligned and above 20 dB they are considered matched or aligned. Table 2.4 shows the relationship of the loss parameters for general load impedances Z.

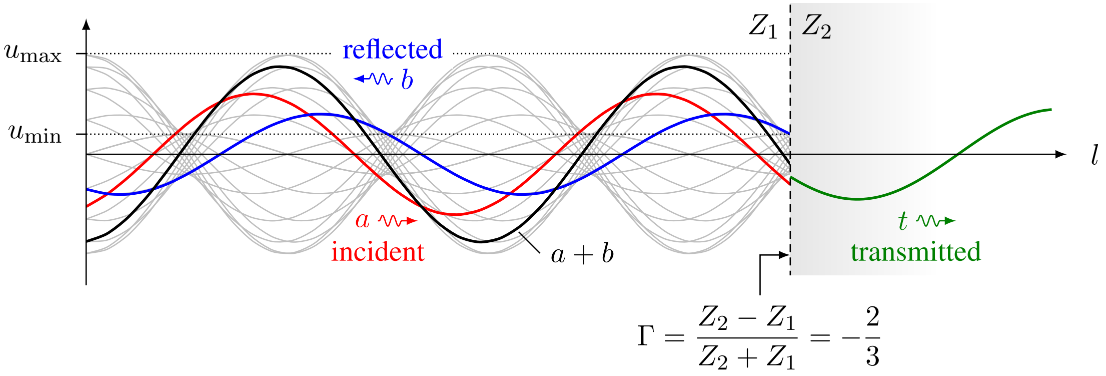


Figure 2.29 The meaning of the voltage standing wave ratio (VSWR).

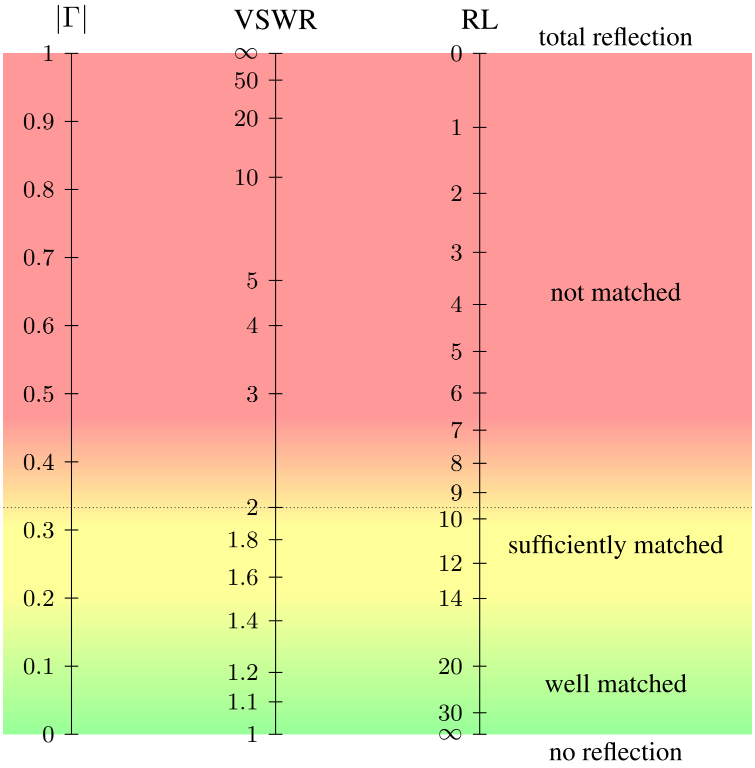


Figure 2.30 Conversion of reflection coefficient, return loss, and VSWR.

	$f(Z)$	$f(\Gamma)$	$f(\text{VSWR})$	$f(\text{RL})$
$Z =$	—	$\dfrac{1+\Gamma}{1-\Gamma}\,Z_0$	—	—
$\Gamma =$	$\dfrac{Z-Z_0}{Z+Z_0}$	—	$\lvert\Gamma\rvert = \dfrac{\text{VSWR}-1}{\text{VSWR}+1}$	$\lvert\Gamma\rvert = 10^{-\text{RL}/20}$
$\text{VSWR} =$	$\dfrac{1+\left\lvert\frac{Z-Z_0}{Z+Z_0}\right\rvert}{1-\left\lvert\frac{Z-Z_0}{Z+Z_0}\right\rvert}$	$\dfrac{1+\lvert\Gamma\rvert}{1-\lvert\Gamma\rvert}$	—	$\dfrac{1+10^{-\text{RL}/20}}{1-10^{-\text{RL}/20}}$
$\text{RL} =$	$20\,\log_{10}\left\lvert\dfrac{Z+Z_0}{Z-Z_0}\right\rvert$	$-20\,\log_{10}\lvert\Gamma\rvert$	$20\,\log_{10}\dfrac{\text{VSWR}+1}{\text{VSWR}-1}$	—

Table 2.4 Conversion of the parameters (valid for $|\Gamma| \leq 1$).

Example: VSWR

Question: What are the return loss and the VSWR when the reflected signal carries 25% of the incident power at the discontinuity of a lossless cable?

Answer: The return loss is

$$|\Gamma_{\text{in}}|^2 = \frac{1}{4} \quad \Rightarrow \quad |\Gamma_{\text{in}}| = \frac{1}{2} \quad \Rightarrow \quad \text{RL} = -20\ \log_{10} |\Gamma_{\text{in}}| = \underline{\underline{6\,\text{dB}}}\,.$$

The standing-wave ratio is

$$\text{VSWR} = \frac{1+|\Gamma_{\text{in}}|}{1-|\Gamma_{\text{in}}|} = \frac{1+\frac{1}{2}}{1-\frac{1}{2}} = \frac{2+1}{2-1} = \frac{3}{1} = \underline{\underline{3}}\,.$$

Insertion Loss

The insertion loss (IL) designates the loss in dB a signal experiences by going through a two-port. IL basically measures the additional loss between a source and a load when inserting the device under test. Hence, the insertion loss can be written as

$$\text{IL} = -20\log_{10}|s_{21}|. \tag{2.63}$$

2.4.2 Properties of Twoports

Twoports as introduced above can have one or several of the following properties:

- **Reciprocity**
  A two-port is reciprocal if the voltage arising at port 2 due to a current applied at port 1 is the same as the voltage appearing at port 1 when the same current is present at port 2. The same holds when voltage and current are exchanged.

  Generally, passive networks (consisting only of R, L, C and transmission lines) are reciprocal. On the other hand, generally, active networks (containing amplifiers, generators etc.) are not reciprocal.

  If a two-port is reciprocal, we have $s_{21} = s_{12}$ of the S-parameters.

- **Symmetry**
  Symmetric two-ports are reciprocal networks with the additional property of $s_{11} = s_{22}$, meaning the input impedance is equal to the output impedance.

- **Lossless**
  For lossless, passive networks the matrix of the scatter parameters needs to be unitary

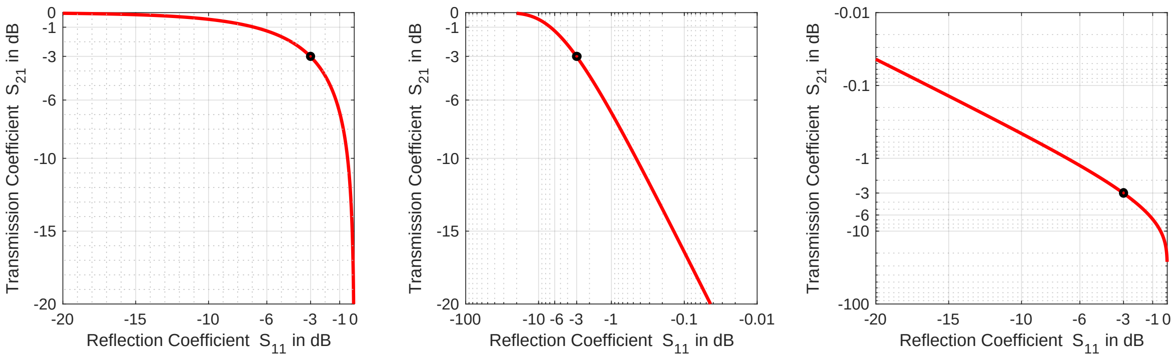
  $$\boldsymbol{S}^H = \boldsymbol{S}^{-1}. \tag{2.64}$$

  For the individual scatter parameters this means

  $$|s_{11}|^2 + |s_{21}|^2 = 1\ , \tag{2.65a}$$

  $$|s_{22}|^2 + |s_{12}|^2 = 1\ , \tag{2.65b}$$

  $$s_{11}^* s_{12} + s_{21}^* s_{22} = 0\ . \tag{2.65c}$$



(a) Linear dB axes (b) Logarithmic dB axis for s_{11} (c) Logarithmic dB axis for s_{21}

Figure 2.31 The relationship between s_{11} and s_{21} in a passive lossless twoport.

In the case of a passive, lossless twoport we can compute the insertion loss as a function of the return loss using Eq. (2.65a). Using s_{21} as the insertion loss IL and s_{11} as the return loss RL, both in dB, we can write

$$\mathrm{IL} = -10 \cdot \log_{10}\left(1 - 10^{\frac{-\mathrm{RL}}{10}}\right). \tag{2.66}$$

Plots of this relationship with different axes are shown in Fig. 2.31. The leftmost chart shows the symmetric relationship between the parameters. The middle chart shows that an improvement of the insertion loss is only marginal, once the return loss is below a certain value, say 10 dB. However, for a limited return loss (expressing limited matching) of 6 dB or less, we get huge differences in the insertion loss.

Cascade of Twoports

If several two-ports are cascaded, the input-reflected wave of the second two-port becomes the output-incident wave of the first two-port and vice versa. For an illustration of this connection, see Fig. 2.32. The scattering matrix of cascaded two-ports cannot be directly computed by the scattering matrices of the individual two-ports. Rather, we need to compute the transfer matrix

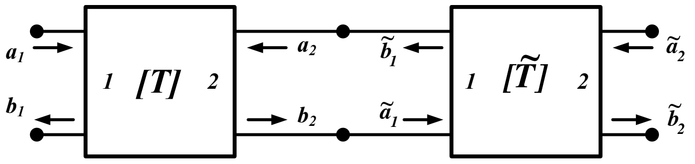


Figure 2.32 Cascaded two-ports.

$$\boldsymbol{T} = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix}, \tag{2.67}$$

which describes the incoming and outcoming waves in the following way,

$$\begin{aligned} \begin{bmatrix} b_1 \\ a_1 \end{bmatrix} &= \boldsymbol{T} \cdot \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \\ &= \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} \begin{bmatrix} a_2 \\ b_2 \end{bmatrix}. \end{aligned} \tag{2.68}$$

The transfer matrices can simply be multiplied for a cascade of two-ports, thus they are also called wave chain matrix. Hence, for the cascade of Fig. 2.32 we get

$$\boldsymbol{T}_{\text{total}} = \boldsymbol{T} \cdot \widetilde{\boldsymbol{T}}. \tag{2.69}$$

Conversions between scattering matrix and transfer matrix are possible using

$$\boldsymbol{S} = \frac{1}{t_{22}} \begin{bmatrix} t_{12} & \det \boldsymbol{T} \\ 1 & -t_{21} \end{bmatrix}, \tag{2.70}$$

$$\boldsymbol{T} = \frac{1}{s_{21}} \begin{bmatrix} -\det \boldsymbol{S} & s_{11} \\ -s_{22} & 1 \end{bmatrix}. \tag{2.71}$$

2.4.3 The Smith Chart

Introduction

The Smith chart is a tool often used in RF design and was invented by Phillip Smith (1905–1987) at the Bell Labs back in the Thirties. The first published report of Smith's work appeared in an IRE article by E. J. Sterba and C. B. Feldman in 1932 [76], reading *"There is another effective way for transforming line impedances by means of short line devices" and a footnote saying "Disclosed to the writers by P. H. Smith".* It would take him another seven years before the reviewers finally accepted a full publication on the matter. In 1939, he introduced his so-called "Transmission Line Calculator" [72] and gave some additional formulations and increased accuracy towards the End of the Second World War in 1944 [73]. At that time, his technique has already been a standard tool for antenna and transmission line matching at Bell Labs and other institutions.

The Smith chart, as his graphical method was called later on, was not an ingenious idea as one could easily assume, but rather Smith's persistent search to find the ultimate solution to all transmission line problems. Although, nowadays, the underlying principle appears to be almost straightforward, because of lack of advanced measurement equipment (e.g. VNA) and deeper insight into the matter at that time, it took him several years and probably hundreds of cases of application to finish his work to his own satisfaction.

Construction

As can be found in Tab. 2.1, the voltage reflection coefficient can be received from the load impedance Z and the characteristic impedance of the transmission line Z_0 according to

$$Z = Z_0 \frac{1+\Gamma}{1-\Gamma} \, . \tag{2.72}$$

More general, by introducing z as the impedance ratio, normalized with respect to the impedance of the transmission line, it follows

$$z \triangleq \frac{Z}{Z_0} = \frac{1+\Gamma}{1-\Gamma} \tag{2.73}$$

and, as can easily be shown, vice versa,

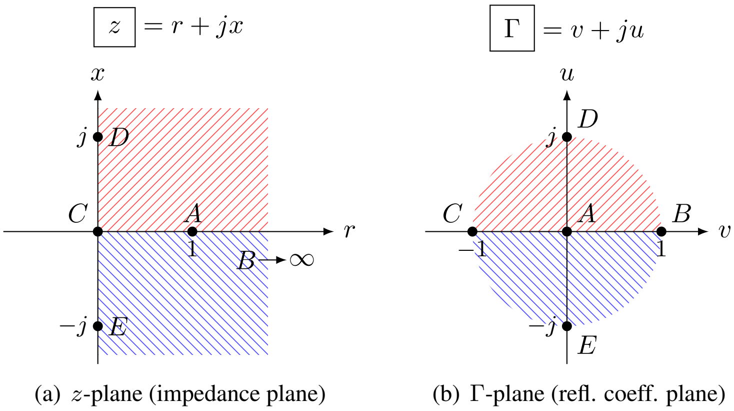
$$\Gamma = \frac{z-1}{z+1} \, . \tag{2.74}$$

Note that both z and Γ can be (or normally are) complex-valued,

$$z = r + jx \; , \tag{2.75a}$$

$$\Gamma = u + jv \; . \tag{2.75b}$$

The quantities z and Γ can then be plotted in their respective complex planes, as shown in Fig. 2.33, where the plot of this Γ-plane was Smith's primary idea.



(a) z-plane (impedance plane) (b) Γ-plane (refl. coeff. plane)

Figure 2.33 The mapping between the z-plane and the Γ-plane.

The region $r > 0$ in the impedance plane is mapped into the interior of the unit circle in the reflection coefficient plane. Also, since Eq. (2.74) is a *bilinear transformation* of z, circles[8] in the z-domain are mapped into circles again in the Γ-domain. Even further, since the mapping is *conformal*, angles between two line segments stay the same after the transformation.

Most interesting are the mappings of constant resistance and constant reactance cases. From Eqs. (2.73) and (2.75) follows

$$z = \frac{1+u+jv}{1-u-jv} = \frac{(1+u+jv)(1-u+jv)}{(1-u-jv)(1-u+jv)} = \frac{(1-u^2)+2jv-v^2}{(1-u)^2+v^2} \; , \tag{2.76}$$

$$r = \operatorname{Re} z = \frac{1-u^2-v^2}{(1-u)^2+v^2} \; , \tag{2.77}$$

$$x = \operatorname{Im} z = \frac{2v}{(1-u)^2+v^2} \; . \tag{2.78}$$

By completing the square, Eq. (2.77) can be written as

$$\left(u - \frac{r}{r+1}\right)^2 + v^2 = \left(\frac{1}{r+1}\right)^2 . \tag{2.79}$$

which represents a circle in the Γ-domain with its center at $(r/(r+1), 0)$ and a radius of $|1/(r+1)|$. The same can be done with Eq. (2.78), resulting in

$$(u-1)^2 + \left(v - \frac{1}{x}\right)^2 = \left(\frac{1}{x}\right)^2 , \tag{2.80}$$

[8] Straight lines can be considered circles with infinite radii, which explains why straight lines (for example the imaginary axis) in the z-plane map into circles in the Γ-plane.

which is also a circle, this time centered at $(1, 1/x)$ and with a radius of $|1/x|$. Variations of both circle types are shown in Fig. 2.34 , where circles of the first kind (constant r) are set in blue and those of the second kind (constant x) in red and the corresponding values for r and x are given.

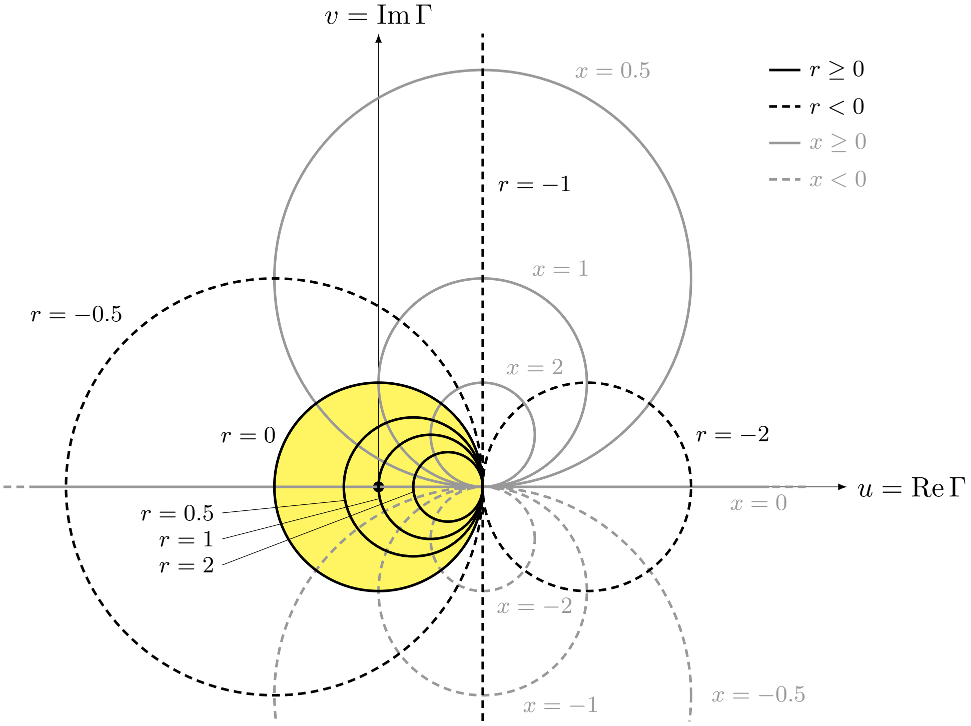
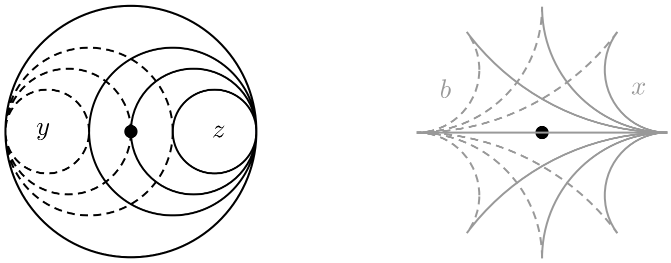


Figure 2.34 Construction of the Smith chart.

The area inside the unit circle (yellow shaded) is known as the Smith chart, e.g. as in Fig. 2.36. As can be seen, all circles of constant positive resistance (black circles) are inside the unit circle, whereas all circles of constant negative resistance (which can only be received with active devices) are outside. Also, since positive reactances are due to inductances (solid gray circles), the upper half is called the inductive area. Thus, the lower half, containing all negative reactances (dashed gray circles), evidently is the capacitive area (see also Fig. 2.33). Fig. 2.36 shows a Smith chart with impedance circles and arcs only.



(a) circles of constant resistance and conductance

(b) circles (arcs) of constant reactance and susceptance

Figure 2.35 Impedance and admittance circles.

If the normalized impedance z is replaced by a normalized admittance $y = Y/Y_0$ (where $Y_0 = 1/Z_0$ etc.),

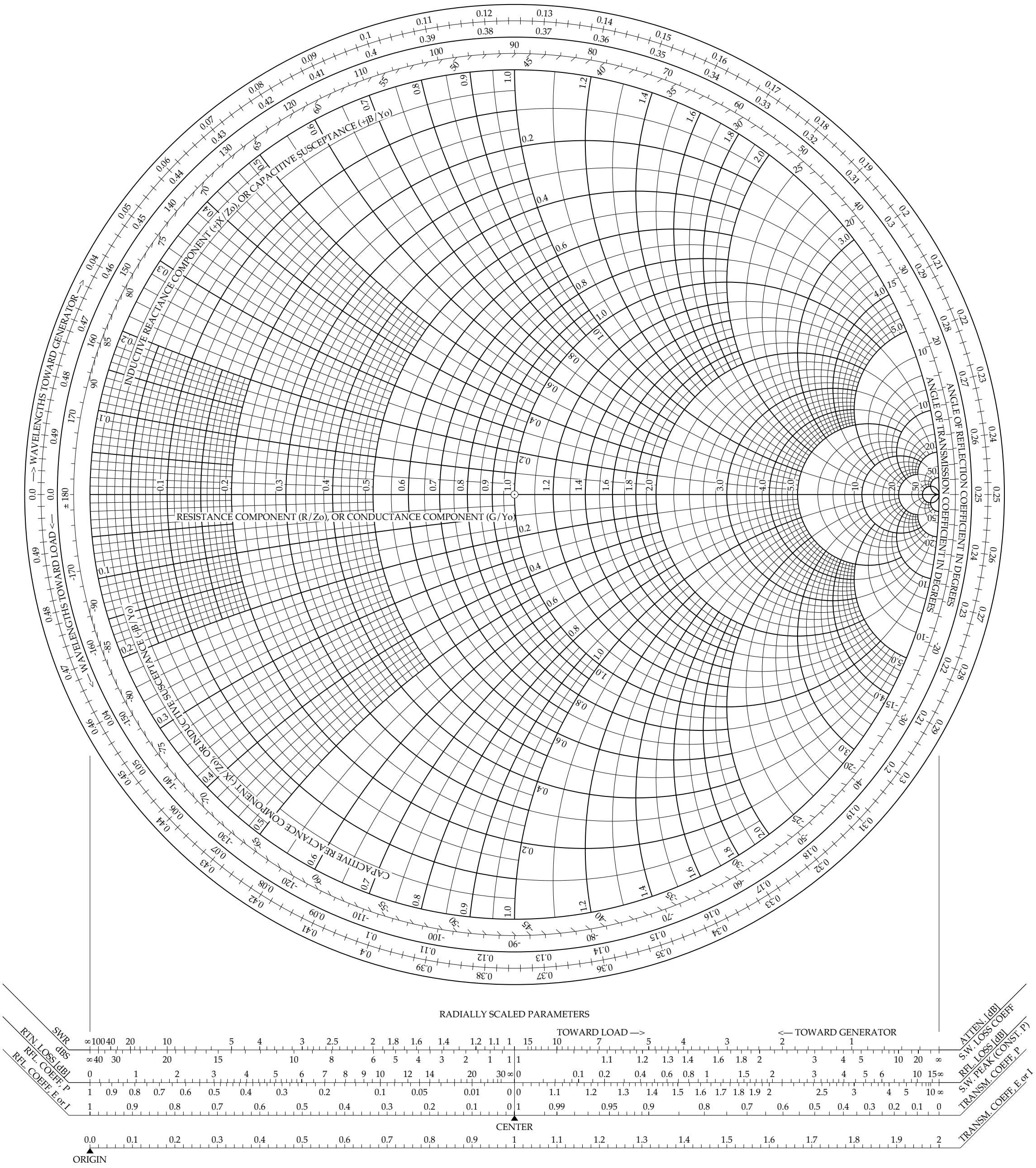


Figure 2.36 The standard impedance Smith chart (by Black Magic Design).

Eq. (2.73) becomes

$$y = \frac{1-\Gamma}{1+\Gamma} \tag{2.81}$$

and solved for Γ, instead of Eq. (2.74), it follows

$$\Gamma = \frac{1-y}{1+y}\ . \tag{2.82}$$

Eq. (2.81) is the same as Eq. (2.73) if Γ is replaced by $-\Gamma$, meaning the admittance circles can be found by rotating the chart around the origin $(0, 0)$ by 180 degrees. Fig. 2.35 shows the relationship between the so obtained impedance and admittance circles.

Whereas the impedance chart is designed for evaluating the effect of a series component, the admittance chart is the right tool for observing the effect of a parallel element. Most modern charts show both kind of circles, e.g., Fig. 2.37.

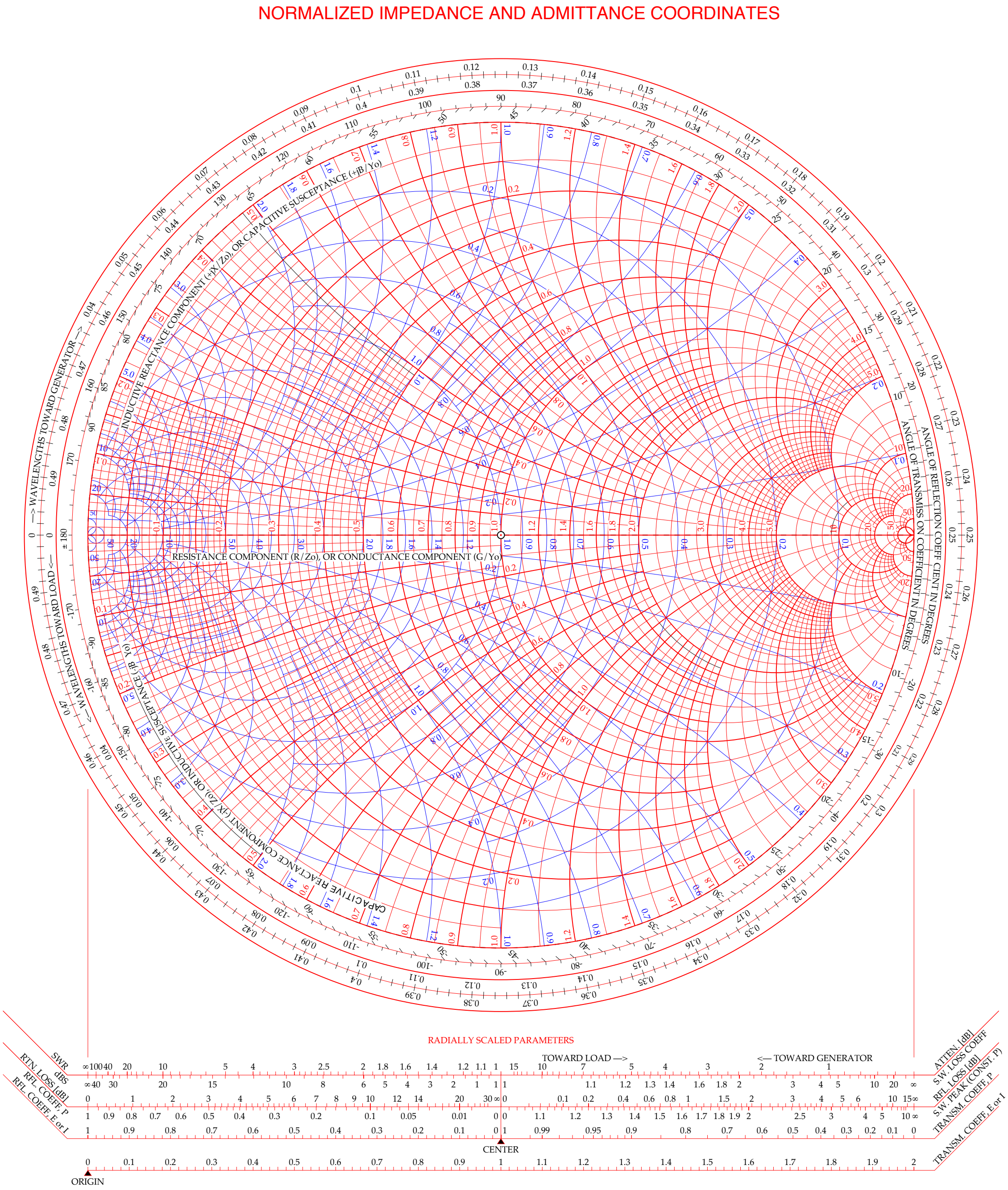

NORMALIZED IMPEDANCE AND ADMITTANCE COORDINATES
INDUCTIVE REACTANCE COMPONENT (+jX/Zo), OR CAPACITIVE SUSCEPTANCE (+jB/Yo)
CAPACITIVE REACTANCE COMPONENT (-jX/Zo), OR INDUCTIVE SUSCEPTANCE (-jB/Yo)
RESISTANCE COMPONENT (R/Zo), OR CONDUCTANCE COMPONENT (G/Yo)
ANGLE OF REFLECTION COEFF CIENT IN DEGREES
ANGLE OF TRANSMISS ON COEFFICIENT IN DEGREES
—> WAVELENGTHS TOWARD GENERATOR —>
<— WAVELENGTHS TOWARD LOAD <—
RADIALLY SCALED PARAMETERS
TOWARD LOAD —>
<— TOWARD GENERATOR
SWR
dBS
RTN. LOSS [dB]
RFL. COEFF, P
RFL. COEFF, E or I
ATTEN. [dB]
S.W. LOSS COEFF
RFL. LOSS [dB]
S.W. PEAK (CONST. P)
TRANSM. COEFF, P
TRANSM. COEFF, E or I
CENTER
ORIGIN

Figure 2.37 Smith chart showing impedance/reactance (red) and admittance/susceptance (blue) circles.

Basic Operations

Fig. 2.38 depicts the directions some fundamental elements will change the reflection factor to. The Smith chart allows the design of impedance matching circuits scaled to a certain system impedance (Z_0, mostly $50\,\Omega$) by simple graphical considerations.

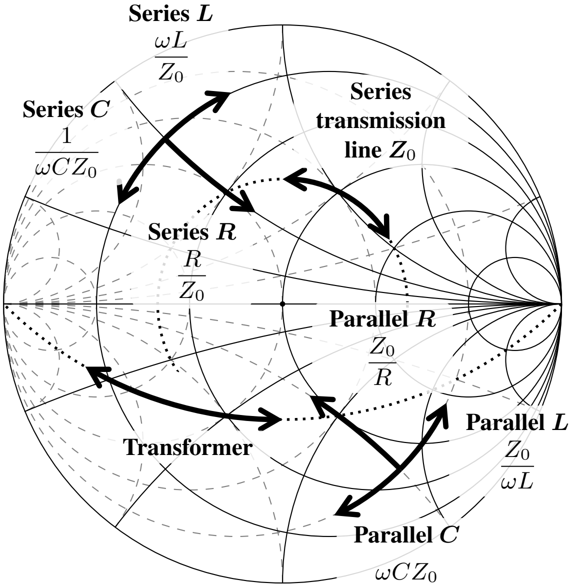


Figure 2.38 Smith chart and driving directions of circuit elements.

Examples: Finding load impedance points

The process of finding the right points in the Smith chart for specific load impedances are best explained by some examples:

- An impedance of $Z = 150\,\Omega$ results in

$$z = \frac{150\,\Omega}{50\,\Omega} = 3$$

in a $50\,\Omega$-system. The point in the Smith chart is then found by looking for the impedance circle $r = 3$ and thereon finding the point where $x = 0$, which is on the horizontal line. Γ (and therefore s_{11}) for purely resistive loads are depicted in Fig. 2.39(a).

- A purely inductive load gives

$$z(\omega) = \frac{j\omega L}{Z_0} \, ,$$

where for $z(\omega = 0) = 0$, thus $\Gamma = -1$ (short), and for $z(\omega \to \infty) = \infty$, therefore $\Gamma = 1$ (open), see Fig. 2.39(b).

- A purely capacitive load results in

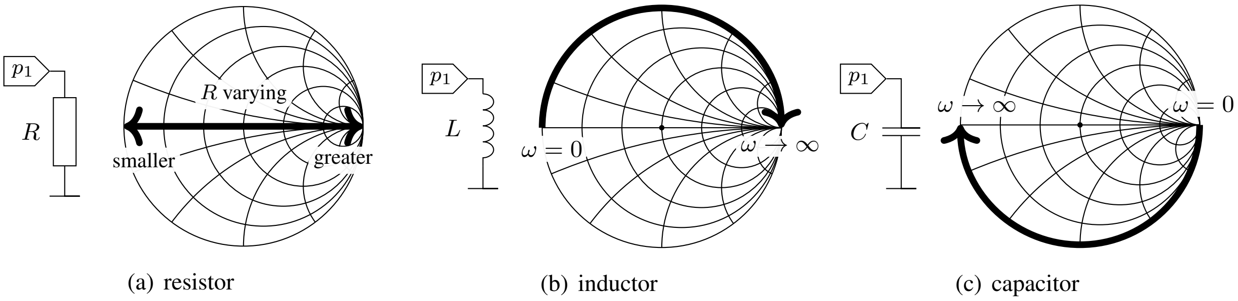
$$z(\omega) = \frac{-j}{\omega C\, Z_0} \, ,$$

for which $\Gamma|_{\omega=0} = 1$ (open end) and $\Gamma|_{\omega=\infty} = -1$ (short circuit), see Fig. 2.39(c).

- A complex impedance of $Z = 10\,\Omega - j20\,\Omega$ results in a normalized impedance of

$$z = \frac{10\,\Omega - j20\,\Omega}{50\,\Omega} = 0.2 - 0.4j\ .$$

This time, the point on the Smith chart is on the impedance circle $r = 0.2$, cut with the reactance circle $x = -0.4$. Evidently, since the point is on the lower half, the impedance is capacitive in this case. Frequency dependent curves of basic circuits are shown in Fig. 2.40. Because the resistive component is always $Z = Z_0 = 50\,\Omega$, the curves are always on $r = 1$ or $1/r = g = 1$ circles.



(a) resistor (b) inductor (c) capacitor

Figure 2.39 Γ for single circuit elements.

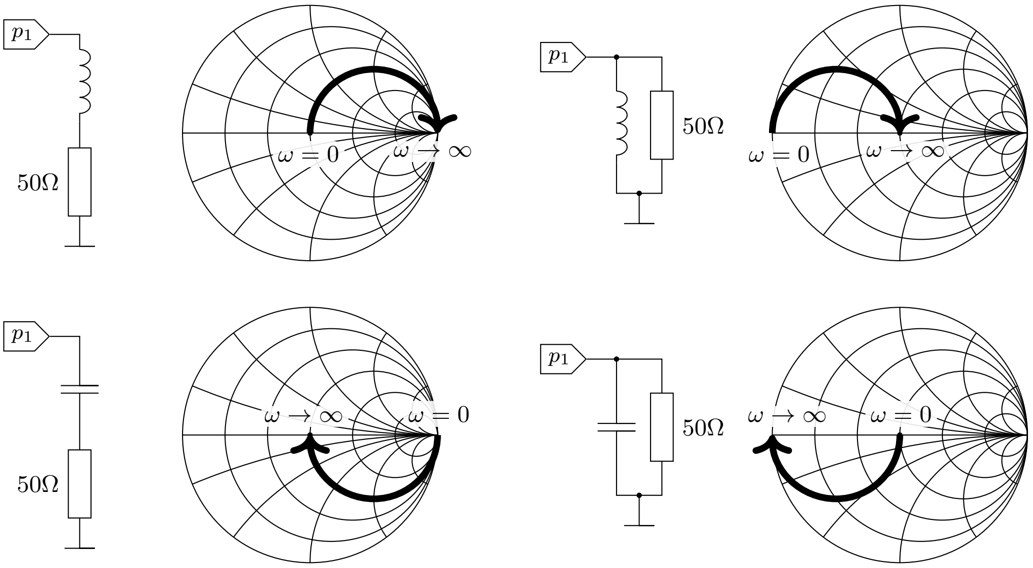
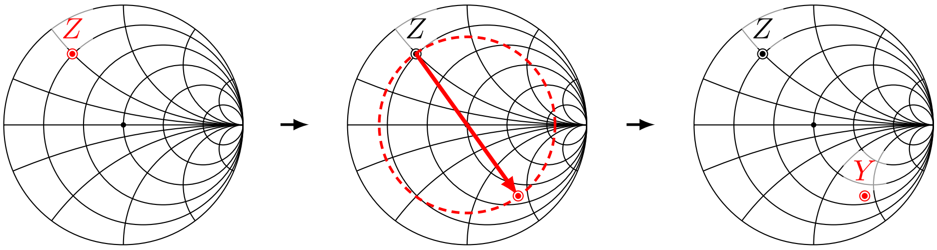


Figure 2.40 Γ for some simple circuits.

Transformations

- **Impedance ↔ admittance:** The normalized admittance can be found by writing the normalized impedance into the bent coordinate system of real and imaginary parts, mirroring that point at the origin and reading the normalized admittance from the bent coordinates (of the impedance grid). Fig. 2.41 illustrates this process. The reverse transformation is exactly the same.



(a) arbitrary impedance

(b) transformation procedure (mirroring)

(c) transformed admittance

Figure 2.41 Transformation impedance-admittance.

It is even simpler if the Smith chart also contains admittance lines: then, instead of reading the impedance from the circles from right to left, the admittance value can be read from the circles from left to right.

- **Impedance ↔ reflection:** From the normalized impedance we get to the reflection by writing the normalized impedance into the bent coordinate system of real and imaginary parts and reading Γ (sometimes r for reflection factor[9]) in polar coordinates. At the outer scale we find angles and the absolute value of Γ can be compared against the linear scale at the top of the diagram of a usual Smith chart. Fig. 2.42 illustrates this process. To transform reflection into the normalized impedance, we reverse the procedure.

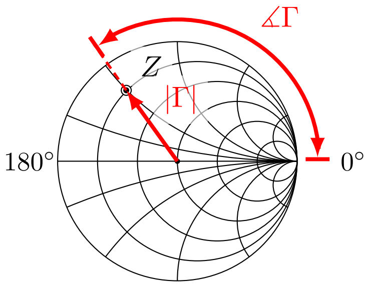


Figure 2.42 Transformation impedance-reflection.

- **Admittance ↔ reflection:** From the combination of the above two procedures it becomes clear how this transformation has to be carried out.

[9] Note that care must be taken not to mistake the reflection coefficient for the real part of the normalized impedance, r, mentioned some pages earlier.

2.4.4 Impedance Matching by Lumped Networks

Returning to the original problem: If a certain load does not 'fit' or 'match' the source, we need to do an impedance matching. Since the reactances of components are frequency dependent, the matching is exact for one frequency only. Depending of the desired bandwidth, there are different matching networks at our disposal.

- **L-network:** The easiest matching network is the L-network. By the rule, the shunt element (in parallel) must be adjacent to the impedance with the higher resistive part. It will be designed in such a way as to decrease the real part of the high resistivity. The series element will subsequently compensate the reactive part. As an example we look at Fig. 2.43.

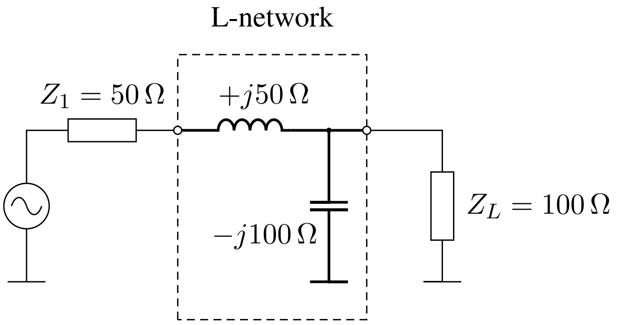
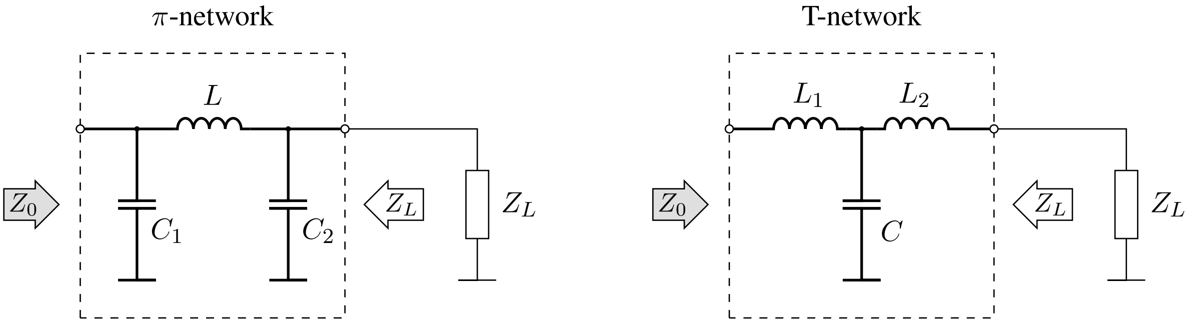


Figure 2.43 Example of impedance matching through an L-network.

- **Pi-network:** A Pi-network (see Fig. 2.44(a)) is a three-element matching circuit and can be considered as a composition of two L-networks. The source impedance is transformed down to a lower virtual impedance and subsequently transformed up to the load impedance. The choice of the virtual impedance is free, thus many different networks that carry out a certain transformation are possible. Decisive for the choice is often the bandwidth desired. In general, the bandwidth is smaller for Pi- and T-networks than for ordinary L-networks.



(a) Matching by using a Pi-network, also called Collins filter.

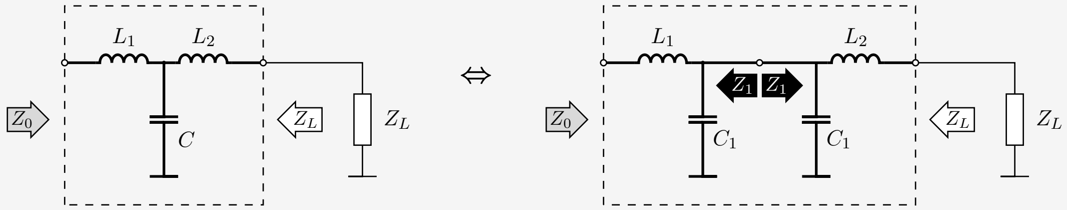
(b) Matching by using a T-network.

Figure 2.44 Narrow-band matching networks.

- **T-network:** The complement to a Pi-network is a T-network (see Fig. 2.44(b)). Again, it can be looked at as the combination of two L-networks, but in inverse order. This time, the source impedance is being transformed to a higher virtual impedance and subsequently transformed down the load impedance.

Explanation: Why Pi/T-networks show narrower bandwidth

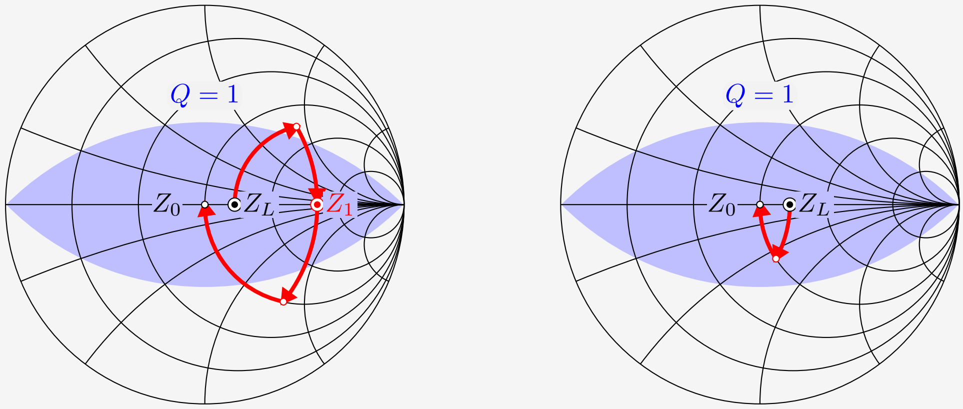
A T-network can be split up into two L-networks (the same holds for Pi-networks):



As shown in Fig. 2.45, below the regions of impedance matching possibilities, the first L-network matches onto a larger impedance, implying $|Z_1| > |Z_0|$, and the second matches from a larger impedance onto a smaller one, thus, $|Z_1| > |Z_L|$. This states that

$$|Z_1| > |Z_0|\,, |Z_L|\ .$$

Thus, the impedance on the left Z_0 is first matched to a larger impedance Z_1, only to be matched to a lower impedance Z_L again. The following graph shows the matching curves in the Smith chart:



Matching by T-network

Matching by L-network

With T-network matching, parts of the matching curves lie in regions of greater Q than for example for an L-network for the same impedances, thus T-networks (and Pi-networks) have a greater overall quality factor Q.

The greater Q states that there is more stored reactive energy compared to a case of lower Q. This is due to the fact that the required reactive energy to match a certain impedance onto another is related to the mismatch of the impedances in the first place. Thus, it takes more reactive energy for more differing impedances, as in the case for $Z_0 \leftrightarrow Z_1$ and $Z_L \leftrightarrow Z_1$, compared to the one of $Z_0 \leftrightarrow Z_L$.

And since the bandwidth of a matching network (or any filter network) is related to the quality factor Q according to

$$B = \frac{f_{\text{upper}} - f_{\text{lower}}}{f_{\text{center}}} \propto \frac{1}{Q}$$

(for large Qs), a large Q results in a narrow bandwidth.

- **Cascade of L-networks:** By cascading L-networks of equal direction, we can achieve matching for a wider bandwidth.

Explanation: Why cascaded L-networks have wider bandwidth

The total quality factor Q_{tot} of cascaded networks is the "parallel circuit" of all separate Q_n (i.e. quality factor of the nth network):

$$Q_{\text{tot}} = Q_1 \parallel Q_2 \parallel \cdots \parallel Q_n = \frac{1}{\frac{1}{Q_1} + \frac{1}{Q_2} + \cdots + \frac{1}{Q_n}}$$

Thus, it is easy to see that the total quality factor of cascaded L-networks (of the same kind) have lower Q than each one by itself.

$$Q_{\text{tot}} < Q_1, Q_2, ..., Q_n$$

Thus, cascaded L-networks have greater bandwidth than single L-networks, and accordignly for every other cascaded network types.

Of course, all these networks can be computed analytically. In practice, however, in order to design matching networks, one tool has proved invaluable which we shall discuss in the following: the Smith chart. It allows the design of matching circuits in a very easy, intuitive way. Fig. 2.45 shows some typical L-type networks to match impedances and their associated possible areas.

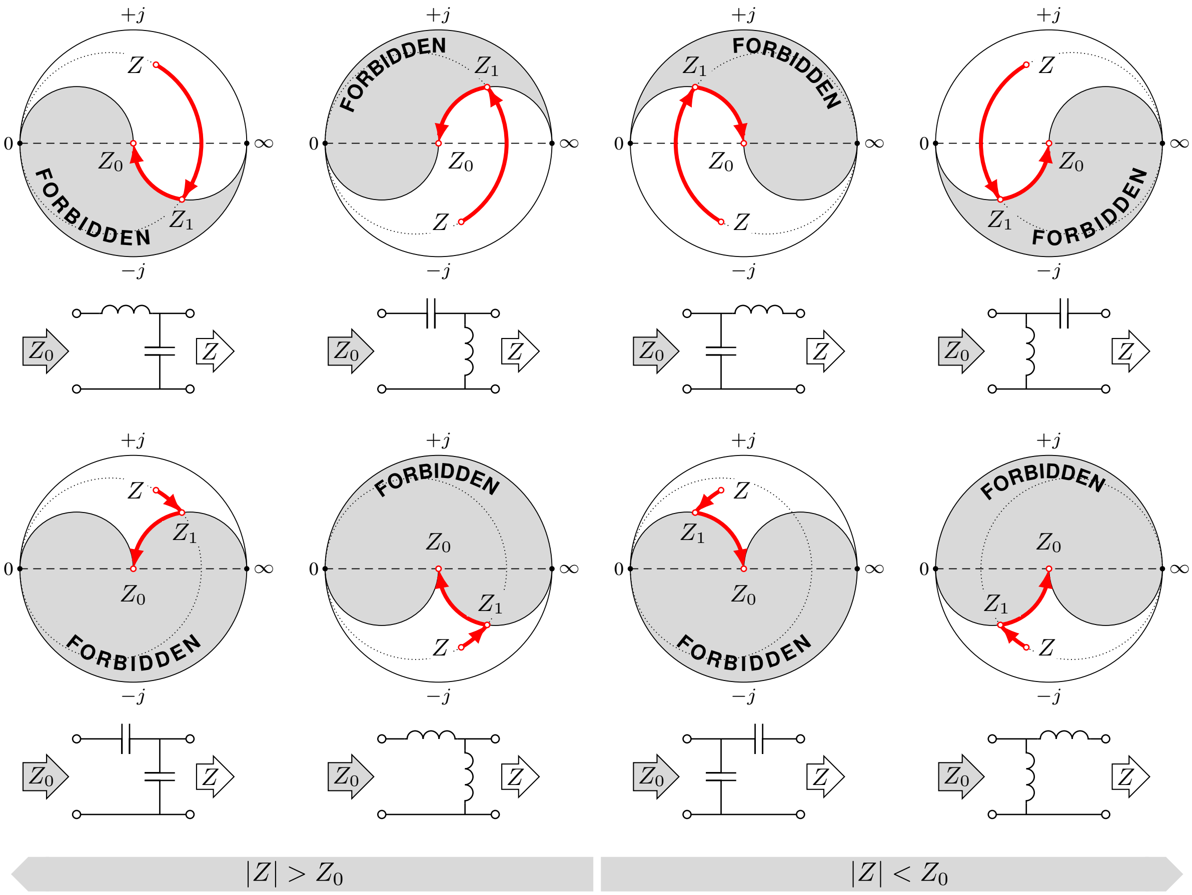
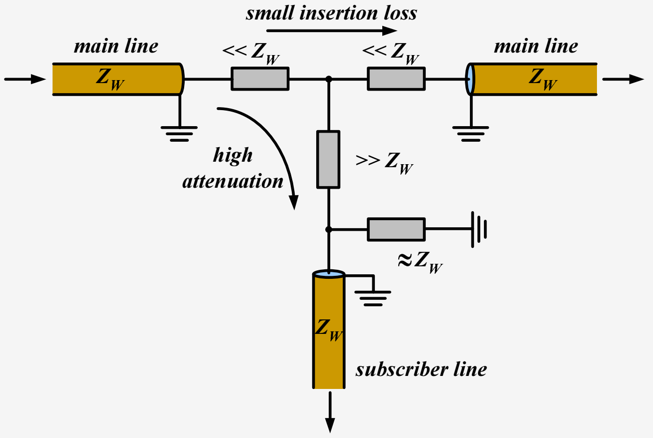


Figure 2.45 Impedances in the shaded regions of the Smith charts are transformable to a purely resistive load Z_0 with the given L-type networks.

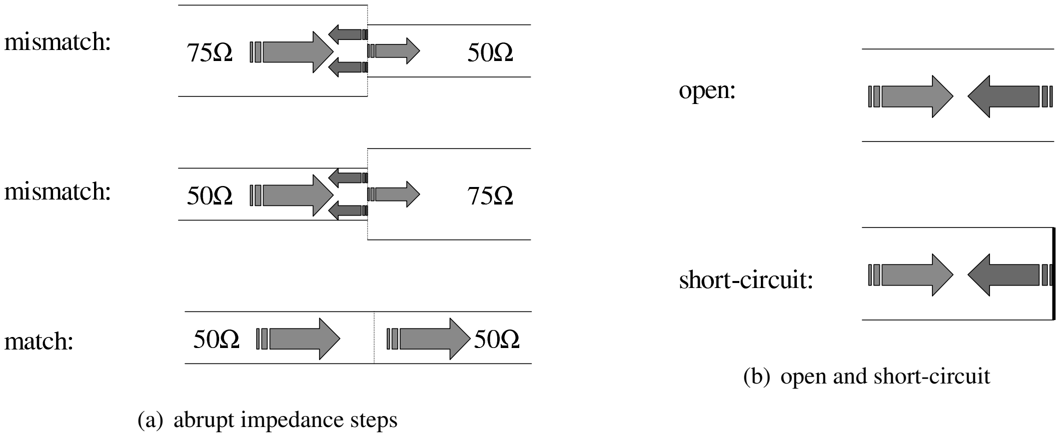
Example: Matching by high attenuation network branches

In some systems, 'matching' (in other words, reducing the reflections) is achieved by hooking up new network branches through high attenuation. An example of this can be seen on the right, where a subscriber's line is connected to a CATV network.
Note: the impedance seen through the line is largely independent of the subscriber's line determination.



2.4.5 Matching of Transmission Lines

All the matching issues of linear two-ports using lumped elements equally apply to cables. In order to maximize the power throughput of a cable, we need to minimize reflections. Cables are like water pipes. If we connect two different types of pipes, water might get lost. Similarly, if we connect two cables of different impedances, some part is always reflected at the discontinuity. Let's have a look at some examples: Connecting a 50-Ω to a 75-Ω cable only allows some part of the power to go through. The rest is reflected and therefore lost. In the inverse case, the reflected power behaves exactly the same way. It is therefore very important to match cable impedances. If two cables of the same impedance are connected, 100 % of the power is transferred. All three cases are shown in Fig. 2.46(a)



(a) abrupt impedance steps

(b) open and short-circuit

Figure 2.46 Reflections in cables due to different cable impedances.

Let us also have a look at two extreme cases as illustrated in Fig. 2.46(b). If a cable is left open, 100 % of the power is reflected. Likewise, if a cable is left short-circuited, 100 % of the power returns to the source. No power transmission is possible in either case.

2.4.6 Matching by Using Transmission Lines

We have seen that connecting cables of different impedances might cause reflections. To counteract this, matching is mandatory. On the other hand, transmission lines might be used instead of discrete components to match two-ports of different characteristic impedances. The reasons for this is that at higher frequencies, parasitics of components may not be easily controlled. Moreover, the size of components (inductor, capacitor) becomes inconvenient. Fig. 2.47 depicts the set-up of impedance matching using a transmission line.

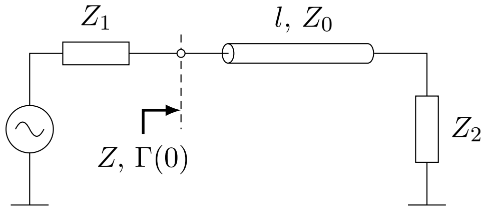


Figure 2.47 Matching by a transmission line.

Transformation of an Impedance Through a Transmission Line

In the following, we designate the beginning of the transmission line (left end) as $z = 0$ and the end as $z = l$. If we put an impedance Z_2 at one end of a transmission line of length l, the incident and reflected wave (U_a and U_b) have to behave according to

$$Z_2 = \frac{U_a(l) + U_b(l)}{I_a(l) - I_b(l)} = \frac{U_a(l) + U_b(l)}{(U_a(l) - U_b(l))/Z_0} = Z_0 \frac{1 + U_b(l)/U_a(l)}{1 - U_b(l)/U_a(l)} = Z_0 \frac{1 + \Gamma(l)}{1 - \Gamma(l)}. \tag{2.83}$$

Since the reflection factor is given by the reflected-to-incident wave ratio

$$\Gamma(z) = \frac{U_b(z)}{U_a(z)}, \tag{2.84}$$

where

$$U_b(z) = U_{b0}\,\mathrm{e}^{\gamma z} = U_{b0}\,\mathrm{e}^{\gamma l}\,\mathrm{e}^{\gamma(z-l)} = U_b(l)\,\mathrm{e}^{\gamma(z-l)}, \tag{2.85a}$$

$$U_a(z) = U_{a0}\,\mathrm{e}^{-\gamma z} = U_{a0}\,\mathrm{e}^{-\gamma l}\,\mathrm{e}^{-\gamma(z-l)} = U_a(l)\,\mathrm{e}^{-\gamma(z-l)}, \tag{2.85b}$$

we can write

$$\Gamma(z) = \frac{U_{b0}}{U_{a0}}\,\mathrm{e}^{2\gamma z} = \Gamma(0)\,\mathrm{e}^{2\gamma z} = \Gamma(l)\,\mathrm{e}^{2\gamma(z-l)}. \tag{2.86}$$

But due to Eq. (2.83) we have at the end of the transmission line

$$\Gamma(l) = \frac{Z_2 - Z_0}{Z_2 + Z_0}. \tag{2.87}$$

The reflection factor for any other place can now be calculated using

$$\Gamma(z) = \frac{Z_2 - Z_0}{Z_2 + Z_0}\,\mathrm{e}^{2\gamma(z-l)}. \tag{2.88}$$

Note the special case at the beginning of the transmission line

$$\Gamma(0) = \frac{Z_2 - Z_0}{Z_2 + Z_0}\,\mathrm{e}^{-2\gamma l}. \tag{2.89}$$

Thus, at the beginning of the transmission line we see the impedance

$$\begin{aligned}
Z &= \frac{1+\Gamma(0)}{1-\Gamma(0)} Z_0 \\
&= \frac{1+\frac{Z_2-Z_0}{Z_2+Z_0}\mathrm{e}^{-2\gamma l}}{1-\frac{Z_2-Z_0}{Z_2+Z_0}\mathrm{e}^{-2\gamma l}} Z_0 \\
&= \frac{(Z_2+Z_0)\,\mathrm{e}^{\gamma l} + (Z_2-Z_0)\,\mathrm{e}^{-\gamma l}}{(Z_2+Z_0)\,\mathrm{e}^{\gamma l} - (Z_2-Z_0)\,\mathrm{e}^{-\gamma l}} Z_0 \\
&= \frac{2Z_2\cosh(\gamma l) + 2Z_0\sinh(\gamma l)}{2Z_2\sinh(\gamma l) + 2Z_0\cosh(\gamma l)} Z_0 \\
&= \frac{Z_2 + Z_0\tanh(\gamma l)}{Z_0 + Z_2\tanh(\gamma l)} Z_0.
\end{aligned} \tag{2.90}$$

For a lossless transmission line, where $\gamma = j\beta$, Eq. (2.90) can also be written as

$$\boxed{Z = \frac{Z_2 + jZ_0\tan(\beta l)}{Z_0 + jZ_2\tan(\beta l)} Z_0\,.} \tag{2.91}$$

Eq. (2.90) allows the transformation of an impedance Z_2 as seen through a transmission line of length l and wave impedance Z_0. The evaluation of Eq. (2.90) is usually only possible using a calculator. For lossless transmission lines, a much easier way to compute the impedance transformation of a transmission line can be found using the Smith chart. Eq. (2.86) indicates that the reflection coefficient is rotated around the origin with angle $2\beta z$ if a transmission line of length z and wave impedance Z_0 is used. Z_0 is also the system impedance of the Smith chart. If another wave impedance Z_w is used, then the transformation must be constructed using a Smith chart with that wave impedance Z_w as its system impedance and all impedances must be drawn on this Smith chart.

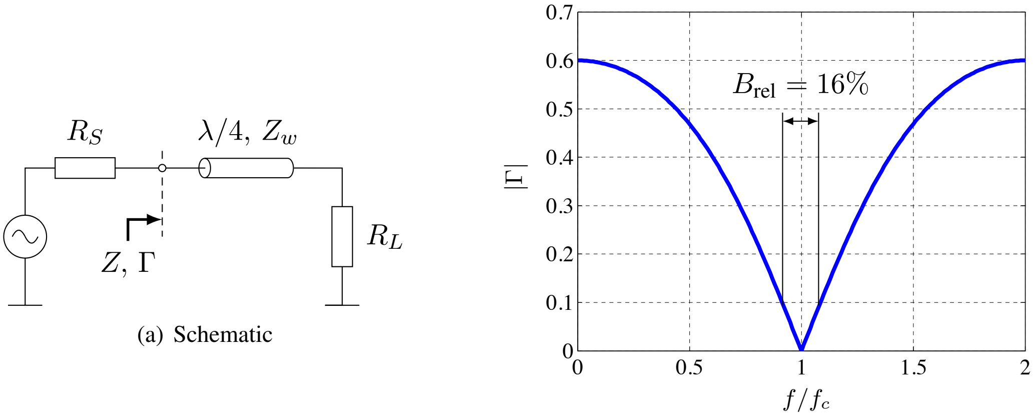
Now, many impedances might be matched using the circuit in Fig. 2.47. We simply have to find the crossing point of the perpendicular in the midpoint of the connection of the two impedances and the real line. This gives the center of a circle that contains both impedance points, thus defines the impedance of the transmission line. The angle from one point to the other determines the length of the transmission line.

Quarter-Wave Impedance transformers

Of special interest to RF design are quarter-wave impedance transformers, often called $\lambda/4$ transformers, see Fig. 2.48(b). By $\lambda/4$ transformers we mean transmission lines that are exactly one quarter of a wavelength for the frequency used. For this wavelength, the rotation of the reflection coefficient is exactly π or 180 °. This means that a real impedance will be transformed into another real impedance. As can be constructed easily from the Smith chart or seen by Eq. (2.91) the input impedance seen through the line is then simply

$$\boxed{Z = \frac{Z_0^2}{Z_2}\,.} \tag{2.92}$$

Of course, the transmission line is only one-quarter of a wavelength for one particular frequency and therefore Eq. (2.92) is only exact at one point. The matching process using a $\lambda/4$ transformer is therefore a narrowband approach, as Fig. 2.48 clearly indicates.

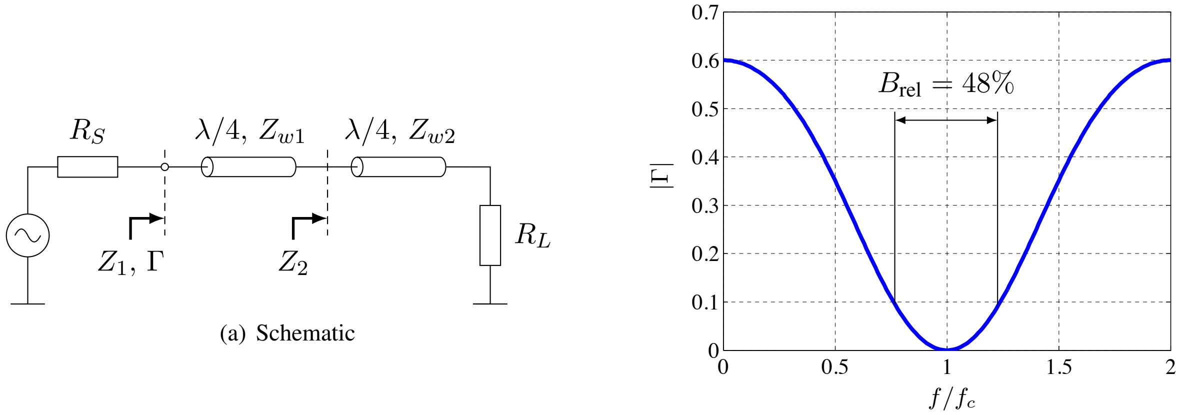


(a) Schematic

(b) Reflection as a function of normalized frequency. The transformer is designed for $R_L/R_S = 4$.

Figure 2.48 Single-stage $\lambda/4$ transformer.

Matching circuits for a wider bandwidth can be constructed by cascading $\lambda/4$ transformers of different wave impedances. Good results are thereby achieved by tapering the wave impedances of the transmission line pieces in such a way that the intermediate virtual impedance between the transmission lines build a geometric series. A circuit diagram and a corresponding frequency response can be seen in Fig. 2.49.



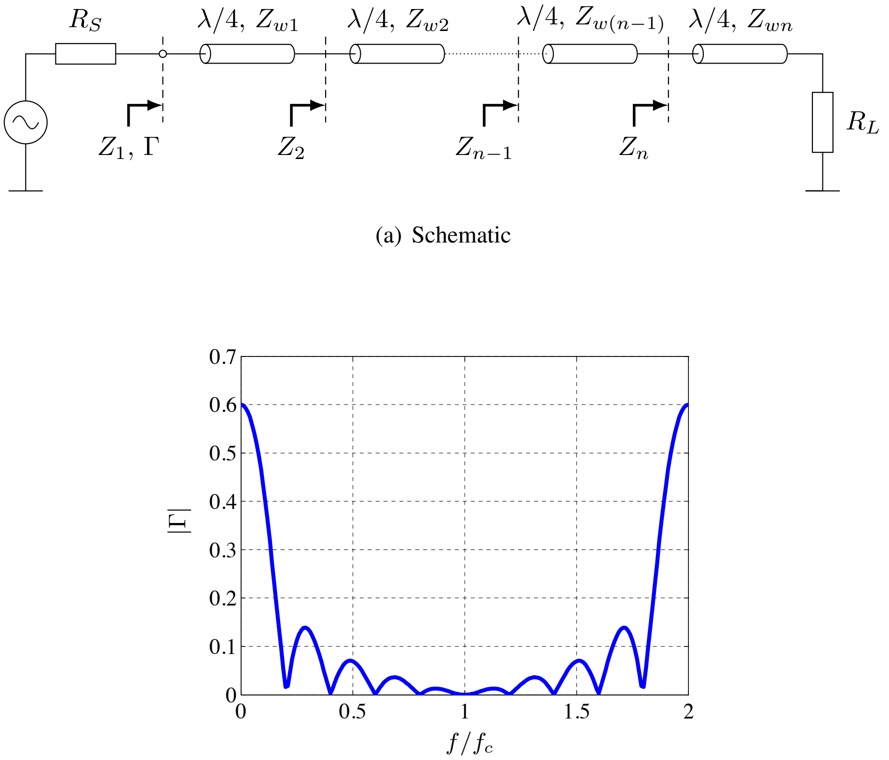
(a) Schematic

(b) Reflection as a function of normalized frequency. The transformer is designed for $R_{\text{L}}/R_{\text{S}} = 4$.

Figure 2.49 Double-stage $\lambda/4$ transformer.

A still wider bandwidth is obtained by cascading more stages. An example of ten such stages and the matching bandwidth obtained is given by Figs. 2.50(a) and 2.50(b), respectively.

It can be shown that the most wideband behavior is obtained if the transformed impedance levels achieved by a certain stage are geometrically increasing (or decreasing), i.e., the impedance after each line follows



(a) Schematic

(b) Reflection as a function of normalized frequency of a multistage-stage (in this case ten) $\lambda/4$ transformer. The transformer is designed for $R_\mathrm{L}/R_\mathrm{S} = 4$.

Figure 2.50 Multistage $\lambda/4$ transformer.

from the impedance of the previous one by multiplying it with a constant factor a. Thus, the factor has to be

$$a = \sqrt[n]{\frac{R_L}{R_S}}. \tag{2.93}$$

Hence, looking eastwards into the transmission line k in Fig. 2.50(a), we see the impedance level

$$Z_k = \left(\sqrt[n]{\frac{R_L}{R_S}}\right)^{k-n-1} \cdot R_L = \left(\frac{R_L}{R_S}\right)^{\frac{k-n-1}{n}} \cdot R_L. \tag{2.94}$$

From that, the wave impedances can easily be computed

$$Z_{wk} = \sqrt{Z_{k+1}Z_k} = \sqrt{\left(\frac{R_L}{R_S}\right)^{\frac{k-n}{n}} \left(\frac{R_L}{R_S}\right)^{\frac{k-n-1}{n}} R_L^2} = \left(\frac{R_L}{R_S}\right)^{\frac{2k-2n-1}{2n}} R_L. \tag{2.95}$$

General Impedance Behavior of Transmission Lines

If we look more closely at Eq. (2.92), we realize that a $\lambda/4$ transformers is an impedance inverter. By setting $Z_2 = 0$ and $Z_2 = \infty$, respectively, we get an infinite and a zero impedance at the input. For general lengths

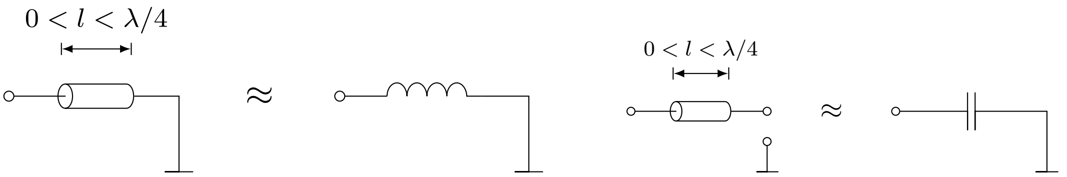


Figure 2.51 For $0 < l < \lambda/4$, the short-circuit line behaves inductive and the open line capacitive.

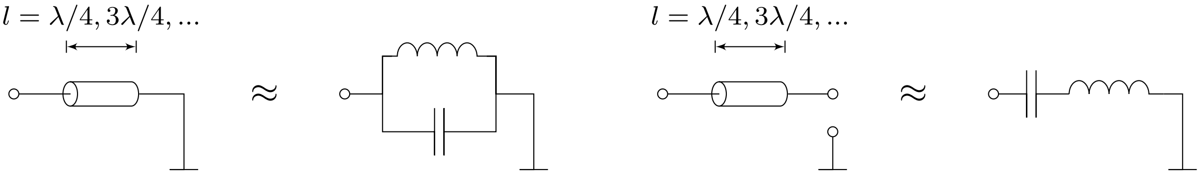

$l = \lambda/4, 3\lambda/4, ...$
≈
$l = \lambda/4, 3\lambda/4, ...$
≈

Figure 2.52 For $l = \lambda/4, 3\lambda/4, 5\lambda/4, \ldots$, the short-circuit line behaves like a parallel resonance and the open line like a series resonsance, and vice versa for $l = \lambda/2, \lambda, 3\lambda/2, \ldots$.

of transmission lines, we can derive the following results, which are displayed in Figs. 2.51 and 2.52 and can be easily seen from the application of Eq. (2.90).

In other words, line stubs that are shorter than $\lambda/4$ behave capacitive if their ends are open and inductive if their ends are short-circuited. The inverse is true for line stubs longer than $\lambda/4$ but shorter than $\lambda/2$. In this case, the line stubs behave inductive if their ends are open and capacitive if their ends are short-circuited.

Other Transmission Line Transformers

Apart from single-stage and multistage $\lambda/4$ transformers, we can build more complicated constructions to handle the matching process, not unlike the case with discrete components. Fig. 2.53 shows a compensated $\lambda/4$ transformer to match a load resistance to a source resistance.

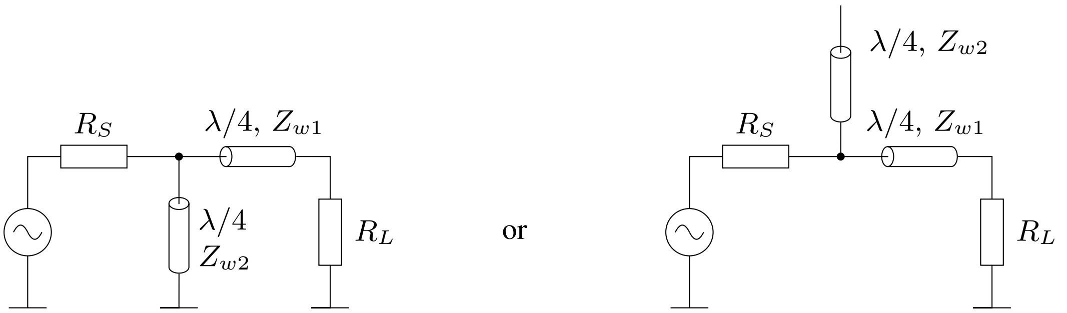
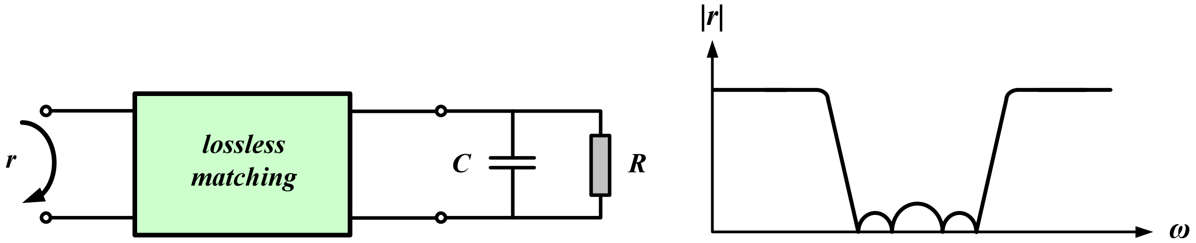


Figure 2.53 Compensated $\lambda/4$ transformer.

Theoretical Limits

Building more and more wideband matching circuits raises the question as to how wideband we can go. When an RC load is matched using a lossless matching circuit according to Fig. 2.54(a), the resulting reflection coefficient $\Gamma(\omega)$, shown in Fig. 2.54(b), is limited by the *Bode-Fano inequality* [8, 21]:

$$\int_0^\infty \ln \frac{1}{|\Gamma(\omega)|}\, d\omega \leq \frac{\pi}{RC}. \tag{2.96}$$



(a) Matching using a lossless matching circuit

(b) Reflection coefficient as a function of frequency

Figure 2.54 Theoretical limit on the reflection coefficient.

2.5 RF Measurements

RF measurement devices are special in comparison with equipment used for audio frequency. In the following, some of the most important devices will be briefly explained. As usual with electrical measurements, we want to either measure signals or systems. For signals, we are interested in the frequency, the power, or the spectral shape. With modulated signals, more parameters such as deviation from ideal value (error vector magnitude), phase noise etc. are of interest. With systems, we are interested in the frequency-dependent amplitude and phase response, but also in the noise figure and region of linearity (e.g., for amplifier).

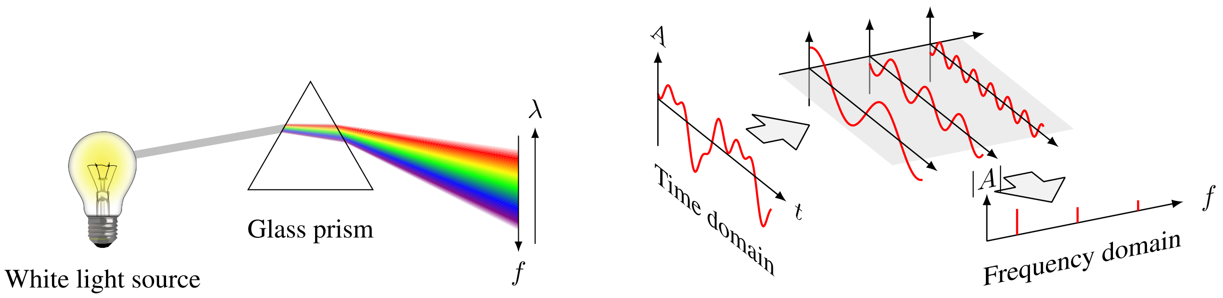
The devices most often used in RF design are the spectrum analyzer and the network analyzer. While the spectrum analyzer investigates an unknown signal with respect to its spectral contents, the network analyzer measures a system or a network, i.e., the corresponding frequency response.

Some special devices or components are needed for some of the measurement equipment described in the following. They shall be described briefly in the next sections.

2.5.1 Spectrum Analysis

As Fig. 2.55 illustrates, spectral analysis can be regarded as the function of a prism: a signal is split into its spectral components. For RF signals however, such a prism does not exist (or rather: it would have to be very large, i.e., much larger than the wavelengths under consideration), but there are two other principles of spectrum analysis for RF, microwave and millimeter-waves in practice:

- **FFT:** sampling a certain duration of the signal and applying the (fast) Fourier transform, where the total signal power is split up in the respective bins, or



(a) Optical spectrum analysis using a prism

(b) RF/microwave frequency analysis, e.g. by FFT

Figure 2.55 Spectrum analysis principles.

- **Filter method:** the signal is fed through a bandpass filter with a (tunable) passband at the frequency under consideration, and the signal energy of the frequency (band) under consideration is directly received as a time signal.

FFT-Based Spectrum Analyzer

The FFT analyzer first samples the input signal and stores it in its memory. In the next step, an FFT is applied to the memory contents. The Fourier transform thus transfers a temporally arranged signal into the frequency domain. This is usually done by means of a microprocessor or a digital signal processor (DSP).

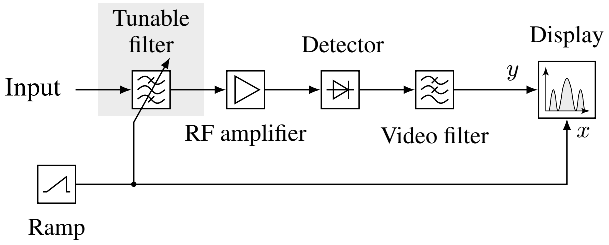
The main disadvantage of FFT-based spectrum analysis is that, with increasing frequency range and signal dynamic, the requirements (e.g. memory storage, ADC dynamic and bandwidth, etc.) grow unduly. Thus, at present, a wide dynamic range of about 100 dB or more can only be achieved up to a few 100 kHz with the FFT-based method. Higher bandwidths inevitably lead to a smaller dynamic range.

However, in contrast to other analyzer methods, the phase information is not lost during the (complex) FFT. Therefore, FFT analyzers are able to measure both the magnitude and the phase of the complex spectrum. This is especially interesting for measurements of filter networks and signal processing applications. It is generally expected that the FFT-based spectrum analyzers will become more interesting for RF measurement applications in the future, when even faster ADC with high dymanic ranges become available.

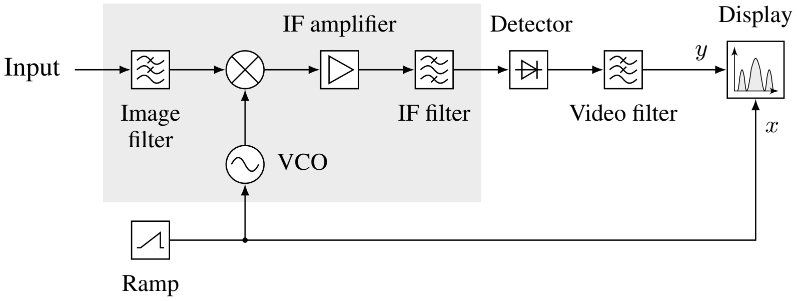
Swept-Tuned Spectrum Analyzer

For RF ranges, swept-tuned analyzers are used. The schematic of such a spectrum analyzer is shown in Fig. 2.57. The underlying principle is different from the FFT-based spectrum analyzer: the spectra is not computed form the signal’s time characteristic, but determined directly by analysis in the frequency domain.

As brought up before, the signal could be split up into frequency bands by a tunable filter (see Fig. 2.56(a)), after which the signal power in that band can be measured by an envelope detector and be displayed at the respective point on the display. In practice however, building such a filter would be too elaborate a task and the filter would become highly complex and expensive.



(a) Tunable filter method (underlying principle)



(b) The heterodyne principle (actual, practical method)

Figure 2.56 Swept filter vs. heterodyne principle.

Fortunately, there is an equivalent principle that can be realized much more easily in practice: instead of tuning a filter and sweep it over the signal, the signal is mixed down into a lower band (called intermediate frequency, IF) and swept past a fixed-tuned filter, as shown in Fig. 2.56, yielding the same result.

The underlying principle is called *heterodyne* principle (we will meet this architecture in radio receivers in Section 6.5.3). This principle has some drawbacks, which have to be taken care of: an image filter is necessary before the mixer to avoid mixing unwanted signals into the IF and also some negative effects due to nonlinearities and finite isolation of the ports of the mixer onto one another have to be considered. Instead of tunig the filter, in the heterodyne case the VCO (driving the mixer) is tuned.

In the end, the IF filter bandwidth and the sweeping velocity (ramp steepness in Fig. 2.56) as well as the video filter bandwidth are parameters which have to be set according to the measurement conditions. The consequences and implications of these parameters as well as some basic rules for their correct setting will be discussed in detail later.

A closer look at an actual spectrum analyzer from 9 kHz to 30 GHz

Fig. 2.57 shows an actual realization of a spectrum analyzer[10] for a total frequency range from 9 kHz to 30 GHz. As shown, the main parts can generally be split up into five sections:

- the two front ends, one for frequencies below 3 GHz and one for those above,
- the IF signal processing,
- the video signal processing (which however overlap somewhat in this case, as some IF processing takes place on the digital side), and finally
- the detection and display part.

[10] R&S FSP Spectrum Analyzer, 9 kHz ... 30 GHz

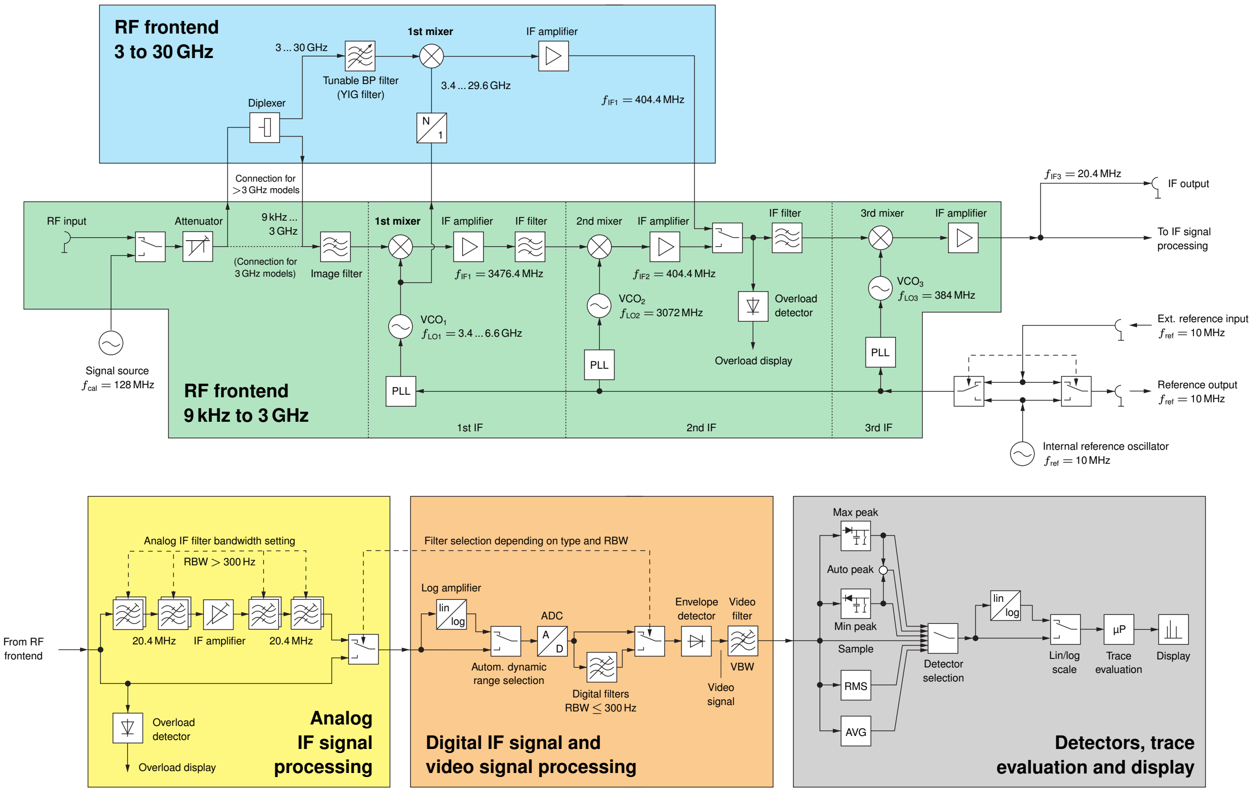


Figure 2.57 Full spectrum analyzer block diagram (adopted from [57], with various alterations).

To simplify the problems of image frequencies (see Section 5.3) the signal is first mixed up to a high IF (3476.4 MHz) by *high-side injection*[11] (HSI).

$$f_{\mathrm{IF}_1} = |f_{\mathrm{LO}} \pm f_{\mathrm{RF}}| \quad \xrightarrow[\text{HSI}]{f_{\mathrm{LO}} > f_{\mathrm{RF}}} \quad \begin{aligned} f_{\mathrm{IF}_1} &= f_{\mathrm{LO}} - f_{\mathrm{RF}} && (2.97) \\ f_{\mathrm{image}} &= f_{\mathrm{LO}} + f_{\mathrm{RF}} > 2 f_{\mathrm{RF}} && (2.98) \end{aligned}$$

That way, the image frequencies can easily be filtered out by a low pass image filter (of relatively low order) right after the attenuator. Because then mixing down to the final low IF directly would create an image frequency problem again, a further IF step (at $f_{\mathrm{IF}_2} = 404.4$ MHz) is used, before the signal is mixed down to the final IF of $f_{\mathrm{IF}_3} = 20.4$ MHz and passed on to IF signal and video signal processing parts.

For frequencies of 3 GHz and above, another frontend is used, where the image problem is solved in a simpler way by a tunable YIG[12] filter (YTF), which can easily be realized for that frequency range and large tunable bandwidths. Thus, in this case, only two IF steps (in total) take place. As shown, in the high frequency front-end, the LO frequency is generated by multiplication (generally with a full multiplicative synthesizer with a PLL). Another way is the so-called *harmonic mixing*, where the nonlinear behaviour of the mixer is used and a second, third or even higher order harmonic of the (strong) LO is used to mix the signal down to the IF. This principle is especially common in millimeter-wave applications, e.g., from 50 GHz to 150 GHz and beyond.

[11] meaning the LO frequency is higher than the RF, thus it comes from the *higher* side

[12] Yittrium Iron Garnet, see Section 5.4.4

The switch from the low frequency to the high frequency front-end can easily be seen from measuring the noise floor of the unconnected spectrum analyzer as shown in Fig. 2.58: somewhere around 3 GHz the noise floor increases by a large step of about 7 dB. This happenes again at 13.2 GHz, where the frequency multiplier changes from factor 2 to 4 and the mixing process changes from low-side to high-side injection. While the step at 3 GHz allways occurs for this device (for all devices with two front-ends at their respective limit frequency), the second step depends on the actual frequency span setting; the spectrum analyzer chooses the frequency multiplication factors and injection sides of the mixing process in order to minimize these effects.

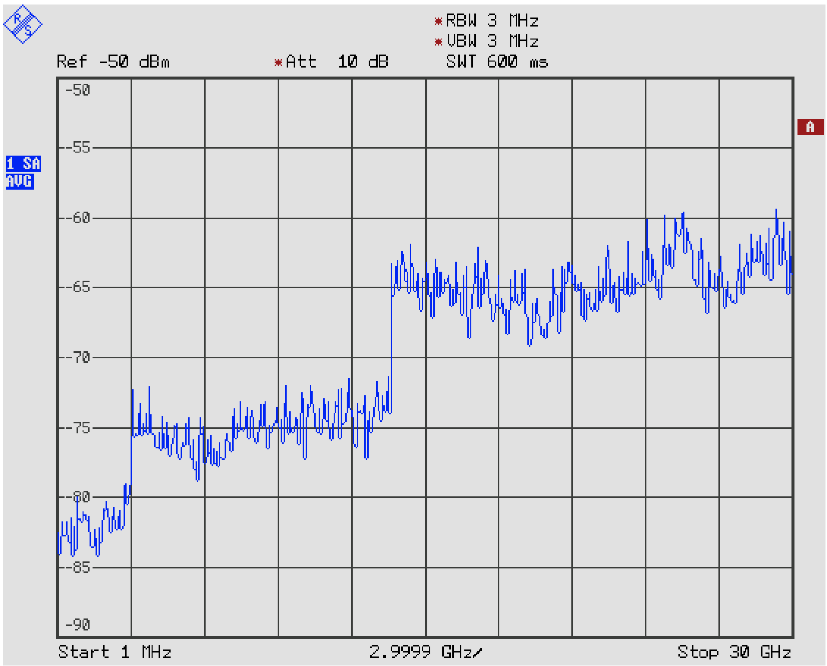


Figure 2.58 Noise floor of the R&S FSP spectrum analyzer (1 MHz ... 30 GHz). Note the rapid jumps at 3 GHz (change of the front-ends) and at 13.2 GHz (change of the frequency multiplier and injection side).

After the front-ends, in the IF signal and video signal processing parts, the actual spectrum analysis takes place: First, the signal bandwidth is limited to the chosen *resolution bandwidth* (RBW) by four filter stages, after which the signal can be amplified by a logarithmic amplifier, to get a higher dynamic range through the ADC, which comes next. After that, the envelope is built digitally and the video signal processing (LP-filtering with the *video bandwidth* VBW) takes place. Also, very small bandwidths (e.g. < 300 Hz) can not be achieved with four analog filters. Thus, in those cases and also in cases where other filter functions (e.g. FFT-based filters) are used, the IF filtering takes place in the digital domain.

The last part, consisting of the detection, evaluation and display functions, is highly dependent on the features available of the actual device at hand. More generously equipped spectrum analyzers offer a full variety of detectors and trace evaluation functions, whereas more economically priced devices have a reduced set of functionalities. Also, in this part the trace is scaled according to all the parameters chosen in the stages before and displayed on the screen.

Parameter Setting

The swept-tuned analyzer method contains several parameters, which have to be set correctly to be able to measure accurately and get a good reading of the measurement values (and also to avoid damaging the device). These parameters include namely:

- **Reference amplitude and attenuation**
  Nowadays spectrum analyzers allow measurements in a very wide signal level range, limited by the inherent noise floor and the maximum permissible input level, e.g. from −150 dBm to 30 dBm, thus a dynamic range of 180 dB and over a large frequency span, from (almost) DC to several dozens of Gigahertz. However, it is not possible to reach the full dynamic range over the full frequency span at once, since different settings are required (such as bias, attenuation, etc.) at different power and frequency ranges, because the dynamic ranges of the components used (e.g. log amplifiers, envelope detectors and ADCs) are much smaller than the full operational range of the spectrum analyzer.

  The main problem (amplitude-wise) is the first mixer. As shown in Section 5.3, a mixer is a semiconductor device, used in the spectrum analyzer to mix the RF signal down to an intermediate frequency (IF). Mixers are (and have to be) nonlinear devices and as such always produce both harmonic and disharmonic[13] distortion besides the wanted signal (which can also be favourable sometimes, for *harmonic mixing*, as previously mentioned). The first mixer in a spectrum analyzer is built in order to have an almost linear RF to IF transfer function and to minimize spurious frequencies over a certain power range and over a wide RF bandwidth—evidently, that becomes harder to achieve as the the frequency and power ranges widen. Thus, the first mixer is both one of the most crucial and one of the most expensive part within a spectrum analyzer. Unfortunately, it is also one of the most sensitive when it comes to the measurement parameter settings.

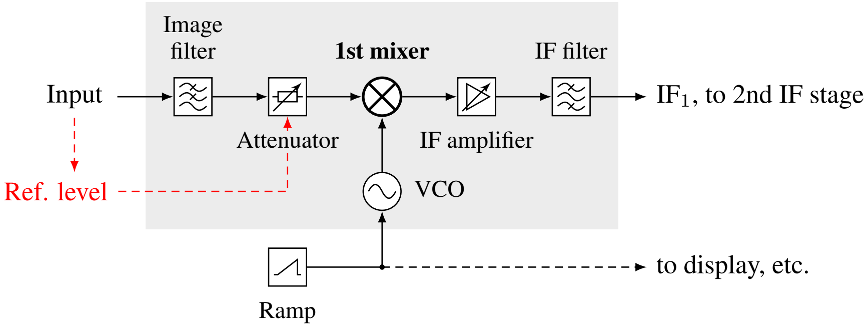


Figure 2.59 The first mixer of a spectrum analyzer.

  To ensure that the first mixer works in the quasi-linear range, the input level should be distinctly below the 1 dB compression point (see Section 5.1.5). To avoid overdriving, the input signal is first attenuated to appropriate levels by the broadband RF attenuator. The attenuation actually required depends on the dynamic range of the mixer (and the subsequent stages) as well as on the specific reference level (i.e. the applied signal power), because evidently

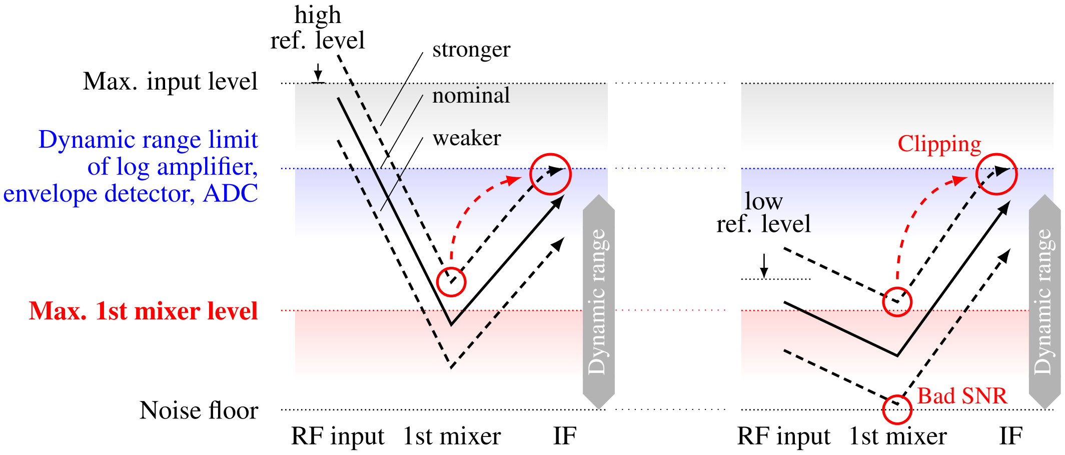
  - a strong signal has to be attenuated more
  - an already pretty weak signal has to be attenuated less (or not at all).

[13] For example when two strong signals are present, besides their harmonics (integer multiplicatives of them) also a signal at their difference frequency $f_{\text{diff}} = |f_1 - f_2|$ and at their sum frequency $f_{\text{add}} = f_1 + f_2$ are produced. The same goes for signals with continuous frequency distributions.

Therefore, it makes sense to couple the attenuation factor setting to the reference level (implying the maximum signal level), which is common for spectrum analyzers of newer generations.

If they have to be set manually, either the maximum signal power is already known to some degree or it should be started with a high attenuation setting and decrease it while watching out for possible overload indication. It should be noted that:

- **Too little attenuation**: if the attenuation is too low or the signal is too strong (too low reference level for coupled settings), the signal amplitude can be clipped (see Fig. 2.60(a) and Fig. 2.61), leading to spurious frequencies. At even higher signal levels and/or low attenuation settings, the mixer can be damaged[14] permanently.
- **Too high attenuation**: The signal level at the mixer input is too low, resulting in an unnecessarily low SNR (which can be seen by a higher noise floor level and unclear reading of the wanted signal), as indicated in Fig. 2.60(b).



(a) Strong signals (b) Weak signals

Figure 2.60 Reference level, mixer power range and IF amplification.

Note 1: Since the IF filter (which selects the signal part you actually want to see on the display) is placed *after* the mixer, seeing only low signal levels on the screen does not necessarly mean that the spectrum under consideration does not contain higher signal levels somewhere else in the spectrum. Thus, even the signal parts that cannot be seen or are of no interest for the measurement can lead to overdriving the mixer. Most of the signal analyzers have an overload detection (see Fig. 2.57) and show that on the display. Especially crucial is DC, which cannot be seen by the spectrum analyzers but some of them do not allow any DC to be present in the signal!

Note 2: The spectrum analyzer might be able to detect the mixer overdrive (the indication flag is commonly called *IF overload*), but generally it cannot protect itself from any damage. The only way to protect the mixer from too powerful signals is to attenuate it appropriately!

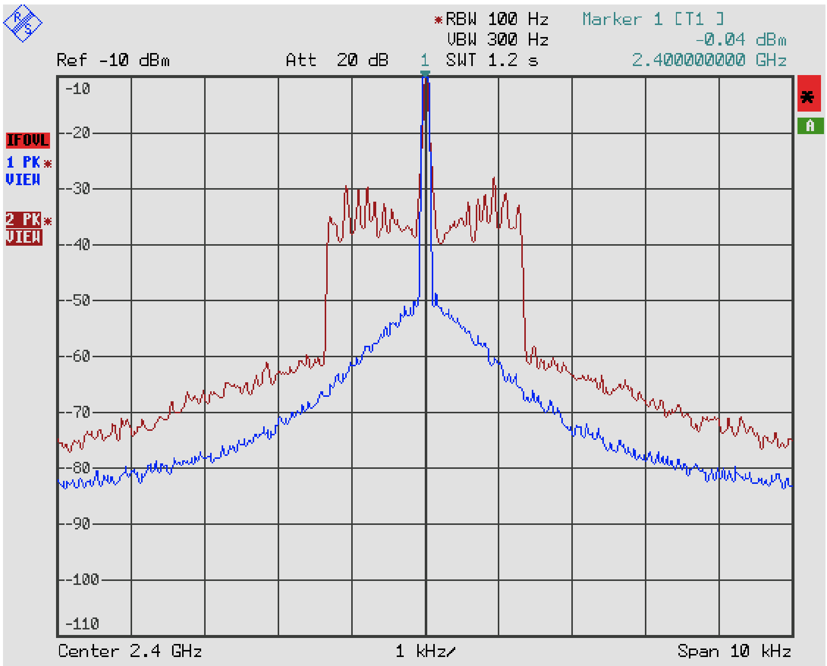
Attenuation increases the noise floor (one-on-one), which is of course a drawback for high attenuation settings. With coupled settings this problem is already optimized, so as to provide the necessary attenuation when strong signals are expected (high reference level setting) and to minimize the attenuation

[14] damaged could also mean, that the mixer looks as if it is working fine, but the level of the produced nonlinearities is higher than it used to be

loss and increased noise floor for low signal levels. Thus, the general rule of thumb for setting the reference amplitude is:

The reference amplitude should be set as low as possible – but as high as necessary.

If the actual signal level is not known, it should be started with a high setting and a wide frequency span, after which the level can slowly be decreased while watching out for the overload indication. Commonly, the attenuator levels can be set in steps of 10 dB, so one should be aware of those immediate level changes. Again, more expensive devices offer 5 dB or even 1 dB attenuator steps. It is also common that the minimum selectable attenuator level is 10 dB (not 0 dB, as might be expected), to always provide enough impedance matching at the mixer's RF input.



```
*RBW 100 Hz     Marker 1 [T1 ]
 VBW 300 Hz           -0.04 dBm
Ref -10 dBm          Att  20 dB    1  SWT 1.2 s     2.400000000 GHz
Center 2.4 GHz                1 kHz/                Span 10 kHz
```

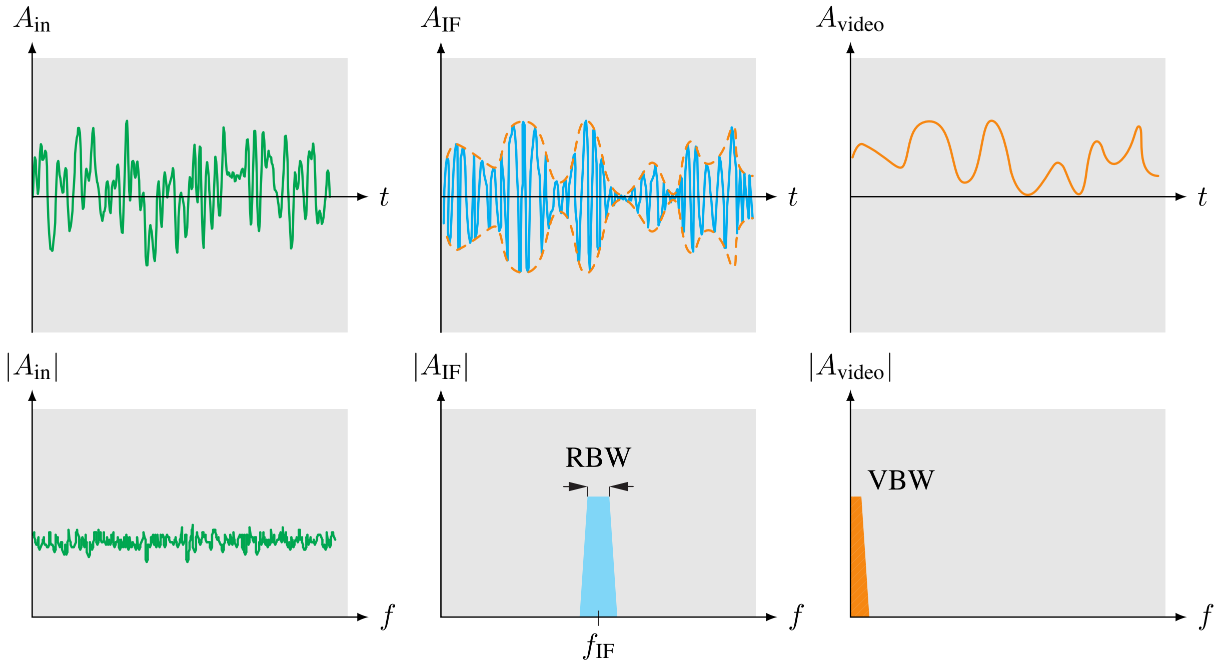
too little attenuation → IF overload

correct attenuation

Figure 2.61 Visible distortion due to IF overload (red) compared to correct attenuation setting (blue).

- **Resolution bandwidth (RBW) and detectors**
  The second crucial parameter is the frequency resolution. Because for swept-tuned spectrum analyzers this is actually a filter bandwidth setting, it is defined by the so-called resolution bandwidth, RBW.

  The RBW defines the frequency range of the signal of which the power will be measured. Fig. 2.62 shows the overall principle: the signal (green) is the input signal bandwidth translated onto the IF frequency, where the PB filter with bandwidth RBW ($= B_{\text{IF}}$) limits its bandwidth onto the bandwidth over which the power is to be measured and indicated at the respective bin on the display. Of this signal the envelope is built, which is then again filtered by the video filter with the bandwidth VBW ($= B_{\text{video}}$). The video signal is the time domain signal of the signal power of the respective frequency and bandwidth RBW (where its fluctuations are filtered by the video filter, for better reading).



Input signal

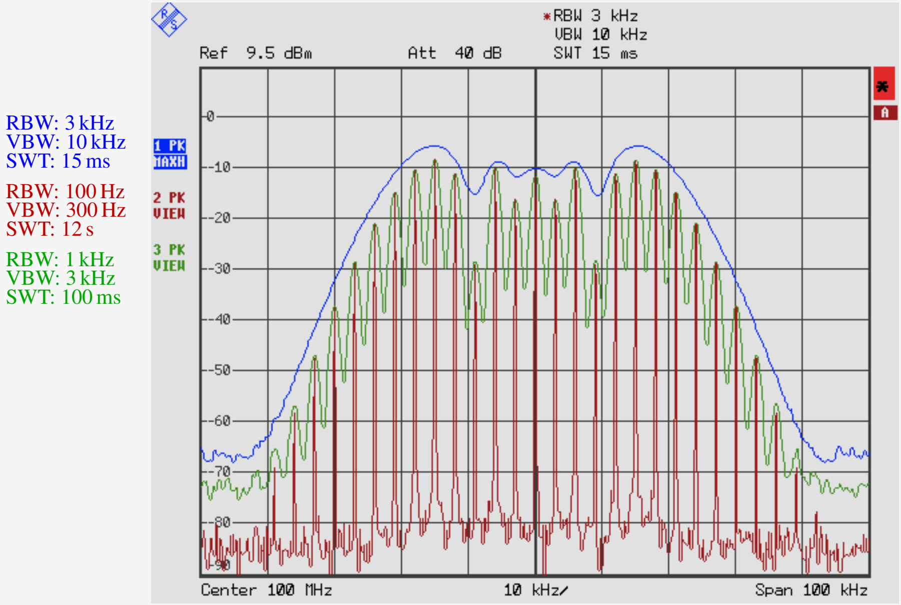
Filtering (RBW), envelope detection

Video signal, video filter (VBW)

Figure 2.62 RBW, Envelope, VBW.

Example: FM-Spectrum

Here is an example showing the implication of the resolution bandwidth setting: shown is an FM spectrum (100 MHz carrier (0 dBm), modulated by 3 kHz (1 V_{pp}) and 20 kHz deviation), measured with three different RBWs (and according VBWs and SWTs, which will be explained later).


RBW: 3 kHz
VBW: 10 kHz
SWT: 15 ms

RBW: 100 Hz
VBW: 300 Hz
SWT: 12 s

RBW: 1 kHz
VBW: 3 kHz
SWT: 100 ms

*RBW 3 kHz
VBW 10 kHz
SWT 15 ms

Ref 9.5 dBm
Att 40 dB

Center 100 MHz
10 kHz/
Span 100 kHz

- The blue line is the result when a large IF bandwidth is used: the peaks that can be seen with a narrow bandwidth (red or green lines) are smudged over each other, meaning that at each point various peaks fell into the band over which the power was measured at each point. Thus, the result is a smooth envelope over the whole spectrum (somewhat too high in amplitude, because each of the peaks is counted several times).

- The green line is something inbetween the other two others: the filter is already narrow enough so that the peaks can easily be distinguished, because the filter function has already dropped enough from one peak to another.
- And finally the fine enough RBW setting (RBW smaller than the distances between the peaks) of the red line shows each single peak in the spectrum, with the amplitudes given according to the respective Bessel equation, as known from communications theory.

- **Video bandwidth (VBW)**
  The video bandwidth solely serves the purpose of making the picture more readable from the display, e.g. by reducing fast fluctuations from short sweep times or by reducing noise in other signal traces. Note that these reductions could also lead to false results. Since the video filtering takes place before the detection[15], both settings also have implications onto one another.

  As a rule of thumb, the video bandwidth should always be larger than the resolution bandwidth, in order not to falsify the result by reducing the displayed signal power—one exception being noise measurements. Again, some spectrum analyzers couple the VBW to the RBW setting, where commonly three coupling ratios are available:

  - for continuous wave (CW) measurements (e.g. sine wave): RBW : VBW = 1:3
  - for pulse measurements: RBW : VBW $\leq$ 1:10
  - for noise measurements: RBW : VBW = up to 10:1

  Another possibility to achieve better reading is to use the trace average function, which just averages the actually displayed trace, but not the whole video signal, as the video filter does.

- **Detectors:**
  Because modern spectrum analyzers use a pixel-based LCD instead of the cathod ray tubes to display the spectrum, both the level and the frequency display is limited. While the finite amplitude resolution can be remedied by using a marker, the finite frequency display resolution is solved by detectors[16] and their functions.

  Depending on the actual frequency span setting, one pixel might contain the information of a relatively large subrange. The actually measured number of points and therefore also the actual frequency step the spectrum analyzer uses, depends on the tuning steps of the 1st local oscillator. Then, the detector chooses what to do with the samples which all will be represented by the single respective pixel.

[15] Note that this detection is not the envelope detection, which appears prior to the video filter, but the detection of the video signal for pixel-based display.

[16] again, not to be confused with the envelope detector

As for the filter types and bandwidths, also the available detectors vary between spectrum analyzers. Depicted in Fig. 2.57 is a set of the most common detectors:

- **sample:** just taking the first of all the samples associated with the respective bin,
- **peak:** taking the minimum or maximum peak (min-peak, max-peak) value of the samples, as well as *auto peak*, which draws a line connecting the two, and
- **RMS and average (AVG):** which calculate the RMS or average of the samples.

However, less professional devices might not have all of these detectors available, the most common are sample and peak detectors, since they do not involve any computations taking place. The whole process from video filtering to the various types of detections is depicted in Fig. 2.63.

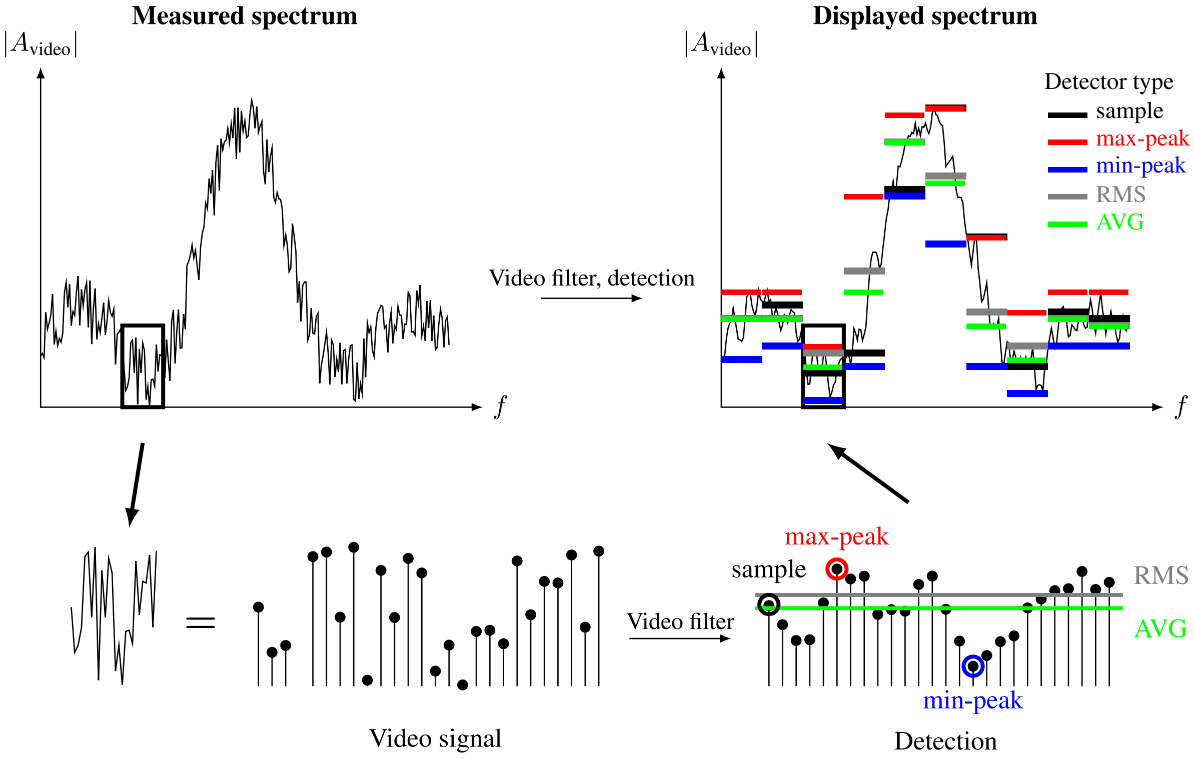


Figure 2.63 Implications of the video bandwidth (VBW) and detection types.

Choosing the right detector type actually depends on the resolution bandwidth and frequency span (and the sweep time) settings as well as the type of measurement actually taking place. Fig. 2.64 shows the implications of the sample and the max-peak detector: since the frequency span is quite large and the RBW is very small, only the max-peak detector ensures that the (very narrow) carrier can be seen on the display. The sample detector however might miss it because one of the samples not showing the carrier is assigned to the pixel. On the other hand, the noise floor of the sample detector is of course much lower than for the max-peak detector.

The RMS detector on the other hand always calculates the true power of each sample sequence assigned to the pixel, including noise. This is favourable for noise and broadband measurements, but might result in wrong values when measuring carrier power in a noisy environment, where the average detector might be more feasible.

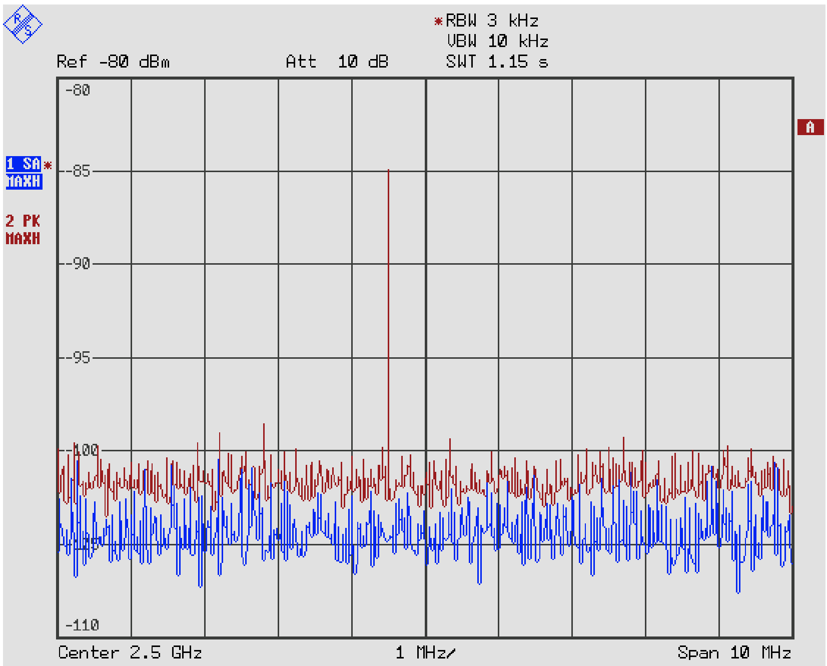


Figure 2.64 An example illustrating the right selection of detectors: the sample detector misses the narrow spurious peak, whereas the peak detector picks it up but also shows an increased noise floor.

Implications on noise measurements

The amplitude of the envelope signal (video signal) is Rayleigh distributed (since the envelope voltage $v_{\text{env}} = \sqrt{v_I^2 + v_Q^2} = \sqrt{\text{Re}^2\, v + \text{j}^2 v}$ is expected to be independently Gaussian distributed in both I and Q or real and imaginary parts and the sum results in a Rayleigh distribution)

$$p_{\text{env}}(v) = \frac{v}{\sigma^2}\, \text{e}^{-v^2/(2\sigma^2)} \; .$$

To measure the noise power with the spectrum analyzer, the most straightforward method is to measure the time-average of the RMS value.

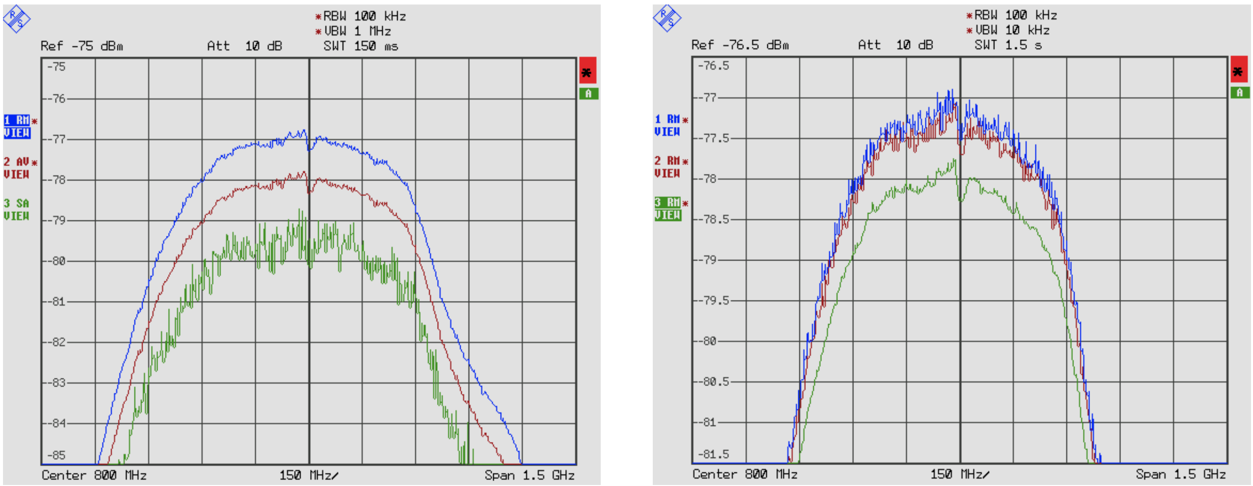
It might also seem feasible to use the square of the average envelope amplitude – but the square of the average is not equal to the average of the square!

$$v_{\text{rms}} = \sqrt{\int_0^\infty v^2 \cdot p_{\text{env}}(v)\, dv} = \sqrt{2}\,\sigma \qquad \leftrightarrow \qquad v_{\text{avg}} = \int_0^\infty v \cdot p_{\text{env}}(v)\, dv = \sqrt{\frac{\pi}{2}}\,\sigma$$

$$20\,\log_{10} \frac{v_{\text{rms}}}{v_{\text{avg}}} = 10\,\log_{10} \frac{2\,\sigma^2}{\pi/2\,\sigma^2} = 10\,\log_{10} \frac{4}{\pi} = 1.05\,\text{dB}$$

This has to be remembered when choosing the right detector type: the RMS detector always calculates the true power inside that window, i.e., including the power of the noise in that signal. But the sample detector displays the value of a single sample, recorded at a defined point in the time window ascribed to the bin (pixel on the display) in question – regardless of all the other points in the same window. Thus, with some averaging it is similar to the average signal resulting from using the average detector. Therefore results differ by 1.05 dB, whereas the results from the average and sample detector come out too low.

Further, if the result is averaged over a narrow video bandwidth (VBW < RBW) using the logarithmic level scale, the displayed average value is an additional 1.45 dB lower. The displayed noise is then 2.5 dB below the RMS value.



(a) Using the average detector results in 1.05 dB too low results, compared to the true RMS detector. Even more severe is averaging (of the sample detector output) in the log scale, then 2.5 dB too low results occur (whereas averaging the sample detector in the lin scale results in the RMS level)

(b) Also, the VBW can reduce the noise power: as long as VBW≫RBW the VBW has no impact (blue), at VBW=RBW the power begins to diminish (red), and for VBW<RBW (green) up to 2.5 dB too low results are received.

Figure 2.65 Implications on the measured noise power by VBW setting and averaging

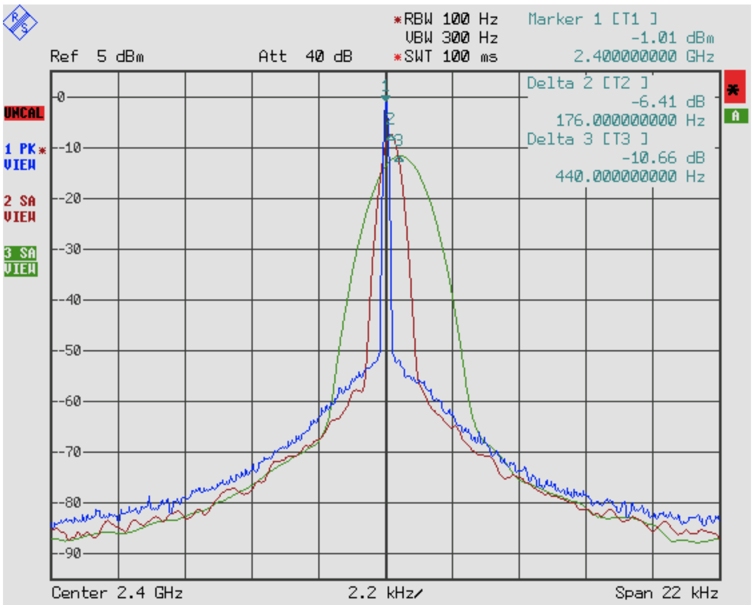
- **Sweep time (SWT)**
  The maximum permissible sweep speed is limited by the transient time of the IF filter and the video filter. If the video bandwidth is larger than the resolution bandwidth, its transient time has no effect and the required transient time increases inversely with the square of the resolution bandwidth. So with a decrease of the resolution bandwidth by the factor n the required minimum sweep time becomes n^2 longer,

$$\min T_{\text{sweep}} \propto \frac{\Delta f}{B_{\text{IF}}^2} \; . \tag{2.99}$$

  The involved proportionality factor depends on the filter type and the permissible transient response error. Analog filters generally require more sweep time, whereas with Gaussian filters a proportionality factor of 1 can be attained. It is generally a good idea to leave the sweep time on autoset, i.e., coupled to the RBW and VBW settings. There are just very few exceptions, where other settings make sense. If the SWT can not be coupled to the filter bandwidths and has to be set manually, again a longer sweep time should be chosen which can then be reduced, speeding up the measurement. Too fast a sweep (therefore too short an SWT) results in stretching the frequency axis, thus showing wrong frequencies on the display, as shown in Fig. 2.66.

Further Literature

The information above was just a very fast introduction into working with the spectrum analyzer and understanding its working principle. However, spectrum analysis can almost be considered an art, especially for more complex signals (e.g., wide-band spread signals such as UMTS or OFDM etc.). There is a variety of literature available to make the reader more familiar with the more elaborate measurement methods and parameter settings, such as the *Fundamentals of Spectrum Analysis* [57] by Christoph Rauscher of R&S or several application notes by all the other measurement device companies, such as Agilent, Anritsu, Rohde&Schwarz and many more.

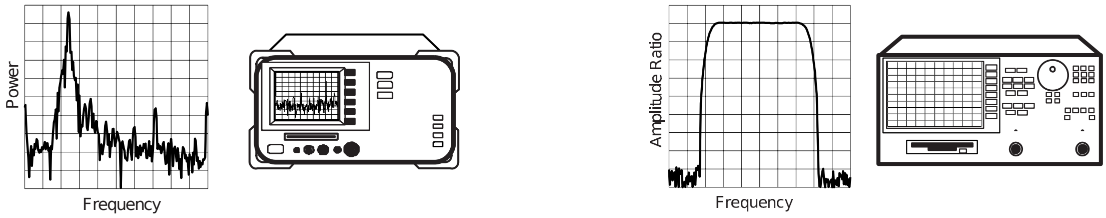


```
*RBW 100 Hz      Marker 1 [T1 ]
 VBW 300 Hz          -1.01 dBm
*SWT 100 ms    2.400000000 GHz
Ref  5 dBm      Att  40 dB

Delta 2 [T2 ]
        -6.41 dB
  176.000000000 Hz
Delta 3 [T3 ]
       -10.66 dB
  440.000000000 Hz

Center 2.4 GHz     2.2 kHz/     Span 22 kHz
```

Figure 2.66 SWT implications.

2.5.2 Network Analysis

Besides spectrum analyzers, network analyzers are the most important measuring devices in general RF and mobile communications technology. They differ both in form (in contrast to the spectrum analyzer, which has one input only, the network analyzer also provides an output, or generally several input/outputs) and function: while the spectrum analyzer is used to measure unknown external signals, the network analyzer uses known signals to measure the unknown properties of a device under test (DUT).



Spectrum Analyzers

- measure signal amplitude characteristics (carrier level, sidebands, harmonics, etc.)
- can demodulate (and measure) complex signal spectrum (real-valued result)
- are receivers only (single channel)
- can be used for scalar component tests with a tracking generator or ext. source
- show low errors, but do not offer error correction

Network Analyzers

- measure properties of components, devices, circuits and sub-assemblies
- display ratioed amplitude and phase (complex result) of frequency or power sweeps
- contains source(s) and receiver(s)
- offer advanced error correction

Figure 2.67 Differences between spectrum and network analyzers.

There are two versions of network analyzers:

- **scalar network analyzer:** consisting of spectrum analyzer and a tracking generator, and
- **vector network analyzer (VNA):** consisting of a spectrum analyzer, an internal tracking generator and a directional signal separation device (directional coupler).

While the difference might seem minor, the VNA can do everything a scalar network analyzer can do, but offers a huge range of additional possibilities. Instead of just detecting the power of the signal as with the spectrum analyzer or scalar network analyzer, the vector network analyzer compares the signals to the reference signal of the tracking generator. Thus, it can not only measure the power (amplitude) or frequency of the signals but also their phases, relative to one another.

The VNA is so much more common in practice, that it is colloquially called just network analyzer[17] (without actually specifying the vector part).

Working Principle

While the wavelengths are different for RF and microwave signals, the overall principle of the network analyzer is the same as testing a lens for its optical properties, as depicted in Fig. 2.68. Some incident light striking the lens is reflected, but an other part of it (most of it) continues through the lens. The same fundamental concept is valid for a network analyzer, which accurately measures the incident, reflected and transmitted energy and compares them to one another, to be able to reveal the properties of the lens (or more generally the DUT).

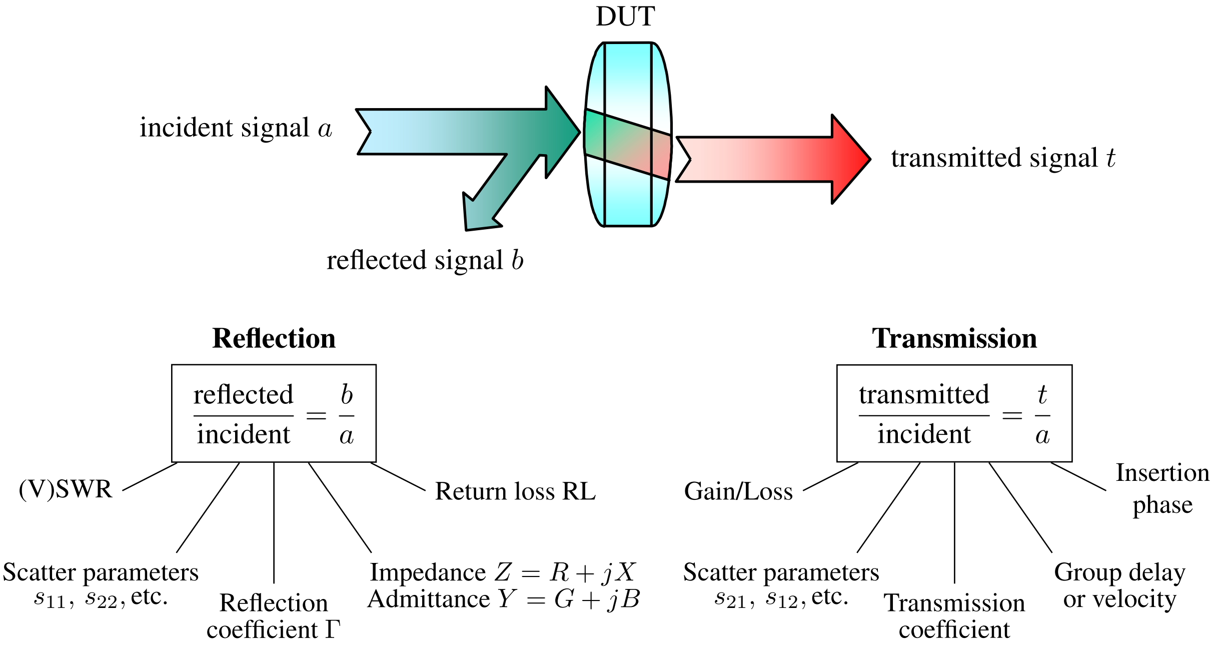

DUT
incident signal a
transmitted signal t
reflected signal b
Reflection
$\frac{\text{reflected}}{\text{incident}} = \frac{b}{a}$
(V)SWR
Return loss RL
Scatter parameters s_{11}, s_{22}, etc.
Reflection coefficient Γ
Impedance $Z = R + jX$
Admittance $Y = G + jB$
Transmission
$\frac{\text{transmitted}}{\text{incident}} = \frac{t}{a}$
Gain/Loss
Insertion phase
Scatter parameters s_{21}, s_{12}, etc.
Transmission coefficient
Group delay or velocity

Figure 2.68 Lightwave analogy to the VNA measurement characterization, original: Agilent.

[17] Note that network analyzer has a very different meaning in IT.

The principle circuit diagram of a two-channel vector network analyzer is shown in Fig. 2.69. Clearly, the directional couplers can be identified. It is used to separate the incident and reflected signals. Their general working principle is covered in Section 4.3.

All three signals (incident, reflected and transmitted) are mixed down into a low IF, where their power is detected, similar to a spectrum analyzer. Also, their phases are compared, enabling complex-valued measurements, such as impedance and admittance, S-parameters, phase and group delays, etc. In contrast, in the scalar network analyzer not all three signals are known down to their phases, making such measurements impossible.

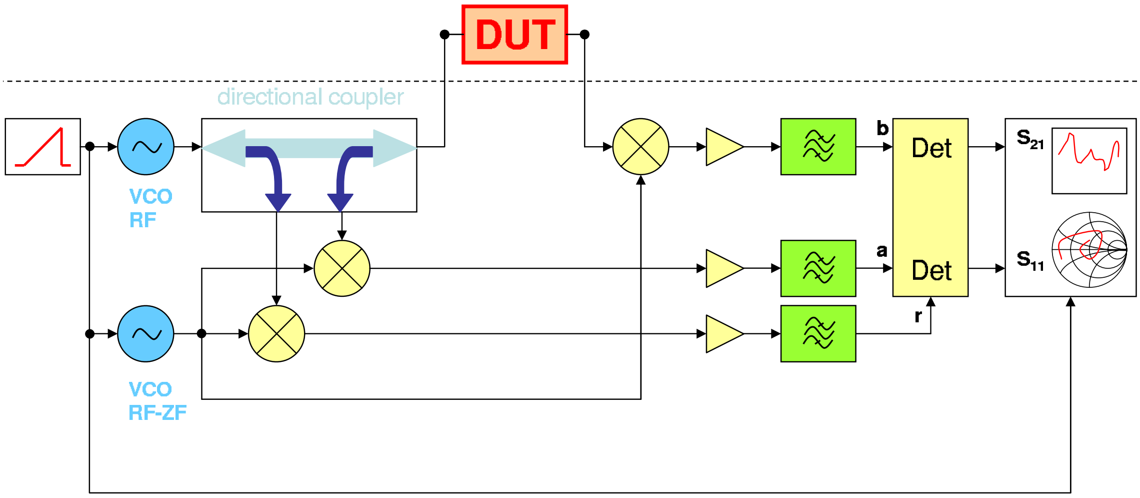
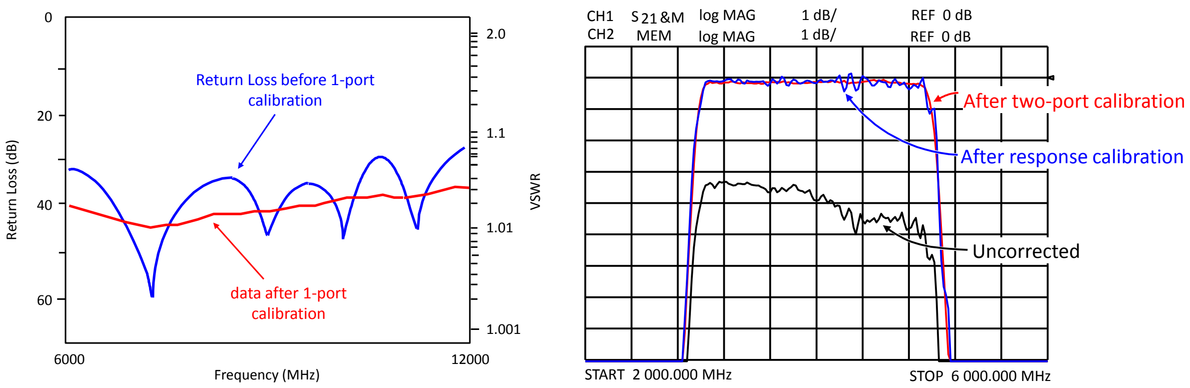


Figure 2.69 Vector network analyzer.

Calibration

The previously mentioned components, especially the directional coupler, but also dividers, mixers, internal transmission lines and even the connectors have tolerances and therefore show some kind of errors compared to their ideal counterparts. While some errors are already accounted for (to some degree) by programming according compensation measures into the devices software, in general these errors have to be corrected for by calibration right before the measurement. The actual calibration process to be used depends on the measurement taking place afterwards, where some processes are simpler and require fewer and less elaborate calibration equipment, others are more involving or have to be done by an according automatic calibration device. Also, the influences of cables used to connect the DUT to the VNA can (or should) be included in the calibration wherever possible. Then, the calibration (or reference) plane is at the end of the cables instead of just the connectors at the device, i.e., the electrical length of the cables is already accounted for. Fig. 2.70 shows two examples of how the calibration could positively influence the measurements and Table 2.5 shows an example of raw and corrected system data.

Note: the calibration is only valid for one parameter setting (i.e., frequency span, IF bandwidth, output power, etc.)! Particularly, if the frequency span is increased after the calibration process, the calibration data is lost! Also, in order to preserve the accuracy of the calibration termination, they should never be used for any other purpose (e.g., as general load termination).



(a) Return loss, VSWR

(b) Measuring filter insertion loss

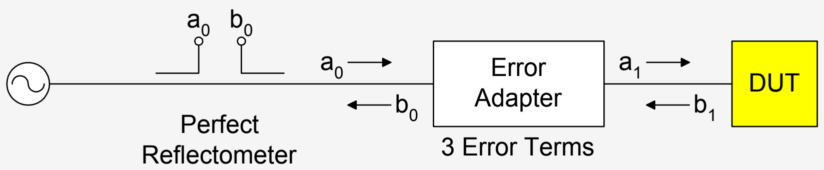
Figure 2.70 VNA calibration examples. Source: Agilent.

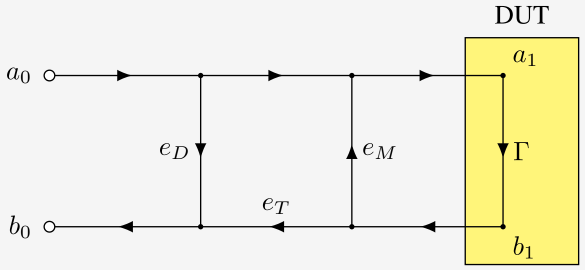
System property	Uncalibrated	Calibrated
Reflection tracking	< 2 dB	≤ 0.04 dB
Directivity	≥ 29 dB	≥ 46 dB
Source match	≥ 22 dB	≥ 38 dB
Transmission tracking	≤ 2 dB	≤ 0.06 dB
Isolation	≥ 130 dB	≥ 130 dB
Load match	≥ 22 dB	≥ 44 dB

Table 2.5 Comparison of typical uncalibrated and calibrated VNA system properties. Source: Rohde&Schwarz.

Example: One-port calibration

VNA calibration procedures are still subject of ongoing research, hundreds of papers have been published in this field of study – and when it is discussed in terms of its mathematical derivation it gets quite complex. However, this is beyond the scope of these lecture notes.
Just as an example, the calibration process is derived for a one-port, 3-term error model. This model only contains the unknown directivity, port match and tracking error. As in most cases, the errors are modeled as a fictitious lumped error adapter between the DUT and an ideal VNA.





e_D: Directivity error
e_M: Port mismatch
e_T: Tracking error
all other arrow factors 1

The measured reflection coefficient is

$$\Gamma_m = e_D + \frac{e_T\,\Gamma}{1-e_M\,\Gamma} = \frac{e_D + (e_T - e_D\,e_m)\,\Gamma}{1-e_M\,\Gamma} = \frac{b_0}{a_0} = \frac{e_D + \Delta_e\,\Gamma}{1-e_M\,\Gamma}\ ,$$

with the error combination $\Delta_e \overset{\triangle}{=} e_T - e_D\,e_M$, whereas the actual reflection coefficient is

$$\Gamma = \frac{\Gamma_m - e_D}{\Gamma_m\,e_M + \Delta_e}\ ,$$

which can be put into the form

$$e_D + \Gamma\,\Gamma_m\,e_M + \Gamma\,\Delta_e = \Gamma_m\ .$$

To solve this equation with the three error variables e_D, e_M and Δ_e, three equations with known $\Gamma_{1,2,3}$ and measured $\Gamma_{m\,1,2,3}$ are necessary:

$$\begin{bmatrix}1 & \Gamma_1\,\Gamma_{m1} & \Gamma_1\\ 1 & \Gamma_2\,\Gamma_{m2} & \Gamma_2\\ 1 & \Gamma_3\,\Gamma_{m3} & \Gamma_3\end{bmatrix}\begin{bmatrix}e_D\\ e_M\\ \Delta_e\end{bmatrix} = \begin{bmatrix}\Gamma_{m1}\\ \Gamma_{m2}\\ \Gamma_{m3}\end{bmatrix}\ ,$$

which is normally simplified by using the OSL (using an o̲pen, s̲hort and l̲oad termination; also called OSM, where the 'M' stands for m̲atch) calibration method, where

$$\text{open:}\quad \Gamma_1 = 1\,,\ \text{short:}\quad \Gamma_2 = -1\,,\ \text{load:}\quad \Gamma_3 = 0$$

which reduces the 3 × 3 matrix equation to $e_D = \Gamma_{m3}$ and a 2 × 2 matrix equation:

$$\begin{bmatrix}\Gamma_{m1} & 1\\ -\Gamma_{m2} & -1\end{bmatrix}\begin{bmatrix}e_M\\ \Delta_e\end{bmatrix} = \begin{bmatrix}\Gamma_{m1}-\Gamma_{m3}\\ \Gamma_{m2}-\Gamma_{m3}\end{bmatrix}\ .$$

Solved for the three errors results

$$\begin{aligned} e_D &= \Gamma_{m3} & & & e_D &= 0\\ e_M &= \frac{\Gamma_{m1}+\Gamma_{m2}-2\Gamma_{m3}}{\Gamma_{m1}-\Gamma_{m2}} & &\xrightarrow[\Gamma_{m3}\to 0]{\Gamma_{m1}=-\Gamma_{m2}\to 1} & e_M &= 0\\ \Delta_e &= \frac{\Gamma_{m1}\,\Gamma_{m3}+\Gamma_{m2}\,\Gamma_{m3}-2\Gamma_{m1}\,\Gamma_{m2}}{\Gamma_{m1}-\Gamma_{m2}}\ , & & & \Delta_e &= 1 \end{aligned}$$

for which, by setting the measured reflection coefficients to the wanted values, zero errors $e_D,\ e_M = 0$ and $e_T = 1$ result.

The one-port, 3-term error model is the simplest of all the error models, but still very common, since the calibration process with the three known port terminations is fairly easy and the terminations can be produced with high accuracy (small tolerances). Nevertheless, the quality (or rather the actual values) of the terminations is very important for a correct calibration.

More complex models, such as the also common two-port, 8-term (reducible to 7-term) VNA error model also require more inputs (in this case also a so-called *thru*, connecting the two ports with the port impedance Z_0 and a zero or known physical extent). Even more accurate models are available, such as the two-port, 15-term error model.

> Generally, in those cases the calibration returns even more values than there are unknowns, meaning the system is over-determined. There are various ways to deal with this, the simplest of course being a (weighted) least-square solution. Also, it is possible to verify the model with the remaining measured values and maybe do an iterating calibration.

Measurements

Network analyzers have become one of the most important measurement tools for characterizing the performance of high-frequency components and devices. The simplest and most straightforward measurements are

- Impedance matching, reflection coefficient, VSWR,
- Transmission coefficient, and
- Group and phase delay,

which are commonly known as the standard linear measurements. Also, the electrical length of the cable can be accounted for by de-embedding with the delay time.

Balanced lines and devices are generally much more complicated to measure with a VNA, since the ports itself are all unbalanced. Therefore, two ports have to be used as a single balanced port, thus a four port VNA is necessary to measure transmission characteristics of a fully balanced device. Also, the VNA needs to calculate the even and odd modes (and the respective impedance) internally, e.g., called *aperture simulator* by Agilent.

One of the hardest properties to be measured with a VNA is crosstalk – especially for multiport devices. This generally also requires the port isolation of the VNA to be calibrated, which is generally avoided since VNAs generally already have uncalibrated isolations high enough for most measurements and careless calibration (or without the required high-quality isolation calibration kit) might even worsen the isolation.

Other advanced measurement methods differ between the VNAs but might include

- mixer measurements,
- pulsed-mode measurements,
- time-domain analysis (via inverse Fourier transform), and
- nonlinear measurements (so-called X-parameters).

Further Reading

The information mentioned above is just an introduction to network analysis measurement. While the parameter setting is not as complex and versatile as for the spectrum analyzers for general return loss and impedance matching properties, they can become quite elaborate for advanced measurements.

For vector network analyzers, too, there exists a vast variety of literature, reaching from general basics (e.g. the *Fundamentals of Vector Network Analysis* by Michael Hiebel of R&S or numerous application notes and white papers by all the other manufacturers), over error analysis and notes and support on calibration to the advanced measurement topics.

2.5.3 Time-Domain Reflectometry

Frequency-domain measurements such as produced by vector network analyzers cannot always be interpreted directly. Very often, it is therefore useful to resort to a time-domain view of the reflections along a circuit. The motivations behind time-domain reflectometry (TDR) are

- time domain shows reflection positions (where to improve),
- relation to nature directly visible (resistive, inductive, capacitive),
- many engineers prefere time domain (logic analyzers, high-speed oscilloscopes).

There is, however, a price to pay. The equipment is usually more expensive. Whereas with a VNA, the detector, since working on an intermediate frequency, is relatively narrow-band, a TDR device needs a wideband detector. Besides, it must generate fast-rising pulses.

Loosely speaking, TDR is like radar. We are sending out signals and observe and analyze what is coming back. In mathematical terms, in TDR a step signal is sent through the circuit, represented by its s-transform

$$X_a(s) = \frac{1}{s}. \tag{2.100}$$

If the load at any position along the circuit is represented by the s-transform of the reflection coefficient $\rho(s)$, we can find the reflected wave by multiplying $X(s)$ by $\rho(s)$, thus

$$X_b(s) = X_a(s)\rho(s) = \frac{\rho(s)}{s}. \tag{2.101}$$

Transforming the sum of incident and reflected wave back into the time domain, we get the step response of the circuit

$$\begin{aligned} g(t) &= \mathcal{L}^{-1}(X_a(s) + X_b(s)) \\ &= \mathcal{L}^{-1}\left(\frac{1}{s} + \frac{\rho(s)}{s}\right). \end{aligned} \tag{2.102}$$

In theory, given the reflection coefficient as a function of s, the step response $g(t)$ could thus be computed. Usually, though, it is measured directly using TDR equipment.

2.6 Further Literature

One of the standard books for RF matters such as matching, scatter parameters, and amplifiers is the excellent textbook by Bowick [9], which is available in paperback form. A very detailed history and lots of applications regarding the Smith chart are provided in the book by the man himself [74]. Many figures of these lecture notes can also be found in [3], accompanied by a useful tutorial text in German.

3 Channel Access and Modulation

3.1 Modulation

3.1.1 Introduction

Modulation is the process of adapting the signal to the communication channel that it shall be carried over. The choice of the modulation type therefore depends on the physical properties of the channel. As we will shortly see, wireline channels prefer different types of modulation schemes to wireless channels. Important properties for the proper selection of the modulation scheme are available bandwidth, maximum power, and signal-to-noise ratio.

Similarly to the analog modulation schemes (known from earlier courses) such as AM (amplitude modulation), PM (phase modulation), and FM (frequency modulation), where the amplitude, the phase, and the frequency, respectively, of an RF carrier are modulated using a continuous-time information signal, there are a variety of digital modulation formats, which differ in the parameters that are modulated (amplitude, phase, frequency, or combinations thereof). Additionally, the modulation formats differ in their alphabet sizes, the simplest showing only two symbol values, more complicated ones up to 256 constellation points.

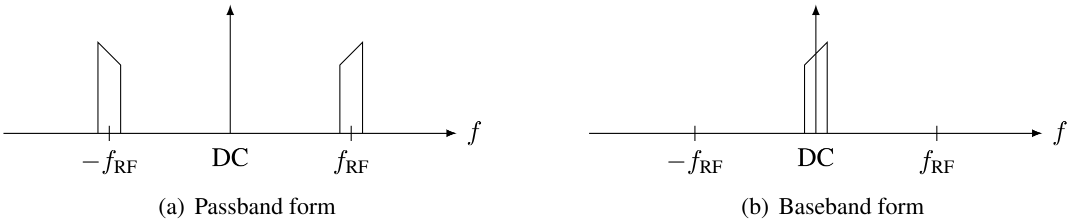
Modems

Modem is the word used to describe the device that transforms data into a format suitable for the channel at hand and back. It is the short version of MOdulator/DEModulator. Most widely known are modems that operate on the telephone channel, also called voiceband modems. They try to overcome the problems associated with data transferred over a medium that was initially designed for voice. Radio modems, on the other hand, have the challenging task to transform data into a format that is robust to an often hostile channel with effects such as fading, multipath propagation etc.

The Complex Baseband Concept

Most RF signals we encounter are passband signals, i.e., their bandwidths are small compared to their carriers. For the construction of the signals, the detection, and often also the simulation, a lowpass equivalent is often far easier to handle. Although an RF signal has a symmetric spectrum (around DC) in the general case, the spectrum is not symmetric around the RF carrier. If we mix the signal down to baseband, we obtain a signal that is no longer symmetric with respect to zero. That means in order to represent the signal properly, we need to resort to the complex domain. Fig. 3.1 shows the passband RF signal in the frequency domain and its corresponding complex baseband representation. The latter is often also referred to as the complex envelope. If $s_{\mathrm{bb}}(t)$ is the complex baseband form of a signal, its corresponding RF signal can be written as

$$s_{\mathrm{RF}}(t) = \mathrm{Re}\left(s_{\mathrm{bb}}(t)\,\mathrm{e}^{j2\pi f_{\mathrm{RF}}t}\right). \tag{3.1}$$



(a) Passband form

(b) Baseband form

Figure 3.1 Signal spectra.

Rather than displaying the continuous signal in the baseband, very often only a vector representation in the form of constellation points is shown. They represent the amplitude and phase of the baseband signal at the exact symbol sampling time.

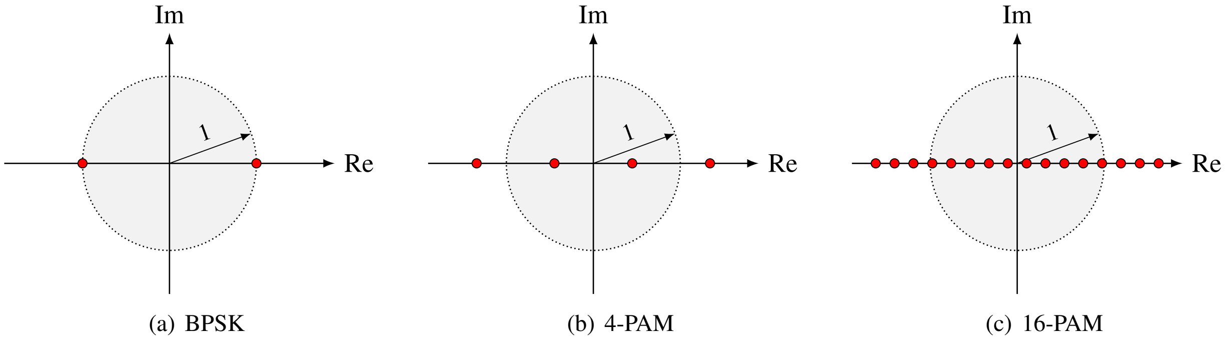
3.1.2 Real-Valued Constellations

Although in general the constellations are complex-valued, some special cases exist that can be represented by a real-valued baseband signal. Assume the phase of a signal can take on two different values, +1 and -1, depending on the symbol to transmit. Such a modulation scheme is called binary phase shift keying (BPSK) and its constellation diagram is shown in Fig. 3.2(a).

BPSK can be regarded as either a modulation scheme that changes between two phases (0 or 180°), or one that changes between two amplitude values (+1 or -1), and might therefore be called 2-PAM (binary pulse amplitude modulation). If we increase the alphabet size and allow more amplitude we end up with a general M-PAM scheme. A higher alphabet size M allows us to transmit more bits per symbol. The number of bits that can be transmitted for every symbol transmitted is

$$N = \log_2 M. \tag{3.2}$$

Two examples of M-PAM schemes are shown in Fig. 3.2(b) and (c).



(a) BPSK

(b) 4-PAM

(c) 16-PAM

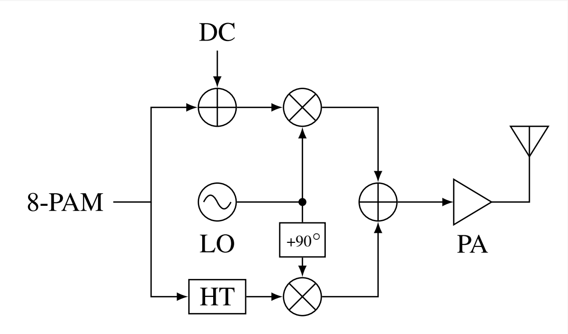
Figure 3.2 Comparison of the BPSK and PAM constellation diagrams.

Pulse-amplitude modulation (PAM) is a digital modulation scheme that is not very bandwidth efficient, since it has a double-sided, redundant spectrum. Making better use of the spectrum can be achieved in three different ways: The alphabet size might be increased, information might be added to the imaginary part of the baseband representation leading to quadrature-amplitude modulation (QAM), or one sideband might be eliminated.

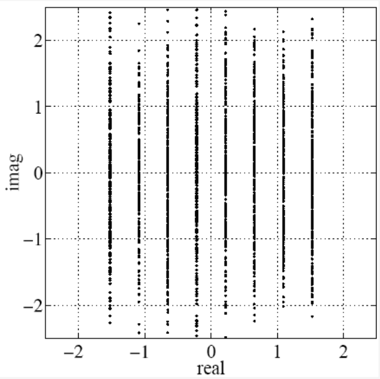
Example: 8-VSB in ATSC (North American HDTV Broadcast Standard)

Vestigial sideband (VSB) modulation is a practical example, where spectral efficiency is enhanced by eliminating one sideband by filtering a PAM signal to essentially half the bandwidth. 8-VSB is the modulation scheme adopted for the high-definition television (HDTV) broadcasting standard in the US and Canada.

An implementation for an 8-VSB signal generator is visualized in the following, along with the resulting constellation diagram. The real part of the VSB signal remains the same as that of an 8-PAM signal except for an additional DC offset, generating an RF carrier which facilitates easy synchronization using a phase-locked loop (PLL) on the receiver end. The imaginary part is generated using the *Hilbert transform* (HT) of the real part of the 8-PAM signal.



Direct-conversion type architecture VSB signal generator.



Constellation diagram: The imaginary part is roughly Gaussian distributed.

In ATSC, the symbol rate is 10.76 Mbaud where symbol includes two bits from the MPEG transport stream which are trellis modulated (a forward error correction coding method) to produce the 3-bit symbols (for $2^3 = 8$ constellations, thus 8-PAM). Such a signal would naturally occupy a spectrum at least 10.76 MHz wide; however, with the VSB technique, the spectrum occupancy is reduced to about 6 MHz.

ATSC is not as spectrally efficient as its European counterpart, DVB-T, which uses OFDM. This comes at the advantage of lower system complexity (relaxed peak-to-average-power issues) and lower sensitivity to interferers and doppler shift.

3.1.3 Complex-Valued Constellations

Rather than starting out with a PAM scheme and filtering off one redundant sideband, we may directly introduce a complex-valued symbol alphabet. The easiest such scheme is QPSK (quaternary phase shift keying). Again, this scheme can be alternatively regarded as 4-QAM (quaternary quadrature amplitude modulation). In QPSK, a constellation diagram of which is depicted in Fig. 3.3, four different phase positions are possible. In the more general case of QAM schemes, we get constellations as displayed in Fig. 3.4.

With QAM schemes, both the amplitude and the phase are varied with the symbol choice. Two constraints usually apply: First, the symbol alphabet size M should be a square, allowing us to arrange the constellation points in a square. Second, M should be a power of two, allowing us to send $\log_2(M)$ bits per symbol. Since both constraints are only possible for $M = 2^{2m}$ where m is a positive integer, only an even number of bits can be packed in the symbol of a square QAM constellation. Other non-square QAM constellations are still

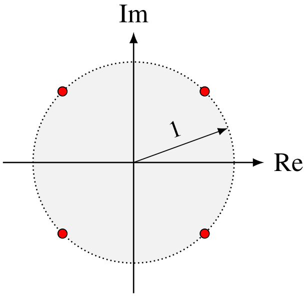
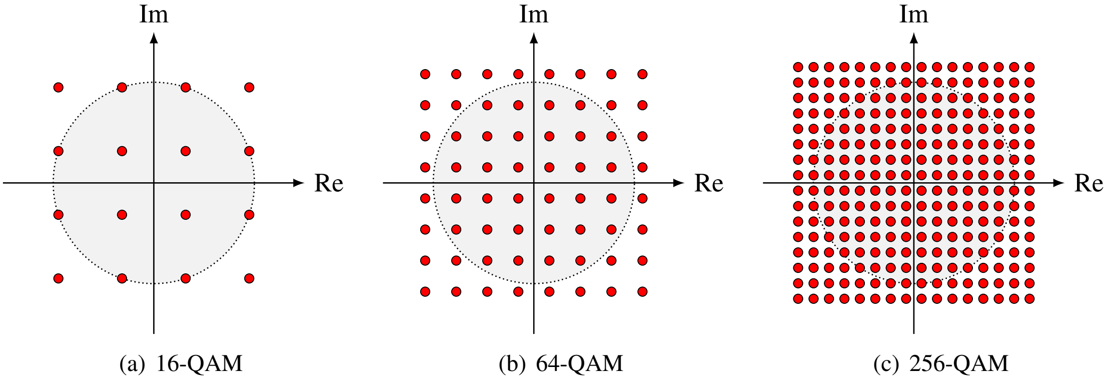


Figure 3.3 Constellation diagram of a QPSK signal.



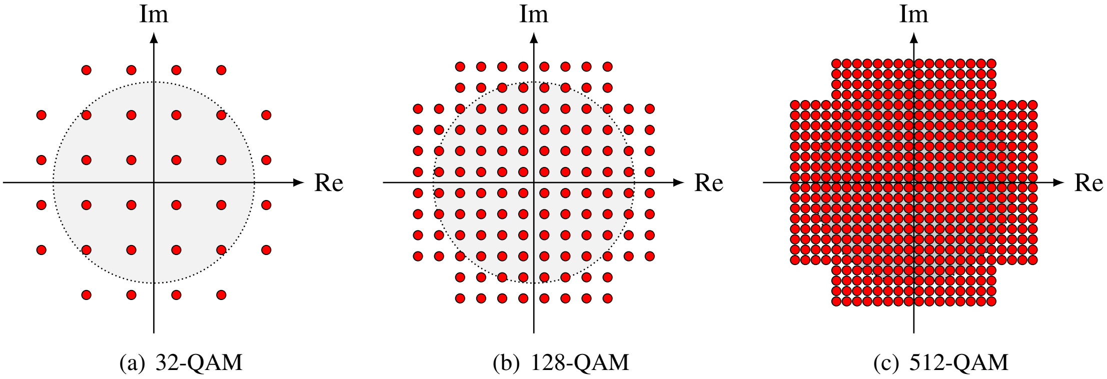
(a) 16-QAM (b) 64-QAM (c) 256-QAM

Figure 3.4 Constellation diagrams of QAM signals.

possible, and are used in current technology, e.g., the voice modem standards V.32 and V.33 use 32-QAM and 128-QAM, respectively, shown in Fig. 3.5. Because an odd power of two can always be written as

$$\begin{aligned} M &= 2^{2m+1} = 2^{2m-2} \cdot 2^3 = (2^{m-1})^2 \cdot (3^2 - 1) \\ &= (2^{m-1} \cdot 3)^2 - (2^{m-1})^2, \end{aligned} \tag{3.3}$$

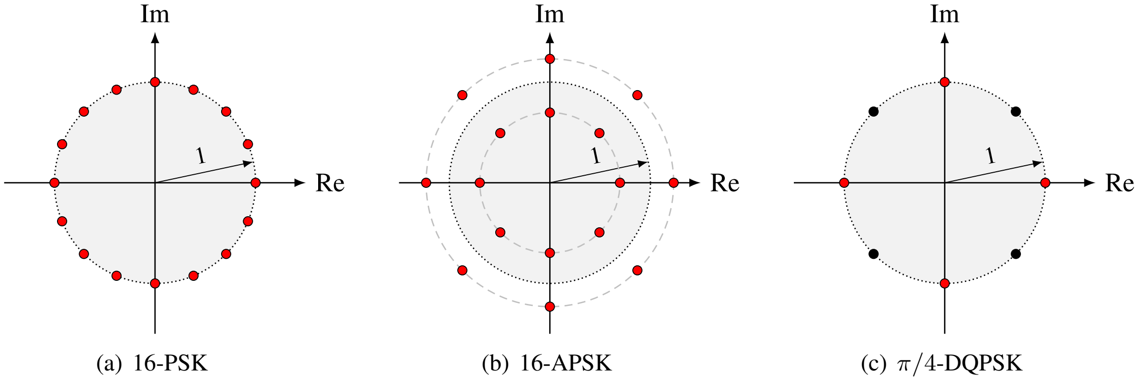
a difference of two squares, each of which is divisible by four for integer values of $m \geq 2$, a constellation can always be built where a large square is pruned by four small corner squares. As an example, for 32-QAM we skip the four corner points of a 36-point square ($m = 2$).



(a) 32-QAM (b) 128-QAM (c) 512-QAM

Figure 3.5 Constellation diagrams of non-square QAM signals.

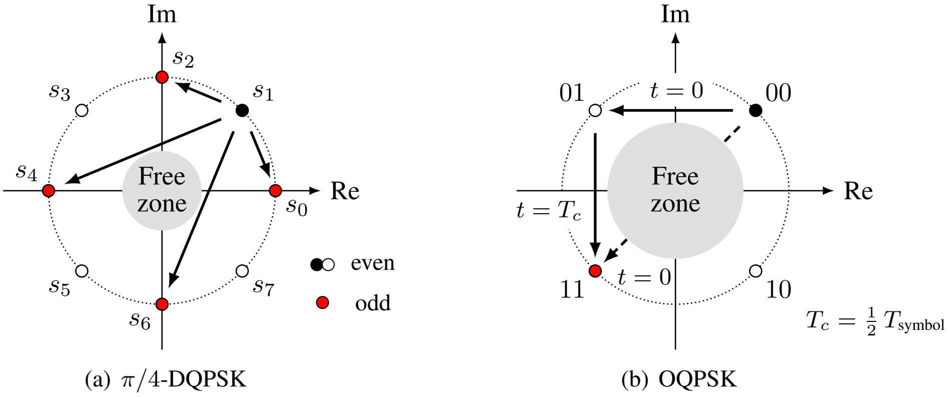
As mentioned, such schemes are popular in voiceband modems. In wireless communications, where the amplitude might fade fast over time, modulation formats that show a constant amplitude are preferred. This leaves the phase as a possible modulation parameter. Such schemes are known under the name of M-PSK (see Fig. 3.6(a)). An unknown gain parameter is no longer a problem, since all the information is contained in the phase of the signal. However, whenever the channel gain does not vary to much or too fast, the amplitude can be modulated in this configuration once again, leading to APSK (see Fig. 3.6(b)). Compared to QAM, the smaller number of power levels can be advantageous, particularly for satellite communications.



(a) 16-PSK (b) 16-APSK (c) $\pi/4$-DQPSK

Figure 3.6 Constellation diagrams of PSK signals.

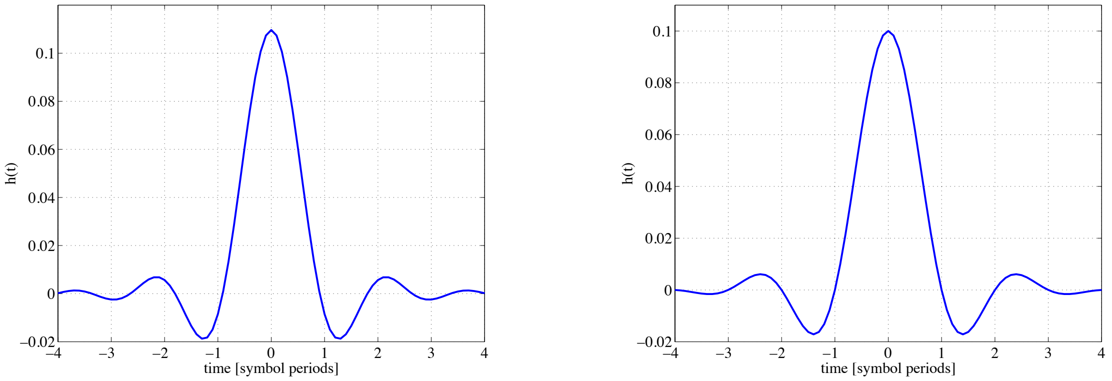
A PSK scheme of special meaning is $\pi/4$-DQPSK, shown in Fig. 3.6(c). It is used in both Asian and European systems for private radio systems such as Tetra and Tetrapol[1]. The D stands for differential and works around the fact that the absolute phase of a QPSK signal is ambiguous. By differentially encoding the signal stream, the absolute phase of the signal is no longer important. Rather, the information is carried in the phase change from symbol to symbol. The $\pi/4$ in the name refers to the phase shift introduced to every other symbol. This decreases the dynamic requirements of the power amplifier at the transmitter, because transitions through the origin are avoided as can be seen in Fig. 3.9. There are two related constellations, that share this property of avoiding the origin: the $\pi/4$-QPSK and the OQPSK modulation scheme, both illustrated in Fig. 3.7. Fig. 3.8 shows the transmit and receive filter responses typically used by QPSK-based constellations. Fig. 3.9 shows the constellations at the point of transmitting and after the receive filter.



(a) $\pi/4$-DQPSK (b) OQPSK

Figure 3.7 Free zone comparison of $\pi/4$-DQPSK and OQPSK.

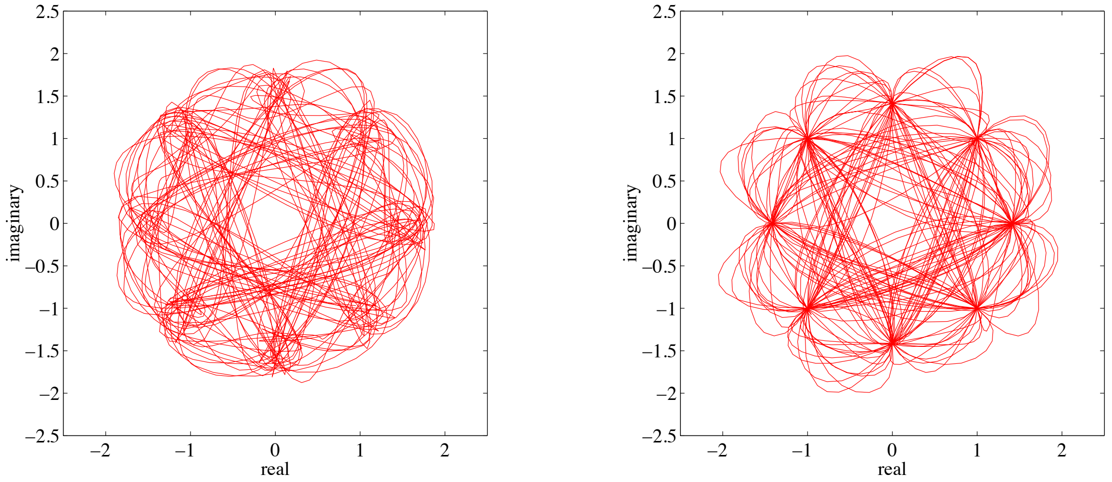
[1] The Polycom system in Switzerland is a Tetrapol system.



(a) Root raised cosine filter, $\rho = 0.35$ (IS-95, TETRA)

(b) Raised cosine filter, $\rho = 0.35$ (IS-95, TETRA)

Figure 3.8 Filter impulse responses used in $\pi/4$-DQPSK.



(a) filtered constellation as transmitted

(b) filtered constellation after receive filter

Figure 3.9 $\pi/4$-DQPSK modulation format.

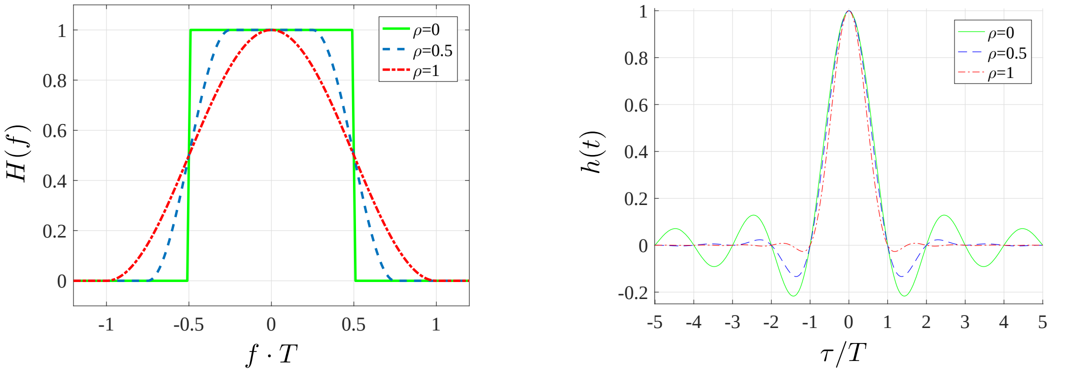
It can clearly be seen from either the impulse response or the resulting constellation that only the overall filter is a Nyquist filter, which does not create intersymbol interference (ISI). Also for other modulation formats, communication signals are often pulse shaped by a raised-cosine filter, whose frequency response is defined by

$$H(f) = \begin{cases} 1, & 0 \leq |f| \leq \frac{1-\rho}{2T} \\ \dfrac{T}{2}\left[1 - \sin\left(\dfrac{\pi(|f|T - \frac{1}{2})}{\rho}\right)\right] = \dfrac{T}{2}\left[1 + \cos\left(\dfrac{\pi T}{\rho}\left[|f| - \dfrac{1-\rho}{2T}\right]\right)\right] & \frac{1-\rho}{2T} \leq |f| \leq \frac{1+\rho}{2T} \\ 0 & \frac{1+\rho}{2T} \leq |f| \end{cases} \tag{3.4}$$

and shown in Fig. 3.10(a). Applying the inverse Fourier transform we get the impulse response

$$\boxed{h(t) = \frac{\sin\left((t/T)\pi\right)}{(t/T)\pi} \cdot \frac{\cos\left(\rho(t/T)\pi\right)}{1 - 4\rho^2(t/T)^2},} \tag{3.5}$$

with ρ being the roll-off factor, $0 \leq \rho \leq 1$. The symbol duration is thereby denoted T. The impulse responses for different ρ are shown in Fig. 3.10(b). Other pulses and their Fourier transform can be found in the appendix.

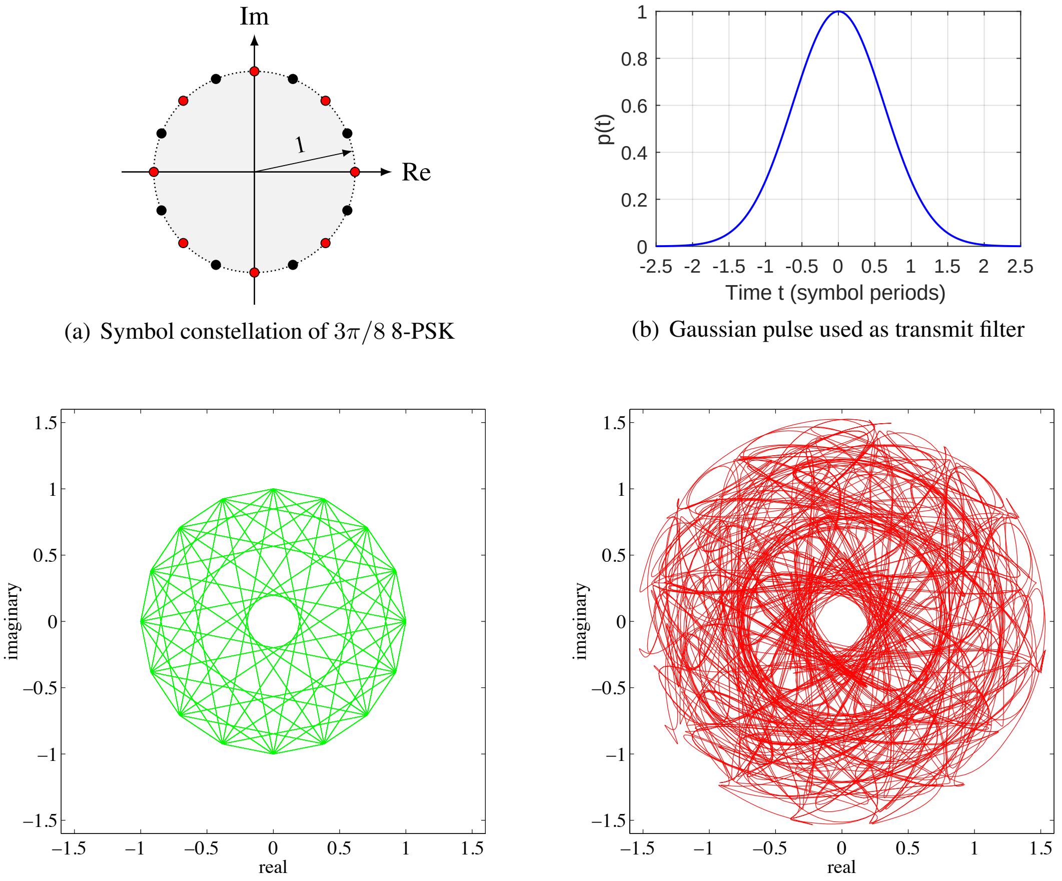


(a) Frequency spectrum

(b) Impulse response for different roll-off factors ρ

Figure 3.10 Spectrum and impulse response of the raised-cosine filter.

The evolution path of GSM (also called 2.5G standards) applies PSK in a special form. Fig. 3.11 shows the EDGE modulation scheme, so-called $3\pi/8$ 8-PSK, which is similar in nature to $\pi/4$-DQPSK.



(a) Symbol constellation of $3\pi/8$ 8-PSK

(b) Gaussian pulse used as transmit filter

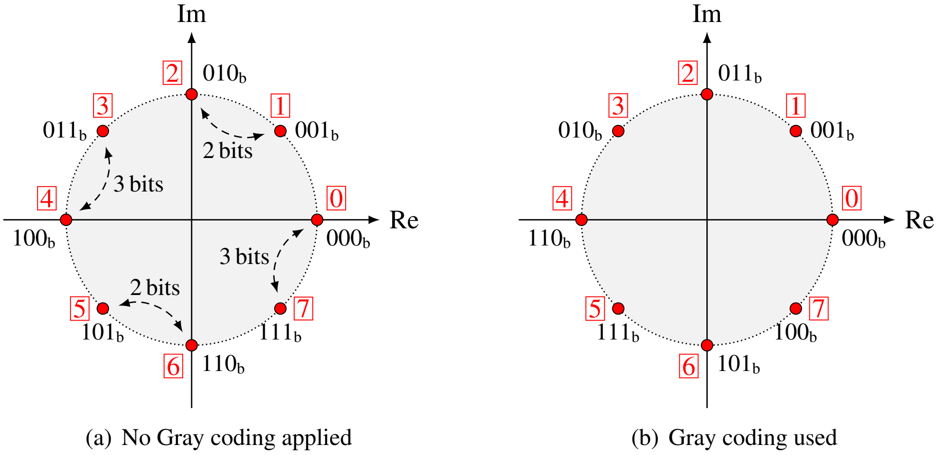
(c) Continuous constellation (unfiltered)

(d) Filtered constellation as transmitted

Figure 3.11 EDGE modulation format.

3.1.4 Gray Coding

In order to minimize the number of bit errors, neighboring symbols of a constellation should consist of as few bit differences as possible, while still being different of course. Such a constellation thus consists of points whose bit assignments are different by one bit only when compared to their immediate neighbors. Such a bit assignment is referred to as Gray coding, after its originator Gray. When a symbol is detected in error, there is a high chance that a neighboring symbol is picked instead. In this case, all but one bit are still correct.



(a) No Gray coding applied (b) Gray coding used

Figure 3.12 Comparison of ordinary and Gray coding of a 8-PSK modulated signal.

3.1.5 Nonlinear Modulation Schemes

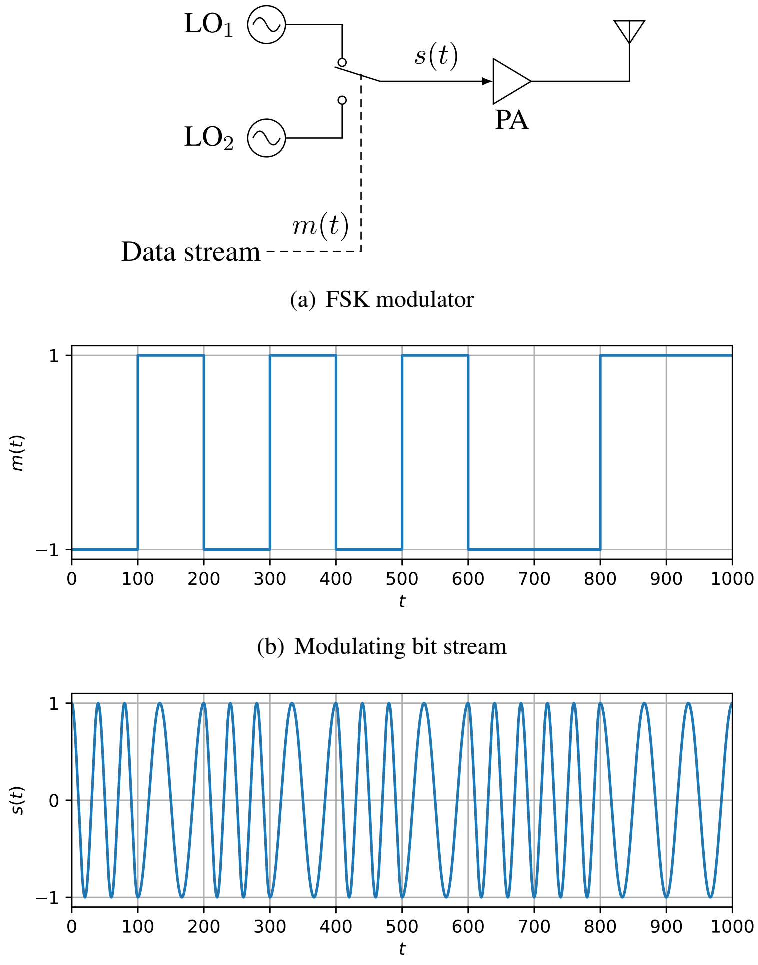
The modulation schemes mentioned so far all belong to the group of linear modulation schemes. Other modulation schemes exist, which shall be treated in the sequel.

Frequency Shift Keying (FSK)

FSK is essentially FM with a quantized modulating signal. Most often, just 2 different levels (binary) are used. In essence, one frequency is transmitted for a binary 0, another for a binary 1 during the time of a symbol. A binary sequence and its FM signal are plotted in Fig. 3.13. The relationship between the maximum frequency deviation and the data rate is called the modulation index

$$h = \frac{\Delta f}{f_\mathrm{d}}. \tag{3.6}$$

The maximum frequency deviation usually refers to the spacing between the smallest and the highest frequency (which is twice of what would be expected from the definition of an analog FM signal). To choose the right modulation index, we have to agree to a compromise: a small modulation index means very bandwidth-efficient operation at a price of harder differentiation between the symbols; a high modulation index may provide enough space to safely decide between the frequencies, but may lead to a prohibitively wide spectrum. For a high data rate for a given spectrum, we need a small modulation index. For noncoherent detection, the minimum modulation index for orthogonal FSK is 1. For coherently detected orthogonal FSK,



(a) FSK modulator

(b) Modulating bit stream

(c) Modulated RF signal

Figure 3.13 Modulated signal using FSK with $h = 1$.

the minimum modulation index is 0.5. The respective receivers for coherent detection and non-coherent detection of an FSK signal are shown in Figs. 3.14 and 3.15.

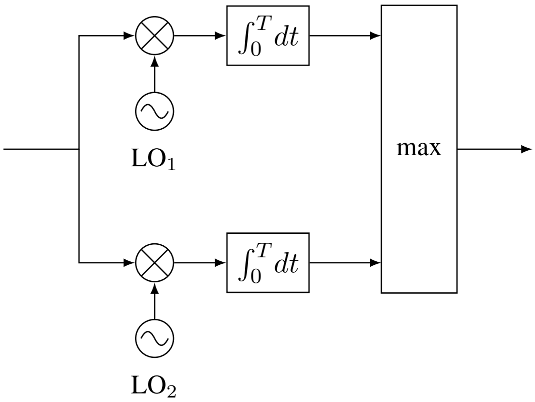


Figure 3.14 Coherent receiver.

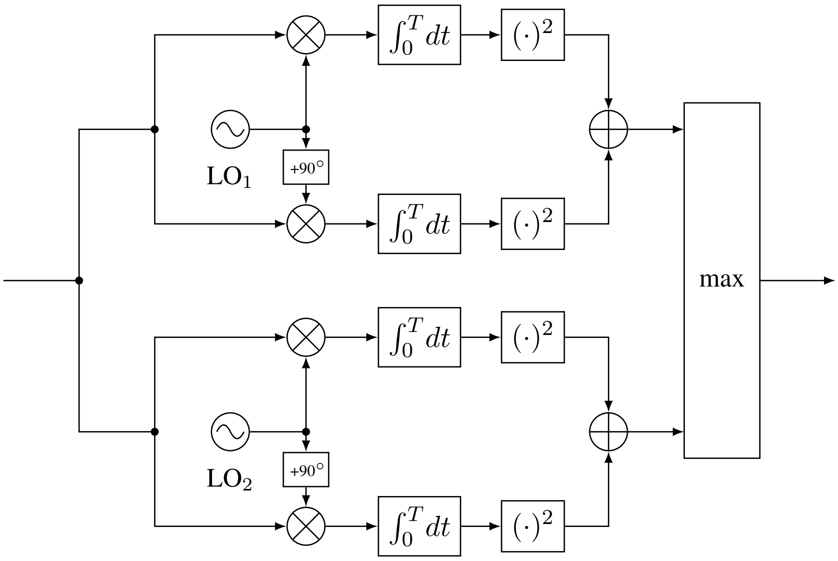
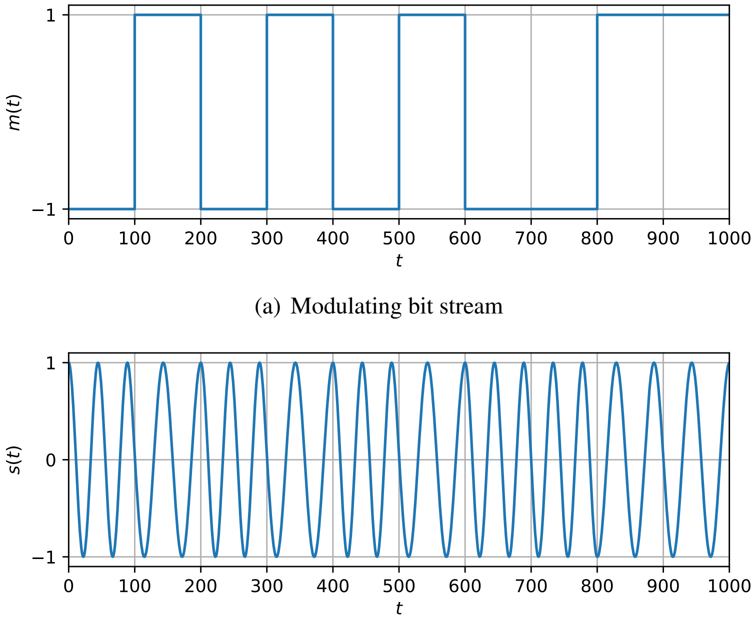


Figure 3.15 Non-coherent receiver.

An FSK scheme with a modulation index of 0.5 is also called fast frequency shift keying (FFSK) or minimum shift keying (MSK). A binary sequence and its FFSK signal are plotted in Fig. 3.16.



(a) Modulating bit stream

(b) Modulated RF signal

Figure 3.16 Modulated signal using FFSK ($h = 0.5$).

Continuous-Phase Modulation (CPM)

An FSK signal could in theory be generated by switching between two local oscillators (LO) running at a fixed but different frequency each. This could mean that the phases of the two LOs are independent to each other, which might cause high phase jumps at the time of switching. Such phase jumps usually lead to a wide spectrum, a behavior usually not accepted. On the other hand, as long as the frequencies are shifted according to the specification of FSK, the phase is a degree of freedom. A possible FSK modulator showing a continuous-phase property can be constructed using a voltage-controlled oscillator (VCO), as shown in Fig. 3.17.

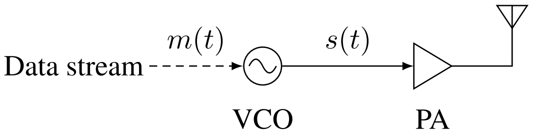


Figure 3.17 Continuous phase modulator (CPM).

Other CPM schemes are in wide use, too. Most of them can be constructed by filtering the data bit stream prior to their input into the VCO.

Gaussian Minimum Shift Keying (GMSK)

Frequency-shift keying (FSK) with a modulation index of $h = 0.5$ is called *minimum-shift keying* (MSK) or *fast frequency-shift keying* (FFSK). GMSK, short for *Gaussian Minimum-Shift Keying* is a special form of MSK, where a Gaussian premodulation filter of a specific bandwidth B (3 dB cutoff frequency) is applied.

The combination of constant envelope and relatively high spectral efficiency made the GMSK modulation scheme highly popular. Both GSM (Global System for Mobile Communication), see Sect. 10.4, and DECT (Digital European Cordless Telephone) use GMSK, with a bandwidth symbol-time product BT of 0.3 and 0.5, respectively. Low BT makes the modulation scheme even more bandwidth-efficient at the price of increased intersymbol interference (ISI).

The baseband representation of a GMSK signal can be expressed as

$$s(t) = A\,\mathrm{e}^{j\varphi(t)} = A \exp\left[j2\pi \sum_{i=0}^{k} \alpha_i \phi(t - iT)\right] \qquad kT \le t < (k+1)T, \tag{3.7}$$

with α_i being the data symbols $\in \{+1, -1\}$ and the phase function

$$\phi(t) = \int_{-\infty}^{t} g(\tau)\, d\tau. \tag{3.8}$$

The Gaussian pulse form is given by

$$g(t) = \frac{1}{8T}\left\{\operatorname{erfc}\left[\sqrt{\frac{2}{\ln 2}}\pi B\left(t - \frac{T}{2}\right)\right] - \operatorname{erfc}\left[\sqrt{\frac{2}{\ln 2}}\pi B\left(t + \frac{T}{2}\right)\right]\right\}, \tag{3.9}$$

with the complementary error function [52]

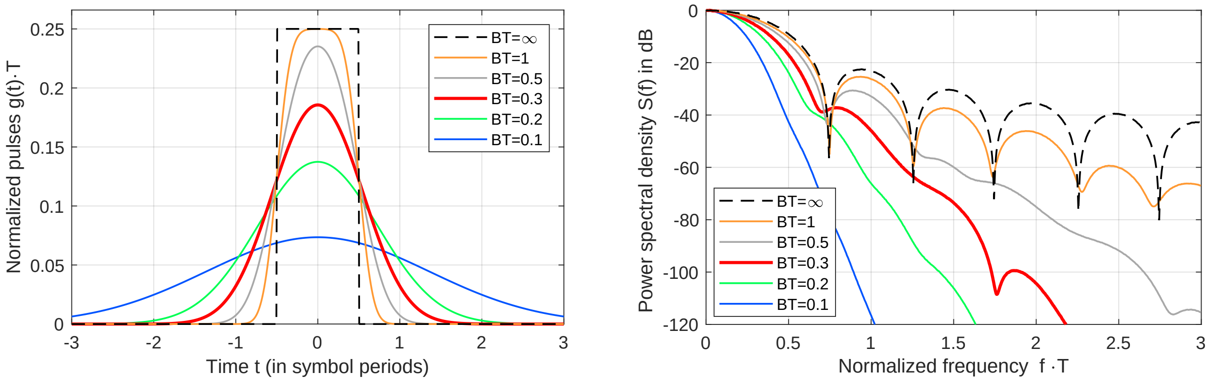
$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} \mathrm{e}^{-t^2}\, dt. \tag{3.10}$$

Solving the integral in (3.8) with (3.9) leads to the phase function [42]

$$\phi(t) = \frac{1}{4} + \frac{1}{8}\left\{\left(t - \frac{T}{2}\right) \cdot \operatorname{erfc}\left[\sqrt{\frac{2}{\ln 2}}\pi B\left(t - \frac{T}{2}\right)\right] - \left(t + \frac{T}{2}\right) \cdot \operatorname{erfc}\left[\sqrt{\frac{2}{\ln 2}}\pi B\left(t + \frac{T}{2}\right)\right]\right\}$$

$$+ \frac{\sqrt{\ln 2}}{8\sqrt{2}BT\pi^{\frac{3}{2}}}\left\{\mathrm{e}^{-\left[\sqrt{\frac{2}{\ln 2}}\pi B(t+\frac{T}{2})\right]^2} - \mathrm{e}^{-\left[\sqrt{\frac{2}{\ln 2}}\pi B(t-\frac{T}{2})\right]^2}\right\}. \quad (3.11)$$

Thus, a GSM transmitter is a simple combination of a Gaussian premodulation filter and an FM modulator. Fig. 3.18 shows the premodulation pulse of a GMSK signal as deployed by GSM, as well as examples with different BT, along with the resulting spectral shapes.



(a) Gauss-filtered modulators for different BT

(b) Normalized spectral shapes of GMSK signals

Figure 3.18 Gaussian premodulation filters and the resulting spectra; $BT = 0.3$ is used in GSM.

The modulation plot itself is simply a circle (constant-envelope scheme). Fig. 3.19 shows a bit pattern and its corresponding GMSK signal in a phase representation.

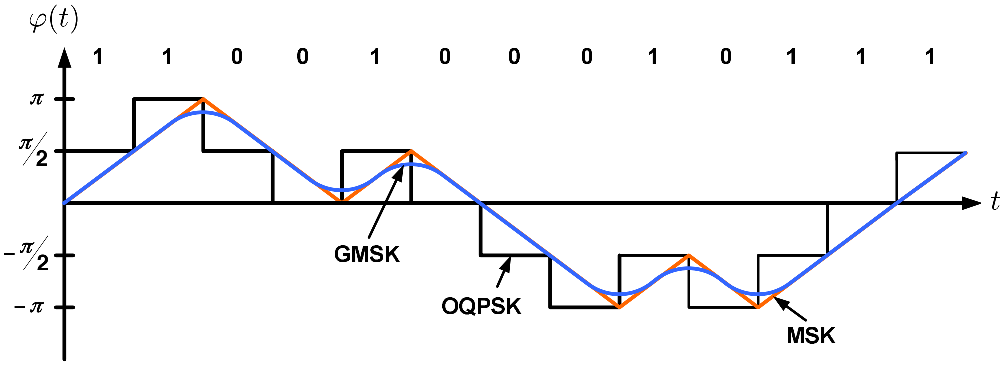


Figure 3.19 Phase course of a GMSK signal.

Since for GMSK with an arbitrary BT there are no exact expressions, Monte-Carlo simulations using Matlab were carried out to simulate the spectra of the respective modulation schemes in Fig. 3.18(b). For MSK and

OQPSK, the exact expressions for their respective spectra are given by [52]

$$P_{\text{OQPSK}} = 2A^2T \left(\frac{\sin 2\pi fT}{2\pi fT}\right)^2, \tag{3.12}$$

$$P_{\text{MSK}} = \frac{16A^2T}{\pi^2} \left(\frac{\cos 2\pi fT}{1 - 16f^2T^2}\right)^2. \tag{3.13}$$

Depending on what one compares, the two schemes have different characteristics in their respective spectrum. While QPSK has a smaller main lobe than MSK for the same bit rate, the side lobes of the latter spectrum fall off faster and MSK has a smaller out-of-band power (power beyond $f \cdot T = 1$).

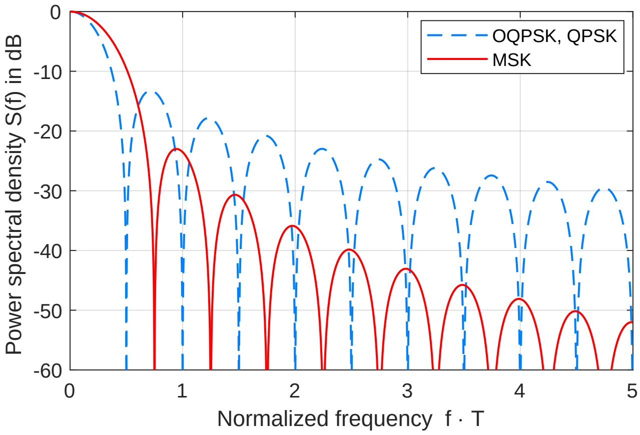


Figure 3.20 Normalized power spectral densities of MSK and QPSK signals.

3.1.6 Performance

Error Vector Magnitude

In reality, the ideal constellation points or rather the trajectories through these points can never be met exactly. The deviation of the actual signal from the ideal signal is often expressed in *error vector magnitude*, or EVM. The EVM is the average length of the difference vector between the real and the ideal signal vectors and is usually expressed as a %, see Fig. 3.21. The origin of EVM is manifold and includes effects such as carrier leak, gain and phase imbalance in the IQ-modulator, carrier noise, digital filter inaccuracy, and so on. Many signal analyzers can measure EVM of different modulation formats.

Error Probability

The quality of transmission in digital communications is most often expressed in some kind of error probability. Error probability is a very general term. It may refer to the symbol transmitted, to the bits that comprise a symbol, or to a whole frame of data bits. In case of bits, it may concern raw bits (in the form transferred over the channel) or information bits, which are the decoded bits after error decoding. Naturally, the latter bit error rate should be smaller than the raw bit error rate.

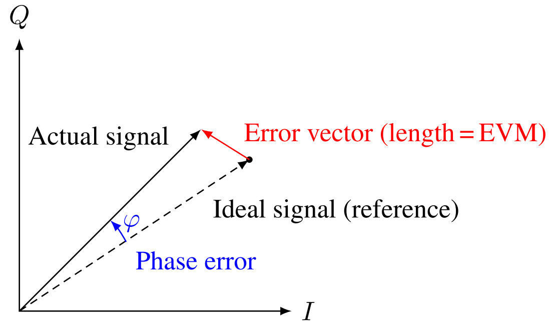


Figure 3.21 Error vector magnitude (EVM).

If we want to find out about such error rates, there are different methods. One method would be to just build a system, let it run for a while and count the errors occurring. This is the most costly method. Furthermore, it has the disadvantage that very small error rates need extremely long running times to verify. A second method is to simulate systems using so-called Monte-Carlo (MC) simulation methods. They receive their name from their approach of essentially letting a random string of data be disturbed by random noise and measuring the resulting error rate. The advantage is that the system does not actually need to be built, just a good model needs to be available. On the other hand, simulations might not run in real-time. Hence, the simulations of low symbol error rates can last even longer. A third method is to compute error probabilities in an analytic way. This method is described in the following.

The easiest error rate to compute is the symbol error probability. Let us consider the BPSK case, which is the easiest of all signaling schemes. As with any binary signaling schemes, the size of the amplitude values has a direct impact on the error probability. In the absence of interference, no error would ever occur, regardless the distance between the constellation points. However, thermal noise at the input to the receiver is always present. The noise is in the form of an additive Gaussian distributed signal with the power $\sigma^2 = N_0/2$. The signal, on the other hand, has the symbol energy E_S. When a positive amplitude is sent, the combined signal with the noise has a distribution $p(x|s = -1)$ like in Fig. 3.22 (lower), otherwise $p(x|s = +1)$ as in Fig. 3.22 (upper). A detector has to decide which symbol has been sent. Although a negative received value is possible when a positive symbol has been sent (and a strong negative noise signal is added), such a scenario is less likely than an original negative signal. It is therefore reasonable to set the decision level in the middle of the two signaling levels (at zero).

That way, the probability of error is the tail of the Gaussian distribution

$$\begin{aligned}
P_E &= \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma}\,\mathrm{e}^{-\frac{(x+\sqrt{E_S})^2}{2\sigma^2}}\,dx \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}\sqrt{\frac{N_0}{2}}}\,\mathrm{e}^{-\frac{(x+\sqrt{E_S})^2}{N_0}}\,dx \\
&= \int_{\sqrt{E_S}}^\infty \frac{1}{\sqrt{2\pi}\sqrt{\frac{N_0}{2}}}\,\mathrm{e}^{-\frac{x^2}{N_0}}\,dx \\
&= Q\left(\sqrt{\frac{2E_S}{N_0}}\right).
\end{aligned} \tag{3.14}$$

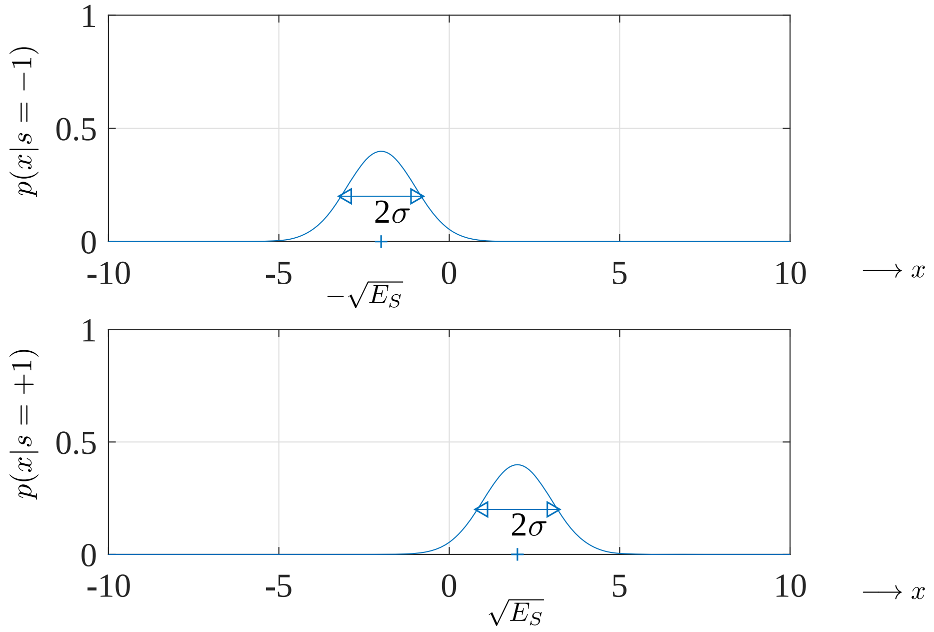


Figure 3.22 Probability density functions of signal after additive noise. Top: Negative symbol sent. Bottom: Positive symbol sent.

$Q(.)$ is called the Q-function or the tail function of the Gaussian distribution. It cannot be written in a direct form, but is often tabulated. Such a table can be found in Appendix A.5.5. Other symbol error probabilities can be derived using a simular line of argument. The results are collected in Table 3.1 and in Fig. 3.23.

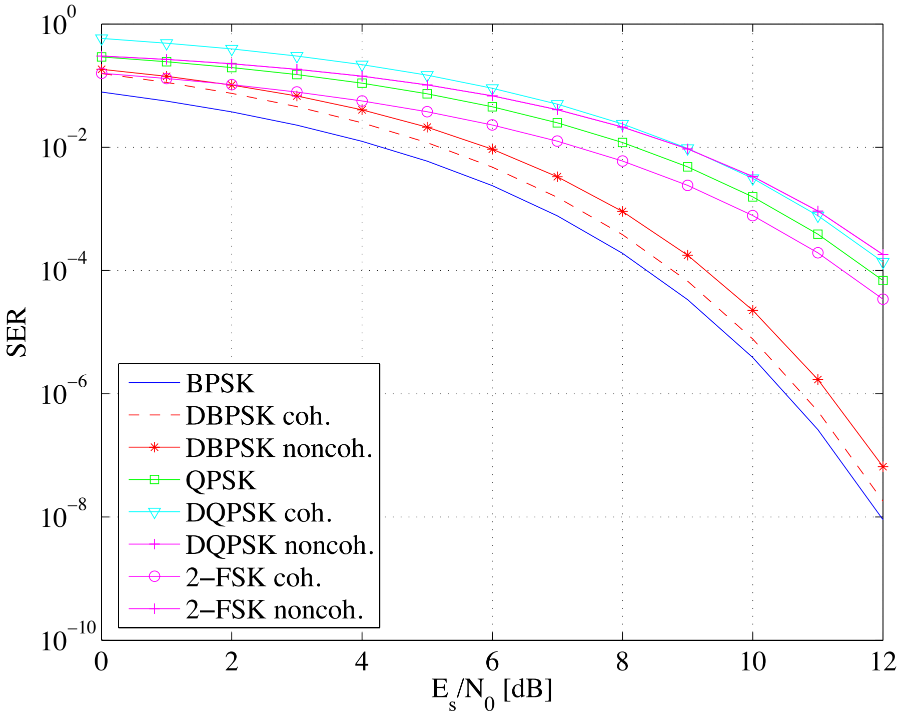


Figure 3.23 SERs of some digital modulation formats.

Modulation scheme	Levels/range	Normalized amplitude	SER (AWGN)	Comments
BPSK	$\pm A$	$A = 1$	$Q\left(\sqrt{\frac{2E_s}{N_0}}\right)$	=BER
DPSK	$\pm A$	$A = 1$	$2Q\left(\sqrt{\frac{2E_s}{N_0}}\right)$	coherent
			$\frac{1}{2}\exp\left(-\frac{E_s}{N_0}\right)$	noncoherent
QPSK	$(\pm 1 \pm j)A$	$A = \frac{1}{\sqrt{2}}$	$1 - \left(1 - Q\left(\sqrt{\frac{E_s}{N_0}}\right)\right)^2$	coherent
DQPSK	$(\pm 1 \pm j)A$	$A = \frac{1}{\sqrt{2}}$	$2\left(1 - \left(1 - Q\left(\sqrt{\frac{E_s}{N_0}}\right)\right)^2\right)$	coherent
			$\frac{1}{2}\exp\left(-\frac{E_s}{2N_0}\right)$	noncoherent
M-PSK	$\frac{1}{M}\sum_{m=1}^{M} A\delta\left(x - e^{j(2m-1)\frac{\pi}{M}}\right),\ m \leq M$	$A = 1$	$\leq 2Q\left(\sqrt{\frac{2E_s}{N_0}}\sin\frac{\pi}{M}\right)$	coherent
M-PAM	$\pm(2m-1)A, \quad m \leq M/2$	$A = \sqrt{\frac{3}{M^2-1}}$	$2\frac{M-1}{M}Q\left(\sqrt{\frac{E_s}{N_0}}\sqrt{\frac{6}{M^2-1}}\right)$	
M-QAM	$(\pm(2m-1) \pm j(2n-1))A,\ m,n \leq \sqrt{M}/2$	$A = \sqrt{\frac{3}{2(M-1)}}$	$1 - \left(1 - 2\frac{\sqrt{M}-1}{\sqrt{M}}Q\left(\sqrt{\frac{E_s}{N_0}}\sqrt{\frac{3}{M-1}}\right)\right)^2$	
2-FSK	$\Delta f = h \cdot \frac{1}{T},\ h_{\min} = 0.5$		$Q\left(\sqrt{\frac{E_s}{N_0}}\right)$	orthogonal, coherent
	$h_{\min} = 1$		$\frac{1}{2}\exp\left(-\frac{E_s}{2N_0}\right)$	orthogonal, noncoherent

Table 3.1 Constellations and SERs in an AWGN channel of some digital modulation formats. Most SERs can be found in [52].

Up to now we have considered symbol error rates. For bit error rates, the computation is often slightly more complicated. For binary signaling, the symbol error rates (SER) and the bit error rates (BER) are identical, since one symbol consists of one bit. It is interesting to note that from Table 3.1 we see that

$$\begin{aligned} \mathrm{SER}_{\mathrm{QPSK}} &= 1 - \left(1 - Q\left(\sqrt{\frac{E_S}{N_0}}\right)\right)^2 \\ &= 1 - \left(1 - Q\left(\sqrt{\frac{2E_b}{N_0}}\right)\right)^2 \\ &\approx 2Q\left(\sqrt{\frac{2E_b}{N_0}}\right) \quad \text{, for small } Q(.). \end{aligned} \tag{3.15}$$

If we Gray-code the QPSK bitstream we can get roughly

$$\mathrm{BER}_{\mathrm{QPSK}} = \frac{1}{2}\mathrm{SER}_{\mathrm{QPSK}} = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) \tag{3.16}$$

which corresponds exactly to the BER of BPSK.

For the evaluation of the SER, we can easily see that the closer the constellation points are, the more likely errors are occurring. Up to a certain extent, errors can be corrected using error-correction codes. On the other hand, using a dense constellation scheme, the information rate is higher. The question is how dense we can make the constellation and therefore increase the information rate before the system breaks down. Naturally, a denser constellation with a constant symbol energy leads to smaller gaps between neighboring symbols, so that errors occur more frequently. Looking at this problem from a slightly different angle reveals that given a certain signal-to-noise ratio, the bit energy decreases if we pack more and more bits into a symbol (by increasing the symbol alphabet or the number of constellation points).

Claude E. Shannon (1916–2001), the famous father of information theory, has proved that given a certain symbol energy to noise density ratio, using an appropriate error correction code, arbitrarily low bit error rates can be achieved if the information rate is below the so-called channel capacity given by

$$C = W \log_2\left(1 + \frac{S}{N}\right), \tag{3.17}$$

where W is the bandwidth, S is the signal power, and N is the average noise power. Eq. (3.17) is also called the Shannon-Hartley theorem. Now when the information rate R is below C, an arbitrarily small error probability results by using a sufficiently complicated coding scheme. For information rates $R > C$, this is not the case. Looking at Eq. (3.17) would lead to the false assumption that no matter how small the symbol power is, the channel capacity might always be increased by just increasing the bandwidth W. However, the noise power N is the product of noise density and bandwidth, hence

$$N = N_0 W. \tag{3.18}$$

For the case in which the information rate is equal to the channel capacity $R = C$ (the limit case), the symbol power can be written as

$$S = E_b C, \tag{3.19}$$

or in words, bit energy E_b times capacity (information rate). Eq. (3.17) can then be written as

$$\frac{C}{W} = \log_2\left(1 + \frac{E_b}{N_0}\frac{C}{W}\right). \tag{3.20}$$

Eq. (3.20) is a central result of information theory.

It states what the maximum information rate per bandwidth is as a function of bit energy over noise power density. Due to the logarithm, Eq. (3.20) cannot be directly solved for the capacity. The curve of the function is, however, given in Fig. 3.24, where also the required E_b/N_0 for many modulation schemes and a symbol error of SER=10^{-5} is shown. The question is now what the minimum E_b/N_0 required is that the capacity is still positive (possibly very small, but finite). To find this value, we write Eq. (3.20) as

$$2^{C/W} = 1 + \frac{E_b}{N_0}\frac{C}{W} \tag{3.21}$$

or

$$\frac{E_b}{N_0} = \frac{2^{C/W} - 1}{\frac{C}{W}}. \tag{3.22}$$

With the rule of Bernoulli/de l'Hôpital for the ratio of terms tending to zero, we get

$$\begin{aligned} \left.\frac{E_b}{N_0}\right|_{\frac{C}{W}\to 0} &= \lim_{\frac{C}{W}\to 0} \frac{2^{C/W} - 1}{\frac{C}{W}} = \ln 2 \cdot \left. 2^{C/W}\right|_{\frac{C}{W}\to 0} \\ &= \ln 2 = 0.6931 = -1.6\,\mathrm{dB}. \end{aligned} \tag{3.23}$$

The bit-energy-to-noise-density ratio must be larger than $-1.6\,$dB in order to result in a positive channel capacity C. Thus, $-1.6\,$dB is the asymptote of the C/W curve.

Spectral Efficiency

A second important criterion besides BER and SER is spectral efficiency. In essence, spectral efficiency expresses how much Eq. (3.20) bandwidth is used per data rate. Bandwidth is a precious commodity, as have numerous auctions over the last few years shown. Spectral efficiency is referred to by two slightly different, albeit not independent effects. First, the term spectral efficiency states how many bits can be transmitted in a given spectral bandwidth. Second, it also defines the roll-off of at the band edges, which overlaps with the next band.

Let us look at the first effect. The spectrum is essentially given by the symbol rate. The higher the symbol rate, the wider the spectrum occupied. By packing more bits into one symbol, we can increase the spectral efficiency, at the price of smaller noise immunity, of course. Without stating the SNR expected of a channel, we can therefore never conclude what the maximal bitrate per bandwidth is.

As to the second effect: Modulation formats that have sharp transitions have typically a wide spectrum. To avoid this, such a signal is usually filtered prior to transmission. For CPM, which are normally filtered prior to modulation (in order to leave them constant envelope), the premodulation filter is the choice to make the modulation bandwidth efficient at the price of increased ISI. Two examples of how different such spectra can look were shown in 3.18(b) and Figs. 3.20.

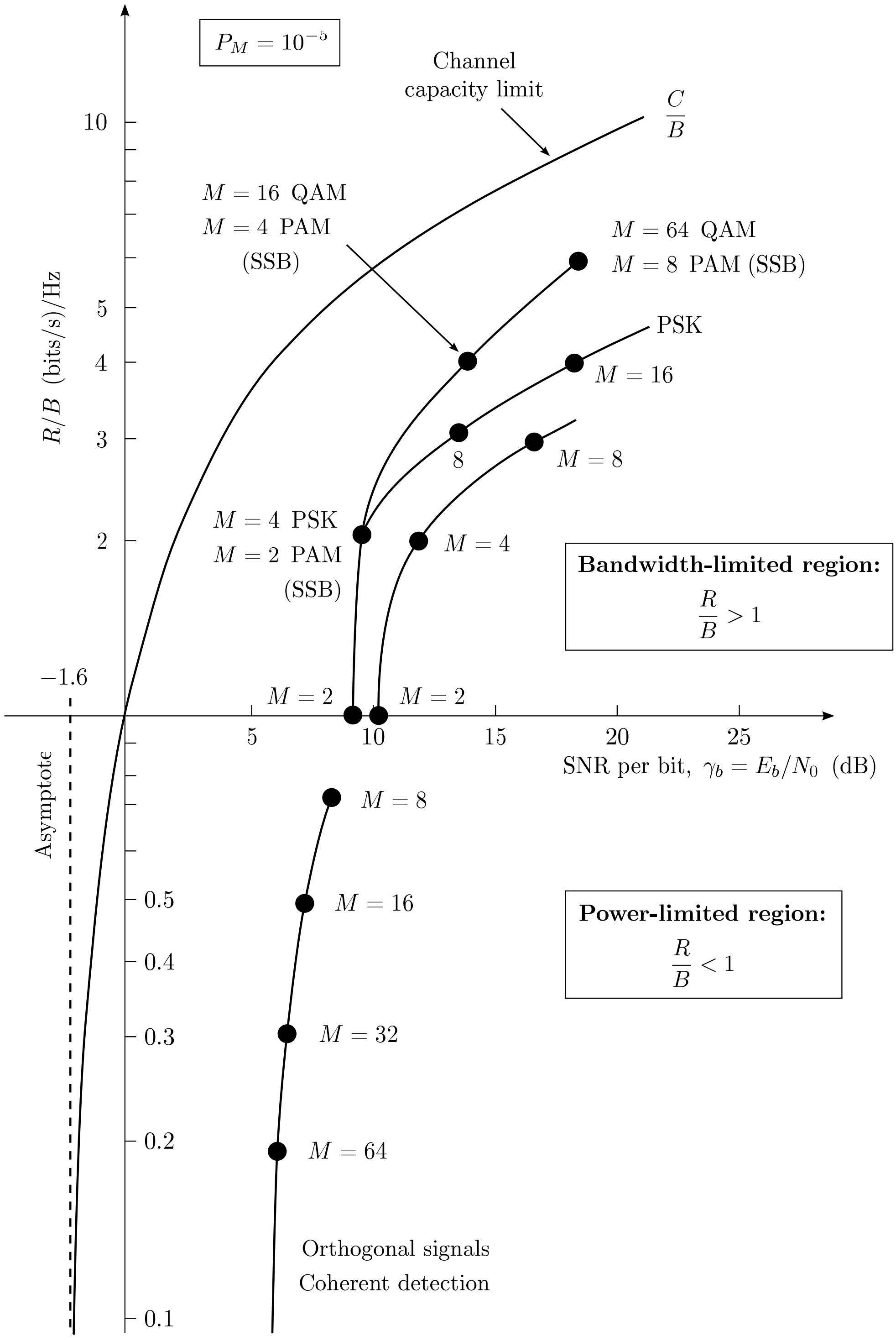


Figure 3.24 Comparison of modulation formats at an SER of 10^{-5} (after [52]).

3.2 Orthogonal Frequency Division Multiplexing (OFDM)

Whereas the usual access to the channel using FDMA requires guard bands between the individual information bands, OFDM makes such a waste of bandwidth unnecessary by employing small bands whose access exploits the property of orthogonality with respect to the subcarriers. This new scheme which is in some way an access scheme as well as a modulation scheme is called *orthogonal frequency division multiplexing* (OFDM). In some applications, this method is also called *discrete multitone modulation* (DMT).

OFDM has become popular both in the wireless world, e.g., IEEE 802.11a/g (WLAN), IEEE 802.16 (WiMAX), DAB, DVB, UWB, LTE, and for 'baseband' transmissions such as xDSL. The division of a band into several subbands has many advantages. If the bandwidth is small enough, the characteristics of each subchannel can be assumed ideal. The subchannel may be approximated as a channel with a flat transfer function across the whole subband, and the noise in all subbands can be modeled as AWGN.

The number of subcarriers of an OFDM system is limited by two considerations. Firstly, the number of subcarriers should be chosen high enough to make the symbol period much longer than the maximum delay of the channel, or, in other words, to make the fading process frequency nonselective (within the subcarrier). The subcarrier bandwidth W/N_c is to be smaller than the coherence bandwidth B_{coh}. Secondly, the subcarrier symbol period N_c/R, where R is the symbol rate of the whole band needs to be short enough in order not to be exposed to the time invariance of the channel, which is characterized by the coherence time T_{coh}. As a consequence of these two limitations, the number of subcarriers N_c should be chosen such as to satisfy

$$\frac{W}{B_{\text{coh}}} \ll N_c \ll WT_{\text{coh}}. \tag{3.24}$$

W is the bandwidth of the whole OFDM-Signal, see also Section 7.10.6.

In OFDM, the data are transmitted on parallel subcarriers using frequency-division multiplexing. The carrier spacing depends on the symbol period and is selected such that each subcarrier is orthogonal to any other subcarrier over a symbol period T_S.

$$\frac{2}{T_S}\int_0^{T_S} \sin(2\pi f_i t)\sin(2\pi f_j t)\,dt = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases} \tag{3.25}$$

In this way, each subcarrier aligns with all the other subcarrier spectral zero-crossing points, as shown in Fig. 3.25. Although the subcarriers spectrally overlap, they do not interfere with each other if the signal is correctly sampled. The basic block diagram of an OFDM system is given in Fig. 3.26.

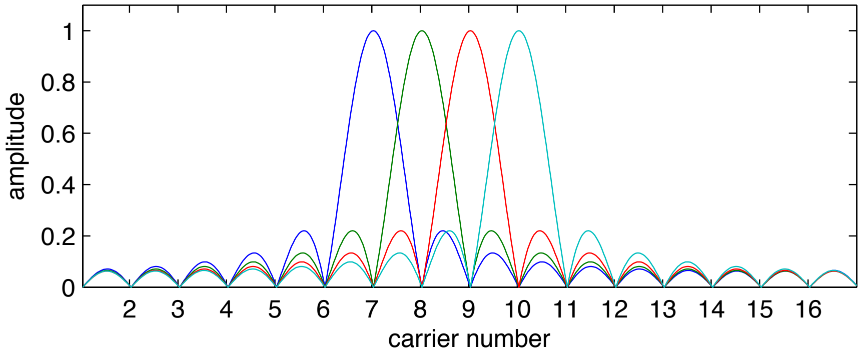


Figure 3.25 Spectral view of the individual carriers in an OFDM system.

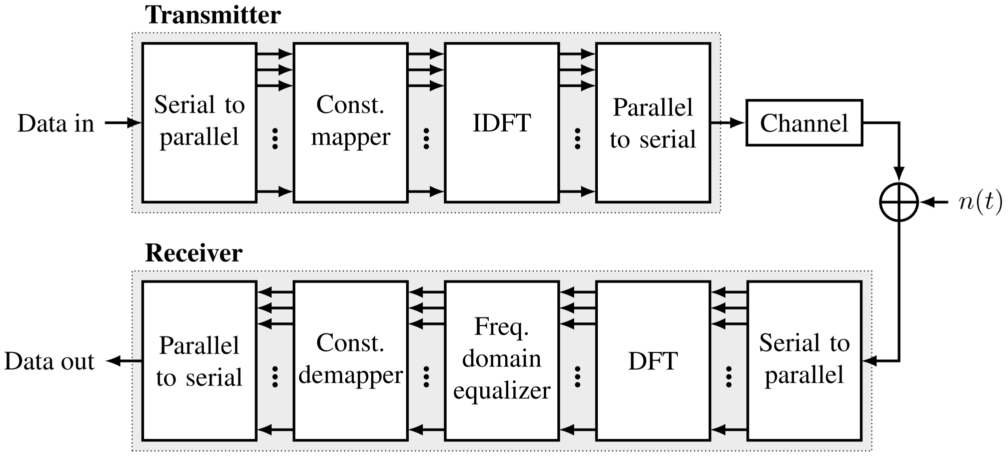


Figure 3.26 Components of an OFDM system.

In order to avoid ICI (interchannel interference) guard intervals are inserted that have the form of a *cyclic prefix*. The length of the cyclic prefix must be larger than or equal to the channel impulse response length (echo duration) in order to maintain the orthogonality between subcarriers, so that no ICI arises.

A further advantage of OFDM technology is the bit-loading of subcarriers according to their individual SNR (adaptive loading), thereby making optimum usage of the spectrum.

3.3 Channel Access

For systems where we only have one single transmitter and one single receiver, channel access is not an issue. When several users have to share the medium, however, we need to think about the best use of the radio channel. The basic configuration of a wireless communication system with several transmitters sharing the same channels can be visualized as shown in Fig. 3.27. Many variants of this picture are possible: several receivers may share the medium, several transmitters and receivers share the medium etc. In order to avoid collisions and use the channel efficiently, we need to define an access protocol. In the wireless-communication world, past government auctions of frequency bands have clearly indicated the need to make optimum use of the bandwidth available.

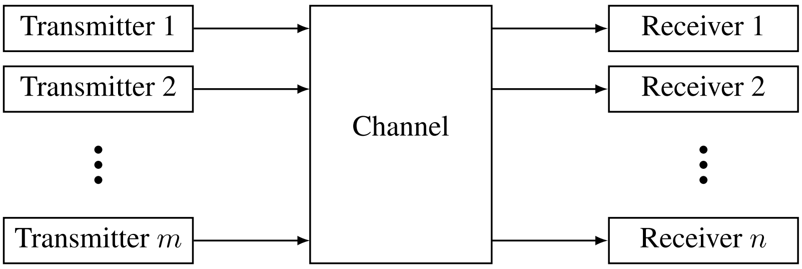
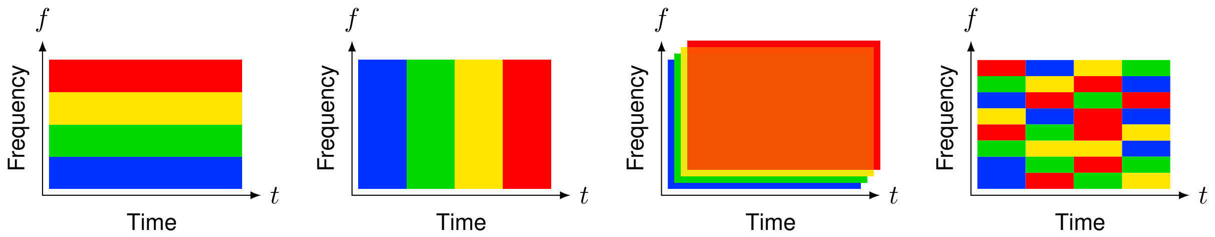


Figure 3.27 Multi-channel (wireless) communication system.

Traditional approaches use TDMA, FDMA, CDMA, and OFDMA, see Figs. 3.28 and 3.29, which shall be explained in due course. All of these multiple access methods provide separate channels to all users. Multiple use of common channels is also possible using *space division multiple access* (SDMA), where the fact is exploited that no two transmitters or receivers have the same physical location. Finally, multiplexing of the channel may be achieved using different polarization schemes (e.g., horizontal and vertical). Such an access scheme is sometimes referred to by the name *polarization division multiple access* (PDMA).



(a) FDMA (b) TDMA (c) CDMA (d) OFDMA

Figure 3.28 Illustrations of the working principles of the four basic channel access methods.

Besides allocating different channels to each user (at least temporarily), the medium can also be shared on a packed-based approach, as gathered in Fig. 3.29 as well. This essentially relies on each user first checking whether the medium is available. If it is, then one or multiple packets are sent; if not, nothing is sent (to avoid a collision) and a new attempt is made at a later time. Based on this principle, multiple techniques exist which try to recover from collisions or avoid them all-together. A common example which uses this principle, rather than a channel-based approach, is wireless LAN (WLAN, IEEE 802.11).

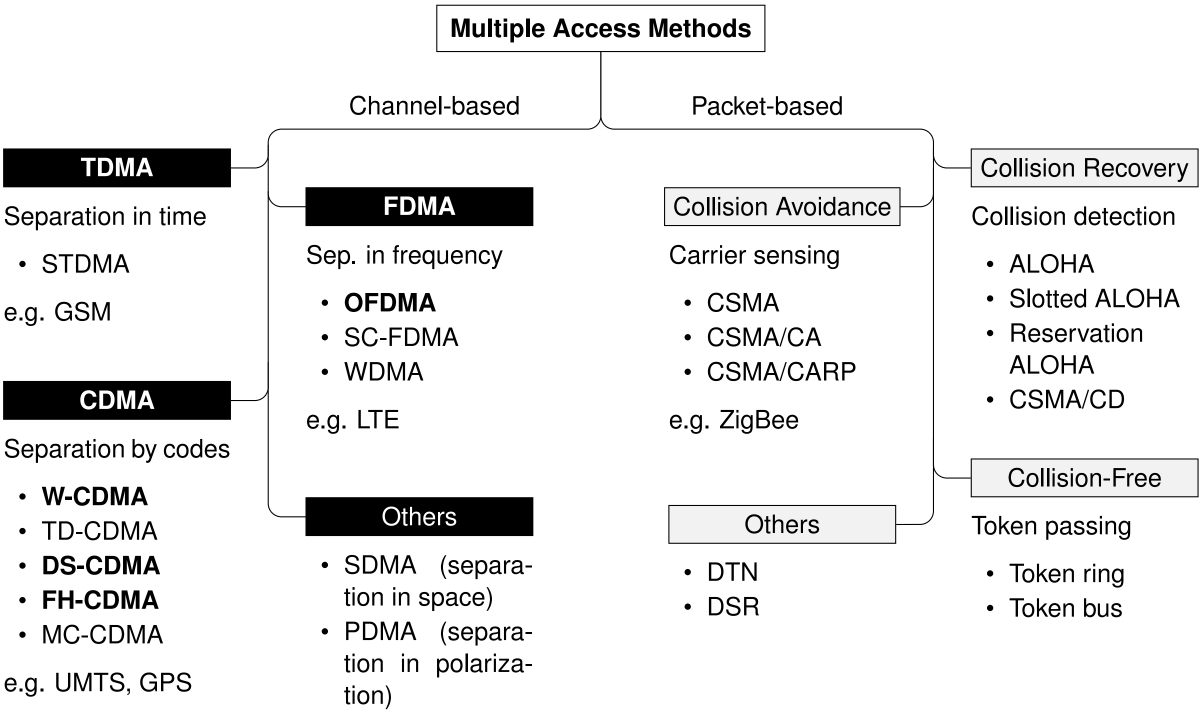


Figure 3.29 Channel- and packet-based multiple access methods.

3.3.1 FDMA

In the first days of telephony, an individual pair of wires was needed for every connection between the different switching centers. In the early 1900s, FDMA (frequency division multiple access) was successfully applied to reduce the number of wires. Many radio systems employ some type of FDMA. This is particularly true for the old analog FM systems. But even systems such as GSM (see Fig. 3.30), which are categorized under TDMA, apply some form of FDMA on top of that in both the uplink and the downlink. This means that some channels differ in their frequencies, some in their assigned time slots, and some in both frequency and time slot.

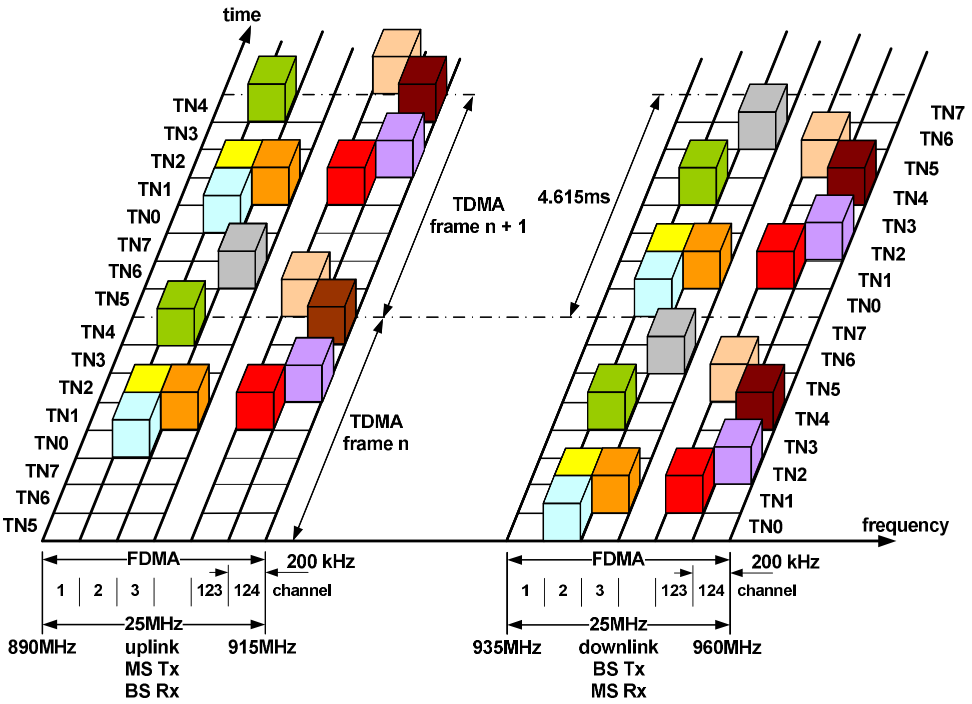


Figure 3.30 TDMA and FDMA as used in GSM.

3.3.2 TDMA

Time division multiple access (TDMA) is what people do (if they are polite) during a conversation. At some periods they talk, at others they listen to other people talking. Only one person is accessing the channel at one given time, so to speak.

ALOHA

In systems such as GSM, the assignment of the time slot to transmit is controlled by a common instance. An example of a system where free access of the time slot is allowed, was invented in Hawaii in 1971 under the name ALOHA[2]. In that scheme, every potential user is allowed to start transmitting at any time. After the transmission of a packet of a constant size, the transmitter waits for the acknowledgment (ACK) of the receiver party. If a negative acknowledgment (NAK) or a timeout occurs (no acknowledgment at all), the

[2] ALOHA is the Hawaiian term for "Hello".

user retransmits the same packet after a random delay. This way, if two users originally transmitted at the same time, thus causing a collision, they will succeed the second time with a very high probability. The optimal channel throughput of the pure ALOHA access scheme can be estimated to be about 18% of the channel capacity, as illustrated in Fig. 3.32.

Better throughput performance can be achieved by employing a common time grid, so that the still random access can only occur at every integer of the frame number. The operation of this so-called *slotted ALOHA* is shown in Fig. 3.31. This extension to the original ALOHA approximately cuts the possibility for collisions in half, thereby doubling the channel throughput to about 37%, as also shown in Fig. 3.32. As can be seen, for very few attempts (far fewer than one per slot), the throughput increases linearly with the number of attempts. However, the performance quickly saturates and then decreases with further increase in attempts (due to the many collisions occuring).

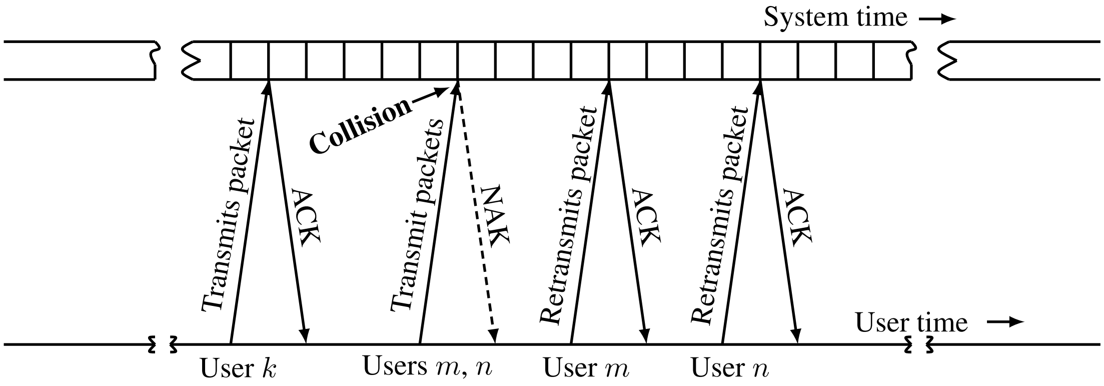


Figure 3.31 Slotted ALOHA operation.

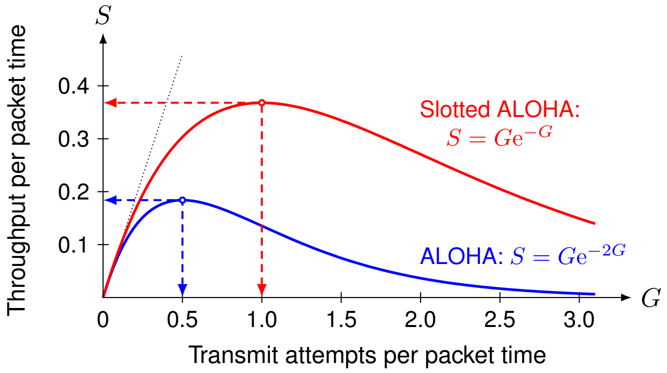
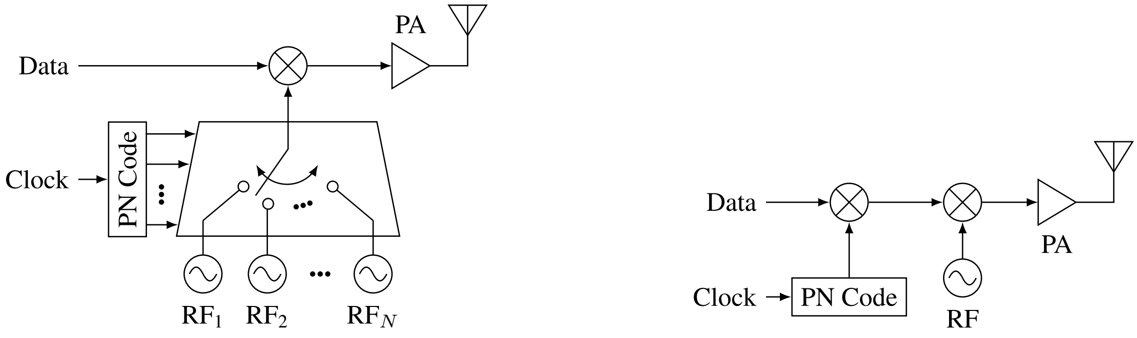


Figure 3.32 Throughput vs. offered traffic for ALOHA systems.

3.3.3 CDMA

CDMA (code division multiple access) is a scheme where all transmitters use the same time and frequency allocation. To distinguish the different users, a code is assigned to each of them. The code is then used to spread a small information bandwidth over the much wider channel bandwidth. CDMA is thus a spread-spectrum (SS) technique. To spread the signal, we can either hop on different frequencies, see Fig. 3.33(a), where we speak of frequency-hopping spread spectrum (FH-SS), or we can directly multiply the data sequence with a much higher chip sequence, see Fig. 3.33(b), which we call direct-sequence spread spectrum (DSSS).



(a) Frequency-hopping spread spectrum (FH-SS)

(b) Direct-sequence spread spectrum (DSSS)

Figure 3.33 Spread spectrum architectures.

There is an ongoing argument whether CDMA systems provide more space to transfer information compared to TDMA and FDMA. Whereas the fundamental quantities that define the channel capacity are independent of the access scheme, there are some differences. If few users use a channel, those who do use the channel get a very good SNR performance in CDMA. In TDMA and FDMA the SNR is independent of the network load. On the other hand, in CDMA, if users are at different distances from a base station, the channel is predominantly assigned to the strong user. The weak-signal user gets a very low SNR. This scenario is called the near-far problem associated with spread spectrum systems. On the other hand, spread-spectrum techniques are suitable for 'blind' assignment of frequency resources to many different users in an uncoordinated way. They do not need assigned time slots and the like. An example is the popular ISM band at 2.4 GHz, where systems like Bluetooth and W-LAN coexist. Further advantages of CDMA are:

- Privacy: Interception is not possible without knowing the code.
- Fading immunity: The resistance against deep fades in the frequency spectrum is much higher.
- Jamming resistance: Narrowband interferer are spread over a wide band at the receiver where only the wanted band will be concentrated back to the signal bandwidth.

In order to collect all energy in a channel with echo (multipath propagation), very often a RAKE receiver as shown in Fig. 3.34 is employed in order to receive a CDMA signal.

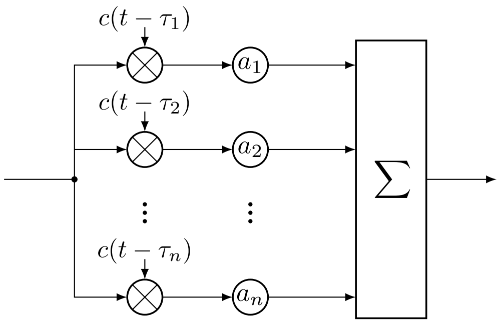


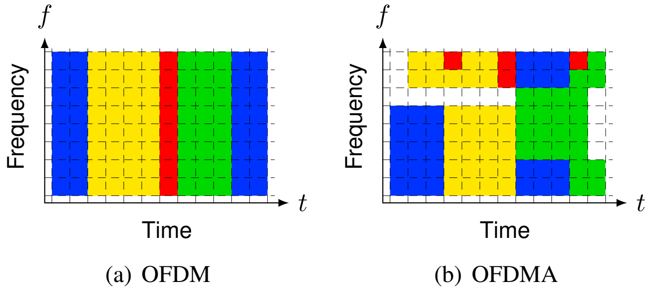
Figure 3.34 RAKE receiver structure.

3.3.4 OFDMA

The most modern multiple-access scheme most likely to be employed in all future generation of mobile communication systems is OFDMA, Orthogonal Frequency Division Multiple Access. The most important example is LTE, the fourth-generation mobile communication standard.

OFDMA sounds a lot like OFDM, and indeed the two are closely related — but not the same: When it comes to providing simultaneous multi-user access, two technologies have a significant difference in the channel allocation mechanism:

- In standard OFDM systems, only a single user can transmit on all of the subcarriers at any given time. In order to support multiple users, time and/or frequency division access techniques (TDMA and FDMA) are used together with OFDM. The major setback to this static multiple access scheme is the fact that the different users see the wireless channel differently is not being utilized. A prominent example of such a system is the 802.11 WLAN family[3], which uses OFDM together with CSMA/CA.
- OFDMA, on the other hand, allows multiple users to transmit simultaneously on the different subcarriers per OFDM symbol; it essentially incorporates the FDMA *and* TDMA capabilities right into OFDM, as illustrated in Fig. 3.35. Thus, OFDMA supports simultaneous low data rate transmission from several users, whereas OFDM can only support one user at given moment. Moreover, the channel can also be dynamically reconfigured into a parallel data high-way for a specific user at any given moment. A further improvement of OFDMA over OFDM is its robustness to fading and interference, since it can assign a specific subset of subcarriers per user and avoid assigning bad channels to a user altogether.



(a) OFDM (b) OFDMA

Figure 3.35 Comparison of OFDM and OFDMA to illustrate their differences.

To summarize: OFDMA is the multi-user OFDM technology, where users can be assigned on both TDMA and FDMA basis dynamically, according to their specific, momentary needs (e.g. data rate, distance and SNR). At any given time, a subset of orthogonal subcarriers is assigned to a particular user. This allows simultaneous low data rate transmission to several users, or combining multiple channels together, to obtain a wide data high-way for a smaller number of users. Since the channels need not be adjacent in spectrum, fading considerations can be incorporated in either scenario, allowing for the greatest flexibility to always use the best non-fading, low-interference channels. Evidently, this flexibility comes at the cost of overall system complexity; OFDMA is clearly the most complex of all classical multiple-access schemes and only the considerable increase in computational power opened the way to apply it to general purpose mobile communication services.

[3] except 802.11ax, to be released in 2019, which is proposed to use OFDMA, 1024-QAM and multi-user MIMO (MU-MIMO) to go beyond 10 Gb/s over the air

3.4 Duplex

One of the key elements of any communication system is the way in which communication is maintained in both directions. Most wireless communication systems are bidirectional, only broadcast systems are unidirectional. It is rather intuitive that it is necessary from a user perspective to be able to talk or exchange data in both directions "‘seemingly simultaneously"’ — however, it is also evident that usually when two people talk at the same time, neither of them understand what the other is saying. Furthermore, different from a regular conversation between people, in data communications, traffic is commonly unbalanced: the downlink generally delivers much larger amounts of data than the uplink; but to achieve a good overall performance, both channels need to be fast and experience low delays.

Hence, some organization and some rules are necessary, to maintain a useful communication link in both directions. Technical terms often used in this context include *simplex, duplex, FDD, and TDD*; we shall briefly discuss what they mean and how they govern some key aspects of the performance of wireless communication systems.

3.4.1 Transmission Control Schemes

There are a variety of different ways of controlling the two-way passage of information. They differ in achievable performance, resource requirements and system complexity. Generally, the simplest systems requiring the least complex circuitry provide more basic performance, whereas more complex systems or systems using more resources are able to provide higher levels of performance. Thus, each scheme has its own advantages and disadvantages.

- **Simplex:** Simplex transmission is the opposite[4] of duplex communication: the exchange of information can only occur in one direction. The typical example of this control scheme is a broadcast system such as radio and TV broadcast.
- **Half duplex:** This is a duplex scheme, where communication is possible in two directions, but only in one direction at a time. If one transmitter is transmitting, the entire communication channel is occupied and the other one must wait until the first stops before transmitting. This form of communication is used for "‘walkie-talkies"’ radios. It is sometimes also counted to simplex communication and should not be confused with TDD.
- **Full duplex:** Full duplex, most commonly referred to simply as duplex, is a scheme whereby transmissions may be sent in both directions "‘seemingly simultaneously"’. However, it is still necessary for the transmissions to be separated in some way, to enable the receivers to receive signals at the same time as transmissions are being made. There are two ways of achieving this: One is to use frequency separation, leading to frequency division duplex, FDD, and the other is to use separation in time, known as time division duplex, TDD. Both of these duplex schemes are covered next.

3.4.2 Frequency Division Duplex (FDD)

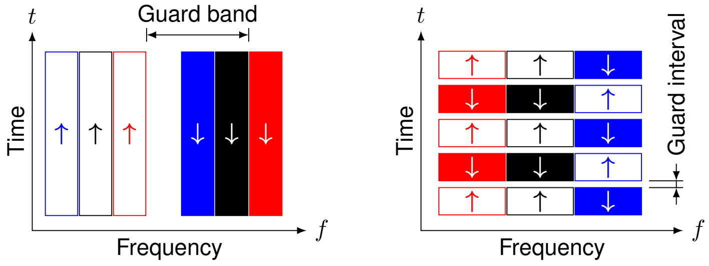
Frequency division duplex, FDD, separates the transmission and reception of signals in the frequency domain: signals are transmitted and received on two different frequencies, as illustrated in Fig. 3.36(a).

[4] The definition of simplex is not always clear; here the definition according to the ANSI (American National Standards Institute) is used. Occasionally simplex may refer to a half duplex scheme as described below.

Using FDD, it is possible to transmit and receive signals simultaneously as the receiver is not tuned to the same frequency as the transmitter as shown. For the FDD scheme to operate satisfactorily, it is necessary that the frequency, i.e. channel separation between the transmission and reception frequencies must be sufficient to enable the receiver not to be unduly affected by the transmitter signal. This is known as the *guard band*, *duplex distance* or also simply as frequency separation.

To avoid *receiver blocking* (i.e. saturating the receiver by strong interferers, most importantly the transmitter signal), often highly selective filters may be required. For cellular systems using FDD, filters are required within the base station and also the handset to ensure sufficient isolation of the transmitter signal without desensitizing the receiver. While cost is not such a significant driver for the base stations, placing a filter into the handsets is more of an issue.

FDD enables true simultaneous transmission and reception of signals. However, essentially two channels are required, which may require spectral resources not always available or at least use the available spectrum inefficiently. Furthermore, it is challenging to incorporate unbalanced traffic into the system design; fourth-generation and later commercial cellular systems are able to accomplish this in real-time, but the increase in system complexity is considerable.



(a) FDD

(b) TDD

Figure 3.36 Illustration of the two full-duplex schemes for an FDMA system.

3.4.3 Time Division Duplex (TDD)

Time Division Duples, TDD, uses only a single frequency and shares the channel between transmission and reception, by separating the two signals in time. Therefore, it is also described as emulating full duplex communication over a half duplex communication link. As the transmission periods are short, the user does not notice any time delays and the communication is still seemingly simultaneous in both directions.

While FDD requires a guard band between the transmitter and receiver frequencies, TDD schemes require a *guard time* or *guard interval* between signal transmission and reception, as shown in Fig. 3.36(b). This delay must be sufficient to allow the signals travelling from the transmitter to the receiver to arrive before a transmission is started at that end and the receiver inhibited. Although this delay is relatively short, when changing between transmission and reception, many times a second, even a small guard time can reduce the efficiency of the system. Especially for systems communicating over short distances, the guard interval is normally small and acceptable. For greater distances it may become an issue. Furthermore, high-performance TDD scheduling requires precise synchronization and makes systems generally more complex than FDD.

On the other side, the spectral requirements are much lower (usually half, since only one channel is needed) and, by choosing the transmission and reception times accordingly, unbalanced traffic can be incorporated well and at moderate increase in complexity, even in real-time. Moreover, for slowly moving mobile systems, the wireless channnel properties of the up- and downlink radio paths are likely to be very similar. Thus, channel estimation is simpler and techniques such as beamforming work well with TDD systems.

3.4.4 Comparison

Both FDD and TDD have their own advantages and disadvantages. Accordingly, they are used for different applications, or in different areas, where the advantages of one over the other can be used to the greatest advantage. Overall, both of the two duplex schemes are both widely used, even though at the moment FDD is somewhat more common than TDD; most likely because of the lower system requirements and the availability of spectral resources. In Table 3.2 the two duplex schemes and their respective advantages and disadvantages are compared.

Attribute	**FDD**	**TDD**
Use of spectrum	Requires separate up- and downlink bands as well as guard bands in-between	Single frequency band for up- and downlink
Unbalanced traffic	Capacity in either direction can only be increased by re-allocating spectrum/channels (cannot easily be adapted except for OFDMA-based systems)	Possibility to adjust the capacity in either direction dynamically by changing the number of slots dedicated to either direction
Distance	No issues with small or large distances	Suited for small distances; guard time increases with distance as signal propagation time increases (approx. 3.3 µs/km)
Latency	No time delays or latency; channels are always "‘open"’	Additional (but often unnoticeable) latency may be added
Equipment costs	Filters (steep, expensive, sometimes bulky) are usually required to prevent receiver blockage.	Typically higher overall system complexity (on the provider’s side), requires better synchronization
Examples	• ADSL & VDSL • Most cellular systems, including GSM, UMTS/WCDMA and LTE • IEEE 802.16 WiMax (fixed)	• TD-LTE (4G/LTE TDD) • TD-UMTS 3G (TD-CDMA, mostly for indoor applications) • TD-SCDMA 3G mob. comm. air interf. • DECT wireless telephony • IEEE 802.16 WiMAX (mobile) • G.fast, a high-speed DSL standard (ITU-T G.9700 and G.9701)

Table 3.2 Comparison of some of the most important characteristics of FDD and TDD schemes.

3.5 Further Literature

Almost any publisher has a book by the title *Digital Communications*. The one from Mc-Graw Hill is the standard book, also referred to as the Proakis [53]; some of its contents and its problems are also treated in a German book by the same author [54]. The one by Prentice-Hall (Bernard Sklar) is very readable [71] and gives a very good overview of the disciplines of digital communications. More constellations for modulation schemes are given in [84].

4 Passive RF Components

4.1 RF Filters

4.1.1 Introduction

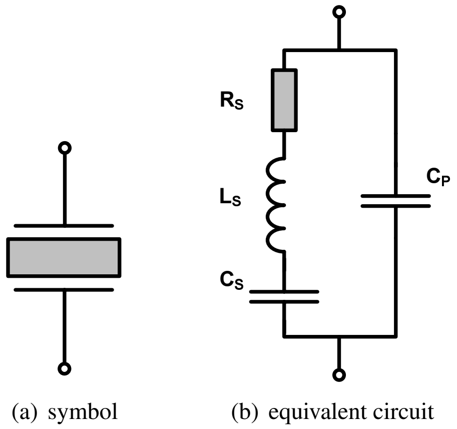
Depending on the frequency range of interest and other parameters of specification, we meet many different types of filter in wireless communications. Of special interest to RF circuits are:

- quartz crystal filters
- ceramic filters
- SAW filters
- transmission line filters

Whereas passive LC filters, active RC filter and many other types are treated elsewhere, in this section we will concentrate on some special types of filter relevant to high frequency circuits.

4.1.2 Quartz Crystal Filters

A quartz crystal is a piezoelectric device. Fig. 4.1 shows the equivalent circuit of a quartz crystal. A quartz crystal has two resonance frequencies: the lower frequency is a series resonance where the impedance is a low resistance. The higher frequency is a parallel resonance where the impedance is a high resistance. Inbetween the two frequencies, the impedance is inductive (phase angle +90 °), below and above both frequencies, the impedance is capacitive (phase angle -90 °). The frequency range a quartz crystal can operate

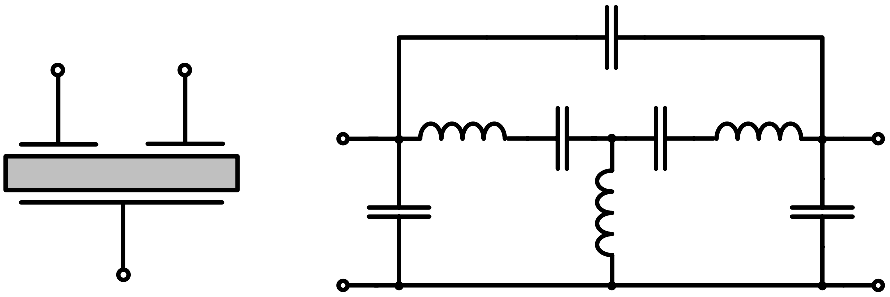


(a) symbol (b) equivalent circuit

Figure 4.1 Quartz crystal symbol and its equivalent circuit.

in is up to 15 MHz in fundamental mode and up to 180 MHz in overtone mode (in that case 7th harmonic). Cheap crystals have a frequency variation of several ppm (parts per million) over aging, temperature, and sample selection. Temperature-compensated crystals include additional components to compensate for the frequency variation over the temperature range, e.g., some thermistor-resistor network driving a varactor diode. The resulting frequency variation may be less than 0.5 ppm over a temperature range of -55 ° to +95 °. An oven-controlled crystal is being operated in a very controlled thermal environment and has therefore a very predictible accuracy. However, size and power consumption prohibit such approaches for wireless communication devices.

A close relative are monolithic crystal filters, which have three electrodes and can be used as a two-port bandpass filter. Bandwidths of below 1 % at frequency ranges between 5 and 350 MHz are possible. Fig. 4.2 shows a monolithic filter equivalent and its possible frequency response.



(a) symbol

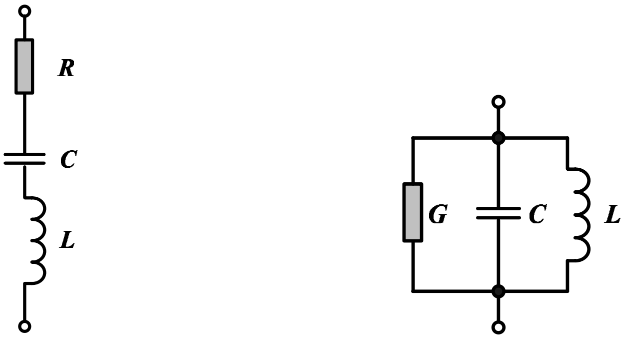
(b) equivalent circuit

Figure 4.2 Monolithic crystal filter.

We have seen that quartz devices are special cases of very high-quality resonators. In order to understand resonators, it is mandatory that we introduce some key concepts and parameters.

Resonance Circuits

Around their resonance frequency, resonators behave either as series resonance circuits, see Fig. 4.3(a), or as parallel resonance circuits, see Fig. 4.3(b). The total impedance of the series resonance circuit can easily



(a) series resonance circuit

(b) parallel resonance circuit

Figure 4.3 Resonance circuits.

be worked out as

$$Z_s = R + j\left(\omega L - \frac{1}{\omega C}\right). \tag{4.1}$$

Similarly, the admittance of the parallel resonance cicruit is

$$Y_p = G + j\left(\omega C - \frac{1}{\omega L}\right). \tag{4.2}$$

The resonance frequency is the frequency for which the reactance of the resonance circuit disappears. For either type we get

$$\boxed{\omega_r = 2\pi f_r = \frac{1}{\sqrt{LC}}.} \tag{4.3}$$

Quality Factor Q

A very important parameter of a resonance circuit is its quality factor Q. In general, any reactive components such as inductors or capacitors have a quality factor. The quality factor is defined as

$$Q = 2\pi\,\frac{\text{energy stored}}{\text{lost energy per cycle}}. \tag{4.4}$$

An inductor stores energy of the value

$$E = \frac{1}{2}LI^2. \tag{4.5}$$

The resistive component dissipates an average of the value

$$P_{\text{avg}} = \frac{1}{2}I^2R. \tag{4.6}$$

Per cycle, this accounts to an energy of

$$E_{\text{loss}} = \frac{1}{2}I^2R \cdot T = \frac{1}{2}\frac{I^2R}{f_r}. \tag{4.7}$$

Hence,

$$Q_{\text{ser}} = 2\pi f_r\frac{L}{R} = \frac{\omega_r L}{R}. \tag{4.8}$$

Using Eq. (4.3), the quality factor of a series resonance circuit can also be expressed as

$$Q_{\text{ser}} = \frac{1}{\omega_r RC}. \tag{4.9}$$

For the parallel circuit, we look at the energy stored in a capacitor

$$E = \frac{1}{2}CU^2. \tag{4.10}$$

Again, the resistive component dissipates on average

$$P_{\text{avg}} = \frac{1}{2}\frac{U^2}{R}. \tag{4.11}$$

Per cycle, this accounts to

$$E_{\text{loss}} = \frac{1}{2}\frac{U^2}{f_r R}. \tag{4.12}$$

Thus,

$$Q_{\text{par}} = 2\pi f_r RC = \omega_r RC. \tag{4.13}$$

With Eq. (4.3) we also have

$$Q_{\text{par}} = \frac{R}{\omega_r L}. \tag{4.14}$$

From these equations we can see, provided our elements L and C have equal resistive losses, that, in order to achieve high Q we should choose

- high inductance and low capacitance for series resonance circuits,
- high capacitance and low inductance for parallel resonance circuits.

Loaded Q

So far we have discussed unloaded Q only, i.e., Q depends on the resonance circuit only. In reality, we always attach some load to our circuit. This load adds to the resistance already present in the resonance circuit. Let us illustrate the computation of the loaded Q for a parallel resonance circuit. If we recall the definition of the unloaded Q of such a circuit

$$Q_0 = \frac{B}{G}, \tag{4.15}$$

where B and G are the susceptance (inverse of reactance) and the conductance (inverse of resistance) of the circuit, respectively, We can define a similar quality factor for the external load (the susceptance remains the same)

$$Q_{\text{ext}} = \frac{B}{G_{\text{load}}}. \tag{4.16}$$

Since both conductances are in parallel, we can write for the loaded Q

$$Q_{\text{loaded}} = \frac{B}{G + G_{\text{load}}}. \tag{4.17}$$

As a consequence, we have

$$\frac{1}{Q_{\text{loaded}}} = \frac{1}{Q_0} + \frac{1}{Q_{\text{ext}}}. \tag{4.18}$$

For every load we get clearly

$$Q_{\text{loaded}} < Q_0, \tag{4.19}$$

i.e., the load can only decrease the loaded Q.

Series and Parallel Conversion

The notion of Q can be very conveniently used to transform a series circuit into a parallel one and vice versa. Setting the definition of Q for the two circuits with respect to each other, we get

$$R_p = R_s\left(1 + Q^2\right), \tag{4.20}$$

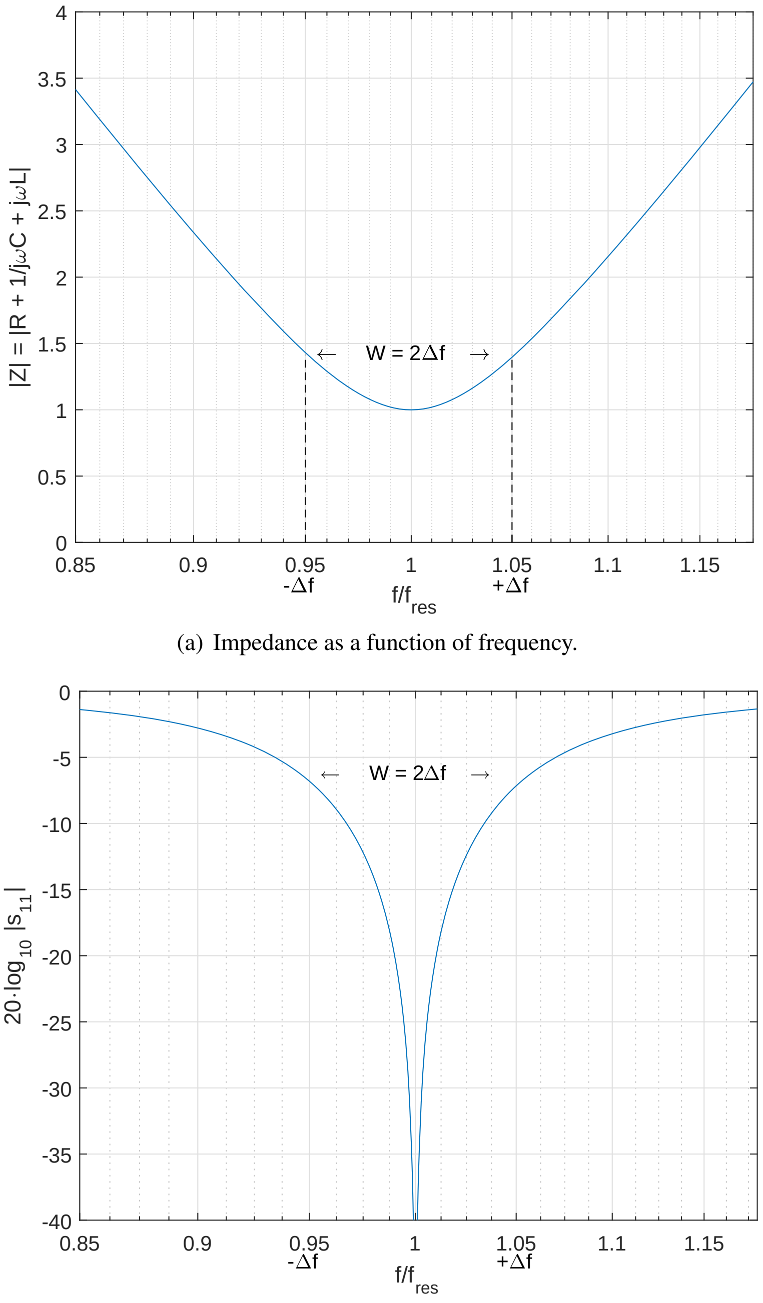
$$L_p = L_s\left(1 + \frac{1}{Q^2}\right), \tag{4.21}$$

$$C_p = C_s\left(\frac{Q^2}{1 + Q^2}\right). \tag{4.22}$$

Relative Bandwidth

The relative resonator bandwidth is the inverse of the quality factor Q. This can be seen if we consider the two points to the left and to the right of the resonance frequency $f \pm \Delta f$ where the reactance part

$$X = j\omega L - \frac{j}{\omega C} = j \cdot \frac{\omega^2 LC - 1}{\omega C} = j \cdot \frac{(2\pi)^2 (f \pm \Delta f)^2 LC - 1}{2\pi (f + \Delta f) C} \approx \pm j \cdot 2 \cdot 2\pi \Delta f \cdot L \tag{4.23}$$



(a) Impedance as a function of frequency.

(b) Return loss as a function of frequency.

Figure 4.4 Relative bandwidth measured with a network analyzer.

no longer cancels and is equal to the resistive part, which itself is known from Eq. (4.8) as

$$R = \frac{2\pi f L}{Q}. \tag{4.24}$$

Hence, by setting $R = |X|$, we get

$$\frac{2\pi f L}{Q} = 2 \cdot 2\pi \Delta f \cdot L, \tag{4.25}$$

and finally the relative bandwidth as

$$W = \frac{2\Delta f}{f} = \frac{1}{Q}. \tag{4.26}$$

Fig. 4.4(a) shows the determination of the relative bandwidth using an impedance scaling of $\sqrt{2}$ or 3 dB. For the given example the relative bandwidth is 0.1 resulting in a Q of only 10. Often resonating circuits are measured with respect to their return loss (reflective measurement). Thus, it is interesting to note the points on the return-loss curve on a VNA measurement where the relative bandwidth can be read. For the following transformation of the impedance into S-parameters we assume[1] that $R = Z_0$, resulting of course in $s_{11} = 0$ at resonance. The general frequency-dependent expression is

$$s_{11}(\omega) = \frac{R + \frac{1}{j\omega C} + j\omega L - Z_0}{R + \frac{1}{j\omega C} + j\omega L + Z_0} = \frac{\frac{1}{j\omega C} + j\omega L}{2R + \frac{1}{j\omega C} + j\omega L} = \frac{1 - \omega^2 LC}{1 - \omega^2 LC + 2j\omega RC}. \tag{4.27}$$

Using

$$\omega = \omega_{\text{res}} + \Delta\omega = \frac{1}{\sqrt{LC}} + \frac{R}{2L}, \tag{4.28}$$

the return loss can be written as

$$\text{RL}(\omega) = 10 \cdot \log_{10} |s_{11}^2(\omega)| = -10 \cdot \log_{10} \left(1 + \left(\frac{2(\frac{1}{\sqrt{LC}} + \frac{R}{2L})RC}{1 - (\frac{1}{\sqrt{LC}} + \frac{R}{2L})^2 LC} \right)^2 \right). \tag{4.29}$$

For cases of high Q, clearly

$$\frac{1}{\sqrt{LC}} \gg \frac{R}{2L}. \tag{4.30}$$

Hence, by neglecting $\frac{R}{2L}$ in the numerator and its squared contribution in the denominator, we can write

$$\text{RL}(\omega_{\text{res}} + \Delta\omega) \approx -10 \cdot \log_{10} \left(1 + \left(\frac{2(\frac{1}{\sqrt{LC}})RC}{1 - (\frac{1}{LC} + \frac{R}{L} \cdot \frac{1}{\sqrt{LC}})LC} \right)^2 \right). \tag{4.31}$$

Simplifying above expression, we get

$$\text{RL}(\omega_{\text{res}} + \Delta\omega) = -10 \cdot \log_{10} (1 + 4) = -7\,\text{dB}. \tag{4.32}$$

4.1.3 Ceramic Filters

Another piezoelectric device is a ceramic filter. Ceramic filters operate at different frequencies ranging from a few kHz up to more than 1 GHz with bandwidths between 0.05 % and 20 %. For their circuit equivalent we can use the same as for quartz crystals.

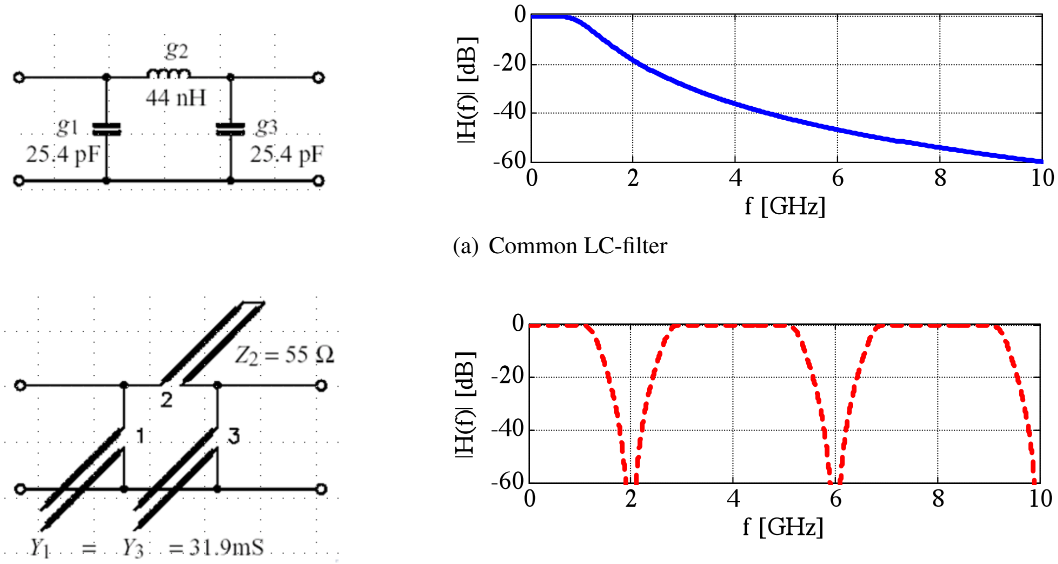
[1] Doing so for a series-resonance circuit will result in a small series resistance any matching circuit should transform the VNA's impedance to. Similarly, in the case of a parallel-resonance circuit the a high-impedance value need be transformed into the VNA's internal impedance. The following derivation applies to either case.

4.1.4 SAW Filters

Surface acoustic wave (SAW) filters show very high Q factors, but are expensive at the same time. They operate by translating the electric signal into an acoustic wave which, in turn, is guided over a mechanical finger setup and experiences selectivity.

4.1.5 Transmission Line Filters

Since lumped-element filters are troublesome for higher frequencies due to their parasitics as shown at the beginning of Chapter 2, distributed-element filters have their advantage at certain frequencies. Additionally, the wavelengths get smaller and thus more practical at higher frequencies. The main difference to lumped-element filters is the periodic nature of the frequency response of transmission-line filters, as can be seen in Fig. 4.5. For the example provided, this means that a low-pass filter has more than one passband.



(a) Common LC-filter

(b) Transmission line filter

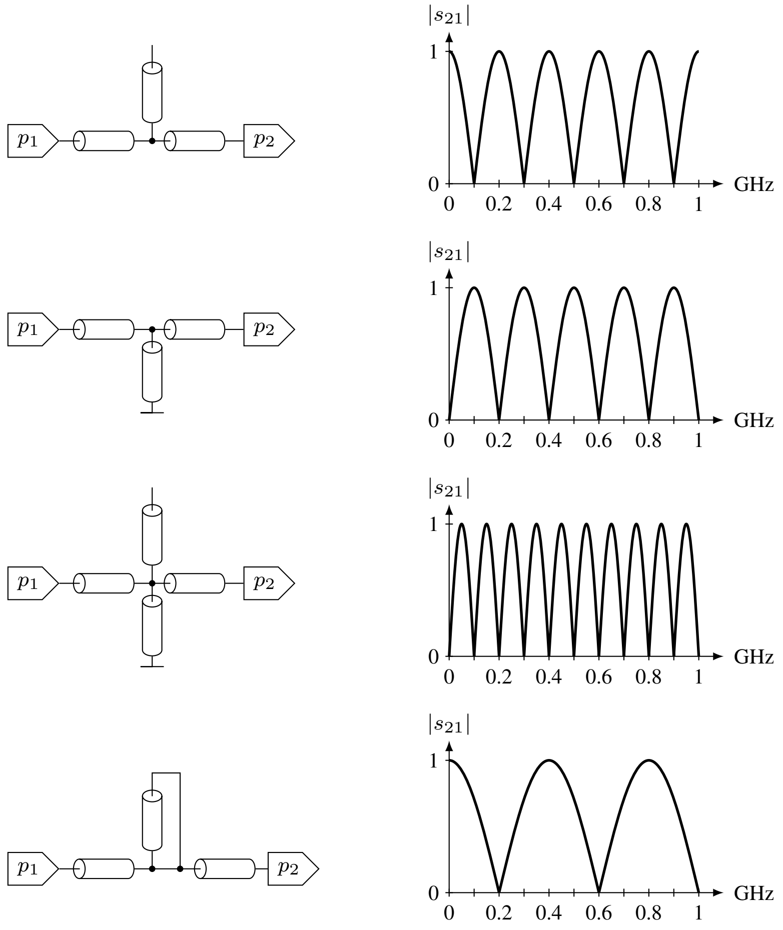
Figure 4.5 Filter periodicity.

Some other filter types that can be realized using transmission lines are illustrated in Fig. 4.6.

In RF design, conductors with certain dimensions with respect to the wavelength are of special interest. The most prominent lengths l for open and short-circuited lines are given in Table 4.1. The given impedance values repeat with extensions of $n \cdot \lambda/2$. The wave impedance Z_w is determined by the width of the microstrip line and by the choice of the material. Generally, its value is taken from diagrams or computed using approximate equations. The propagation velocity v of electromagnetic waves in a homogenous material can be calculated as

$$v = \frac{c}{\sqrt{\varepsilon_r \mu_r}}, \tag{4.33}$$

where ε_r is the dielectric permittivity of the dielectric, μ_r is the relative permeability, and c is the speed of light in vacuum.



All lines are of $\lambda/4$ length at $f = 100\,\text{MHz}$.

Figure 4.6 Some transmission line filters and their transfer functions.

	Z_{open}	Z_{short}
$l < \lambda/4$	capacitive	inductive
$l = \lambda/4$	0 (series resonance)	∞ (parallel resonance)
$\lambda/4 < l < \lambda/2$	inductive	capacitive
$l = \lambda/2$	∞ (parallel resonance)	0 (series resonance)
in general	$Z = -jZ_w \cdot \cot\left(\frac{2\pi l}{\lambda}\right)$	$Z = jZ_w \cdot \tan\left(\frac{2\pi l}{\lambda}\right)$

Table 4.1 Impedance of open and short transmission lines.

4.1.6 Microstrip Filters

Microstrip filters are filters using the copper conductor on a PCB board. For microstrip lines, Eq. (4.33) has to be adapted, since the insulator is partly given by the PCB material and partly by the air above it. For narrow lines, edge capacitances may no longer be neglected leading to a different effective permittivity and hence different propagation velocity. Thus, the ratio of the wavelength in space λ_0 and the wavelength in the microstrip line λ_m, determined by the according use of a diagram, is used to compute the effective propagation velocity

$$v_m = \frac{c}{\lambda_0/\lambda_m} \ . \tag{4.34}$$

Sometimes, the effective permittivity

$$\varepsilon_{\text{eff}} = \left(\frac{\lambda_0}{\lambda_m}\right)^2 \tag{4.35}$$

is taken to describe this effect. At the end of each line, there are also capacitive effects that lead to an extension of the effective line length. To compensate for this effect, the physical length of such a line has to be reduced in the design. The reduction value can be inferred from diagrams, too. Normally, an iterative simulation with optimized design will lead to the best result.

Richards Transformation

The input impedances of transmission lines show a periodic behavior. This can be seen by employing Eq. (2.91) to a quarterwave stub line ($Z_2 = \infty$) of length $l = \frac{v_m}{4f_0}$, observed at a frequency of $f = \frac{v_m}{\lambda}$

$$Z = \frac{Z_0}{j\tan\beta l} = \frac{Z_0}{j\tan\frac{2\pi}{\lambda}l} = \frac{Z_0}{j\tan\frac{\pi}{2}\frac{f}{f_0}}. \tag{4.36}$$

A design exploiting line characteristics for filtering purposes needs to consider this circumstance, since any frequency response becomes periodic, too. A lowpass, e.g., gets additional zeros at its $\lambda/4$ frequency and odd multiples thereof. They cause an increased steepness for lowpass filters near the $\lambda/4$ frequency. Classical filter theory has produced standard design procedures for LC filters. Tables with normalized element values g_i can easily be found in books. For transmission-line filters, a transformation of the corner frequency using the Richards transformation becomes necessary. The Richards transformation

$$\Omega = \tan\left(\frac{\pi f}{2f_0}\right) \tag{4.37}$$

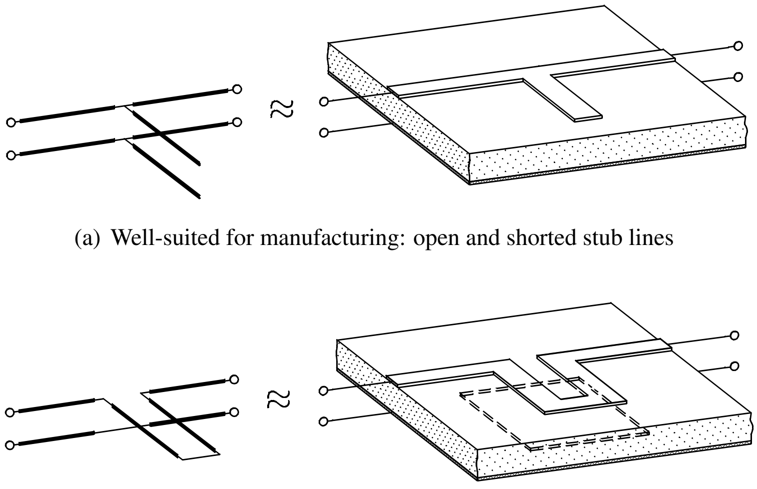
carries out a 'prewarping' of the frequency axis. The parameter f is the frequency for the specification of the implemented transmission-line filter. Ω is the corresponding normalized frequency of the prototype-filter specification in a lumped-element design. The resonance frequency f_0 is determined by the wavelength λ_m and the propagation velocity v_m of the transmission line

$$f_0 = \frac{v_m}{\lambda_m}. \tag{4.38}$$

Kuroda Transformation

As we will see in the following design example, some configurations of line elements are not practical for implementation. An example is shown in Fig. 4.7. Kuroda therefore suggests to transform (using transmission lines) those configurations into more practical ones. In a way we use transmission lines to transform unpractical transmission lines into more practical transmission lines.

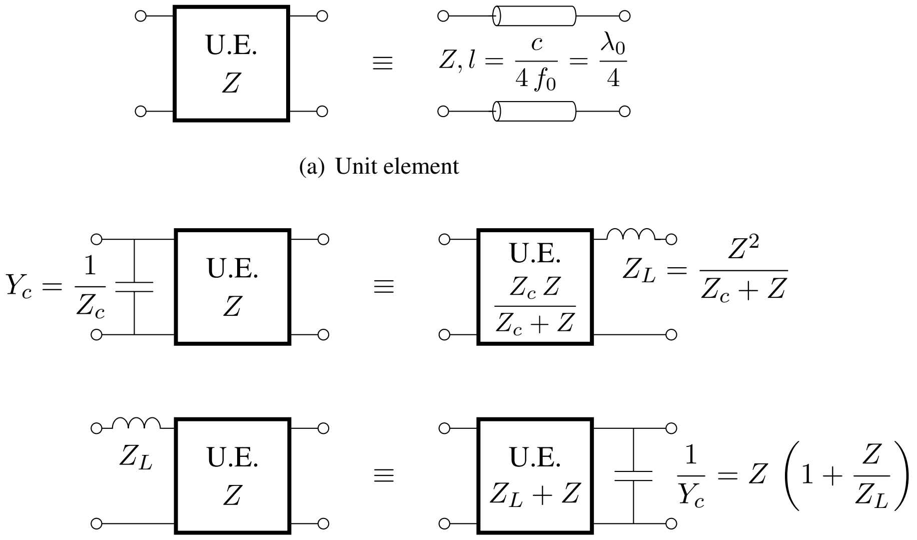
The Kuroda transformation describes the change an impedance experiences if it is connected to a line of length $\lambda/4$. Such a line is called a *unit element* (UE). A unit element is an important design element. Without its existence, all elements would have to be connected at a single point. Furthermore, a unit element allows the transformation of short-circuited, ground-free lines (series inductances) into open, grounded lines (parallel capacitances). This may ease the realization of a circuit considerably. A unit element can perform the transformation according to Fig. 4.8.



(a) Well-suited for manufacturing: open and shorted stub lines

(b) Difficult to manifacture: ground-less lines

Figure 4.7 Manufacturing problems of some microstrip filters.


U.E.
Z
$\equiv$
$Z, l = \frac{c}{4f_0} = \frac{\lambda_0}{4}$
$Y_c = \frac{1}{Z_c}$
U.E.
Z
$\equiv$
U.E.
$\frac{Z_c Z}{Z_c + Z}$
$Z_L = \frac{Z^2}{Z_c + Z}$
Z_L
U.E.
Z
$\equiv$
U.E.
$Z_L + Z$
$\frac{1}{Y_c} = Z\left(1 + \frac{Z}{Z_L}\right)$

(a) Unit element

(b) Kuroda identities

Figure 4.8 Kuroda transformations (after [3]).

Design example: Transmission-line lowpass

Since for the standard lowpass the corner frequency is normalized to $\Omega_c = 1$, the equivalent line elements g_i have to be divided by

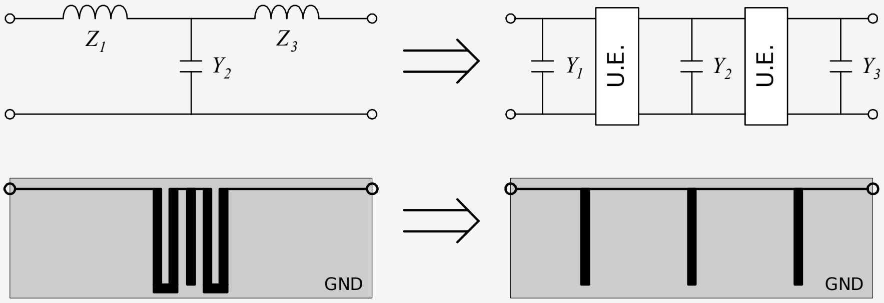
$$\Omega'_c = \tan\frac{\pi f_c}{2f_0}\,.$$

This gives the impedance Z_i of the short-circuited transmission line and the admittance of an open transmission line as

$$Z_i = \frac{g_i}{\Omega'_c} \qquad \text{and} \qquad Y_i = \frac{g_i}{\Omega'_c}\,,$$

respectively. The resulting transmission line lowpass filter can still not be realized this way.

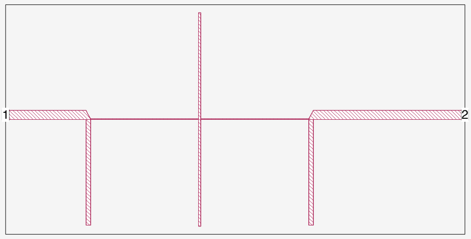
It has to be transformed using the Kuroda transformation into a form which only consists of parallel capacitors and unit elements:



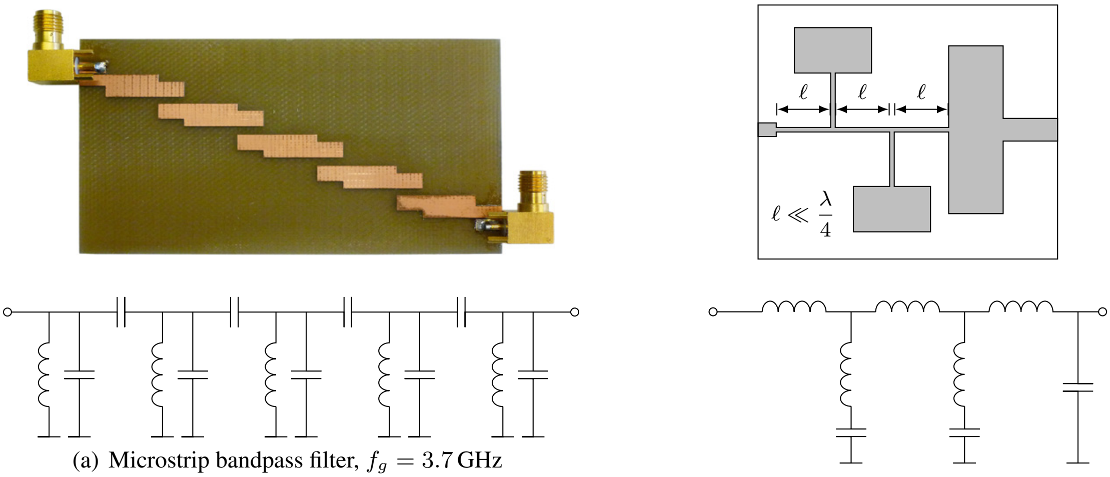
The original circuit can be thought of with an additional unit element at its input and output with the normalized impedance $Z = 1$. The system impedance is usually $Z_0 = 50\Omega$. From that, the denormalization of the line impedances can be computed as

$$Z_{ir} = \frac{Z_0}{Y_i} \qquad \text{and} \qquad Z_{ir} = Z_i \cdot Z_0,$$

respectively. The layout of the final PCB design would look like this:



Some microstrip filters and their equivalent LC circuits are shown in Fig. 4.9.



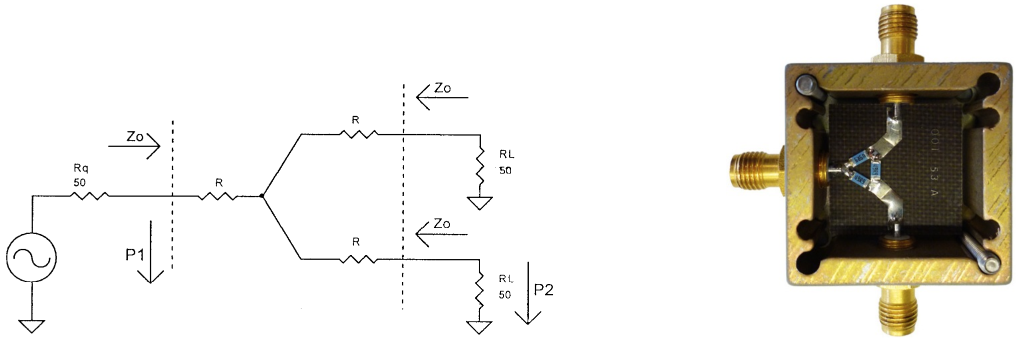
(a) Microstrip bandpass filter, $f_g = 3.7\,\text{GHz}$

(b) Microstrip low pass filter

Figure 4.9 Some microstrip filters and their lumped element circuits.

4.2 Power Splitters and Combiners

Whereas in audio circuits a signal combiner can be as simple as a connection of nodes (e.g., current adder in front of an opamp), in RF circuits this does not work. Since we work with power signals, we cannot simply add voltages or currents, we must always maintain proper impedance match. A power splitter or divider can in principle be built using resistors only, see Fig. 4.10.

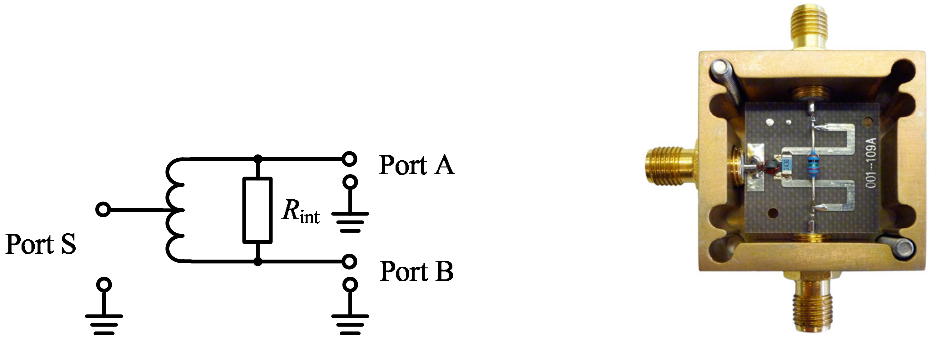


(a) Schematic (star network form)

(b) Broadband resistor splitter/combiner (△ network) (Mini-Circuits ZFRSC-42, Freq. 0 – 4.2 GHz)

Figure 4.10 Power splitter/combiner using resistors.

However, half of the power will be used to heat up the internal resistors. Hence, a 6 dB attenuation will be experienced by each output port. If the whole power is to be split between the two output ports, a structure as shown in Fig. 4.11 is applied. For the perfect matched case, no current flows through the internal resistor. Thus, only a 3 dB attenuation is noticed at the output ports. Instead of a discrete-type transformer, transmission lines can also be used. One such possibility leads to the Wilkinson divider, which is displayed in Fig. 4.12. The scattering matrix of a Wilkinson divider at mid-band (where the quarterwave transmission lines are of the right length) is

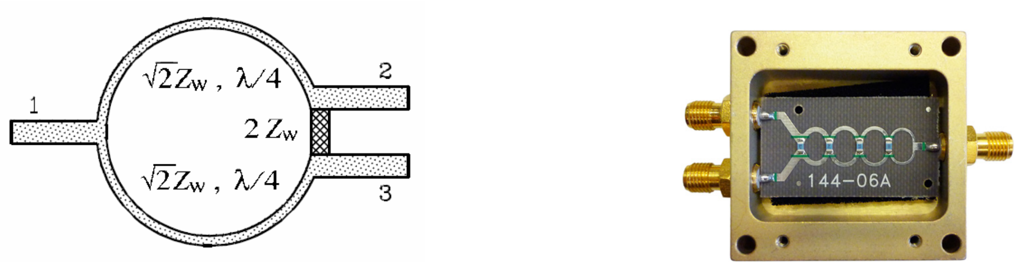


(a) Schematic

(b) Broadband splitter using transformer/Wilkinson combination (Mini-Circuits ZFSC-2-2500, Freq. 10 MHz – 2.5 GHz)

Figure 4.11 Lossless splitter/combiner divider.

$$\boldsymbol{S} = \frac{-j}{\sqrt{2}} \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}. \tag{4.39}$$



(a) Schematic, Source: [3].

(b) Broadband multi-stage Wilkinson splitter/combiner (Mini-Circuits ZN2PD-9G, Freq. 1.7 – 9 GHz)

Figure 4.12 Wilkinson divider.

A power splitter or divider can also be used as a power combiner. The previous input is now used as the output, the former outputs are now the inputs to be combined. The internal resistor uses up half the power even for a 'lossless' device as shown in Fig. 4.11. In that sense, 'lossless' is true only for splitting powers, not for combining, except for the case where the signals in port 1 and port 2 are completely correlated. Ideally, the device shows perfect isolation between port A and port B. The isolation deteriorates if the matching at the ports is no longer ideal. A mismatch at port S, for example, produces a reflected wave to reach a signal coming from port A to reach port B via port S.

4.3 Directional Couplers

A directional coupler is an element which can measure direction-dependent power transmission. This is particularly useful if we have to tell the incident from the reflected wave, e.g., to measure the result of an impedance matching activity. To do this, we have to couple off a certain amount of power. This part is usually small in order not to disturb the main power flow too much. An often-met directional coupler is a planar-line coupler, the layout of which is shown in Fig. 4.13. The gap between the two transmission lines of one quarter wavelength determines the coupling factor k. It is dependent on the even-mode and on the odd-mode line impedance

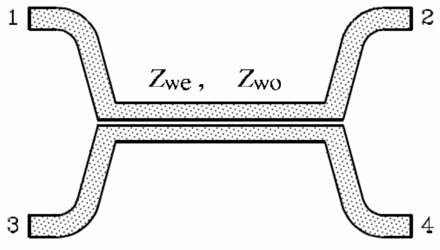


Figure 4.13 Directional coupler using coupled microstrip lines. Source: [3].

$$k = \frac{Z_{we} - Z_{wo}}{Z_{we} + Z_{wo}}. \tag{4.40}$$

Hence, if the k^2th part is coupled off, the remaining power on the main path is

$$\kappa = \sqrt{1 - k^2}. \tag{4.41}$$

The scattering matrix at mid-band can now be written as

$$\boldsymbol{S}=\begin{bmatrix}0 & -j\kappa & k & 0\\ -j\kappa & 0 & 0 & k\\ k & 0 & 0 & -j\kappa\\ 0 & k & -j\kappa & 0\end{bmatrix}. \tag{4.42}$$

In reality, there are no zeros in the matrix due to parasitic effects and asymmetries. The most important parameters of a real directional coupler are:

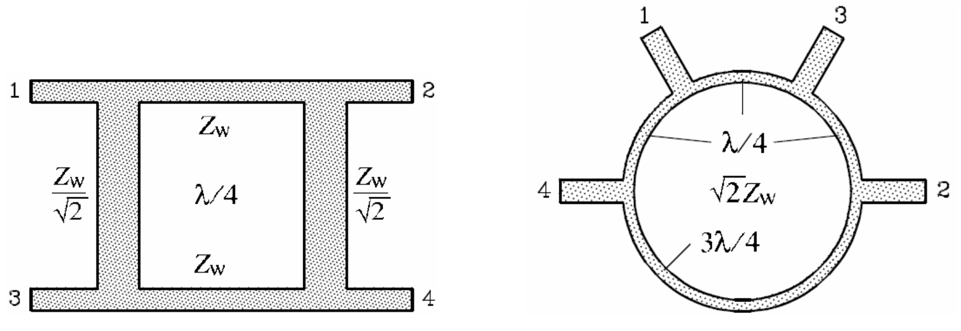
return loss	$\mathrm{RL}=-20\log_{10}\lvert s_{11}\rvert\,,$
through loss	$T=-20\log_{10}\lvert s_{21}\rvert\,,$
coupling	$C=-20\log_{10}\lvert s_{31}\rvert\,,$
isolation	$I=-20\log_{10}\lvert s_{41}\rvert\,,$
directivity	$D=I-C.$

If a considerable part (such as 3 dB) is to be coupled off, other structures are available. One possible structure is the branchline coupler, see Fig. 4.14(a). Its scattering matrix at mid-band is

$$\boldsymbol{S}=\frac{-1}{\sqrt{2}}\begin{bmatrix}0 & 0 & j & 1\\ 0 & 0 & 1 & j\\ j & 1 & 0 & 0\\ 1 & j & 0 & 0\end{bmatrix}. \tag{4.43}$$

Yet another example of a coupler is the so-called ratrace, which is displayed in Fig. 4.14(b). Its scattering matrix at mid-band is

$$\boldsymbol{S}=\frac{-j}{\sqrt{2}}\begin{bmatrix}0 & 0 & 1 & 1\\ 0 & 0 & 1 & -1\\ 1 & 1 & 0 & 0\\ 1 & -1 & 0 & 0\end{bmatrix}. \tag{4.44}$$



(a) Branch-line coupler. Source: [3].

(b) Rat-race divider. Source: [3].

Figure 4.14 Other coupler forms.

4.4 Circulators

Taking the directional coupler one step further, we get a circulator. A circulator, see Fig. 4.15(a) and (b), is an element with at least three ports, each of which transfers its power only to "‘the next port"’ but not to any other (i.e. "‘not to the previous one"’).

A perfect circulator (lossless, $s_{21} = s_{32} = s_{13} = 1$) can be described by the scattering matrix

$$\boldsymbol{S} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} . \tag{4.45}$$

Circulators are non-reciprocal devices which usually consist of a permanent magnet and ferrite material.

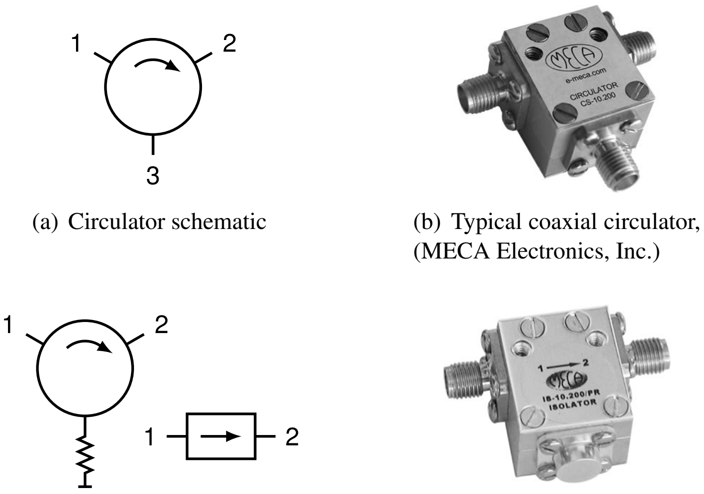
4.5 Isolators

Directly related to circulators are isolators: By adequately matching one of the ports of a three-port circulator, we can build a two-port device which isolates one port from the other in one direction, while connecting them in the other. Often, isolators look just like circulators, where one port has been removed (depending on the power that needs to be handled, heat sinks are also commonly found on that termination), see Fig. 4.15(c) and (d).

The scatter matrix of an optimal isolator is

$$\boldsymbol{S} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} . \tag{4.46}$$

Thus, isolators constitute the simplest form of non-reciprocal devices. They are commonly found in devices with a substantial RF/microwave power output, in order to protect the last stages from strong reflections when the output is not well matched to a load.



(a) Circulator schematic

(b) Typical coaxial circulator, (MECA Electronics, Inc.)

(c) Isolator schematics

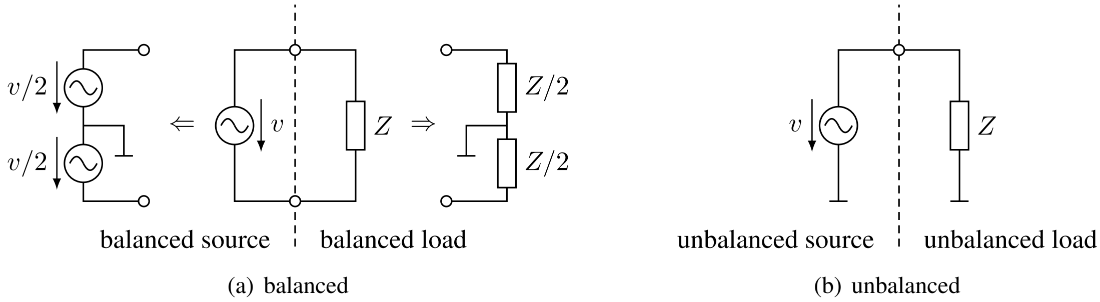
(d) Typical coaxial isolator, (MECA Electronics, Inc.)

Figure 4.15 Comparison of circulators and isolators.

4.6 Baluns

4.6.1 Balanced vs. Unbalanced

A balanced circuit has its midpoint as the ground potential. Opposed to this, an unbalanced circuit has one side grounded. An unbalanced line is a line whose ground is at zero potential relative to distant objects.



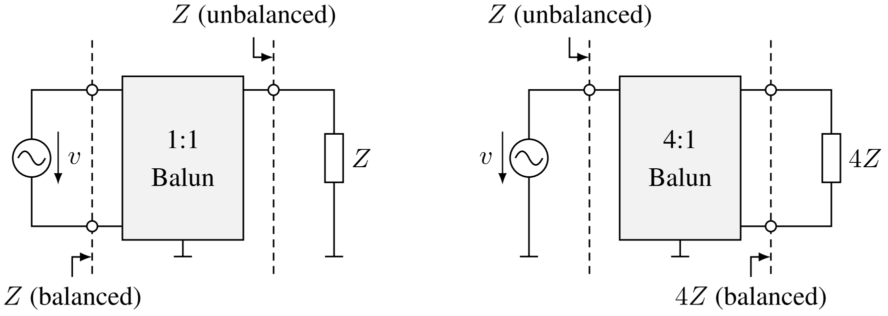
(a) balanced

(b) unbalanced

Figure 4.16 Balanced and unbalanced elements, i.e., sources and loads.

A coaxial cable is the best example of an unbalanced line; another example of an unbalanced line is a single microstrip line. Examples of balanced lines are twisted pair cables and coplanar strips (consisting of two strips). They are characterized by asymmetric potential relative to distant objects. Balanced circuits or parts of circuits are used when extra spurious noise suppression is needed. Any unwanted interferer influences both signals in a balanced design, so that such a common-mode noise is easily cancelled, since the interference only changes the potential on both lines, but not the difference. Hence, the biggest advantage of a balanced design is its *common-mode rejection* capability, whereas its biggest disadvantage is its increased complexity, because of the use of multiple lines and cables.

When changing from balanced to unbalanced elements and back, a balun (abbreviation for **bal**anced-to-**un**balanced transformer) is needed, as shown in Fig. 4.17(a). Furthermore, as depicted in Fig. 4.17(b), in addition to the balanced-unbalanced conversion, they can often also carry out impedance transformations.



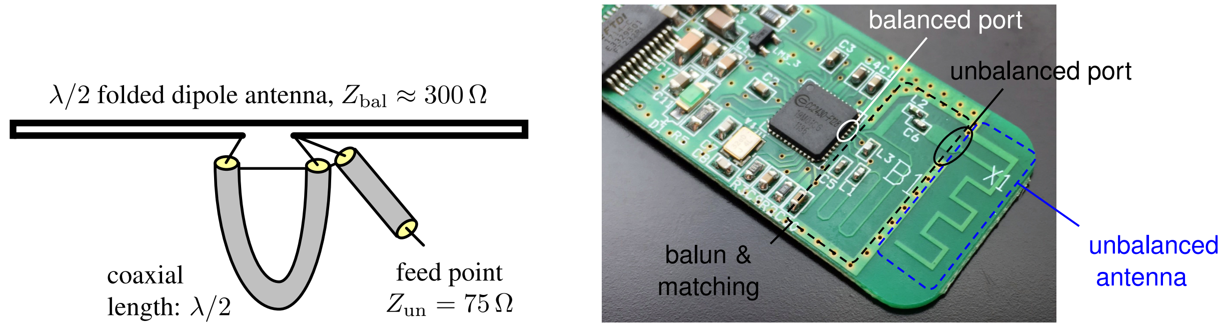
(a) 1:1 balanced to unbalanced

(b) 4:1 unbalanced to balanced

Figure 4.17 Baluns are needed to switch from balanced to unbalanced elements and back; often they also provide means of impedance transformation.

4.6.2 Realizations of Baluns

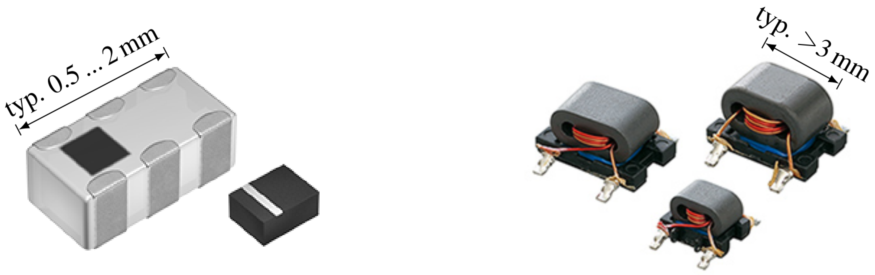
Baluns can be realized using transmission lines (see Fig. 4.18), (discrete or integrated) reactive elements, and magnetic transformers (see Fig. 4.19). Transmission-line type baluns (TL baluns) are often realized by amateur radio operators and wherever space availability allows or power handling requirements demand them; they are very simple and allow high power throughput but are usually narrowband and occupy a considerable amount of space, since the lines have to be large with respect to the wavelength. Most commercially available miniature baluns consist of reactive elements in form of multilayer hybrid integrated circuits; they can be multi- or broadband and very cost effective. Larger and usually more efficient types use wire-wound transformers, which tend to be more expensive but can often lead to broadband solutions.


$\lambda/2$ folded dipole antenna, $Z_{\text{bal}} \approx 300\,\Omega$
coaxial
length: $\lambda/2$
feed point
$Z_{\text{un}} = 75\,\Omega$
balanced port
unbalanced port
balun &
matching
unbalanced
antenna

(a) 4:1 TL balun using $\lambda/2$ coaxial line, typical configuration of amateur radio operators to connect balanced dipole antennas to an unbalanced receiver

(b) TL balun integrated into a PCB layout, connecting balanced RF chip to unbalanced antenna (by courtesy of electronics4u)

Figure 4.18 Typical transmission line-based baluns.



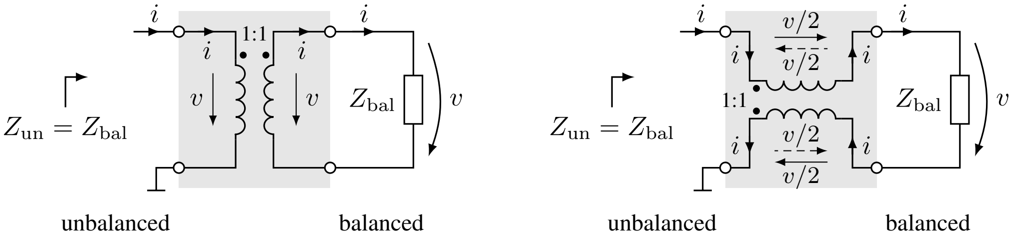
(a) Ceramic multilayer integrated baluns (TDK)

(b) SMD transformer baluns (Murata)

Figure 4.19 Examples of commercial baluns: Cermanic hybrids and wire-wound transformers.

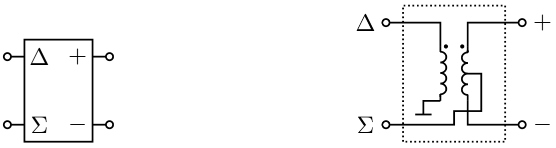
The transformer can be used as a voltage balun, see Fig. 4.20(a), or as a current balun, see Fig. 4.20(b). The voltage balun is commonly used in low-frequency circuits, as it works effectively if the load impedance is symmetric with respect to ground. The *mixed-mode converter*, see Fig. 4.21, is a similar, yet more general element, which besides the difference also provides the sum of two signals.

The choke (current) balun, on the other hand, is very popular for higher-frequency circuits, since it is not as dependent on the symmetry of the load. However, it works well only for relatively small load impedances (until about 200 Ω), i.e., a certain current must flow. The large advantage of the current balun can be seen by observing the current, which is forced to be equal on either side of the transformer, and the voltage. On



(a) Voltage balun (b) Current (choke) balun

Figure 4.20 The simplest 1:1 voltage and current baluns made of transformers.



(a) Mixed-mode converter symbol (b) Realization using transformers

Figure 4.21 The mixed-mode converter can be interpreted as a more general Balun: it provides both the sum (Σ) *and* the difference (Δ) of two input signals ($+/-$) and vice versa.

either side, a voltage is induced which is inverse in direction from the one determined by the current. They cancel each other out. Hence, there is no resulting voltage across the windings. Thus, exactly the same impedance is seen at the input, regardless of any parasitics across the windings.

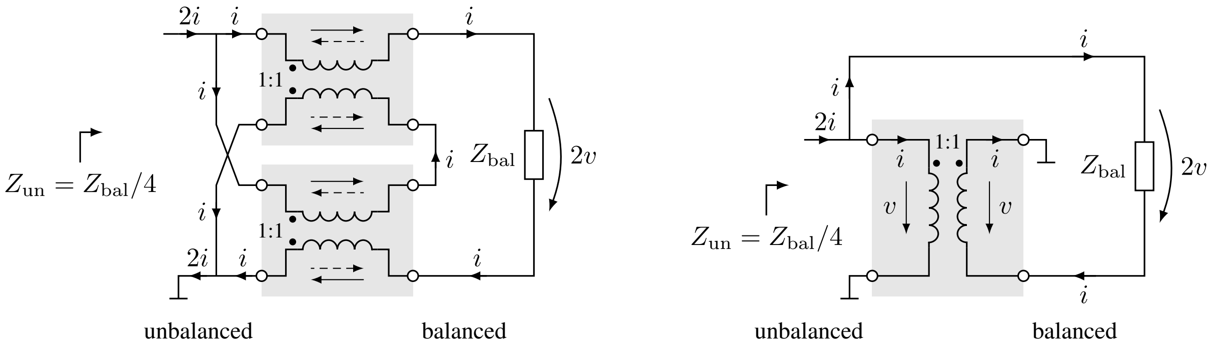
Many further baluns exist, both voltage- and current-types, with different impedance transformation ratios and other characteristics. The standard nomenclature specifies the impedance ratio as $Z_{\mathrm{bal}} : Z_{\mathrm{un}}$ (in the same sequence as in the name “balun”), where commonly the balanced impedance Z_{bal} is larger than or equal to the unbalanced impedance Z_{un} (typically the receiver). In Fig. 4.22, two additional examples for 4:1 impedance ratios are shown: the 4:1 current balun Fig. 4.22(a), also referred to as a *Guanella Current Balun*, named after its Swiss inventor Gustav Guanella[2], and the corresponding 4:1 voltage balun, also called *Ruthroff Voltage Balun*, after the American electrical engineer Clyde L. Ruthroff [64], working at Bell Labs, in Holmdel NJ, at the time (around the 1950s). More on transformer-based baluns and others can be found in [44] (in German).

To save space on the PCB and costs, discrete-type LC baluns are also used. They tend to be efficient but narrowband. Two possible solutions are shown in Fig. 4.23. It can easily be shown that, in order to obtain the desired balancing condition $v_1 = -v_2$ and the real-valued input impedance $Z_{\mathrm{un}} = R_{\mathrm{un}}$ at a single design frequency ω_0, the component values for the schematic shown in Fig. 4.23(a) have to be chosen according to

$$C = \frac{Y_0}{\omega_0} \quad \text{and} \quad L = \frac{Z_0}{\omega_0}\,, \tag{4.47}$$

where $Z_0 = \sqrt{R_{\mathrm{bal}} R_{\mathrm{un}}}$ is the characteristic impedance of the balun and $Y_0 = 1/Z_0$ the corresponding admittance. If the balanced load R_{bal} is symmetrically grounded internally, the lower capacitor and inductor

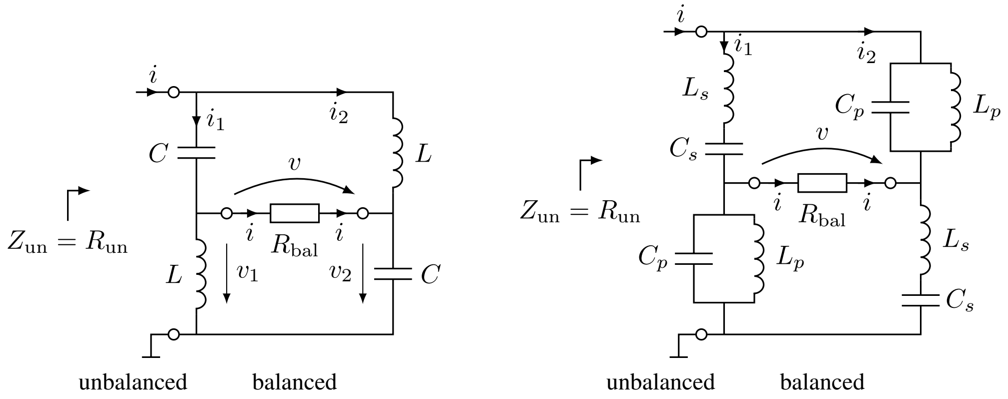
[2] Gustav Guanella, 1909–1982, born in Chur, graduated from the Swiss Federal Institute of Technology in Zurich (ETH Zurich) in 1933. He worked most of his career at Brown, Boveri & Cie (BBC), in Baden, Switzerland, where he became head of a department involved in high-frequency electronics product development. In his lifetime, Guanella was inventor or co-inventor of more than 200 patents; he personally held patents for both the 1:1 and 4:1 current baluns, in Figs. 4.20(b) and 4.22(a). He was also the brother in-law of Albert Hofmann, the discoverer and pioneer of LSD, with whom together he was honored with the title Dr. sc. tech. (honoris causa) in 1969.



(a) 4:1 Guanella current balun (b) 4:1 Ruthroff voltage balun

Figure 4.22 Common 4:1 baluns realized using transformers.

can be omitted to reduce complexity and cost. Then, however, the exact phase relations between the symmetric port can no longer be retained for the reduced discrete balun. As shown in Fig. 4.23(b), this balun



(a) single-band (b) dual-band

Figure 4.23 Discrete-type LC baluns.

type can be made dual-band, i.e., designed to be matched at two potentially different balanced loads $R_{\text{bal},1}$ and $R_{\text{bal},2}$, at two angular design frequencies $\omega_2 > \omega_1$. The generalization is straightforward and simply takes place by replacing the capacitors C by series and the inductors L by parallel resonators, respectively. The series and parallel capacitance and inductance values, respectively, then have to be chosen according to

$$L_s = \frac{1}{\gamma}\left(\frac{Z_2}{\omega_1} + \frac{Z_1}{\omega_2}\right) \tag{4.48a}$$

$$C_s = \gamma\left(\frac{Y_2}{\omega_1} \parallel \frac{Y_1}{\omega_2}\right) \tag{4.48b}$$

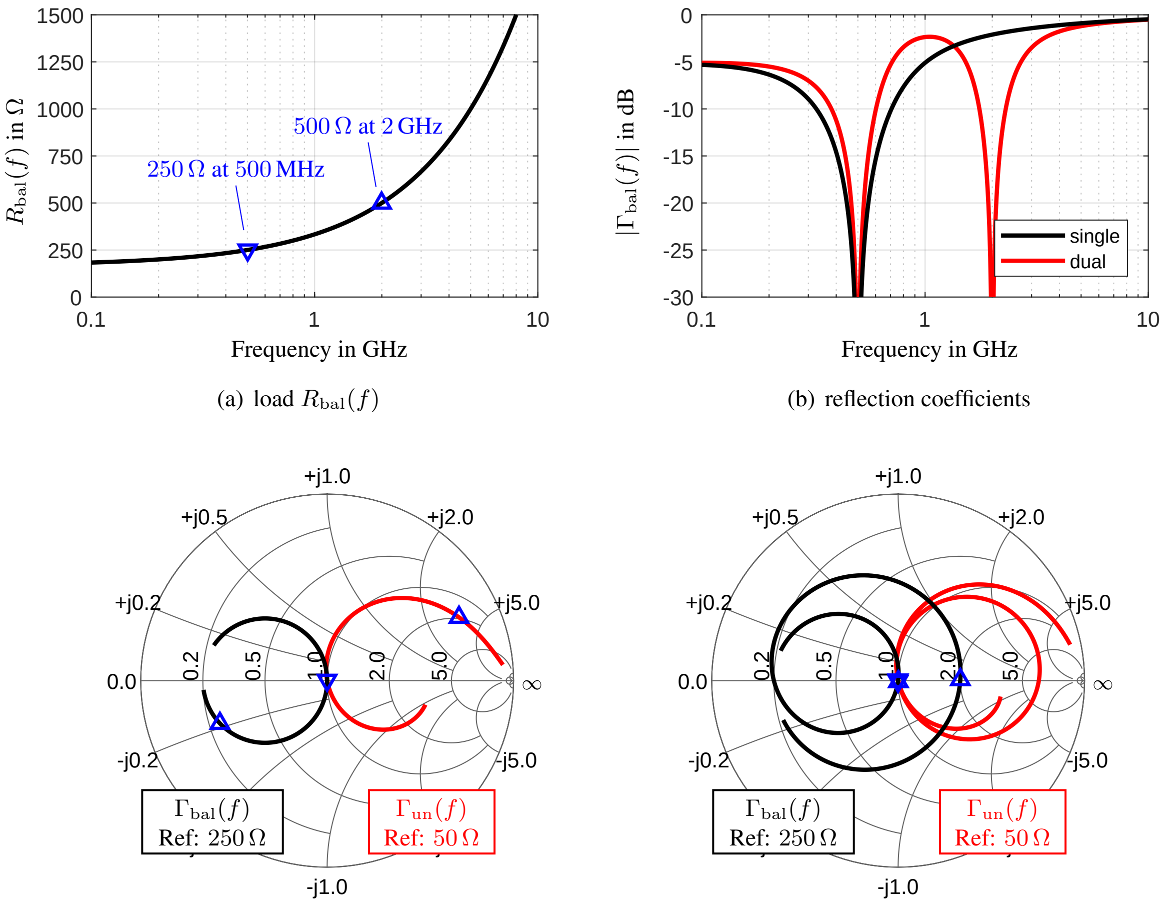
$$L_p = \gamma\left(\frac{Z_2}{\omega_1} \parallel \frac{Z_1}{\omega_2}\right) \tag{4.48c}$$

$$C_p = \frac{1}{\gamma}\left(\frac{Y_2}{\omega_1} + \frac{Y_1}{\omega_2}\right) \tag{4.48d}$$

where $\gamma = \omega_2/\omega_1 - \omega_1/\omega_2$ is a normalized deviation of the two design frequencies and $Z_n = \sqrt{R_{\text{bal},n}R_{\text{un}}}$ are the characteristic impedances of the balun at angular frequency ω_n, for $n = 1, 2$, respectively.

Similar design rules can be derived for two different unbalanced impedances instead of two different balanced loads; however, that is of less practical relevance.

Fig. 4.24 shows a single- and a dual-band design of discrete baluns, which match a frequency-dependent balanced load at a single or two frequencies to a constant unbalanced reference impedance of $50\,\Omega$. As can be seen, the perfect balanced matching is successful at the design points; the resulting bandwidth is larger for the single-band balun. In the smith charts it is visible how the dual-band design provides two rotations, where the second one leads to the point $2 \times 250\,\Omega = 500\,\Omega$, as desired.



(a) load $R_{\mathrm{bal}}(f)$

(b) reflection coefficients

(c) single-band balun

(d) dual-band balun

Figure 4.24 Comparison of the single- and dual-band discrete baluns: The frequency-dependent balanced load (a) is matched to an unbalanced reference impedance of $R_{\mathrm{un}} = 50\,\Omega$ at the frequencies $f_1 = 0.5$ and $f_2 = 2\,\mathrm{GHz}$, where the impedances are $250\,\Omega$ and $500\,\Omega$, respectively. The resulting reflection coefficients are compared in (b) and separately plotted in (c) and (d).

5 Active RF Components

Many key components can be seen repeatedly throughout a receiver and transmitter chain of a wireless system: amplifiers, mixers, oscillators, synthesizers, filters, antennas, and others. Except for antennas, which are covered in Chapter 7, these key components are treated in a general way here before used in combination to build transmitter and receiver systems in Chapter 6. We concentrate on specification and system understanding rather than on design and implementation.

5.1 Amplifiers

5.1.1 Introduction

Amplifiers constitute a group of the most important components of any wireless communication system; they usually appear both close to the end of a transmitter as well as close to the beginning of a receiver. We distinguish between two main groups (with many inbetween flavors) of amplifiers:

- Low-Noise Amplifier (LNA) and
- Power Amplifier (PA).

In a wireless communication system, LNAs are typically used close to the receiving antenna, where power levels are (very) low and low noise figures are paramount. Also, gains need to be sufficiently high to relax the requirements on the noise figures of the following stages, as will be discussed. Contrary to that, PAs are commonly used just before the transmitting antenna, at power levels way above thermal noise, such that noise figures are not even a design parameter. Here, maximum power output, power efficiency, and possibly linearity are the parameters of interest.

Evidently, for different amplifier types different design parameters are important. In the following, we discuss the most important parameters that are needed either for LNAs or for PAs: gain, stability, noise behavior, and nonlinear effects. A more extensive list of parameters can also be found in [43].

5.1.2 Gain

An amplifier is typically a two-port device and from its S-parameters the following three fundamental parameters are identified:

- The *gain* G, also called *forward gain* or *power gain* of an amplifier, is usually $|s_{21}|^2$ given in dB, hence

  $$G = 20 \log_{10} |s_{21}| \,. \tag{5.1}$$

  It refers to the small-signal gain, i.e., how many times stronger an input signal of interest shows up at the output of the device, while disregarding any bias. *Gain flatness* expresses the variation of the amplifier gain over the frequency range the amplifier is specified for.

- The complementary parameter to forward gain is *isolation*, usually given by

$$I = -20\log_{10}|s_{12}|\,, \tag{5.2}$$

resulting in a positive dB figure[1]. Isolation expresses how a signal applied to the output is attenuated back to the input of an amplifier.

- Lastly, the term *active directivity* expresses the difference between isolation and forward gain in dB:

$$D = -20\log_{10}\left(|s_{21}|\,|s_{12}|\right) = I - G\,. \tag{5.3}$$

Directivity is an indicator of source-load isolation.

According to these definitions, it is generally desirable that all these three parameters G, I, and D be large.

Two-Port Gain Definitions

The previous three parameters only describe the gain element with respect to the reference impedances, with which the S-parameters have been measured. However, the performance of an amplifier in a system depends on the properties of the entire system, most importantly on the impedance mismatch at the input and at the output, as depicted in Fig. 5.1. Whereas the forward gain G corresponds to s_{21} and expresses how much a signal is amplified if both ends of the amplifier are connected to the reference impedance (mostly $50\,\Omega$), additional gain definitions state the performance of an amplifier when considering the source and load as well. The definitions of the powers used to formulate these gains are illustrated in Table 5.1.

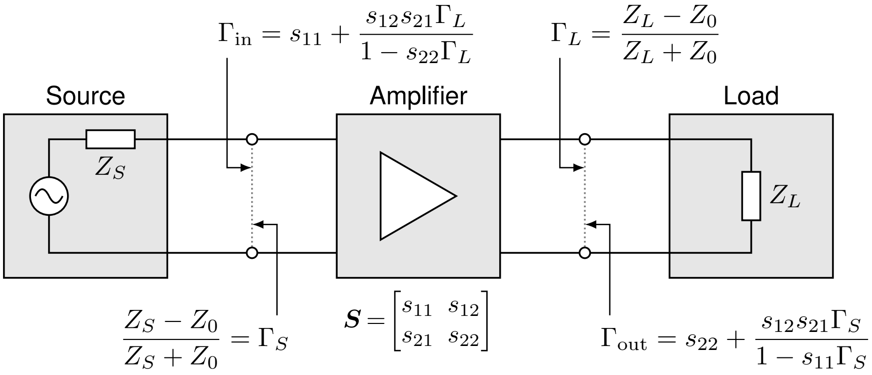

$\Gamma_\text{in} = s_{11} + \dfrac{s_{12}s_{21}\Gamma_L}{1 - s_{22}\Gamma_L}$
$\Gamma_L = \dfrac{Z_L - Z_0}{Z_L + Z_0}$
Source
Amplifier
Load
Z_S
Z_L
$\dfrac{Z_S - Z_0}{Z_S + Z_0} = \Gamma_S$
$\boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}$
$\Gamma_\text{out} = s_{22} + \dfrac{s_{12}s_{21}\Gamma_S}{1 - s_{11}\Gamma_S}$

Figure 5.1 Reflection coefficients relevant to compute the three two-port gains operating gain, available gain, and transducer gain of the amplifier.

- The *power gain*, also referred to as *operating gain* or *actual gain*, is the ratio of the actual power P_L dissipated in the load Z_L and the power P_in delivered to the input of the two-port network:

$$G_P = \frac{P_L}{P_\text{in}} = \frac{|s_{21}|^2\,\frac{1-|\Gamma_L|^2}{|1-s_{22}\Gamma_L|^2}}{1 - |\Gamma_\text{in}|^2} = \frac{|s_{21}|^2\left(1 - |\Gamma_L|^2\right)}{\left(1 - |\Gamma_\text{in}|^2\right)|1 - s_{22}\Gamma_L|^2}\,. \tag{5.4}$$

It describes the increase in *actually delivered* power from the input to the output, i.e., how much more power is absorbed by the load compared to how much power is absorbed by the input of the amplifier.

[1] Caution: Some datasheets and textbooks use a different (old) definition without the negative sign.

Note that this gain is independent[2] of the source impedance Z_S or source reflection coefficient Γ_S, but generally is a function of the load reflection coefficient: $G_P = G_P(\Gamma_L)$:

- The input power P_in is generally lower than the maximum power the source could deliver with optimal input matching, $\max P_S$. However, this simply results in a correspondingly lower maximum output power at the amplifier. Thus, this influence is cancelled out during the division.
- The actual power delivered to the load P_L, usually is again lower than what could be delivered if the output was perfectly matched, $\max P_\text{out}$. However, this circumstance does not cancel out and therefore remains in the final expression of the power gain.

- The *available gain* is the ratio of the power available at the output of the two-port network and the power available from that particular source at the amplifier input:

$$G_A = \frac{\max P_\text{out}}{\max P_S} = \frac{|s_{21}|^2 \frac{|1-\Gamma_\text{in}\Gamma_S|^2}{|1-s_{11}\Gamma_S|^2(1-|\Gamma_\text{out}|^2)}}{\frac{|1-\Gamma_\text{in}\Gamma_S|^2}{1-|\Gamma_S|^2}} = \frac{\left(1-|\Gamma_S|^2\right)|s_{21}|^2}{|1-s_{11}\Gamma_S|^2\left(1-|\Gamma_\text{out}|^2\right)} \,. \tag{5.5}$$

It describes the increase in *available* power from the input to the output; the ratio of the powers that could potentially be extracted from the source and from the amplifier, if either were matched accordingly. Thus, both powers implicitly assume perfect conjugate matching at the source and the load ports and depends on Z_S (the bottle-neck of the source) but not on Z_L; thus, $G_A = G_A(\Gamma_S)$.

Note that the available gain and power gain expressions are very similar; by exchanging source and load reflection coefficients $\Gamma_S \leftrightarrow \Gamma_L$ as well as input and output reflection coefficients $\Gamma_\text{in} \leftrightarrow \Gamma_\text{out}$, they can be transformed into each other. What the power gain says about the dependence on load matching conditions, the available gain says about the dependence on source matching conditions.

It may seem that the available gain is of less practical value than the power gain. However, this is not entirely correct, since the available gain is of crucial importance when considering the noise contributions in cascaded systems.

- An important special case of the available gain is the *maximum available gain*, also often abbreviated MAG, the ratio of the maximum power available at the output, $\max P_\text{out}$, and the power actually delivered to the amplifier input, P_in:

$$G_M = \frac{\max P_\text{out}}{P_\text{in}} = \frac{|s_{21}|^2 \frac{|1-\Gamma_\text{in}\Gamma_S|^2}{|1-s_{11}\Gamma_S|^2(1-|\Gamma_\text{out}|^2)}}{1-|\Gamma_\text{in}|^2} = \frac{|s_{21}|}{|s_{12}|}\left(K-\sqrt{K^2-1}\right) \,. \tag{5.6}$$

K is *Rollet's stability factor*, as defined later in Eq. (5.23). According to that definition, K depends only on the S-parameters of the amplifier itself. Thus, the same holds true for the MAG; it is an intrinsic property of the amplifier independent of the actual source and load conditions.

Note that the maximal available gain according to Eq. (5.6) is only properly defined[3] for $K \geq 1$, implying an unconditionally stable amplifier[4], as will be discussed in Section 5.1.3. If this is not the case, we have to settle for the *maximum stable gain* given by $G_{MS} = G_M|_{K=1} = |s_{21}| \,/\, |s_{12}|$.

[2] The characteristics of some active devices themselves may depend on the source impedance Z_S.

[3] For very large stability factors $K \gg 1$, it is numerically more stable to use the (equivalent) factor $1/(K+\sqrt{K^2+1})$ instead of $(K-\sqrt{K^2-1})$.

[4] Instability generally leads to oscillation, independent of the input, which might therefore be considered "infinite" gain, but does not really make any practical sense.

- The *transducer gain* is defined as the ratio of the power delivered to the load and the power available from the source:

$$\boxed{G_T = \frac{P_L}{\max P_S} = \frac{|s_{21}|^2 \, \frac{1-|\Gamma_L|^2}{|1-s_{22}\Gamma_L|^2}}{\frac{|1-\Gamma_{\text{in}}\Gamma_S|^2}{1-|\Gamma_S|^2}} = \frac{(1-|\Gamma_S|^2)|s_{21}|^2(1-|\Gamma_L|^2)}{|(1-s_{11}\Gamma_S)(1-s_{22}\Gamma_L)-s_{21}s_{12}\Gamma_S\Gamma_L|^2} \,.} \tag{5.7}$$

It states how much of power actually is delivered to the load, P_L, under both non-ideal source and load matching conditions, compared to to the maximum power that would be available from the source, had it been connected to a conjugate-matched load, $\max P_S$.

Often, the transducer gain is the most important performance measure of the amplifier within an entire system. Essentially, it describes how effective the amplifier is in extracting the available power from the source, increasing it, and then delivering it to the load. Thus, the transducer gain depends on both source and load impedance matching conditions, meaning $G_T = G_T(\Gamma_S, \Gamma_L)$.

For most amplifiers, it is true that $|s_{12}| \ll |s_{21}|$. Under the assumption that $s_{12} = 0$ (implying perfect isolation and directivity), the input and output of the amplifier can be separated and the *unilateral*[5] transducer gain can be formulated from Eq. (5.7) as

$$G_{TU} = \frac{P_L}{\max P_S} = \frac{1-|\Gamma_S|^2}{|1-s_{11}\Gamma_S|^2}\,|s_{21}|^2\,\frac{1-|\Gamma_L|^2}{|1-s_{22}\Gamma_L|^2} \,. \tag{5.8}$$

Furthermore, for complex conjugate matching, $\Gamma_S = \Gamma_{\text{in}}^* = s_{11}^*$ and $\Gamma_L = \Gamma_{\text{out}}^* = s_{22}^*$, we get

$$\boxed{G_{TU,\max} = \frac{1}{1-|s_{11}|^2}\,|s_{21}|^2\,\frac{1}{1-|s_{22}|^2} \,.} \tag{5.9}$$

The first and the last factor in Eq. (5.9) designate the gain obtained by matching the input and the output, respectively.

Finally, from the unilateral transducer gain expression Eq. (5.8) it can easily be seen, that by setting $\Gamma_S = \Gamma_L = 0$, implying both source and load are matched to the reference impedance Z_0 on either side of the amplifier, we get

$$G_{TU,\text{opt}} = G_{TU}|_{\Gamma_S=\Gamma_L=0} = |s_{21}|^2 = G \,. \tag{5.10}$$

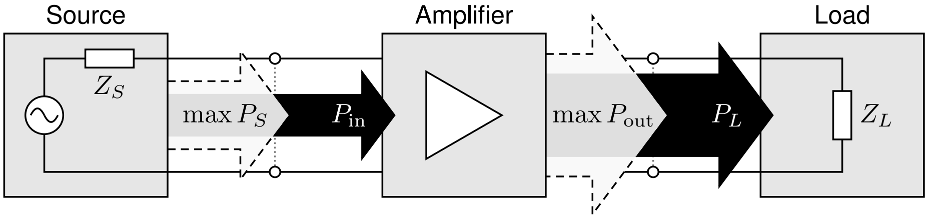
Note that this gain can in general be exceeded in many other cases, where $\Gamma_S \neq 0$ and $\Gamma_L \neq 0$.

Two-Port Gain Comparison

Clearly, from the definitions of (5.5) and (5.6), it follows that $G_M \geq G_A$ and $G_M \geq G_P$. Similarly, from (5.4) and (5.7), it can be concluded that $G_T \leq G_P$ and $G_T \leq G_A$. However, no such comparative statement can be made about G_A and G_P. Table 5.1 lists all four two-port gain definitions according to their constituents.

So then, it is a reasonable question to ask: Which one of these gain definitions should be used, when specifying or designing an amplifier? Which one do vendors use to specify the performances of their components? The truth is, often we do not know exactly; but for well-designed amplifiers, the gain values generally do not differ significantly.

[5] "unilateral" means no interaction between the output and input ports



	P_L	$\le$	$\max P_{\text{out}}$
P_{in}	**Power Gain** $G_P = \dfrac{P_L}{P_{\text{in}}}$ $G_P(\Gamma_L) = \dfrac{\lvert s_{21}\rvert^2 \left(1-\lvert\Gamma_L\rvert^2\right)}{\left(1-\lvert\Gamma_{\text{in}}\rvert^2\right)\lvert 1-s_{22}\Gamma_L\rvert^2}$	$\le$	**Max. Available Gain (MAG)** $G_M = \dfrac{\max P_{\text{out}}}{P_{\text{in}}}$ $G_M = \dfrac{\lvert s_{21}\rvert}{\lvert s_{12}\rvert}\left(K-\sqrt{K^2-1}\right) \quad \forall K \geq 1$
$\le$	$\ge$		$\ge$
$\max P_S$	**Transducer Gain** $G_T = \dfrac{P_L}{\max P_S}$ $G_T(\Gamma_S, \Gamma_L) = \dfrac{(1-\lvert\Gamma_S\rvert^2)\lvert s_{21}\rvert^2(1-\lvert\Gamma_L\rvert^2)}{\lvert(1-s_{11}\Gamma_S)(1-s_{22}\Gamma_L)-s_{21}s_{12}\Gamma_S\Gamma_L\rvert^2}$	$\le$	**Available Gain** $G_A = \dfrac{\max P_{\text{out}}}{\max P_S}$ $G_A(\Gamma_S) = \dfrac{\left(1-\lvert\Gamma_S\rvert^2\right)\lvert s_{21}\rvert^2}{\lvert 1-s_{11}\Gamma_S\rvert^2\left(1-\lvert\Gamma_{\text{out}}\rvert^2\right)}$

Table 5.1 Comparison of the two-port gains and the considered powers.

Some notable special cases include:

- Whenever building a circuit using (50 Ω) coaxial cables and components, such as for measurement setups, G_T is usually the only gain of interest, as we only care about the answer to the question "How much power do we get compared to what we could have gotten without the amplifier?". Thus, for many practical applications, the transducer gain is often the most important.

- When designing an amplifier from scratch (e.g. starting from a single transistor), stability considerations are central, which relate to the MAG G_M. Additionally, a large gap of G_T to G_A points to a significant improvement potential when matching the output better, and a large gap between G_P and G_T similarly to a significant potential in improvement by matching the input better.

- To assess the power transfer efficiency of a passive power delivery network with $G \leq 1$, the operating gain G_P is of central interest, since it expresses the actual efficiency $P_{\text{out}}/P_{\text{in}} = P_L/P_{\text{in}}$ at which the transfer takes place. Similar considerations are sometimes made for instrumentation amplifiers, where the power extracted from the source cannot or should not be maximized.

- In the unique case where the amplifier in the system is terminated by the characteristic impedance Z_0 used to measure the original S-parameter, follows $\Gamma_S = \Gamma_L = 0$ and the power and available gains reduce to

$$G_P\big|_{\Gamma_S=\Gamma_L=0} = \frac{|s_{21}|^2}{1-|s_{11}|^2} \qquad \text{and} \qquad G_A\big|_{\Gamma_S=\Gamma_L=0} = \frac{|s_{21}|^2}{1-|s_{22}|^2}\,, \tag{5.11}$$

  respectively, while the transducer gain becomes the optimally matched unilateral transducer gain according to Eq. (5.10). The MAG, being independent of Γ_L and Γ_S, is still given by Eq. (5.6) and exceeds all three.

Amplifier and Gains

The S-parameters (with respect to $50\,\Omega$) of a Qorvo QPL9503 LNA at $f = 1\,\text{GHz}$ and for a supply voltage of $3.3\,\text{V}$ are:

$$\boldsymbol{S} = \begin{bmatrix} 0.50\,\angle -115.1^\circ & 0.031\,\angle 10.8^\circ \\ 9.51\,\angle 73.9^\circ & 0.153\,\angle -62.9^\circ \end{bmatrix}$$

Suppose, we have a $75\,\Omega$ generator and a $25\,\Omega$ load that we want to use it with. The source, load, input, and output reflection coefficients are, respectively:

$$\Gamma_S = 0.2 \qquad \Gamma_{\text{in}} = 0.59\angle - 111.6^\circ$$
$$\Gamma_L = -0.333 \qquad \Gamma_{\text{out}} = 0.114\angle - 45.4^\circ$$

From the S-parameters follows $G = 19.56\,\text{dB}$ and $G_M = 21.1\,\text{dB}$ and with the aforementioned reflection coefficients we can calculate the two-port gains, which turn out to be:

$$G_A = 19.05\,\text{dB} \qquad G_T = 18.25\,\text{dB} \qquad G_P = 20.7\,\text{dB}$$

In the optimum case, the *transducer gain* can be equal to the *maximum available gain* (MAG), $G_T = G_M$; thus, we are currently missing about $3\,\text{dB}$ gain. The *available gain* from this source with Γ_S, is not much higher than what we currently get: $G_A - G_T = 0.8\,\text{dB}$. However, the *power gain* to this load Γ_L promises almost 2.5 dB more. Since G_T is always smaller than both, we cannot achieve the power gain, because the available gain is lower. Thus, it is desirable to enhance the source reflection (e.g. via matching network), to increase the available power. If the amplifier were *unilateral* ($S_{12} = 0$), the best choice would be a conjugate match $\Gamma_S = S_{11}^*$. Doing so, we obtain a new available gain of $G'_A = 21.1\,\text{dB}$, thereby exceeding the power gain (which remains at $G'_P = G_P = 20.7\,\text{dB}$), and the transducer gain increases to $G'_T = 20.6\,\text{dB}$. Now, to further increase G_T, we would have to enhance Γ_L next. Once again, using the conjugate match at the output, $\Gamma_L = S_{22}^*$, we obtain

$$G''_A = 21.1\,\text{dB} \qquad G''_T = 20.88\,\text{dB} \qquad G''_P = 20.9\,\text{dB}$$

This is close to and yet still not quite the optimum, as $G''_T < G_M$. Why not? Because, since $S_{12} \neq 0$, the optimum source and load reflection coefficients are not the conjugate matches of the S-parameters of the amplifier, but of the input and output reflection coefficients: $\Gamma_S = \Gamma_{\text{in}}^*$ and $\Gamma_L = \Gamma_{\text{out}}^*$. Each of those, however, is a (complex, bilinear) function of the other. Thus, we cannot easily find a solution that fits both; we have to use a numerical technique, for example an iterative process to stepwise approach the equilibrium, where we use the previous conjugate matches as starting points. Indeed, at the points

$$Z_S^0 = 35.11\angle 50.4 \xrightarrow{\text{opt}} Z_S^{\text{opt}} = 31.38\angle 55.4$$

$$Z_L^0 = 57.34\angle 15.6 \xrightarrow{\text{opt}} Z_L^{\text{opt}} = 38.47\angle 27.8$$

we finally find $G_T = G_A = G_P = G_M = 21.1\,\text{dB}$. Of course, in practice, a difference of up to a dB or two is usually good enough. The problems associated with trying to design the matching networks (as well as loss and bandwidth considerations) most likely outweigh the advantages by far.

Gain Circles

So-called gain circles illustrate the situation where the input or output are not matched in the Smith chart. Under the condition that s_{12} is zero, the gain term due to input matching[6] of Eqs. (5.5) and (5.7)

$$g = \frac{1 - |\Gamma_S|^2}{|1 - s_{11}\Gamma_S|^2} \tag{5.12}$$

can be isolated. Note that g is given as a linear term. In the following, we want to find the geometrical locations for which constant gains g result. To do so, we define $\Gamma_S \triangleq x + jy$ and $s_{11} \triangleq a + jb$. Eq. (5.12) can now be written as

$$1 - (x^2 + y^2) = g|1 - (a + jb)(x + jy)|^2. \tag{5.13}$$

Carrying out the multiplications, we end up with

$$\begin{aligned} 1 - x^2 - y^2 &= g\left((1 - ax + by)^2 + (bx + ay)^2\right) \\ &= g\left(1 + a^2x^2 + b^2y^2 - 2ax + 2by - 2abxy + b^2x^2 + a^2y^2 + 2abxy\right) \\ &= g\left(1 + x^2(a^2 + b^2) + y^2(a^2 + b^2) - 2ax + 2by\right). \end{aligned} \tag{5.14}$$

Exchanging some terms, we can write

$$x^2(g(a^2 + b^2) + 1) - 2agx + y^2(g(a^2 + b^2) + 1) + 2bgy + g - 1 = 0, \tag{5.15}$$

or by redefining new terms,

$$x^2 - 2Cx + y^2 - 2Dx = E, \tag{5.16}$$

where

$$C \triangleq \frac{ag}{g(a^2 + b^2) + 1}, \quad D \triangleq \frac{-bg}{g(a^2 + b^2) + 1}, \quad E \triangleq \frac{1 - g}{g(a^2 + b^2) + 1}. \tag{5.17}$$

Completing the square leads to

$$\begin{aligned} x^2 - 2Cx + C^2 + y^2 - 2Dx + D^2 &= E + C^2 + D^2 \\ (x - C)^2 + (y - D)^2 &= E + C^2 + D^2 \end{aligned} \tag{5.18}$$

It can now be seen that the geometrical locations for constant gains g are circles with their centers on a line between the origin and s_{11}^*. More specifically, the center of the gain circle with gain g is given by

$$\Gamma_S = \frac{g \cdot s_{11}^*}{g|s_{11}|^2 + 1}. \tag{5.19}$$

The corresponding radius is given by

$$\rho_s = \sqrt{E + C^2 + D^2} = \frac{\sqrt{g|s_{11}|^2 + 1 - g}}{g|s_{11}|^2 + 1}. \tag{5.20}$$

Fig. 5.2 shows the circles of constant gain resulting from input mismatch for $s_{11} = -0.5 - j0.2$. Similar circles can be drawn for output mismatch. In summary, amplifiers may fail to achieve the maximum gain due to one or several of the following reasons:

- failure of matching at input or output
- matching uses lossy elements
- potential instability (in this case, the maximum stable gain is relevant) .

[6] The corresponding situation with respect to output matching can be obtained by exchanging s_{11} and s_{22}, and Γ_S and Γ_L, respectively.

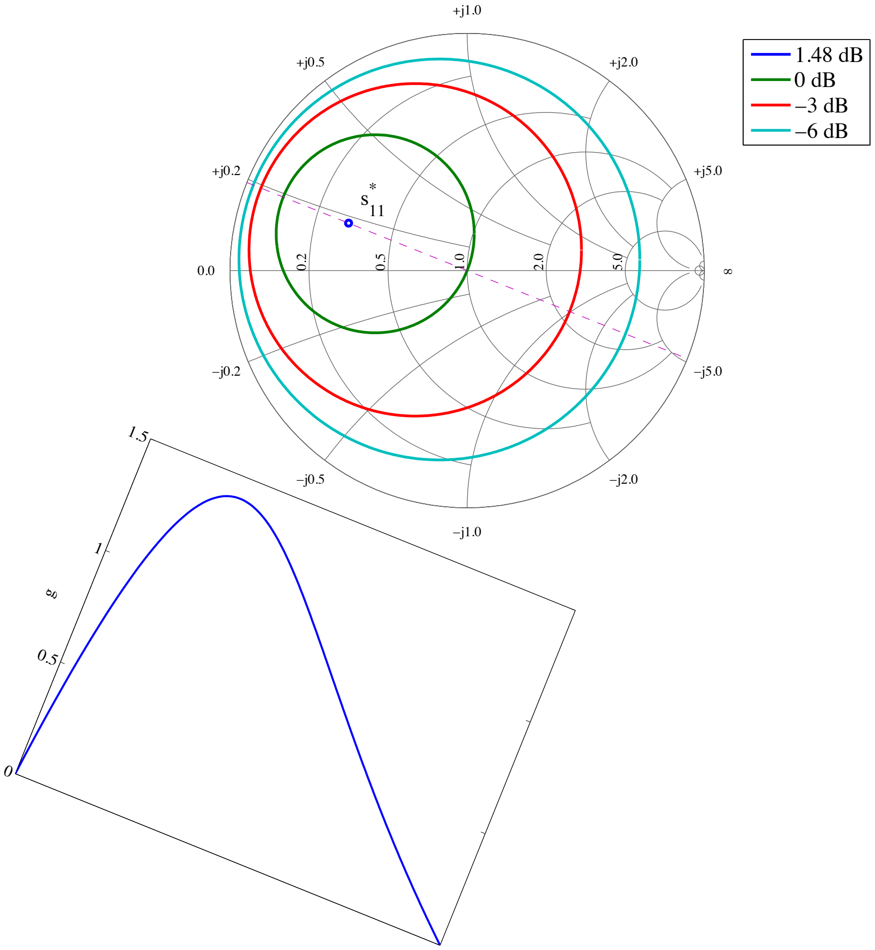


Figure 5.2 Gain circles of an amplifier (top) and corresponding gains obtained from source reflection coefficients sitting on the line through the origin and s_{11}^* (bottom).

5.1.3 Stability

A prerequisite of an amplifier is its stability at the operational point. We distinguish between *conditional* and *unconditional* stability:

- An amplifier is said to be conditionally stable if it oscillates under particular load or source impedance conditions.

- If an amplifier is unconditionally stable, on the other hand, it will not oscillate no matter what load or source impedances are connected. An amplifier is unconditionally stable if neither input nor output impedance have a negative real part at any frequency.

Stability can be proven by showing that

$$K > 1, \tag{5.21}$$

$$|\Delta| < 1, \tag{5.22}$$

where

$$K = \frac{1 - |s_{11}|^2 - |s_{22}|^2 + |\Delta|^2}{2|s_{12}s_{21}|}, \tag{5.23}$$

$$\Delta = s_{11}s_{22} - s_{12}s_{21}. \tag{5.24}$$

K is also called *Rollett's stability factor*, Eq. (5.21) is referred to as *Rollett's condition*. Both are named after their discoverer John Rollett [60, 61], who worked at Marconi's Wireless Telegraph Company at the time and had been inspired by the findings of Samuel Mason (known for *Mason's rule*). The two conditions of Eqs. (5.23) and (5.24) are also known as the $K - \Delta$ Test.

For a long time, these two simultaneous conditions had to be checked for stability. In 1992, Edward and Sinsky [16] found a single necessary and sufficient condition for stability

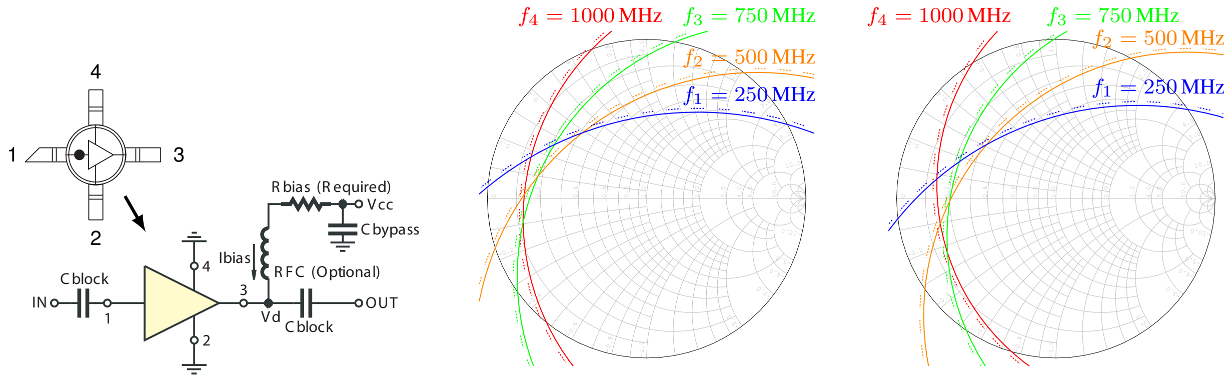
$$\mu > 1 \tag{5.25}$$

where

$$\mu = \frac{1 - |s_{11}|^2}{|s_{22} - s_{11}^*\Delta| + |s_{21}s_{12}|}, \tag{5.26}$$

$$\Delta = s_{11}s_{22} - s_{12}s_{21} \qquad \text{(again)}. \tag{5.27}$$

Unconditionally stable amplifiers have a $\mu > 1$ in their region of operation. For conditionally stable amplifiers, stability circles in the Smith chart enclose the 'no-go' region. The stability circles are given by either $|\Gamma_{\text{in}}| = 1$ or $|\Gamma_{\text{out}}| = 1$. When input and output matching circuits are designed, the crossing of these stability circles must be avoided. Fig. 5.3 shows a typical behavior of such stability circles with increasing frequencies $f_1 < f_2 < f_3 < f_4$. To keep the design stable, the source reflection coefficient Γ_S cannot lay in the shaded regions shown in (b); similarly, the load reflection coefficient Γ_L should not be in the shaded region of (c).



(a) Schematic (b) Input stability (c) Output stability

Figure 5.3 Stability circles of the conditionally stable amplifier MAR-8SM+ by Mini Circuits. The dashed lines indicate the circle areas where loads lead to potential instabilities (forbidden regions).

5.1.4 Noise Considerations

Noise has many causes: Among the most prominent sources are thermal noise, quantization noise (ADC), flicker noise ($1/f$-noise), distortions, crosstalk, etc. One of the most common adverse effects in wireless communications is thermal noise, also called Johnson noise. It is always present, but may not always play the dominant role. The spectrum of noise as met in electronic systems is depicted in Fig. 5.4. In the flat region, the noise is modeled very well by a Gaussian distributed (see Fig. 5.5) white process, whose spectrum is depicted in Fig. 5.6(a). The normalized noise signal in the time domain is shown in Fig. 5.6(b).

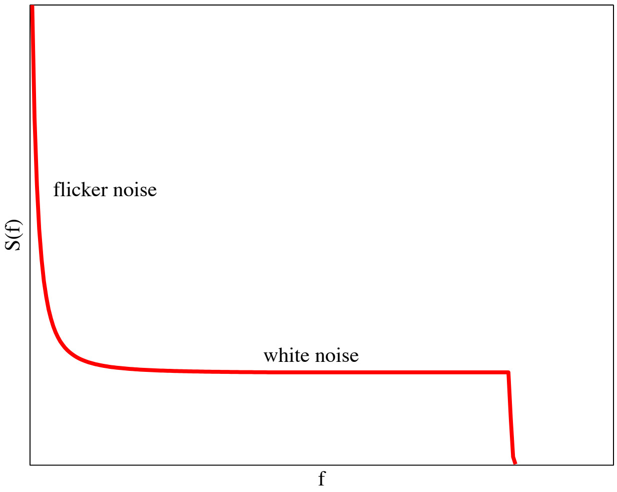


Figure 5.4 Spectrum of noise in electronic systems.

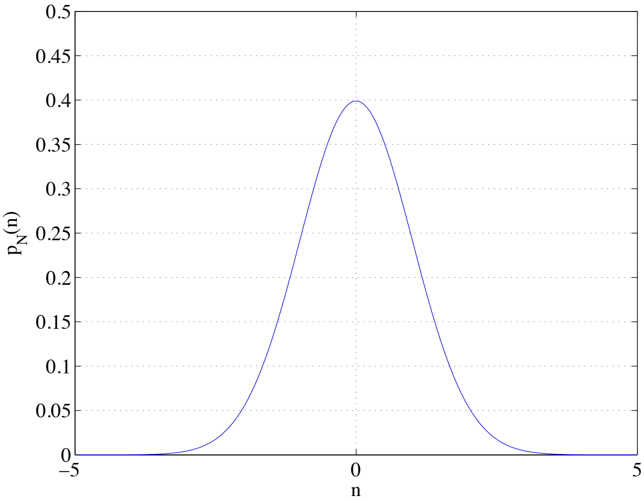
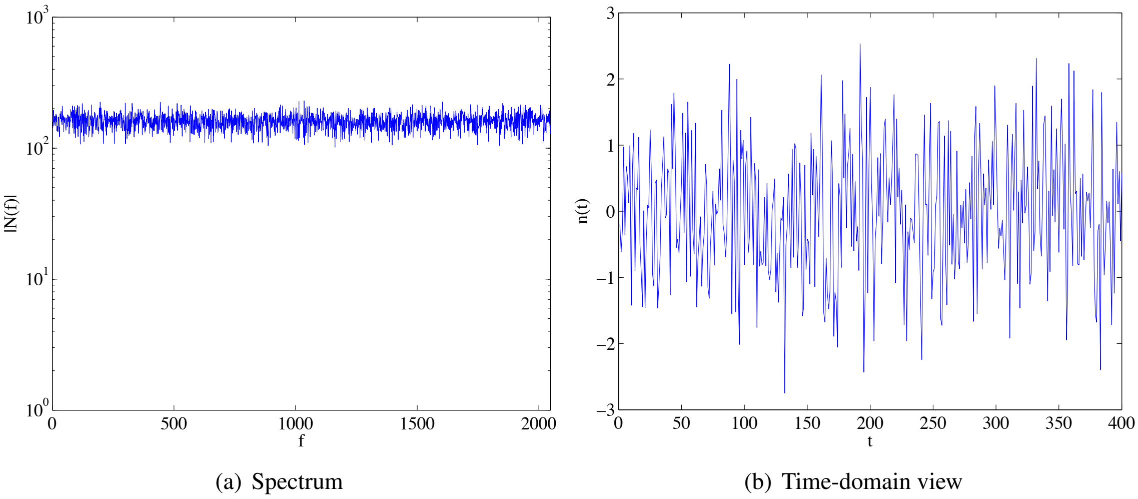


Figure 5.5 Probability density function of a Gaussian signal with $\sigma^2 = 1$.

Thermal Noise

Any object that is hotter than absolute zero, emits electromagnetic radiation. In the regions where wireless circuits transmit and receive signals, the noise power generated by a component is directly proportional to



(a) Spectrum

(b) Time-domain view

Figure 5.6 White noise.

the bandwidth, hence

$$\boxed{P_N = kT_E B}, \tag{5.28}$$

where $k = 1.38 \cdot 10^{-23}$ is the Boltzmann constant, T_E is the effective system temperature, and B is the bandwidth. The effective system temperature T_E is an important parameter of a device. It is defined as follows: Imagine a $50\,\Omega$ source connected to the input of our device. Let the source be cooled to absolute zero, thus no noise would be produced by this source. All the noise measured at the output would be internally generated noise. Now, the temperature to which the source would have to be heated in order to increase the output noise of the device by 3 dB is called the effective temperature T_E.

Noise Factor

The noise factor F is the ratio of the total output noise power to the noise power caused by the source when the source temperature is at $T_0 = 290\,\text{K}$. Without internal noise this would be

$$N_{\text{out, source}} = GkT_0B, \tag{5.29}$$

where G is the gain of the device. Note that for room temperature ($T_0 = 290\,\text{K}$), the constant kT_0 on the decibel scale is given by

$$\boxed{kT_0 = -174\,\text{dBm/Hz}}, \tag{5.30}$$

a value that shall prove to be of practical consequences. For the total output noise, we also have to consider the internally generated noise, hence

$$N_{\text{out, total}} = Gk(T_0 + T_E)B. \tag{5.31}$$

Clearly,

$$F = \frac{N_{\text{out, total}}}{N_{\text{out, source}}} = 1 + \frac{T_E}{T_0}. \tag{5.32}$$

From a system point of view, we can regard the noise factor as the degradation of the SNR at a block's output compared to its input SNR

$$F = \frac{S_{\text{in}}/N_{\text{in}}}{S_{\text{out}}/N_{\text{out}}} \geq 1\,. \tag{5.33}$$

Noise Figure

The noise figure NF is the noise factor in dB, hence

$$\mathrm{NF} = 10\log_{10} F = 10\log_{10}\left(1 + \frac{T_E}{T_0}\right). \tag{5.34}$$

Alternatively, the noise figure of a device can be regarded as the decrease of the SNR at an amplifier's input to the SNR at its output, all units in dB.

By cascading two devices with gain G_1 and G_2, respectively, and noise factors F_1 and F_2, respectively, we can evaluate the overall noise behavior. As stated above, the noise power at the output of the first device is

$$N_{1,\,\text{total}} = G_1 kB(T_0 + T_{E_1}). \tag{5.35}$$

because we assume a source at the input of the first device with an equivalent noise temperature T_0. $N_{1,\,\text{total}}$ is now the input noise of the second device. The noise power at the output of the second device is then

$$N_{2,\,\text{total}} = G_2 G_1 kB(T_0 + T_{E_1}) + G_2 kBT_{E_2}. \tag{5.36}$$

The overall noise factor is thus

$$F = \frac{G_2 G_1 kB(T_0 + T_{E_1}) + G_2 kBT_{E_2}}{G_2 G_1 kBT_0} = \underbrace{1 + \frac{T_{E_1}}{T_0}}_{F_1} + \underbrace{\frac{T_{E_2}}{G_1 T_0}}_{\frac{F_2 - 1}{G_1}}. \tag{5.37}$$

The calculation of the noise figure (factor) of a cascade of amplifiers as shown in Fig. 5.7 is one of the essential skills of an RF engineer. Assume a cascade of amplifiers with gain G_k and noise figure F_k, where k is the integer designating the stage number. Both gain and noise factors are given in the linear power domain (not as dB numbers). The overall gain of K cascaded amplifiers is, of course, simply

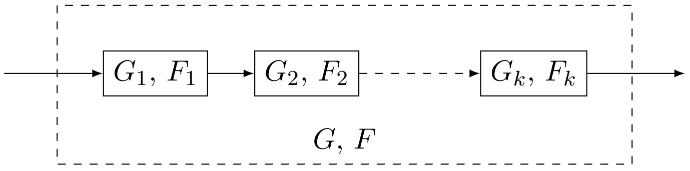


Figure 5.7 Gain and noise figure of a cascade of blocks.

$$G = G_1 \cdot G_2 \cdots G_K. \tag{5.38}$$

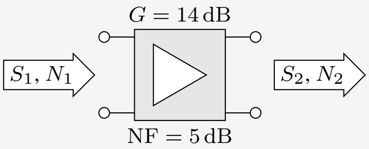
The noise factor is computed as

$$\boxed{F = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + \ldots + \frac{F_K - 1}{G_1 G_2 \cdots G_{K-1}}}\,. \tag{5.39}$$

Eq. (5.39) is also known as *Friis' formula*. The corresponding noise figure in dB is obtained via Eq. (5.34).

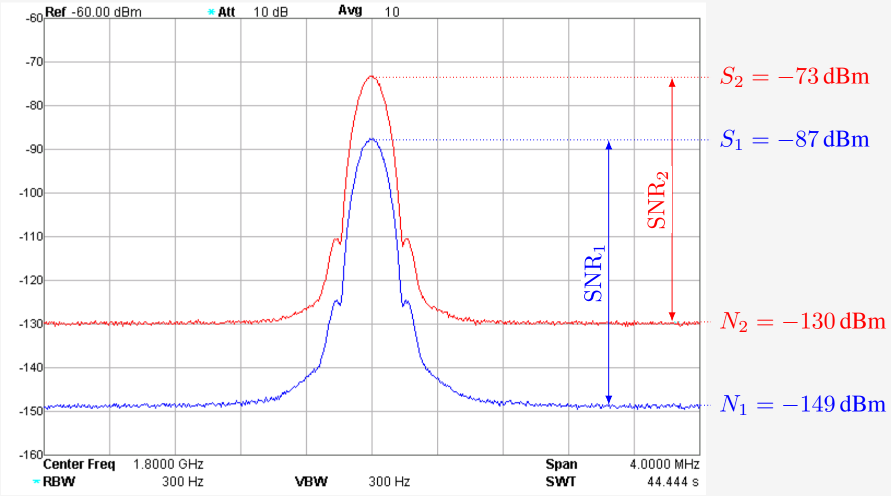
Example: Gain and Noise Figure of an Amplifier

A GSM signal (GMSK, BT $= 0.3$, 200 kHz, see Section 10.4) centered around 1.8 GHz is amplified by an amplifier with $G = 14\,\text{dB}$ gain and a noise figure of $\text{NF} = 5\,\text{dB}$.



The measured spectra before and after amplification are shown below. Note that an RBW of $B = 300\,\text{Hz}$ was used, which increases the detected noise floor level to

$$N_{\text{min}} = N_0 + 10\,\log_{10} B = -174\,\text{dBm/Hz} + 24.7\,\text{dB-Hz} \approx -149\,\text{dBm}\,.$$



As can be seen, while the signal level increases by the gain of the amplifier,

$$S_2 = S_1 + G = -87\,\text{dBm} + 14\,\text{dB} = -73\,\text{dBm}\,,$$

the noise floor level increases from $N_1 = N_{\text{min}} = -149\,\text{dBm}$ by the gain *and* the noise figure of the amplifier:

$$N_2 = N_1 + G + \text{NF} = -149\,\text{dBm} + 14\,\text{dB} + 5\,\text{dB} = -130\,\text{dBm}\,.$$

Thus, the noise level increased even more than the signal level and the respective SNRs before and after amplification are

$$\begin{aligned}\text{SNR}_1 &= S_1 - N_1 = -87\,\text{dBm} - (-149\,\text{dBm}) = 62\,\text{dB}\\ \text{SNR}_2 &= S_2 - N_2 = -73\,\text{dBm} - (-130\,\text{dBm}) = 57\,\text{dB}\,,\end{aligned}$$

meaning the amplification decreased the SNR by the noise figure of the amplifier:

$$\Delta\text{SNR} = \text{SNR}_2 - \text{SNR}_1 = -5\,\text{dB} = -\text{NF}\,.$$

Note that in general such measurements are affected by the noise figure of the measurement equipment as well, according to Friis' formula Eq. (5.39); in this case, that is negligible.

Noise Figure of a Passive Component

The noise figure of a passive component, e.g., cable, attenuator, or filter, is always its loss. Rather than multiplying the noise figure of the next stage by its gain, it is divided by the loss. Hence, the noise factor of a cascade of two amplifiers with some passive component with loss L inbetween, as depicted in Fig. 5.8, is given by

$$\boxed{F = F_1 + \frac{L-1}{G_1} + \frac{(F_2-1)L}{G_1}}. \tag{5.40}$$

The overall gain in this case is

$$G = \frac{G_1 \cdot G_2}{L}. \tag{5.41}$$

As can be seen from Friis' formula, the first device should have the lowest noise figure and a high gain, since the noise figure of any subsequent stage will be divided by the gain of the preceding stages. If saturation is not a concern, amplifiers should always be placed before an attenuator.

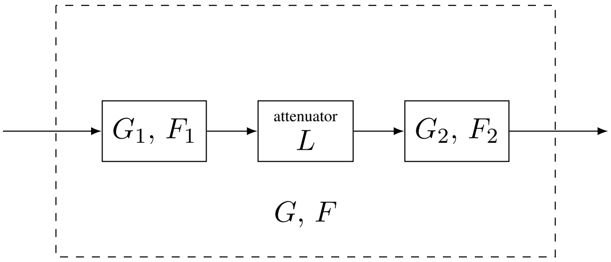


Figure 5.8 Gain and noise figure of a cascade with lossy blocks.

Flicker Noise

Another source of noise plays only a role at lower frequencies. This is so-called *flicker noise* or $1/f$ noise. There is an increase of the noise factor when going to lower frequencies,

$$F = F_0 \cdot \left(\frac{f_c}{f} + 1\right), \tag{5.42}$$

see Fig. 5.9. The corner frequency depends on the semiconductor technology. Some typical values can be found in Table 5.2.

Element	**Corner frequency f_c**
Si-bipolar transistors	100 Hz to 1 kHz
Si-MOSFET	100 Hz to 1 MHz
GaAs-MESFET	1 MHz to 50 MHz

Table 5.2 Corner frequencies of some semiconductor technologies.

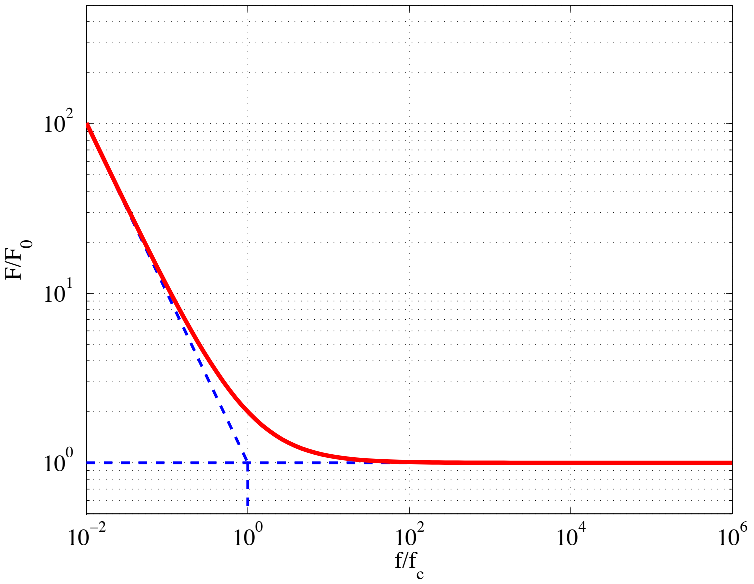


Figure 5.9 Flicker noise influence on noise spectrum.

Noise Matching vs. Power Matching

There is one important remark on the noise figure yet to be made. Further up we have discussed the right matching to achieve the highest power transfer. This matching point, however, is not necessarily the optimum point with respect to the noise figure. The two optimum points (the one for optimum power transfer and the one for optimum noise figure) may be collocated in the Smith chart, but are seldom identical. Thus, if the noise figure is important, it is important to match the circuit for the optimum noise figure and not its complex conjugate. Similarly to the circles of constant gain, there are noise circles which indicate the match to achieve a certain noise figure.

Noise Circles

Similar to gain circles, we can draw noise circles that show points of equal noise figure in a Smith Chart, i.e., the dependency of the noise figure on the input matching condition. The noise figure of an amplifier in function of its input reflection coefficient Γ is given by

$$F = F_{\text{min}} + \frac{4R_n}{Z_0} \frac{|\Gamma - \Gamma_{\text{opt}}|^2}{|1 + \Gamma_{\text{opt}}|^2(1 - |\Gamma|^2)} \tag{5.43}$$

where F_{min} is the minimal noise figure (linear), Γ_{opt} is the optimal reflection coefficient, and R_n is the noise resistance. Fig. 5.10 shows an example of noise figure circles. In this example, the best input match with respect to the noise figure is given for $\Gamma_{\text{opt}} = -0.19 + 0.15j$.

Noise Measurements

In order to measure the noise figure of a component, dedicated measurement equipment is normally deployed. The noise figure could in principle be calculated by measuring the device's output noise and know-

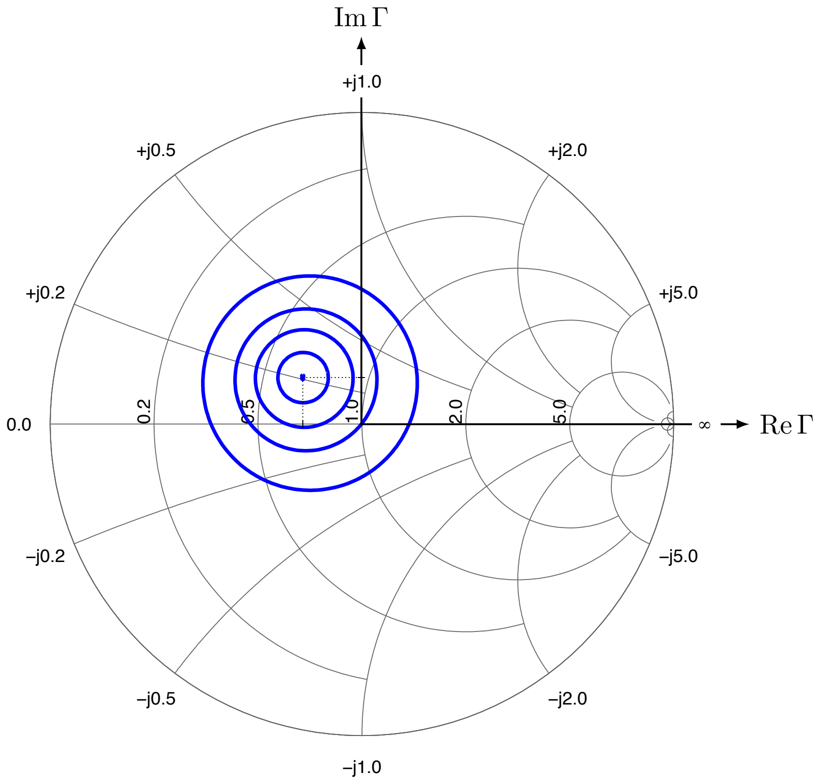


Figure 5.10 Examples of noise circles.

ing its gain and bandwidth. The latter two parameters are not usually known with great precision, so this method is rather inaccurate.

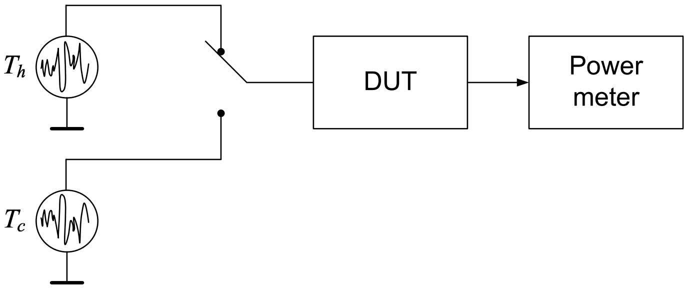


Figure 5.11 Measurement of the noise temperature using the Y-factor. T_h indicates a 'hot' source with a higher-level noise power, while T_c indicates a 'cold' source with a lower-level corresponding noise power.

Instead, the noise figure is measured using a so-called Y-factor measurement. To do so, the component (DUT) is alternately attached to two sources of different noise temperatures (simply two different noise levels), see Fig. 5.11. Now, the power detected at the output of the DUT when connected to the 'hot' source[7] is

$$P_h = kBG(T_h + T_E). \tag{5.44}$$

[7] The terms *hot source* and *cold source* are a relict of the time when the noise source was a resistor put in a heated chamber or in a cooled bath, respectively.

Likewise, the power detected at the output of the DUT when connected to the 'cold' source is

$$P_c = kBG(T_c + T_E). \tag{5.45}$$

The Y-factor is now given as the ratio of Eqs. (5.44) and (5.45)

$$Y = \frac{P_h}{P_c} = \frac{T_h + T_E}{T_c + T_E}. \tag{5.46}$$

The Y-factor can be measured either using a precision power meter, or a precision attenuator, with which the result of one measurement is attenuated until the reading of both measurements are equal. The Y-factor in this case is the difference in the attenuation of the two measurements. Solving Eq. (5.46) for the equivalent noise temperature of the DUT T_E yields independently of the G and the bandwidth B

$$T_E = \frac{T_h - YT_c}{Y - 1}. \tag{5.47}$$

Very often, the relationship between noise temperatures (hot and cold) of a source is given by the *excess noise ratio* (ENR) when the cold temperature is $T_c = T_0$:

$$\text{ENR} = \frac{T_h - T_c}{T_0} = \frac{T_h - T_0}{T_0}. \tag{5.48}$$

Using Eq. (5.32), we can now find

$$F = \frac{\text{ENR}}{Y - 1}. \tag{5.49}$$

5.1.5 Nonlinear Effects

An ideal amplifier would, of course, be linear regardless of the input level. Naturally, no real amplifier can amplify signals at any power level. The supply voltage is an upper bound of what the output voltage can be. At that point, any further increase of the input voltage is basically clipped; this behavior is referred to *gain compression* or *saturation*. Normally, nonlinear effects can be observed much before that point. Further important nonlinear effects to consider are *harmonic distortion* (generation of multiples of the fundamental signal) and *intermodulation* (products of two- and more input signals), which will be briefly discussed in the following.

Gain Compression, Harmonic and Intermodulation Distortion

The *1 dB compression point* is the input power of an amplifier at which the output is 1 dB lower than it is supposed to be, if it were ideal. See Fig. 5.12 for an illustration. Gain compression can be explained using a cubic nonlinearity as the transfer function of a real amplifier

$$v_{\text{out}} = a_1 v_{\text{in}} - a_3 v_{\text{in}}^3. \tag{5.50}$$

Using the trigonometric identity

$$\sin^3 \alpha = \frac{3}{4}\sin\alpha - \frac{1}{4}\sin 3\alpha, \tag{5.51}$$

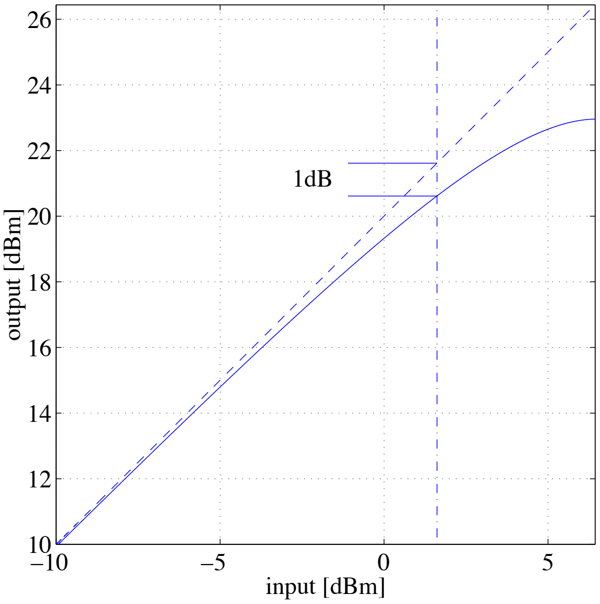


Figure 5.12 Output power vs. input power of an amplifier with gain $G = 20\,\text{dB}$. The 1 dB compression point is at 1.6 dBm.

we find the output for an input signal $v_{\text{in}} = A \sin \omega t$ as

$$v_{\text{out}} = A\left(a_1 - \frac{3}{4}a_3A^2\right)\sin\omega t + \frac{1}{4}a_3A^3\sin 3\omega t. \tag{5.52}$$

Whereas the first part of Eq. (5.52) builds the gain compression, the second is the third-order contribution. In this simple case, the term created is called *harmonic distortion*. Since the frequency is much higher (three times) than the fundamental, it can easily be filtered away. If we have a more complicated input signal consisting of several frequency components, we do not only produce harmonic distortion, but also *intermodulation distortion*, terms that contain sums and differences of multiples of the input frequencies. They may be much more difficult to filter, since they often fall in the band of interest. Consider as an example two input frequencies f_1 and f_2. The output corresponding to an input signal of the form $v_{\text{in}} = A_1 \cos\omega_1 t + A_2 \cos\omega_2 t$ is

$$\begin{aligned}
v_{\text{out}} = {} & a_1A_1\cos\omega_1 t + a_1A_2\cos\omega_2 t - \frac{3}{4}a_3A_1^3\cos\omega_1 t - \frac{3}{4}a_3A_2^3\cos\omega_2 t \\
& - \frac{1}{4}a_3A_1^3\cos 3\omega_1 t - \frac{1}{4}a_3A_2^3\cos 3\omega_2 t - 3a_3A_1^2A_2\cos^2\omega_1 t\cos\omega_2 t - 3a_3A_1A_2^2\cos\omega_1 t\cos^2\omega_2 t \\
= {} & a_1A_1\cos\omega_1 t + a_1A_2\cos\omega_2 t - \frac{3}{4}a_3A_1^3\cos\omega_1 t - \frac{3}{4}a_3A_2^3\cos\omega_2 t \\
& - \frac{1}{4}a_3A_1^3\cos 3\omega_1 t - \frac{1}{4}a_3A_2^3\cos 3\omega_2 t \\
& - \frac{3}{2}a_3A_1^2A_2\cos\omega_2 t - \frac{3}{2}a_3A_1^2A_2\cos 2\omega_1 t\cos\omega_2 t \\
& - \frac{3}{2}a_3A_1A_2^2\cos\omega_1 t - \frac{3}{2}a_3A_1A_2^2\cos\omega_1 t\cos 2\omega_2 t.
\end{aligned} \tag{5.53}$$

Using trigonometric identities,

$$\cos 2\omega_1 t \cos \omega_2 t = \frac{1}{2}\cos(2\omega_1 + \omega_2)t + \frac{1}{2}\cos(2\omega_1 - \omega_2)t \tag{5.54}$$

$$\cos \omega_1 t \cos 2\omega_2 t = \frac{1}{2}\cos(\omega_1 + 2\omega_2)t + \frac{1}{2}\cos(\omega_1 - 2\omega_2)t\ , \tag{5.55}$$

we end up with

$$\begin{aligned} v_\mathrm{out} = {} & (a_1A_1 - \frac{3}{4}a_3A_1^3 - \frac{3}{2}a_3A_1A_2^2)\cos\omega_1 t + (a_1A_2 - \frac{3}{4}a_3A_2^3 - \frac{3}{2}a_3A_1^2A_2)\cos\omega_2 t \\ & - \frac{1}{4}a_3A_1^3\cos 3\omega_1 t - \frac{1}{4}a_3A_2^3\cos 3\omega_2 t \\ & - \frac{3}{4}a_3A_1^2A_2\cos(2\omega_1+\omega_2)t - \frac{3}{4}a_3A_1^2A_2\cos(2\omega_1-\omega_2)t \\ & - \frac{3}{4}a_3A_1A_2^2\cos(\omega_1+2\omega_2)t - \frac{3}{4}a_3A_1^2A_2\cos(\omega_1-2\omega_2)t\ . \end{aligned} \tag{5.56}$$

The third-order distortions not only contain the frequency components $3f_1$ and $3f_2$, but also $2f_1 - f_2$ and $2f_2 - f_1$, as illustrated in Fig. 5.13. If f_1 and f_2 are close, these distortions will be very close either to f_1 or f_2.

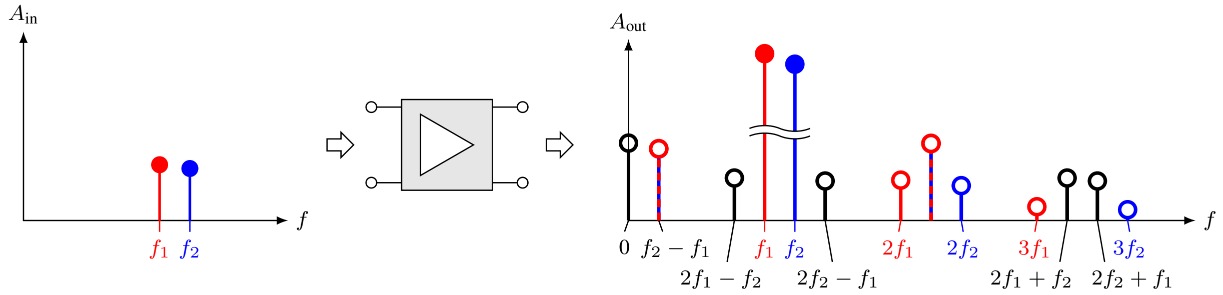


Figure 5.13 Intermodulation distortion due to second- and third-order two-tone products.

Third-Order Intercept Points (IIP$_3$ and OIP$_3$)

As has been demonstrated by Eq. (5.52), the term at the third harmonic grows with the cube of the input level. In a logarithmic plot, the slope of the third-order distortion is three times that of the fundamental. Although for normal operation the third-order harmonic and intermodulation products will be lower than the fundamental, at some point for high input levels the two slopes will cross. This is called the *input related third-order intercept point* or short IIP$_3$. The corresponding point on the output axis is called the *output related third-order intercept point* or short OIP$_3$. This point is ficticious in that it cannot be measured directly, since the 1 dB compression point, which is usually 10-15 dB below the IIP$_3$, prevents the output signal from reaching such levels. Nevertheless, IIP$_3$ is an important amplifier parameter often listed in data sheets, because it allows the derivation of other distortion-related figures of merit. Whereas IIP$_3$ is a parameter given for a certain device, the third-order harmonic distortion HD$_3$ and the IM$_3$ products depend on the actual input powers. Some simple geometrical analysis based on Fig. 5.14 reveals that (for dB values

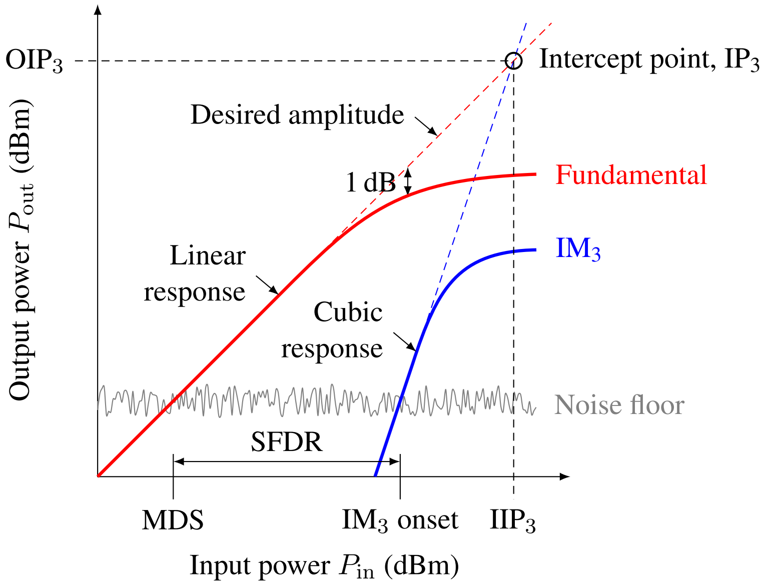


Figure 5.14 Third-order intercept point, spurious-free dynamic range and gain compression.

of the following variables)

$$\boxed{\begin{aligned}\text{IM}_3 &= P_{\text{out}} - 2(\text{IIP}_3 - P_{\text{in}})\\ &= P_{\text{out}} - 2(\text{OIP}_3 - G - (P_{\text{out}} - G))\\ &= 3P_{\text{out}} - 2\text{OIP}_3.\end{aligned}} \tag{5.57}$$

Note that all figures are given in dB or dBm, respectively. Although the IP$_3$ point might be established using a single sinusoidal excitation, it is far more common to measure the IM$_3$ products using two frequency sources (so-called two-tone tests) with the same amplitude but different frequencies and extrapolating to find the (virtual) intercept point.

A similar figure as the IIP$_3$ also exists for second-order effects, the second-order intercept point IIP$_2$, although second-order effects are usually not as strong as third-order effects. Both parameters can also be given as output-related intercept points, OIP$_2$ and OIP$_3$, which are simply the figures obtained by multiplying IIP$_2$ and IIP$_3$ by the gain of the stage. For a cascade of stages, the output-related intercept points can be computed using (the equations are for a three-element cascade, the variables given in linear values)

$$\frac{1}{\sqrt{\text{OIP}_{2,\text{total}}}} = \frac{1}{\sqrt{G_2G_3\text{OIP}_{2,1}}} + \frac{1}{\sqrt{G_3\text{OIP}_{2,2}}} + \frac{1}{\sqrt{\text{OIP}_{2,3}}}, \tag{5.58}$$

where OIP$_{2,k}$ denotes the output-related second-order intercept points of the kth stage, and

$$\boxed{\frac{1}{\text{OIP}_{3,\text{total}}} = \frac{1}{G_2G_3\text{OIP}_{3,1}} + \frac{1}{G_3\text{OIP}_{3,2}} + \frac{1}{\text{OIP}_{3,3}},} \tag{5.59}$$

where OIP$_{3,k}$ denotes the output-related third-order intercept points of the kth stage. The respective input-related intercept points can easily be derived using

$$\mathrm{IIP}_{2,\mathrm{total}} = \frac{\mathrm{OIP}_{2,\mathrm{total}}}{G_1 G_2 G_3}, \tag{5.60}$$

$$\mathrm{IIP}_{3,\mathrm{total}} = \frac{\mathrm{OIP}_{3,\mathrm{total}}}{G_1 G_2 G_3}. \tag{5.61}$$

Note that the variables in Eqs. (5.58) to (5.61) are given in linear values, not in dB.

Spurious-Free Dynamic Range (SFDR)

The *spurious-free dynamic range* is a very important parameter that combines some of the fundamental limitations of amplifiers. SFDR expresses the dynamic range an amplifier can operate over, limited at the bottom by the noise floor and, at the top, by the third-order distortions. In other words, for very weak signals, we have the thermal noise floor, which, together with the SNR needed, determines the minimal detectable signal. If the input signal gets so large that the nonlinear distortions grow larger than thermal noise, the SNR is dominated by the distortions, which means that for every dB of increased input power, the SNR is decreased by 3 dB. The relationship of the different parameters is shown in Fig. 5.14. The parameter MDS designates the minimum discernible signal, which is simply a signal at the thermal noise floor. Using a geometrical argumentation we can see that the spurious-free dynamic range can be computed using

$$\begin{aligned}\mathrm{SFDR} &= \frac{2}{3}(\mathrm{IIP}_3 - N) \\ &= \frac{2}{3}(\mathrm{IIP}_3 - (-174\,\mathrm{dBm} + 10\log_{10} B + \mathrm{NF})).\end{aligned} \tag{5.62}$$

Sometimes, a related figure of merit is given, the *receiver factor*, which is bandwidth independent and expresses the difference between the third-order intercept point and the noise figure

$$\mathrm{RxF} = \mathrm{IIP}_3 - \mathrm{NF}. \tag{5.63}$$

Efficiency

The efficiency of an amplifier expresses the power in the signal delivered to the load divided by the DC power used and depends heavily on the implementation of the amplifier, i.e., the biasing point. These so-called amplifier classes are illustrated in Fig. 5.15. In wireless communications, for all low-noise applications and applications that prefer linearity over efficiency, still class-A and -AB amplifiers are used. However, in most mobile applications, where efficiency is of greater importance (since nonlinear effects can often be mitigated by filtering and signal processing), class-C, -E and -F amplifiers are the most commonly found. Table 5.3 lists the typical efficiencies of some of the more well-known amplifier types.

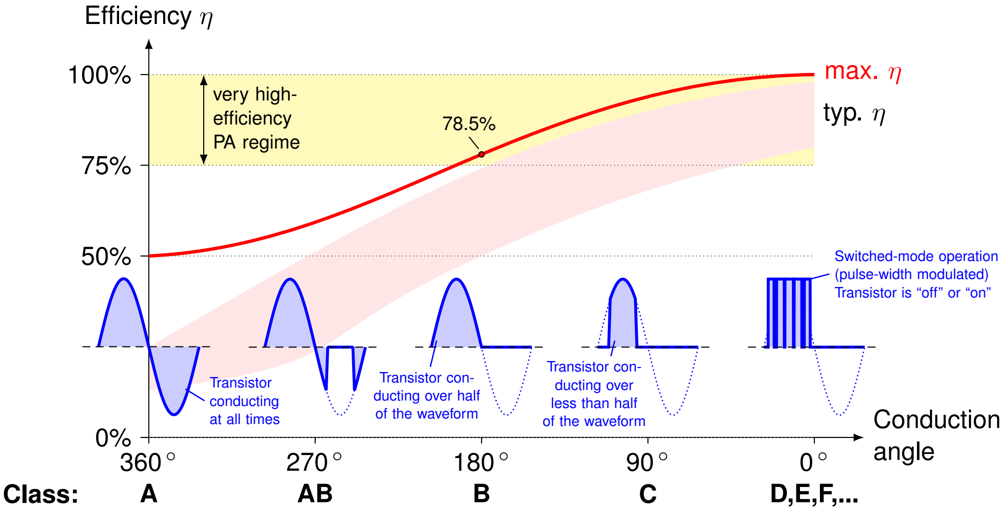


Figure 5.15 Amplifier classes, their conductance angles and maximum efficiencies.

Class	Cond. Angle	Max. Eff.	Typ. Eff.	RF Usage	Remarks
A	360°	50%	15-25%	LNA	strong bias, best linearity
AB	180–360°	78%	25-50%	LNA	biased push-pull
B	180°	78%	<75%	PA	unbiased push-pull
C	<180°	<100%	50-90%	PA	highly nonlinear
D	switched	100%	>90%	–	switched push-pull (PWM)
E	switched	100%	>70%	PA	inductive, low-pass-filtered output
F	<180°	100%	>80%	PA	inductively-biased, harmonic-tuned
F^{-1}	<180°	100%	$\geq$F	PA	"‘inverse-F"’ = inversely harmonic-tuned
EF	switched	100%		PA	harmonic-tuned switched
G				–	adaptive power supply (audio only)
H				–	adaptive power supply (audio only)
I	switched	100%		PA	Interleaved PWM amplifier
J		78%	60-75%	PA	2nd-harm. enhanced comb. of B/C, E & F
S	switched		(<40%)	(PA)	D + $\Delta\Sigma$-modulator (Ex-Sony, research for RF)
T	switched	80–90%		–	Tripath proprietary, enh. AB + D (audio only)

Table 5.3 Conduction angles as well as maximum and typical efficiencies of some amplifier types. Typically, linear amplifiers are inefficient and only used as LNAs, whereas efficient PAs produce harmonics due to their nonlinear behavior. Some amplifier classes are not commonly used for RF/microwave applications (but e.g. for audio).

5.2 Diodes

RF design has many uses for diodes. Depending on these uses, different diodes are chosen. The ordinary PN diode as used in AF circuits has some distinct disadvantages if used in RF circuits. More dedicated diodes are used instead.

5.2.1 Varactor diodes

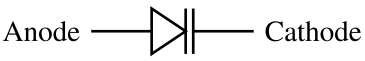


Figure 5.16 Symbol of a varactor diode.

The capacitance in the equivalent circuit diagram of a PN diode depends on the width of the depletion region. The depletion region, in turn, depends on the bias U_b of the diode, for ordinary diodes showing an abrupt doping profile, resulting in a diode capacitance of

$$C = C_0 \left(1 - \frac{U_b}{\Phi_b}\right)^{-\frac{1}{2}}, \tag{5.64}$$

where Φ_b is the barrier potential. The more reverse biased (negative U_b) a diode is, the wider the depletion

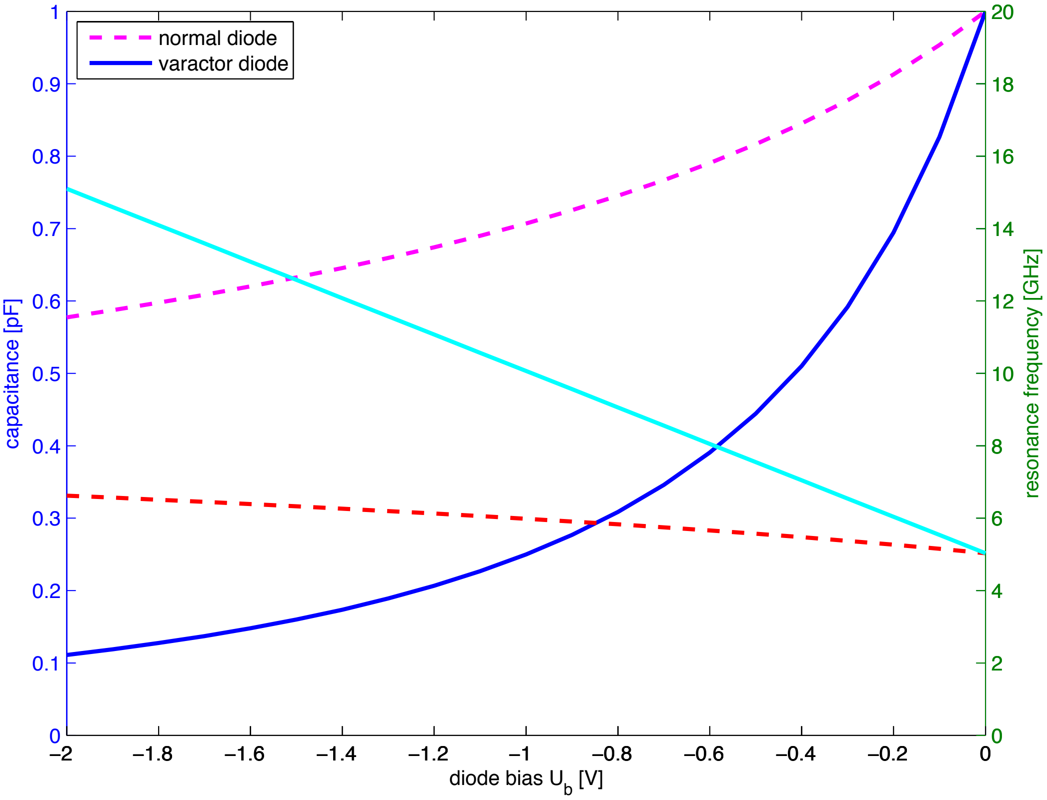


Figure 5.17 Capacitance and resonance frequency related to a varactor diode. For simplicity $n = 2$, $C_0 = 1\,\text{pF}$, $\Phi_b = 1\,\text{V}$, $L = 1\,\text{nH}$.

region is. And, the wider the depletion region is, the smaller the capacitance is. Most capacitance, and still not conducting, is achieved for zero bias. If the doping profile is tuned for maximum capacitance ratio, we get the varactor diode, which has a hyperabrupt doping profile. In this case, the capacitance of the diode can be modelled as

$$\boxed{C = C_0 \left(1 - \frac{U_b}{\Phi_b}\right)^{-n},} \tag{5.65}$$

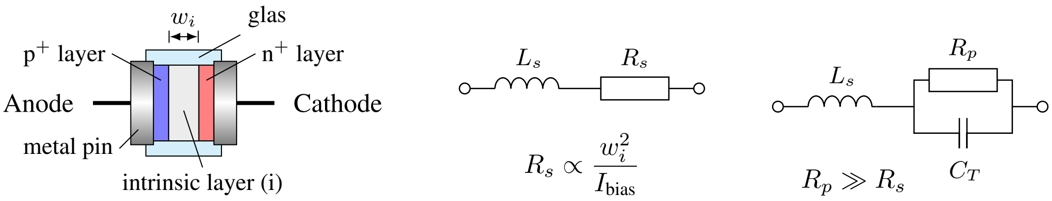
with the exponent n close to 2. Typical values for C_0 vary from 1 pF to 25 pF for silicon varactors. Varactor diodes, also called varicap diodes, and whose symbol is shown in Fig. 5.16, are used in tunable filters, matching circuits, or, most often, in oscillators, making them tunable by adjusting a negative DC voltage bias. A VCO can essentially be built around a varactor diode. Since the resonance frequency

$$f = \frac{1}{2\pi\sqrt{LC}} \tag{5.66}$$

is inversely proportional to the square root of C, we can use this to make a linear tuning range. For a doping profile with $n = 2$ we can show that the resonance frequency of an LC (where the C is built by the varactor diode) is a linear function of the bias voltage, see also Fig. 5.17.

5.2.2 PIN diodes

A PIN diode, see Fig. 5.18, consists of a third region sandwiched between the heavily doped p and n region. This intrinsic region, hence the letter 'i', consists of a large layer with no or little doping. The intrinsic zone exhibits little capacitance in the case of no bias, and a high resistance. The conductivity can become high if a forward bias is applied. This variation, from typically 5–10 kΩ in the no-bias case down to 1–10Ω in the forward-bias situation, make the PIN diode a prime candidate for switches, tunable attenuators, and mixers, see also Fig. 5.19. Its limited recovery time needed for the recombination of p and n carrier make it a rather slow element. This is, however, often desired in RF switching. The RF signal, even when rather high, does not change the bias situation, because the PIN diode is simply too slow.



(a) cross section (b) forward bias (c) reverse bias

Figure 5.18 PIN diode buildup and equivalent circuits.

5.2.3 Schottky diodes

The Schottky diode, sometimes referred to as Schottky-barrier diode, or hot-carrier diode, does not have a PN junction, but a metal-semiconductor junction. The metal side acts as the anode, and n-type semiconductor acts as the cathode of the diode. This typical Schottky barrier results in both very fast switching and

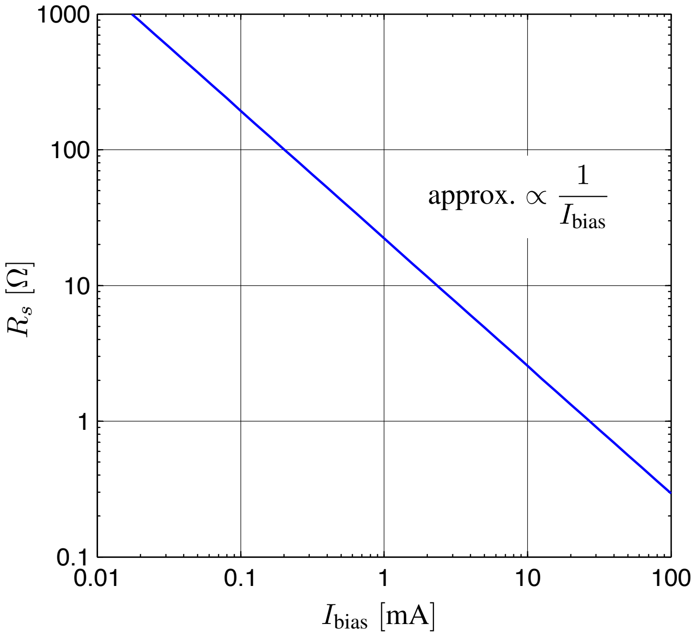


Figure 5.19 Resistance vs. bias current of a Microsemi UM9552 PIN diode.

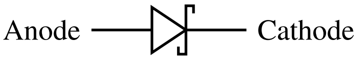


Figure 5.20 Symbol of a Schottky diode.

low forward voltage drop. While the ordinary diode symbol is often used for the Schottky diode as well, a special symbol for it is also in use, see Fig. 5.20.

The lower forward voltage drop compared to ordinary PN diodes is useful for efficient RF detectors. Furthermore, since the Schottky diode only has majority carriers, which are electrons in the n-type diode, the n-type Schottky diode can be very fast, since electrons can be quickly transferred to the conduction band. The Schottky diode has a rather high reverse leakage current, which may pose a problem in certain high-impedance circuits. Trade-offs between some of these parameters can be achieved by either choosing an n-type or a p-type Schottky diode. Diodes using a p-type silicon have a very low barrier, making them well suited for self-biased detectors. Such semiconductors also have very low breakdown voltages (limiting the maximum power) and high parasitic series resistances. Lower series resistances can be achieved by n-type diodes, but they have higher barrier heights. Thus, n-type Schottky diodes are used in biased detectors.

The equivalent circuit of a Schottky diode is given in Fig. 5.21. L_p and C_p are case parasitics. R_s is a chip

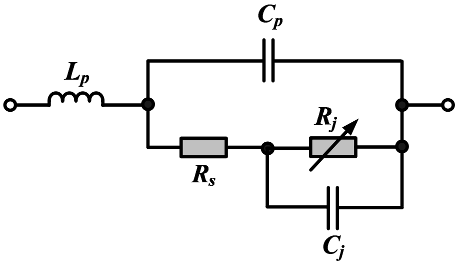


Figure 5.21 Equivalent circuit of a Schottky diode.

parasitic. The junction capacitance C_j and resistance R_j depend on the bias, as we will shortly see. The current through the diode is given by

$$i_d = I_0 \left(\mathrm{e}^{\frac{u_d}{nU_T}} - 1 \right), \tag{5.67}$$

with an n (ideality factor) close to one, and the thermal voltage given by

$$U_T = \frac{kT}{q} \approx 26\,\mathrm{mV}, \tag{5.68}$$

at room temparature. By differentiating Eq. (5.67) with respect to the voltage applied, we note for the inverse of the differential resistance

$$\frac{1}{R_d} = \frac{di_d}{du_d} = \frac{1}{nU_T} I_0 \, \mathrm{e}^{\frac{u_d}{nU_T}}. \tag{5.69}$$

Together with Eq. (5.67) the differential resistance is

$$R_d = \frac{nU_T}{i_d + I_0}. \tag{5.70}$$

I_0 is usually small, so R_d is inversely proportional to any bias we might give the diode. Increasing the current through the diode is thus a popular method to match the diode.

When building a diode detector (for historical reason also called video detector), we face several challenges. For high frequencies, C_j shorts out the junction resistance. All power is now transferred into heat in R_s. Depending on the type of the diode, a little bias may help increasing the sensitivity. Usually such a forward bias is in the region of a couple of μA, but there are diodes (in particular p-type ones) that are used in a zero-bias fashion, also called self-biased mode. The detector diode does not present the RF signal the required load. This problem can be solved in three different ways: Either a forced-matching condition is achieved by placing a 50 Ω resistor in front of the diode. This comes at the price of lower sensitivity, but is rather wideband. Better sensitivity at the price of lower bandwidth is achieved by a tuned detector (placing a matching circuit in front of the diode). As mentioned, increasing the bias may also help to match the diode.

Since the junction resistance is a function of the current through the diode (high current means low resistance), for high currents, the diode seems less and less existent (at least for the positive half-wave). For a detector circuit according to Fig. 5.22 this means that there is a region for the input voltage, where the output DC voltage is essentially proportional to the peak input voltage. Hence, for high RF power, the detector works as a peak detector. The detector is also said to work in the linear region, where the DC output voltage is proportional to the RF input voltage. However, the so-called quadratic region, where the DC output voltage is proportional to the RF input power, is more interesting to the RF designer. This is the case for low input signal (usually up to roughly -20 dBm). In order to analyze the situation, we develop a Taylor series expansion of the voltage-current transfer function given in Eq. (5.67). For a voltage u_d around an operating point $u_{d,0}$

$$u_d = u_{d,0} + \partial u, \tag{5.71}$$

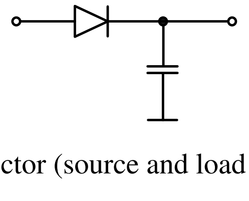


Figure 5.22 Diode detector (source and load impedances not shown).

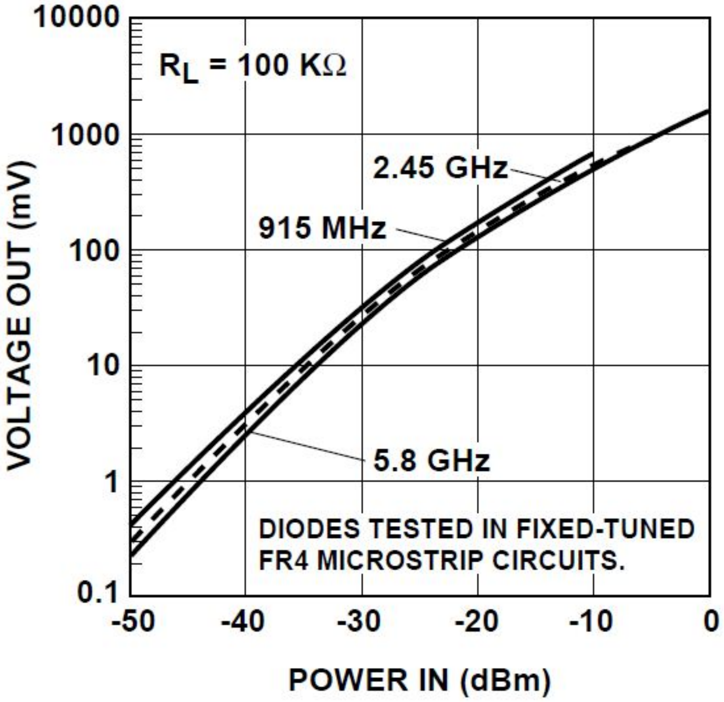


Figure 5.23 Detector curve for a detector based on the diode HSMS-2850. Source: Datasheet by HP.

we can develop the current

$$i_d = i_{d,0} + \partial i = i_{d,0} + \frac{di_d}{du_d}\partial u + \frac{1}{2}\frac{d^2 i_d}{du_d^2}\partial u^2\bigg|_{u_d=u_{d,0}}. \tag{5.72}$$

For an RF signal $\partial u = U_s \cos\omega t$, the diode current becomes

$$i_d = i_{d,0} + \partial i = i_{d,0} + \frac{di_d}{du_d}U_s\cos\omega t + \frac{1}{2}\frac{d^2 i_d}{du_d^2}U_s^2\frac{1+\cos 2\omega t}{2}\bigg|_{u_d=u_{d,0}}. \tag{5.73}$$

The DC part can be identified as

$$\partial i_{d,0} = \frac{1}{4}\frac{d^2 i_d}{du_d^2}U_s^2. \tag{5.74}$$

We now realize that the DC signal is proportional to the squared signal voltage, hence to the power. Fig. 5.23 shows the response curve of a diode detector. The quadratic region for lower signal power might be slightly expanded to higher signals by loading the diode on the DC side. This, however, reduces the sensitivity.

The power absorbed in the diode is

$$P_s = \frac{U_s^2}{2R_d}, \tag{5.75}$$

where R_d is given by Eq. (5.70). By comparing Eqs. (5.74) and (5.75), we can write

$$\partial i_{d,0} = \frac{1}{2}\frac{d^2 i_d}{du_d^2}R_d P_s = \beta_i P_s, \tag{5.76}$$

where β_i [A/W] is called the current sensitivity of the detector. Taking the second derivative of the original transfer function given by Eq. (5.67),

$$\begin{aligned} \frac{d^2 i_d}{du_d^2} &= \left(\frac{1}{nU_T}\right)^2 I_0 \, \mathrm{e}^{\frac{u_d}{nU_T}} \\ &= \frac{1}{R_d} \cdot \frac{1}{nU_T}, \end{aligned} \tag{5.77}$$

the current sensitivity becomes

$$\beta_i = \frac{1}{2} \frac{d^2 i_d}{du_d^2} R_d = \frac{1}{2nU_T}. \tag{5.78}$$

For the ideal Schottky diode ($n = 1$) this is at room temperature β_i =20 A/W. Very often, the voltage sensitivity is looked for. It can be written as

$$\beta_u = \beta_i R_d = \frac{1}{2(i_{d,0} + I_0)}. \tag{5.79}$$

The diode detector is certainly a very simple and thus cheap circuit, but it has its shortcomings. It is not very sensitive. We can see this by defining the noise equivalent power (NEP) as the power that produces the same current in the diode as the noise current given by

$$\langle i_n^2 \rangle = \frac{4kT\Delta f}{R_d}. \tag{5.80}$$

The detected current is

$$i_s = \beta_i \text{NEP} = \frac{1}{2nU_T} \text{NEP}. \tag{5.81}$$

Setting the noise current equal to the detected current results in

$$\frac{1}{2nU_T} \text{NEP} = \sqrt{\frac{4kT\Delta f}{R_d}} \tag{5.82}$$

or, for $n = 1$,

$$\text{NEP} = 4U_T \sqrt{\frac{kT\Delta f}{R_d}}. \tag{5.83}$$

Using Eqs. (5.68) and (5.70), we get

$$\text{NEP} = 4kT \sqrt{\frac{(i_{d,0} + I_0)\Delta f}{q}}. \tag{5.84}$$

If we compare this value with the thermal noise per Hz, which is -174 dBm, for a typical diode value of $I_0 = 100\,\text{nA}$, this is -109 dBm, more than 60 dB less sensitive! This difference is decreased somewhat by moving to higher bandwidths, since the noise current only grows with the square root of the bandwidth.

5.2.4 Tunnel diodes

Tunnel diodes have very thin depletion layers owing to their heavy doping. Thus, the physical effect of tunneling through this thin layer may occur. This results in a non-monotonous voltage-to-current function, see Fig. 5.25. At some points of the curve, a negative differential resistance is obtained, making the tunnel diode usable in oscillator circuits.

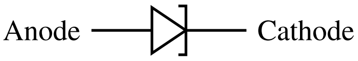


Figure 5.24 Symbol of a tunnel diode.

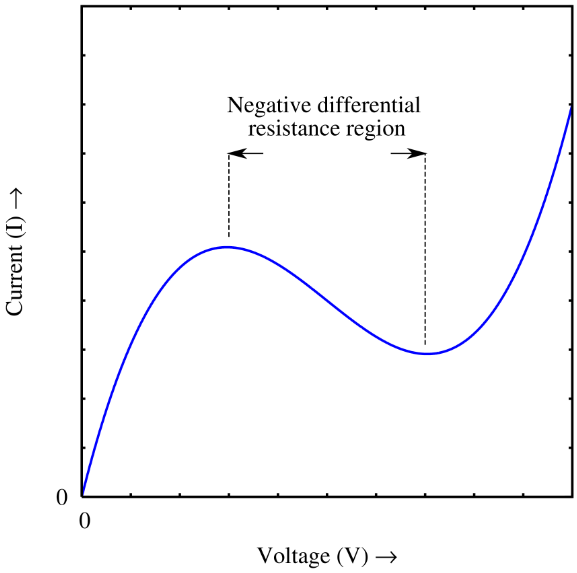


Figure 5.25 Current as a function of voltage with a negative differential resistance region. Source: Wikipedia.

5.2.5 Gunn diodes

A similar negative differential resistance region is obtained with a Gunn diode, although due to a different physical effect, which is seen in gallium arsenide (GaAs) or gallium nitride (GaN). In these semiconductors, there is a third band above the normally top conduction band. This band exhibits lower electron mobility than the conduction band. If, due to a high forward bias, electrons are transferred to this high-energy band, the conductance suffers, producing the negative differential resistance effect. Strictly speaking, the Gunn diode is not a diode, since it has no PN regions. Rather, there is only n-doped material.

5.2.6 Impatt or Avalanche diodes

Impact avalanche transit time (impatt) or simply avalanche diodes show a similar behavior to the Zener diodes, in that they break down and conduct at a specified reverse bias voltage. The reason for this is different, however. Avalanche diodes can be used for noise generators.

5.2.7 Step-recovery diodes

A PN diode similar to a varactor diode is the step-recovery diode. By changing from the forward-bias situation to the reverse-bias situation, the charge held between the doped regions produces a high impulsive current, thus producing many harmonics. A step-recovery diode is therefore often used for frequency multiplication, e.g., triplers.

5.3 Mixers

A mixer is a nonlinear[8] device that can multiply two input signal in a desired way. It can either achieve this directly, if the output is just the multiplication of the two input signals, or indirectly, e.g., the mixer has a square transfer function. In the latter case, the two input signals are provided as a sum to the input of the mixer, which produced the square of each signal (undesired) and the sum and difference of the frequency components.

Consider a signal $s(t)$ with the spectrum $s(t) \circ\!\!-\!\!\bullet\, S(\omega)$. Multiplying it with a sinusoidal carrier ω_0 in the time domain is equivalent to the convolution of the respective spectra in the frequency domain:

$$s(t) \times \underbrace{\cos(\omega_0 t)}_{\frac{1}{2}\left(\mathrm{e}^{j\omega_0 t} + \mathrm{e}^{-j\omega_0 t}\right)} \quad \circ\!\!-\!\!\bullet \quad S(\omega) * \frac{1}{2}\Big(\delta(\omega + \omega_0) + \delta(\omega - \omega_0)\Big) = \frac{1}{2}\Big(S(\omega + \omega_0) + S(\omega - \omega_0)\Big) \tag{5.85}$$

Thus, mixing the signal with a carrier leads to translation by the carrier frequency in the spectrum. In more general terms, mixers always produce the sum and the difference of the input frequencies:

$$f_{\mathrm{out}} = f_{\mathrm{in}\,1} \pm f_{\mathrm{in}\,2} \tag{5.86}$$

The nomenclature for the thee mixer ports are

- the *Local Oscillator* (LO) port,
- the *Radio Frequency* (RF) port, and
- the *Intermediate Frequency* (IF) port,

as shown in Fig. 5.26. The LO port is the only port which is just an input. It is typically driven with either a sinusoidal continuous wave (CW) signal or a square wave signal; the choice depends on the application and the mixer. Conceptually, the LO signal acts as the "gate" of the mixer in the sense that the mixer can be considered "ON" when the LO is a large amplitude and "OFF" otherwise. For that purpose, the LO signal typically needs to carry substantial amount of power, commonly 0 ... 13 dBm or even more.

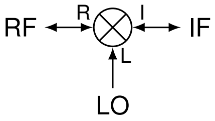


Figure 5.26 Nomenclature and signal flow of the three mixer ports.

The other two ports of the mixer, the RF and IF ports, can serve interchangeably as either the second input or the output, depending on the actual configuration and the application. They are not the same, however, as they typically have different bandwidth/frequency and signal power range limitations.

Most often, mixers are used in one of two distinguishable operating modes:

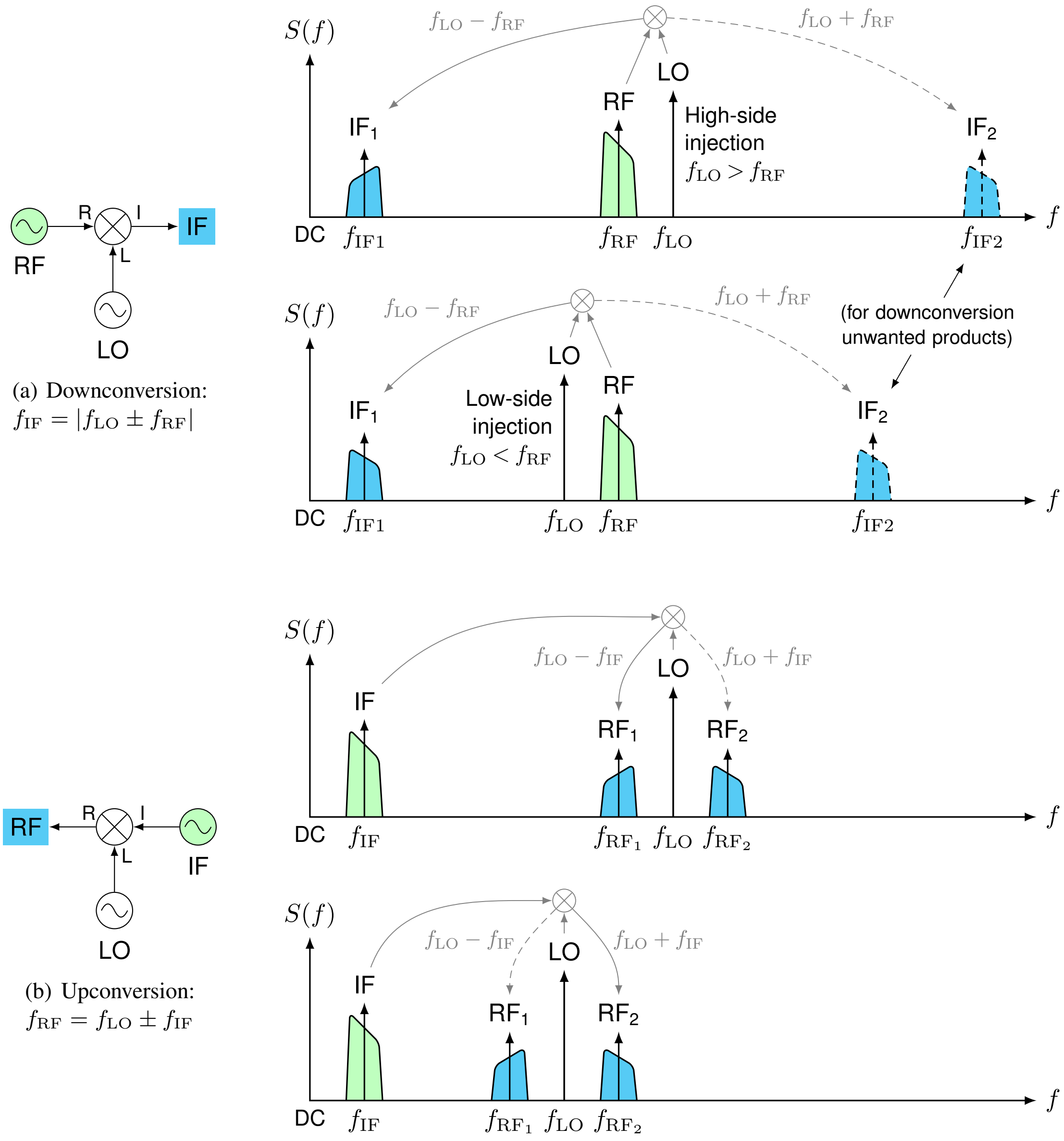
[8] Alternatively, instead of *nonlinear time-invariant*, mixers can also be considered *linear time-variant* devices. Either way, mixers are non-LTI components.

- When the desired output frequency is lower than the two input frequencies, the process is called *downconversion* and the RF port serves as second input, while the IF port is the output. The relationship between input and output frequencies is given by:

$$f_{\mathrm{IF}} = |f_{\mathrm{LO}} - f_{\mathrm{RF}}| \;. \tag{5.87}$$

  As illustrated in Fig. 5.27(a), there are other (unwanted) products, which usually need to be filtered away by an IF filter, as shown in Fig. 5.28. Depending on whether the LO frequency is higher or lower than the RF frequency, the process is called *high-side injection* or *low-side injection*.

- When the desired output frequency is higher than the second input frequency, the process is called *upconversion* and the IF port serves as input, while the RF port is the output. This is illustrated in


$S(f)$
$f_{\mathrm{LO}} - f_{\mathrm{RF}}$
$f_{\mathrm{LO}} + f_{\mathrm{RF}}$
LO
RF
High-side injection
$f_{\mathrm{LO}} > f_{\mathrm{RF}}$
IF$_1$
IF$_2$
DC
f_{IF1}
f_{RF}
f_{LO}
f_{IF2}
f
RF
R
I
IF
L
LO
(for downconversion unwanted products)
Low-side injection
$f_{\mathrm{LO}} < f_{\mathrm{RF}}$
$f_{\mathrm{LO}} - f_{\mathrm{IF}}$
$f_{\mathrm{LO}} + f_{\mathrm{IF}}$
IF
RF$_1$
RF$_2$
f_{IF}
$f_{\mathrm{RF_1}}$
$f_{\mathrm{RF_2}}$

(a) Downconversion: $f_{\mathrm{IF}} = |f_{\mathrm{LO}} \pm f_{\mathrm{RF}}|$

(b) Upconversion: $f_{\mathrm{RF}} = f_{\mathrm{LO}} \pm f_{\mathrm{IF}}$

Figure 5.27 Illustrations of the mixer downconversion and upconversion principles.

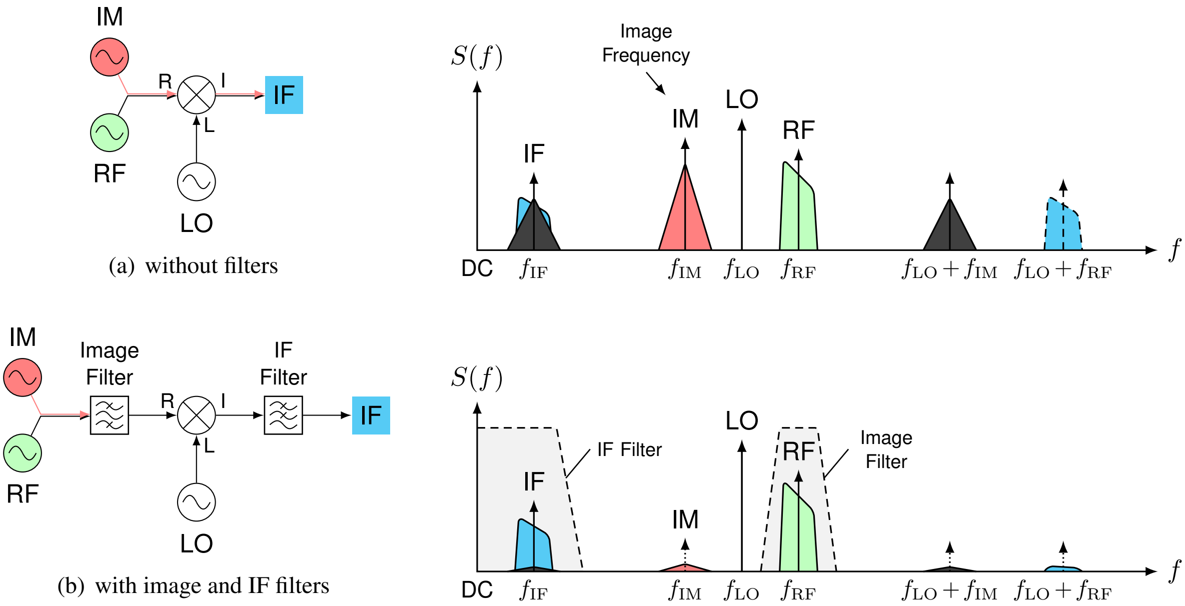
Fig. 5.27(b). This process essentially translates the (two-sided) spectrum of the (real) input signal by the LO frequency and the output frequencies (one or both of them) are given by:

$$f_{\mathrm{RF}} = f_{\mathrm{LO}} \pm f_{\mathrm{IF}} \tag{5.88}$$

Take careful note of the shape ("orientation") of the spectra; depending on the choice of LO frequency, the RF spectrum appears mirrored about the desired RF frequency, similar to the IF spectrum in the downconversion case.

A mixer can thus up-convert or down-convert a signal, when undesired components are filtered off accordingly. This filtering process happens after mixing, by the means of an *IF filter*; in many cases a low-pass would suffice but bandpasses are commonly used to avoid DC issues.

In the downconversion operating mode, an additional filter can be required. Fig. 5.27 illustrates the frequencies of the RF signal, the local oscillator (LO) and the mixing products in the presence of a so-called image frequency, which would also be translated into the IF band. Thus, this image frequency needs be filtered away prior to the mixing process, by an *image filter* (also called image rejection filter), typically a bandpass. Alternatives to image filtering are the use of an image-rejection mixer or the application of a direct-conversion receiver, as shown in Section 6.7.



(a) without filters

(b) with image and IF filters

Figure 5.28 Solving the image problem commonly occurring when using mixers for downconversion.

5.3.1 Additional Mixer Terminology

Besides the limitations on the frequency and power ranges at the RF, LO, and IF ports, respectively, essential parameters of a mixer circuit are similar to the one for amplifiers. Where they are equal in definition, we omit repetition, and concentrate on the differences instead: Instead of gain or amplification, mixers inhibit conversion loss, and the isolation (5.2) is generalized to three ports.

Conversion Loss

For passive mixers, the output power at the IF port is usually lower than the RF power. Instead of this negative gain, the usual terminology is to call this attenuation *conversion loss* (a positive value in dB). Note that the powers occur at different frequencies and are not as straightforward to measure as for amplifiers or attenuators (e.g. using a simple VNA).

Port Isolation

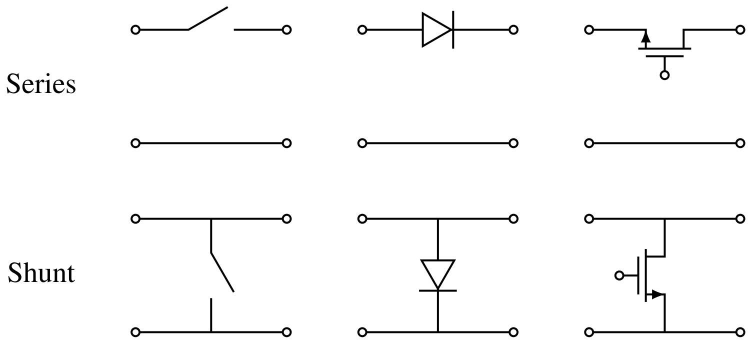
For the three ports of a mixer device, three different isolation figures are usually specified:

- LO-RF isolation,
- RF-IF isolation, and
- LO-IF isolation.

For mixers, where the RF signal is often weak compared to the LO, the second parameter (RF-IF isolation) may not be specified. There are different grades of isolation depending on the type of mixer used.

Mixer Types

There are many ways to build a mixer. The simplest mixer is a *single diode mixer*, also called an *unbalanced mixer*. As the name says, such a mixer consists of a single diode. This is simple, but isolation is very poor between all port; i.e., at the output (IF) port, we find both the original RF signal and the LO signal, besides the desired mixing terms. Moreover, besides the input and the desired terms, typically many other nonlinear intermodulation products appear.

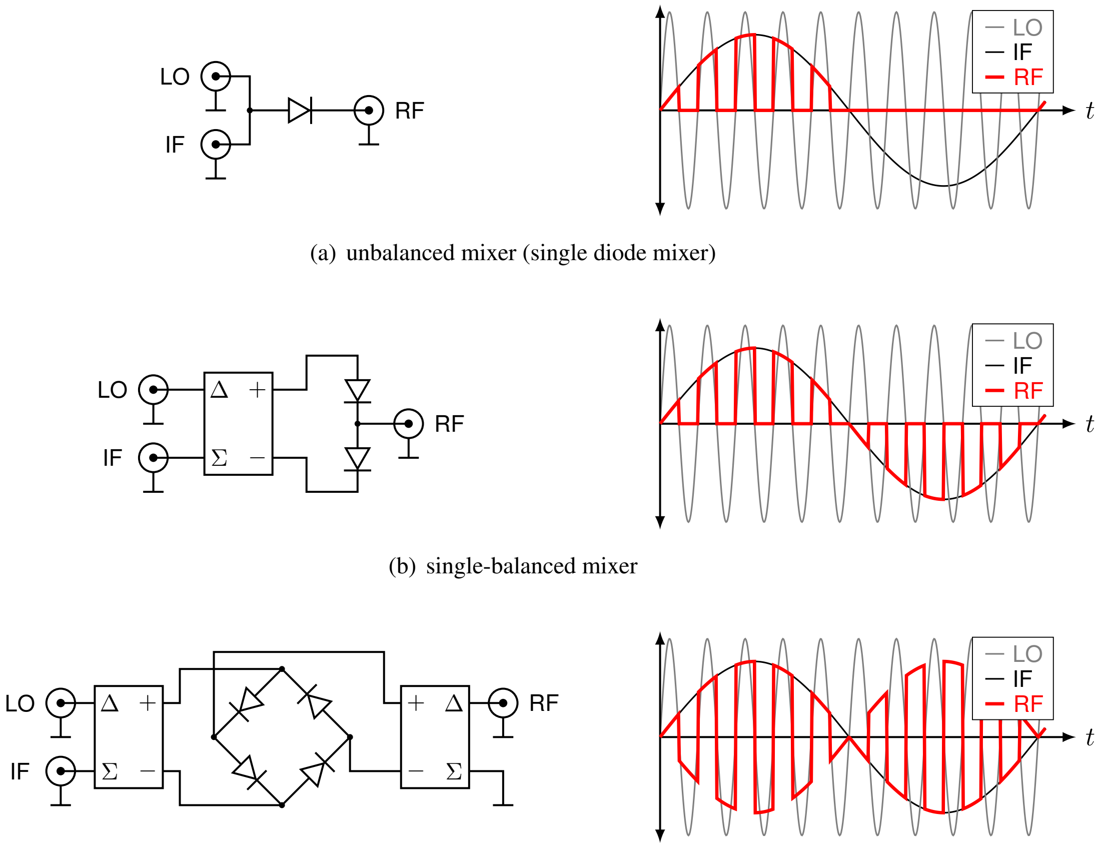


(a) ideal switches (b) using diodes (c) using FETs

Figure 5.29 Switches and their (passive) diode and (active) FET implementations.

Balanced mixers offer several advantages over their unbalanced counterparts: inherent isolation among all mixer ports, cancellation of many/most intermodulation products, common-mode signal rejection, and lower conversion loss (i.e. improved conversion efficiency). The notion of "balancing" refers to the transformation of the typically unbalanced inputs (i.e. the signals are defined against ground) to internal balanced signals (see Section 4.6 on baluns). Besides baluns or hybrids, such implementations consist of at least two diodes instead of one; high performance mixers are designed using four or even eight diodes. As a rule of thumb, more diodes lead to a better mixer but require higher LO power to operate.

- A *single-balanced mixer* (often just called *balanced mixer*) can be realized using two diodes, as shown in Fig. 5.30(b). It essentially includes a balun (or rather a *mixed-mode converter*, see Fig. 4.21) at the LO port and, thus, provides considerably improved LO-RF isolation. Furthermore, intermodulation products are reduced by as much as 50%. However, the RF-IF isolation still remains poor.



(a) unbalanced mixer (single diode mixer)

(b) single-balanced mixer

(c) double-balanced mixer

Figure 5.30 Unbalanced, single- and double-balanced mixers and their signals.

- Two single-balanced mixers can be combined to form a *double-balanced mixer* (DBM), which then contains two baluns and four diodes, as shown in Fig. 5.30(c). Besides the LO port, usually also the RF port is balanced, while the IF remains unbalanced. Such a configuration leads to considerably improved isolation on all ports, reduces intermodulation by up to 75% and further lowers conversion loss (since odd even-order harmonics are avoided altogether). Note that the diodes are arranged in a circular manner, unlike for full-bridge rectifiers.

- Finally, the combination of two double-balanced mixers leads to a *doubly double-balanced mixer*, also known as *triple-balanced mixer*. As the name suggests, this implementation uses balanced LO, RF and IF signals and consists of eight diodes and for unbalanced inputs and five baluns/hybrids. This type of mixer offers the highest isolations (apart from active mixers) as well as lowest intermodulations of any of the mixers. The conversion loss is sometimes higher than for the double-balanced mixers, due to increased loss in the balun and matching circuits.

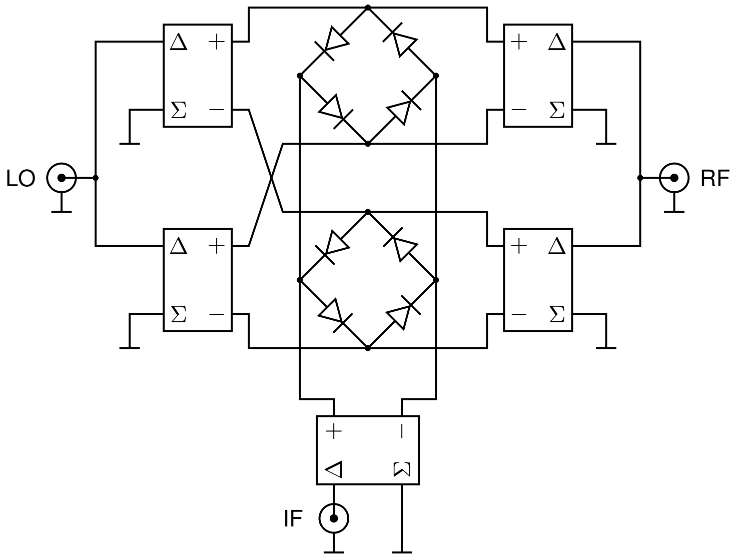


Figure 5.31 Triple-balanced (doubly double-balanced) mixer implementation.

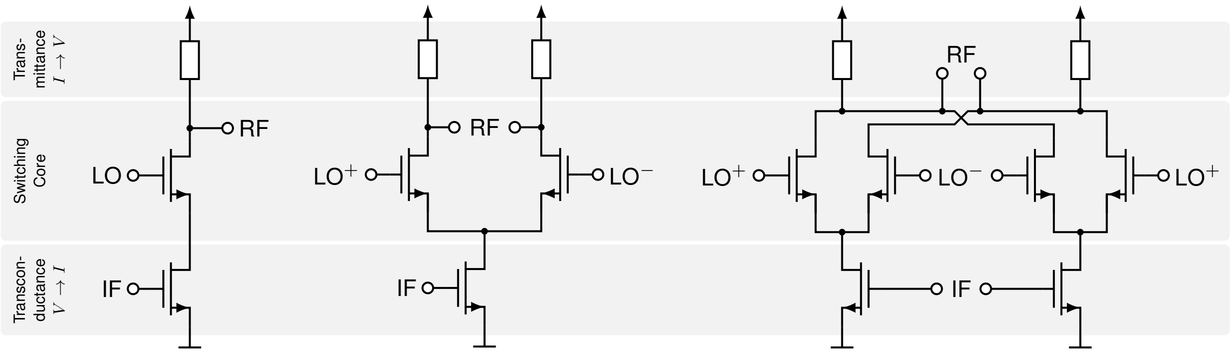
Table 5.4 compares some of the typical characteristics of each of these mixer types.

Mixer Type	**Advantages**	**Disadvantages**
Unbalanced mixer (single diode mixer)	• Simplest mixer	• No port isolation • Min. conversion loss 3.9 dB • Highest spurious and intermodulation products
Single-balanced mixer	• LO-IF (or RF-IF) isolation • up to 50% reduced intermod. • lower conversion loss	• typ. 3 dB higher LO required
Double-balanced mixer (DBM)	• LO-IF and RF-IF isolation • up to 75% reduced intermod. • increased linearity • lower conversion loss • enhanced bandwidth • DC IF/RF	• typ. 6 dB higher LO required • diodes need to be well matched • more involved design • LO and IF/RF cannot overlap
Triple-balanced mixer (TBM)	• excellent isolation (all ports) • lowest intermodulation • most linear behavior • LO, IF/RF can overlap • widest bandwidths	• typ. higher conversion loss than DBM • increased complexity and cost • high LO level (+9 dB) required • no DC IF/RF

Table 5.4 The typical characteristics of the mixer types.

Active Mixers

While this is a passive implementation of a mixer using diodes, Fig. 5.32(c) shows a mixer using active FETs in a so-called Gilbert cell.



(a) Unbalanced mixer (b) Single-balanced mixer (c) Gilbert cell double-balanced mixer

Figure 5.32 FET mixer implementations.

Noise figure

Although noise figure is used with much the same consequences as for amplifiers, there is one subtlety: there are two definitions for the noise figure of a mixer in common usage, the *single sideband noise figure* (SSB NF) and the *double sideband noise figure* (DSB NF). SSB NF assumes a signal from only one sideband, but noise from both sidebands. This is relevant for heterodyne architectures. On the other hand, the DSB NF is used in direct-conversion receivers and includes both signal and noise from both sidebands. Generally, the DSB NF is 3 dB smaller than the SSB NF.

Third-order intermodulation

Although every mixer lives from a nonlinear effect, usually a second-order effect, there are undesired nonlinear effects, mostly third-order intermodulation. The concept of intercept points introduced with amplifier is still valid, although the calculations are usually performed using two RF input frequencies of the same power. This is referred to as two-tone intermodulation. The most troublesome third-order output frequencies are then

$$f_{\mathrm{IF,\,3rd}} = f_{\mathrm{LO}} \pm (2f_{\mathrm{RF}_1} - f_{\mathrm{RF}_2}), \tag{5.89}$$

$$f_{\mathrm{IF,\,3rd}} = f_{\mathrm{LO}} \pm (2f_{\mathrm{RF}_2} - f_{\mathrm{RF}_1}). \tag{5.90}$$

5.4 Oscillators

Although we do not want to get involved in the design of oscillators themselves, it is necessary to understand some key concepts as well as unique design parameters and specifications to successfully choose oscillators for a particular system design. In other words, we do not care *why* the circuit oscillates, but we want to understand the *quality* of its output, its influences and possibilities of corrections and control.

5.4.1 Design Parameters

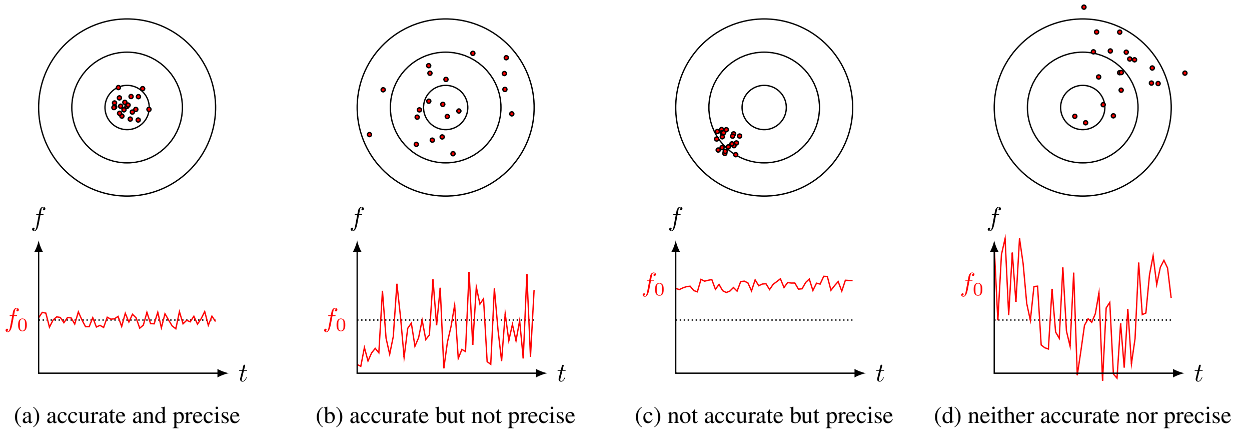
For the design of wireless systems the following oscillator parameters are important:

- actual output frequency,
- frequency stability (e.g. temperature drift),
- phase noise,
- settling time (at frequency changes), and
- harmonics.

However, the most important characteristics of an oscillator can be described by its

- **accuracy,** comparing ist mean output to the actually desired frequency over a longer time, and
- **precision**, regarding the output frequency deviation (over shorter time) from the mean frequency,

of an oscillator. The two quality characteristics are visualized in Fig. 5.33, where the temporal output of an oscillator is compared to a marksman's sequence of bullet holes on a target.



(a) accurate and precise (b) accurate but not precise (c) not accurate but precise (d) neither accurate nor precise

Figure 5.33 Accuracy vs. precision.

Whereas the actual output frequency offset is accounted for in the accuracy criterion, the frequency stability is included in the precision. Depending on the actual time under consideration, phase noise adds to both characteristics: over a short time span it increases the frequency deviation from the mean frequency, whereas for long time spans it can account for a drift in the mean of the frequency. Settling time and harmonics are not accounted for, in this description.

Of course, generally the case as in Fig. 5.33(a) is desirable, but most oscillators perform more like the cases (c) or (d). As will be discussed briefly, frequency standards (e.g. a rubidium frequency standard) itself show a characteristic like (b), but can in combination with a precise (but generally not accurate) oscillator (c) achieve the desired performance (a).

Accuracy

Obviously, the actual output frequency of the oscillator is very important for most applications. An inaccurate frequency can result in timing problems, interference with other channels, inhibit reception, and many more. Generally, the output frequency of oscillators is a function of various variables, such as the temperature of the environment, age of the device, pressure, gravity, sometimes even the output impedance matching[9], etc.

It is custom to specify the frequency accuracy in ppm (parts per million, a factor of 10^{-6}) or ppb (parts per billion, a factor of 10^{-9}), often combining several (or all) effects, so that it is difficult for the user to tell what parts are due to sample variance, aging, and temperature drift, respectively.

Precision

The other figure of merit of an oscillator besides its accuracy is precision. While accuracy specifies the mean frequency (deviation) of the oscillator, therefore, what frequency the oscillator produces in general, the precision states the short time deviation of this frequency.

Depending on the time span in which the precision and accuracy is measured, some of the effects having an impact on the precision over short times also affect the accuracy when observed over longer times. The most crucial effect in this area is the phase noise, as described later. Also other effects can diminish the precision, most importantly temperature dependence, but also sensitivity to mechanical vibration or output power variations.

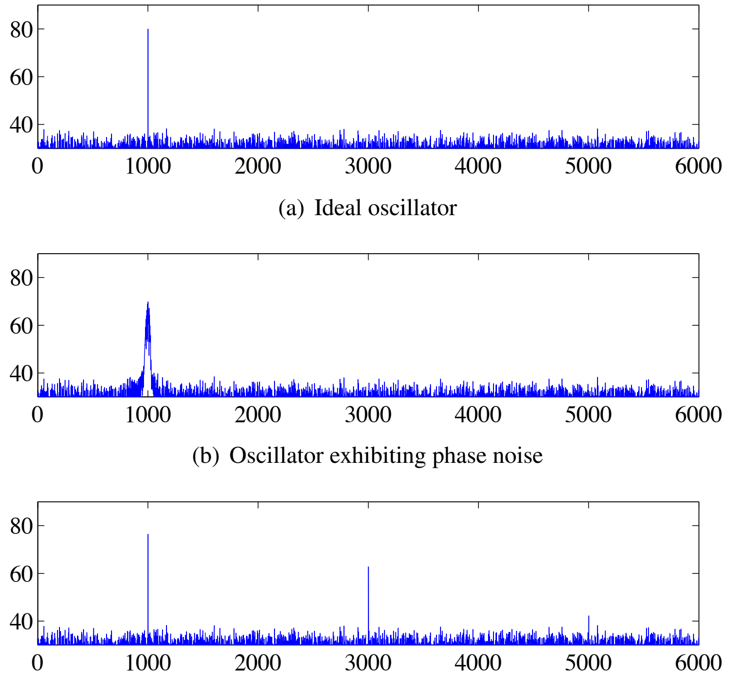
Phase Noise

Since it can affect both accuracy and precision of an oscillator, phase noise is one of the parameters a designer of a communication system is most concerned about. Phase noise is the name used for describing short-term random frequency fluctuations. This non-ideal behavior and nonlinear effects are illustrated in Fig. 5.34.

Although there are many possibilities to quantify phase noise of an oscillator, it is usually specified in the frequency domain. Very often, it is given as single-sideband phase noise in dBc/Hz at a given frequency offset from the carrier. Most definitions state the phase noise $S_c(f)$ as the ratio of power density in one sideband per Hz bandwidth at an offset frequency f away from the carrier and the total signal power. The phase noise stems from thermal noise in the amplifying elements of the oscillator and decay with -20 dB per decade, as shown in Fig. 5.35(a). Since most devices have some amount of flicker noise at low frequencies, depicted in Fig. 5.35(b), the overall spectral shape of the phase noise will be as shown in Fig. 5.35(c). An example of an actual phase-noise measurement of a reference oscillator and its corresponding PLL output can be seen in Fig. 5.42.

The reason why phase noise is so important is illustrated in Fig. 5.36. The process of frequency components mixed by the phase noise part of a LO is called *reciprocal mixing*. When a close, strong interferer is mixed by a noisy LO signal, the resulting IF part may overlap the spectrum of interest. This can degrade the SNR of the receiver, since extra noise is coming in. We speak of desensitization of the receiver due to reciprocal mixing.

[9] This is normally avoided by inserting an isolator at the output of the oscillator, which however results in power loss, increased noise and can even lead to nonlinear interference.

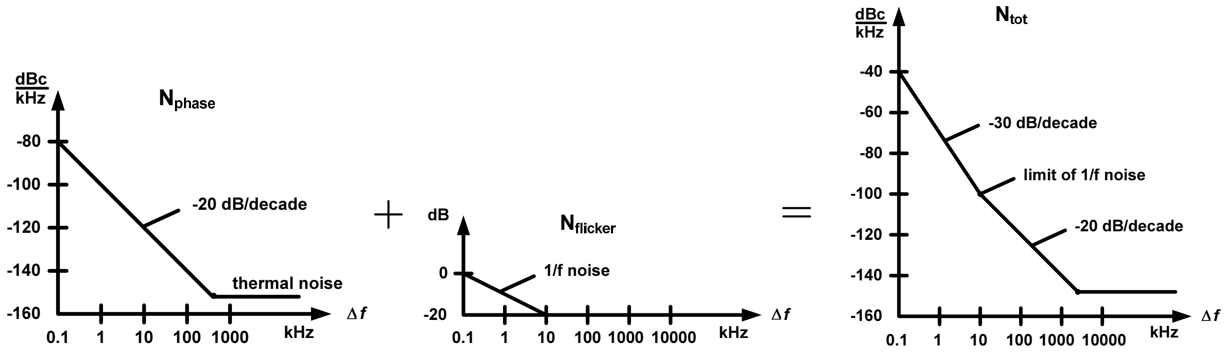


(a) Ideal oscillator

(b) Oscillator exhibiting phase noise

(c) oscillator exhibiting odd harmonics (e.g., clipping)

Figure 5.34 Spectra of oscillator signals.



(a) theoretical phase noise decay

(b) flicker noise spectrum

(c) overall phase noise spectrum

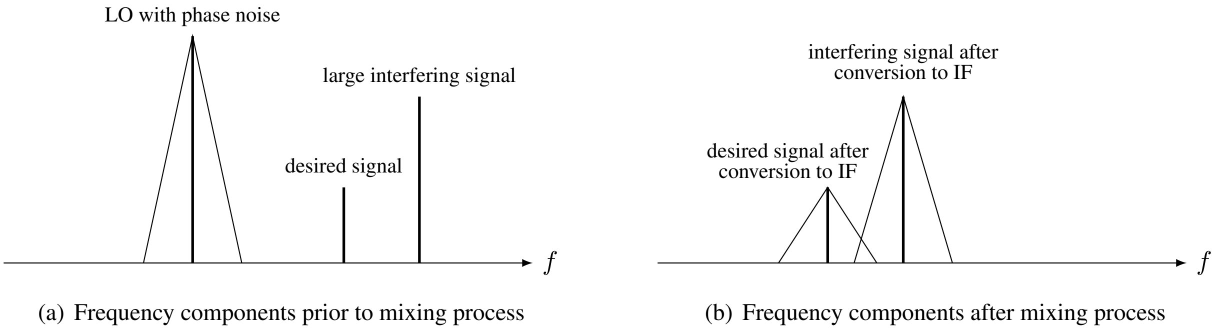
Figure 5.35 Construction of the total phase noise spectrum.

5.4.2 Constant Frequency Oscillators

Quartz Crystal Oscillators

The base of almost all oscillators used in wireless equipments is a quartz crystal oscillator, abbreviated[10] XO, i.e., an oscillator using a quartz crystal resonator as the frequency-selective part. Why a quartz crys-

[10] The abbreviation XO instead of CO has its origin most likely from its military history, because the US Army sponsored most of the research, especially in the early days. There, the abbreviation CO already exists and stands for *commanding officer*. Because "crystal" sounds somewhat like "xtal" (which is nowadays also a common nomenclature), XO was used and later made its way out into the nonmilitary world.


LO with phase noise
large interfering signal
desired signal
f
interfering signal after conversion to IF
desired signal after conversion to IF
f

(a) Frequency components prior to mixing process

(b) Frequency components after mixing process

Figure 5.36 Reciprocal mixing.

tal resonator? Because quartz is the only material known that possesses all of the following properties combined:

- piezoelectric (coupling between mechanical and electrical properties)
- temperature-compensated (even zero temp. coefficient) cuts exist
- stress-compensated cut exist
- high Q / low loss
- easy to process (hard but not brittle)
- abundant in nature and easy to grow in large quantities with high purity and perfection, thus low cost

Cheap quartz crystals show ppm figures of about 20 ppm, which does not seem a lot, but considering 1 GHz of carrier frequency, 20 ppm are already 20 kHz (e.g., almost one channel off in the old days of 25 kHz FM channels). Thus, quartz crystal oscillators are generally not very accurate but compared to other oscillators they show only low phase noise, already making them pretty precise.

However, quartz crystal oscillators are also highly affected by changes in the environment, especially change in temperature (but also mechanical vibration, pressure etc.). Thus, research into XOs, their properties and enhancements, has come a long way, as listed in Table 5.5, and has provided various ways to compensate the temperature-related effects. The three most common are:

- **TCXO** (temperature-compensated XO): A temperature sensor is used to generate a correction voltage, which is then applied to a variable tuning reactance (e.g., a varactor) in the oscillator network. This allows to compensate some of the crystal's $f(T)$ characteristics: Analog TCXOs can provide ten to twenty times improved accuracy and even greater stability improvement over uncompensated XOs.
- **MCXO** (microcontroller-compensated XO): The crystal is excited in both the fundamental as well as the third harmonic overtone mode frequency, from which then the multiplied fundamental is subtracted to receive the so-called *beat frequency* $f_\beta = f_3 - 3\,f_1$. For SC-cut quartz resonators, this beat frequency is a monotonic and nearly linear function of the temperature, which can easily be processed by the microcomputer, together with an adequate control circuit allowing for temperature compensation without needing an actual (external) temperature sensor.
- **OCXO** (oven-controlled XO): For an even better performance the crystal oscillator (along with other temperature sensitive components) is put inside a fully temperature-controlled environment, generally at temperatures somewhat higher than room temperature. This is then called an *oven-controlled*

crystal oscillator, OCXO. Oven-controlled oscillators can provide over a thousand times f vs. T improvement over uncompensated XOs in terms of accuracy and much more in terms of precision, making them highly long-term stable. This of course comes at the cost of decreased electrical efficiency compared to uncompensated crystal oscillators, because of the power needed for the heating process. Also, OCXOs tend to be more bulky and expensive and often have a shorter life-time than other oscillators.

The overall performance of quartz oscillators can be visualized according to picture Fig. 5.33(d): the actual frequency is not hit that accurate, and the precision is highly affected by the environment. However, when the temperature is controlled, the picture changes to Fig. 5.33(c), where the actual frequency is still not very accurate, but it stays the same (high precision) and suffers little phase noise.

If cheap oscillators with such coarse accuracy are employed, very often some kind of frequency correction is mandatory. The reference frequency can be derived from other, relatively more accurate signals, for example from transmitting basestations within a wireless system, it can be derived from a GPS-locked oscillator (GPS Disciplined Oscillator, GPSDO) or it can be obtained directly from a frequency standard, when frequency accuracy is crucial and nothing else is easily available (e.g., in GPS satellites).

5.4.3 Frequency Standards

> *Of all measurement quantities, frequency represents the one that can be determined with by far the highest degree of accuracy.* Fritz Riehle [58]

In fact, the progress in frequency measurements achieved over the past hundred years allows performing measurements of other physical quantities with unprecedented precision, whenever they can be transformed into a some kind of frequency measurements (e.g., lengths or distances).

Year	Milestone
1880	Piezoelectric effect discovered by Jacques and Pierre Curie
1905	First hydrothermal growth of quartz in a laboratory by G. Spezia
1917	First application of piezoelectric effect, in sonar
1918	First use of piezoelectric crystal in an oscillator
1925	K. S. Van Dyke describes the equivalent circuit for a quartz crystal resonator
1926	First quartz crystal controlled broadcast station (Cambridge, Massachusetts)
1927	First temperature compensated (4ppm/°C around 40°C at 50 kHz) quartz cut discovered, amateur radio operators build their own and sell them
1927	First quartz crystal clock built
1934	First practical temperature compensated cut (the AT-cut) developed
1949	Contoured, high-Q, high stability AT-cuts developed
1956	First commercially grown cultured quartz available
1956	First TCXO described
1972	Miniature quartz tuning fork developed; quartz watches available
1974	The SC-cut (and TS/TTC-cut) predicted; verified in 1976
1982	First MCXO with dual c-mode self-temperature sensing

Table 5.5 One century of quartz crystal technology history.

Frequency is undoubtedly closely intertwined with time; it is the inverse of the period of the repetitiveness, the time before the oscillation starts over. Thus, being able to measure frequency accurately is equivalent to being able to tell time exactly. For centuries, frequency standards were based on celestial observations, for example, the earth's rotation rate or the duration of one orbit of the earth about the sun. Time has essentially been defined by the duration of a day and been broken down into smaller, more useful parts, such as hours, minutes, and eventually seconds and beyond, by using faster oscillators, counting those periods (time intervals) and correcting the frequency of the oscillator according to the celestial obersvations when needed.

Thus, in order to be able to tell time accurately down to very fine resolutions, other (faster) frequency standards are desirable. Historically, the development went from pendulum clocks, to spring-based clocks for maritime navigation, and starting in the early 20th century, quartz crystal resonators. However, each of these frequency standards had its limitations: The earth's rotation frequency varies in time, and the frequency stability of mechanical resonators and quartz crystals are limited by environmental effects such as changes in temperature and manufacturing tolerances.

The realization that an atom could be an ideal frequency standard because, as far as we know, one atom is exactly identical to every other atom of the same species, is attributed to none other than James Clerk Maxwell. He stipulated that, if it would be possible to build a device that registers the frequency of a natural oscillation of an atom (e.g., the oscillations of an electron about the atom's core), all such devices would run at exactly the same frequency[11], independent of comparison. Thus, the overall idea of an atomic clock or frequency standard is relatively simple to state: Take a sample of atoms and build an apparatus that resonates to excite one of its natural oscillations. This apparatus then is an atomic frequency standard; to obtain an atomic clock, simply count cycles of this oscillatory signal.

As simple as the idea sounds, at the early stage, only natural oscillations of atoms in the optical frequency region were known. They were known to be very fine lines in the spectrum (the oscillation occurs only in a very narrow neighborhood of the resonant frequency, equivalent to a very high quality factor Q), which is a desired property for accuracy, but proved to be too difficult to observe reliably. Eventually, hyperfine transitions of alkali metals were observed and proposed for use as frequency standards; most importantly rubidium (Rb) at 6.834682610904 GHz and caesium (Cs) at 9.19263177 GHz proved very promising.

Nowadays, the second as defined according to SI, corresponds to exactly the frequency of this hyperfine transition of caesium. Three historic milestones comprise the definition:

- 1967, during the 13. General Conference on Weights and Measures (GCWM), it was decided that, henceforth, the SI second is defined as:

  The duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium-133 atom.

- 1977 it was added that the atomic clocks have to be corrected for relativistic effects (gravitational time dilation due to altitude). Essentially the output of each atomic clock as to be corrected down to mean sea level (MSL), where the second is actually defined.

- 1997 it was refined that:

  This definition refers to a caesium atom at rest at a temperature of 0 K.

[11] Except for relativistic effects, which he might or might not have been aware of, according to different historians and scientists.

> In practice, this means that the frequencies of all primary frequency standards have to be corrected for the shift due to black-body radiation caused by elevated ambient temperature. The definition was finally amended this way in 1999.

Due to the finite width of the spectral line of the hyperfine transition, and due to many other influences and finite measurement accuracy in general, even measuring time according to this definition is not perfect and time can only be measured with finite accuracy. For this reason, time as we know it is currently derived from a weighted combination of over 400 atomic clocks[12] operated by over 50 institutions all over the world. Their combined frequency information is corrected according to multi-frequency GNSS-based (GPS and GLONASS) comparisons for relativistic effects and used to apply corrections to EAL (*Echelle Atomique Libre*, the free atomic scale), via a special iterative algorithm. It essentially relies on H masers for short-term and on Cs clocks for long-term stability. From EAL, TAI (*temps atomique international*, international atomic time) is derived, from which UTC (coordinated universal time) is obtained by adding an integral number of seconds (leap seconds), to account for irregular rotation of the earth[13]. The TAI frequency is further monitored by comparison to thirteen primary frequency standards, including eleven Cs fountain clocks. Its accuracy is commonly quoted as about 2 to 5×10^{-16}, with a precision of about 3×10^{-16} when averaging over a month. Research efforts to increase the long-term stability are put into both developing better time-scale algorithms as well as more accurate & precise frequency standards. In the USA, the NIST-F2 Cs fountain clock should achieve an accuracy of 1.7×10^{-16} soon. However, optical atomic clocks currently under development are expected to surpass fountain clocks and achieve accuracies in the order of 10^{-17} or even 10^{-18}. Thus, currently, our time is prone to an error of about one second in 100 million years and might eventually reach errors of only one second in 30 billion years.

In wireless communication and related areas, and for many applications, what is referred to as “frequency standards” are the three currently commercially available types for atomic frequency standards:

- the rubidium (Rb) standard,
- the caesium (Cs) beam standard, and
- hydrogen masers (active and passive).

They are the reference devices capable of producing (long-term) stable, well known frequencies with a very high accuracy. Fig. 5.37 shows the working frequencies or frequency bands of the most common frequency references, from Earth’s rotation over pendulum clocks to the microwave and optical frequency standards.

The rubidium frequency standard is the most widely used frequency standard today. It is generally the most compact, low-power and most easily transportable as well as least expensive of all the atomic frequency standards. Rb standards are commonly used to control the frequency of television, radio and cell phone base stations, test equipments and many more. However, in many such applications, GPS Disciplined Oscillators (GPSDO) are making their way into the market. They provide similar if not better performance (as long as they are GPS-locked), while requiring less maintenance.

Caesium beam frequency standards are currently the most accurate commercially available frequency standards; they can relatively easily give accuracies of 10^{-12} and beyond. They do not suffer from frequency aging, i.e. change in frequency due to their age, even though their caesium tubes commonly have life-times of only three years (high-precision) or ten years (medium-precision). Their newest addition is the Chip Scale

[12] With about 87%, the large majority are commercial Cs beam clocks (predominantly HP/Agilent/Symmetricom/Microsemi 5071A, with high-performance Cs tubes) or active H masers. On average a little over 10% of the clocks are used with the maximum weight at all times; usually all of those are H masers.

[13] The rotation of Earth is slowing down and as a result the days becoming longer

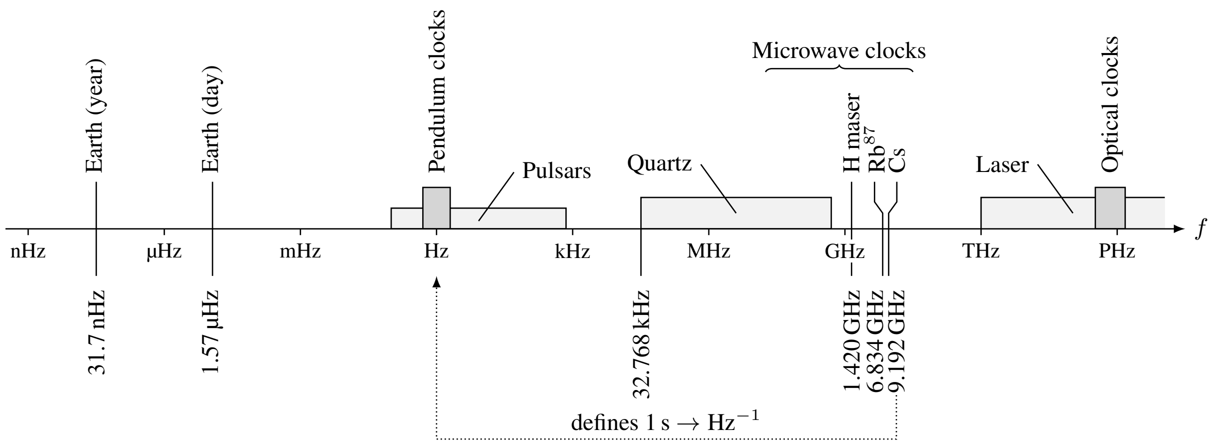


Figure 5.37 Frequencies (ranges) of the most common frequency standards.

Atomic Clock (CSAC). What had started as a US DARPA research program to manufacture a microchip-sized atomic clock for use in portable military equipment for location and battlespace situational awareness whenever GPS is not available in 2001, has recently (2012) become available for civilian applications as well. At about $40 \times 35 \times 10\,\mathrm{mm}$, it is considerably smaller than common rack-mount Cs beams, and at 35 g it is almost a thousand times lighter, as well.

Finally, hydrogen masers (H masers), are currently the most long-term stable commercially available frequency standards; many of them achieve stabilities of beyond 10^{-16} over a few days or months. However, this comes at a price: H masers are commonly the bulkiest, heaviest and most expensive of all of the three atomic standards. It has been a particular challenge to get them ready for applications in space, as they are also know to be susceptable to environmental influences.

Table 5.6 summarizes some of the characteristics of these atomic frequency standards and other constant frequency oscillators for comparison. Fig. 5.38 compares their performance visually, by plotting accuracy vs. short-term precision. Required accuracy levels for some communication and timing standards are given as well, in addition to current and expected accuracy levels of time measurements in general.

Type	**Accuracy**	**Aging (10 years)**	**Power Cons.**	**Weight**	**Cost** ($)
XO (uncomp.)	1 ... 100 ppm	10 ... 20 ppm	$\leq$ 100 mW	0.01 ... 10 g	0.1 ... 10
TCXO	0.1 ... 10 ppm	1 ... 5 ppm	$\leq$ 200 mW	0.03 ... 50 g	0.5 ... 100
MCXO	1 ... 500 ppb	1 ... 3 ppm	$\leq$ 200 mW	0.5 ... 50 g	1 ... 100
OCXO	0.1 ... 50 ppb	0.01 ... 1 ppm	1 ... 5 W	50 ... 500 g	0.1 ... 2 k
Rb	$10^{-11} \ldots 10^{-9}$	$10^{-10} \ldots 10^{-9}$	1 ... 20 W	0.1 ... 2.5 kg	0.2 ... 10 k
Cs chip-size (CSAC)	$5 \times 10^{-11} \ldots 5 \times 10^{-10}$	$\leq 10^{-7}$	$\leq$ 0.5 W	35 g	$\approx$ 5.5 k
Cs beam	$2 \times 10^{-14} \ldots 10^{-12}$	none	50 ... 100 W	10 ... 30 kg	20 k ... 100 k
H maser (passive)	$5 \times 10^{-13} \ldots 10^{-11}$	$10^{-13} \ldots 10^{-11}$	50 ... 100 W	30 ... 50 kg	$\approx$ 50 k
H maser (active)	$3 \times 10^{-13} \ldots \times 10^{-12}$	$10^{-13} \ldots 10^{-12}$	$\geq$ 100 W	$\approx$ 100 kg	$\approx$ 200 k
GPSDO	$10^{-12} \ldots 10^{-11}$	10^{-11}/none	1 ... 30 W	0.2 ... 5 kg	0.3 ... 15 k

Table 5.6 Comparison of the typical characteristics of commercially[14] available constant frequency oscillator and frequency standards.

[14] Cs fountain clocks are not yet commercially available; though there are on-going developments trying to change that. However, fountain clocks will remain very expensive and large, as the resonance line width (essentially narrowing down the stability) depends on the tossing height, i.e. the height of the fountain (usually $> 1\,\mathrm{m}$).

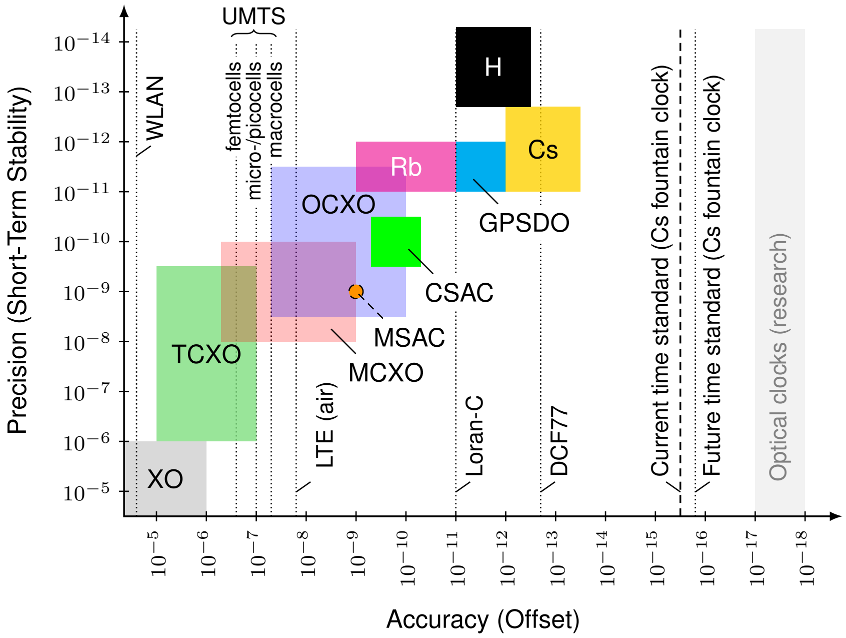
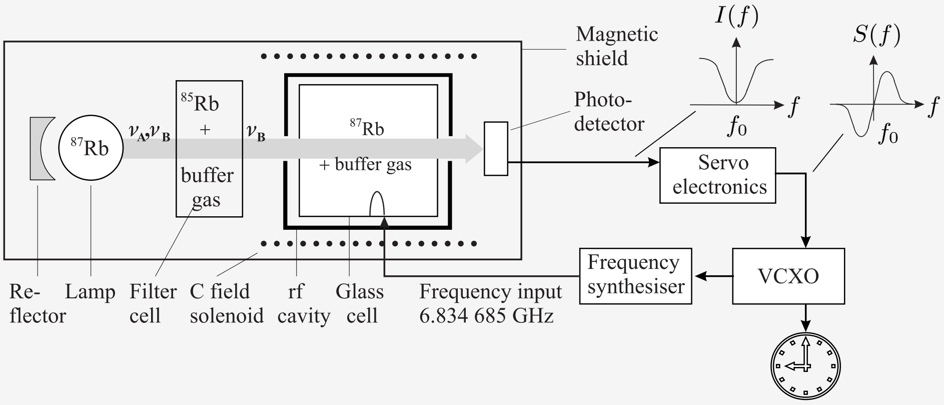


Figure 5.38 Typical accuracies and precisions of constant frequency oscillators and atomic frequency standards, as well as timing accuracy requirements (for transmitters) of some signaling standards.

Working principle of the rubidium frequency standard

The rubidium frequency standard working schematic is shown below: the light of a ^{87}Rb vapor discharge lamp is prefiltered by a ^{85}Rb vapor absorption filter, passed through a ^{87}Rb vapor absorption filter and the remaining light energy is measured by a photo detector.



The remaining light is minimal, when the ^{87}Rb vapor absorption filter resonates between the so-called *hyperfine-split ground state* of the ^{87}Rb isotope, which happens at the given frequency of 6.834 685 GHz (and increased temperature), which is applied from outside. Because the actual level of the minimum is unknown (or rather depends on the actual condition of the lamp and the filters etc.) the control mechanism of the PLL delivering the 6.83 GHz has to always be shifting back and forth over the bell shaped absorption curve, leading to quite high phase noise, but due to the quantum physical nature of the energy states, the minimum always remains at the same frequency, making it long-term stable.

Thus, the Rb oscillator itself can be considered to perform as illustrated in Fig. 5.33(b). But when a TCXO, MCXO or OCXO (iself behaving as Fig. 5.33(c)) is carefully adjusted (tracked) by the mean of the Rb oscillator, an almost ideal performance as symbolized in Fig. 5.33(a) can be achieved.

Atomic Standards used in Space

GPS satellites are equipped with a variety of different Rb and Cs frequency standards; currently, seven different types of frequency standards are operational[15]:

- Block II/IIA satellites (launched from 1989 to 1997) contain two Cs and two Rb clocks.
- Block IIR/M satellites (launched from 1997 to 2009) use three Rb clocks.
- Block IIF satellites (launched from 2010 to 2016) contain one Cs and two Rb clocks each.

The Rb standards initially proved more reliable, while the Cs standards were more accurate and stable, whenever they were fully operational. For the satellite versions Block IIF, improved Rb standards with enhanced stability (using Xenon lamp buffer gas and a thin-film spectral filter) were deployed. In addition, the used Cs standards had been substantially improved compared to the ones of Block II/IIA satellites as well, leading to better short-term stability and increased lifetime. However, it is the Rb clocks in the IIF satellites that are said to perform the best in the GPS constellation.

Current and future GLONASS satellites are also equipped with Rb and Cs frequency standards:

- GLONASS-M satellites (launched from 2001 to 2017) were launched with three Cs standards.
- GLONASS-K1 satellites (launched from 2016 on) and GLONASS-K2 satellites (to be launched from 2018) contain two Rb and two Cs atomic standards, each. The clocks in K2 should be almost a decade more stable than the ones of K1, which is a substantial improvement and might even exceed the performance of the GPS standards.

Contrary to that, the European GALILEO satellite navigation system uses a combination of a Rb frequency standards and passive hydrogen masers: the reliability, accuracy and stability of the Rb frequency standard is thereby enhanced with the even better (four times higher) stability of the H maser and, for redundancy reasons, two items of each are included in each satellite. Under normal conditions, the operating H maser produces the reference frequency, from which the navigation signal is generated. Should it encounter any problem, an instantaneous switchover to one of the Rb clock is performed. Both frequency standards are provided by the Swiss company SpectraTime[16] in Neuchatel[17]. They also provide the frequency standards for the Chinese (BeiDou) and Indian navigation satellite system (IRNSS) and are the self-proclaimed "world's largest Swiss supplier of highly reliable atomic clocks in space".

In press briefing in January 2017, it was announced that three Rb standards and six H masers failed (due to tolerance issues, among other possible reasons). However, due to the quadruple redundancy of the timing subsystem, the overall performance of GALILEO has not been not affected.

5.4.4 VCO – Voltage-Controlled Oscillators

Most wireless applications require that oscillators be tunable, meaning their output frequency has to be adjustable according to some function of a control input. Usually, this control input requires a voltage signal, making these oscillators so-called voltage-controlled oscillators (VCO). Current-controlled oscillators can also be built, but are not as commonly used in RF systems. The most important performance parameters of VCOs are:

[15] InsideGNSS, Nov/Dec 2012, pp. 28-29

[16] previously know as Tekelec Neuchatel Time (from 1995 to 2003) and Temex Neuchatel Time (from 2003 to 2007)

[17] in partnership with Selex Galileo from Edinburgh, Scotland; now Leonardo S.p.A., in Rome, Italy

- **Tuning range:** This is the desired tuning range of the output frequency. Of course the tuning range of the oscillator has to span over the whole frequency range used by the application in the end. But also, the tuning has to be achievable with the given voltage range of the application device. This requires sufficient sensitivity on the control signal, called VCO gain.
  The tuning range has to be chosen (somewhat) wider than the actual frequency span of the application, to account for process and temperature variations. On the other hand, if the sensitivity of the frequency on the control voltage is high (to achieve a wide tuning range), noise on the control input will create increased phase noise. Thus, the tuning range should be chosen as wide as necessary, but as narrow as possible, to achieve best overall performance in the end.
- **Tuning linearity:** Ideally, the output frequency would be a linear function of the control input. However, this is never the case in real-life applications; practical VCOs always exhibit a nonlinear tuning characteristic. This may have to be considered, especially for closed-loop applications, such as PLLs.
- **Spectral purity:** There always appear some sidebands around the wanted oscillation frequency, due to imperfect phase noise (or jitter) performance as well as unwanted nonlinearities (e.g., clipping, etc.). Both phase noise and harmonic levels are used to quantify the spectral purity of an oscillator.
- **Efficiency:** The VCO is usually the most power consuming part in a PLL system – maybe even of the whole receiver. This is especially important for mobile communication applications. However, low-power VCO design involves trade-offs in almost all other parameters.

Integrated VCOs

While in the early 80's, when mobile communications was at its beginning, all RF circuits were built using GaAs semiconductor technology, today's RF front-ends (especially for the consumer marked) are more and more realized in CMOS, which is much less expensive and opens the way for single-chip solutions (often called *system on a chip*, SOC). Thus, in nowadays mobile phones, only the power amplifiers are still made of GaAs semiconductor material.

Single-chip wireless transceivers include both analog and digital signals on the same substrate. Digital circuits inject noise into the substrate through their switching nature. The noise can then couple between blocks over the substrate and the power supply lines as well as over the signal lines. Noise coupled to the VCO control line (as well as to the power supply and bias lines) is converted into phase noise through FM, AM and AM-FM conversion principles. Thus, fully integrated communication systems have to be designed very carefully in order to minimized these effects – where, of course, suitable trade-offs have to be found in practice. Therefore, it comes as no surprise that there is still a lot of (basic) research going on in this area, especially for multiband and broadband applications.

Actual realizations of such VCOs are beyond the scope of this lecture. However, it should be kept in mind that all practical oscillators are non-ideal, meaning that they might produce considerable amounts of harmonic, disharmonic and spurious signal parts as well as phase noise – depending on the output power, size, temperature and other influences.

Gunn Oscillators (Relaxation Oscillators)

Gunn oscillators utilize the bulk negative conductance properties of the semiconductor material (GaAs, GaN or InP) to convert DC into RF, microwave and millimeter-wave frequencies using some kind of resonator (mostly waveguide and dielectric resonators). Gunn oscillators are available for low to medium output power in the range of a few GHz to over 100 GHz.

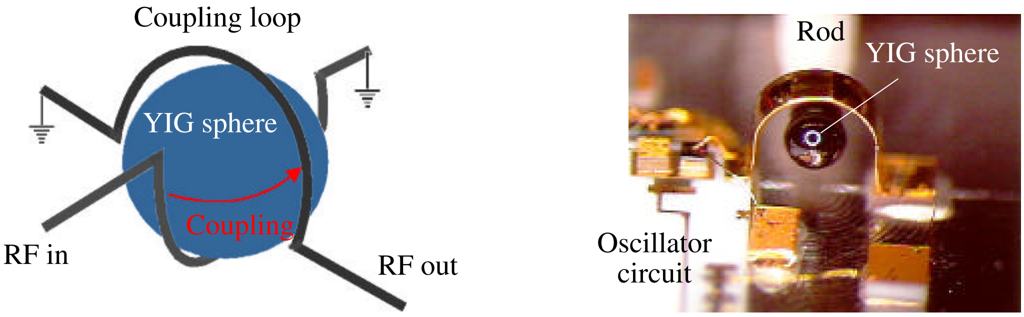
The Gunn diode has to be driven in reverse bias mode, where with the bias current the actual output frequency can be adjusted. One of their main problems is that due to insufficient power supply voltage stability, Gunn oscillators can produce ripples on the supply lines, resulting in both AM and FM phase noise. Also, they feature almost linear frequency-to-voltage tuning but suffer from high temperature dependency (negative, in the order of some MHz/K). Also, for their spectral purity the resonator is of high importance, making them sensitive to mechanical influences.

Gunn oscillators are used for many purposes, especially as medium power RF source in sensor and measuring applications of frequency regions reaching from 6 to about 24 GHz, such as local oscillators, transmit and receive oscillators for radio communications; military, commercial and police radar sources as well as various sensors for measuring velocity, proximity, fluid levels; detecting directions, collision avoidance and intelligent cruise control, and others.

Yttrium Iron Garnet Oscillators (YIG, YTO)

For wideband electrically tunable oscillators (e.g. in modern high quality spectrum analyzers), Yttrium Iron Garnet (YIG) or varactor resonators are used. YIG is a synthetic ferrimagnetic garnet (dt. *Granat*) with the chemical composition $Y_3Fe_2(FeO_4)_3$, Yittrium (Y) being a rare earth element. The YIG resonator is a ferrite sphere of single crystal YIG material, fractions of a millimeter in diameter, whose ferrimagnetic resonance can be tuned over a wide frequency band (several octaves) by varying the biasing DC magnetic field.

Typically, a YTO (YIG Tunable Oscillator) consists of a YIG sphere, two coils (one for the main magnetic field and one for fine adjustments such as phase locking or for FM purposes), a heater or temperature control (optional, to control frequency drift of the YIG sphere and due to the temperature-dependent coil resistance, increasing the overall performance) and a microwave oscillator circuit coupled to the sphere via loop-shaped lines around the sphere (both single and double loop configurations are possible, as shown in Fig. 5.39). The relationship between frequency and current is the YTO's tuning sensitivity, expressed in MHz/mA. For example: if the oscillator has a 20 MHz/mA tuning sensitivity, then 100 mA will tune the oscillator to 2.0 GHz, whereas 500 mA will put it to 10 GHz output frequency. YIG resonators show very high Q-factors, providing very low phase noise, and can be tuned in wide bands (e.g., 3 to 50 GHz with a single resonator). For even wider tuning ranges, several coupled YIG filter stages are used, each stage consisting of a sphere and its loop(s).



(a) YIG sphere coupling principle (with two loops, also single loop oscillator circuits are possible)

(b) YIG resonator in place in the oscillator: the YIG sphere (on top of a ceramic rod) under the (single) loop line

Figure 5.39 YIG sphere coupling and placed in the oscillator.
Source: `www.microlambdawireless.com` with additions.

The high performance (e.g., low phase noise) and convenient size of YIG resonators for applications in microwave integrated circuits make it an excellent choice in a large number of applications, such as filters, multipliers, discriminators, limiters, and oscillators in frequency ranges extending from about 500 MHz to over 50 GHz. Unfortunately, building YIG oscillators requires very specific skills and experience. And since they are manufactured by just a hand full companies worldwide, very little information is available about their actual internal workings and specialties.

5.4.5 NCO – Numerically Controlled Oscillators

A *numerically controlled oscillator*, NCO, (sometimes called *digitally controlled oscillator*, DCO) is a device which approximates an oscillation (most commonly but not necessarily a sine wave) based on some time base (reference clock) with a finite digital resolution.

As shown in Fig. 5.40, an NCO generally consists of two parts:

- a phase accumulator, adding the output values at each clock edge
- a phase-to-amplitude converter, providing the corresponding amplitude value to the added up phase.

Also, optionally an interpolation filter is used to smooth the output signal, thus increasing the accuracy while reducing phase noise.

The heart of the NCO is the phase-to-amplitude converter, usually a look-up table, LUT (sometimes in combination with an iterative algorithm to reduce the size of the LUT), which approximates the function looked for, usually sine and cosine wave. Two bit register lengths, usually denoted M and N define the resolution of the NCO:

- M bit-width of the phase address of the look-up table (LUT)
- N bit-width of the DAC actually building the signal.

Whereas the pure NCO delivers varying digital signal values over time, for some (or rather most) applications an analog signal is required. There are several methods to accomplish this, the simplest and most common being *direct digital synthesis* (DDS), just consisting of an NCO, DAC and an interpolation filter (LP), as discussed in Section 5.5.6.

The full performance depends on both M and N, where to utilize the full dynamic range of the DAC $M > N$ and the look-up table increases exponentially as N increases. Instead of a large LUT[18], usually the CORDIC algorithm, an iterative algorithm to compute sine and cosine values, is used.

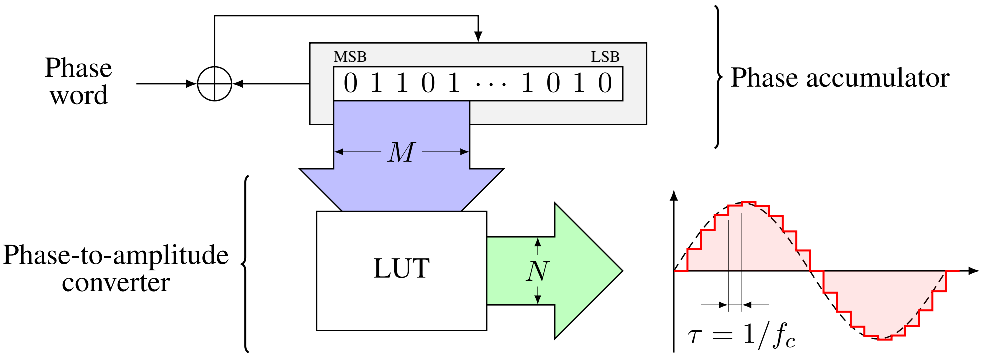


Figure 5.40 Numerically controlled oscillator.

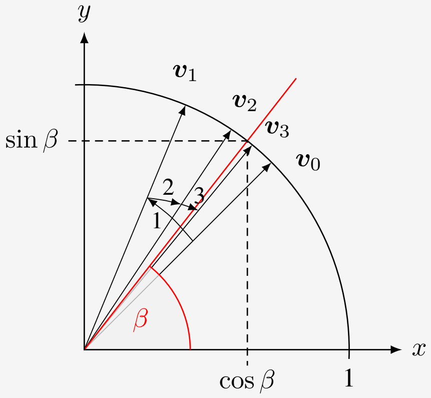
[18] generally LUTs are more efficient when N is small, but it is hard to synthesize a LUT with a wide address bus

The CORDIC Algorithm

CORDIC stands for COordinate Rotation DIgital Computer and was first described by Jack E. Volder (it is also known as Volder's algorithm) during the Cold War in 1959, to replace the old analog resolver in the B-58 bomber's navigation computer. At first calculating only trigonometric functions, it was later generalized to calculate hyperbolic and exponential functions as well as logarithms and square roots.

As will be derived, by doing the computation iteratively, only addition, subtraction, bit-shift and table-lookup operations are necessary, making CORDIC well-suited for applications on FPGAs and (smaller) microcontrollers, where hardware multipliers are unavailable. When multipliers are available, table-lookup methods and power series are generally faster than CORDIC.

Underlying principle: to calculate the sine and cosine values of some angle β, a polar vector $\boldsymbol{v}$ is rotated iteratively around the unit circle, where its x- and y-coordinate values converge to $x = \cos\beta$ and $y = \sin\beta$.



The rotation is done by multiplication with rotation matrix $\boldsymbol{R}_i$ in every iteration i, according to

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} = \boldsymbol{v}_i = \boldsymbol{R}_i\, \boldsymbol{v}_{i-1} ,$$

with

$$\boldsymbol{R}_i = \begin{bmatrix} \cos\gamma_i & -\sin\gamma_i \\ \sin\gamma_i & \cos\gamma_i \end{bmatrix} = \frac{1}{\sqrt{1+\tan^2\gamma_i}} \begin{bmatrix} 1 & -\tan\gamma_i \\ \tan\gamma_i & 1 \end{bmatrix} ,$$

where the trigonometric identities

$$\cos\gamma = \frac{1}{\sqrt{1+\tan^2\gamma}} \qquad \text{and} \qquad \sin\gamma = \frac{\tan\gamma}{\sqrt{1+\tan^2\gamma}}$$

have been applied. Thus, it follows

$$\begin{aligned} \boldsymbol{v}_i &= \frac{1}{\sqrt{1+\tan^2\gamma_i}} \begin{bmatrix} 1 & -\tan\gamma_i \\ \tan\gamma_i & 1 \end{bmatrix} \begin{bmatrix} x_{i-1} \\ y_{i-1} \end{bmatrix} \\ &= \frac{1}{\sqrt{1+\tan^2\gamma_i}} \begin{bmatrix} x_{i-1} - y_{i-1}\cdot\tan\gamma_i \\ x_{i-1}\cdot\tan\gamma_i + y_{i-1} \end{bmatrix} . \end{aligned}$$

Restricting the angles γ_i so that $\tan\gamma_i$ only takes on the values $\pm 2^{-i}$, the multiplication with the tangent can be replaced by a division by a power of two, which in digital hardware can be realized using a bit-shift operation:

$$\boldsymbol{v}_i = \frac{1}{\sqrt{1+2^{-2i}}} \begin{bmatrix} x_{i-1} - y_{i-1}\cdot\sigma_i\, 2^{-i} \\ x_{i-1}\cdot\sigma_i\, 2^{-i} + y_{i-1} \end{bmatrix} , \quad \text{with} \quad \sigma_i = \begin{cases} 1 & \beta_i \geq 0 \\ -1 & \beta_i < 0 \end{cases}$$

being the sign of the remaining angle β_i for the ith iteration.

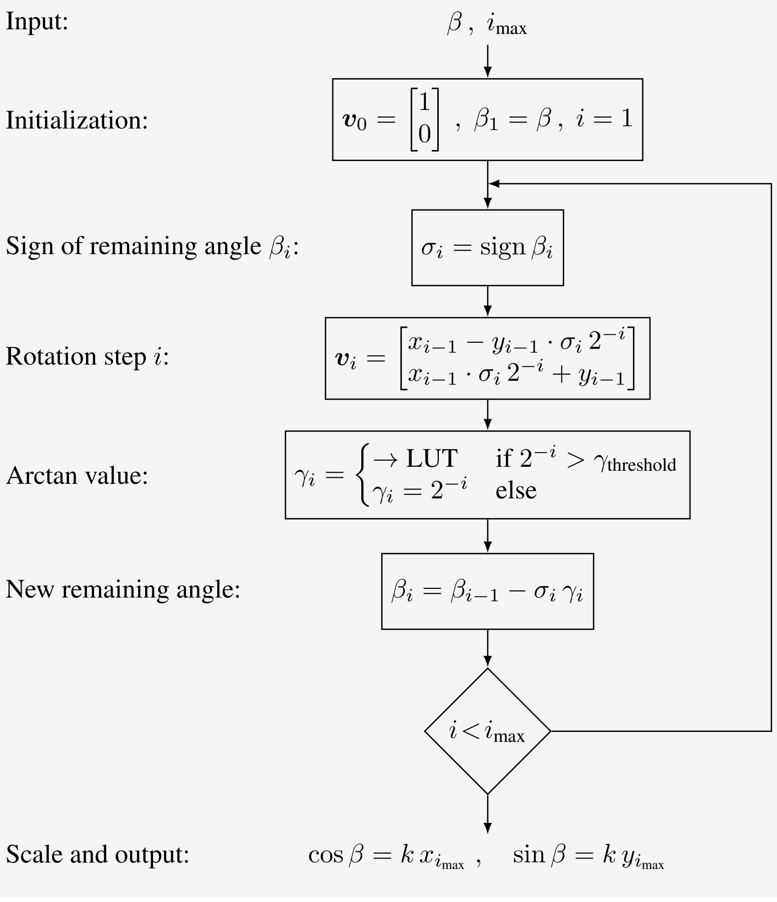
Further, the fractional factor converges to

$$\lim_{i_{\max}\to\infty}\prod_{i=0}^{i_{\max}}\frac{1}{\sqrt{1+2^{-2i}}}\approx 0.6072529...=k$$

for many iterations i (high i_{max}), which can be precomputed and stored as LUT (varying for different i) or as a single constant and applied only once, at the end (simpler, but decreases the convergence a bit).
At last, in every iteration the remaining angle β_i for the next iteration has to be calculated:

$$\beta_i=\beta_{i-1}-\sigma_i\,\gamma_i=\beta_{i-1}-\sigma_i\,\arctan 2^{-i}\ ,$$

where the values of $\gamma_i=\arctan 2^{-i}$ can be computed in advance and be stored in a LUT. Since for small angles $\arctan\alpha\approx\alpha$, the size of the LUT can be reduced in those regions (slowing down the convergence somewhat). The full CORDIC algorithm for sine and cosine of the angle $\beta\in[-\pi/2,\pi/2]$ therefore is:


Input:
$\beta\ ,\ i_{\text{max}}$
Initialization:
$\boldsymbol{v}_0=\begin{bmatrix}1\\0\end{bmatrix}\ ,\ \beta_1=\beta\ ,\ i=1$
Sign of remaining angle β_i:
$\sigma_i=\operatorname{sign}\beta_i$
Rotation step i:
$\boldsymbol{v}_i=\begin{bmatrix}x_{i-1}-y_{i-1}\cdot\sigma_i\,2^{-i}\\x_{i-1}\cdot\sigma_i\,2^{-i}+y_{i-1}\end{bmatrix}$
Arctan value:
$\gamma_i=\begin{cases}\to\text{LUT} & \text{if } 2^{-i}>\gamma_{\text{threshold}}\\ \gamma_i=2^{-i} & \text{else}\end{cases}$
New remaining angle:
$\beta_i=\beta_{i-1}-\sigma_i\,\gamma_i$
$i<i_{\text{max}}$
Scale and output:
$\cos\beta=k\,x_{i_{\max}}\ ,\quad \sin\beta=k\,y_{i_{\max}}$

There are many more and more elaborate methods used in advanced NCOs, including Delta-Sigma and others. NCOs offer several advantages over other types of oscillators, namely in terms of agility, accuracy, stability and reliability. They are used in various communications applications including 3G wireless (which made them so ubiquitous), multilevel FSK/PSK modulators and demodulators, SDRs, digital PLLs and even radar systems.

5.5 PLLs and Synthesizers

5.5.1 Introduction

A precise treatment of all issues of a phase-locked loop is beyond this lecture. In fact, whole books have been written on just this topic. Since the analysis of the loop and loop filter are covered in other lectures, we will only briefly touch the topic of PLLs with emphasis on topologies.

Many applications nowadays run above 1 GHz, for which stable and reasonably accurate oscillators on the order of, say, 1 ppm are not readily available. For lower frequencies such as 100 MHz, this would not pose a problem. But how can higher frequency signals be derived from an existing lower frequency reference?

5.5.2 Phase-Locked Loop (PLL)

A PLL is a control loop. In the example given in Fig. 5.41, a reference signal of 100 MHz is assumed. A voltage-controlled oscillator (VCO) outputs a frequency of around 3.2 GHz, initially not necessarily very accurate, either due to the VCO itself or due to the control voltage at the input of the VCO. The output of the VCO is now frequency divided by 32, which results in a signal around 100 MHz, in the region of the accurate reference signal. Both ’100 MHz’-signals are input to a phase discriminator, whose output is a function of the phase difference of its input signals. This signal, in turn, is filtered by the loop filter and controls the VCO. Hence, if the frequency of the 3.2 GHz signal is, say, too low, there will be a frequency offset and hence a phase offset at 100 MHz, which will be detected by the phase discriminator, filtered by the loop filter, pulling the VCO up to the right frequency. The filter order determines the loop order. The choice of the order of the loops depends on the locking speed desired and on the phase response needed. Phase discriminators can be either in the form of a multiplicator or digital (XOR gates).

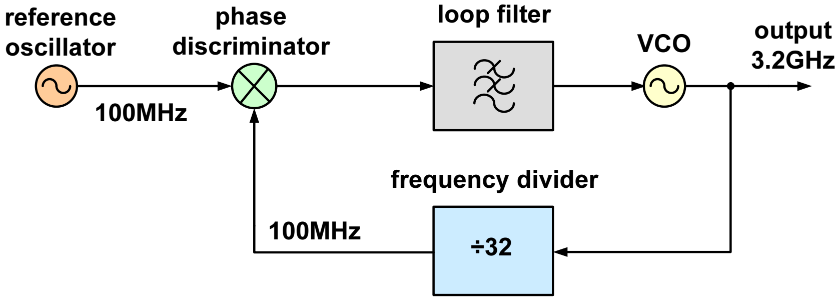


Figure 5.41 Phase-locked loop.

5.5.3 Integer-N Synthesizer

In the last section we have seen the principal operation of a PLL. In order to use the PLL for many different frequencies, we need to incorporate a programmable divider rather than a fixed one. Fig. 5.43 shows the simplest of all frequency synthesizers, where the output frequency can be any of

$$f = N \cdot f_R, \qquad N \text{ integer.} \tag{5.91}$$

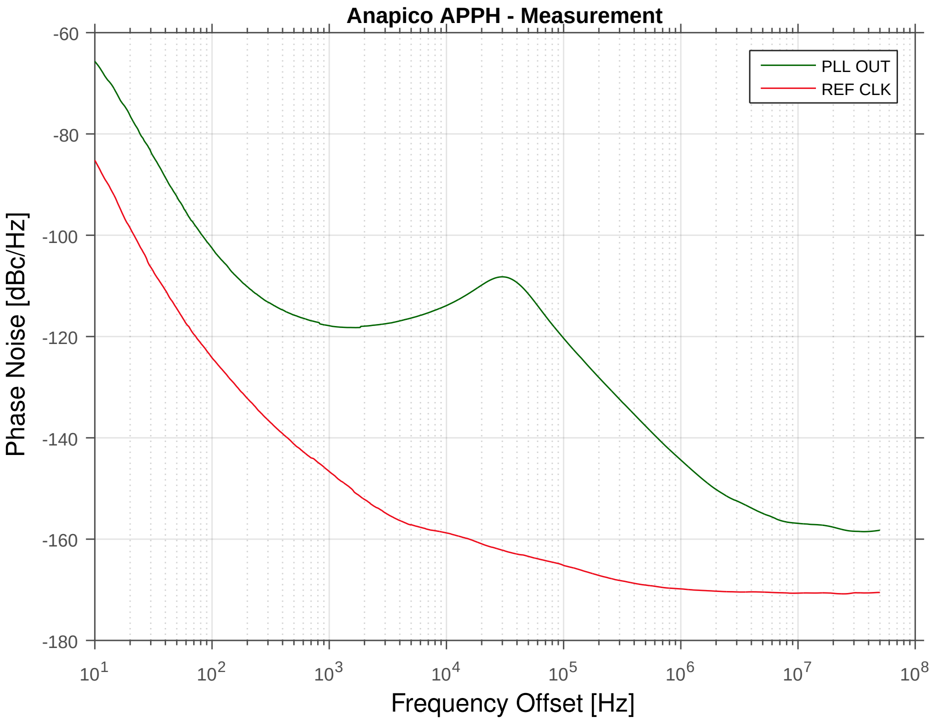


Figure 5.42 Example of phase-noise measurement. The factor 10 in the division of the reference clock can be seen by the 20 dB increase in the phase noise close to the carrier. Also, the influence of the loop-filter bandwidth can be clearly observed in the PLL output.

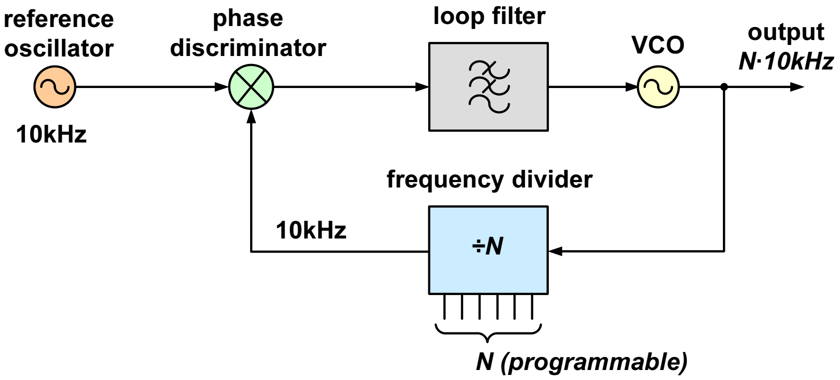


Figure 5.43 Integer-N synthesizer.

To allow for the frequency grid needed, f_R must be made small. But a small reference frequency means a large frequency acquisition time (settling time upon change of N). Besides, the phase noise of the reference oscillators appears multiplied by N^2 at the output.

5.5.4 Fractional-N Synthesizer (Dual Modulus)

We can avoid such low reference frequencies by using non-integer frequency divisions. To this end, the frequency is divided alternately by two different ratios, for example by $N = 10$ and $N = 11$. The synthesizer

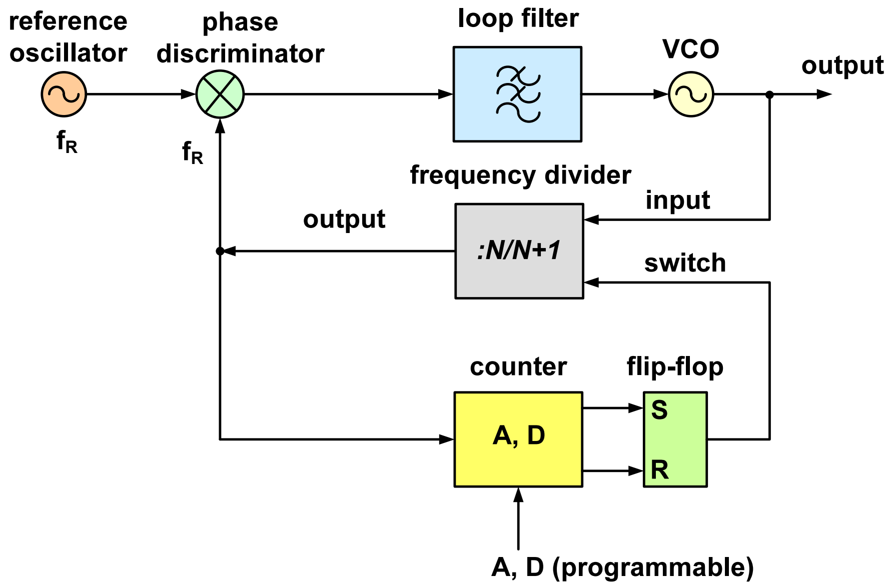


Figure 5.44 Fractional-N dual-modulus synthesizer.

given in Fig. 5.44 works as follows: We start the loop operation by dividing through $N + 1$. We perform a total of A such divisions before we switch the divider to division by N. The counter counts on to D, resets the divider to division by $N + 1$, and resets itself to zero. The number of input pulses T (at the divider) for such a cycle as just described is

$$T = (N + 1)A + N(D - A) = A + DN. \tag{5.92}$$

During that time, the counter has registered D pulses. The average ratio of input and output frequency of the divider is thus

$$R = \frac{T}{D} = N + \frac{A}{D}. \tag{5.93}$$

A/D is not an integer and allows a frequency grid resolution much smaller than f_R. An example: for $N = 10$, $D = 100$, and $A = 0 \ldots 100$, we get ratios of $R = 10 \ldots 11$ in steps of 0.01.

The disadvantage of this solution is that the frequency of the output of the divider varies over one cycle. To overcome this problem, the time constant of the loop filter needs to be large.

5.5.5 Dual-modulus integer-N synthesizer

The dual-modulus concept is also used for integer-N synthesizer, particularly above 1 GHz where only binary dividers or $N/(N + 1)$ dividers are available. The function of the synthesizer given in Fig. 5.45 is as follows: The two programmable parameters are M and A, with $M \geq A$. We start by dividing by $N+1$. The second counter counts to A and subsequently switches the divider to N. The first counter counts to M and switches the divider back to $N+1$ before it resets both itself and the second counter. When the first counter resets everything, it has only registered one pulse at its output. A whole cycle contains T input impulses

$$T = (N + 1)A + N(M - A) = A + MN. \tag{5.94}$$

T is the divider ratio. For $N = 10$, $M = 100$, and $A = 0 \ldots$ 100, we can program every integer ratio between 1000 and 1100 using ’slow’ programmable counters.

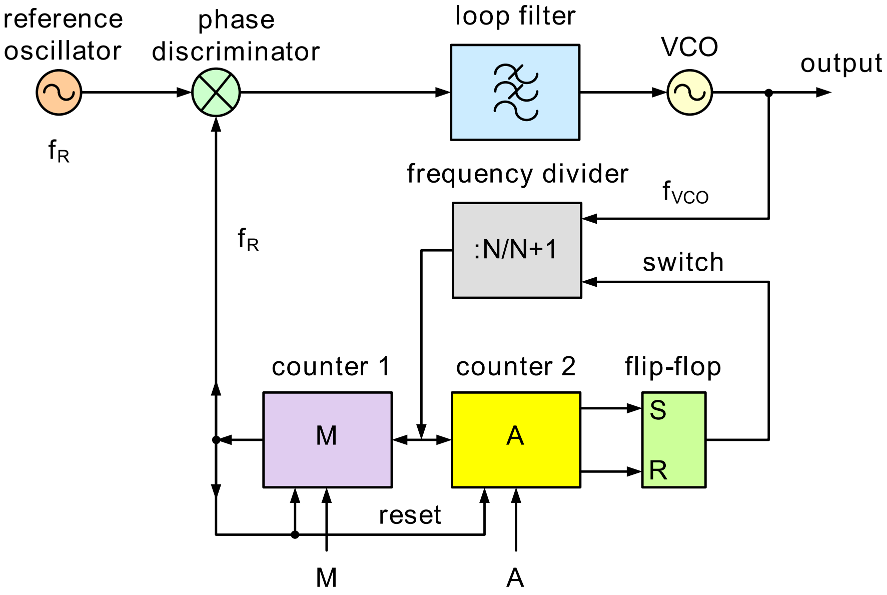


Figure 5.45 Integer-N dual-modulus synthesizer.

5.5.6 Direct Digital Synthesizer (DDS)

The best synthesizer in terms of phase noise and acquisition time is the *direct digital synthesizer*. A block diagram of a DDS is shown in Fig. 5.46. There is basically no settling time, so that the DDS may be directly modulated using an information signal. The phase noise behavior is given by clock jitter of the sampling frequency, divided by the number of sampling points per period. However, the necessary sampling frequency might be very high, so that a combination of DDS and a classic PLL is needed.

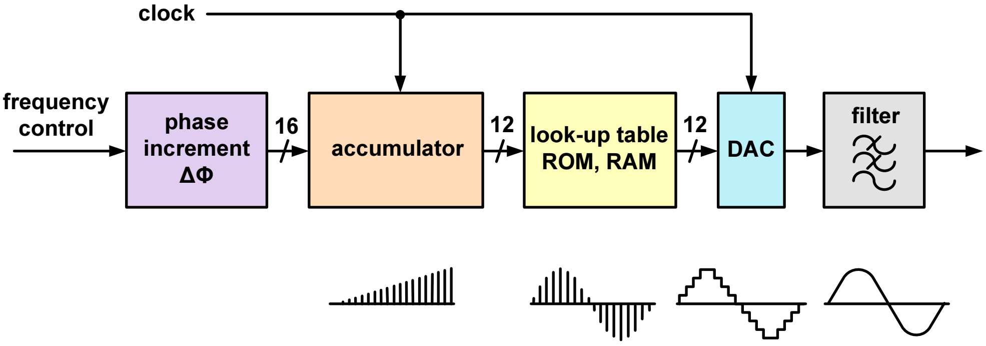


Figure 5.46 Direct digital synthesizer.

5.6 Further literature

An excellent source of many more details of topics covered in this chapter is the book written by W. Bächtold [3], from where a lot of figures are borrowed.

6 Transmitter and Receiver Architectures

Equipped with an understanding of the components listed in Chapters 4 and 5 we are now ready to analyze wireless transmitters and receivers.

6.1 AM Transmitter

The simplest form of a wireless transmitter that transmits some useful information could be produced by switching on and off an oscillator that is hooked to an antenna. Such a transmitter is provided by Fig. 6.1. This would produce a modulation scheme known as on-off keying, which is only used to transmit Morse code. The next simplest thing is an AM transmitter, see Fig. 6.2. An AM transmitter simply modulates its RF carrier by the audio frequency, which results in a double-sideband spectrum, shown in Fig. 6.3(a) and (b).

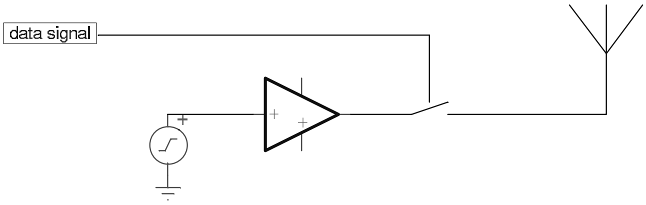


Figure 6.1 On-off keying transmitter.

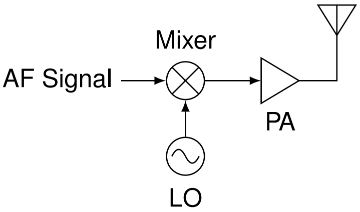
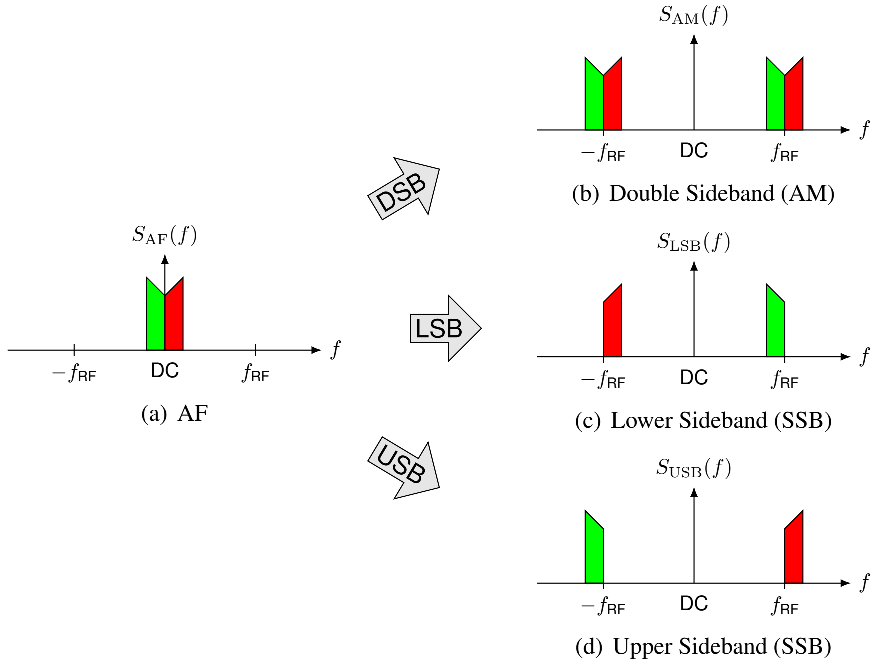


Figure 6.2 AM transmitter.

6.2 SSB Transmitter

Naturally, AM transmitters are not very spectrum efficient, since they transmit the full double sideband, a costly redundancy. There are several architectures that allow the transmission of one AF sideband only. Possible SSB spectra are shown in Fig. 6.3(c) and (d). The lower sideband (LSB) is sometimes called the *inverted sideband* to account for the inverse order of frequency components compared to the baseband components. Similarly, the upper sideband (USB) is referred to as the *erect sideband*.



(a) AF

(b) Double Sideband (AM)

(c) Lower Sideband (SSB)

(d) Upper Sideband (SSB)

Figure 6.3 AM (double sideband, DSB) and single sideband (SSB) spectra.

6.2.1 The Filter Method

With the filter method, the AF signal is modulated onto an IF carrier, at which point one of the sideband is filtered away. The resulting signal is then further up-converted using a second mixer, before being amplified and transmitted. Fig. 6.4 illustrates this process.

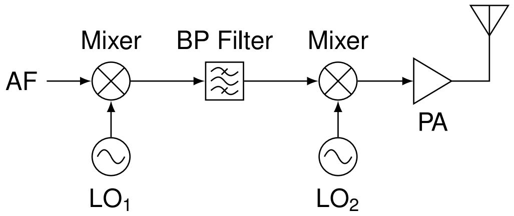


Figure 6.4 SSB generation: the filter method.

6.2.2 The Phasing Method

The phasing method, also called Hartly modulator, uses its two mixers in a quadrature way rather than in a heterodyne way, i.e., the LO signal to the mixers is at the same frequency (the RF) but phase-shifted by 90°. See Fig. 6.5 for a circuit diagram. A second phase shifter is necessary, either at the AF band or at the RF before the signals are added up. The phase shift of the LO signal is relatively simple. However, the phase shift for the AF signal is often cumbersome (large relative bandwidth). An equivalent phase shift at RF is simpler but this approach is less flexible if the RF should be changed. By adding two phase-shifted signals (or subtracting), one of the two sideband cancels, leaving an SSB signal.

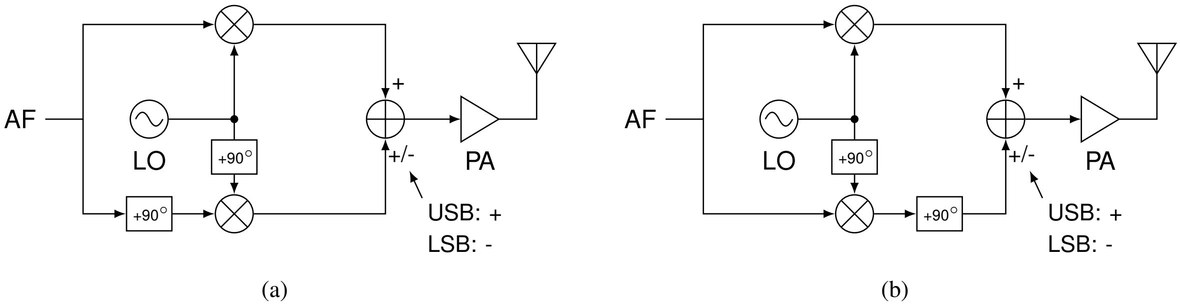


Figure 6.5 SSB generation: the phasing method. (a) 90 ° phase shift at AF; (b) 90 ° phase shift at RF.

6.2.3 The Weaver Method

After the filter method and the phasing method to generate an SSB signal, the Weaver method is often also referred to as the *third method* [83]. Fig. 6.6 contains a Weaver transmitter. The Weaver method forms a very nice compromise between the former two methods. It gets rid of the highly selective filter (as used in the filter method), as well as the wideband phase shift network (which were needed in the phasing method). All it needs is the 90 ° phase shifts at the LOs. This is usually relatively easy to achieve, e.g., using cascaded Flip-Flops as shown in Fig. 6.7. An alternative is to use a PLL with a phase discriminator being a multiplier (output will be zero if input signals show 90 ° phase difference).

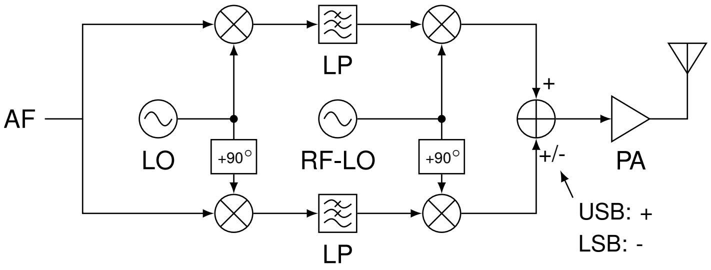


Figure 6.6 SSB generation: the Weaver method.

Mixers using a 90 ° shifted version of the same LO are often also referred to as image-rejection mixers, since they can be used in receivers, when an image-rejection filter would be too costly.

6.3 I/Q up-converter

As we will shortly see with receiver architectures, the task of converting the input frequency to the output frequency may happen in one, two, or several conversion steps. Although the image problem is relaxed with transmitters (we have, in essence, control over the spectrum), the choice of the number of conversion steps remains. Recently, however, direct up-converters have become popular. Their modulators/mixers can directly be fed by a digitally modulated signal, which is present in its complex baseband representation. Such a transmitter is depicted in Fig. 6.8. Cartesian coordinates such as the I/Q representation can always be transformed into polar coordinates representing amplitude and phase. Thus, an alternative transmitter may be built using a polar modulator, see Fig. 6.9.

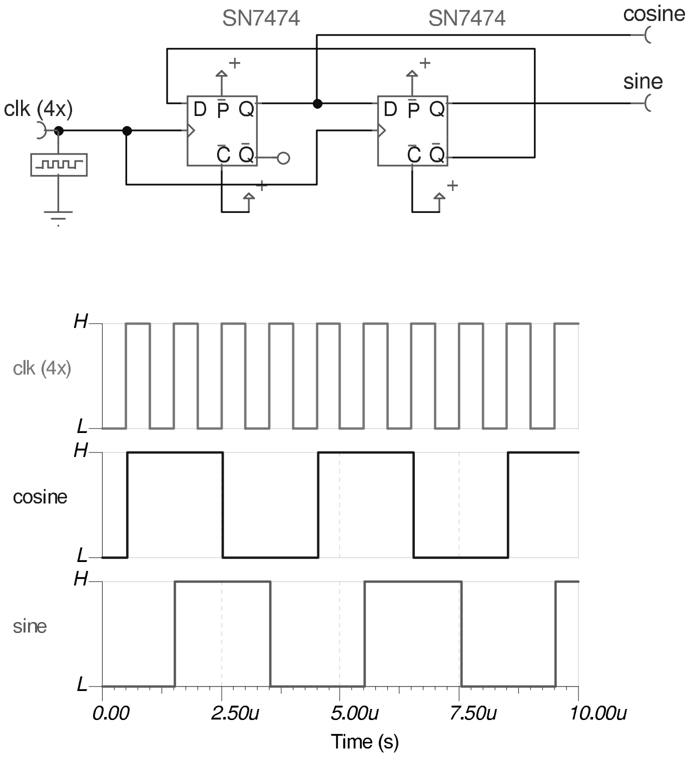


Figure 6.7 Generation of 90° shifted LO signals. Top: circuit diagram; bottom: signal time chart.

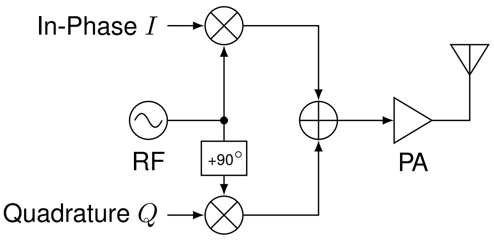


Figure 6.8 Transmitter using direct up-conversion of a complex baseband signal.

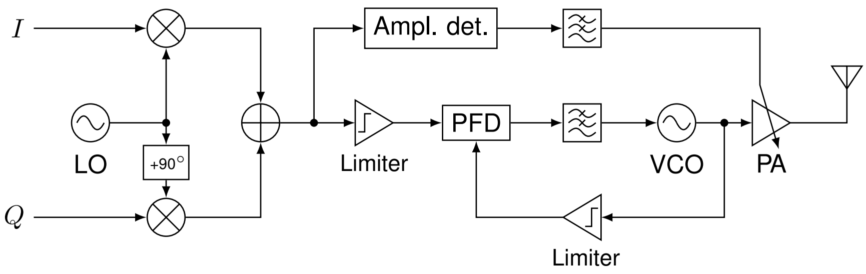


Figure 6.9 Polar modulator, e.g. for EDGE.

On the other hand, if there is no information in the amplitude, i.e., a CPM scheme, we can have very interesting architectures based on *open-loop modulation* and *modulated synthesis* [13]. These approaches are shown in Figs. 6.10 and 6.11, respectively.

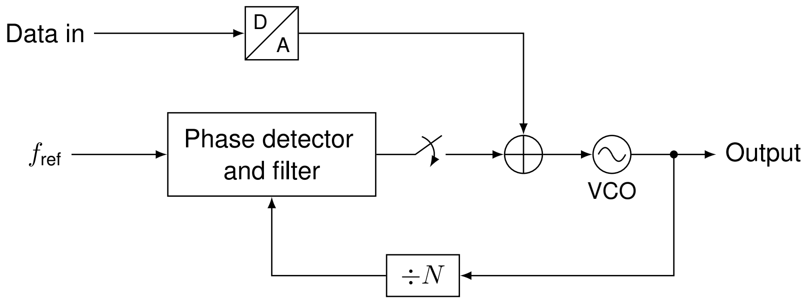


Figure 6.10 Open-loop modulation.

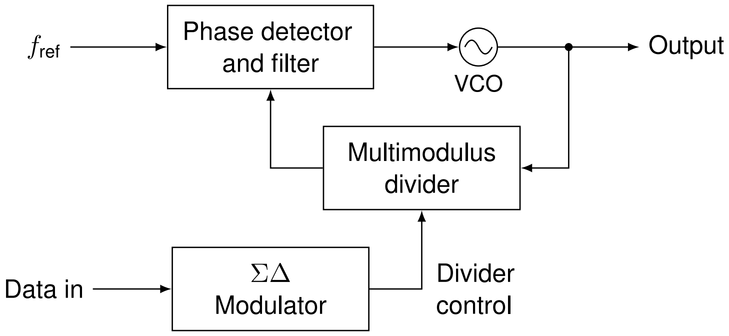


Figure 6.11 Modulated synthesis.

6.4 Receiver Specifications

It is much more difficult to build a good receiver than a good transmitter. The reason for this is mainly that we cannot really influence the kind of interference we will face. It is much easier not to generate spurious signals (in the transmitter design) than to actually cope with them. One of the main challenges of receiver design is the huge dynamic range we may have to cover. Many receivers employ some automatic gain control (AGC). This works around the problem of amplifier saturation, but the selectivity problem with high-level interferers close to the signal of interest remains. Sharp filters are necessary to counterfeit such situations.

On the other hand, many neat tricks in receiver design have arisen over the past years. Apart from the very simple receivers, we will always find (in different order and possibly repeatedly) LNA, filter, mixer, and some detector, which comprise a receiver, but the order and the number of these components may vary widely.

Similarly to the design parameters of components, design parameters for whole receivers and transmitters exist. Among the ones for receivers are:

- sensitivity
- selectivity
- power consumption
- complexity

6.5 Early Radio Sets

6.5.1 The Diode Detector

The easiest of all wireless receivers is the diode detector, schematically shown in Fig. 6.12. The diode detector consists of only a handful components and works, if carefully selected, even without the use of external power.[1] An efficient antenna brings its signal through a tunable resonance circuit (LC) to the diode detector, which rectifies the signal, in other word, passes on the envelope, filtered by a capacitor, to the earphone. In the old days, the diode diode was a crystal, which was tapped by a thin wire, looking like a cat's whisker. The famous crystal detector was therefore often referred to as a cat's whisker detector.

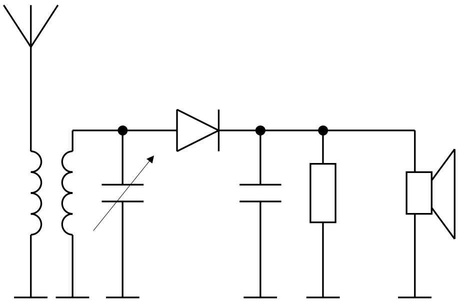


Figure 6.12 Diode detector.

6.5.2 Tuned RF Sets

The next step in the development of wireless receivers was to add amplifiers both at RF and at audio. The amplifiers of these early radios were, of course, built of vacuum tubes. Throughout the chain of the receiver, several tuning stages were placed, which had to be tuned simultaneously, hence its name *tuned RF sets* (TRF). This was highly impractical, and often charts had to be employed to correctly tune all stages.

6.5.3 The Regenerative Receiver

Edwin Armstrong (1890–1954) originally came up with a receiver concept called *regenerative receiver*. Since his most successful design, the superhet, was protected by patent rights, the regenerative receiver was the way out for many competitors. This receiver is basically an amplifier with a lot of feedback, so much that it almost oscillates, but just about does not. In the presence of an incoming signal on that frequency, it oscillates and regenerates again, when the signal disappears. The oscillation activity is thus modulated by the radio signal. One of the drawbacks of such a receiver is its sensitivity to a changing environment (temperature change may make it start oscillating).

A variant, the *superregenerative receiver* solves this problem by frequently opening the feedback to stop any oscillation present. This process is called *quenching*. The receiver starts oscillating once every quenching period, but it does so slightly earlier when an external signal is present at the oscillating frequency.

[1] For a batteryless circuit, long efficient antennas and high-impedance earphones are a must.

6.6 The Superheterodyne Receiver

Armstrong again was the inventor of the superheterodyne receiver[2], called *superhet* for short. In 1917 the superhet, whose general outline is given in Fig. 6.13 was a massive innovation to avoid the problem of insufficient gain of tube amplifiers above 1-2 MHz. Up to this day, the superhet has undiminished popularity, however, due to other reasons. Gain is no longer a problem.

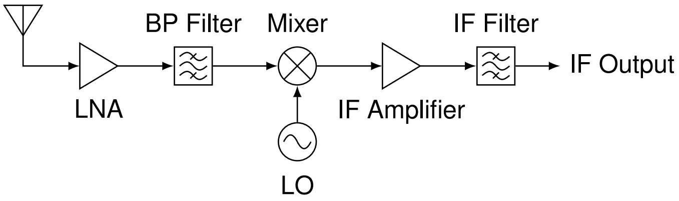


Figure 6.13 Superhet receiver.

The main disadvantage of a superhet is the image problem (but not as in: other people frown at you). To see where the problem comes from, consider Fig. 6.14. You remember, a mixer produces sum and differences of its input frequencies. To down-convert a signal, we have two possibilities, since the same IF frequency can be generated using two different LO frequencies. But this also means that another RF than the signal of interest can fall onto the IF chosen. This 'other' RF is the image. It must be filtered off prior to the mixing process. Otherwise, it will end up on top of the wanted signal at the same IF. We speak of high-side injection if the LO frequency is higher than the RF, and of low-side injection if the LO frequency is lower than the input RF. The main benefit of low-side injection is the smaller bandwidth required of the image filter. The advantage of high-side injection is the smaller relative tuning range of the LO signal, the drawbacks are the inversion of the signal spectrum (which is not always an issue) and the higher LO frequency needed.

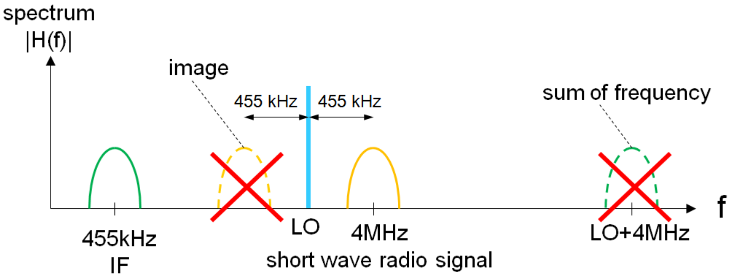


Figure 6.14 Spectrum of signals in a superhet receiver (low-side injection) to illustrate image problem.

Compared to other modern alternatives, the superhet exhibits the following advantages:

- many standard IF bandpass filters available,
- flicker noise is not an issue,
- good dynamic range (AGC needed).

Disadvantages, on the other hand, are:

- additional components for external filters,

[2] As is often the case in history, some clever inventor who paved the way gets forgotten. It was in fact Reginald Fessenden, who came up with the concept of a heterodyne receiver, which converted the incoming RF signal directly to audio, an approach that finds its revival in today's direct-conversion receivers.

- image and spurious signals,
- power consumption.

The concept of the superhet can even be carried further. Rather than using one IF only, there are double-conversion and even triple-conversion receivers. The additional complexity is usually compensated for by easing the requirements on the image filters. See Fig. 6.15 for an extended superhet form including AGC.

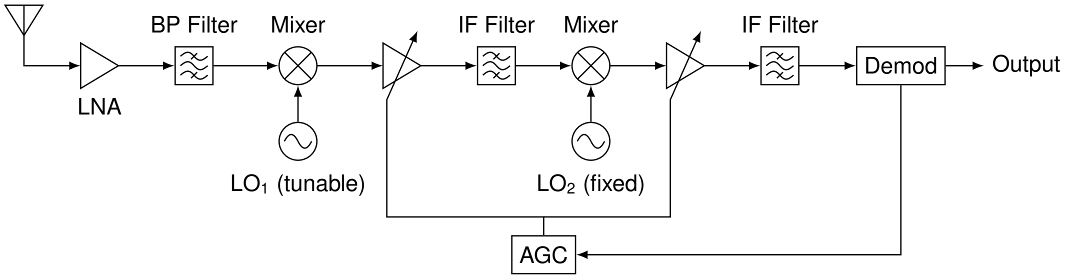


Figure 6.15 Dual-conversion superhet receiver with AGC.

6.7 The Direct-Conversion Receiver

In more recent times, one particular architecture has gained importance, the so-called *direct-conversion receiver*. Despite some shortcomings which make it unusable for certain applications, the direct-conversion receiver can solve many classical problems originally solved with a superhet design. A typical block diagram of a direct-conversion receiver is given in Fig. 6.16.

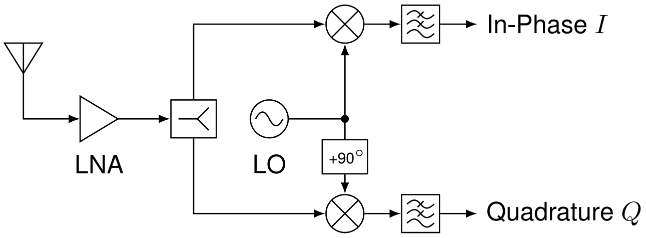


Figure 6.16 Direct-conversion receiver.

Compared to the superhet design, the direct-conversion receiver has both advantages and disadvantages. Advantages are:

- simple receiver
- baseband filtering possibly digital

Some disadvantages are:

- flicker noise
- self-reception causing DC offset
- image-rejection mixer with precision needed
- second-order effects

The main benefit and the reason why this approach is becoming so popular is the fact that the image problem is no longer one, see Fig. 6.17.

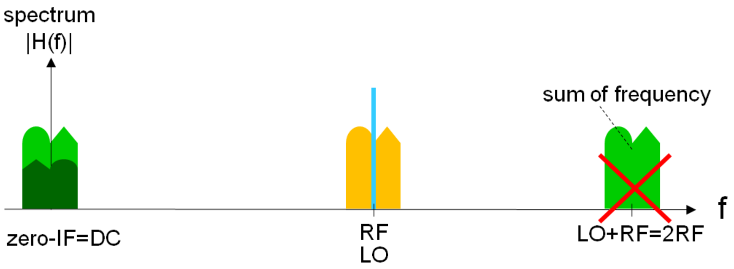
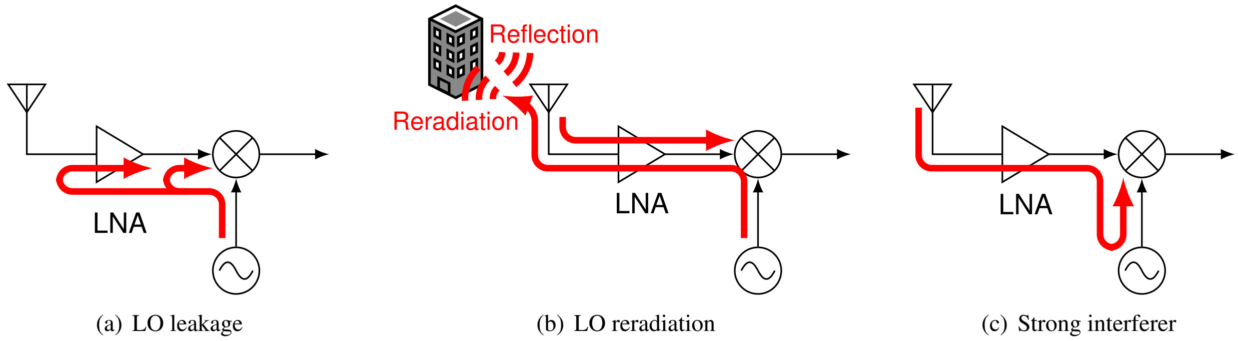


Figure 6.17 Spectrum of signals in a direct-conversion receiver.

Because the LO frequency is at the RF frequency, there is no image. The signal of interest is its own image, so to speak. As long as we have a certain amount of I/Q balance (phase shift close to 90°and equal gain), the image is suppressed (dark area in the baseband). However, as mentioned above, so-called self-reception due to the LO signal leaking into the RF port is a problem, see also Fig. 6.18. The resulting DC offset has to be cancelled using complicated DC adjusting techniques or by AC coupling. The latter approach destroys some of the spectrum of the desired signal. Because the LO frequency is the same as the RF frequency, an IF of zero is produced. A direct-conversion receiver is therefore often referred to as a zero-IF receiver. It mixes the signal down to baseband.



(a) LO leakage (b) LO reradiation (c) Strong interferer

Figure 6.18 Origins of DC problems of direct conversion receivers.

6.7.1 I/Q Imbalance

As stated above, the image suppression of a direct conversion receiver depends on the gain and phase imbalance of the two mixer paths. By expressing the two mixer paths with gain imbalance G and phase imbalance ϕ

$$I(t) = G \cdot \cos(\omega t + \phi), \tag{6.1}$$

$$Q(t) = \cos\left(\omega t + \frac{\pi}{2}\right), \tag{6.2}$$

we can find the RF signal $s_{\text{RF}}(t)$ by modulating it with the carrier signal

$$\begin{aligned}s_{\text{RF}}(t) &= G\cdot\cos(\omega t+\phi)\cos(\omega_c t)-\cos(\omega t+\phi)\cos(\omega_c t)\\ &= \underbrace{\frac{1}{2}G\cos(\omega_c t+\omega t+\phi)-\frac{1}{2}\sin\left(\omega_c t+\omega t+\frac{\pi}{2}\right)}_{s_{\text{USB}}(t)}\\ &\qquad + \underbrace{\frac{1}{2}G\cos(-\omega_c t+\omega t+\phi)-\frac{1}{2}\sin\left(\omega_c t-\omega t-\frac{\pi}{2}\right)}_{s_{\text{LSB}}(t)},\end{aligned} \tag{6.3}$$

where the first and the second half have been identified as the upper-sideband and the lower-sideband signal, respectively. Using more trigonometric identities we can form

$$\begin{aligned}s_{\text{USB}}(t) &= \frac{1}{2}G\cos(\omega_c t+\omega t+\phi)-\frac{1}{2}\cos(\omega_c t+\omega t)\\ &= \frac{1}{2}G\cos(\omega_c t+\omega t)\cos\phi-\frac{1}{2}G\sin(\omega_c t+\omega t)\sin\phi-\frac{1}{2}\cos(\omega_c t+\omega t) \qquad (6.4a)\\ s_{\text{LSB}}(t) &= \frac{1}{2}G\cos(-\omega_c t+\omega t+\phi)-\frac{1}{2}\cos(\omega_c t-\omega t)\\ &= \frac{1}{2}G\cos(-\omega_c t+\omega t)\cos\phi-\frac{1}{2}G\sin(-\omega_c t+\omega t)\sin\phi+\frac{1}{2}\cos(\omega_c t-\omega t). \qquad (6.4b)\end{aligned}$$

The suppression between the two sidebands in dB is

$$\boxed{A[\text{dB}] = 10\log_{10}\left(\frac{|s_{\text{LSB}}(t)|^2}{|s_{\text{USB}}(t)|^2}\right) = 10\log_{10}\left(\frac{G^2-2G\cos\phi+1}{G^2+2G\cos\phi+1}\right),} \tag{6.5}$$

where G is the amplitude ratio of the two paths and ϕ is the deviation from the desired 90° phase shift between the two LO paths. Fig. 6.19 visualizes this equation.

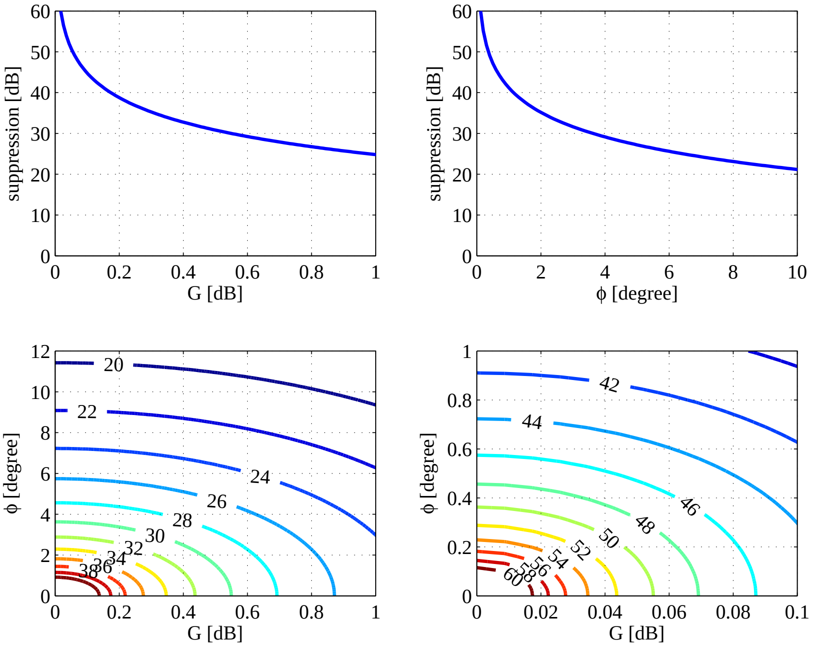
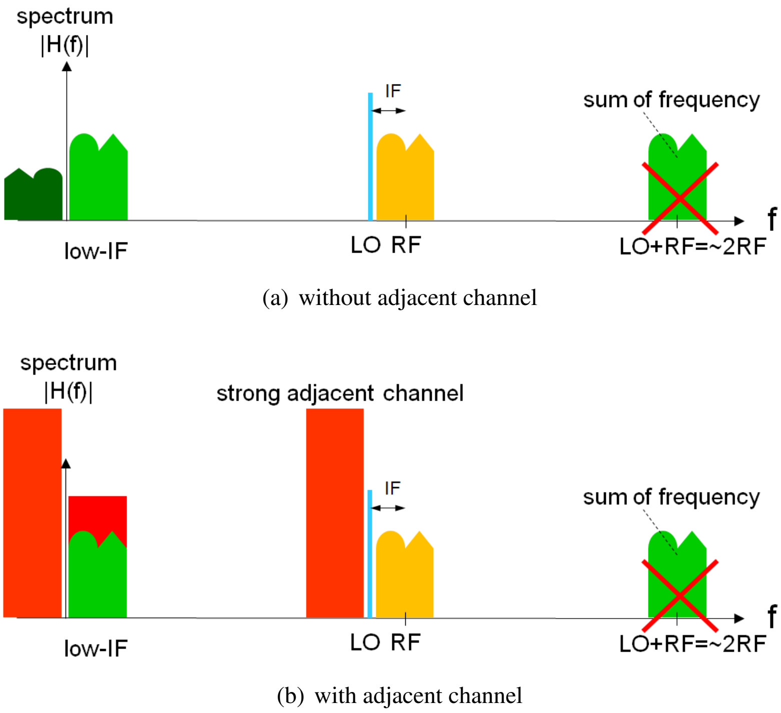


Figure 6.19 Image suppression as a function of gain and phase imbalance.

6.8 Low-IF Receiver

The above-mentioned problems have made people come up with something called a low-IF receiver, where rather than down-converting to zero IF, the signal is mixed down just far enough that the DC point is no longer part of the spectrum of the signal of interest, see Fig. 6.20(a). Since the spectrum is not symmetric, we need again a complex signal to represent it. A receiver as given in Fig. 6.16 is capable of down-converting the signal to a low IF, as long as the LO is set to the right frequency. Now, the image to fight is the adjacent channel which might increase the dynamic requirements of any analog-to-digital converter. of the signal of interest, see Fig. 6.20(b).



(a) without adjacent channel

(b) with adjacent channel

Figure 6.20 Spectra of signals in a low-IF receiver: The adjacent channel becomes a problem.

A real-valued bandpass filter as shown in Fig. 6.21 cannot distinguish between the signal of interest and the adjacent channel sitting at the exact negative frequency of the wanted signal.

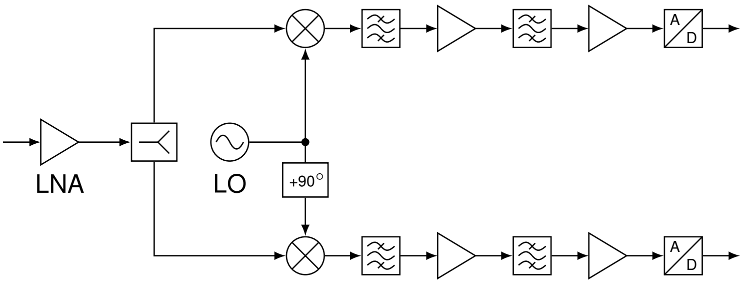


Figure 6.21 Real bandpass filters.

Such an adjacent signal can, however, be filtered using complex-valued impulse responses, which are implemented using a complex polyphase filter, see Fig. 6.22.

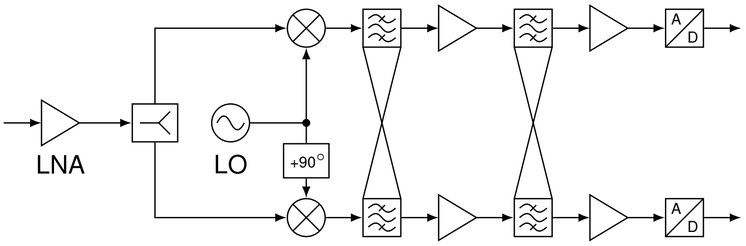


Figure 6.22 Complex polyphase filters.

Furthermore, if we are not careful with balancing I and Q, a strong adjacent channel (can be much stronger than our wanted RF signal) can leak into our low IF and still dwarf our signal.

6.9 Noise Considerations

Depending on the area within electronics, noise is modeled differently. Even in a single circuit there might be different aspects and different methods to describe noise, as illustrated in Fig. 6.23. Besides the concept of using effective noise temperatures and noise figures (NF), it is common to compute noise voltage, in particular in high-impedance networks. In order to compute the NF of a receiver chain, when some stages are defined with alternative noise parameters we develop some transformation operations. Given a certain noise power as

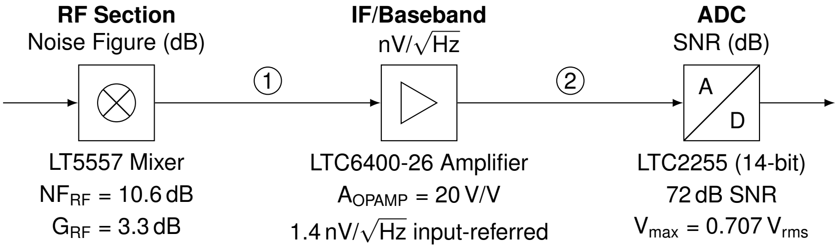


Figure 6.23 Example of receiver chain for numerical evaluation (according to the DN439 by Linear Technology, now part of Analog Devices).

$$P_N = kT_E B, \tag{6.6}$$

or rather, the noise power density (per Hz bandwidth) as

$$S_N = kT_E, \tag{6.7}$$

we can model the noise source with a voltage of

$$U_N = \sqrt{S_N \cdot 4R} = \sqrt{kT_E \cdot 4R}. \tag{6.8}$$

The factor 4 comes from the fact that noise voltage is usually given with respect to the open-circuit terminals of a resistor, while the noise power is measured in a matched source-load situation. The noise voltage produced corresponding to $T_0 = 290\,\mathrm{K}$ for a $R = 50\,\Omega$ load is

$$U_{N,50} = \sqrt{kBT_0 \cdot 4 \cdot 50} = 0.91\,\frac{\mathrm{nV}}{\sqrt{\mathrm{Hz}}}. \tag{6.9}$$

Due to the voltage divider consisting of source and load impedance we only get half the value. However, the voltage is amplified by the gain G given in dB and the noise figure NF also provided in dB. It is thus

$$U_{N_1} = 10^{\frac{G+\mathrm{NF}}{20}} \cdot \frac{U_{N,50}}{2}. \tag{6.10}$$

If at IF or baseband we find ourselves in a voltage regime, the corresponding voltage after the next stage with a linear voltage gain A is computed by taking into account the uncorrelated noise density $U_{N,\mathrm{IF}}$ from the current stage, resulting in

$$U_{N_2} = A \cdot \sqrt{U_{N_1}^2 + U_{N,\mathrm{IF}}^2}\,. \tag{6.11}$$

The noise relating to A-to-D conversion is characterized using yet different parameters. Depending on the crest factor of the signals converted the resulting $\mathrm{SNR}_{\mathrm{ADC}}$ of the ADC is usually given around 6 dB per bit. Assuming a constant spread of noise power over the Nyquist band defined through f_s (which for $\Sigma\Delta$ converters might not be the case), we can quantify the noise density using the maximum voltage $U_{\max}$ at the ADC as

$$U_{N,\mathrm{ADC}} = U_{\max} \cdot 10^{-\frac{\mathrm{SNR}_{\mathrm{ADC}}}{20}} \cdot \frac{1}{\sqrt{\frac{f_s}{2}}}. \tag{6.12}$$

The total noise figure in dB can now be expressed in terms of the noise density contributions of the different stages

$$\mathrm{NF}_{\mathrm{total}} = 20\log_{10}\left(\frac{2 \cdot \sqrt{U_{N_2}^2 + U_{N,\mathrm{ADC}}^2}}{A \cdot U_{N,50}}\right) - G. \tag{6.13}$$

In the following, we shall evaluate the numerical example after Design Note 439 of Linear Technologies, which is given in Fig. 6.23. To do so we are going through Eqs. (6.7) to (6.13) and obtain

$$U_{N_1} = 10^{\frac{3.3+10.6}{20}} \cdot \frac{0.91 \cdot 10^{-9}}{2} = 2.25\,\frac{\mathrm{nV}}{\sqrt{\mathrm{Hz}}}, \tag{6.14}$$

$$U_{N_2} = 20 \cdot \sqrt{2.25^2 + 1.4^2}\,\frac{\mathrm{nV}}{\sqrt{\mathrm{Hz}}} = 53\,\frac{\mathrm{nV}}{\sqrt{\mathrm{Hz}}}, \tag{6.15}$$

$$U_{N,\mathrm{ADC}} = 0.707 \cdot 10^{-\frac{72}{20}} \cdot \frac{1}{\sqrt{\frac{125 \cdot 10^6}{2}}} = 22.5\,\frac{\mathrm{nV}}{\sqrt{\mathrm{Hz}}}, \tag{6.16}$$

$$\mathrm{NF}_{\mathrm{total}} = 20\log_{10}\left(\frac{2 \cdot \sqrt{53^2 + 22.5^2}}{20 \cdot 0.91}\right) - 3.3 = 12.7\,\mathrm{dB}. \tag{6.17}$$

6.10 Further Literature

Good books to read on the topics of RF architectures are the ones by Hagen [29] and Sabin/Schoenike [65].

7 Antennas and Propagation

7.1 The Wireless Channel

The physical medium that transports the signal from the transmitter to the receiver is called the communication channel. In wireless communications, this is simply the atmosphere or free space (satellite communications). Additionally, there are wavelength-dependent phenomena such as earth-guided waves, reflected and diffracted waves etc. The interface between the electric circuit that produces the signal and the medium is the antenna, and as such will be considered as part of the wireless channel, since loss and other closely related phenomena occur in both parts the antenna and the channel.

Any channel (and this is not restricted to the wireless one) modifies the signal in ways more or less unpredictable to the receiver. These modifications are called noise. We distinguish between additive and multiplicative noise, as shown in Fig. 7.1. The additive noise arises from noise sources within the receiver itself, such as thermal noise, flicker noise, shot noise, and also from external noise sources such as atmospheric effects, cosmic radiation, and interference from other transmitters or electric devices.

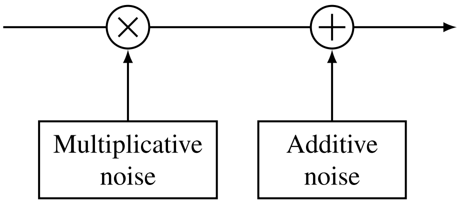


Figure 7.1 Two types of noise in the wireless channel.

The multiplicative noise sources can be found among:

- directional properties of transmitter and receiver antennas
- reflection (ground, walls, and hills)
- absorption (walls, foliage, atmosphere)
- scattering (from rough surfaces such as sea, ground, foliage)
- diffraction (from edges such as buildings and hills)
- refraction (layers of different permittivity)

It is common to further divide the multiplicative processes into three types of fading (see Fig. 7.2):

- *path loss*,
- *shadowing* or *slow fading*, and
- *fast fading* or *multipath fading*.

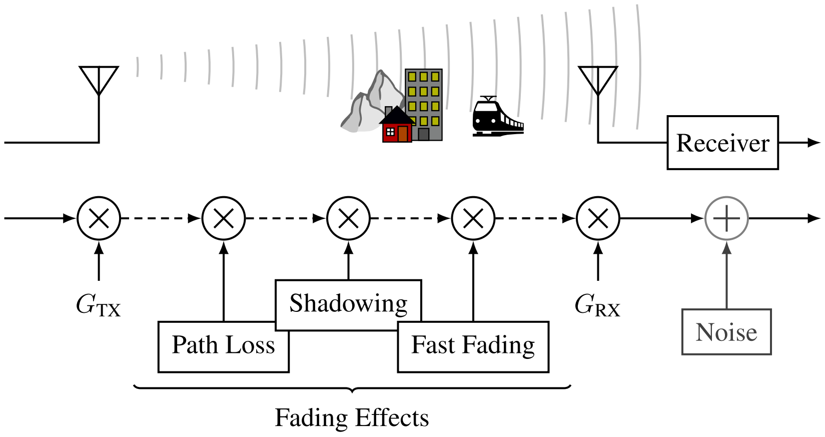


Figure 7.2 Additive and multiplicative [22] noise in the wireless channel.

The latter two processes are time-varying and depend heavily on the exact positions of the transmitter and receiver antennas. An example is given by Fig. 7.3. It shows the variation of three multiplicative effects as a mobile receiver moves away from a transmitting base station. The path loss is a deterministic effect, which models the higher attenuation at a larger distance. The shadowing process arises due to different degrees of obstruction in the path, whereas the fast-fading process occurs due to constructive and destructive interference between multiple waves reaching the mobile receiver.

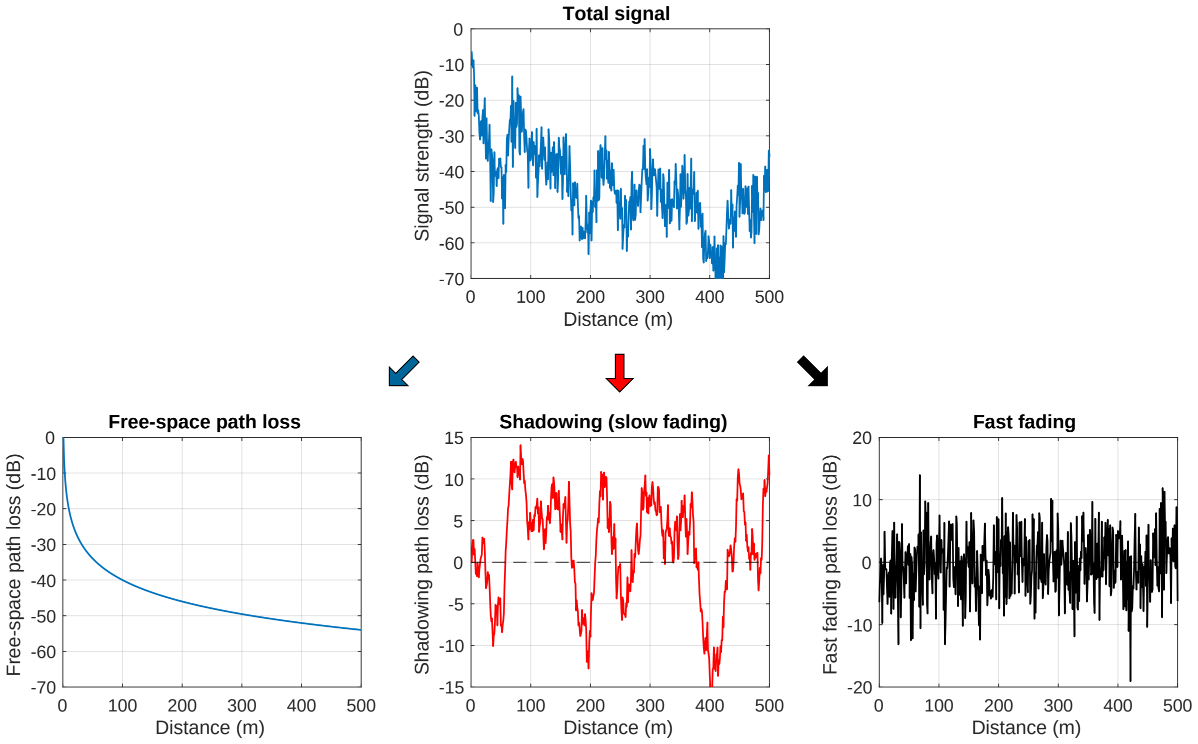


Figure 7.3 Three subtypes of multiplicative noise in the wireless channel. Each signal strength is plotted as a function of distance between transmitter and receiver.

7.2 Introduction to Antennas

Antennas are the interface between the transmitter (or receiver) and the transfer medium. Phil Smith (the inventor of the Smith chart) calls an antenna

> *...a component that matches the [transmission] line to space.*

There is little chance that antennas are ever going to be optimized out or integrated into ever-shrinking chips, since physics dictates its size. So with the growing market in mobile communications, the future of antennas is indeed very bright. The challenge to reduce it in size, however, remains. As operating frequencies increase, smaller form factors are possible. Competition in the design of reasonably good antennas increases as size becomes more important than sensitivity. On the other hand, building an antenna with higher gain directly relaxes the specification on the noise figure of the receiver chain and makes cheaper receivers possible. Hence, the antenna remains a key object in receiver optimization.

7.2.1 Comparison with Transmission Lines

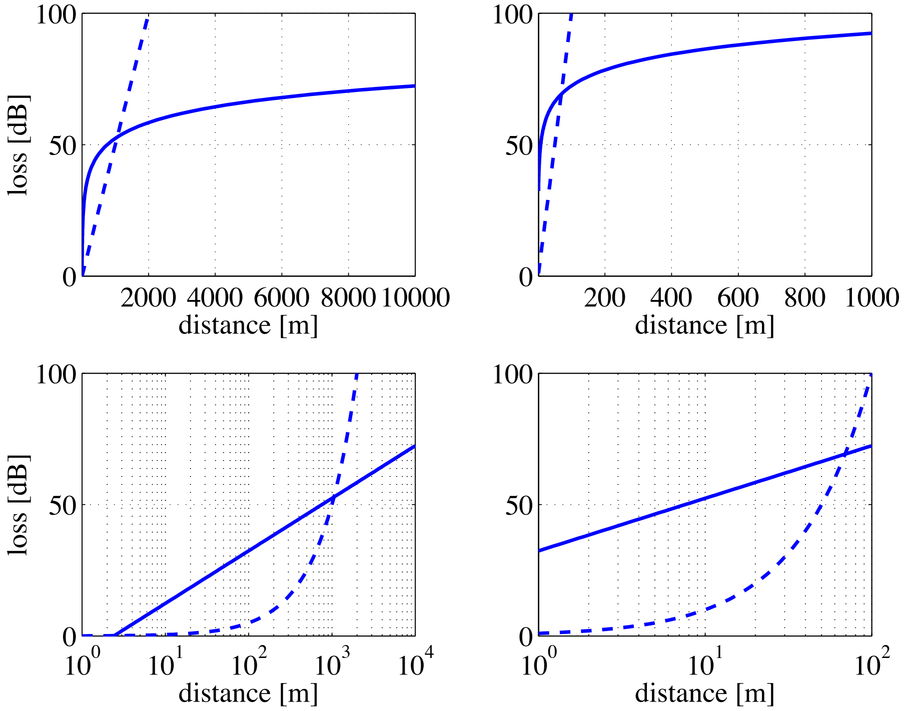


Figure 7.4 Comparison of antenna loss (solid line) and cable loss (dashed line). Left: operating frequency $f = 10$ MHz. right: $f = 1$ GHz. The top and bottom curves are identical except for the logarithmic view of the distances in the lower plots.

To start out, let us think about the question: ”Is a transmission line (e.g., a coaxial cable, RG-58C/U or similar) or an antenna more efficient in carrying a signal over a distance?” As we will shortly derive, the path loss of an antenna can be written as

$$L\ [\text{dB}] = -147.6 + 20\log_{10} f + 20\log_{10} d. \tag{7.1}$$

The loss in dB of free space only grows with the logarithm of the distance. The path loss in dB of a transmission line, however, is proportional to the distance. For 1 GHz the typical attenuation of an RG-58C/U is 1 dB/m, compared to only 0.05 dB/m at 10 MHz. Clearly, as Fig. 7.4 indicates, since the antenna loss in dB is proportional to $\log_{10} d$ but transmission line losses to d, at some point the antenna loss will be favorable. In the examples given at 10 MHz, wireless transmission is superior to wired transmission after one kilometer, while at 1 GHz, wireless is already better for less than 100 m.

7.2.2 Maxwell Equations

The most important fundamentals of this chapter are the Maxwell equations, named after their discoverer James Clerk Maxwell (1831–1879). However, we do not want to cover this rather theoretical aspect in detail and seek a somewhat more practical access to their implications for antenna and propagation. Just for reference, the Maxwell equations[1] are listed:

$$\operatorname{div} \boldsymbol{D} = \nabla \cdot \boldsymbol{D} = \rho_Q, \tag{7.2}$$

$$\operatorname{div} \boldsymbol{B} = \nabla \cdot \boldsymbol{B} = 0, \tag{7.3}$$

$$\operatorname{curl} \boldsymbol{E} = \nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{7.4}$$

$$\operatorname{curl} \boldsymbol{H} = \nabla \times \boldsymbol{H} = \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}, \tag{7.5}$$

with the relations of isotropic, lossless, dielectric and magnetic materials being

$$\boldsymbol{D} = \varepsilon \boldsymbol{E} = \varepsilon_0 \varepsilon_r \boldsymbol{E}, \tag{7.6}$$

$$\boldsymbol{B} = \mu \boldsymbol{H} = \mu_0 \mu_r \boldsymbol{H}. \tag{7.7}$$

Note that $\boldsymbol{E}$ and $\boldsymbol{H}$ denote the electric and magnetic field strength vectors, while $\boldsymbol{D}$ and $\boldsymbol{B}$ denote the electric displacement field and the magnetic flux density (magnetic induction) vectors, respectively.

More details on Maxwell's equations are contained in [22] and [35]. Eqs. (7.2) to (7.5) may be formulated in words [66]:

- *electric field lines may either start and end on charges, or are continuous*
- *magnetic field lines are continuous*
- *an electric field is produced by a time-varying magnetic field*
- *a magnetic field is produced by a time-varying electric field or by a current*

The last two equations, Eqs. (7.4) and (7.5), are referred to as Maxwell's curl equations.

[1] The operations used for the differential form of the Maxwell equations are operators in the vector field, $\operatorname{div} \boldsymbol{D} = \frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z}$ yields a scalar, while $\operatorname{curl} \boldsymbol{E} = (\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}, \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x}, \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y})$ gives as a result another vector field.

7.2.3 Electric and Magnetic Fields

An electromagnetic field is generated by an accelerated electric charge. Similar to a stone thrown into a lake generating a wave pattern, any disturbance of the electromagnetic field propagates. In accordance with Maxwell's equations (cf. Eqs. (7.2) to (7.5)), charges moved at a constant speed do not produce radiation. If, however, the charges are carriers of an RF oscillating signal, they radiate. To see how an antenna radiates, let us consider a setup of an open transmission line along the x dimension as in Fig. 7.5.

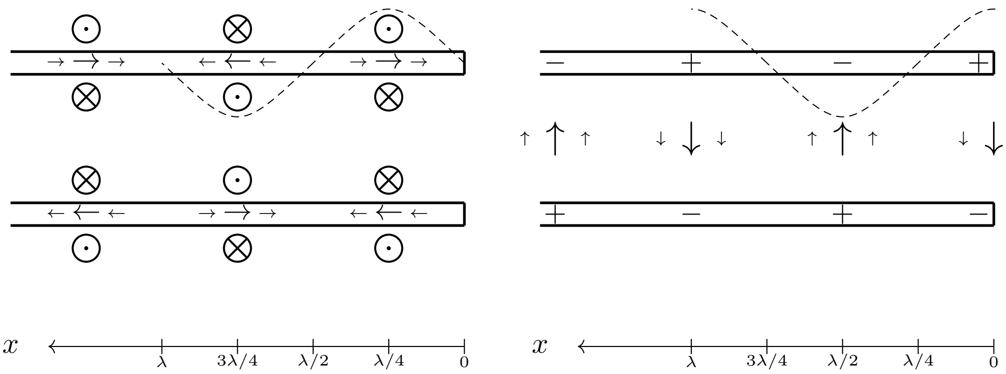


Figure 7.5 Open transmission line with standing-wave pattern. Left: Currents and magnetic field. The circled cross and dot represent arrow tail and head, respectively, of the magnetic field component. Right: Voltages and electric field.

Imagine that a standing-wave pattern has been established. The current at the open end must be zero, the voltage has a maximum at this point. Inbetween the two wires, the magnetic fields due to the opposite currents in the wires reinforce. Outside, from far away, there is no magnetic field, since the currents cancel in their influence on the magnetic field. Similarly with the voltage and the electric field. Inbetween the wires, we find electric field lines reaching from positive to negative charges. Outside (sufficiently far away), the charges cancel and we do not see an electric field. Thus, an open two-pole wire in first approximation (an assumption for which is that the wire spacing is much smaller than the wavelength) does not radiate.

Now, we bend the last quarter wavelength of the transmission line as in Fig. 7.6. The current of both bent ends now travels in the same direction, thus adds constructively to produce a magnetic field outside the line ends. Similarly, the electric field is no longer confined to the interior of the transmission line, but reaches from one end to the other over a distance of half a wavelength. Thus far, we have only considered one time

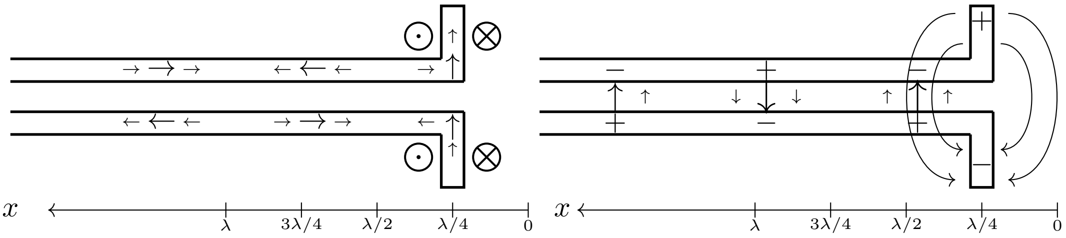
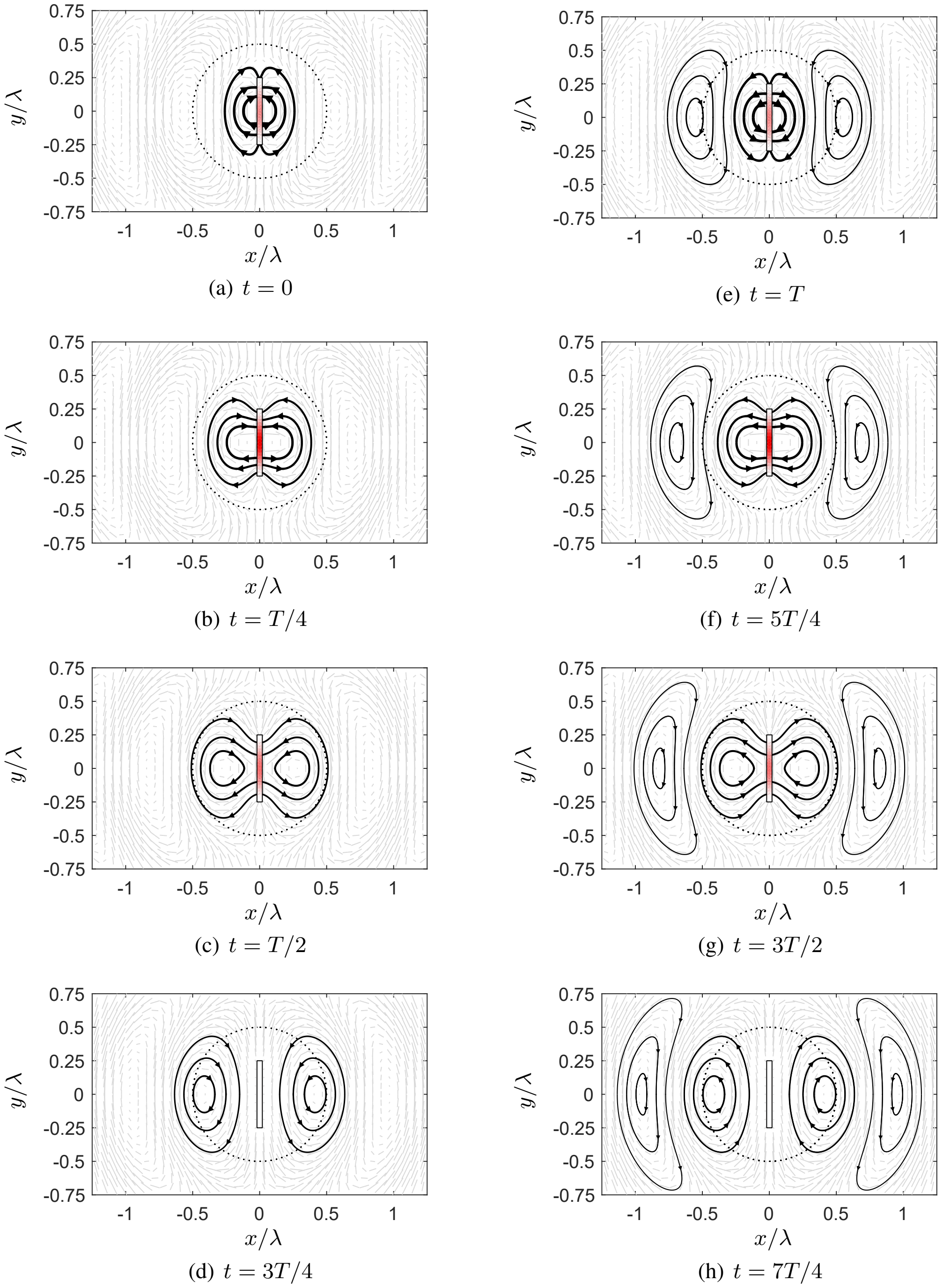


Figure 7.6 Open transmission line with last quarter-wavelength piece bent outwards. Left: Currents and magnetic field. The circled cross and dot represent arrow tail and head, respectively, of the magnetic field component. Right: Voltages and electric field.

instant. However, the standing wave pattern oscillates in time. Fig. 7.7 shows the formation of close E-field lines at times, when the voltage over the antenna disappears.



(a) $t = 0$

(e) $t = T$

(b) $t = T/4$

(f) $t = 5T/4$

(c) $t = T/2$

(g) $t = 3T/2$

(d) $t = 3T/4$

(h) $t = 7T/4$

Figure 7.7 Generation of a propagating E-field by an oscillating dipole.

What we have just developed is the $\lambda/2$-dipole. More on the dipole will be said when we talk about the different antenna types.

The space around the antenna is divided into a *near field* and a *far field*, the border radius of which is roughly given by

$$R = \frac{2L^2}{\lambda}, \tag{7.8}$$

where L designates the diameter of the antenna or the diameter of the smallest sphere containing the antenna, and λ is the wavelength. The *near field*, also called the *Fresnel region*, of an antenna is very difficult to compute, since it contains both radiating and reactive energy. The latter oscillates towards and away from the antenna. Thus, other objects sitting in the near field of an antenna can significantly change the impedance and radiation characteristics. In the *far field*, also called *Fraunhofer region*, we only find radiating energy (no reactive components). The far field is therefore easier to compute and most analysis (link budget etc.) assume the far field. The wave fronts in this region appear as spherical waves. In the far field of an antenna, the electric and magnetic components stand in a special relationship to each other. To see this we look at the one-dimensional wave equation similar to the Telegraph equation. The two Maxwell equations for Ampere's law and Faraday's law are in vacuum

$$\nabla \times \boldsymbol{H} = \varepsilon_0 \frac{\partial \boldsymbol{E}}{\partial t} \tag{7.9}$$

and

$$\nabla \times \boldsymbol{E} = -\mu_0 \frac{\partial \boldsymbol{H}}{\partial t}, \tag{7.10}$$

respectively. In the far field we assume a TEM wave with E_x, H_y propagating in the z direction. Hence, Eqs. (7.9) and (7.10) reduce to

$$\frac{\partial H_y}{\partial z} = -\varepsilon_0 \frac{\partial E_x}{\partial t}, \tag{7.11}$$

$$\frac{\partial E_x}{\partial z} = -\mu_0 \frac{\partial H_y}{\partial t}. \tag{7.12}$$

Now, further differentiation of Eq. (7.11) with respect to t and Eq. (7.12) with respect to z leads to

$$\frac{\partial}{\partial t}\frac{\partial H_y}{\partial z} = -\varepsilon_0 \frac{\partial^2 E_x}{\partial t^2}, \tag{7.13}$$

$$\frac{\partial^2 E_x}{\partial z^2} = -\mu_0 \frac{\partial}{\partial z}\frac{\partial H_y}{\partial t}. \tag{7.14}$$

Eq. (7.13) can now be used in Eq. (7.14) to yield

$$\frac{\partial^2 E_x}{\partial z^2} = -\varepsilon_0 \mu_0 \frac{\partial^2 E_x}{\partial t^2}. \tag{7.15}$$

Considering forward-travelling waves only, a solution to this one-dimensional wave equation is provided by

$$E(z,t) = E^+ \mathrm{e}^{-jkz} \mathrm{e}^{j\omega t}, \tag{7.16}$$

where k is the wavenumber. From Eq. (7.15) we see that

$$-k^2 E(z,t) = -\omega^2 \varepsilon_0 \mu_0 E(z,t), \tag{7.17}$$

so that the wavenumber can be written as

$$k = \omega\sqrt{\varepsilon_0 \mu_0}. \tag{7.18}$$

Similarly of course,

$$H(z,t) = H^{+}\mathrm{e}^{-jkz}\mathrm{e}^{j\omega t}. \tag{7.19}$$

Eq. (7.13) now becomes

$$-jkH^{+}\mathrm{e}^{-jkz}\mathrm{e}^{j\omega z} = -j\varepsilon_0\omega E^{+}\mathrm{e}^{-jkz}\mathrm{e}^{j\omega z}. \tag{7.20}$$

The division of electric and magnetic field delivers[2]

$$\boxed{Z = \frac{|\boldsymbol{E}|}{|\boldsymbol{H}|} = \frac{E^{+}}{H^{+}} = \frac{k}{\varepsilon_0\omega} = \sqrt{\frac{\mu_0}{\varepsilon_0}} = 120\pi\,\Omega = 377\,\Omega}. \tag{7.21}$$

Z is called the free-space impedance. For far-field observations, it is thus always sufficient to compute or measure either the electric or the magnetic field and compute the other one using Eq. (7.21). The relationship between electric field and power density is similar to the one for voltage and power in ordinary circuits, and likewise the analogy between magnetic field and current. Hence, we can write the power density S as

$$\boxed{S = \frac{|\boldsymbol{E}|^2}{Z} = |\boldsymbol{H}|^2 Z = |\boldsymbol{E}||\boldsymbol{H}|}. \tag{7.22}$$

Now the power is simply the power density times the area A over which the power density is collected

$$P = SA = \frac{|\boldsymbol{E}|^2}{Z}A = |\boldsymbol{H}|^2 ZA. \tag{7.23}$$

We will see in Section 7.3.2 exactly how the area is determined.

7.2.4 Reciprocity

Interestingly, the properties or characteristics of an antenna remain the same, regardless whether the antenna is used to transmit or to receive. This is called the principle of *reciprocity*.

7.3 Antenna Characterization

Before we look at different types of antenna, we have to find measures to characterize antennas. Besides several physical properties such as size and shape, many electrical and field characteristics can be distinguished for different antenna types. This section shall contain a list of parameters and their respective description.

7.3.1 Radiation Pattern, Beamwidth, and Directivity

Before we can quantify some radiation pattern parameters, we need to define a sensible reference coordinate system. The spherical coordinate system with (r, θ, ϕ) has proven very useful. The whole reference coordinate system is shown in Fig. 7.8. The antenna is usually situated at the origin of the xyz-system. The radius r determines the distance of the measurement point from the antenna. The angle ϕ is the so-called azimuth angle, while θ describes the elevation angle. The radiation pattern always refers to the far field of an

[2] This result can also be obtained from the following comparison. Since in free-space we have a TEM wave similar to the situation in a transmission line we can also consider the impedance as $Z = \sqrt{\frac{L}{C}} = \sqrt{\frac{L/l}{C/l}} = \sqrt{\frac{\mu_0}{\varepsilon_0}}$.

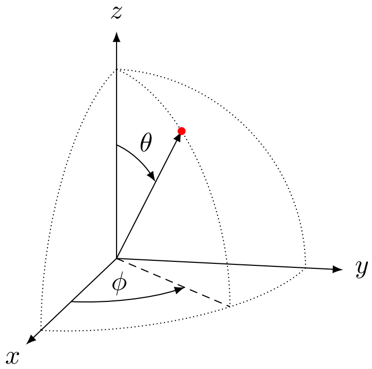


Figure 7.8 Reference coordinate system for antenna radiation measurements.

antenna. It displays the power radiated from an antenna per unit solid angle. This is the radiation intensity U, which is the power density at a given point multiplied by the squared distance r^2

$$U = r^2 S. \tag{7.24}$$

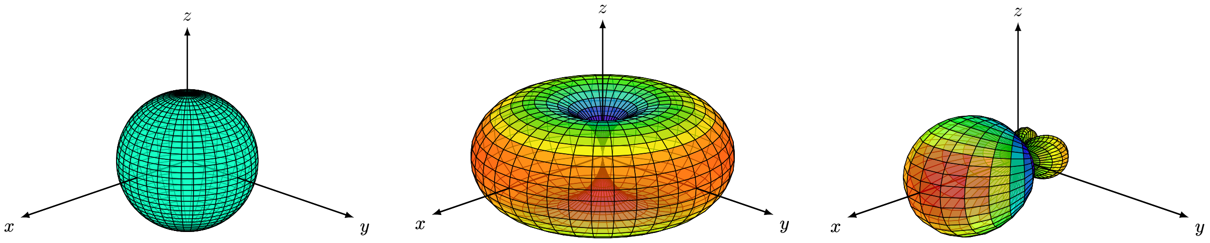
This way, the radiation pattern is independent of the distance it is evaluated. At this point, we have to introduce a theoretical reference antenna: the *isotropic radiator*. The isotropic radiator is a fictive antenna that radiates equally in all directions, as shown in Fig. 7.9(a). It cannot be built in practice, but it is a useful concept, as it facilitates the comparison with other, more practical antennas. For the isotropic antenna, the power density is given by

$$\boxed{S = \frac{P_{\text{rad}}}{4\pi r^2}}, \tag{7.25}$$

hence, the radiation intensity is

$$U = r^2 S = \frac{P_{\text{rad}}}{4\pi}. \tag{7.26}$$

P_{rad} is the radiation power. Most real antenna radiation patterns are functions of the angles θ and ϕ. An example of the radiation pattern of a Hertzian dipole (infinitesimally small dipole) is given in Fig. 7.9(b).The normalized radiation function is in this case simply given by $U/U_{\text{max}} = \sin^2\theta$.



(a) Isotropic (hypothetical) (b) Omnidirectional (e.g. Hertzian dipole) (c) Directive (e.g. Yagi-Uda antenna)

Figure 7.9 The three types of radiation patterns.

A dipole is an example of an *omnidirectional antenna*, radiating the same amount of power in any direction in one plane around the antenna. Such an antenna is desirable for mobile handsets for example, since at transmission the directional position of the handset is unknown.

Finally, Fig. 7.9(c) shows the typical radiation pattern of a *directional antenna*, which has a dominant power radiation in one direction. Directional antennas are very practical for base-stations, when only a certain sector of a cell has to be served. The directivity of an antenna is defined by its ratio of maximum power density (in the direction it radiates the most power) and the average power density.

Since three-dimensional plots with both parameters (elevation θ and azimuth ϕ) varied are not always practical, there are different parameterized two-dimensional plots of the radiation pattern. Two main plots are in wide use: The *principal plane cut* also known as the *θ cut* or the *elevation cut*, this antenna measurement type fixes the value of ϕ (parameter). The angle θ is then swept to cover an entire sweep of the upper hemisphere ($-\pi/2$ to $\pi/2$). The second type is the so-called *conic cut*, also known as the *ϕ cut* or the *azimuth cut*. This antenna measurement type fixes the value of θ (parameter). The angle ϕ is then swept to cover the full circle ($-\pi$ to π).

In order to compare radiation patterns, some parameters are defined (for an illustration consult Fig. 7.10, which can be either the elevation cut or the azimuth cut):

- The *half-power beamwidth* (HPBW) or simply the *beamwidth* is the angle from one half-power point to the other one in the main lobe.
- The *front-back ratio* is the ratio between the peak amplitudes of the main lobe and the back lobe.
- The *sidelobe level* is the amplitude of the largest sidelobe in relation to the peak of the main lobe.

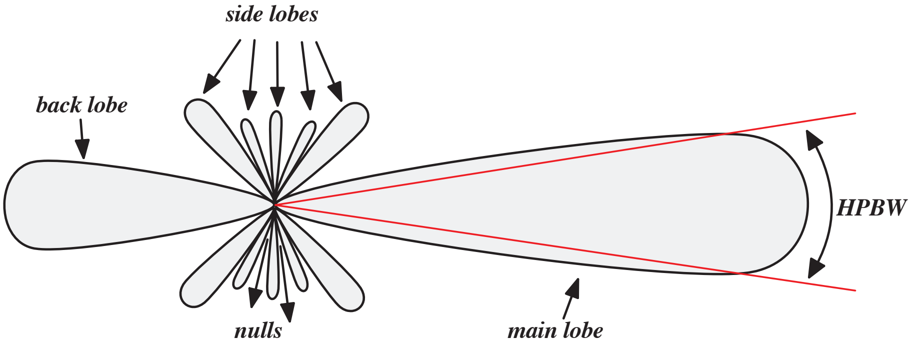


Figure 7.10 Some parameters differentiating radiation patterns.

The *directivity* of an antenna is a function of the two angles θ and ϕ, namely the ratio of radiation intensity in the direction indicated over the mean radiation intensity in all directions

$$D(\theta, \phi) = \frac{U(\theta, \phi)}{\bar{U}} = \frac{4\pi U(\theta, \phi)}{P_{\text{rad}}}. \tag{7.27}$$

If no directions are specified, the directivity refers to the angles of maximum radiation intensity

$$D_{\max} = \frac{4\pi U_{\max}}{P_{\text{rad}}}. \tag{7.28}$$

In terms of directivity, we distinguish three types of antennas: the isotropic, the omnidirectional, and the directional antenna. The *isotropic antenna* is a ficticious object defined for ease of reference and radiates the same power in any given direction. Now imagine placing a sphere around an isotropic antenna. The total power that crosses the sphere surfaces is always the same—the power cannot escape—regardless of the size of the sphere. Hence, if we place spheres of increasing radii around an isotropic radiator, the power density is the power radiated divided by the surface area of the antenna and is thus related to the distance away from the antenna by

$$S = \frac{P_{\text{rad}}}{4\pi r^2}.$$

which was shown in Eq. (7.25).

7.3.2 Antenna Area

The effective antenna area, often called *effective aperture*, is the area that the power density (around the antenna) has to be multiplied with to get the power delivered to an optimally tuned receiver, hence

$$P_{\text{rx}} = A_{\text{eff}}\, S. \tag{7.29}$$

A physical interpretation of the antenna area makes sense only for some antenna types. Typical examples are horn antennas, where we can see the opening area, or dish antennas, which are very large compared to λ and have a visible area as well.

In most cases, the effective aperture is smaller or equal to the physical aperture, i.e. $A_{\text{eff}} \leq A_{\text{phys}}$, as one might expect. However, the effective aperture can exceed the physical area: for example for loop antennas the physical aperture often has very little to do with the effective aperture. For other antenna types, such as dipole antennas, it is often difficult to see a physical area in the first place. Nevertheless, the concept is valid for all antenna types and it is very common to use the effective aperture of any antenna, irrespective of its shape or type.

In general, the effective antenna area is given by

$$\boxed{A_{\text{eff}} = G\, A_{\text{eff}\,\text{iso}} = \frac{\lambda^2 G}{4\pi}}\,, \tag{7.30}$$

where λ is the (free-space) wavelength and G is the antenna gain, a parameter we shall meet in due course and which basically states how much larger (or smaller) the effective aperture of an antenna is, compared to the hypothetical isotropic radiator, $A_{\text{eff}\,\text{iso}} = \lambda^2/(4\pi)$.

7.3.3 Effective Length

The maximum voltage that an antenna can produce from a surrounding electromagnetic field is proportional to the intensity of that field. For example using the electric field we can state:

$$V_0 \propto E\,. \tag{7.31}$$

The proportionality constant is called *effective length* of the antenna, l_{eff}, since the electric field in V/m has to be multiplied by a length in m to obtain a voltage in V. Since the electric field is a vector quantity, it is also often used in vector form and then called *vector effective length*, $\boldsymbol{l}_{\text{eff}}$:

$$V_0 = l_{\text{eff}}\, E \tag{7.32a}$$

or

$$V_0 = \boldsymbol{l}_{\text{eff}} \cdot \boldsymbol{E}\,. \tag{7.32b}$$

Note that these relations are similar but not the same, as the latter requires the antenna to be aligned to the electric field for the scalar product to be maximized. Therefore, that formulation already accounts for any misalignment, most importantly polarization mismatch.

For some antenna types, the effective length can be calculated or at least estimated directly. For linear antennas (regular wire antennas such as dipoles and monopoles), the effective length is given by the normalized current integral on the structure (the physical length l_{phys}):

$$l_{\text{eff}} = \int_{l_{\text{phys}}} \frac{I(kx)}{I(kx_0)}\, dx \tag{7.33}$$

where $k = 2\pi/\lambda$ is the wave number and x_0 is the feed point position on l_{phys}. For example for a typical center-fed ($x_0 = 0$) half-wave dipole with a cosine current distribution, we obtain:

$$l_{\text{eff}}^{\lambda/2\text{-dipole}} = \int_{-\lambda/4}^{\lambda/4} \frac{I(kx)}{I(0)}\, dx = \int_{-\lambda/4}^{\lambda/4} \frac{I_0\, \cos\frac{2\pi x}{\lambda}}{I_0}\, dx = \left[\frac{\lambda\, \sin(2\pi x/\lambda)}{2\pi}\right]_{-\lambda/4}^{\lambda/4} = \frac{\lambda}{\pi}\,. \tag{7.34}$$

Thus, the effective length of a half-wave dipole is somewhat smaller than its physical length, i.e. $l_{\text{eff}}^{\lambda/2\text{-dipole}} = \lambda/\pi < l_{\text{phys}}^{\lambda/2\text{-dipole}} = \lambda/2$. This is due to the fact that the current density is not constant along the entire antenna; the parts where the current is larger radiate (or receive) more, whereas the areas with lower currents are less effective.

As mentioned, V_0 is the maximum voltage at the antenna terminals; thus, the open-connected voltage. Usually, particularly for all receiving cases, we want to match the antenna so as to transfer the maximum amount of power to the receiver, leading to the voltage

$$V = \frac{V_0}{2} = \frac{l_{\text{eff}}}{2}\, E \tag{7.35}$$

and the corresponding power

$$P_{\text{rx}} = \frac{|V|^2}{R_{\text{rad}}} = \frac{l_{\text{eff}}^2}{4}\,\frac{|E|^2}{R_{\text{rad}}}\,. \tag{7.36}$$

By equating this relation with the formulation using the effective aperture (7.29) we find

$$A_{\text{eff}}\, S = \frac{l_{\text{eff}}^2}{4}\,\frac{|E|^2}{R_{\text{rad}}}\,. \tag{7.37}$$

Moreover, using the free-space wave impedance $Z = 120\,\pi$ (in Ohms) and following Eq. (7.22), the power density S can be replaced by $S = |E|^2/Z$ and we find the correspondences of the effective aperture and length of an antenna as follows:

$$\boxed{A_{\text{eff}} = \frac{l_{\text{eff}}^2}{4}\,\frac{Z}{R_{\text{rad}}} \quad \Leftrightarrow \quad l_{\text{eff}} = 2\sqrt{A_{\text{eff}}\,\frac{R_{\text{rad}}}{Z}}} \tag{7.38}$$

In other words, the effective aperture is not the squared effective length, but rather the squared half of the length, multiplied by a factor given by the quotient of the antenna radiation resistance and the free-space wavelength.

7.3.4 Antenna Factor

Yet another parameter sometimes used to characterize an antenna is the *antenna factor*, which describes the ratio of the electric field and the voltage related to 50 Ω

$$K = \frac{E}{V}. \tag{7.39}$$

Other symbols used for the antenna factor are AF or A_F and it is usually given in dB/m. From Eq. (7.35) we get

$$K = \frac{2}{l_{\text{eff}}} \tag{7.40}$$

or from Eqs. (7.35) and (7.30)

$$K = \sqrt{\frac{Z}{A_{\text{eff}} R_{\text{rad}}}} = \sqrt{\frac{120\pi \cdot 4\pi}{\lambda^2 G R_{\text{rad}}}} = \sqrt{\frac{480\Omega}{G R_{\text{rad}}}} \frac{\pi f}{c} \tag{7.41}$$

or in dB and for $R = 50\,\Omega$

$$\begin{aligned} K_{\text{dB/m}} &= 20 \log_{10} \sqrt{\frac{480}{50}} \frac{\pi}{c} + 20 \log_{10} f - G_{\text{dBi}} \\ &= 20 \log_{10} f_{\text{MHz}} - 29.78 - G_{\text{dBi}}. \end{aligned} \tag{7.42}$$

7.3.5 Polarization

The electric field, the magnetic field and the Poynting vector (direction of propagation) build a right-hand system, i.e., the direction of the propagation is always perpendicular to each the electric field and the magnetic field, see also Fig. 7.11. In fact, the absolute value of the Poynting vector $\boldsymbol{S}$ is the power density, which we have met in Eq. (7.22). From the relationship between the E- and H-fields and the Poynting vector,

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H}^*, \tag{7.43}$$

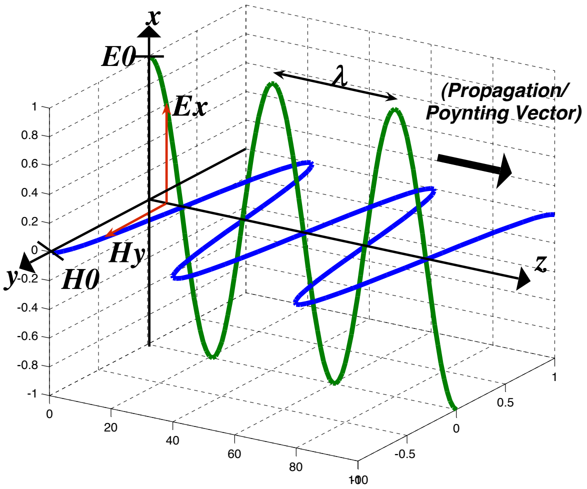


Figure 7.11 E-field, H-field, and direction of propagation.

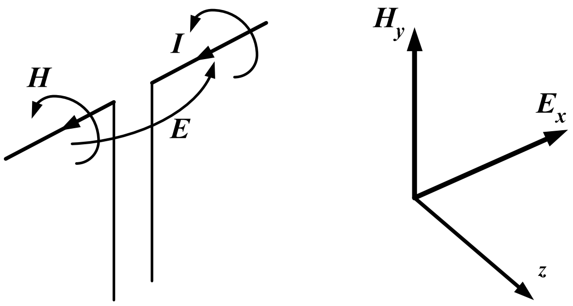
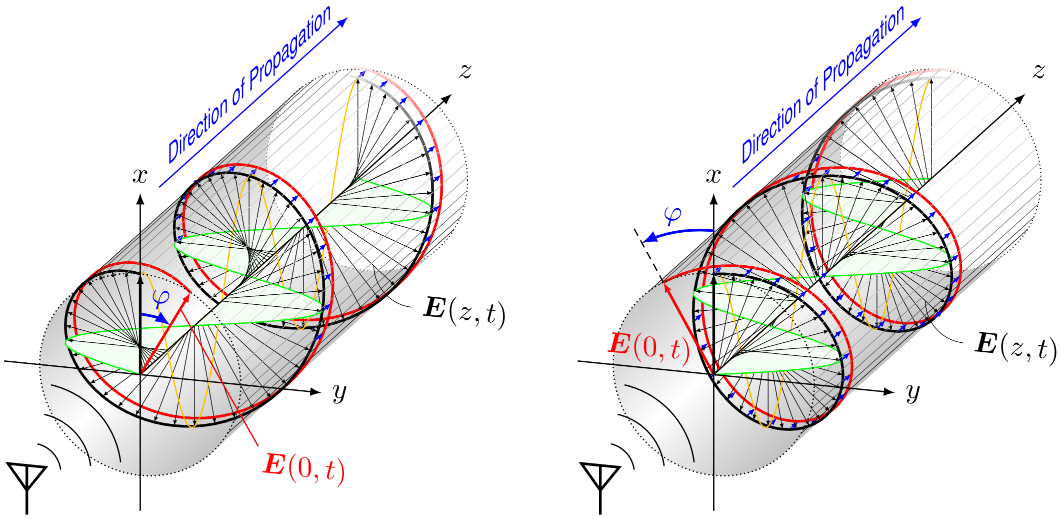


Figure 7.12 Polarization of a dipole antenna.



(a) RHCP

(b) LHCP

Figure 7.13 Illustrations of the two circularly polarized wave types, from the perspective of a transmitting antenna: In each z-plane, the field vector revolves around the axis of propagation; for (a) RHCP in the "‘right"’ and for (b) LHCP in the "‘left"’ direction, respectively. Along the direction of propagation, the tip of the field vectors form a spiral which moves forward as time progresses.

we now see that Eq. (7.22) was simply a consequence of the fact that the E- and H-field components are orthogonal to each other.

The direction of the electric field relative to the earth is called the polarization of an antenna. The polarization can be vertical, horizontal, right-hand circular, and left-hand circular. Circular polarization is a rotation of the polarization direction as the wave propagates. It can either be produced by phase-delayed second feed point of the antenna or by a circularly polarized antenna such as a helical antenna. Most patch antennas as used in GPS receivers achieve circular polarization with only one feed point slightly off the diagonal and an asymmetric side-length design. Note that the definition of right-hand and left-hand circular polarization is somewhat confusing. The engineering convention of circular polarization states that a right-hand circularly polarized wave is defined as one whose electric field is rotating clockwise ("‘to the right"’) as seen by a transmitting antenna. Likewise, a wave is left-hand circularly polarized if the electric field is rotating counterclockwise as it propagates away from the transmitting antenna. Examples of linear polarization and circular polarization are given by Fig. 7.12 and Fig. 7.13, respectively.

7.3.6 Antenna Resistance, Efficiency, and Gain

An antenna cannot radiate more power than is transmitted to it. The fraction of what is radiated is called efficiency

$$\eta = \frac{P_{\text{rad}}}{P_{\text{input}}}, \tag{7.44}$$

where P_{input} is the power input to the antenna (perfect matching assumed). The efficiency is never 100 %. Some power is always dissipated due to Ohmic losses. The power radiated is given by the radiation resistance R_{rad}, which should be as large as possible compared to the loss resistance R_{loss}. See Fig. 7.14 for an equivalent circuit of an antenna including loss resistance and radiation resistance. The efficiency of an antenna is usually given [5] by

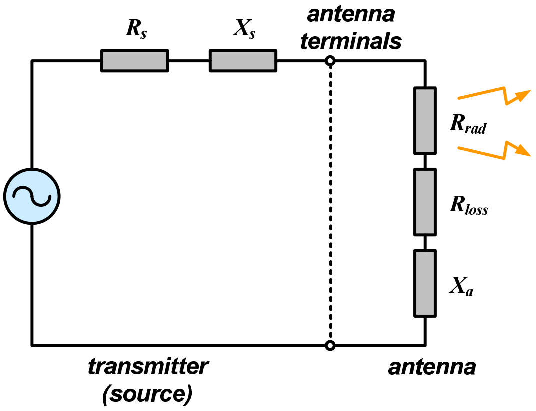


Figure 7.14 Equivalent circuit of a transmitting antenna.

$$\boxed{\eta = \frac{P_{\text{rad}}}{P_{\text{input}}} = \frac{R_{\text{rad}} \cdot I^2}{(R_{\text{rad}} + R_{\text{loss}}) \cdot I^2} = \frac{R_{\text{rad}}}{R_{\text{rad}} + R_{\text{loss}}}}. \tag{7.45}$$

The radiation resistance is a bit like the internal resistance of a source and cannot easily be controlled. It depends a lot on the environment of the antenna such as proximity to conducting surfaces. Although high antenna efficiency is a goal to achieve, an even more important objective is to match the antenna, otherwise a lot of power is reflected and does not even reach the antenna. In order to match the antenna, it is important to know the impedance at the feeding point of the antenna, which is not automatically given by the radiation and loss resistances but depends heavily on the position of the feed.

The combination of efficiency and (maximum) directivity of an antenna (given by the parameter D) is called antenna gain

$$G = \eta \cdot D. \tag{7.46}$$

If we characterize the gain of an antenna and use it in calculations such as link budget analysis, we usually assume that the antenna is operated in its favorable direction, so the gain always refers to the main lobe. When comparing the directivity or gain to that of an isotropic radiator, the gain of an antenna receives the unit [dBi]. Very often, gain is compared to the $\lambda/2$-dipole, which already shows some directivity. The gain then has the unit [dBd] and is smaller by 2.15 dB, i.e.,

$$0\,\text{dBd} = 2.15\,\text{dBi}. \tag{7.47}$$

7.3.7 Bandwidth

The bandwidth of an antenna is the frequency range over which the antenna impedance remains roughly constant. As a parameter, the bandwidth is often defined as the range over which the power gain does not vary more than 3 dB, or for which the VSWR is below 2:1. Often, the bandwidth is specified as a percentage of the operating frequency. Obviously, for wide-band signals or if several signals in different frequency bands have to be transmitted/received through the same antenna, the bandwidth should be wide. A small bandwidth, on the other hand, offers a lot of selection and might relax the specification of proceeding filters. However, some pitfalls are lurking here. For a small bandwidth, the antenna is much more prone to mismatch due to environmental change, e.g., the proximity of a mobile phone to its user's body, hand, and head.

7.4 Antenna Types

7.4.1 Dipole

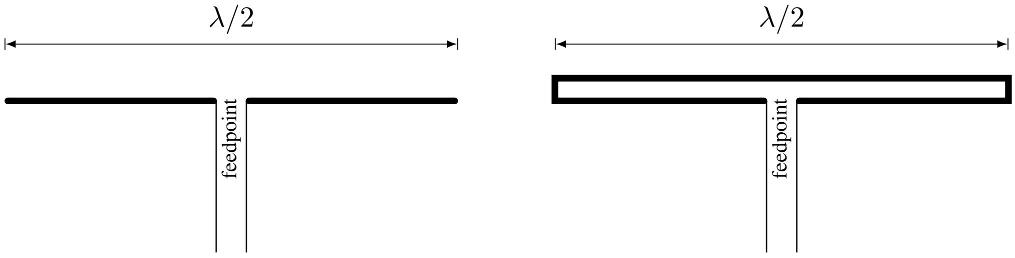
Dipoles belong to the most prominent antennas. Often the length is a multiple of half the wavelength. On the other hand, very short dipoles of length $l \ll \lambda$ are considered for computational models. An infinitesimally small dipole is also called a Hertzian dipole after Heinrich Hertz, pioneer in the wireless field. It can be shown [5] that the far-field distribution ($kr \gg 1$) of a small dipole of length l and current I_0 looks as

$$E_\vartheta = jZ_0 \frac{kI_0 l \mathrm{e}^{-jkr}}{8\pi r} \sin\vartheta, \tag{7.48a}$$

$$H_\varphi = j \frac{kI_0 l \mathrm{e}^{-jkr}}{8\pi r} \sin\vartheta. \tag{7.48b}$$

Other field components vanish due to the dependency on a higher order of r, i.e., $1/r^2$ or even $1/r^3$.

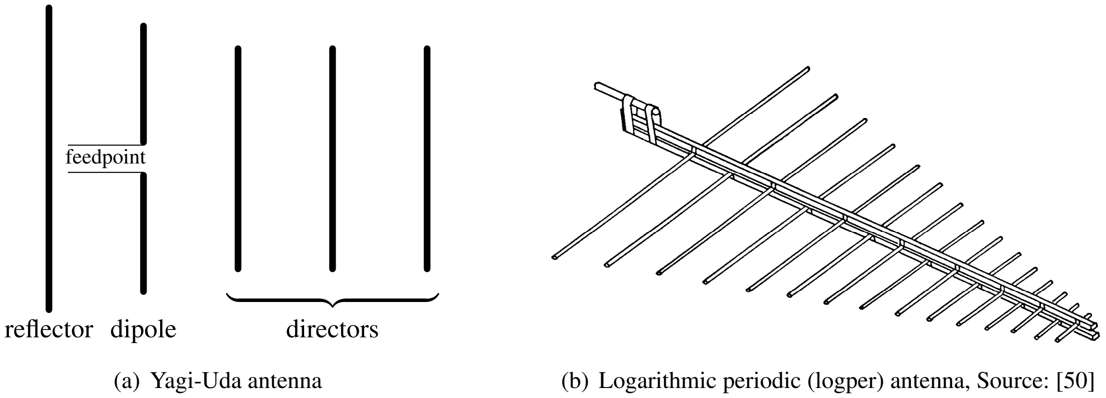
Of special interest is the so-called half-wave dipole, see Fig. 7.15(a). It is the smallest self-resonant antenna structure, i.e., its conductor is one-half wavelength long at the operating frequency. Thus, the half-wave dipole has a sinusoidal distribution of its current along the conductor. The mechanical length of a dipole is usually made slightly shorter than half a wavelength to make it exactly resonant. The impedance is then 73 Ω. A dipole must be fed by balanced lines. Another variant of the dipole is the *folded dipole*, which looks like a very flat loop. Also very often used together with the dipole is the Yagi-Uda antenna, which looks like an array of dipoles of decreasing size. In fact, only one of the conductors is a dipole (usually the second but last, see Fig. 7.16(a)), the one behind the dipole is a reflector (slightly larger in physical size than the



(a) Normal dipole (b) Folded dipole

Figure 7.15 Dipole antennas.

dipole) and the ones in front are directors (usually smaller in size than the dipole), which make the whole setup highly directive. This antenna is often used for television broadcasting in the VHF and UHF regions. Closely related in structure and appearance to the Yagi antenna is the logarithmic periodic antenna (logper antenna), see Fig. 7.16(b). The difference is that the logper antenna is a wideband antenna, whereas the Yagi antenna usually concentrates on a smaller band.



(a) Yagi-Uda antenna

(b) Logarithmic periodic (logper) antenna, Source: [50]

Figure 7.16 Larger antennas (arrays) consisting of dipoles.

7.4.2 Monopole

If the symmetric arrangement of a dipole is imitated by one half-side and a virtual mirror, we get a monopole, see Fig. 7.17. The mirror is thereby a ground plane perpendicular to the monopole. In other words, a monopole is a dipole with half of its structure replaced by an image created by the ground plan. The length of a resonant monopole is therefore $\lambda/4$. A monopole has the additional advantage that it can be fed in an unbalanced way. Since the monopole can concentrate on one hemisphere only, its gain is 3 dBd.

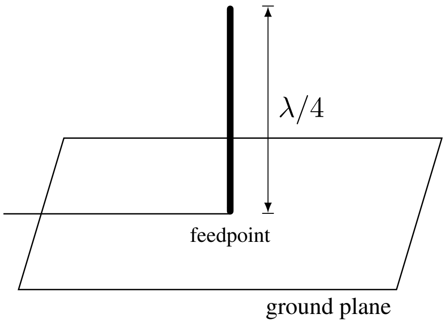


Figure 7.17 Monopole antenna.

7.4.3 Loop Antenna

Although resonant loops, where the circumference corresponds to a wavelength or odd multiples thereof, are in use, most applications deploy a small loop, where the current distribution can be considered constant.

In the latter case the loop antenna with current I_0 lying in the horizontal plan can be considered as the dual form of the vertical Hertzian dipole, as the field distributions in the far field ($kr \gg 1$) show [5]

$$H_\vartheta = \frac{k^2 I_0 a\, \mathrm{e}^{-jkr}}{4r} \sin\vartheta, \tag{7.49a}$$

$$E_\varphi = Z_0 \frac{k^2 I_0 a\, \mathrm{e}^{-jkr}}{4r} \sin\vartheta. \tag{7.49b}$$

Again, other field components vanish due to the dependency on a higher order of r, i.e., $1/r^2$ or even $1/r^3$. A comparison of the wave propagation of small loops and resonant loops is carried out in the following. The radiation patterns are shown in Fig. 7.18. It is interesting to note that while the small loop antenna has a null in the propagation of the null axis, the resonant loop propagates in this direction most strongly.

Small loop antennas (see Fig. 7.19), despite their inefficiency, have found wide application in personal communication devices where space is tight, e.g., pagers. They are mostly built as rectangular metal strip

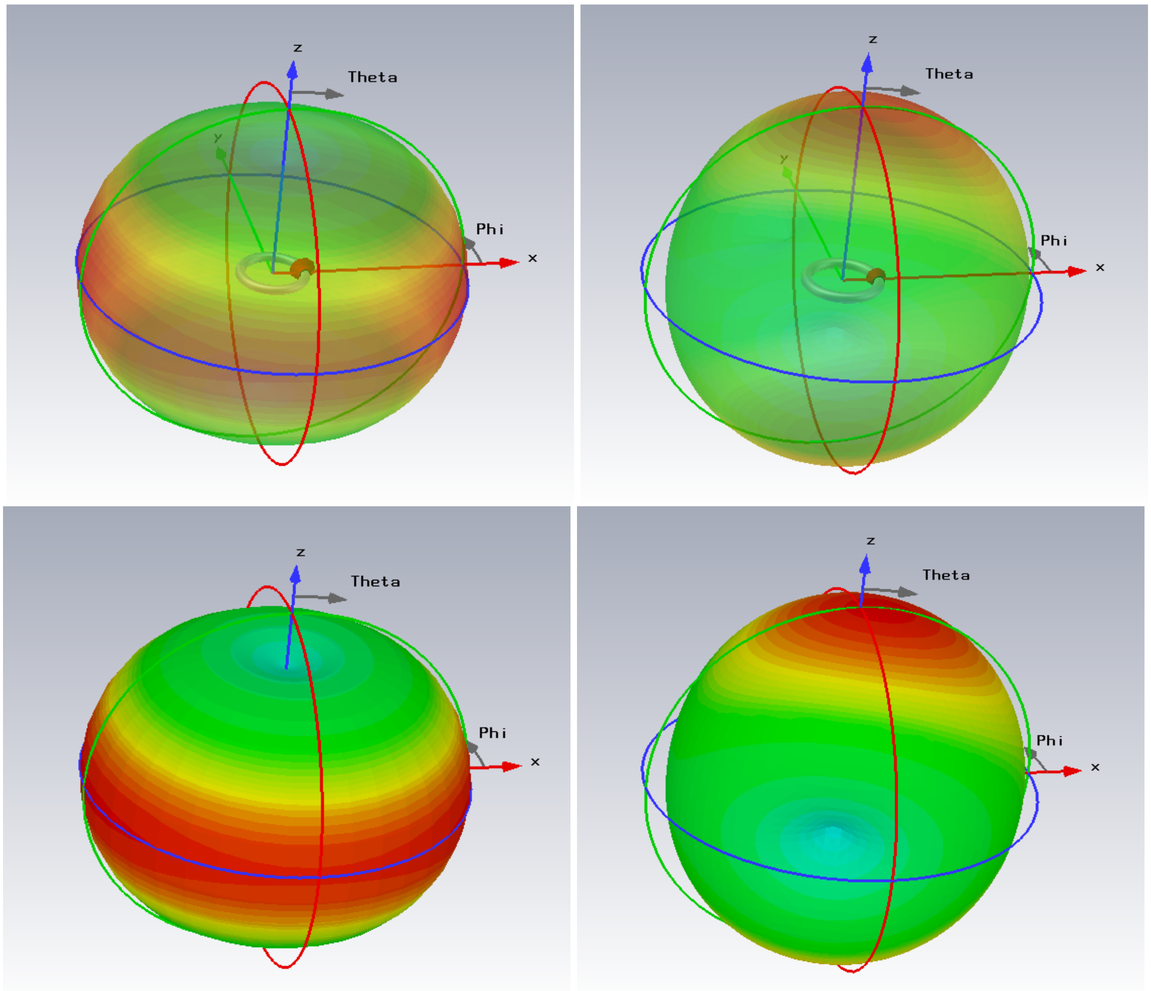


Figure 7.18 Simulated farfield radiation patterns of the loop antenna, operated at different frequencies to make it appear electrically small or resonant, respectively. Figures in the left column: Farfield radiation pattern of the electrically small loop (at 3 GHz). Figures in the right column: Farfield radiation pattern of the electrically large loop (at 30 GHz). The top and bottom figures represent the same simulation result with a different degree of transparency to see the loop structure.

loops or as printed circuits. The problem of finding an optimal shape within a limited space equally applies to either form. A loop antenna is considered small if the loop diameter is small enough with respect to the wavelength for the current to be assumed constant around its circumference. The radiation resistance is proportional to the squared aperture of the loop antenna

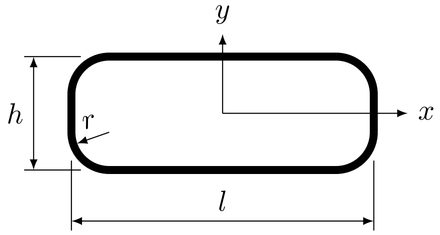


Figure 7.19 Drawing of rectangular loop antenna shape with rounded corners.

$$R_{\text{rad}} = 31171\,\Omega \cdot \left(\frac{A}{\lambda^2}\right)^2 = k_1 \cdot A^2, \tag{7.50}$$

where A denotes the aperture or area of the loop antenna and λ is the wavelength of the frequency under consideration. The loss resistance is the Ohmic resistance of the metal and therefore proportional to the circumference c of the loop

$$R_{\text{loss}} = k_2 \cdot c. \tag{7.51}$$

Thus, Eq. (7.45) can be written as (with $k = k_2/k_1$)

$$\eta = \frac{A^2}{A^2 + k \cdot c} = \left(1 + k \cdot \frac{c}{A^2}\right)^{-1}. \tag{7.52}$$

k_1, k_2, and k are constants. The optimal shape of a loop antenna under an area constraint is derived in [40].

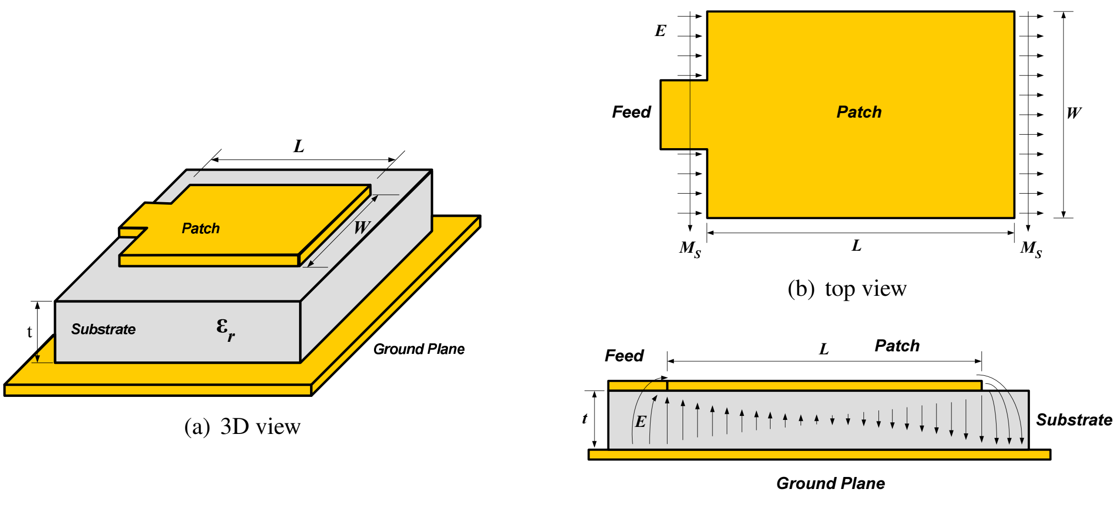
Sometimes the small loop antenna is called a magnetic dipole, since in the near-field the loop stores most of its energy in the magnetic field, in contrast to short dipoles, which store their energy in the near-field in the electric field. Loop antennas may be equipped with ferromagnetic material inside the loop leading to so-called ferrite rods.

7.4.4 Patch Antenna

Patch antennas or microstrip antennas are becoming increasingly popular for use in mobile communications and GPS equipment. They consist of a square or round patch of conductor over a ground plane. Between the patch and the ground plane, there is usually dielectric substrate; hence, the implementation of a patch antenna on a PCB is straightforward. For geometry and fields of a patch antenna see Fig. 7.20. Because the patch antenna is a resonant antenna (usually one side L is roughly $\lambda/2$), its bandwidth is very limited, which makes it difficult for systems that support dual-band or even triple-band modes.

One side corresponds to $\lambda/2$. Due to edge effects (the field is extended somewhat making the structure longer), the length of the antenna is to be chosen slightly smaller than $\lambda/2$ [77],

$$L \approx 0.49\,\frac{\lambda}{\sqrt{\varepsilon_r}}, \tag{7.53}$$



(a) 3D view

(b) top view

(c) upright projection

Figure 7.20 Patch antenna.

where λ is the wavelength and ε_r is the relative permittivity of the PCB (substrate). The width W of the patch antenna is given by impedance considerations [77]. The impedance of a patch antenna is given as

$$Z = 90 \frac{\varepsilon_r^2}{\varepsilon_r - 1} \left(\frac{L}{W} \right)^2 \Omega, \tag{7.54}$$

to be measured in the middle of the "W"-edge as shown on top of Fig. 7.20. This type of excitation is the most simple and referred to as *direct feeding*. This is just one of four general feeding types for patch antennas, the others being coaxial feeding, proximity feeding, and aperture coupling. All four are illustrated in Fig. 7.21.

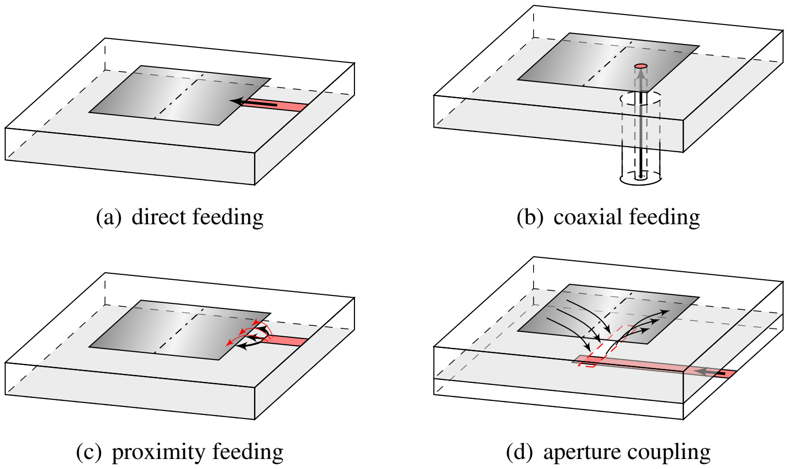
- Direct feeding: As mentioned, this simply means connecting the patch edge directly to a microstrip feeding line, as depicted in Fig. 7.21(a). The feeding structure and the patch antenna are placed on the same layer, above a ground plane, accounting for the lowest manufacturing complexity and costs. On the drawback side, we have a narrow bandwidth, radiation pattern perturbation by the feeding structure, and only limited impedance-matching capabilities. The relative bandwidth of the direct-fed patch follows the well-known rule of thumb [4]:

$$B \propto \frac{h_{\text{sub}}}{\sqrt{\varepsilon_{r\,\text{eff}}}} \tag{7.55}$$

  Therefore, to increase the impedance bandwidth of a patch antenna, thick substrates with low (relative) dielectric permittivity should be used. This makes the antenna large, however.

- Coaxial feeding: A commonly found method is coaxial feeding, where the patch is connected by a coaxial probe (or via) from below. An evident problem is this hole through the ground plane, which tends to increase manufacturing costs, especially for robust ceramic high-permittivity materials. It might therefore be more expensive to manufacture and has the same narrow bandwidth. However, advantages exist in the form of little to no spurious radiation, more impedance matching capabilities.

  Since the feeding structure is not on the same layer as the patch, thicker substrates can be considered than with direct feeding, without lowering the efficiency of the feeding lines.



(a) direct feeding

(b) coaxial feeding

(c) proximity feeding

(d) aperture coupling

Figure 7.21 The four basic (linear polarized) patch antenna excitation techniques.

The exact position of the feeding point determines the impedance of the antenna. It varies from high-Ohmic levels (R_{edge} typ. 300 Ω) on the patch edge to zero at its center. The progression in between can by approximated by [23]

$$R_{\text{feed}} = R_{\text{edge}} \cos^2 \frac{\pi \, x_{\text{feed}}}{L} \tag{7.56}$$

where L is the patch length and x_{feed} the distance of the feed position to the patch edge.

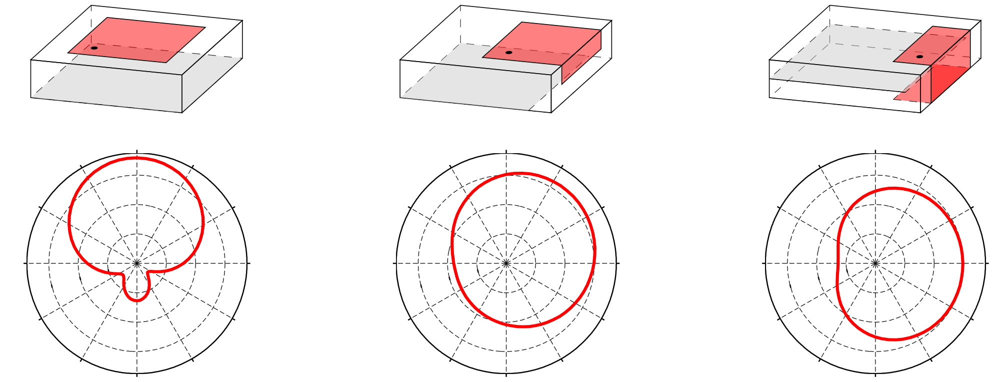
- Proximity feeding: For patch antennas, proximity feeding means excitation by coplanar or parallel (stacked) capacitive coupling. Naturally, the strength of the coupling is determined by the length and width of the gap between the feed line and the patch. Both, the impedance matching capabilities and the wider bandwidth are improved with proximity feeding.

- Aperture coupling: Still better performance in terms of matching and wide bandwidth can be obtained by so-called aperture coupling. Additionally, it shows very good cross-polarization isolation. However, this method takes up more space and is more expensive due to the additional layer needed in the substrate.

  In this case, the feeding is a two-step process: first, a primary aperture is excited, which is most commonly done by proximity effects from a line stub. Then, in the second step, this aperture excites the second element, in this case the patch antenna. The overall coupling is much looser than with direct or coaxial feeding. Thus, the impedance bandwidth of the aperture coupled patch antenna is known to be much broader (e.g.factor 2-3 [63] or 10-30% [6] are reported) than the one of direct-fed patch antennas. The primary aperture is most commonly a slot in the ground plane. It is normally not excited at resonance, since this would produce radiation towards the patch which then would just be reflected or diffracted. [63] Aperture coupled patch antennas are known to show very good cross polarization isolation. On the same time, with two slots (e.g.beside each other or in cross form) the patch can be excited independently in two directions or also be circularly polarized. [51]

The patch itself does not necessarily have to sit on one surface; it can be bended or wrapped around the edges. Clearly, this has some influence on the radiation pattern as can be seen in Fig. 7.22.

Other forms of microstrip antennas are also possible, e.g., printed dipoles etc. Similar in structure, but partly contrary in operation, are so-called slot antennas. Instead of a copper patch sitting above the ground plane, there is a hole, or rather a slot in the ground plane, which radiates.



(a) normal (centered) patch (b) offset patch (c) wrapped patch

Figure 7.22 Radiation pattern of normal, offset and wrapped patch antenna configurations.

7.5 Summary of Antennas

In Table 7.1 some of the more important antennas and their parameters are collected.

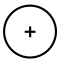
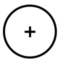
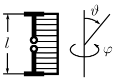
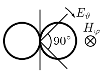
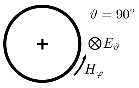
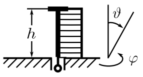
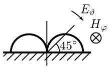
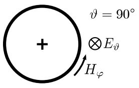
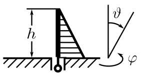
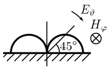
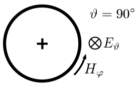
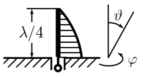
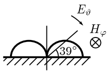
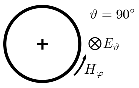
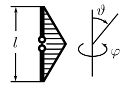
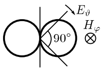
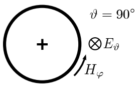
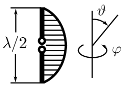
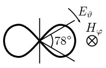
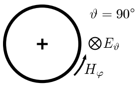
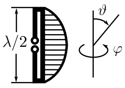
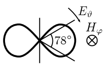
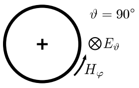
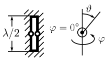
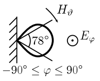
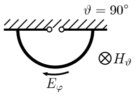
Antenna type	Diagram, assignment	Directivity, gain linear (dBi)	Effective antenna area	Effective height	Radiation resistance	Elevation cut (3 dB-range)	Azimuth cut
isotropic antenna	fictitious	1 (0)	$\frac{\lambda^2}{4\pi}$	–	–		
Hertzian dipole, dipole with end capacitance		1.5 (1.8)	$\frac{3\lambda^2}{8\pi}$	l	$80\,\Omega\left(\frac{\pi l}{\lambda}\right)^2$		
short antenna $h \ll \lambda$, with roof capacitance on a conductive plane		3 (4.7)	$\frac{3\lambda^2}{4\pi}$	$2h$	$160\,\Omega\left(\frac{\pi h}{\lambda}\right)^2$		
short antenna on a conductive plane $h \ll \lambda$		3 (4.7)	$\frac{3\lambda^2}{4\pi}$	h	$40\,\Omega\left(\frac{\pi h}{\lambda}\right)^2$		
$\lambda/4$ antenna on a conductive plane		3.28 (5.1)	$0.26\,\lambda^2$	$\frac{\lambda}{\pi}$	$40\,\Omega$		
short dipole $l \ll \lambda$		1.5 (1.8)	$\frac{3\lambda^2}{8\pi}$	$\frac{l}{2}$	$20\,\Omega\left(\frac{\pi l}{\lambda}\right)^2$		
$\lambda/2$ dipole		1.64 (2.1)	$0.13\,\lambda^2$	$\frac{\lambda}{\pi}$	$73\,\Omega$		
$\lambda/2$ folded dipole		1.64 (2.1)	$0.13\,\lambda^2$	$\frac{2\lambda}{\pi}$	$290\,\Omega$		
slot antenna, radiating into half space		3.28 (5.1)	$0.26\,\lambda^2$	$1.18\,\lambda$	$\approx 500\,\Omega$	 H_ϑ 78° E_φ $-90° \le \varphi \le 90°$	

Table 7.1 Antenna parameters for some widely used antennas. Translated version of [41] with omissions.

7.6 Antenna Measurement

In order to measure an antenna with respect to its directivity, an anechoic chamber can be of use. In such a room, special absorber pyramids avoid reflections and foster free-space propagation. Fig. 7.23 shows the setup of the measurement equipment inside such an anechoic chamber. An RF generator drives a transmitter antenna, usually a device with known gain. The device under test (DUT) is then used as the receiver antenna and turned slowly by a controlled turntable. For every angle the received power is registered and later drawn in a polar plot, e.g., Fig. 7.24.

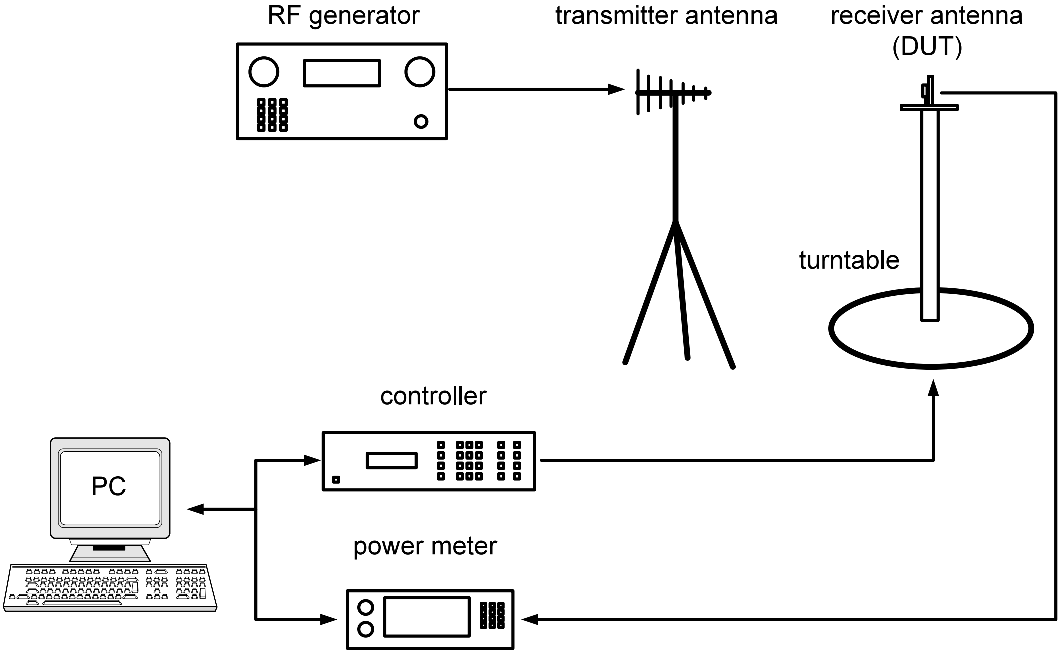


Figure 7.23 Measurement setup inside an anechoic chamber.

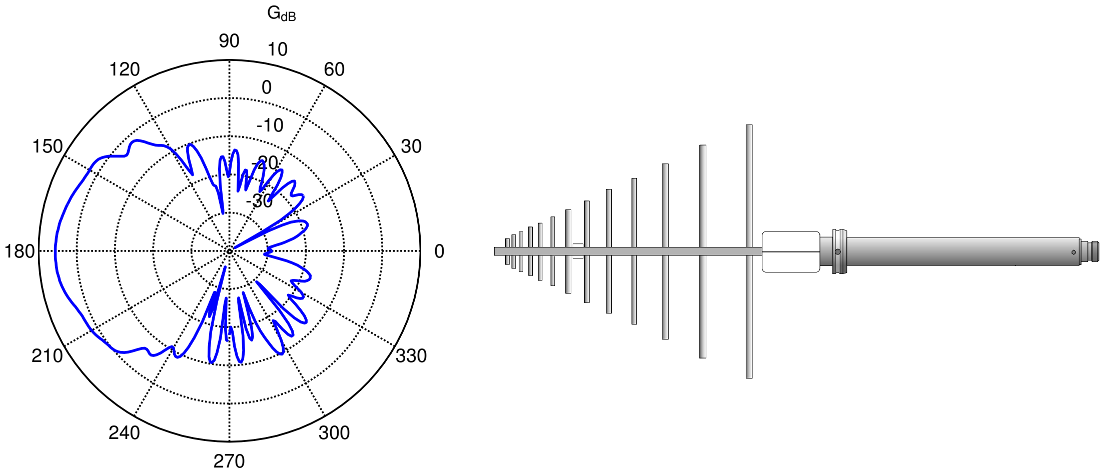


Figure 7.24 Example of a polar plot (antenna pattern) of a Schwarzbeck USLP9142 logper antenna at 5 GHz (horizontally polarized, azimuth cut), as measured in the anechoic chamber.

7.7 Mechanisms of Wireless Propagation

In the context of wireless communication, the term *propagation* describes the effects of the advancement of the electromagnetic wave bearing the information to transmit. We have four principle models for wireless propagation to distinguish: ground waves, sky waves, free-space waves, and open-field waves. In order to understand the propagation mechanism we need to have a short glimpse at the layers above the earth surface. They are illustrated in Fig. 7.25.

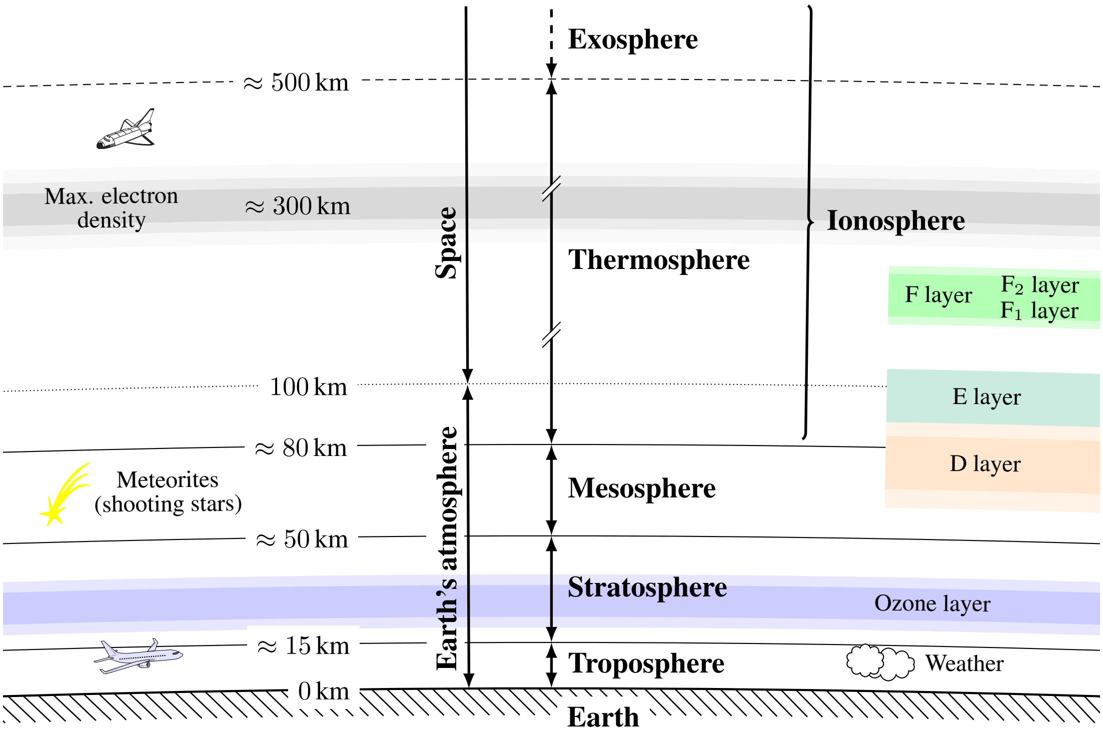
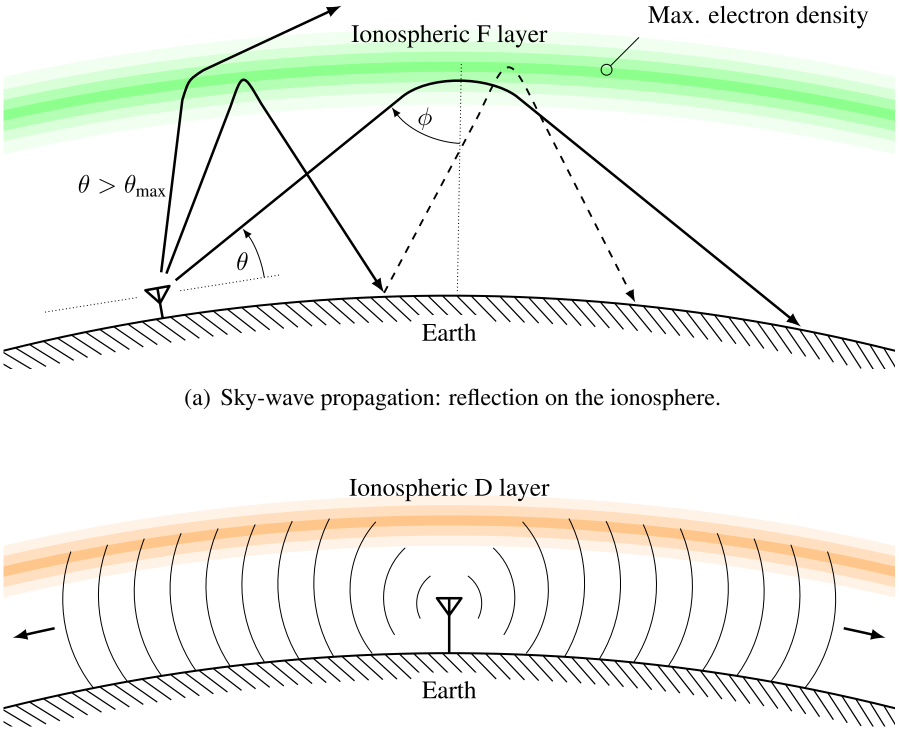


Figure 7.25 Zones around the earth.

Ground waves or *surface waves* occur at low frequencies, up to a few MHz, and for vertical polarization (horizontally polarized fields are short-circuited by the earth). Such waves travel over the earth's surface attenuated by the absorption of the conducting earth. Through the curvature of the earth, the waves can make it to points which have no direct line of sight. The higher the frequency, the more the wave gets absorbed by the ground.

Sky waves or *room waves* occur at frequencies between 3 and 30 MHz. They describe the propagation mode by which waves are reflected from the ionosphere, a thin-air layer high above the surface of the earth, a process illustrated in Fig. 7.26(a). The ionosphere is ionized by the sunlight at which point it builds a good reflector for the wavelength indicated above. The reflection quality is therefore heavily dependent on the time of day, season, longitude on earth, and the sunspot cycle. Using sky waves, transmission virtually around the earth has been reported (amateur radio). The ionosphere may build a waveguide for the signal to propagate according to Fig. 7.26(b).



(a) Sky-wave propagation: reflection on the ionosphere.

(b) Earth-ionosphere waveguide: the ionosphere and ground serve as waveguide walls.

Figure 7.26 Propagation modes making use of the ionosphere.

7.7.1 Free-Space Propagation

The *free-space propagation* is the easiest propagation mechanism to understand. It assumes that no reflections occur. Imagine a setup of two antennas in free space (no obstacles inbetween) as in Fig. 7.27. The transmitting antenna (marked by 'T') transmits at power P_T. If the directivity was 0 dB, the power would distribute equally over the sphere with radius r, or $S = P_T/(4\pi r^2)$. Due to a larger directivity, we get for the power density at a point r away in the direction of the main lobe

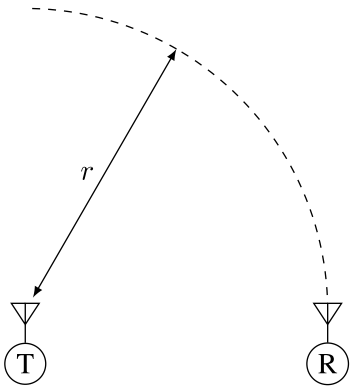


Figure 7.27 Free-space setup with two antennas.

$$S = \frac{P_T G_T}{4\pi r^2}. \tag{7.57}$$

The product $P_T G_T$ is usually called the *equivalent isotropic radiated power* (EIRP), which is essentially the radiated power relative to an isotropic radiator. Sometimes ERP is used, which refers to the power relative to a dipole. The receiving antenna (marked by 'R') now captures the power density over an area A_R, the effective area or aperture. The total received power is therefore

$$P_R = \frac{P_T G_T}{4\pi r^2} A_R. \tag{7.58}$$

With Eq. (7.30), we get for the power relation in the free-space case

$$\frac{P_R}{P_T} = G_T G_R \left(\frac{\lambda}{4\pi r}\right)^2 . \tag{7.59}$$

This is the *Friis transmission formula*. It is valid only if the main lobes of the antennas are adjusted exactly, if the polarizations match, and if both antennas are impedance matched. The path loss is usually given as the inverse of Eq. (7.59) (positive dB numbers), when the antenna gains can be isolated. Thus, the free-space path loss is

$$L = \frac{P_T G_T G_R}{P_R} = \left(\frac{4\pi r}{\lambda}\right)^2 = \left(\frac{4\pi r f}{c}\right)^2 . \tag{7.60}$$

In decibels, Eq. (7.60) can be expressed as

$$\boxed{L\,[\text{dB}] = -147.6 + 20\log_{10} r + 20\log_{10} f}\,. \tag{7.61}$$

The free-space path loss increases (gets worse) with increasing frequency because of the smaller capture area of the antenna.

In practice, the free-space model is usually optimistic, because other effects are neglected (e.g., absorption by raindrops etc.). Still, it can serve as a practical minimum of the path loss for a given distance before other, more involved methods are applied.

7.7.2 Open-Field Propagation

A more realistic scenario of propagation is suggested by the open-field model, also called plane-earth model. The model is given by Fig. 7.28. Both transmitting and receiving antennas are situated at a certain height (h_b for the base station and h_m for the mobile) above a flat reflecting ground. Propagation now takes place via two paths, a direct path of length

$$r_1 = \sqrt{r^2 + (h_b - h_m)^2}, \tag{7.62}$$

and a reflected path, whose length may be inferred by applying image theory, considering the reflected ray as coming from an image of the transmitter in the ground, as

$$r_2 = \sqrt{r^2 + (h_b + h_m)^2}. \tag{7.63}$$

The interference from the reflection may be constructive, destructive, or part of either. Which one, depends on the path length difference between r_1 and r_2 with respect to the wavelength, hence, we are interested in

$$\begin{aligned}\frac{r_2 - r_1}{\lambda} &= \frac{1}{\lambda}\left(\sqrt{r^2 + (h_b + h_m)^2} - \sqrt{r^2 + (h_b - h_m)^2}\right)\\ &= \frac{r}{\lambda}\left(\sqrt{1 + \left(\frac{h_b + h_m}{r}\right)^2} - \sqrt{1 + \left(\frac{h_b - h_m}{r}\right)^2}\right).\end{aligned} \tag{7.64}$$

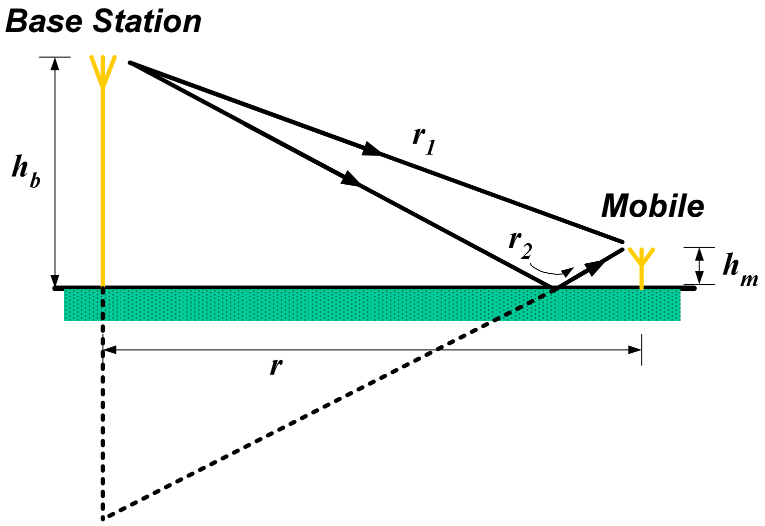


Figure 7.28 Open-field propagation model.

For small antenna heights compared to the length r, we can use the first-order approximation

$$(1+x)^n \approx 1+nx, \qquad x \ll 1. \tag{7.65}$$

Doing so we get

$$\frac{r_2-r_1}{\lambda} \approx \frac{2h_bh_m}{\lambda r}. \tag{7.66}$$

If we now assume that the direct and reflected waves have approximately the same amplitude (the path lengths are almost the same and the reflection coefficient R is close to minus one, which is strictly true only for low grazing angles $\gamma = \tan^{-1}(\frac{h_b+h_m}{d})$, we get for the combined loss the multiplicative factor (in addition to the free-space loss)

$$\left|1+R\exp\left(j2\pi\frac{r_2-r_1}{\lambda}\right)\right| \approx \left|1+R\exp\left(j\frac{4\pi h_bh_m}{\lambda r}\right)\right| \approx \left|1-\exp\left(j\frac{4\pi h_bh_m}{\lambda r}\right)\right|. \tag{7.67}$$

The minus sign for the reflection coefficient R in the above equations stems from the fact that polarization flips at the reflection (at least for incident angles close to grazing). We get, together with the free-space path loss, the open-field path loss as

$$\boxed{L = \left(\frac{4\pi r}{\lambda}\right)^2 \left|1-\exp\left(j\frac{4\pi h_bh_m}{\lambda r}\right)\right|^{-2} = 4\left(\frac{\pi r}{\lambda}\right)^2 \sin^{-2}\left(\frac{2\pi h_bh_m}{\lambda r}\right)}. \tag{7.68}$$

By observing the open-field path in Fig. 7.29, we see that the loss function for the open field has many nulls up to a certain point. After that, it decreases much faster than the $20\log_{10}(.)$ we know from the free-space case. In fact, since exponential functions can be approximated using

$$\exp(x) \approx 1+x, \qquad x \ll 1, \tag{7.69}$$

we get for large distances an open-field path loss of

$$L = \left(\frac{4\pi r}{\lambda}\right)^2 \left(\frac{\lambda r}{4\pi h_bh_m}\right)^2 = \frac{r^4}{h_b^2h_m^2}. \tag{7.70}$$

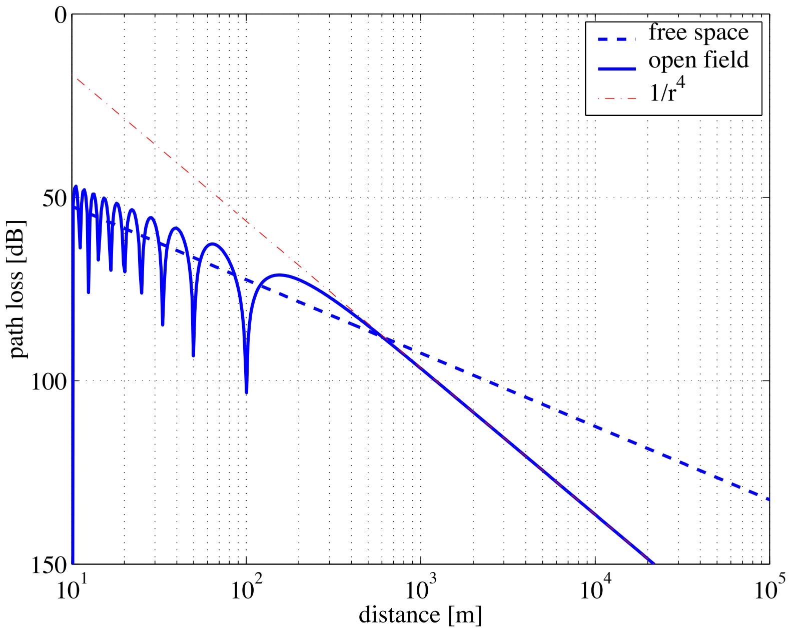


Figure 7.29 Open-field path loss (solid line). In this case $h_b = 20\,\text{m}$, $h_m = 1.5\,\text{m}$, and $f = 1\,\text{GHz}$.

As opposed to the free-space loss, which grows with r^2, the open-field loss grows with r^4. Furthermore, the loss is now independent of the operating frequency. The point where the curve starts going down with r^4 is roughly at the distance of the last local maximum,

$$r = \frac{4h_b h_m}{\lambda}. \tag{7.71}$$

We have seen that the path-loss exponent in the free-space model was 2 (in logarithmic form $20\log_{10}(r)$) and in the open-field model 4 (or $40\log_{10}(r)$). Other, empirically found path-loss exponents are in use, see for example Table 7.2. If we now want to estimate the power P_2 received at a certain point in a certain distance d_2 from the knowledge of the power P_1 at a different distance d_1, we can write down the relation

Environment	**Path-loss exponent** n
free space	2
open field (long distance)	4
cellular radio, urban area	$2.7 - 4$
shadowed urban cellular radio	$5 - 6$
in building, line-of-sight	$1.6 - 1.8$
in building, obstructed	$4 - 6$

Table 7.2 Path-loss exponents for different environments. Source: [24, p. 362].

$$\frac{P_2}{P_1} = \left(\frac{d_1}{d_2}\right)^n. \tag{7.72}$$

7.7.3 Diffraction

Besides reflexion, which we have met in Section 7.7.2, and absorption, which does not contribute to the models used here, there are other mechanisms of propagation, which we shall shortly describe. Diffraction (*dt. Beugung*) is the explanation to the question of how objects that are not in the direct line of sight can be reached by electromagnetic waves. Shadows are never completely sharp, some energy does always propagate into the shadow region.

We can understand the reason by looking at Huygen's principle: First, at a given instant in time, each element of a wavefront can be regarded as the center of a secondary disturbance, giving rise to spherical wavelets. Second, the position of the wavefront at any later time is the envelope of all such wavelets. We can apply this principle to our problem of a plane wavefront hitting a wall-top (so called *knife-edge diffraction*), which is illustrated in Fig. 7.30. Wavefronts impinging on the top of the absorbing screen become curved by the edge in such a way, that waves seem to emerge from points close to the edge, filling the shadow region with diffracted wavelets.

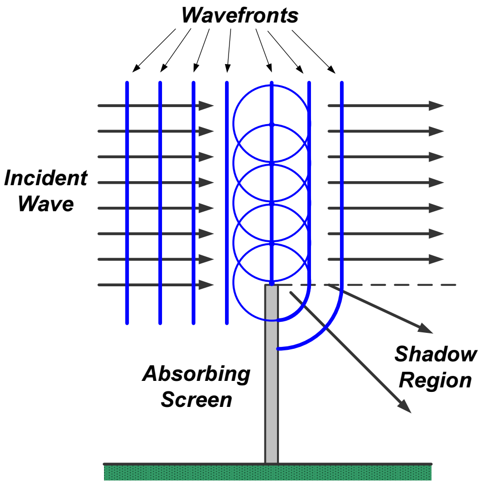


Figure 7.30 Huygen's principle for knife-edge diffraction.

If Huygen's principle is applied in a quantitative way, taking into account the respective amplitudes and phases of an infinite number of secondary sources above the edge, we get the propagation loss inside the shadow region

$$L = 20\log_{10}\left|\frac{E_i}{E_d}\right| = -10\log_{10}|F(v)|, \tag{7.73}$$

where E_i and E_d are the impinging and diffracted fields, respectively, and

$$F(v) = \frac{1+j}{2}\int_v^\infty \exp\left(-j\frac{\pi t^2}{2}\right)dt \tag{7.74}$$

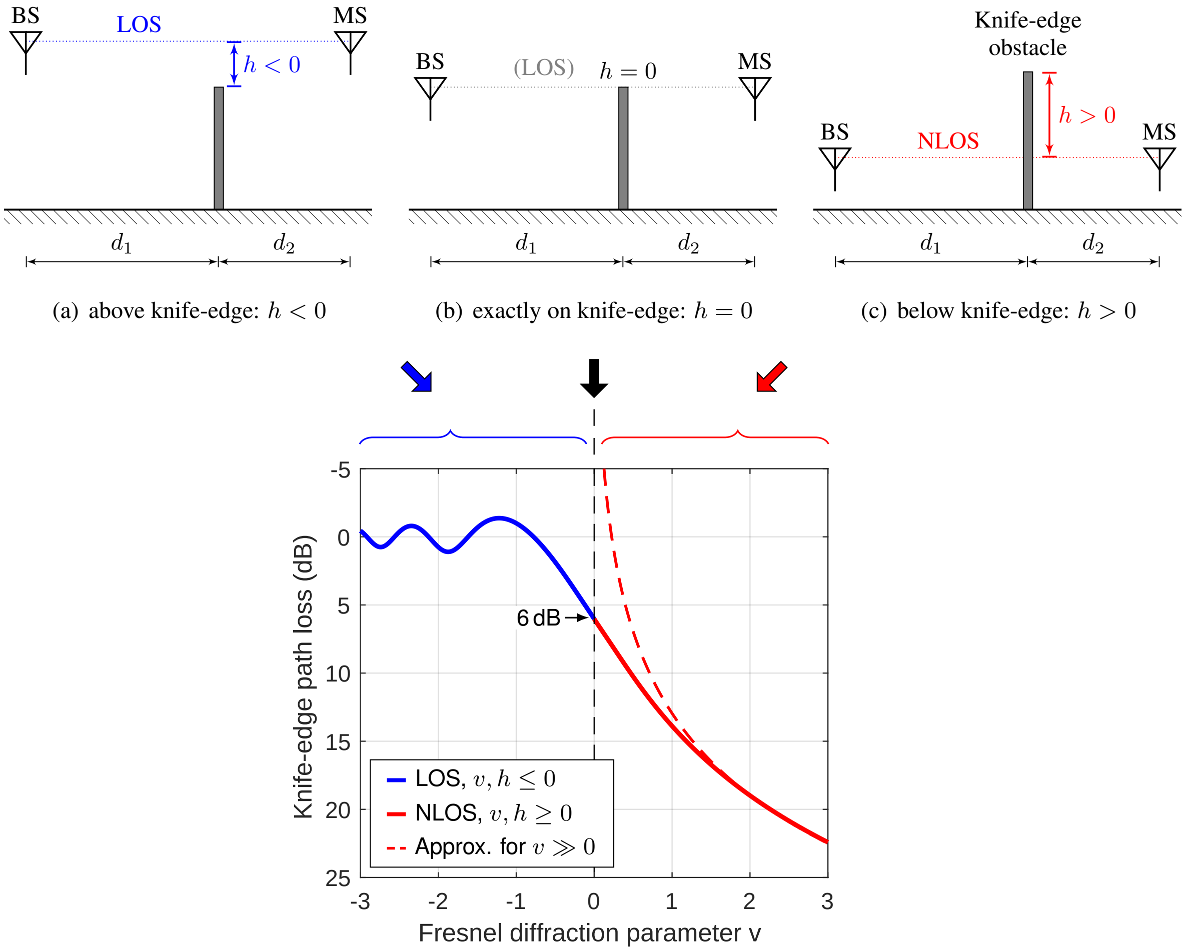
is a function of the Fresnel integrals, see for example [66], and the diffraction parameter given by

$$\boxed{v = h\sqrt{\frac{2(d_1+d_2)}{\lambda d_1 d_2}}} \tag{7.75}$$

with d_1, d_2 being the respective distances of the transmitter and receiver to the knife edge and h the excess height of the knife edge. In fact, the distances d_1 and d_2 and the height h refer to a coordinate system in which both transmitter and receiver sit on a baseline. Any situation can be transformed into such a setup using a simple coordinate transformation, provided the distances d_1 and d_2 are large compared to h, which is usually the case.

Let us consider three practical situations: when there is a line-of-sight (LOS), meaning there is clearance between line connecting the two antennas and the knife edge case, Fig. 7.31(a), h is negative and diffraction liss is usually negligible. At the point where the obstacle covers the space up to exactly the LOS line (i.e. the excess height is zero), illustrated in Fig. 7.31(b), a diffraction loss of 6 dB occurs. Lastly, when the obstacle obstructs the LOS line (leading to a non-line-of-sight condition, NLOS), h is positive and increased diffraction loss occurs, as shown in Fig. 7.31(c). Note that this loss occurs in addition to the free-space loss[3] for a given distance. For values of $v > 1$, which refers to points well inside the shadow region, Eq. (7.73) may be approximated with accuracy better than 1 dB by

$$L \approx 20\log_{10}(\sqrt{2}\pi v) \approx 20\log_{10}\frac{v}{0.225} \qquad v > 1. \tag{7.76}$$



(a) above knife-edge: $h < 0$

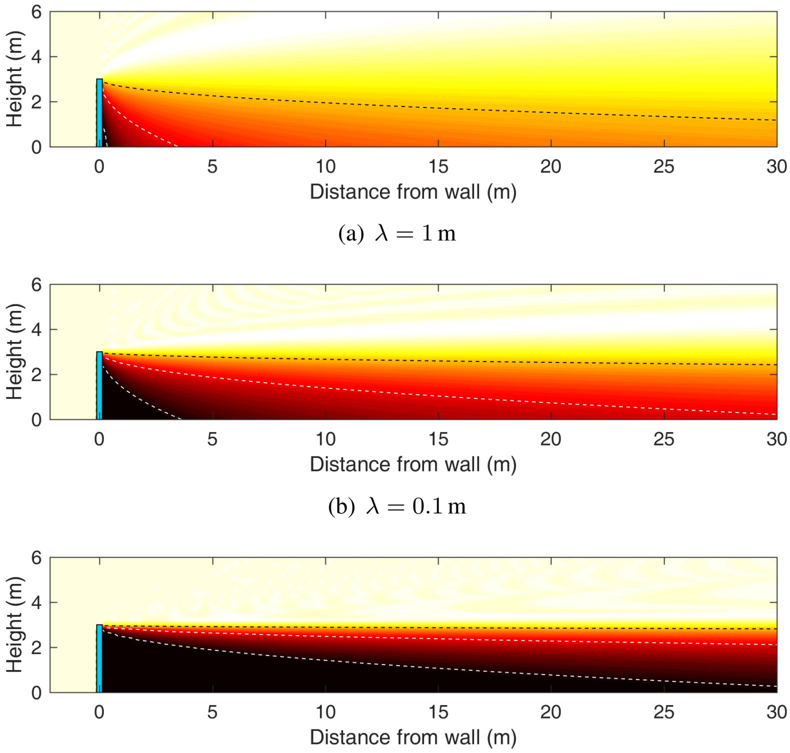
(b) exactly on knife-edge: $h = 0$

(c) below knife-edge: $h > 0$

Figure 7.31 Examples of knife-edge propagation due to partial obstruction by a wall and the involved diffraction loss, calculated using exact Fresnel integrals, according to Eq. (7.74), and the approximation for large v Eq. (7.76).

[3] The diffraction edge prevents any reflections in this case, so that the free-space model is appropriate.

It has to be emphasized that for all other cases no closed formula can be given and it is best to infer the corresponding loss from Fig. 7.31. For negative values v (negative excess height as seen in Fig. 7.31(a)), which means there is a direct line of sight, there can still be some loss. For the limit case of $v = 0$, i.e., transmitter and receiver are at the same height as the diffraction edge (the excess height is zero), the loss is 6 dB.



(a) $\lambda = 1\,\mathrm{m}$

(b) $\lambda = 0.1\,\mathrm{m}$

(c) $\lambda = 0.01\,\mathrm{m}$

Figure 7.32 Diffraction loss due to a wall (3 m tall) for different wavelengths.

To illustrate the dependence of the diffraction loss by the wavelength, Fig. 7.32 shows the result with λ as a parameter. Clearly, when going towards optics, the 'illumination' of shaded areas gets more difficult as would please our intuition.

7.7.4 Fresnel Zones

Sometimes, the expression *Fresnel zone* is used to indicate the proximity of obstacles to the direct line of sight. To visualize Fresnel zones, we need to derive ellipses that have the transmitter and the receiver in their focus points. The first Fresnel zone is of special interest and has the following derivation: The ellipse curvature consists of the points through which the path is longer by $\lambda/2$ compared to the direct line. Hence, we can construct the ellipse by choosing the longer and shorter axes as a and b and the distance between transmitter and receiver as $2c$, as seen in Fig. 7.33. Now the problem of computing the ellipse axes of the first Fresnel zone is working out a and b, given the distance $d = 2c$ and the wavelength λ. For an ellipse, every path via a point on the curve is of the same length, hence also the one via the end points of the ellipse

$$l = c + a + (a - c) = 2a. \tag{7.77}$$

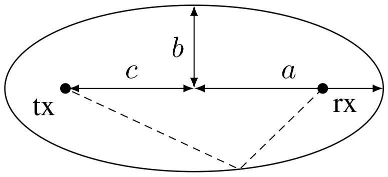


Figure 7.33 Construction of the first Fresnel zone.

The direct path is $2c$. Thus,

$$2a = 2c + \frac{\lambda}{2} \tag{7.78}$$

or

$$a = c + \frac{\lambda}{4} = \frac{d}{2} + \frac{\lambda}{4}. \tag{7.79}$$

The ellipse equation (linear excentricity) $c^2 = a^2 - b^2$ delivers

$$b = \sqrt{\frac{c\lambda}{2} + \frac{\lambda^2}{16}}. \tag{7.80}$$

For small wavelengths $\lambda \ll c$, we have

$$b \approx \sqrt{\frac{c\lambda}{2}}. \tag{7.81}$$

We have now derived the major and minor axes of the first Fresnel ellipse. In a wireless transmission system, one usually tries to keep the first Fresnel zone free of obstacles. More generally, the major and minor axes of the n-th order ellipse can be derived as

$$a = \frac{d}{2} + \frac{n\lambda}{4}, \tag{7.82}$$

$$b = \sqrt{\frac{cn\lambda}{2}}. \tag{7.83}$$

Since usually $\lambda \ll d$, we can write Eq. (7.82) as

$$a \approx c = \frac{d}{2}. \tag{7.84}$$

The minor axis b is the clearance height in the middle of the transmitter and receiver. In the following, we want to derive the clearance height at every position between transmitter and receiver, such that the distance to a potential obstacle is d_1 and d_2 and not necessarily $\frac{d}{2}$, see Fig. 7.34(a). We start by stating that for every pair (x, y) the ellipse equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{7.85}$$

must be satisfied. Solving Eq. (7.85) for y yields

$$y = b\sqrt{\frac{a^2 - x^2}{a^2}}\ . \tag{7.86}$$

With $d_1 = c + x$ and $d_2 = c - x$ we get

$$d_1 d_2 = c^2 - x^2 = a^2 - b^2 - x^2\ . \tag{7.87}$$

Using the fact that $b \ll d$ and Eqs. (7.83) to (7.84) in Eq. (7.86), we get

$$y = b\sqrt{\frac{d_1 d_2}{c^2}} = \sqrt{\frac{cn\lambda}{2}}\sqrt{\frac{d_1 d_2}{c^2}} = \sqrt{\frac{n\lambda d_1 d_2}{2c}}\ . \tag{7.88}$$

Now, c is of course the mean of the two distances, $c = \frac{d_1+d_2}{2}$, hence

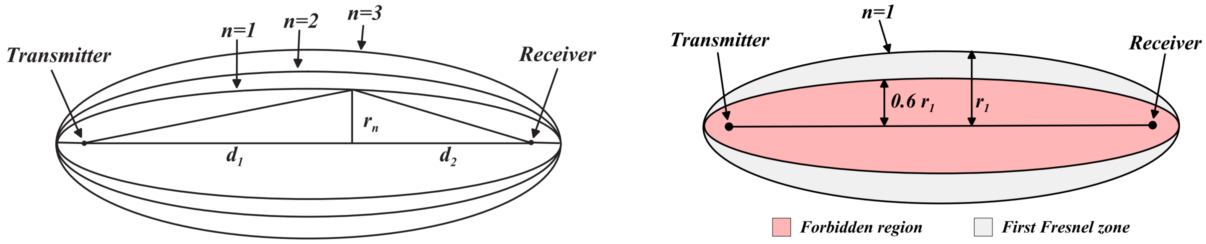
$$y = \sqrt{\frac{n\lambda d_1 d_2}{d_1 + d_2}}. \tag{7.89}$$

Due to the symmetry around the line tx–rx we can speak of the Fresnel radius rather than y. Thus, the nth Fresnel zone is given in approximation by the radius

$$\boxed{r_n \approx \sqrt{\frac{n\lambda d_1 d_2}{d_1 + d_2}}}, \tag{7.90}$$

for which the ellipsoids as shown in Fig. 7.34(a) result. Usually, one tries to avoid the placement of obstacles in the first Fresnel zone. By expressing the diffraction parameter as a function of the first Fresnel zone, we realize that

$$v = \frac{h}{r_1}\sqrt{2}. \tag{7.91}$$



(a) The first three fresnel zones.

(b) 60 % of the first Fresnel zone clearance defines significant obstructions.

Figure 7.34 Fresnel zones.

Now, as long as an obstruction is outside the circle whose radius is given by 60 % of the first Fresnel zone, the diffraction parameter is around $v = -0.85$. Referring to Fig. 7.31, the diffraction loss is then 0 dB. This clearance is often used as the criterion to decide whether an object is a significant obstruction (see also Fig. 7.34(b)). With the exception of the first Fresnel zone, all zones cancel each other in their influence (pairwise). The reception thus stays approximately constant if all but the first zone are shaded.

7.7.5 Scattering

In Section 7.7.2, we have assumed total reflection, so-called *specular reflection*. Frequently, however, when surfaces are not smooth, an impinging wave does not get reflected in one exact direction, but gets scattered, see Fig. 7.35. One possible technique to model scattering (*dt. Streuung*) is due to Beckmann and Spizzichino [7], who multiply the reflection coefficient R in Eq. (7.67) by a roughness factor f, which depends on the angle of incidence θ and on the standard deviation of the surface height σ_s. This roughness factor is given as

$$f = \exp\left(-\frac{1}{2}\left(\frac{4\pi\sigma_s\cos\theta}{\lambda}\right)^2\right), \tag{7.92}$$

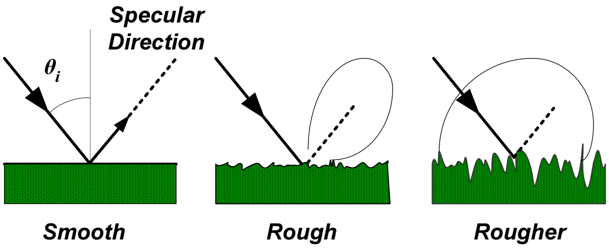


Figure 7.35 Scattering due to surface roughness.

which results in the plot as provided in Fig. 7.36.

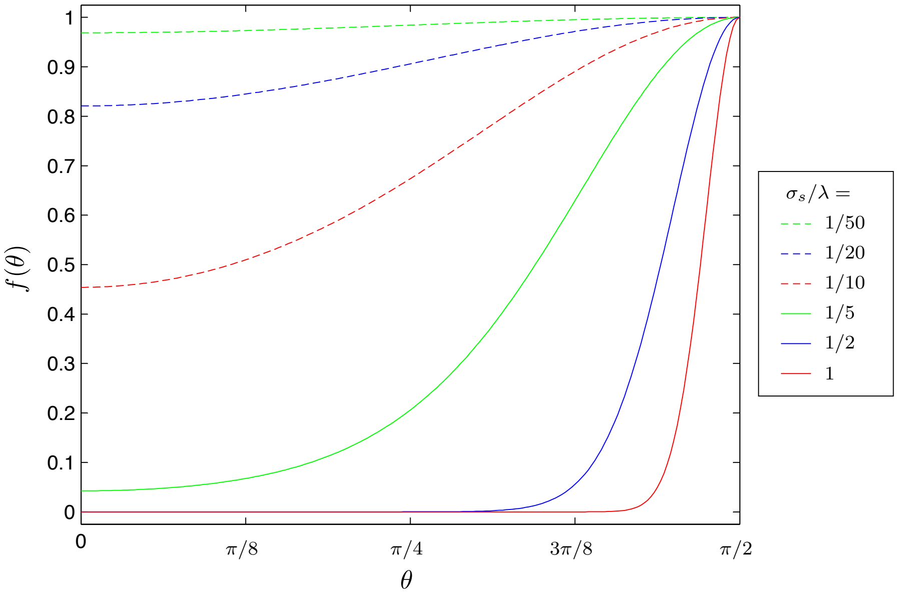


Figure 7.36 Roughness factor.

7.7.6 Further Losses

In addition to the loss factor treated so far, we find further attenuation in the wave propagation due to absorption caused by interaction with gas molecules, most importantly oxygen (O_2) and water vapor (H_2O), as shown in Fig. 7.37. It is well-known that the water molecule possesses an electric dipole moment, which interacts with the electric fields. Lesser known, but equally important, is the fact that the oxygen molecule possesses a permanent magnetic dipole moment associated to aligned electron spins of its constituent atoms. Hence, oxygen molecules interact with magnetic fields. Interestingly, the molecule of nitrogen (N_2), the most abundant gas in earth's atmosphere, neither has a significant electric nor magnetic dipole moment, and, thus, does not interact with electromagnetic waves in the microwave spectrum at all.

As can be seen by the graphs in Fig. 7.37, the first local maximum of the absorption at microwave frequencies is at 22.2 GHz, due to water. The second local maximum of the absorption is due to a family of 26 resonance lines of the oxygen molecule around 60 GHz; the separate lines only become visible at very high altitudes (above sea level, a.s.l.). At such low gas pressure levels, the interaction between the molecules is reduced

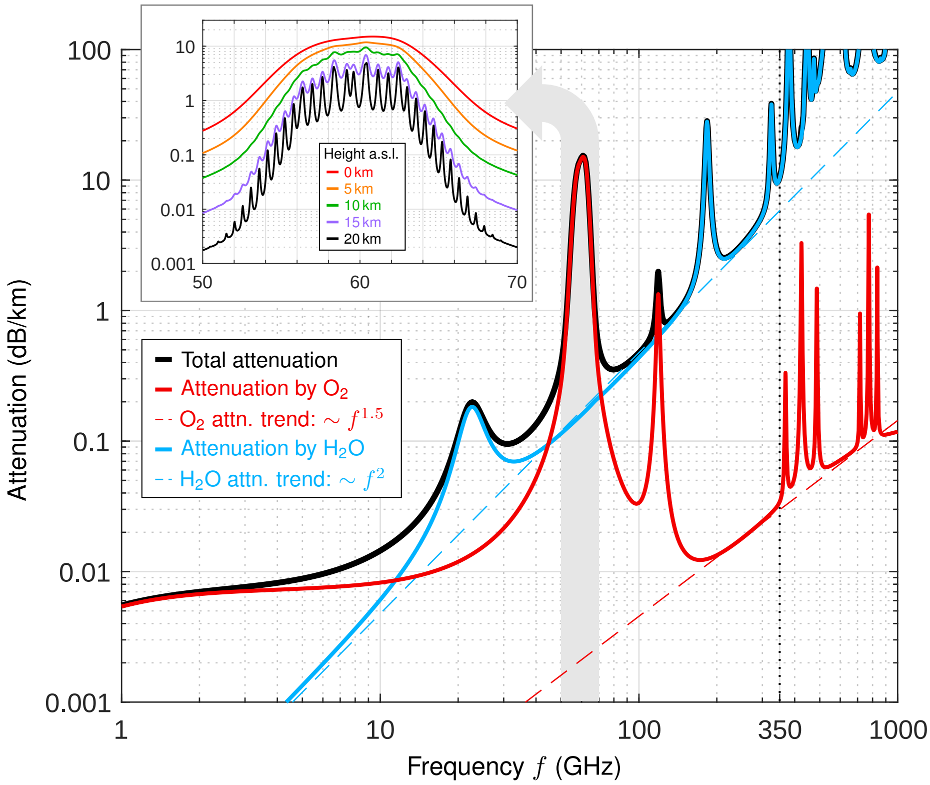


Figure 7.37 Attenuation due to molecular absorption of oxygen and water vapor (at sea level) in the microwave spectrum.

considerably, such that the contributions of all energy transitions become separable. The third resonance is due to oxygen again, this time a single sharp resonance line around 119 GHz. After this frequency, water has two more resonances around 183 and 321 GHz, respectively, before both oxygen and water absorption becomes difficult to predict above 350 GHz.

Note that the actual attenuation depends on the amount of water in the atmosphere. Moreover, if undissolved water is in the air (rain), this attenuation has to be added to the free space loss and the two attenuation values due to water vapor and oxygen. Table 7.3 lists the attenuation due to rain in more detail for some practical frequencies.

Rain type	Intensity	450 MHz	1 GHz	3 GHz	10 GHz
Mizzle (drizzle)	0.25 mm/h	$2.2 \cdot 10^{-8}$	$1.5 \cdot 10^{-6}$	$1.5 \cdot 10^{-4}$	0.02
Light rain	5 mm/h	$1.0 \cdot 10^{-6}$	$2.0 \cdot 10^{-5}$	$1.0 \cdot 10^{-3}$	0.08
Medium rain	12.5 mm/h	$3.0 \cdot 10^{-6}$	$7.0 \cdot 10^{-5}$	$3.0 \cdot 10^{-3}$	0.28
Heavy rain	25 mm/h	$7.5 \cdot 10^{-6}$	$1.5 \cdot 10^{-4}$	$1.0 \cdot 10^{-2}$	0.6
Shower	50 mm/h	$1.0 \cdot 10^{-5}$	$3.0 \cdot 10^{-4}$	$2.0 \cdot 10^{-2}$	1.5

Table 7.3 Attenuation (in dB/km) due to rain. Source: [39].

Whereas the loss factor (due to power-density thinning for larger distances) in dB grows logarithmically with distance, atmospheric absorption in dB grows linearly (it is the same for every meter of atmosphere no matter how far away), i.e., the latter effect dominates for some specific frequencies and larger distances, as can be seen in Fig. 7.38.

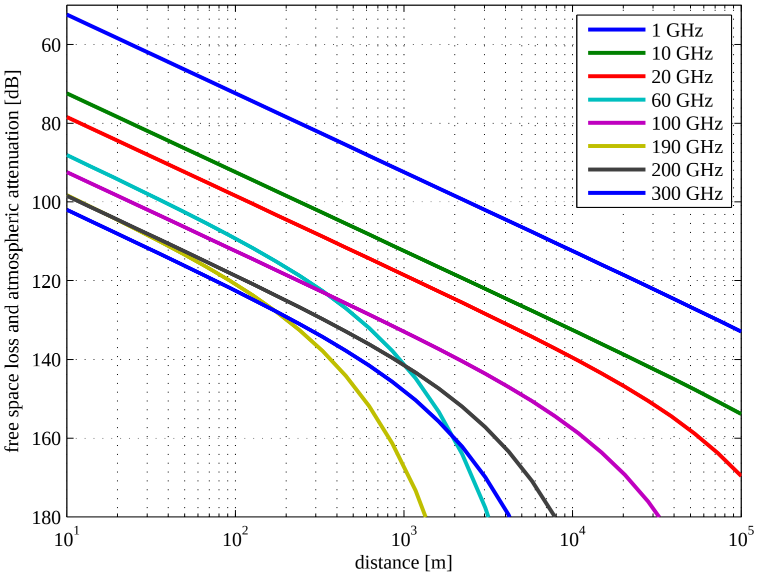


Figure 7.38 Combined attenuation due to distance loss and atmospheric absorption.

7.8 Empirical Path-Loss Models

Very often it might not be practical to determine the loss introduced by the wireless channel by the mechanisms covered so far, partly because the environment is not known exactly or because the mathematics get untractable. In these cases it is useful to have empirical path-loss models. The word *empirical* relates to the fact that such models have been extracted using actual measurements without considering the physical justification.

One of the first introduced such model was the one by Egli [17] in his publication 'Radio propagation above 40 MC over irregular terrain'[4] published in the Proceedings of the IRE[5] Egli realised that the open-field model shows a path-loss exponent of four and a dependence on the antenna heights. Additionally, he argued that a dependence on the frequency of f^2 should be reintroduced like in the free-space case. Because in his reasoning there was partly physical justification involved, this approach would also become known as a semi-empirical method. Egli's results of fitting a model with measurements taken around American cities were at the time provided in nomograms, which Delisle *et al.* [14] later translated into an equation (loss in dB) of the form

$$L = 40\log_{10} R + 20\log_{10} f_{\text{MHz}} - 20\log_{10} h_b + \begin{cases} 76.3 - 10\log_{10} h_m & \text{for } h_m < 10\,\text{m} \\ 85.9 - 20\log_{10} h_m & \text{for } h_m > 10\,\text{m} \end{cases}, \tag{7.93}$$

with a small discontinuity at $h_m = 10\,\text{m}$. Note that in Eq. (7.93) the unit of R is km and the unit of f_{MHz} is MHz.

[4] Note that MC was once used as the abbreviation for Megacycles per second which today is MHz of course.

[5] The IRE (Institute of Radio Engineers) is today known as the IEEE.

The most widely cited model is the Okumura-Hata model, which is a fully empirical model. It is based on measurements taken by Okumura in Tokyo [47] and equations fitted by Hata some years later [31]. By restricting ourselves to frequencies above 400 MHz, a base-station antenna height of 30 to 200 m, a vehicular antenna height below 10 m and a distance R_{km} between 1 and 20 km, which is no real restriction for most modern wireless systems, we can write these equations (loss in dB) as

$$L_p = 69.55 + 26.16 \log_{10} f_{\text{MHz}} - 13.82 \log_{10} h_b + (44.9 - 6.55 \log_{10} h_b) \log_{10} R_{\text{km}} - a(h_m), \tag{7.94}$$

where the coefficient a depends on the surrounded built-up area and is given as

$$a(h_m) = \begin{cases} 8.29(\log_{10}(1.54h_m))^2 - 1.1 & \text{for large cities and } f_{\text{MHz}} < 300\,\text{MHz} \\ 3.2(\log_{10}(11.75h_m))^2 - 4.97 & \text{for large cities and } f_{\text{MHz}} \geq 300\,\text{MHz} \\ (1.1 \log_{10} f_{\text{MHz}} - 0.7)h_m - (1.56 \log_{10} f_{\text{MHz}} - 0.8) & \text{for medium-small cities.} \end{cases} \tag{7.95}$$

For areas outside cities, the mobile antenna height h_m has no effect. Therefore Hata gives the loss in dB as

$$a(h_m) = \begin{cases} 2(\log_{10}(f_{\text{MHz}}/28))^2 + 5.4 & \text{for suburban areas} \\ 4.78(\log_{10} f_{\text{MHz}})^2 - 18.33 \log_{10} f_{\text{MHz}} + 40.94 & \text{for open areas.} \end{cases} \tag{7.96}$$

Note that in Eqs. (7.94) to (7.96) the units of R_{km} is km and the unit of f_{MHz} is MHz. This equation is the foundation of many commercial network-planning software. Other empiric models are known such as the Lee model, the Ibrahim and Parsons model, and still others. A comparison of different path-loss models is shown in Fig. 7.39, where it can be seen that the deterministic models usually result in too optimistic values as compared to results based on real measurements (empiric models).

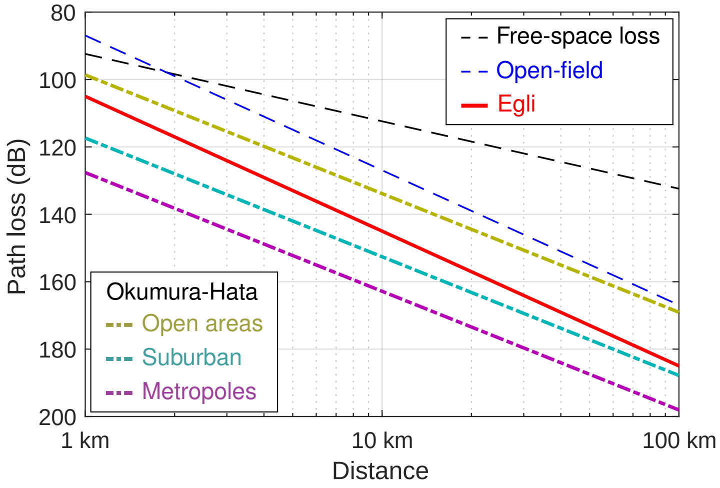


Figure 7.39 Comparison of different deterministic and empiric path-loss models (for $f = 1\,\text{GHz}$, $h_b = 30\,\text{m}$, $h_m = 1.5\,\text{m}$).

For indoor environments, there are empirical models as well. One such model is the COST231 Multi-Wall. It consists of an approximation of the loss in dB by

$$L = L_{\text{free space}} + A + B \cdot n_{w_1} + C \cdot n_{w_2} + D \cdot n^{\left(\frac{n+2}{n+1} - E\right)}, \tag{7.97}$$

where n is the number of traversed floors not thicker than 30 cm, n_{w_1} the number of light internal walls, and n_{w_2} the number of concrete/brick internal walls, respectively. The terms for A to E are frequency dependent.

7.9 Link Budget

A task frequently performed by a wireless engineer is the calculation of the link budget of a system. The question is usually one of the following:

- Given all parameters of the transmitter and the operating distance, what sensitivity of the receiver is needed (noise figure)?
- Given the operating distance and the receiver specification what is the minimum transmission power required?
- Given all parameters of the transmitter and receiver, what is the maximum operating range of the system?

The link budget is simply the summation of the dB values of all losses and gains acquired on the way from the transmitter to the receiver.

$$\boxed{P = N + B + \mathrm{NF} + \mathrm{SNR} + L - G_T - G_R} \tag{7.98}$$

with

N_0	noise power density (i.e. normally $N_0 = -174$dBm/Hz),
B	bandwidth used,
NF	noise figure of the receiver,
SNR	required SNR at the receiver end,
L	total loss in the transmission (e.g. path loss, cable loss, etc.),
G_T, G_R	gains of the transmitter and receiver antenna.

Example: Link budget

For a system operating at 1 GHz, what is the required transmission power to cover 10 km of range (free-space model assumed) if the receiver needs at least an SNR of 20 dB in a 25 kHz bandwidth with a total receiver chain noise figure of 5 dB, and both transmitter and receiver use half-wave dipoles?

Let us first list the gain/loss components in [dB]:

noise power density	$N_0 = -174\,\mathrm{dBm}$ (per Hz)
bandwidth	$B = 10\log_{10} 25000 = 44\,\mathrm{dB}$
noise figure	$\mathrm{NF} = 5\,\mathrm{dB}$
signal-to-noise ratio	$\mathrm{SNR} = 20\,\mathrm{dB}$
path loss	$L = -10\log_{10}\left(\frac{\lambda}{4\pi r}\right)^2 = -20\log_{10}\frac{c}{4\pi r f} = 112.4\,\mathrm{dB}$
receiver antenna gain	$G_R = 2.1\,\mathrm{dBi}$
transmitter antenna gain	$G_T = 2.1\,\mathrm{dBi}$

All we have to do now is add up the figures (with the right signs):

$$P = N_0 + B + \mathrm{NF} + \mathrm{SNR} + L - G_T - G_R = 3.2\,\mathrm{dBm} = \underline{\underline{2.1\,\mathrm{mW}}}\,.$$

7.10 Fading Channels

7.10.1 Shadowing

Until now, we have covered the first of the multiplicative channel effects only, namely path loss. In our analysis so far we have assumed at most two paths, a direct one and a reflected one. In reality, however, there is usually a multitude of paths, making the problem hard to determine. Also, very often the exact surroundings may not be known. The approach sought is often statistical modeling. A complicated environment with several objects between transmitter and receiver make the outcome of a path-loss analysis highly dependent on the exact location of transmitter and receiver. Placing the receiver slightly to one side might produce a high difference in path loss, not predicted by our former models. This phenomenon is called *slow fading* or *shadowing*. A typical variation of path loss can be seen in Fig. 7.40. The probability distribution of the underlying signal powers is log-normal, i.e., the loss expressed in dB has a normal distribution. The variation of the path loss occurs over distances comparable to object sizes, i.e., widths of buildings and hills and is called *location variability* σ_L. The reason for the log-normal distribution can easily be seen by considering the composition of the path loss by several independent effects on the way. Thus, the loss can be written as

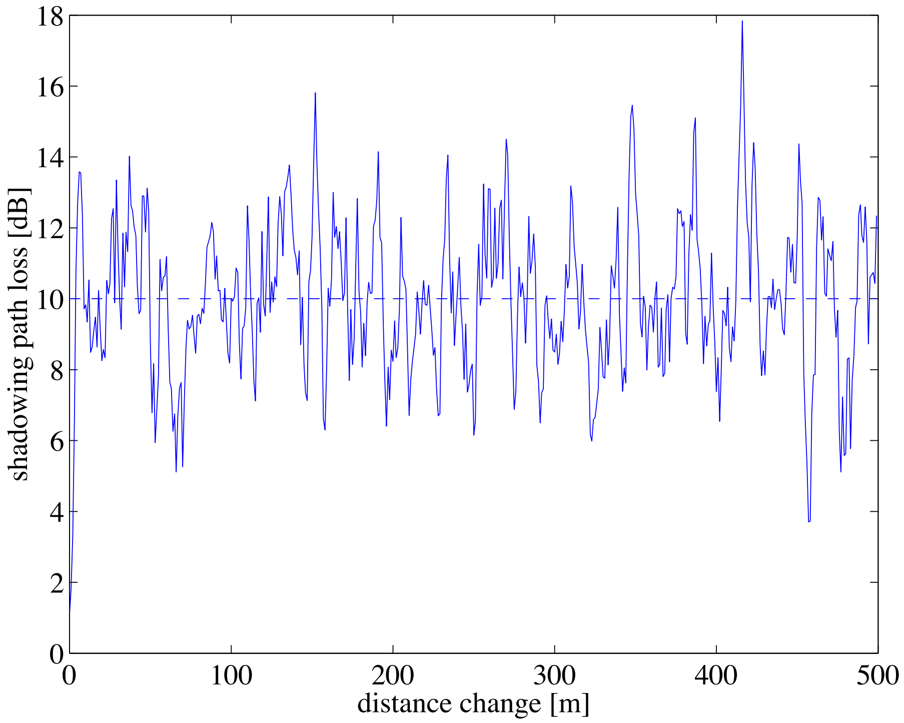


Figure 7.40 Typical shadowing path loss due to location variation of mobile receiver. Current example: median path loss = 10 dB (dashed line); path loss standard deviation (location variability) $\sigma_L = 5\,\text{dB}$.

$$L = L_1 \cdot L_2 \cdot L_3 \cdots L_N \tag{7.99}$$

in linear form, or in dB form as

$$L\,[\text{dB}] = L_1\,[\text{dB}] + L_2\,[\text{dB}] + L_3\,[\text{dB}] + \ldots + L_N\,[\text{dB}]. \tag{7.100}$$

If all of the N contributions are independent random variables, the central limit theorem dictates that $L\,[\mathrm{dB}]$ is a Gaussian random variable.

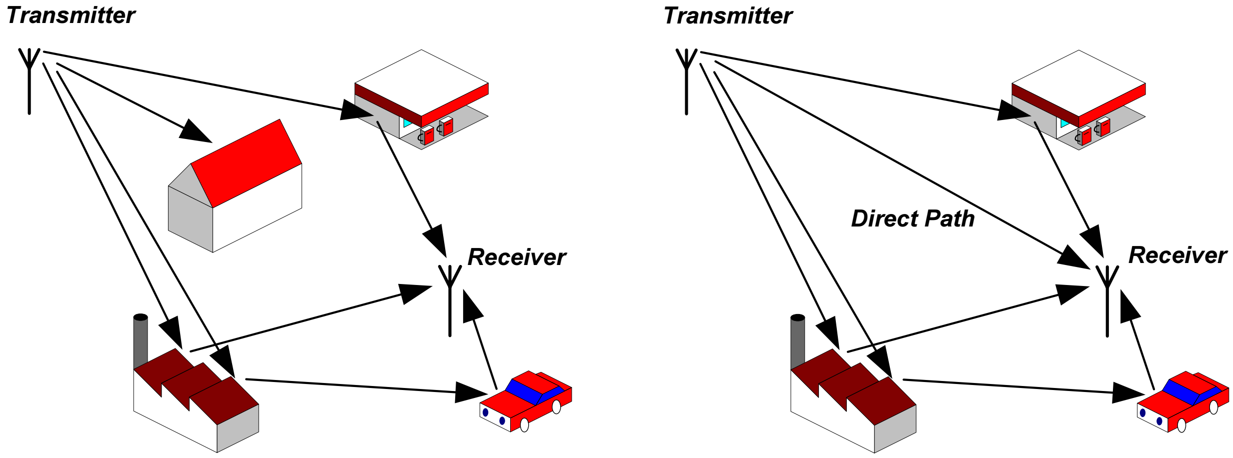
Shadowing has some impact on the link budget. If only the median path loss is used to evaluate a certain range covered, the real path loss exceeds the maximal value for reliable communications for 50 % of the cases. Hence, a certain *fade margin* has to be included in the link budget to make sure that most (you can never be 100 % sure but you can get very close to it) path losses incurred lead to successful communications. Thus, the fade margin to be included depends on the location variability and the percentage in desired successful communications.

7.10.2 Fast Fading

An even more dramatic effect than shadowing is *fast fading*, since it occurs on a much smaller time scale or with a much shorter correlation over location change. Since the relative bandwidth of a signal is usually very small, we analyze all fading effects in their complex baseband representation. The fading effects can then be conveniently modeled by complex phasors, indicating that fading changes amplitude and phase rapidly over time.

Flat Fading

With *flat fading*, we mean that there is only one path to consider (or at least several paths come in at such a short delay that they can be treated as one). The path loss parameter can therefore be accurately modeled by one time-varying complex scalar (amplitude and phase). Since such a multiplication influences the whole signal spectrum in the same way, flat fading is often also called *frequency nonselective fading*. As with shadowing, the path loss parameter has a statistical distribution. The kind of distribution depends on which one of two main cases we find: First, there is no line-of-sight path and the signal is the composition of a large number of random reflections, or, second, the random reflections are superimposed by a line-of-sight path.



(a) Non-line-of-sight multipath propagation

(b) Line-of-sight multipath propagation

Figure 7.41 The two types of multipath propagation.

Let us consider the first case, which is illustrated by Fig. 7.41(a). Note that as opposed to shadowing, where the composition of one path consisted of several multiplicative effects, we have now several paths adding up to a fluctuating, complex path-loss factor. Since the real part and the imaginary part of the complex path-loss factor (remember we model everything in the baseband) are independent processes, each of the part is a zero-mean Gaussian random variable. It can be shown that the corresponding distribution of the phase, essentially the argument of real and imaginary part, is uniformly distributed over $[0 \ldots 2\pi]$ and the amplitude r has a Rayleigh distribution given by its probability-density function (pdf)

$$p(r) = \frac{r}{\sigma^2}\,\mathrm{e}^{-\frac{r^2}{2\sigma^2}}. \tag{7.101}$$

The pdf of the Rayleigh distribution is depicted by Fig. 7.42. Such a fading process is therefore termed Rayleigh fading. It concerns the loss factor L in the linear domain (not the dB domain).

In the second case, which is illustrated by Fig. 7.41(b), we have a strong line-of-sight signal in addition to the scatterer. The distribution of the amplitude is a Ricean distribution given as

$$p(r) = \frac{r}{\sigma^2}\,\mathrm{e}^{-(r^2+s^2)/2\sigma^2} I_0\left(\frac{rs}{\sigma^2}\right), \tag{7.102}$$

where $s^2 = m_1^2 + m_2^2$ is the noncentrality parameter consisting of the I- and Q-contribution of the signal present and I_0 is the modified Bessel function of the first kind and degree zero. The pdf of the Rice distribution is also depicted by Fig. 7.42. As opposed to the Rayleigh distribution, the Rice distribution looks much more symmetrical, due to the influence of the direct path. Still, no negative values are possible, so it is slightly asymmetrical. For large values of s, the Rice distribution can be accurately modeled by a Gaussian distribution.

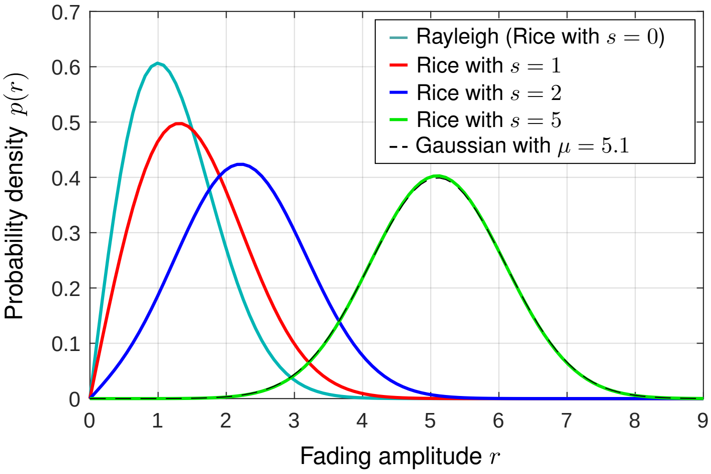


Figure 7.42 Probability density functions of the Rayleigh and the Rice distributions. Both distributions are normalized to $\sigma = 1$.

The technique to evaluate the path loss due to fast fading is for either case the same as for shadowing: Specify the needed percentage (probability of successful communication) and evaluate the minimum amplitude via the cumulative density function (or the tail function). This amplitude then directly converts into a fast-fading power loss (or fade margin) to put into the link budget.

One direct application of the Rayleigh distribution given by Eq. (7.101) of a flat fading process is of direct interest to the communications engineer: the BER curve of a BPSK signal in a flat Rayleigh-fading channel.

We have seen in Chapter 3 that the BER of a BPSK signal is of the form

$$P_E(E_s/N_0) = P_E(\gamma) = Q\left(\sqrt{\frac{2E_s}{N_0}}\right). \tag{7.103}$$

and is clearly a function of the SNR as given by $\gamma = \text{SNR} = E_s/N_0$. The underlying assumption at the point of derivation was that the symbol energy was constant, leading to a constant γ. Now when there is fading, this is no longer true. It can be shown that if the amplitude is Rayleigh distributed, the SNR has a chi-square distribution with two degrees of freedom, which is an exponential distribution given by

$$p(\gamma) = \frac{1}{\Gamma}\,\mathrm{e}^{-\frac{\gamma}{\Gamma}}, \tag{7.104}$$

where Γ is the average SNR of the fading process. We now have to evaluate a tail function, whose argument is exponentially distributed. The BER is then given as the expectation of the different BERs for varying SNR according to their distribution given by Eq. (7.104). Hence,

$$\begin{aligned} P_E &= E\left\{P_E(\gamma)\right\} \\ &= \int_0^\infty \frac{1}{\Gamma}\,\mathrm{e}^{-\frac{\gamma}{\Gamma}} Q(\sqrt{2\gamma})d\gamma \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}} \int_{\sqrt{2\gamma}}^\infty \mathrm{e}^{-\frac{x^2}{2}} dx\, \frac{1}{\Gamma}\,\mathrm{e}^{-\frac{\gamma}{\Gamma}} d\gamma \\ &= \frac{1}{\sqrt{2\pi}\Gamma} \int_0^\infty \int_0^{\frac{x^2}{2}} \mathrm{e}^{-\frac{\gamma}{\Gamma}} d\gamma\ \mathrm{e}^{-\frac{x^2}{2}} dx. \end{aligned} \tag{7.105}$$

The last equality is obtained by changing the integration order. The inner integral can now easily be computed. Hence,

$$\begin{aligned} P_E &= \frac{1}{\sqrt{2\pi}\Gamma} \int_0^\infty \Gamma\left(1 - \mathrm{e}^{-\frac{x^2}{2\Gamma}}\right) \mathrm{e}^{-\frac{x^2}{2}} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \mathrm{e}^{-\frac{x^2}{2}} - \mathrm{e}^{-\frac{x^2}{2}(1+\frac{1}{\Gamma})} dx \\ &= \frac{1}{2} - \frac{1}{2}\sqrt{\left(1 + \frac{1}{\Gamma}\right)^{-1}} \\ &= \frac{1}{2}\left(1 - \sqrt{\frac{\Gamma}{1+\Gamma}}\right). \end{aligned} \tag{7.106}$$

The result can be visualized in Fig. 7.43. The BER of a flat Rayleigh fading channel does not improve as fast as in the AWGN case when increasing the SNR. SERs/BERs of other binary modulation schemes are given in Table 7.4 without derivation.

The above modeling of a channel using Rayleigh and Rice distributions is in fact only half the truth. In addition to this, second-order statistics (essentially how fast the channel changes) have to be incorporated to make a channel model more realistic. The mechanism behind the fact that the fading process is correlated rather than white can be characterized by the *coherence time*, the time during which a channel is considered constant, which is inversely proportional to the *Doppler spread*, a channel characteristics that indicates how wide the frequency occupation of a signal is spread due to a moving transmitter, receiver, or both.

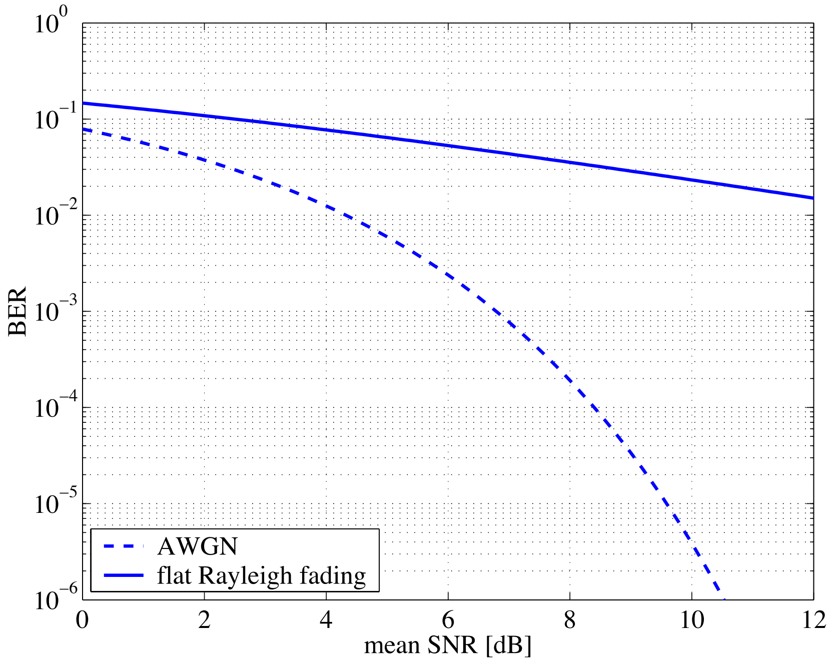


Figure 7.43 Bit-error rates of a BPSK signal in the AWGN case and in a flat Rayleigh-fading channel, respectively.

Modulation	SER/BER (Rayleigh fading (flat))
BPSK	$\frac{1}{2}\left(1-\sqrt{\frac{\Gamma}{1+\Gamma}}\right)$
DPSK (noncoherent)	$\frac{1}{2(1+\Gamma)}$
2-FSK (orthogonal, coherent)	$\frac{1}{2}\left(1-\sqrt{\frac{\Gamma}{2+\Gamma}}\right)$
2-FSK (orthogonal, noncoherent)	$\frac{1}{2+\Gamma}$

Table 7.4 SERs for some digital modulation formats in a flat Rayleigh-fading channel.

Frequency-Selective Fading

As the delays between the individual paths of a multipath signal get larger with respect to the symbol duration, the fading process can no longer be considered flat, or frequency non-selective. In fact, with larger delays between rays, the channel becomes a wideband fast-fading channel or a *frequency-selective fading* channel. Such a channel can often be modeled by a baseband FIR filter, whose complex-valued taps are independently faded according to the statistics (either Rayleigh or Rice distributed as seen in the last section), see Fig. 7.44.

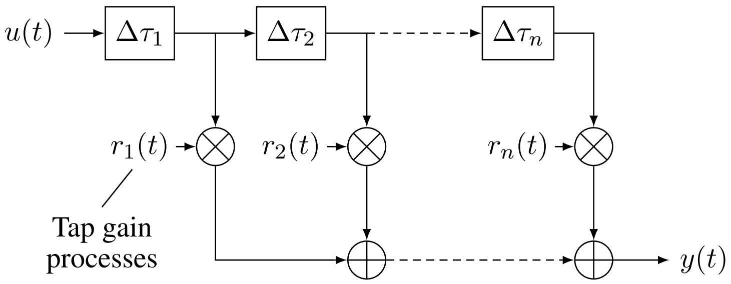


Figure 7.44 Frequency-selective fading channel model.

The relationship between input and output is given by the convolution

$$y(t) = (h * u)(t) = \int h(\tau)u(t-\tau)d\tau, \tag{7.107}$$

where the impulse response of the wireless channel is

$$h(t) = \sum_{k=1}^{n} r_k(t)\delta(t-t_k). \tag{7.108}$$

The channel impulse response is often abbreviated as CIR in the literature. Its Fourier transform is called the channel transfer function (CTF). Note that the channel taps are both complex-valued and time-variant. The mean powers of the taps

$$P_k = E\{|r_k|^2\} \tag{7.109}$$

are usually given by a channel profile together with the corresponding delays, the so-called *power delay profile* (see Fig. 7.45), which is further characterized by the following parameters:

- *total excess delay*: the delay between the first and the last arriving tap response, essentially the amount by which the duration of a transmitted symbol is extended by the channel,

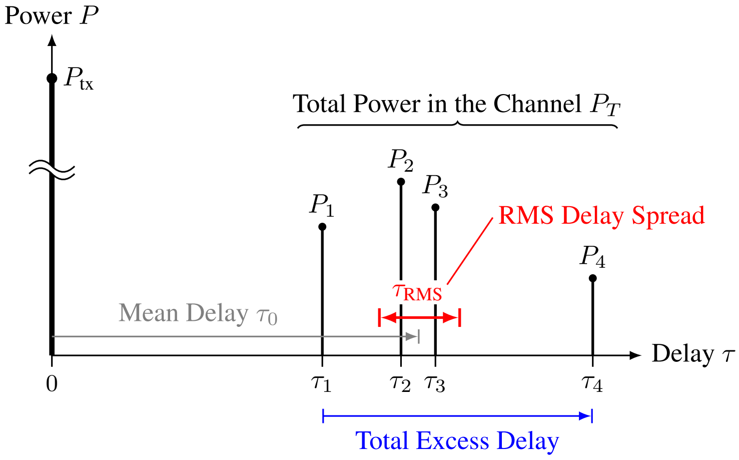


Figure 7.45 Power delay profile.

- *mean delay* defined by

$$\tau_0 = \frac{1}{P_T}\sum_k P_k\tau_k, \tag{7.110}$$

where, compared to Fig. 7.44, we have

$$\tau_k = \sum_{l=1}^{k} \Delta\tau_l, \tag{7.111}$$

- *RMS delay spread* defined by

$$\tau_{\text{RMS}} = \sqrt{\frac{1}{P_T}\sum_k P_k(\tau_k - \tau_0)^2}, \tag{7.112}$$

where the total power in the channel is given by

$$P_T = \sum_k P_k. \tag{7.113}$$

The last parameter, the delay spread, is a good indicator of how frequency selective the channel is. It is inversely proportional to the *coherence bandwidth*. If the delay spread of a channel is much smaller than the symbol duration, the channel may be considered narrowband (flat fading). On the other hand, large delay spreads compared to symbol lengths mean that intersymbol interference (ISI) occurs, i.e., echos of a previous symbol overlap with the current incoming symbol.

7.10.3 Mobile Environment

In a truly mobile communication environment we face two challenges: Firstly, mountains, hills, and buildings lead to excessive delay spread, which in turn introduce frequency-selective fading. Secondly, changing the user position makes the channel response highly variable. An RF receiver usually needs to address both problems, the first one by providing a flexible structure that can essentially invert the channel, and the second one by providing an algorithm to track the channel changes. Such an equalizer structure is shown in Fig. 7.46.

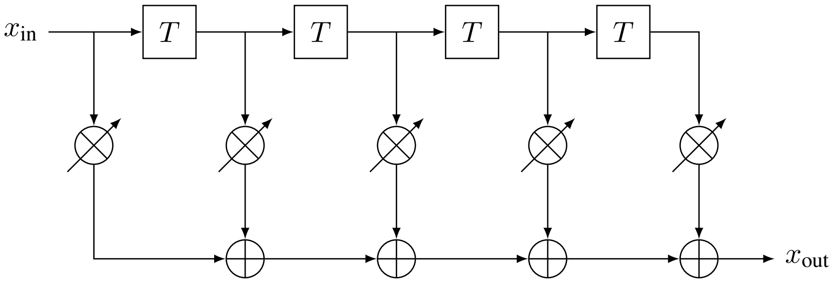


Figure 7.46 Block diagram of a channel equalizer.

7.10.4 Relationship of Fading Parameters

In the last sections, we have encountered different types of fading. These types can be characterized by different parameters, some of which are related to each other.

The coherence time T_{coh}, which characterizes the time over which the channel transfer function stays essentially the same, and the Doppler spread, which describes the spectral broadening, are inversely related, thus

$$T_{\text{coh}} \propto \frac{1}{f_d}. \tag{7.114}$$

On the other hand, the coherence bandwidth B_{coh} determines the span over which the channel appears flat. Its inverse parameter, the delay spread τ stands for the time broadening. Again these two parameters relate to each other as

$$B_{\text{coh}} \propto \frac{1}{\tau}. \tag{7.115}$$

In summary, for a given communication system with symbol time T and bandwidth B we can now build Table 7.5. Note that a fading channel can be at the same time one of the top two types and one of the bottom two types, hence, any one of four combinations is possible. In fact, the proportionality factors in Eqs. (7.114) and (7.115) depend on the exact behavior of the channel impulse response and its variation in time, respectively. Rules of thumb are given by [56]

Fading type	**Condition**	**Alternative formulation of condition**
Flat fading	$B_{\text{coh}} \gg B$	$\tau \ll T$
Frequency-selective fading	$B_{\text{coh}} < B$	$\tau > T$
Slow fading	$T_{\text{coh}} \gg T$	$f_d \ll B$
Fast fading	$T_{\text{coh}} < T$	$f_d > B$

Table 7.5 Fading types.

$$B_{\text{coh}} \approx \frac{1}{50\tau}, \tag{7.116}$$

if the frequency correlation function is above 0.9, and

$$B_{\text{coh}} \approx \frac{1}{5\tau}, \tag{7.117}$$

if the frequency correlation function is above 0.5. Sometimes, the relationship

$$B_{\text{coh}} \approx \frac{1}{2\pi\tau} \tag{7.118}$$

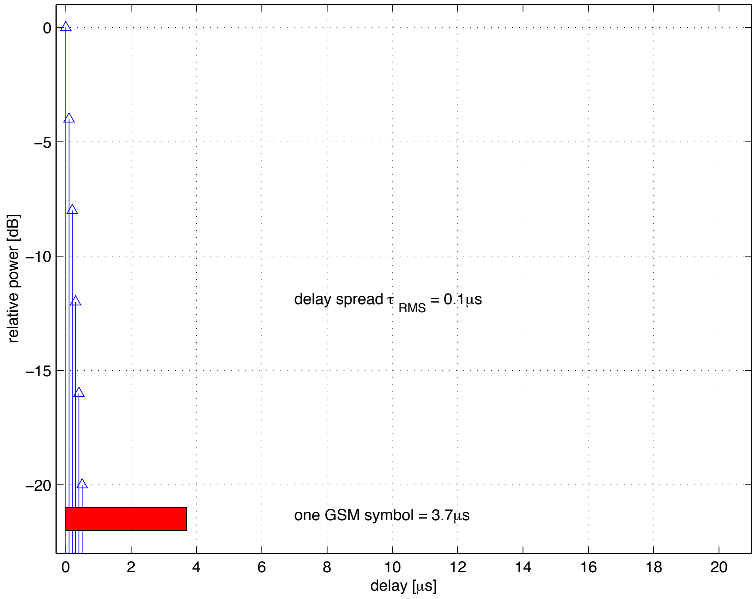
is also used. Very similar rules can be found for the relationship given by Eq. (7.114). Rappaport [56] states that

$$T_{\text{coh}} \approx \frac{9}{16\pi f_d} \tag{7.119}$$

for the time correlation to be above 0.5.

7.10.5 Examples of Channel Models

Channel models against which radio equipment are tested, are often given by standardization bodies. In the following, we list three typical channel models as standardized by ETSI for use in the GSM system. The 'x' stands for the speed, which does not, however, affect the tap model as such, but the underlying fading process.



Tap no.	Relative delay [μs]	Av. rel. power [dB]
1	0	0
2	0.1	−4
3	0.2	−8
4	0.3	−12
5	0.4	−16
6	0.5	−20

Figure 7.47 Rural area (RAx) propagation model for GSM.

In Figs. 7.47 to 7.49 values for comparison with the symbol length used in GSM are given in both graphical and tabular form.

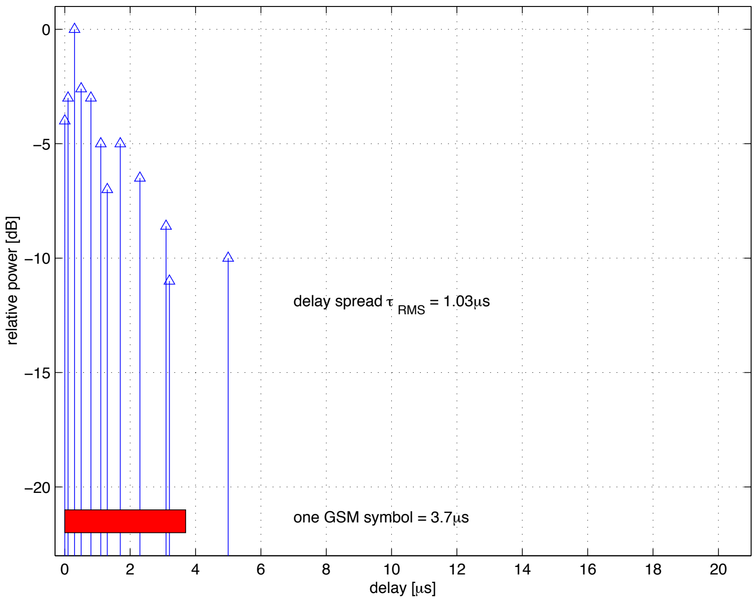
7.10.6 Ways Out

If we have to design a wireless communication system in such a hostile environment as mentioned above, we can either accept the situation and deploy a channel equalizer, or try to seek the situation where we do not need one. If we look at a typical snapshot of a channel frequency response, e.g., Fig. 7.50, we realize that the channel looks quite flat as long as we only use a narrow part (subband) of it, indicated by the region between the dashed lines. We then say that the bandwidth (of the subband) is smaller than the coherence bandwidth

$$W_{\mathrm{sub}} \ll B_{\mathrm{coh}}. \tag{7.120}$$

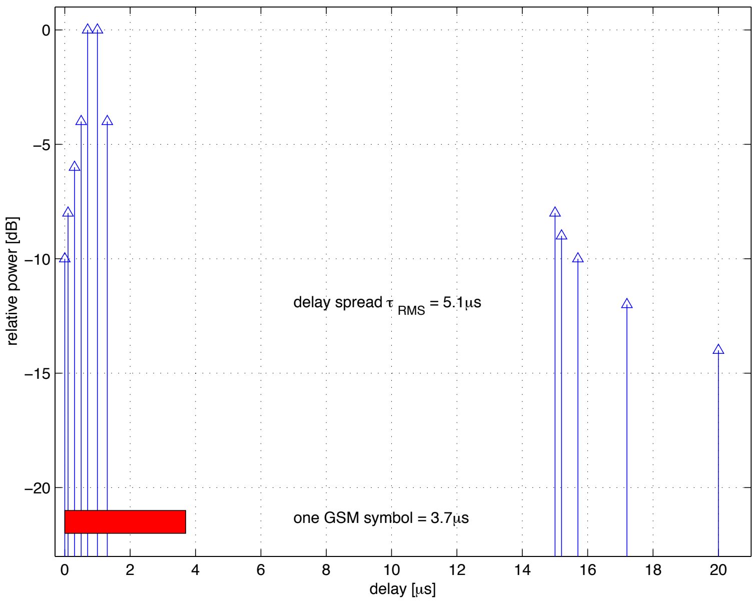
The idea of dividing the available bandwidth into a multitude of subbands that are smaller than the coherence bandwidth was realized by Chang [11] in the 60s. All we need to 'equalize' the channel in each 'flat' subband, is a variable factor, a complex-valued factor, in order to restore amplitude and phase. Why does this factor need to be variable? Well, because each subband may show a different attenuation to be compensated for.

Moreover, the attenuation may change, and so does the complex-valued factor we have to multiply the subband with. We better make sure the situation does not change within the duration of a symbol, otherwise



Tap no.	Relative delay [µs]	Av. rel. power [dB]
1	0	−4
2	0.1	−3
3	0.3	0
4	0.5	−2.6
5	0.8	−3
6	1.1	−5
7	1.3	−7
8	1.7	−5
9	2.3	−6.5
10	3.1	−8.6
11	3.2	−11
12	5	−10

Figure 7.48 Typical urban (TUx) propagation model for GSM.



Tap no.	Relative delay [µs]	Av. rel. power [dB]
1	0	−10
2	0.1	−8
3	0.3	−6
4	0.5	−4
5	0.7	0
6	1	0
7	1.3	−4
8	15	−8
9	15	−9
10	16	−10
11	17	−12
12	20	−14

Figure 7.49 Hilly terrain (HTx) propagation model for GSM.

we cannot determine the compensation factor. Hence, we have a second condition on the choice of our

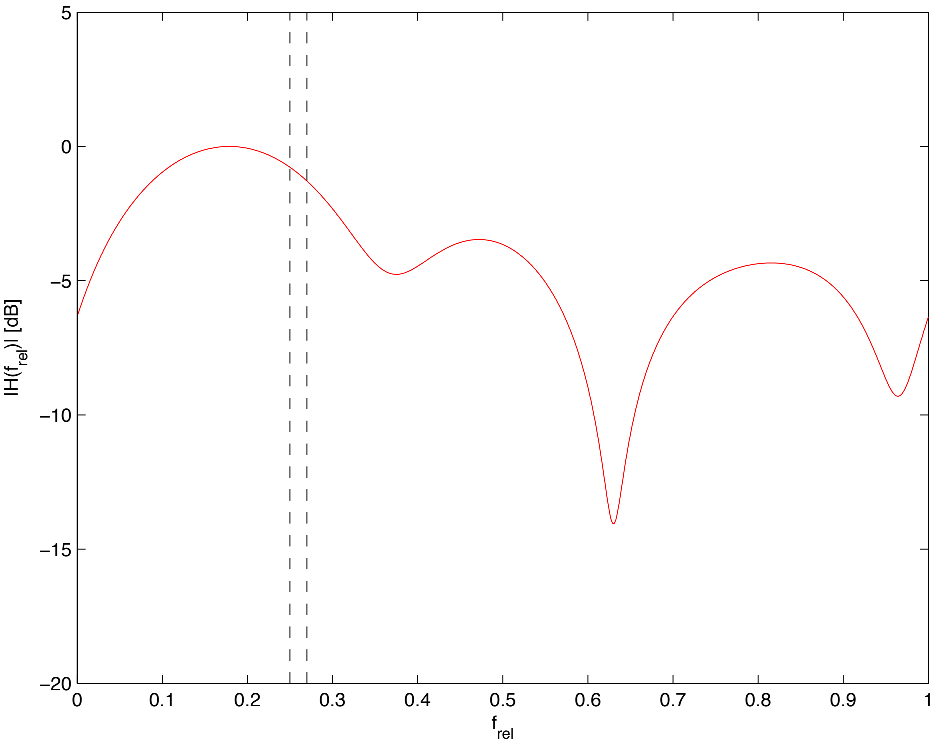


Figure 7.50 Frequency response snapshot of a frequency-selective fading channel.

system. The duration of a symbol within a subband shall be much shorter than the coherence time

$$T_{\text{sub}} \ll T_{\text{coh}}. \tag{7.121}$$

Dividing a bandwidth W into N_c subbands delivers

$$W_{\text{sub}} = \frac{W}{N_c} \tag{7.122}$$

and

$$T_{\text{sub}} = \frac{1}{W_{\text{sub}}} = \frac{N_c}{W}. \tag{7.123}$$

Eqs. (7.120) and (7.121) now deliver

$$\frac{W}{N_c} \ll B_{\text{coh}}, \tag{7.124}$$

$$\frac{N_c}{W} \ll T_{\text{coh}}. \tag{7.125}$$

Combined into one inequality we finally get

$$\frac{W}{B_{\text{coh}}} \ll N_c \ll WT_{\text{coh}}. \tag{7.126}$$

We thus get lower and upper limits for the choice of the number of subchannels N_c. This consideration was already used when OFDM was introduced in Section 3.2.

7.11 Further Literature

The bible for antenna designers is by Balanis [5]. Many other books treat antennas in a chapter or two, but never in the detail provided by Balanis. A good source is also the ARRL Antenna Book [2], which is as good as the more famous ARRL Handbook [1] printed every year. Fundamentals are treated and described in a very understandable manner, although the types of antennas are somewhat restricted due to their application to radio amateur bands. A German book is the one by Kark [33]. A neat work combining antenna theory and propagation models has been published by Saunders [66].

Part II

Applications

8 Radar Systems

8.1 Introduction

There are applications where radio frequency signals are used not solely to communicate but to measure distances. GPS, or more generically speaking GNSS, is currently the most prominent system applying this approach. There are many other systems that use RF signals to range objects. They are usually called RADAR, which is short for *radio detection and ranging*. An illustration of a typical RADAR system is shown in Fig. 8.1.

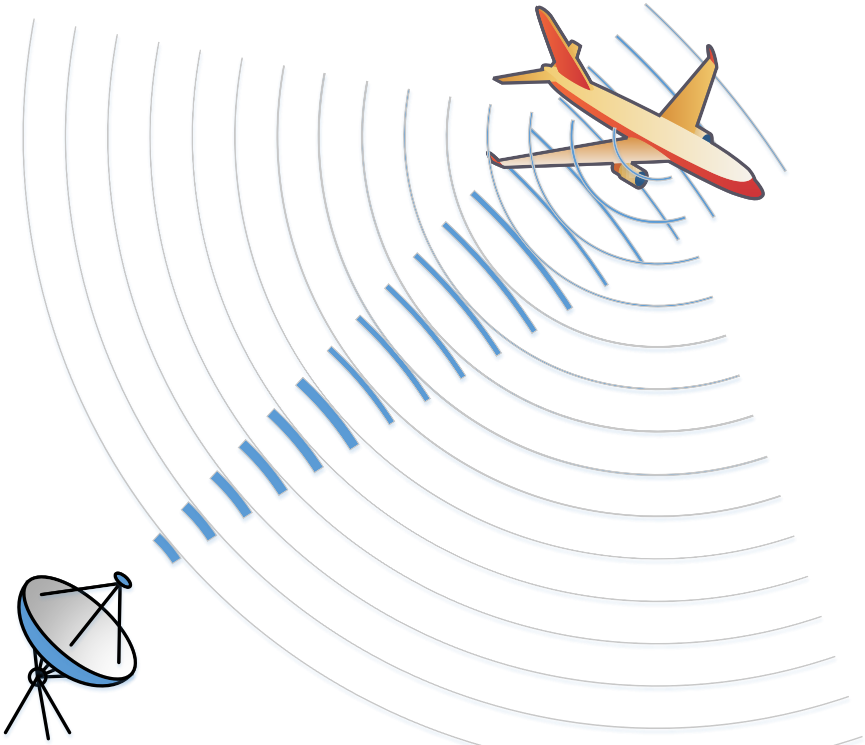


Figure 8.1 RADAR principle with reflected field from an airplane.

Radar systems can be divided into two groups, impulse radar, which is mainly used for military and aviatic purposes, and CW radar, the most famous example of which is speed radar used by police forces all over the world, although lately more and more replaced by more precise laser measurements.

8.2 Frequency Bands

Frequencies up to 30 MHz have the advantage that ground wave propagation and ionospheric refraction and reflection facilitate over-the-horizon usage. However, large antenna arrangements and spectrum scarcity make these frequency bands rather unattractive. VHF and UHF ranges are used for Radar, although maybe not with the same popularity as the L band and higher bands. Table 8.1 shows some of the most popular frequency bands used in radar.

Frequency	Band name	Radar specific
30–300 MHz	VHF	138–144 MHz 216–225 MHz
300–1000 MHz	UHF	420–450 MHz 890–942 MHz
1–2 GHz	L	1.215–1.4 GHz
2–4 GHz	S	2.3–2.5 GHz 2.7–3.7 GHz
4–8 GHz	C	5.25–5.925 GHz
8–12 GHz	X	8.5–10.68 GHz
12–18 GHz	Ku	13.4–14.0 GHz 15.7–17.7 GHz
18–27 GHz	K	24.05–24.25 GHz
27–40 GHz	Ka	33.4–36.0 GHz
40–75 GHz	V	59-64 GHz
75–120 GHz	W	76–81 GHz 92–100 GHz
110–300 GHz	millimeter waves	126–142 GHz 144–149 GHz 231–235 GHz 238–248 GHz

Table 8.1 Radar-specific bands (depending on region, here IEEE Region 2).

8.3 The Radar Equation

Regardless of the kind of pulses that are transmitted, the power received in any radar system follows the so-called *Radar equation*, which describes the dependency of the received power to the distance r and the radar cross section (RCS) A_{obj} of the object. The latter is not only depending on the physical size, but also on the shape and the material of the object. Although in radar systems, line-of-sight is usually a prerequisite, the dependency of the received power to the distance is not to the power of 2 but to the power of 4. The reason for that is that we have to consider two links in a multiplicative way, one to the object and one from the object. Both distance dependencies are of quadratic nature, hence the power of 4 in the product. In mathematical terms the radar equation can be obtained by considering the power density at the object in the free-space propagation model according to Eq. (7.57)

$$S_1 = \frac{P_{\text{tx}_1} G_{\text{tx}_1}}{4\pi r^2}, \tag{8.1}$$

where P_{tx_1} and G_{tx_1} are the transmitted power and the gain of the transmitter antenna, respectively. The power "received" by the object is thus proportional to the object size

$$P_{\text{rx}_2} = S_1 \cdot A_{\text{obj}}. \tag{8.2}$$

Similarly, on the reflected path back from the object to the transmitter we have again the power density

$$\begin{aligned} S_2 &= \frac{P_{\text{rx}_2} G_{\text{obj}}}{4\pi r^2} \\ &= \frac{S_1}{4\pi r^2} \cdot A_{\text{obj}} \cdot G_{\text{obj}} \\ &= \frac{S_1}{4\pi r^2} \cdot \sigma, \end{aligned} \tag{8.3}$$

where we call σ the radar cross section (RCS) of an object. The RCS of an object describes the size and form involved in the reflection process. Finally, the reflected power back at the antenna is above power density multiplied with the area of the radar antenna A_{eff}

$$\begin{aligned} P_{\text{rx}_1} &= S_2 \cdot A_{\text{eff}} \\ &= S_2 \cdot \frac{\lambda^2 \cdot G_{\text{rx/tx}}}{4\pi} \\ &= \frac{\lambda^2}{4\pi} \cdot G^2_{\text{rx/tx}} \cdot \frac{P_{\text{tx}_1}}{(4\pi r^2)^2} \cdot \sigma \\ &\boxed{P_{\text{rx}_1} = \frac{\lambda^2}{(4\pi)^3} \cdot G^2_{\text{rx/tx}} \cdot \frac{P_{\text{tx}_1}}{r^4} \cdot \sigma}. \end{aligned} \tag{8.4}$$

We can clearly see that the received power decreases with the fourth power of the distance, a characteristic very typical for radar systems.

8.4 Modes of Operation

Radar system are divided into different categories. One main grouping divides them into imaging and non-imaging radars. Imaging radar systems produce measurements on a map, e.g., airport surveillance, weather radar. Non-imaging radar systems produce one-dimensional measurement results, e.g., altimeter, speed traffic enforcement. The latter classification is carried out according to the display of the result. Yet another classification distinguishes systems whose transmitter and receiver are collocated (monostatic radar) or separated (bistatic radar).

8.4.1 Impulse Radar

Mainly used in aeronautic and military applications, the impulse radar allows the most sensitive identification of objects flying or swimming around us. The most common radar waveform is a train of narrow, rectangular-shape pulses modulating a carrier frequency. The distance (or range) r to the target object is then determined by measuring the time t it takes the pulse to travel (at approximately the speed of light c_0) to the target and back [75]:

$$r = \frac{c_0\, t}{2}, \tag{8.5}$$

$$r\ [\mathrm{km}] \approx 0.15\, t\ [\mu\mathrm{s}]. \tag{8.6}$$

In order to get the desired resolution, the pulses must often be very short. On the other hand, for each transmitted pulse a sufficient length of time must elapse to allow any echo signals to return and be dectected before the next pulse is transmitted. The rate at which the pulses may be transmitted is determined by the longest range at which targets are expected. If the pulse repetition is too high, *multiple-time-around* echoes [75] resulting in ambiguous measuring ranges might result. The maximum unambiguous range, prohibiting any multiple-time-around echoes is

$$r_{\max} = \frac{c_0}{2\, f_p}, \tag{8.7}$$

where f_p is the pulse repetition frequency.

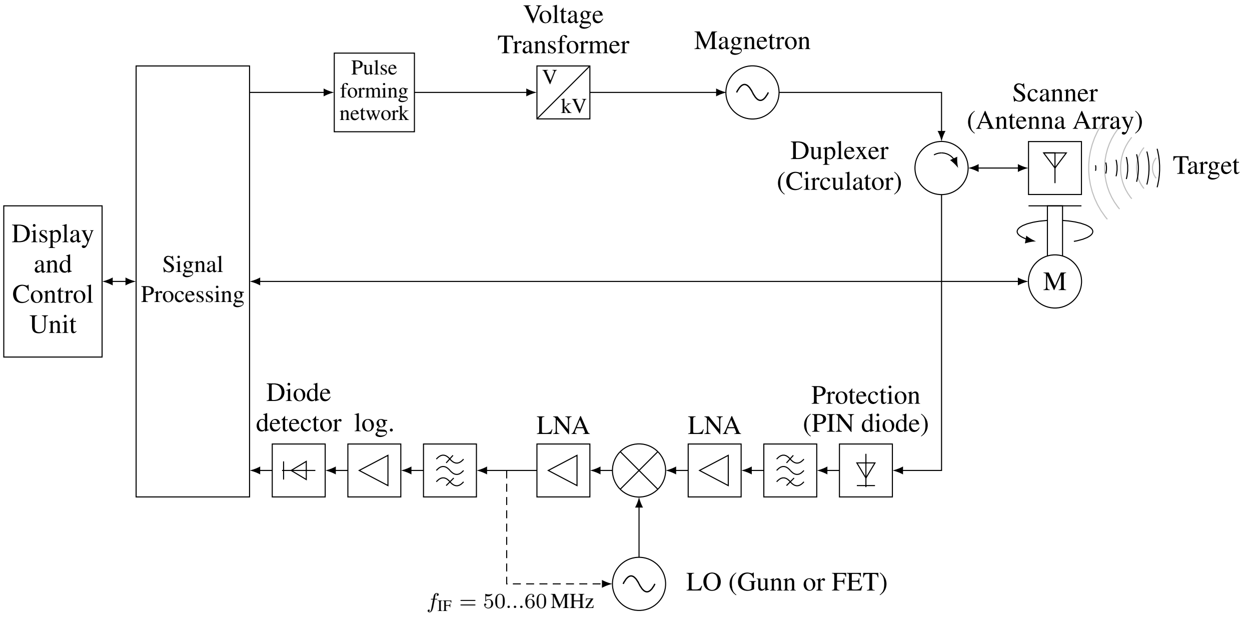


Figure 8.2 Typical marine radar system configuration (non-coherent reception).

Marine radars are permitted to operate in the X-band between 9.3 and 9.5 GHz and in the S-band from 2.9 to 3.1 GHz [75]. Due to limited duration in time, the peak power must be high in order to be able to 'see' the received signal clear enough (and be able to distinguish it from clutter). This is normally achieved by using high peak power magnetrons in the transmitter, newer (and more expensive) coherent transceiver designs also use systems containing RF transistors, travelling-wave-tubes (TWT) and klystrons. The peak transmit power is typically in the range of between 2 to 10 kW for pleasure craft systems and 2 to 60 kW for commercial vessels (and even more for military operations). In addition, the radar antennas (scanner) usually have a gain of 25 to 33 dBi, with horizontal beam widths ranging from around 2.5° to less than 1° and a 25° wide beam in the vertical and rotate at 20 to 50 rpm.

To be able to measure the time delays most accurately, the pulse slopes need to be as steep as possible. The ITU requirements[1] restrict the rise time to $\geq 10\,\text{ns}$ to limit out-of-band interference. The fall time is usually even longer, due to the electromagnetic inertia of the magnetron [75].

8.4.2 CW Radar

Almost 200 years ago, the Austrian physicist Christian Doppler (1803–1853), described in theory what is today known as the *Doppler effect*: if the source or destination of some wave transmission system is moved in the direction of the propagation between source and destination, the waves get compressed or expanded. Thus, the frequency of the obtained signal may change slightly over the frequency transmitted. In the case of the radar, the radar object is both the destination (target) and the source (reflection). The Doppler frequency shift is thus twice the usual factor, namely

$$f_{\text{Doppler}} = 2\,\frac{f \cdot v}{c - v} \approx 2\,\frac{f \cdot v}{c} \quad (\text{if } v \ll c), \tag{8.8}$$

where f is the frequency of operation, v is the speed in the direction of the propagation, and c is the speed of light. If the target is moving towards the radar, the Doppler shift f_d will be positive, otherwise negative.

8.4.3 FMCW Radar

With CW radar, the signal can either be a single frequency signal or can be varied in its frequency, so-called *frequency-modulation continuous wave* (FMCW) radar. A block diagram of a radar system is shown in Fig. 8.3. The linear frequency modulation (LFM) method employs periodic frequency ramps as its transmitted measurement signal. The current transmission frequency $f_T(t)$ varies linearly from $f_{\min}$ to $f_{\max}$ during the measurement time $0 \leq t \leq T_m$ as shown in Fig. 8.4. The slope of the ramp during the measurement period T_m is

$$\frac{d}{dt} f_T(t) = \frac{f_{\max} - f_{\min}}{T_m} = \frac{\Delta f}{T_m}, \tag{8.9}$$

where $\Delta f = f_{\max} - f_{\min}$ is the frequency difference of the ramp. The current frequency for the increasing and decreasing ramp, corresponding to Fig. 8.4(b), is

$$f_{T\uparrow}(t) = f_{\min} + \Delta f\,\frac{t}{T_m}, \tag{8.10}$$

$$f_{T\downarrow}(t) = f_{\min} + \Delta f\,\frac{T_m - t}{T_m}\ . \tag{8.11}$$

[1] Additionally, the International Maritime Organisations International Convention for Safety of Life at Seas (IMO SOLAS) gives some minimum detection performances in clear conditions: e.g. range measurements have to be within 30 m accuracy or withing 1% of the maximum ranging scale used and withing 1° bearing (azimuth angle). The more technical details are defined within IEC 60945 (such as minimum shock, vibration and corrosion resitivities) and IEC 61162, which defines the messages used for navigation and radiocommunications equipment to interchange digital data. Minimum antenna parameters are specified in IEC 62388 and 62252 [75].

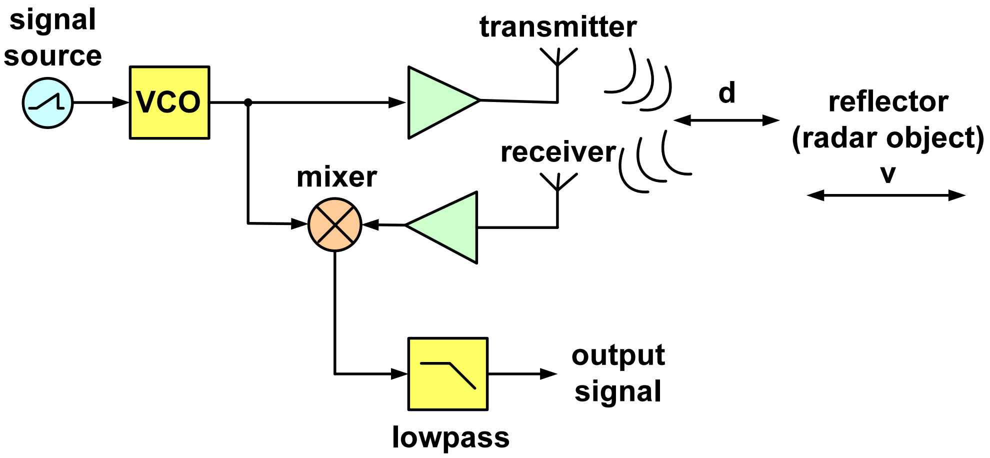


Figure 8.3 FMCW radar principle.

Note that although the ramps defined in the above equations follow one after the other (not simultaneously), we consider $0 \leq t \leq T_m$ in either equation. Thus, the transmitted signal has the form

$$s_T(t) = \cos(2\pi f_T(t) \cdot t). \tag{8.12}$$

This signal leaves the transmit antenna, hits a target at distance d, where a small fraction is reflected back to the receiver, where in turn it is received after a time of flight of $\tau = 2d/c$. In the receiver, the delayed received signal is mixed with the current transmitted signal producing sum and difference frequency. Whereas the sum frequency is high and not interesting, the difference frequency f_{diff} is typically in the AF band. The frequency is constant during each slope and has the values

$$f_{\text{diff}\uparrow} = f_{T\uparrow}(t) - f_{R\uparrow}(t) = f_{T\uparrow}(t) - f_{T\uparrow}(t-\tau) = \Delta f\, \frac{\tau}{T_m}, \tag{8.13}$$

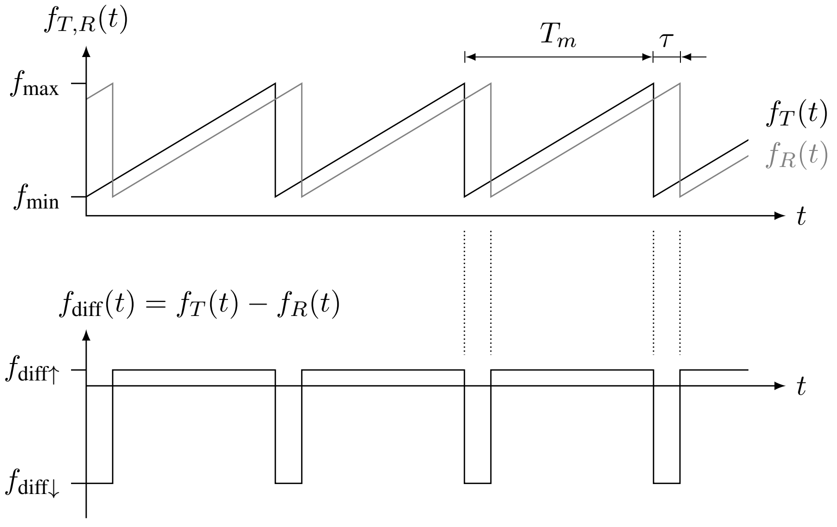
$$f_{\text{diff}\downarrow} = f_{T\uparrow}(t) - f_{R\uparrow}(t) = f_{T\downarrow}(t) - f_{T\downarrow}(t-\tau) = -\Delta f\, \frac{\tau}{T_m}. \tag{8.14}$$

Discontinuities arise at the ends of the ramps as the slopes suddenly change. The frequency course at these points depends on the form of the frequency exciter (sawtooth or triangle). This effect can be observed at the bottom of Fig. 8.4. For a sawtooth signal, the difference frequency during the time τ results in a value of $f'_{\text{diff}\uparrow} = f_{\text{diff}\uparrow} - \Delta f = -\Delta f\,(1 - \tau/T_m)$, which is generally much higher (in its absolute value) than $f_{\text{diff}\uparrow}$ and is thus attenuated by the AF filter.

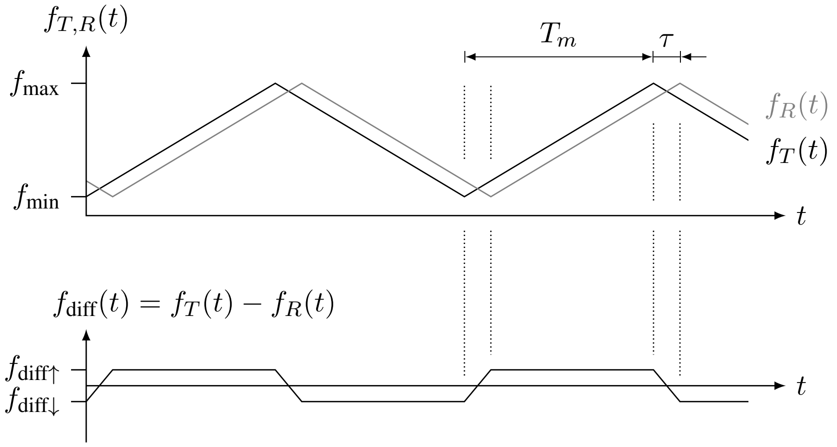
If a radar target is moving with an angle α relative to the antenna and the speed v, the received frequency is additionally shifted by the Doppler frequency

$$f_{\text{Doppler}} = 2 \cdot \frac{f_T \cdot v \cdot \cos\alpha}{c}\ . \tag{8.15}$$

The Doppler shift is experienced in both the forward and the return direction, hence the factor two. The total frequency shift depends on the distance d between the antenna and the target, and the target speed (resulting in a Doppler shift). Note that the sign of the frequency shift of the first part depends on the slope of the ramp (up or down), whereas the sign of the frequency shift of the second part depends on the movement direction. For a movement in the positive velocity direction v (away from the radar observer), we get the negative sign


$f_{T,R}(t)$
T_m
τ
$f_{\max}$
$f_{\min}$
$f_T(t)$
$f_R(t)$
t
$f_{\text{diff}}(t) = f_T(t) - f_R(t)$
$f_{\text{diff}\uparrow}$
$f_{\text{diff}\downarrow}$
t

(a) Sawtooth


$f_{T,R}(t)$
T_m
τ
$f_{\max}$
$f_{\min}$
$f_R(t)$
$f_T(t)$
t
$f_{\text{diff}}(t) = f_T(t) - f_R(t)$
$f_{\text{diff}\uparrow}$
$f_{\text{diff}\downarrow}$
t

(b) Triangle

Figure 8.4 Signal forms of the two most common LFM-CW radar methods.

for the Doppler shift frequency in Eq. (8.16). Thus,

$$\begin{aligned} f_{\text{diff,tot}\uparrow} &= f_{\text{diff}\uparrow} - f_{\text{Doppler}} \\ &= \frac{\Delta f \tau}{T_m} - \frac{2 f_T\, v}{c} \cdot \cos\alpha, \end{aligned} \tag{8.16}$$

$$\begin{aligned} f_{\text{diff,tot}\downarrow} &= f_{\text{diff}\downarrow} - f_{\text{Doppler}} \\ &= -\frac{\Delta f \tau}{T_m} - \frac{2 f_T\, v}{c} \cdot \cos\alpha. \end{aligned} \tag{8.17}$$

The negative sign for the Doppler frequency means that a velocity vector away from the measurement antennas decreases the total difference frequency. We can now solve the two equations for our variables of interest, v and d. Adding Eqs. (8.16) and (8.17) results in

$$f_{\text{diff,tot}\uparrow} + f_{\text{diff,tot}\downarrow} = -\frac{4 f_T\, v}{c} \cdot \cos\alpha, \tag{8.18}$$

or

$$\boxed{v = -\frac{f_{\text{diff,tot}\uparrow} + f_{\text{diff,tot}\downarrow}}{4f_T \cos\alpha} \cdot c.} \tag{8.19}$$

Likewise, subtracting Eq. (8.17) from Eq. (8.16) results in

$$f_{\text{diff,tot}\uparrow} - f_{\text{diff,tot}\downarrow} = \frac{2\Delta f \tau}{T_m} = \frac{2\Delta f}{T_m} \cdot \frac{2d}{c}, \tag{8.20}$$

or

$$\boxed{d = \frac{f_{\text{diff,tot}\uparrow} - f_{\text{diff,tot}\downarrow}}{4\Delta f} \cdot T_m \cdot c.} \tag{8.21}$$

Note, that the signs of $f_{\text{diff,tot}\uparrow}$ and in particular of $f_{\text{diff,tot}\downarrow}$ might be negative. In order to properly determine the sign, an IQ mixer is needed. By making assumptions on the sign, a single mixer might suffice.

8.5 Radar Applications

There are many types of radar in use today. In the following we can't but list the most important ones. The list is not exhaustive.

8.5.1 Surveillance

We all know radar probably best from its application in controlling air traffic near airports. The antenna is a rotating antenna yielding a fan-shaped beam, i.e., small azimuth angle and large elevation angle. Surveillance radars usually use the L- or S-band in a primary radar mode, in which range, azimuth angle, and radar cross-section are obtained [38]. The secondary radar uses a so-called "information friend or foe" (IFF) concept. In the secondary radar the airplane answers actively to give information on aircraft altitude, speed, and direction of flight. The interrogation frequency from ground to air is at 1.03 GHz, whereas the answering frequency is at 1.09 GHz. The observation display for the surveillance radar is a circular cathode ray tube (CRT), in which the radius of the bright spot shows the range of the aircraft, the angle of the spot is equivalent to the azimuth angle and the brightness itself depends on the radar cross section (RCS).

8.5.2 Military

In military applications radar systems are used for locating and tracking missiles. Such systems are much taller than the civilian systems and frequently use parabolic dish antennas with a diameter of up to 30 m and a pulse power of 1 MW. Also, the signal processing is more sophisticated using chirp modulation within the pulse to improve the range resolution.

8.5.3 Weather

Weather forecasts all over the world use data collected by weather radar systems. X-band or still higher frequencies allow reflections due to water in the atmosphere. Modern systems apply pulse-Doppler radar in order to measure both intensity and speed of rain droplets.

8.5.4 Traffic Enforcement

Speed checking as used by the police around the world is often carried out using CW Doppler radar. Small antennas with a Gunn diode oscillator are used as a transmitter. The receiver antenna is either another small antenna or the same as the transmitter antenna.

8.5.5 Altimeters

An airplane scans the ground continuously using FMCW radar to obtain the flying altitude. Since the movement of the airplane is perpendicular to the transmission of the radar waves, the speed of the airplane does not cause a Doppler shift in the received signal.

8.5.6 Ground-Penetrating Radar

Ground-penetrating radar (GPR) is a non-destructive method to scan the ground using electro-magnetic waves in the UHF/VHF range. Often, short pulses (very wideband) are used. GPR is used to detect objects, e.g., land mines, changes in material, cracks and so on. The frequency used is a compromise between penetration depth and resolution.

8.5.7 Door Openers and Other Applications

Sensors at doors to automatically open the door if somebody is approaching use principles such as infrared, ultrasound, and more and more radar (CW and FMCW). Modules are cheap, transmit very little power and have a reaching distance of a couple of meters only.

8.5.8 Harmonic Radar

Harmonic radar, sometimes referred to as harmonic RFID (1-bit RFID system), is a special type of radar where the reflected signal shows a multiple of the transmitted frequency. This helps suppress self-reception as well as reflections from any other objects or clutter nearby. Most often, the multiple is set to be the second or third harmonic and is generated by a nonlinear device on the radar object side. The radar system is thus often also called a non-linear junction detector, the radar object being essentially a diode. Since other nonlinear devices rectify, multiply and reflect incoming signals as well, spurious responses are produced by electronics parts in the vicinity. Even a rusty nail may produce a spur. On the plus side, the objects are passive and operate without a battery. They essentially scatter back part of the power produced by the transmitter.

In contrast to other radar systems, with this kind of radar only directional information may be retrieved but no distance estimates (except for the binary information within range vs. not within range). The frequencies most often used are around 888 MHz or 915 MHz. Applications can be found in the detection of eavesdropping devices, anti-theft systems for shopping malls, car distance radar, metal reradiation radar, locating of avalanche victims, and tracking of animals such as bugs (real alive ones), frogs, and birds.

An application that has been in existence since the eighties is the the Recco system for finding persons buried in an avalanche. The original patent was filed by people from the Royal Institute of Technology, Stockholm, around 1980, and even determines the diode used (Schottky-barrier-diode HP 5082-2835). The system is popular with a lot of textile manufacturers that embed the passive device into the lining of coats as used by skiers. The transmitter sends a signal of 1.5 to 5 W power at a frequency of 917 MHz or 915 MHz (in Europe) and looks for the reflected signal at twice the frequency. A transmitter/detector uses a 5-element Yagi antenna for the transmission and a 4-element patch antenna for the reception of the signal. Although distance ranges of up to 70 m have been reported, in snow this reduces quickly to below 10 m.

In the following decades, patents have been filed essentially describing the same system but with a completely different application in mind, the locating of golf balls gone off green. Several possible diodes are mentioned in the patent application. The commercial name for it is RadarGolf and it is interesting to see that not only do they operate at 915 MHz, too, but they exhibit exactly the same equivalent circuit, consisting of a Schottky diode put on a dipole with an additional DC path short-circuiting the diode at low-frequency for ESD protection.

When concentrating on tracking small animals, some interesting papers are [59], [48], [12], [55], and [78]. Riley and Smith [59] give a good analysis of the link budget and achieve a range of up to 900 m with a 9.4 GHz signal! However, they made use of a 25 kW transmitter used in naval radar navigation. O'Neil et al. [48] report their practical experience making measurements using the Recco system for tracking insects, except for the tags. Colpitts and Boiteau [12] carried out extensive simulations (all for a transmission frequency of 9.41 GHz) to optimize the tag, MoM simulation using NEC for the antenna, CST Microwave Studio for the inductive loop, and ADS for the harmonic balance simulation including a SPICE model of the diode. Also, they experimented with one of the dipole ends to be the beetle. Psychoudakis et al. [55] work at 6 GHz with a superhet receiver approach with an IF around 2 GHz and the use of a fractal antenna.

8.5.9 Passive Radar

Passive radar systems do not transmit actively. They use signals, such as broadcast signals, that are already in the air. By coherently observing signals at different receiver stations, moving objects, which due to their reflections change the received signal at the respective observation station, may be detected in speed and location.

9 Global Navigation Satellite Systems (GNSS)

9.1 Introduction

The term GNSS refers to any of the four global navigation satellite systems currently in space, namely GPS, Glonass, Galileo and Beidou. Many of the physical aspects are valid for all of them. Since the American GPS was the first such system operating and is still the most wide-spread, we will focus on GPS in this chapter.

NAVSTAR-GPS (Navigation System with Timing and Ranging Global Positioning System), for short GPS, is a system that allows a user to locate his/her position within an accuracy of a few meters, everywhere on the world, provided he/she has an appropriate receiver. Besides position, also accurate time and many other derived parameters such as speed and acceleration are available to the user, see Fig. 9.1.

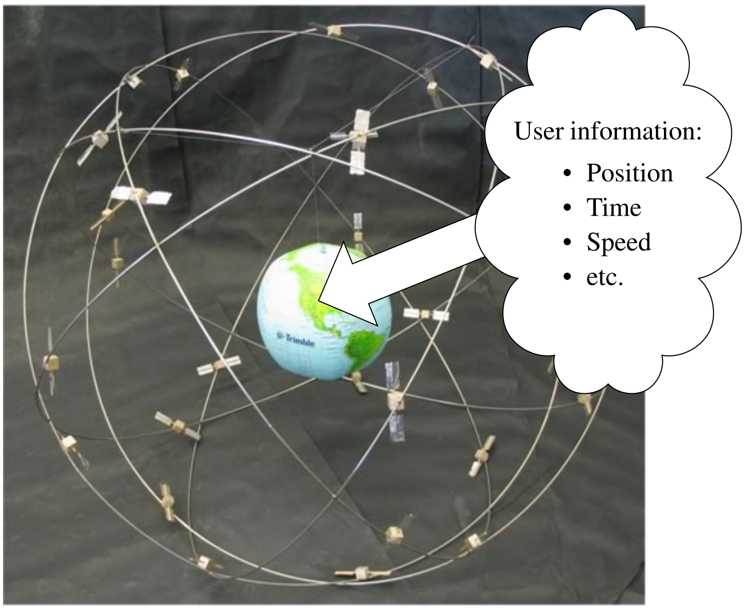


Figure 9.1 Basic idea of GPS.

This model has been built by the lab assistant and is hanging in front of the Lab 2.101. The earth and the orbits are built in a scale of 1:45 Mio. The SV, however, would be smaller than 1 µm at that scale. They are, therefore, shown at a larger scale.

9.1.1 Motivation

The Global Positioning System (GPS) has gained importance at a rapid pace. Although originally developed for the US military, GPS has proven invaluable for a multitude of civilian applications and, therefore, quickly made its way into the consumer market, where numerous products provide some location-based capability based on the reception of GPS. Each application demands specific performance from the GPS receiver, and the associated requirements often vary widely. For the RF engineer, a solid knowledge of the key functionality of GPS is therefore vital. Besides, GPS combines many technologies of the last decades, such as CDMA, PLL tracking, satellite transmission, low-noise receiver design, so that it constitutes an ideal playing field to try and exemplify these technologies.

9.1.2 History

Although the human desire for navigation emerged long ago, satellite-based systems, similar to GPS, only appeared in the early 1970s. The GPS program started in 1973 by its approval. In 1978, the first space vehicle (SV) or satellite was in orbit. In 1993 the GPS system facilitated 24 SV in orbit and was fully operational. Already in the following year, the system was approved for aviation use by the US Federal Aviation Agency (FAA). Although the system was and still is developed by the GPS Joint Program Office (JPO) under the command of the Department of Defense (DOD), it is now intended to consider the needs of civil users for future improvements of the system. In the year 2000, the intentional jittering of precision, called SA (selective availability), was switched off, and even after the terrorist attack of 9/11 (September 2001), the US government has confirmed its intention to leave in place the high-precision access for everyone.

9.1.3 Goals of GPS

The goals of GPS can be listed as follows:

- 3-dim. positioning of constant and moving objects on earth or in proximity
- determination of speed of objects
- time information
- unlimited number of simultaneous users
- independence of weather and climate
- high immunity to arbitrary and intentional disturbances
- high positioning precision (RMS of 30 m) and unambiguous results
- high speed precision (RMS of 0.3 m/s)
- high time precision (RMS of 10 ns)
- time to first fix: some minutes; reacquisition: less than 30 s

9.1.4 Requirements to RF signals

The three main requirements of GPS with respect to its RF signal are:

- high bandwidth (frequencies over 1 GHz)
- small antennas (high frequencies)
- low propagation loss (low frequencies)

The compromise to satisfy these partly contradictory requirements was to use an L-band frequency.

9.2 Concept

The underlying idea of GPS is *distance ranging*. This technique is very well known by almost anybody to estimate the distance of lightning. By knowing that the speed of light is around $v_l = 300'000$ km/s compared to only $v_s = 330$ m/s for the speed of sound, we may estimate the distance to lightning by counting the seconds between the light flash and the following thunder noise, and multiplying this figure by 330 m.

Similarly, we can measure the distances to different satellites by looking at their transmitted signals. Once we know these distances and the positions of the satellites, we can work out our own position by computing the cross-section of the spheres around the satellites with the respective distances as radii. This idea is illustrated in Fig. 9.2. Principally, the cross-section of three spheres would determine the user position (up to an ambiguity that can usually be resolved). Since the difference in the time base between the satellites and the user is unknown, we get one further degree of uncertainty, which adds one more satellites to the number needed. Four satellites in line of sight means, that a complete three-dimensional position fix (called 3D-fix) can be obtained. If more satellites are in view, the receiver may use this information to estimate the position more accurately. On the other hand, if only three satellites are observable, the receiver can assume its altitude to be on the surface of the earth. Applying such an assumption (2D-fix), the receiver can still compute longitude and latitude. Obviously, a 2D-fix will not work in an airplane. On the other hand, a plane is rarely in the position to only view three satellites.

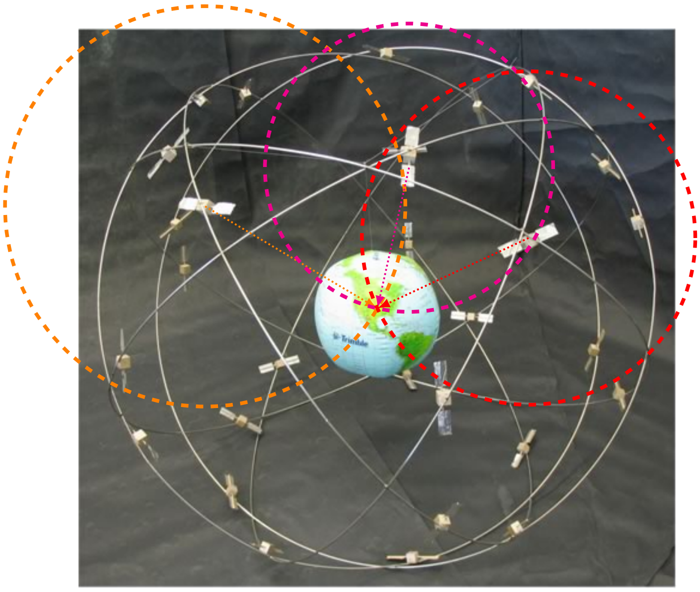


Figure 9.2 Cross-section of spheres around satellites.

9.3 System Overview

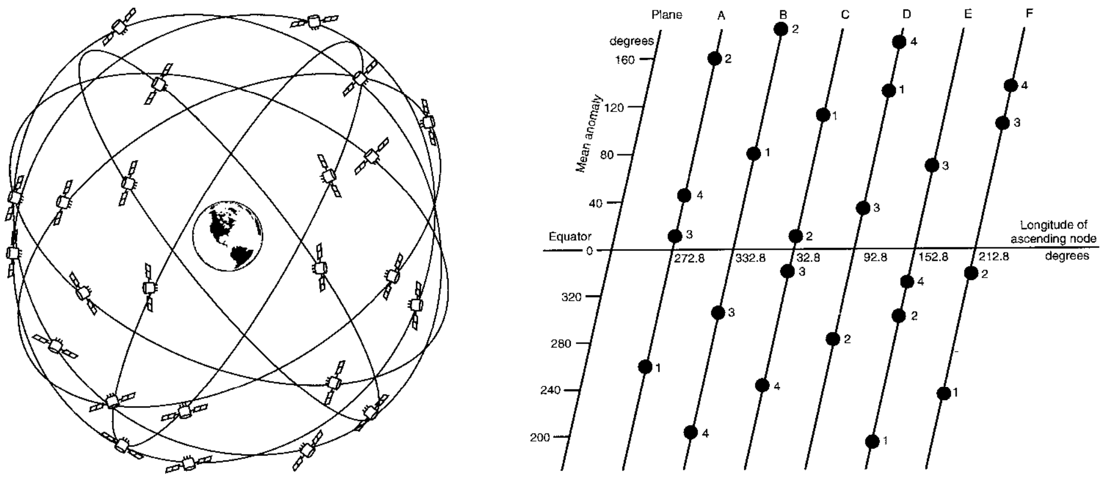
GPS consists of three system segments:

1. the satellite constellation,
2. the control network,
3. the user equipment.

Each of these components shall be discussed in the following.

9.3.1 Satellite Constellation

The GPS satellites (currently around 30 active) circulate the earth in six different orbits, see Fig. 9.3. The orbits are so-called MEO (medium earth orbit) and are at a distance of 20’192 km above the earth’s surface. At this altitude, the time for one full evolution around the earth is 12 h. Strictly speaking, the orbital period of a SV is 11 h 57 min 57.26 s, which corresponds exactly to one-half of a sidereal day. A sidereal day is the time it takes the earth to revolve exactly once around itself. It takes slightly more (namely a solar day or 24 h) to reach the same angle towards the sun, since the earth moves around the sun at the same time.



(a) GPS SV constellation

(b) GPS SVs distributed on 6 tilted planes

Figure 9.3 GPS SV constellation and planes. Source: [32].

The orbits are distributed according to Fig. 9.3 and lie on planes with a difference of 60 ° with respect to each other and an inclination angle of 55 ° relative to the equator. There are a total of 6 orbits, numbered from ’A’ through ’F’, each with orbital positions 1 to 4 (in some cases 5 or even 6).

Other interesting facts about the satellites are their weight of around 845 kg, the size of their solar panel being 7.25 m^2, and their life expectancy of 7.5 years, although some satellites have reached more than twice this life expectancy. The age of a satellite can also be seen by its type designation. Valid designations are listed in Table 9.1.

Type designation	Description
I	Initial and validative type
II	Operational SVs
IIA	Advanced
IIR	Replenishment
IIR-M	Replenishment (with new M code)
IIF	Follow-on
IIIA	With new L1C code

Table 9.1 SV type designations.

9.3.2 Control Segment

The control segment, officially called *Operational Control Segment* (OCS), is responsible for maintaining the satellites and their proper functioning. This includes the control of the orbital positions of the SVs and monitoring their condition (called *health*). The OCS usually updates ephemeris, almanac, and other data messages in the navigation message, also called the payload of the satellite signal, once per day.

The *master control station* (MCS) is located at Falcon Air Force Base, Colorado Springs. A total of five monitoring stations can be found throughout the world. The collected data from these monitoring stations are sent to the MCS, which, in turn, transfers them up to the SVs via S-band once they are in sight.

The most important tasks of the control segment are:

- observation of SV movement and computation of ephemerides,
- surveillance of SV clocks,
- time synchronization of SVs,
- transmission of navigation data to the satellite.

9.3.3 User Segment

The user segment is the least specified segment. Whereas the former two segments are strictly regulated and operated by one body, the user segment is not restricted in any form. Whoever wants to build equipment is free to do so, at his own specification. Two different services are provided by GPS: the *precise positioning service* (PPS) and the *standard positioning service* (SPS). The former is intended for military and governmental use and can only be accessed through a several-week-long spreading code, which may be obtained for special civilian use from the DOD. Further cryptographic techniques prevent spoofing—the impersonation by an adversary in an attempt to fool the system. This feature is called anti-spoofing (AS) and applies only to the PPS. The SPS, on the other hand, can be used by everybody, free of charge. Its precision is limited, but accurate enough for most user purposes, particularly since the DOD switched off SA.

Whereas twenty, or even fewer, years ago, GPS was used in expensive equipment only, it can be an add-on in many devices for only a few dollars these days. Many low-cost manufactures have seized the opportunity to enter the market with small form-factor modules that take a GPS signal from an antenna and convert it into a serial data stream containing the position, which is then offered to some higher controlling system.

9.4 Signal Characteristics

In the following, we will mainly focus on the signal characteristics of the C/A signal, which is the signal chosen to provide the SPS.

9.4.1 Overview

The purpose of the signal sent out by the SVs is two-fold. On the one hand, it is used to estimate the distance from the user to the satellite, and, on the other, data (navigation message) is transmitted over this path.

The data rate of this signal is very low, only 50 bits/s are transmitted containing the following information:

- SV time and synchronization signals,
- ephemerides (precise orbit data of the transmitting SV),
- almanac (coarse orbit data of all SVs),
- time correction signals,
- ionospheric data,
- health information of SVs.

The transmission of the whole information takes 12.5 minutes, and is repeated after that period. All GPS satellites use a set of the same radio frequencies in the L-band. The use of different frequencies allows a user to estimate ionospheric delay, which is different for different frequencies. For low-end consumer devices, only one frequency is used. This RF carrier is called L1 and is located at 1.57542 GHz. Although the P(Y) code is also transmitted on this frequency, we will ignore its existence due to its orthogonality on the carrier and concentrate on the C/A code, which is the signal used for most commercial GPS receivers. Since all SVs transmit on the same frequency at the same time, the different signals must be distinguished using a unique CDMA signature.

All frequencies generated inside the SV are derived from very accurate atomic time standards. The stability of the time standard is better than $2 \cdot 10^{-13}$ [49]. The fundamental reference frequency[1] is 10.23 MHz, from which all necessary frequencies can be derived, see Fig. 9.4. If we multiply the fundamental frequency

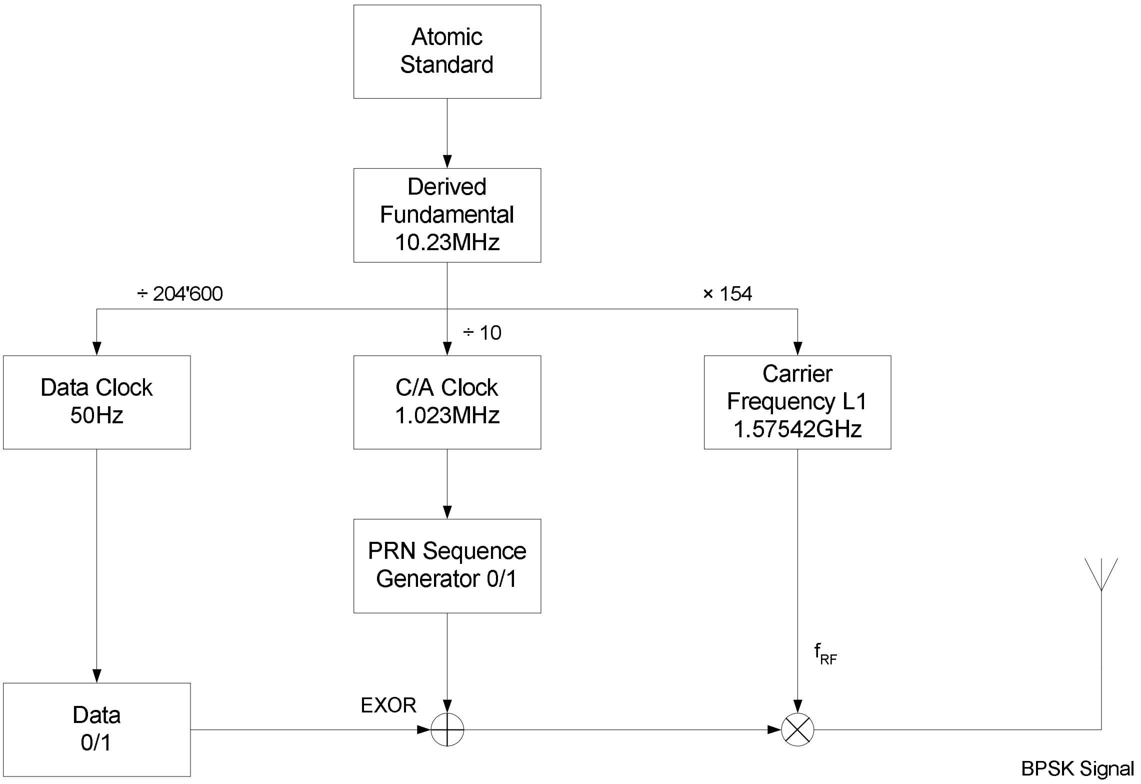


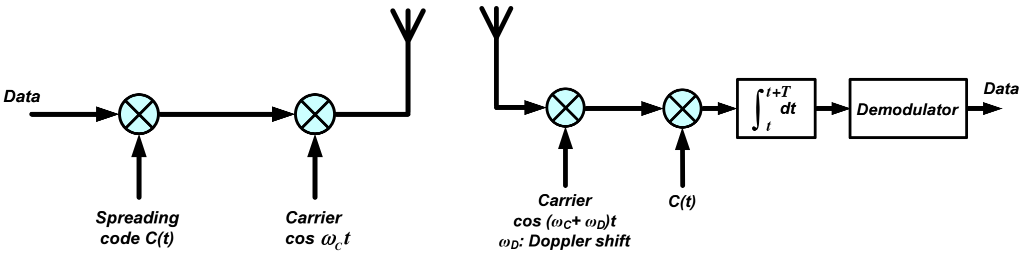
Figure 9.4 Block diagram of clock derivation and signal generation in a SV.

[1] In order to account for relativistic effects of a different time base at high speed and at lower gravitational exposure, the exact reference frequency is in fact 10.229999995433 MHz.

by 154, we get the empty RF carrier. This carrier is now modulated using a spread-spectrum sequence, resulting in a BPSK modulation scheme. This modulation scheme is therefore often referred to as direct sequence spread spectrum (DSSS). The sequence used in the C/A case is a PRN of length 1023 chips at a chipping rate of 1.023 MHz (one tenth of the fundamental frequency). The whole PRN sequence thus lasts exactly 1 ms and is repeated thereafter. Twenty such PRN sequences bear the information of 1 bit. The PRN sequence is simply multiplied by ±1 to carry the bit stream.

9.4.2 Spread-Spectrum Operation

GPS is based on spread-spectrum techniques, which allow one to receive signals buried deep in thermal noise. The basic configuration of transmitters and receivers operating in spread-spectrum techniques can be seen in Fig. 9.5. The resulting signals can be viewed in Fig. 9.6.



(a) Transmitter (b) Receiver

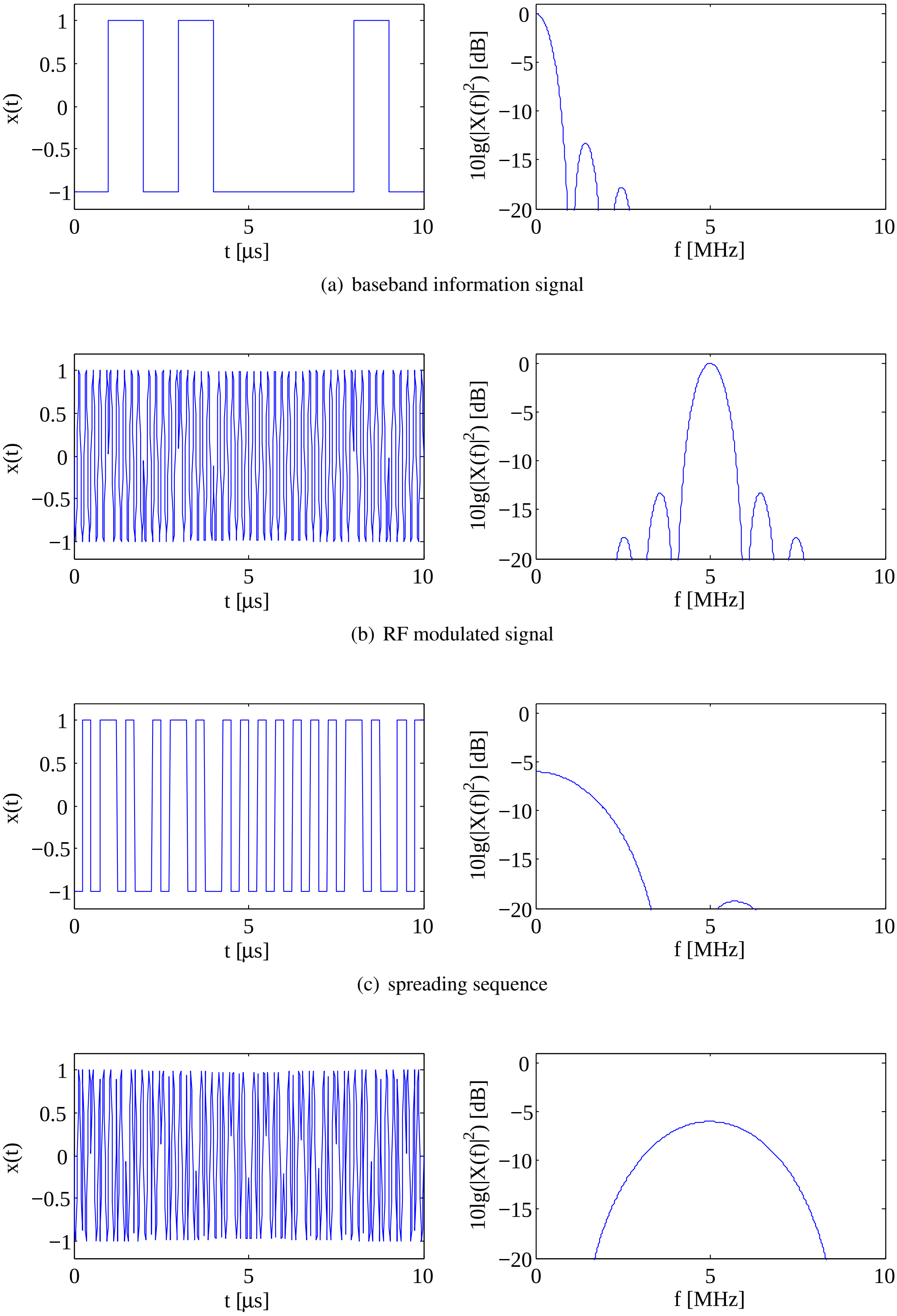
Figure 9.5 Transmitter and receiver for spread-spectrum signals.

9.4.3 Generation of the PRN Sequences

Each of the SV transmits a unique PRN sequence, which should have low autocorrelation sidepeaks and low crosscorrelation with respect to each other. For the C/A codes, the choice of PRN sequences is the use of so-called Gold codes. The generation of the Gold codes used for GPS are shown in Fig. 9.7.

The fundamental building blocks are two 10-bit shift registers, G1 and G2, which each generate a maximum-length pseudonoise sequence with a length of $2^{10} - 1 = 1023$. Such a maximum-length sequence is defined by the feedback taps, of which there are only two in G1, and six in G2. The first register is loaded with an XORed combination of the feedback taps. These feedback taps are fixed and stay the same for every SV. The resulting sequence from such maximum-length shift register is always the same, regardless of the initial state except for the all-zero state, which will generate a zero sequence and must therefore be avoided. The time delay of the sequence is, however, determined by the starting state. For reproducible delays, the states are all set to ones, i.e., initially, both shift registers are stuffed with ones.

To produce a Gold code, the outputs of the two m-sequences G1 and G2 are simply XORed. The Gold code for every SV is made distinguishable using a different delay of the second code G2. Alternatively, two different taps of G2 may be XORed, producing the same result. The complete G2 delay or, alternatively, G2 tap selection and the first 10 chips of every Gold code sequence are shown in Table 9.2.



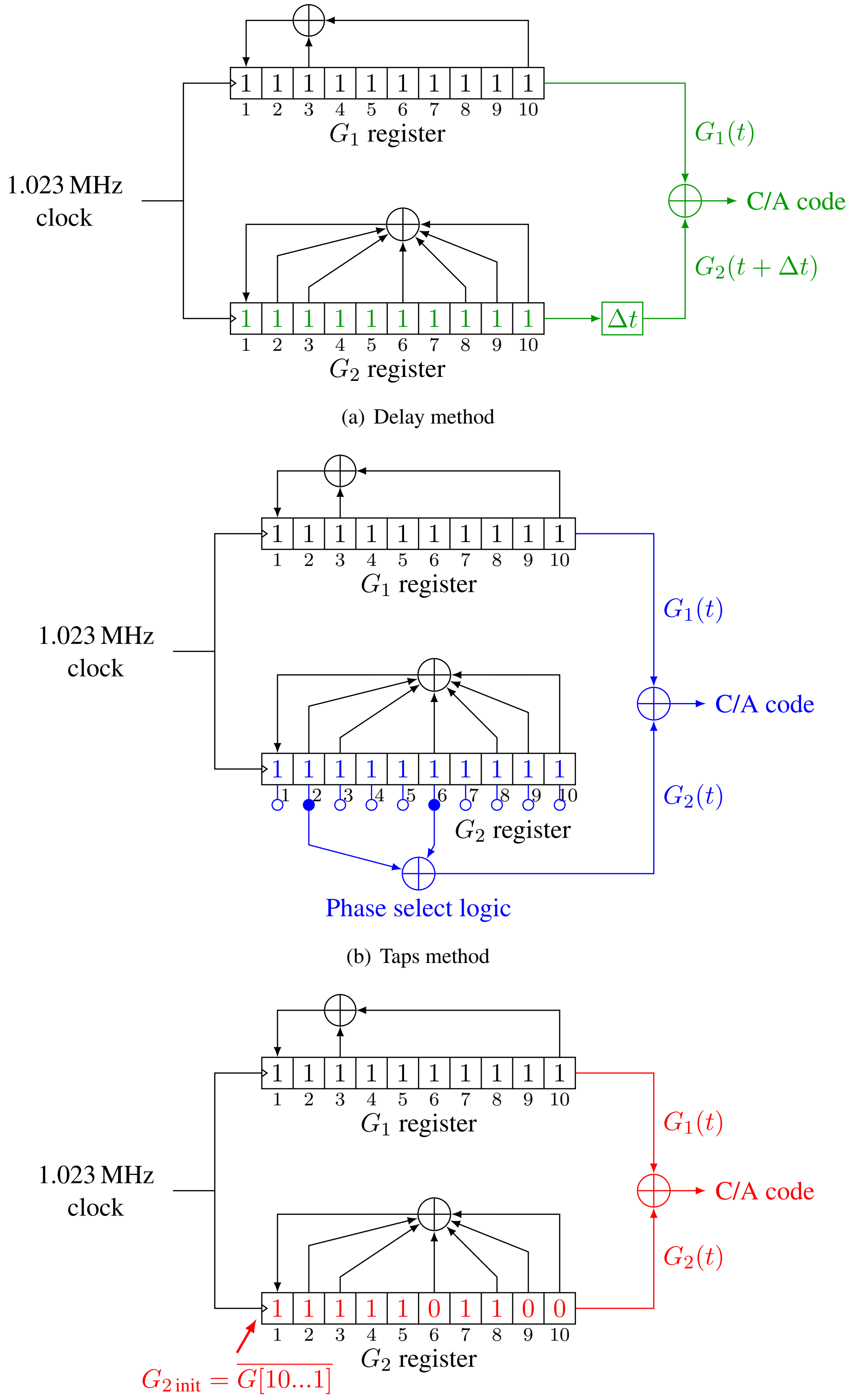
(a) baseband information signal

(b) RF modulated signal

(c) spreading sequence

(d) RF spread signal

Figure 9.6 Time-domain (left) and frequency-domain signals (right, in dB) at different stages of a spread-spectrum system.



(a) Delay method

(b) Taps method

(c) Init method

Figure 9.7 Gold Code generation

SV PRN	G2 delay	G2 taps	First 10 C/A chips	First 10 C/A chips (octal)
1	5	$2\oplus6$	1100100000	1440
2	6	$3\oplus7$	1110010000	1620
3	7	$4\oplus8$	1111001000	1710
4	8	$5\oplus9$	1111100100	1744
5	17	$1\oplus9$	1001011011	1133
6	18	$2\oplus10$	1100101101	1455
7	139	$1\oplus8$	1001011001	1131
8	140	$2\oplus9$	1100101100	1454
9	141	$3\oplus10$	1110010110	1626
10	251	$2\oplus3$	1101000100	1504
11	252	$3\oplus4$	1110100010	1642
12	254	$5\oplus6$	1111101000	1750
13	255	$6\oplus7$	1111110100	1764
14	256	$7\oplus8$	1111111010	1772
15	257	$8\oplus9$	1111111101	1775
16	258	$9\oplus10$	1111111110	1776
17	469	$1\oplus4$	1001101110	1156
18	470	$2\oplus5$	1100110111	1467
19	471	$3\oplus6$	1110011011	1633
20	472	$4\oplus7$	1111001101	1715
21	473	$5\oplus8$	1111100110	1746
22	474	$6\oplus9$	1111110011	1763
23	509	$1\oplus3$	1000110011	1063
24	512	$4\oplus6$	1111000110	1706
25	513	$5\oplus7$	1111100011	1743
26	514	$6\oplus8$	1111110001	1761
27	515	$7\oplus9$	1111111000	1770
28	516	$8\oplus10$	1111111100	1774
29	859	$1\oplus6$	1001010111	1127
30	860	$2\oplus7$	1100101011	1453
31	861	$3\oplus8$	1110010101	1625
32	862	$4\oplus9$	1111001010	1712

Table 9.2 Characterization of Gold sequences used for C/A codes.

9.4.4 Power Levels

The maximum distance between the satellites and the user is 25785 km (when the SV is at the horizon). The free-space loss can then be calculated as

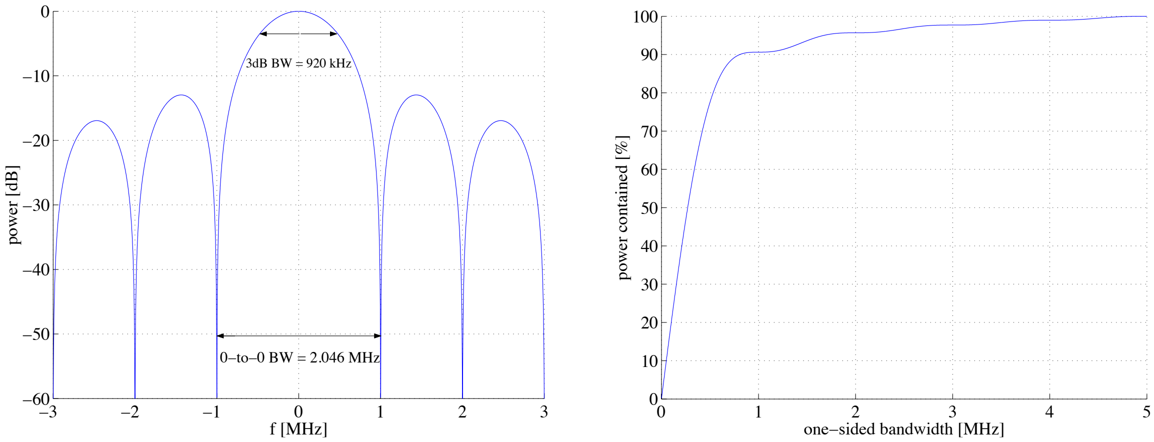
$$L\,[\mathrm{dB}] = -147.6 + 20\log_{10}(1.57542 \cdot 10^9) + 20\log_{10}(2.5785 \cdot 10^7) = 184.6\,\mathrm{dB}. \tag{9.1}$$

The transmitter antenna gain is specified at 13.4 dB. The transmitted power is roughly 22 W or 43.4 dBm. Polarization mismatch (around 3 dB) is compensated by the receiver antenna gain. An additional attenuation in the atmosphere of around 2 dB makes the receiver power approximately

$$P_{\mathrm{rx}} = 43.4 + 13.4 - 184.6 - 2 \approx -130\,\mathrm{dBm}. \tag{9.2}$$

9.4.5 Signal Spectrum

The bandwidth of a GPS C/A signal is very often given as roughly 2 MHz. Strictly speaking, this refers to the null-to-null bandwidth, which is exactly 2.046 MHz, due to a chipping frequency of $f_{\text{chip}} = 1.023\,\text{MHz}$. Other definitions of bandwidth lead to other values. For ease of reference, the spectral contents of a C/A signal is shown in Fig. 9.8(a).



(a) Associated bandwidths

(b) Cumulative power density as a function of bandwidth included in the receiver filter.

Figure 9.8 Properties of the GPS C/A signal.

As can be seen from Fig. 9.8(b), roughly 90 % of the C/A signal are contained in the null-to-null bandwidth. This bandwidth is therefore often cited as the minimum required bandwidth for GPS reception. In the following section, simulation results shall provide further insight into the influence of the IF filter bandwidth to the signal reception parameters.

9.4.6 Noise

As mentioned in Chapter 5, the spectral noise power density is obtained through the system temperature T (absolute, in Kelvin) and the Boltzmann constant $k = 1.38 \cdot 10^{-23}$ J/K

$$N_0 = kT \qquad \text{(in dBm/Hz)}\,. \tag{9.3}$$

The spectral density is taken as N_0 for a one-sided spectrum (positive frequencies only) and as $N_0/2$ for a two-sided spectrum. The noise power will be $N_0 B$, with B denoting the bandwidth, in any case. In GPS, because most of the signal is contained in the null-to-null bandwidth of 2.046 MHz, we have a spreading gain of

$$G_{\text{spreading}} = 10\log_{10}(2.046 \cdot 10^6) = 63.1\,\text{dB-Hz}\,, \tag{9.4}$$

i.e., for -130 dBm signal and roughly -110 dBm noise, a -20 dB SNR at the receiver input transfers to a 43.1 dB-Hz *carrier-to-noise-density ratio* C/N_0.

Alternatively, we might get this number (or slightly more, since -110 dBm is in fact computed for a slightly wider band) by expressing N_0 in the logarithmic domain

$$N_0 = 10\log_{10}(kT) = -204\,\text{dBW/Hz} = -174\,\text{dBm/Hz}. \tag{9.5}$$

So we can directly relate the received signal power to this number to get

$$G_{\text{spreading}} \approx -130\,\text{dBm} - (-174\,\text{dBm/Hz}) = 44\,\text{dB-Hz} \ . \tag{9.6}$$

9.4.7 Correlation Properties of PRN Sequences

Maximum-Length Sequences (m-Sequences)

Maximum-length sequences have very interesting autocorrelation properties: they have one peak for exact alignment, and the same low level for misalignment of one or more chips. Since the sequences always consist of an odd number of chips, the autocorrelation functions can never be zero. If the autocorrelation function is taken for a relative shift of below ±1 chip (just a slight misadjustment), we can observe the correlation triangle. In summary, for a code period of $2^n - 1$, for perfect alignment we have a peak of $2^n - 1$, and for misalignment outside one chip an autocorrelation value of -1. For an m-sequence of length 10, the autocorrelation function is shown in Fig. 9.9.

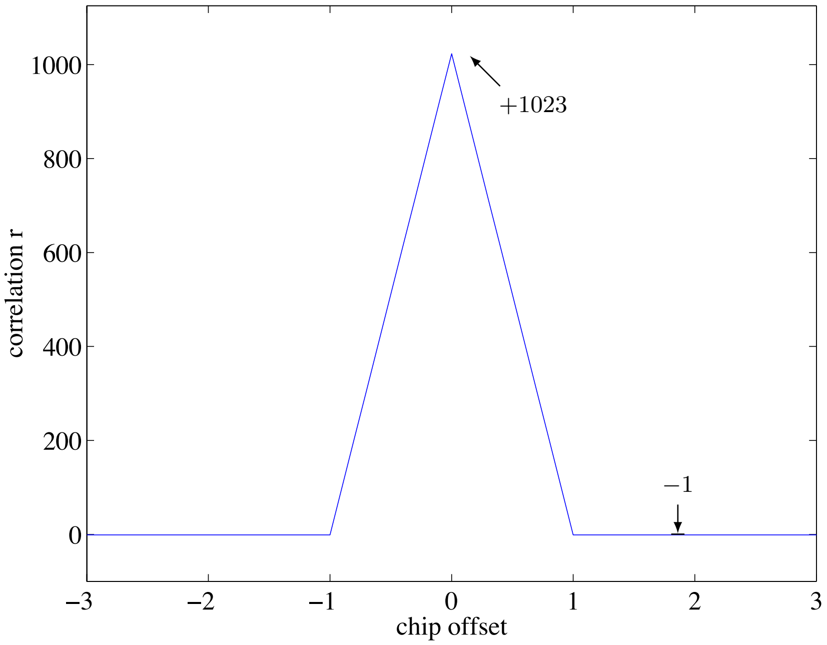
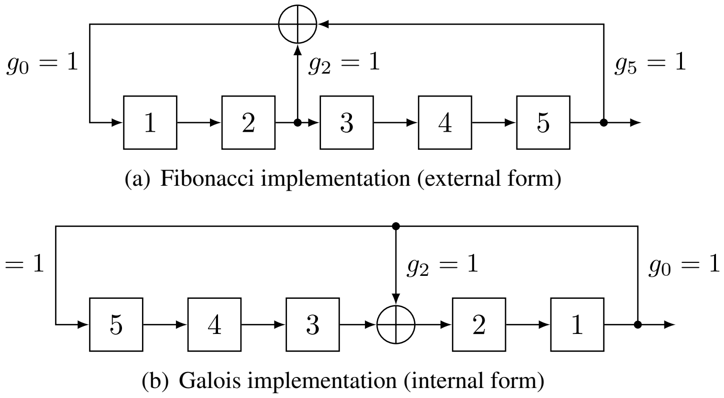


Figure 9.9 Autocorrelation function of an m-sequence.

Similar to different but equivalent architectures of IIR filters, there are two well-known forms of linear feedback shift registers: the Fibonacci form and the Galois form. They are shown in Fig. 9.10. The description of the feedback takes different forms in the literature. The feedback used in Fig. 9.10 can be written as $[5, 2]$ or as a polynomial $x^5 + x^2 + 1$. Note that some authors reverse the order of the feedback shift register, so that the feedback polynomial needs a different interpretation.

In order to generate a maximum-length sequence, the feedback polynomial needs to be irreducible. Such polynomials are given in Table A.10. Note that the number of feedback taps is always even.



(a) Fibonacci implementation (external form)

(b) Galois implementation (internal form)

Figure 9.10 Example of equivalent forms corresponding to a feedback polynomial of $x^5 + x^2 + 1$

Gold Sequences

Unfortunately, there are not enough m-sequences of that length available. Hence, in GPS, Gold codes are used instead. They are named after their discoverer, mathematician and coding researcher Robert Gold[2], who first published results concerning crosscorrelation properties of new codes, built using combinations of m-sequences, and their applications as spreading functions. [25, 26]

Compared to maximum-length sequences, Gold codes have only limited autocorrelation and crosscorrelation capabilities. Apart from the main peak (which is 1023 for this length of code), the correlation functions show rather high side-peaks of $+63$ and -65, respectively, which results in a dynamic range of

$$20\log_{10}\frac{1023}{65} = 23.9\,\text{dB}. \tag{9.7}$$

Autocorrelation and crosscorrelation functions are shown in Fig. 9.11.

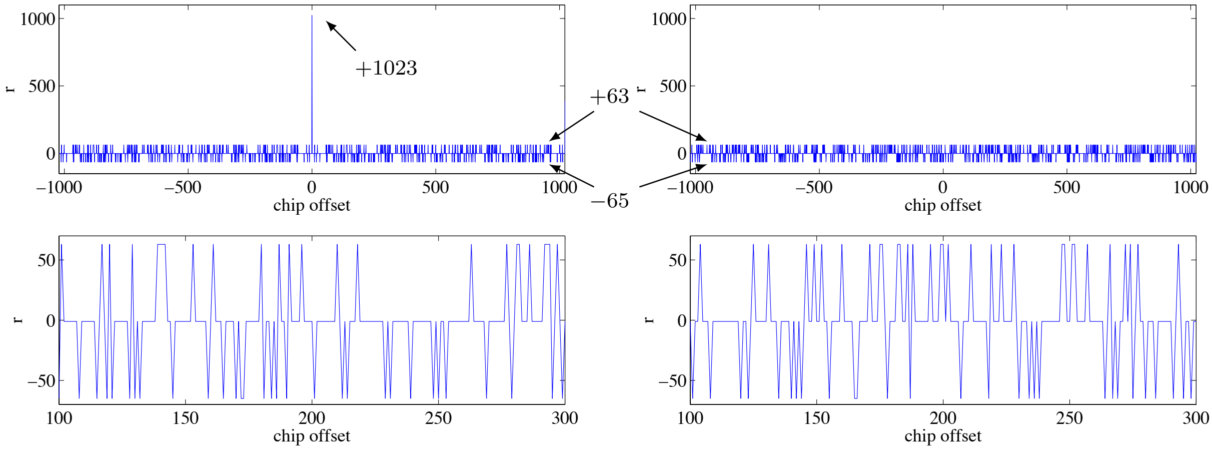
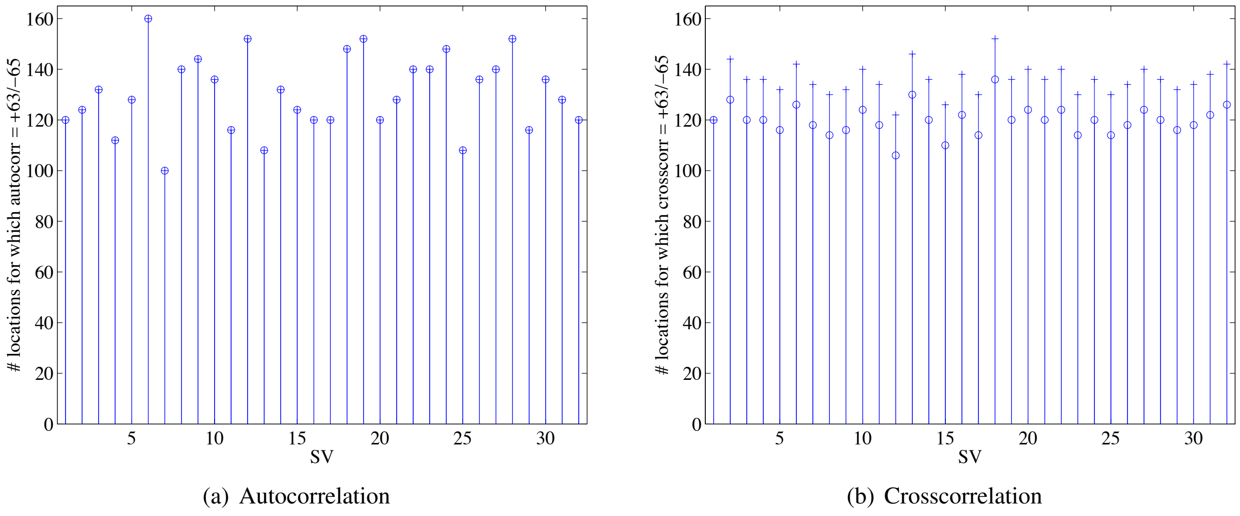


Figure 9.11 Typical autocorrelation function (left) and crosscorrelation function (right) of the Gold codes with two different zoom sizes.

[2] Dr. Robert Gold is still active in research and development in the field of tactical and commercial communications with his company Robert Gold Comm Systems, RGCS, see www.rgcsystems.com.



(a) Autocorrelation

(b) Crosscorrelation

Figure 9.12 Peak histograms of the processed Gold sequences; $\circ = -65$, $+ = +63$.

Since false locking to sidepeaks of the correlation function can occur, it is of interest to know the frequency of such side-peaks, which is an inherent feature of the Gold codes. The auto- as well as the crosscorrelation function can take three different values: -1, -65, and $+63$. The relative occurrence can be seen from Figs. 9.12(a) and Figs. 9.12(b). Interesting facts are that the numbers of -65 and $+63$ are the same for any autocorrelation but different to the respective numbers for a different sequence. This is not the case for crosscorrelation functions.

9.4.8 Evaluation and Simulation of C/N_0

One of the most important figure to characterize a GPS receiver is the signal-to-noise power density, C/N_0, which is measured in dB-Hz. It consists of the signal power after the despreading (integrator output) and the thermal noise density. In the literature, the thermal noise density is usually designated by the double-sided noise density $N_0/2 = kT/2$, where $k = 1.38 \cdot 10^{-23}$ is the Boltzmann constant and T is the system temperature. The reason for this is that by integrating over the frequencies of interest, we get for the total noise power

$$P_{\mathrm{n}} = \int_{-\infty}^{\infty} \frac{N_0}{2} \, |H(f)|^2 \, df = N_0 B. \tag{9.8}$$

where $H(f)$ is the transfer function of the receiver filter. For GPS low-end receivers usually a bandwidth of around 2 to 2.4 MHz is assumed. This results in a noise power of

$$P_{\mathrm{n}} = N_0 B = kTB = -110\,\mathrm{dBm}. \tag{9.9}$$

For the simulation it is important to take the sampling rate into consideration, otherwise the noise generator distributes the assigned noise power over the whole Nyquist bandwidth. Proper noise generation can be done in two ways: First, by generating the noise samples, sampled once per chip, and then upsample them to the sampling rate of choice, whilst keeping constant the total noise power, or second, by filtering a generated noise signal down to its appropriate bandwidth. Fig. 9.13 shows the noise bandwidth usually considered.

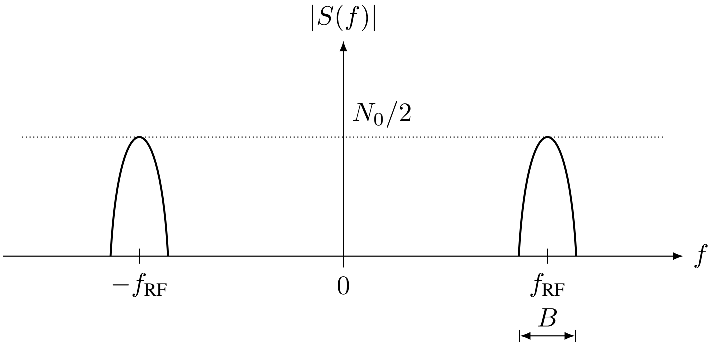


Figure 9.13 Noise density in bandwidth of interest.

9.4.9 Message payload

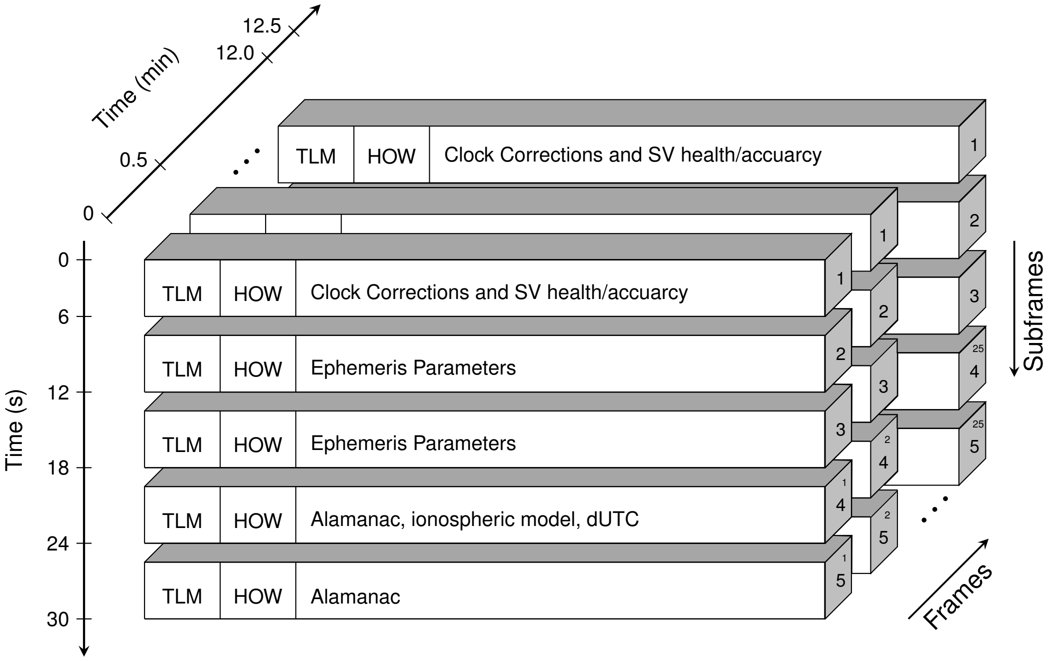


Figure 9.14 Subframe structure of a GPS navigation message [45].

In GPS there is a message component, called the navigation message, buried in each C/A-signal. With a moderate bitrate of 50 bits/s only, 20 consecutive copies of the same polarity of PRN sequences indicate a 0 or a 1 of the bitstream. The structure of this bitstream (which is individual to every SV) is shown in Fig. 9.14. In total we have a 1500-bit-long frame, corresponding to 30 s duration, divided into 5 subframes. Thus, every subframe contains of 300 bits, and every subframe contains 10 words of length 30 bits each. The first three subframes are repeated in each frame containing the clock corrections, satellite status and ephemeris parameters. Ephemeris data describe in an exact manner the orbit of the SV that transmits this data (such that we can navigate with it). In subframes 4 and 5 the almanac is transmitted. Almanac data describe in a rough manner the orbits of all SVs. All subframes begin with two special words, the telemetry (TLM) and handover word (HOW). The TLM contains an 8-bit preamble to find the start of the subframe.

9.5 Receiver Architecture

A principle receiver circuit diagram is shown in Fig. 9.15. As examples, several generations of commercial implementations are depicted in Figs. 9.16(a) to 9.17(c).

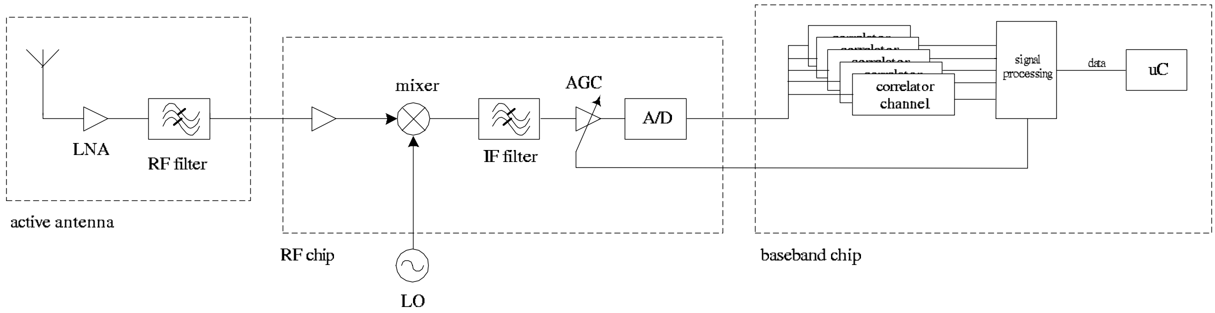
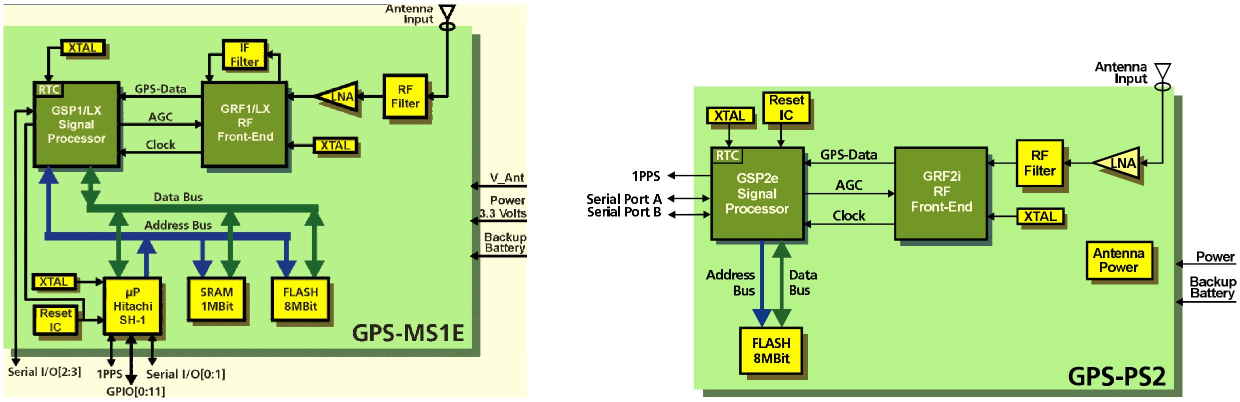


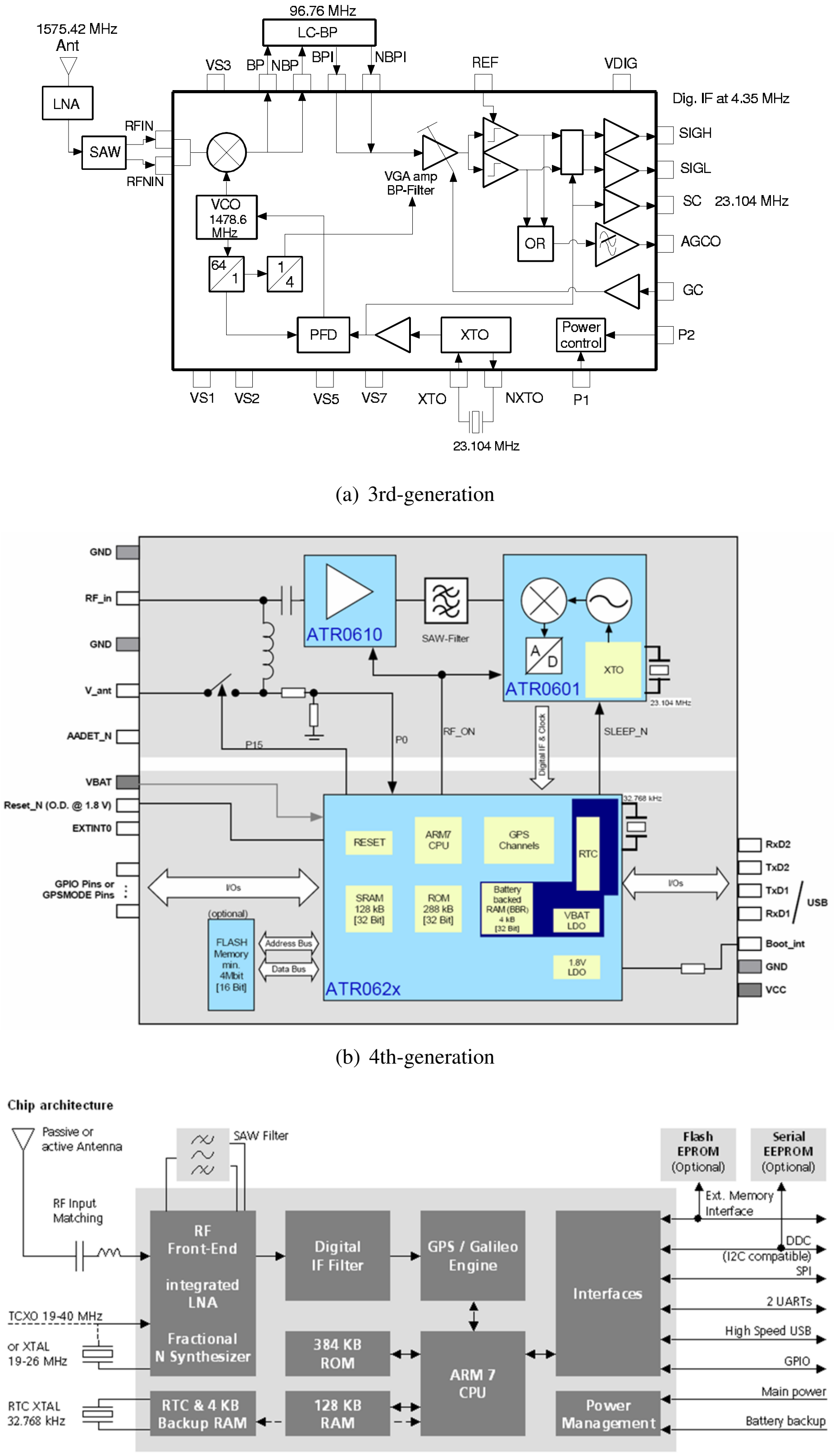
Figure 9.15 Typical architecture of handheld GPS receiver.



(a) 1st-generation

(b) 2nd-generation

Figure 9.16 1st- and 2nd-generation GPS receiver by u-blox (Switzerland).



(a) 3rd-generation

(b) 4th-generation

(c) 5th-generation

Figure 9.17 3rd-, 4th- and 5th-generation GPS receiver by u-blox (Switzerland) based on proprietary chip set.

9.5.1 Antenna

The RF signal coming from an SV is right-hand circularly polarized (RHCP). The receiver antenna should ideally accommodate for this polarization mode, although the use of horizontal polarized antennas is possible with an additional loss of 3 dB. Besides, right-hand polarized antennas suppress the first reflection in a multipath environment, since any reflection changes the right-hand polarization into a left-hand one. For the same reason, GPS antennas should show low back lobes, since reflection often originates from the ground. In general, a GPS antenna should show an omnidirectional pattern with respect to the northing angle and cover the hemisphere above the ground as equally as possible (hemispherical gain coverage) whilst rejecting signals from below the horizon.

The bandwidth of the antenna required depends on the kind of GPS signals used. If only L1 is used (1.57542 GHz), low bandwidth is required. If L2 (1.2276 GHz) needs to be received, too, plus possibly the Russian GPS equivalent called GLONASS (operating at frequencies of 1.246 to 1.257 GHz and 1.602 to 1.616 GHz), either wide bandwidth or several narrow bands are required from the antenna.

For hand-held applications, the use of small helical antennas or microstrip patch antennas is established. An example of a commercial active antenna is shown in Fig. 9.18. A data sheet excerpt is given by Table 9.3.

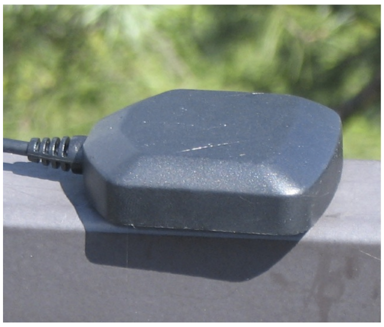


Figure 9.18 Active GPS antenna with housing. Dimensions: 46 x 39 x 12,5 mm. Usually a strong magnet at the bottom helps to attach such an antenna to a car-roof top or a similar metall surface.

Such an antenna needs a power supply for its built-in LNA, usually provided by some bias-T network as shown in Fig. 9.19.

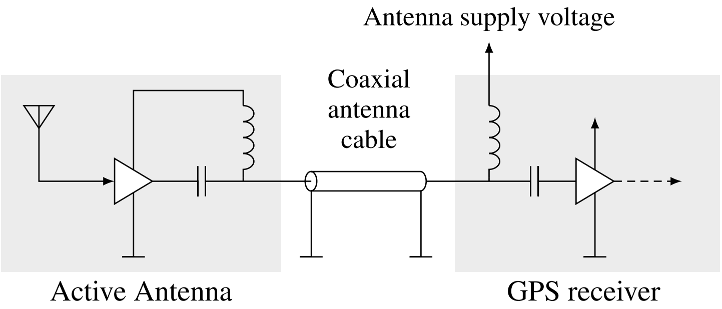


Figure 9.19 Active antenna setup.

Mechanical characteristics	
Dimensions ($l \times w \times h$)	46 × 39 × 12.5 mm
Weight (without cable)	35 g
Cable length	6 m

Electric characteristics	
Voltage	5 V ± 0.5 V
Current dissipation	20 mA typ., 25 mA max.
Temperature range	-40 to +100 °C
Humidity	40 to 95 %

RF characteristics	
Frequency range	L1 (1575.42 ± 1.023 MHz)
Impedance	50 Ω
VSWR	1.5 typ. / 2.5 max.
3dB bandwidth	max. 45 MHz
Gain (including cable)	24 dBi ± 5 dBi

Filter characteristics	
Attenuation (typ.)	40 dB (at ± 95 MHz)

Antenna characteristics	
Gain (zenith)	3.0 dBi typ.
Gain (horizon)	-10 dBi min.
Axial ratio	3 dB typ.

Table 9.3 Characteristics of the NAiS GPS active antenna.

9.5.2 LNA

Very often, the first LNA is directly integrated into a common housing with the anntena. There are two main reasons for doing so; first, the gain should precede the attenuation of a long cable; second, such a cable will pick up radiation easily, so a higher signal level is preferred.

At this point, the level is high enough to pass the signal through a possibly lossy bandpass, whose main objective is the suppression of the image frequency. Further gain stages may follow.

9.5.3 Down-Conversion

The signal is then downconverted to an IF that must be high enough to take into account the bandwidth of the GPS signal as well as make the job of the image filter not too hard. Alternatively, an I/Q mixer can downconvert the signal to a much lower IF, but then two ADCs are needed for in-phase and quadrature components. The frequency accuracy of the LO is essential in the determination of the search space. The Doppler range of the satellite accounts for about 4 ppm of frequency accuracy. If cheap quartz crystals are used, the frequency uncertainty might be around 20 ppm, totalling in around 24 ppm or around 38 kHz of deviation from the assumed RF. The acquisition strategy has to take this into consideration.

9.5.4 AGC

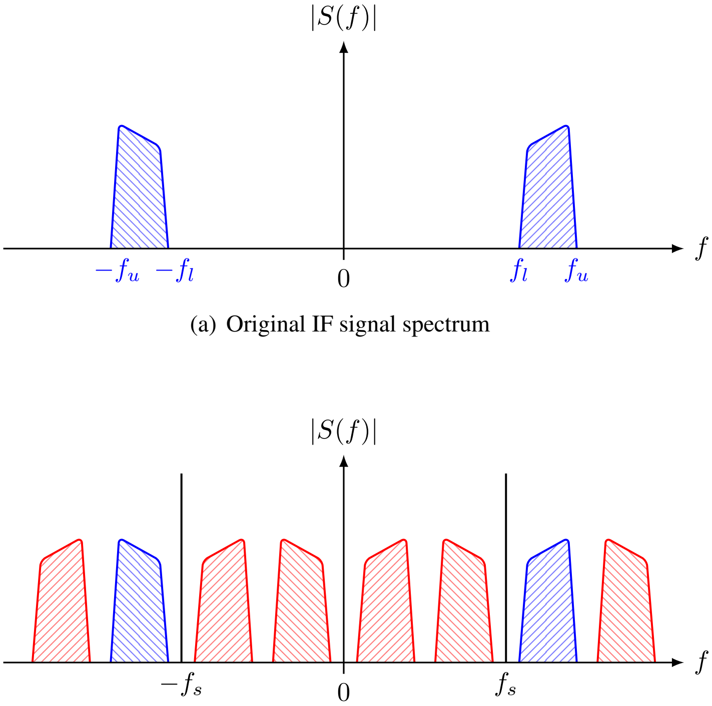
In normal wireless radio equipment, the AGC serves the purpose of adapting the system, in particular the following ADC stage, to changing levels of the signal due to fading and shadowing. In GPS, the signal of interest is deeply buried in the thermal noise, so changing levels of the GPS signal do not affect the level seen by the ADC. However, in many GPS receivers, an AGC stage is provided to account for different gain in previous, external stages, such as active antennas, which are unknown or could potentially change through simple unplugging. The requirements to the AGC are thus very relaxed, no fast tracking capability is needed.

9.5.5 ADC

In the absence of jamming signals, the analog-to-digital converter does not need to provide a large dynamic range as is often the case in wireless receivers. Very often, 1-2 bits are used in cheap GPS receivers. The sampling rate, however, needs to accomodate a rather high signal bandwidth. If a zero-IF signal is sampled with a lowpass bandwidth of around 1 MHz, the sampling rate of the ADC must be above 2 MS/s. If an IF is used, the sampling rate is usually much higher, except for the cases where subsampling is employed.

Subsampling

Subsampling is the process of sampling a signal at a lower frequency than twice the highest frequency occuring in the signal of interest. The purpose of subsampling is, on the one hand, to keep the sampling rate low and, on the other hand, to mix down the signal from the first IF to the second IF, instead of using another mixer. This apparent violation of the Nyquist's sampling theorem produces aliases, the lowest of which is usually taken for further processing. In order to keep the different aliases apart, the Nyquist theorem with respect to the bandwidth must be followed, i.e., the sampling frequency, albeit lower than twice the highest frequency, must be higher than twice the *lowpass bandwidth* of the signal sampled. Besides, the sampling frequency must be chosen carefully with respect to the IF, in order to allow easy postprocessing, see also Fig. 9.20.

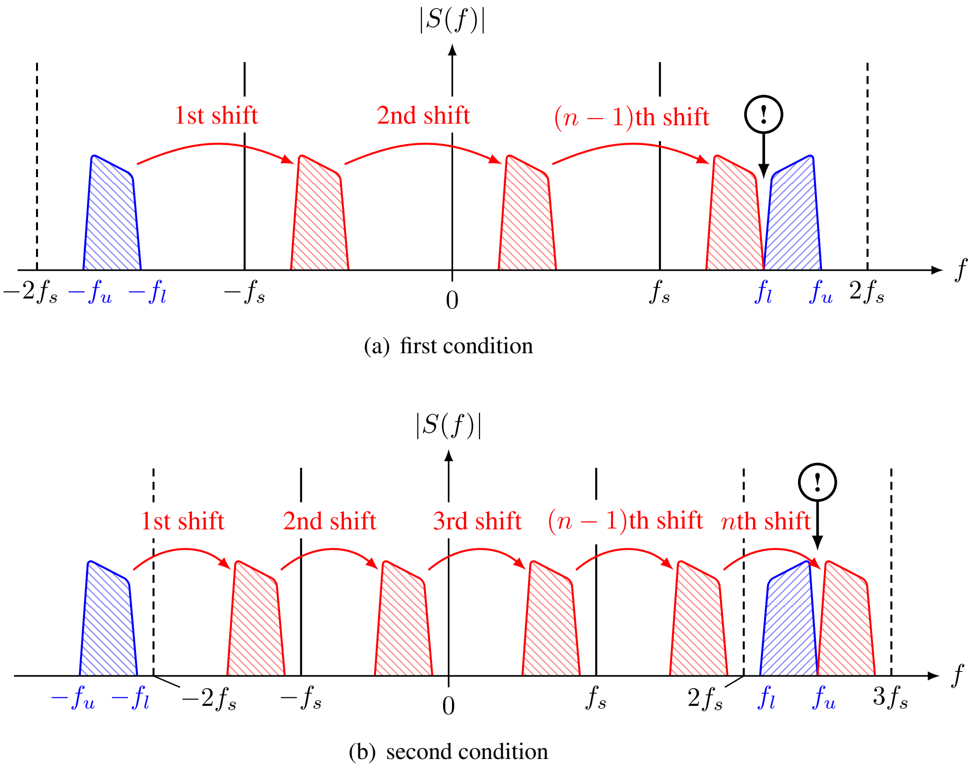


(a) Original IF signal spectrum

(b) Subsampled signal spectrum

Figure 9.20 Consequences in the spectrum resulting from subsampling.

In the following we want to derive the conditions of obtaining overlapping-free aliases of the spectrum limited by f_l and f_u, with f_l and f_u being the lower and the upper frequency of the bandpass signal,



(a) first condition

(b) second condition

Figure 9.21 The two subsampling conditions.

respectively. The integer n is denoted the subsampling factor. The original spectrum (including positive and negative frequencies for real signals) is repeated with shifts of

$$\Delta f = k f_s, \tag{9.10}$$

where k is any negative or positive integer. By starting out with a high sampling rate f_s (no spectra are yet overlapping), lowering f_s produces a rake of spectra that starts overlapping with the existing spectrum. Fig. 9.21 explains how we have to shift the aliased spectrum between other aliases. In other words, we must shift our spectra far enough but not too far. Essentially, we must avoid an overlapping situation when shifting the negative spectra $(n-1)$ times to the right with the original spectrum, hence

$$-f_l + (n-1)f_s \leq f_l \quad \Rightarrow \quad f_s \leq \frac{2f_l}{n-1}. \tag{9.11a}$$

On the other hand, also one of the replica is shifted by $n \cdot f_s$. This alias should be outside the original spectrum, thus

$$-f_u + nf_s \geq f_u \quad \Rightarrow \quad f_s \geq \frac{2f_u}{n}. \tag{9.11b}$$

Combining Eqs. (9.11a) and (9.11b) we get

$$\frac{2f_u}{n} \leq f_s \leq \frac{2f_l}{n-1}, \tag{9.12}$$

which is an important result for the choice of the sampling rate f_s. For any choice of the integer n, a necessary condition for f_s to exist is

$$\frac{2f_u}{n} \leq \frac{2f_l}{n-1}, \tag{9.13}$$

which can be reworked into

$$n \leq \frac{f_u}{f_u - f_l}. \tag{9.14}$$

Since n is a positive integer,

$$1 \leq n \leq \left\lfloor \frac{f_u}{f_u - f_l} \right\rfloor. \tag{9.15}$$

Denoting $B \overset{\triangle}{=} f_u - f_l$, we can easily see from Eq. (9.15) that the maximum value for n is $\lfloor f_u/B \rfloor$. Hence, due to Eq. (9.12), the sampling rate satisfies

$$\frac{2f_u}{\lfloor \frac{f_u}{B} \rfloor} \leq f_s. \tag{9.16}$$

A lower bound for the sampling rate can be constructed using

$$\frac{2f_u}{\frac{f_u}{B}} \leq \frac{2f_u}{\lfloor \frac{f_u}{B} \rfloor} \leq f_s. \tag{9.17}$$

This is, of course, the well-known Nyquist theorem in its bandpass formulation

$$2B \leq f_s. \tag{9.18}$$

Note that depending on the situation of the actual bandpass signal in the frequency domain, it may not be possible to find a valid f_s equal to twice the bandwidth B. This is only possible for so-called integer band positioning, i.e., if the lower band frequency f_l sits at an integer value of the bandwidth $B = f_u - f_l$.

In an earlier GPS receiver of the company *u-blox*, a signal at an IF of 47.65 MHz was subsampled at 38.194 MHz. The resulting aliased spectrum can be seen in Fig. 9.20. In this case, the alias at the frequency $f = 47.65 - 38.194 = 9.456$ MHz was used as the new digital IF. In terms of the rules given above, we can now verify this numerical example. For a bandwidth B of, say, 10 MHz we get

$$f_l = \text{IF} - \frac{B}{2} = 42.65\,\text{MHz}, \tag{9.19a}$$

$$f_u = \text{IF} + \frac{B}{2} = 52.65\,\text{MHz}. \tag{9.19b}$$

The possible ranges for f_s are obtained by letting n be an integer from 1 to 5:

$$n = 1: \quad 105.3\,\text{MHz} \leq f_s, \tag{9.20a}$$

$$n = 2: \quad 52.65\,\text{MHz} \leq f_s \leq 85.3\,\text{MHz}, \tag{9.20b}$$

$$n = 3: \quad 35.1\,\text{MHz} \leq f_s \leq 42.65\,\text{MHz}, \tag{9.20c}$$

$$n = 4: \quad 26.325\,\text{MHz} \leq f_s \leq 28.433\,\text{MHz}, \tag{9.20d}$$

$$n = 5: \quad 21.06\,\text{MHz} \leq f_s \leq 21.325\,\text{MHz}. \tag{9.20e}$$

We can see that the actual sampling frequency chosen lies in the interval for $n = 3$. In fact, choosing a higher n (equal to 4 or 5) would result in a still smaller sampling frequency. However, the requirements on the digital-IF filter get tougher.

9.5.6 Baseband Signal Processing

Most receivers have several channels, each of which can acquire and track a different satellite. The most important parts of the baseband signal processing stage are the correlators. They multiply the signal by the spreading sequence of the different SVs. Once the pseoudorange has been determined, the signals are passed to a general-purpose microcontroller, which then computes the navigation solution. The final position may then be displayed or conveyed to some higher instance, e.g., a wireless telephone.

9.6 GPS and the Theory of Relativity

Einstein's theories of relativity have already been around and studied for a quite a while, when GPS was conceived. As we will shortly see, both special and general relativity have an influence on the timing relation between the satellite clocks and the user clocks and must therefore be accounted for.

9.6.1 Special Relativity

Special relativity dictates that the time slows down in an object moving at speed v relative to the time of a standing observer. More precisely, the so-called time dilation can be written as

$$t_1 = t_E \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}. \tag{9.21}$$

The speed v on a circular orbit with radius R around the earth is given by the equilibrium of gravitational and centripetal forces

$$\frac{mv^2}{R} = \frac{GmM}{R^2}, \tag{9.22}$$

hence,

$$v = \sqrt{\frac{GM}{R}}. \tag{9.23}$$

Furthermore, we have a gravitational acceleration at the earth surface (distance R_E from the earth center) of

$$g = \frac{GM}{R_E^2}, \tag{9.24}$$

so Eq. (9.23) can be expressed as

$$v = \sqrt{g \frac{R_E^2}{R}}. \tag{9.25}$$

Inserting Eq. (9.25) into Eq. (9.21) gives us the time dilation as a function of the orbit radius of a satellite

$$t_1 = t_E \frac{1}{\sqrt{1 - \frac{g}{c^2} \cdot \frac{R_E^2}{R}}}. \tag{9.26}$$

The resulting curve is given in Fig. 9.22.

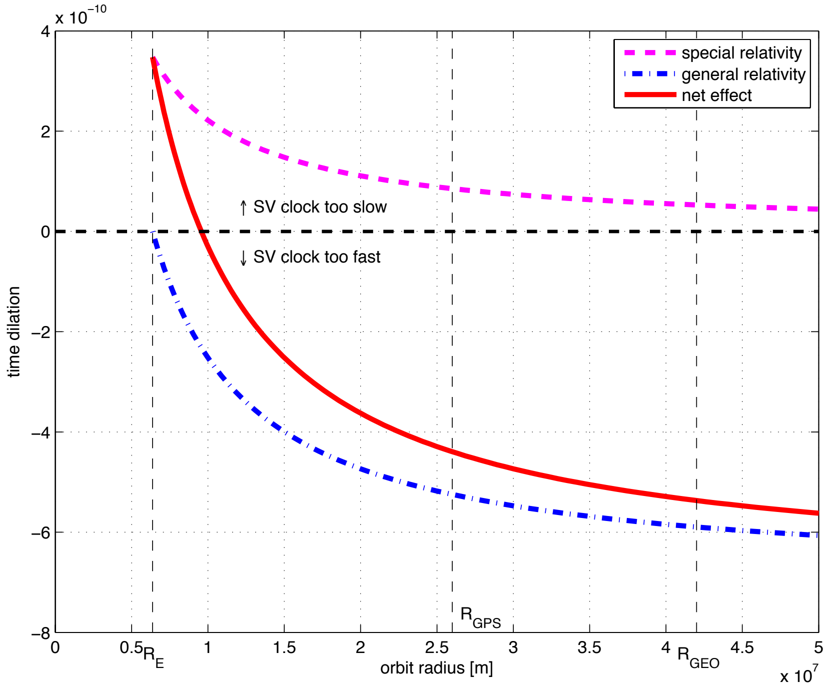


Figure 9.22 Relativistic effects as a function of orbit radius.

9.6.2 General Relativity

The general theory of relativity on the other hand predicts that any clock exposed to a gravitational field is slowed down compared to a point in zero gravitation. As we move away from the earth, the clock speeds up. Generally, the time dilation (compared to a point in zero gravitation) is

$$t_2 = t_0 \frac{1}{\sqrt{1 - \frac{2GM}{Rc^2}}} = t_0 \frac{1}{\sqrt{1 - \frac{g}{c^2} \cdot 2\frac{R_E^2}{R}}}, \tag{9.27}$$

where Eq. (9.24) was invoked. On the earth surface the time factor is calculated as

$$t_E = t_0 \frac{1}{\sqrt{1 - \frac{g}{c^2} \cdot 2R_E}}, \tag{9.28}$$

Hence relative to the earth clock, we can express the time in the satellite as

$$t_2 = t_E \frac{\sqrt{1 - \frac{g}{c^2} \cdot 2R_E}}{\sqrt{1 - \frac{g}{c^2} \cdot 2\frac{R_E^2}{R}}}, \tag{9.29}$$

for $R \geq R_E$. Again, Fig. 9.22 shows the time dilation as a function of the orbit radius.

9.6.3 Net Effect

In GPS, where the SVs are on an orbit of around 26’000 km, the time dilation due to the special relativity makes the satellite clock tick lag by 7.2 µs per day. On the other hand, due to the lower gravitational field (general relativity), the clock leads by 45.9 µs per day. As a net effect, the clock bias would increase by 38.7 µs every day. During that time the SV covers 15 cm and the earth rotates by 17 mm (at the equator). The effect must therefore be compensated for by choosing a slightly slower clock of 10.229999995433 MHz instead of 10.23 MHz.

As a side comment it is interesting to work out the orbiting distance of a satellite where special and general relativity cancel, i.e., the net time dilation is zero. We must have

$$\frac{t_1}{t_E} \cdot \frac{t_2}{t_E} = 1. \tag{9.30}$$

Hence,

$$\frac{1}{\sqrt{1 - \frac{g}{c^2} \cdot \frac{R_E^2}{R}}} \cdot \frac{\sqrt{1 - \frac{g}{c^2} \cdot 2R_E}}{\sqrt{1 - \frac{g}{c^2} \cdot 2\frac{R_E^2}{R}}} = 1. \tag{9.31}$$

By approximating

$$\sqrt{(1-x)} = (1-x)^{\frac{1}{2}} \approx 1 - \frac{x}{2}, \tag{9.32}$$

we can rewrite Eq. (9.31) as

$$\frac{1}{1 - \frac{g}{2c^2} \cdot \frac{R_E^2}{R}} \cdot \frac{1 - \frac{g}{2c^2} \cdot 2R_E}{1 - \frac{g}{2c^2} \cdot 2\frac{R_E^2}{R}} = 1. \tag{9.33}$$

Since for all x_k above, $x_k \ll 1$ we can write

$$\frac{1}{1-x_1} \cdot \frac{1-x_2}{1-x_3} \approx \frac{1-x_2}{1-x_1-x_3} = 1. \tag{9.34}$$

Clearly, we must have

$$x_2 = x_1 + x_3, \tag{9.35}$$

or, in other words,

$$2R_E = \frac{R_E^2}{R} + 2\frac{R_E^2}{R} = 3\frac{R_E^2}{R}. \tag{9.36}$$

Finally, we have the orbit radius at which the relativistic effects cancel as

$$R = \frac{3}{2} R_E. \tag{9.37}$$

Hence, the height h above the earth surface at which the relativistic effects cancel is

$$h = R - R_E = \frac{1}{2} R_E. \tag{9.38}$$

This result can also be seen in Fig. 9.22.

9.7 Acquisition

Most GPS receivers operate in two different modes: the acquisition mode, where a coarse synchronization process searches for the satellite signals, and a tracking mode, in which the SV signals found are followed in order to track fluctuations of carrier and code phase.

The acquisition is the tiresome job of searching a three-dimensional space for the correlation peak. One dimension is the SV number (at least if the receiver has no *a-priori* knowledge). The second dimension is the frequency space, since an unknown Doppler shift of the SV and an inaccurate frequency reference in the receiver produce a high frequency offset uncertainty. The third dimension is the code phase.

The correlation function can be written as

$$y = f(\tau) = \int_{t_0}^{t_0+T_{\text{seq}}} r(t) \cdot c(t+\tau)dt, \tag{9.39}$$

with $r(t)$ and $c(t)$ being the received signal and the code sequence, respectively. The generic correlator channel is shown in Fig. 9.23. Note that for both the I and the Q channel, three equidistantly spaced code phases are correlated. They are denoted by E (early), P (prompt), and L (late), respectively. Later in the explanation of the tracking mode, we will see that the code phase is adjusted such as to give equal early and late correlation and maximum prompt correlation.

Among the challenges of the acquistion process are the two mechanisms by which the acquisition time is influenced by the signal strength. Not only does the number of correlation units (measured in ms) grow by a factor of 10 for every 10 dB less signal power, but with the growing correlation duration also the grid resolution in the frequency dimension needs to be finer, leading to another factor of 10 for every 10 dB less signal power. Fig. 9.24 shows the relationship between signal level and correlation time. Obviously there is a need for parallelism here.

Many modern GPS receivers have 8 or 12 such channels in parallel to facilitate a faster search process. For every channel, which searches for a different satellite, all possible code phases, usually with a resolution

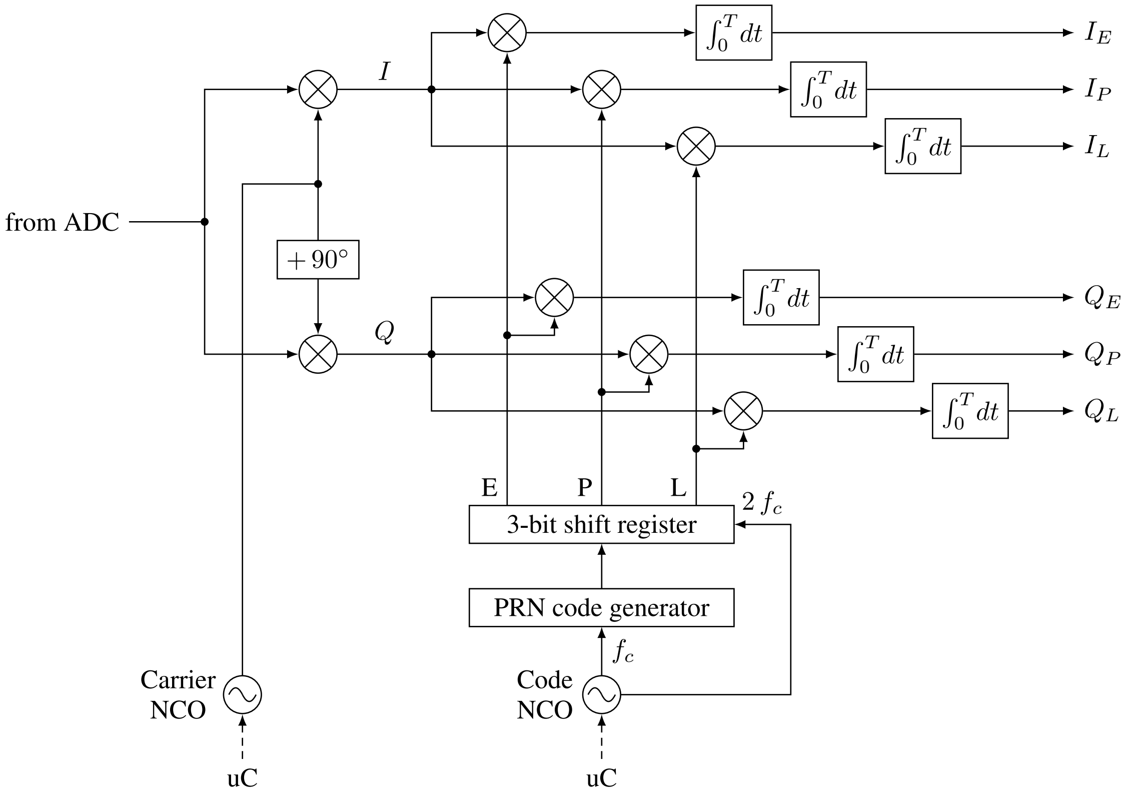


Figure 9.23 Generic correlator channel.

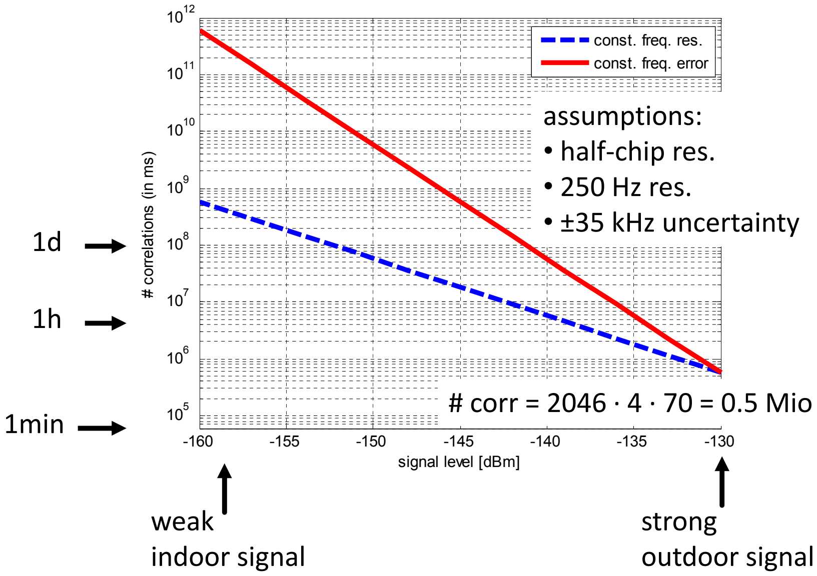


Figure 9.24 Correlation complexity as a function of signal level. The assumptions are shown for the strong signal and the duration times on the left indicate the time needed for a real-time correlator of a single channel.

of one half chip, have to be correlated until a peak is found. Although the search space consists of many frequencies, the carrier phase is still an unknown, and may even change during the duration of a code transmission. Traditionally, the search space of low-end receivers without prior knowledge is large. We assume that half-chip resolution is required for the acquisition process, and the frequency uncertainty is 20 ppm. This adds 30 kHz frequency uncertainty to the max. Doppler frequency of roughly 5 kHz. The search range is therefore ±35 kHz. Accepting the loss of 2.5 dB for half-chip resolution and applying the same loss to the frequency resolution means a resolution of 800 Hz for a dwell time of 1 ms, or in general 80 % of 1 over the dwell time. To play safe, a resolution of 500 Hz is assumed.

For a cold start (SVs in view are not known), the search space is comprised of the following dimensions:

- SV number, range: 1 to 28 (or more), size: 28 to 32
- code phase, range: 0 to 2045 (half-chips), size: 2046
- frequency, range: -35 kHz to +35 kHz, size: 141

For every satellite, the search space is therefore $N = 2046 \cdot 141 = 288486$.

The phase relation between the SV transmitter and the receiver is arbitrary due to a residual frequency offset. Hence, the peak may show up at either the I or the Q channel, or in both. Furthermore, the sign of the peak is unknown due to a possible carrier phase shift of $-180\,°$. Therefore, the signals at the output of the integrator are combined into

$$s = I_P^2 + Q_P^2. \tag{9.40}$$

The signal s can then be constantly compared to a threshold. If the threshold is exceeded, a correlation peak has been registered. If no peak is found after one full search through the code phases, the frequency has to be changed until either a peak correlation is found, or the search through the entire assigned frequency window has been completed. The correlator may then change its search to a different satellite.

The duration of the acquisition process heavily depends on the knowledge. If nothing is known, up to 32 SVs have to be searched for, over both the entire code-phase window and the frequency window. This is known as a cold start. Since the whole navigation message needs to be downloaded, such a cold start may last up to 30 minutes, particularly for older devices. If the time and rough location are known, the SV in sight can be computed from a recent almanac. This is known as a warm start. A warm start should not last more than one or two minutes. A hot start is when the receiver was switched off for less than 2 hours. In such a case reacquisition may only last a couple of seconds.

Most conventional receivers do not provide the memory to capture a certain amount of data that can be searched over the whole space at a very high speed. Thus, for every point in the two-dimensional space of code phase and frequency offset, new data need to be sampled. If one integration lasts over one code period of 1 ms, the entire code phase search at resolution of 1 chip lasts 1 second! And this concerns one frequency offset point only! It becomes clear from this that the desired resolution greatly affects acquisition time. On the other hand, coarse resolution results in performance loss.

9.7.1 Resolution Issues

A link budget usually considers losses due to propagation, noise figures, and implementation losses. The latter is often not further detailed. In the following, we will shortly reveal details of losses due to finite resolutions. The dependence of the carrier-to-noise-density ratio C/N_0 on the chip position and frequency offset is derived and given below.

Chip Resolution

The reception of a direct-sequence spread spectrum system relies on proper alignment of the despreading sequence with respect to the received signal. The correlation of the received signal and the C/A-code (despreading sequence) is positive for position shifts of up to ±1 chips. The correlation result depending on the code position is the well-known triangle with a slope magnitude of one. Hence, the loss of C/N_0 compared to a perfectly aligned despreading sequence is given by

$$\boxed{\text{loss [dB]} = 20\log_{10}\left(1 - \frac{|\Delta t|}{T_{\text{chip}}}\right)}. \tag{9.41}$$

A plot of this relationship is given in Fig. 9.25.

Frequency Resolution

If the Doppler shift of the SV under acquisition and the local oscillator frequency offset are unknown, the correlation of the spreading sequence needs to be carried out for different frequency shifts. Kaplan [32], e.g., considers Doppler bins of size $2/(3T)$, which for the dwell time $T = 1\,\text{ms}$ corresponds to 667 Hz. For the establishment of the relationship between frequency offset in transmitter and receiver, we assume a perfectly aligned code phase. Let us describe the signal at IF by

$$s(t) = G_i(t) \cdot \cos(2\pi f_{\text{IF}} t), \tag{9.42}$$

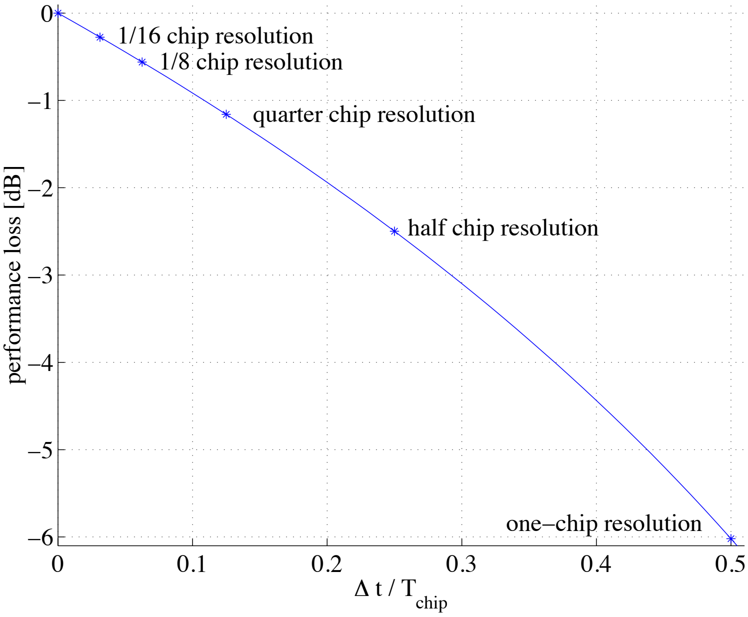


Figure 9.25 Performance loss as a function of code position offset (for an infinite bandwidth).

where $G_i(t)$ is the code function. By multiplying the signal with a duplicate of the code function and an image-rejection IF strip (with frequency offset f_{offset}), we get the baseband signals

$$\begin{aligned} b_{\text{I}}(t) &= G_i^2(t) \cdot \cos(2\pi f_{\text{IF}} t) \cdot \cos(2\pi (f_{\text{IF}} + f_{\text{offset}}) t) \\ &= \frac{1}{2}\cos(2\pi f_{\text{offset}} t) + \frac{1}{2}\cos(2\pi (2 f_{\text{IF}} + f_{\text{offset}}) t), \end{aligned} \tag{9.43a}$$

$$\begin{aligned} b_{\text{Q}}(t) &= G_i^2(t) \cdot \cos(2\pi f_{\text{IF}} t) \cdot \sin(2\pi (f_{\text{IF}} + f_{\text{offset}}) t) \\ &= \frac{1}{2}\sin(2\pi f_{\text{offset}} t) + \frac{1}{2}\sin(2\pi (2 f_{\text{IF}} + f_{\text{offset}}) t), \end{aligned} \tag{9.43b}$$

where we used the fact that $G_i^2(t) = 1$. If $1/f_{\text{IF}}$ is small compared to the integrator time, we get at the output of the integrate-and-dump block

$$d_{\text{I}} = \int_0^T b_{\text{I}}(t)\,dt = \frac{\sin(2\pi f_{\text{offset}} T)}{4\pi f_{\text{offset}}}, \tag{9.44a}$$

$$d_{\text{Q}} = \int_0^T b_{\text{Q}}(t)\,dt = \frac{1 - \cos(2\pi f_{\text{offset}} T)}{4\pi f_{\text{offset}}}. \tag{9.44b}$$

The squared values of these signals are depicted in Fig. 9.26. As can be seen from the figure, for an integrator signal to be independent of phase offset and resistant to a wider range of frequency offsets, the power of both in-phase and quadrature components needs to be considered. Eqs. (9.44a) and (9.44b) squared and summed yield

$$d_{\text{I}}^2 + d_{\text{Q}}^2 = \frac{1 - \cos(2\pi f_{\text{offset}} T)}{8\pi^2 f_{\text{offset}}^2}. \tag{9.45}$$

Compared to the value for no frequency offset, which is $T^2/4$, we can now quantify the loss as

$$\text{loss [dB]} = 10\log_{10}\frac{1 - \cos(2\pi f_{\text{offset}} T)}{2\pi^2 f_{\text{offset}}^2 T^2} = 10\log_{10}\frac{\sin^2(\pi f_{\text{offset}} T)}{(\pi f_{\text{offset}} T)^2} \tag{9.46}$$

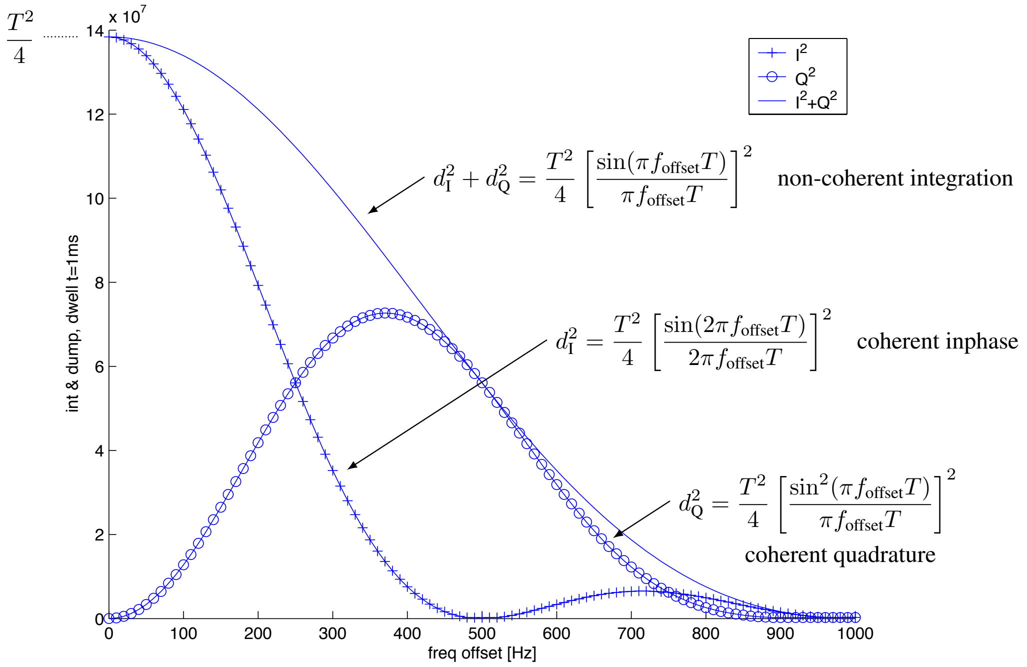


Figure 9.26 Output of integrate-and-dump block as a function of frequency offset for an integration time (dwell time) of 1 ms.

or, equivalently,

$$\boxed{\text{loss [dB]} = 20\log_{10}\left|\frac{\sin(\pi f_{\text{offset}}T)}{\pi f_{\text{offset}}T}\right|}. \tag{9.47}$$

A plot of this relationship is given in Fig. 9.27. The potential loss of the example given above (667 Hz-wide bins which means that the frequency error is 333.5 Hz at most) is therefore between 1.5 and 2 dB.

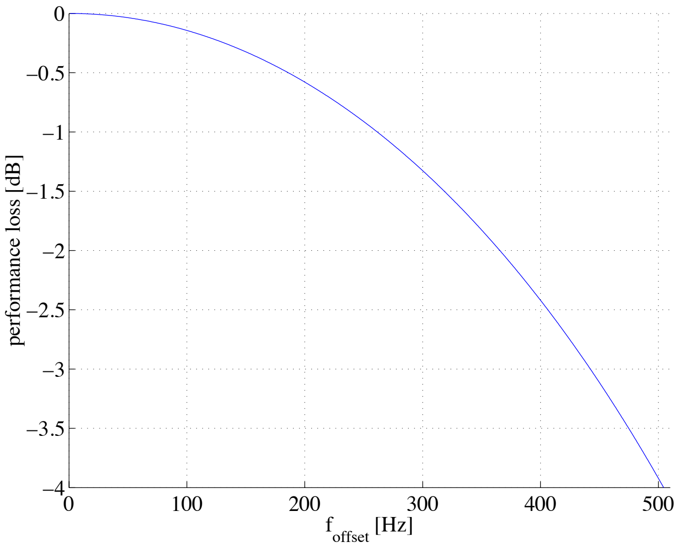
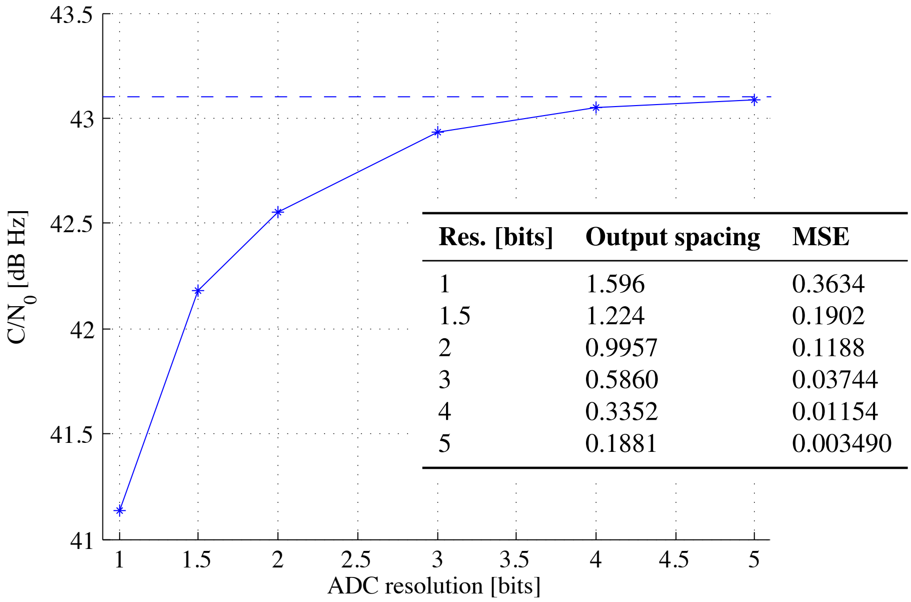


Figure 9.27 Performance loss as a function of frequency offset for an integration time of 1 ms.

9.7.2 A-to-D Conversion

A-to-D Conversion with N-bit Resolution

In general, the solution to the problem of finding the optimum levels for a quantizer is known as the Lloyd-Max quantizer. The quantizer values and corresponding mean-squared errors for uniform level spacing and a Gaussian distribution are given by [52]. The most useful results are duplicated in the legend of Fig. 9.28. Monte-Carlo simulations were carried out in order to investigate the influence of finite precision of the A-to-D conversion. The results are depicted in Fig. 9.28. As can be seen, the carrier-to-noise ratio does not benefit from more than 4 bits spent for the ADC, which confirms the findings of an earlier report.



Res. [bits]	Output spacing	MSE
1	1.596	0.3634
1.5	1.224	0.1902
2	0.9957	0.1188
3	0.5860	0.03744
4	0.3352	0.01154
5	0.1881	0.003490

Figure 9.28 Carrier-to-signal ratio as a function of ADC resolution.

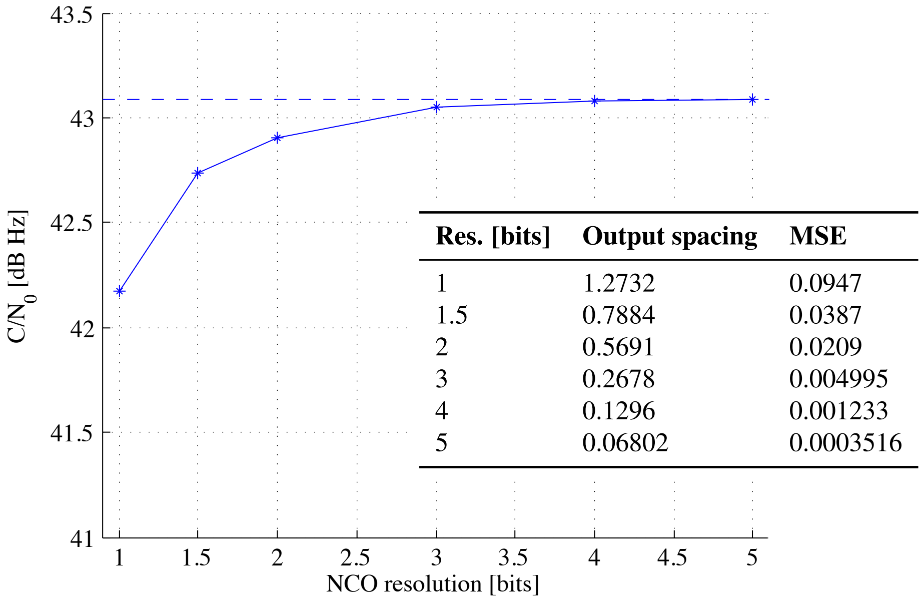
Finite NCO Resolution

Similarly to the finite resolution of the received signal, the amplitude quantization of the NCO has an influence on the performance (signal-to-noise ratio). The waveforms of the inphase and quadrature component of an image-rejection mixer are both sinusoidal. The Lloyd-Max quantizer problem has therefore to be solved for the amplitude density of a sine wave, which can be written as

$$p_x(x) = \begin{cases} \dfrac{1}{\pi\sqrt{1-x^2}}, & -1 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases} \tag{9.48}$$

The optimum quantization levels and the corresponding MSE have been computed numerically and are summarized in the legend of Fig. 9.29.

Interestingly, for the 1.5 bit quantizer, the optimal percentage of the signal in the dead zone is different from that for a Gaussian signal. For the sine wave, the MMSE is obtained for 25.8 % zeros. This is due



Res. [bits]	Output spacing	MSE
1	1.2732	0.0947
1.5	0.7884	0.0387
2	0.5691	0.0209
3	0.2678	0.004995
4	0.1296	0.001233
5	0.06802	0.0003516

Figure 9.29 Carrier-to-signal ratio as a function of NCO resolution.

to the different fourth-order cumulant of a sine wave (the amplitudes are concentrated around two points rather than around zero as in the Gaussian case). Monte-Carlo simulation results for different quantization resolutions are shown in Fig. 9.29.

9.8 Tracking

After the frequency and code phase parameters have been obtained from the acquisition process, they are passed on to the tracking mode.

Three basic parameters are tracked:

- the code phase,
- the carrier frequency, and
- the carrier phase.

The latter two are often tracked using a PLL, whereas the code phase is tracked using an independent loop, usually implemented with a DLL (delay-locked loop), as shown in Fig. 9.30.

Let us have a look at the DLL in more detail. The DLL is based around the correlation triangle, whose peak occurs for exact alignment (prompt) of the code sequences. Since the peak's height itself depends on the signal strength, one correlation result alone can never reveal the direction of misalignment. In fact, the correlation result is evaluated at two positions before and after the main peak, usually spaced by one chip duration (denoted by the E-L spacing in Fig. 9.31). The correlation window are referred to as early and late gate, respectively. Fig. 9.31 illustrates this behavior.

By means of a control loop, the results from early and late-gate correlation are balanced by shifting all three correlator times (early, late, and prompt) together forth and back. When early and late-gate correlations

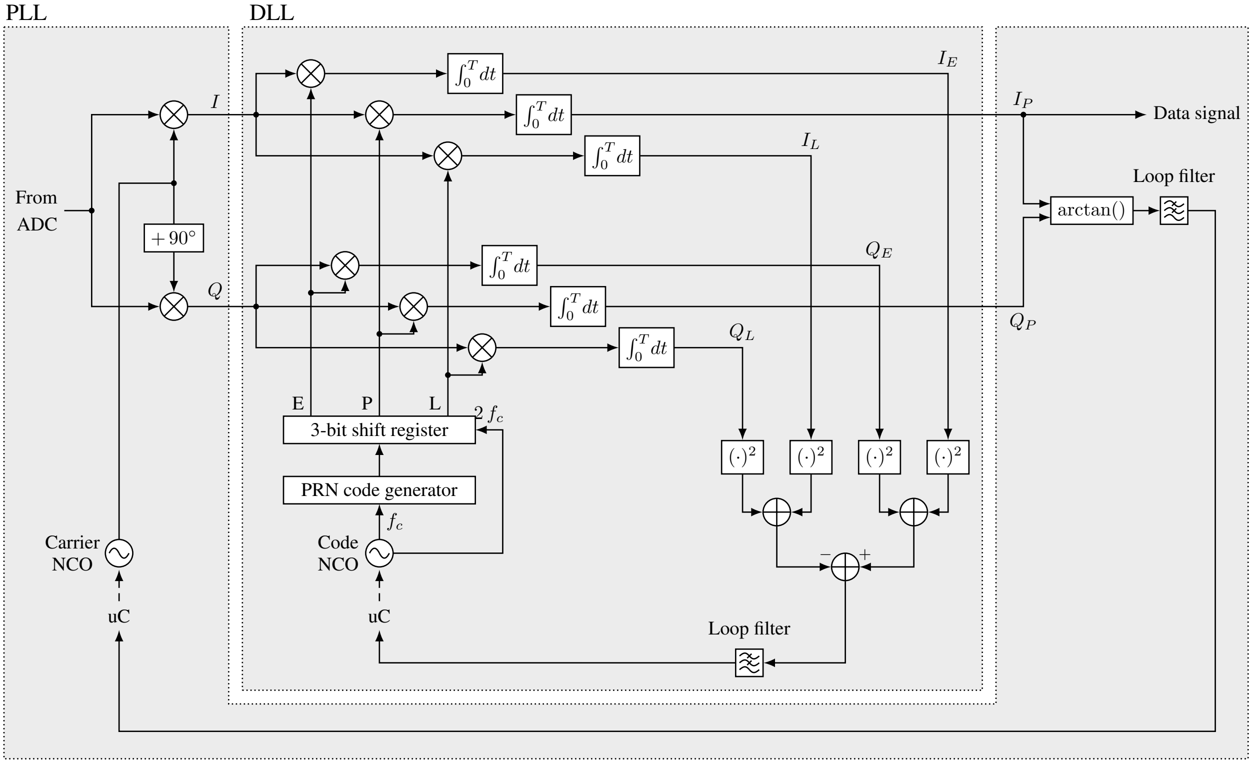
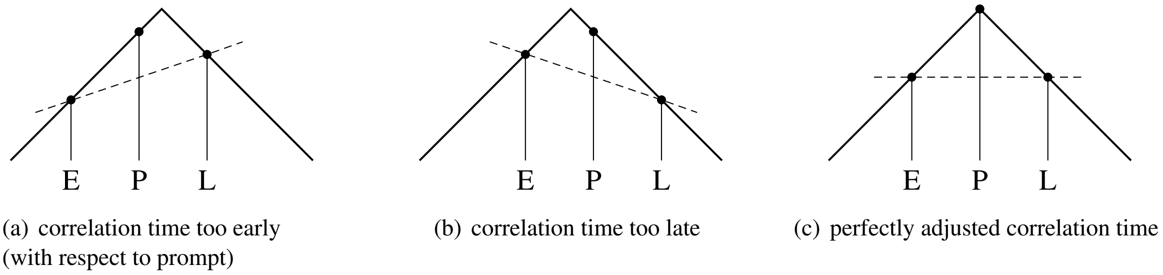


Figure 9.30 Code phase tracking loop (DLL) and carrier phase tracking loop (PLL) of a GPS tracking receiver.



(a) correlation time too early (with respect to prompt)

(b) correlation time too late

(c) perfectly adjusted correlation time

Figure 9.31 Early/late-gate correlation of the DLL (E = early, P = prompt, L = late).

are equal, the prompt correlation will be maximum, and its timing alignment indicates the correct PRN position.

- In picture a) correlation occurs too early, i.e., the result of the early correlation (E) is smaller than that of the late correlation (L).
- As a result, the late correlation result is higher and will retard the correlation times of all three positions (E, P, and L). The contrary is the case in picture b).
- If the correlation time is adjusted correctly, the correlation results of E and L are equal as can be seen in picture c) and no further changing of the correlation time is necessary.

9.9 Performance and Interference

The general precision of positioning that can be achieved with a GPS-L1 receiver depends on many different aspects such as

- global location (the constellation is not optimized for usage close to the poles)
- local location (inside urban canyons, i.e., large buildings standing close together, only few SV might be visible)
- dynamic movements of the user, which makes tracking a tougher task
- quality of the signal (the deliberate jittering of the signal called selective availability or SA has been switched off by the US in the year 2000)
- complexity of the receiver (phase tracking allows to achieve centimeter accuracy levels)
- other interference sources such as jammers
- reduced ephemerides precision (5m)
- refraction in troposphere/ionosphere
- multipath propagation
- poor geometry of the satellites

9.9.1 Multipath Propagation

In Fig. 9.32 we can see the three main propagation possibilities we have in receiving a satellite's signal. Vehicle #1 receives the signal directly from the satellite and nothing else. For some positions on earth, in our case Vehicle #2, the direct path is obstructed, but the signal might propagate through a reflection. Obviously this introduces some error into the pseudorange estimation of that particular satellite. Finally, the situation for Vehicle #3 is the combined of the two former situations, a direct and a reflected path. This is called a multipath situation and although the direct path indicates the correct pseudorange to the SV, an estimation error is introduced owing to the DLL operation and the deformation of the correlation peak. Fig. 9.33 shows the effect of a reflected path on the correlation peak.

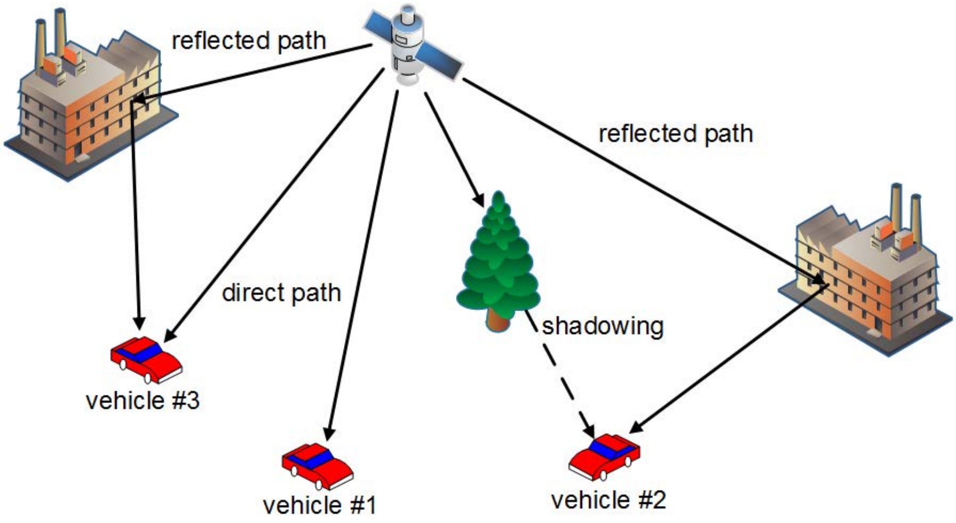
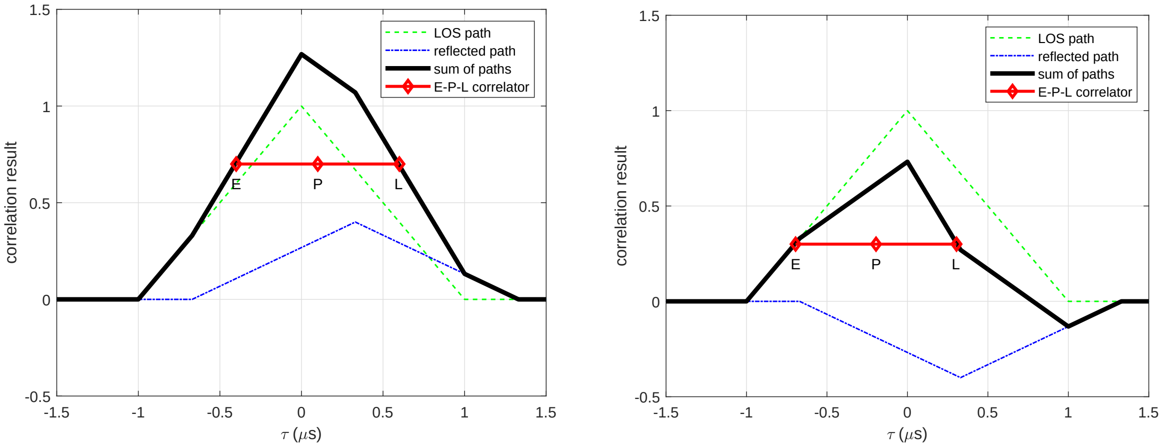


Figure 9.32 Different propagation-paths setups of a GNSS signal.



(a) No phase difference of direct and reflected paths.

(b) Reflected path with 180° phase shift from direct path.

Figure 9.33 Deformation of the correlation peak due to multipath for an early-late spacing of 1 chip.

Note that the peak of the correlation triangle (strictly speaking, it is no longer a triangle) remains at the correct position. Because in the DLL the position of the prompt correlator is exactly in the middle of the early and late correlator positions, we get an estimation error. In the example chosen, the reflected path has 40% of the amplitude of the direct path. The relative delay between the paths is 100 m, which corresponds to one third of a chip. If the relative carrier phase between the two paths is zero as shown in Fig. 9.33(a), the correlation peaks show in the same direction, and we get a positive bias of 0.1 chip (corresponding to 30 m) in the pseudorange estimation process. If, however, the relative carrier phase between the two paths is 180° as shown in Fig. 9.33(b), the correlation peaks show in opposite directions, and we get a negative bias in the pseudorange estimation process. The resulting error in this case is -0.195 chip (corresponding to -58.5 m).

In the following we want to analyze the envelope of the potential pseudorange error for different parameter settings. We leave the multipath delay (between direct path and reflected path) τ as a free parameter. The amplitude a of the reflected path is half that of the direct path, hence $a = \pm\frac{1}{2}$. We can also say that the MDR (multipath-to-direct-path ratio) is -6 dB. Note that as explained in the previous example, the phase of the reflected path is crucial. It determines a positive or a negative pseudorange error. We model a 180° phase shift by a negative sign of a. Furthermore, we designate the parameter s as half the E-L spacing. In effect, s corresponds to the spacing between E (early) and P (prompt). For most E-L spacings we distinguish three zones for the multipath delay τ. If the delay is small, the pseudorange error is proportional to τ,

$$\Delta t = \frac{a}{1+a}\tau \qquad \text{for } 0 \le \tau < (1+a)s. \tag{9.49}$$

Observing Fig. 9.34 we realize that for this zone a narrow correlator spacing does not help. In the second zone, the error is constant, i.e., independent of the multipath delay, but proportional to the correlator spacing,

$$\Delta t = as \qquad \text{for } (1+a)s \le \tau < T-(1-a)s. \tag{9.50}$$

Finally, in the third zone, there is a linear decrease of the influence of the path delay τ for large delays,

$$\Delta t = \frac{a}{2-a}(T+s-\tau) \qquad \text{for } T-(1-a)s \le \tau < T+s. \tag{9.51}$$

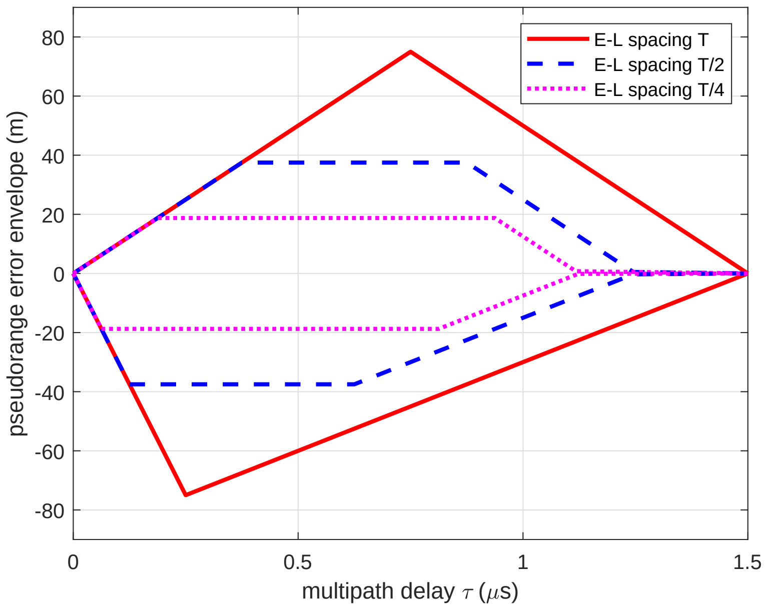


Figure 9.34 Pseudorange error envelope for different E-L spacings and an MDR of -6 dB.

The beginning and the end of each zone is depending on the correlator spacing but also on the phase of the reflected path. A real-world measurement of a correlation triangle in a multipath environment has been carried out in [45] and is shown in Fig. 9.35.

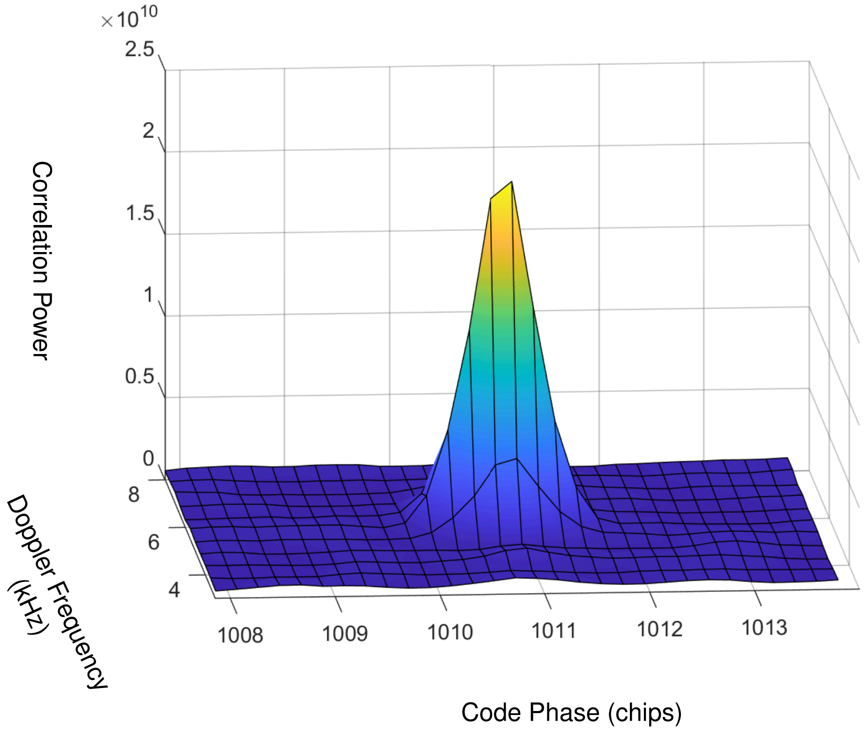


Figure 9.35 Real-world measurement of a correlation triangle in a multipath environment.

9.9.2 Refraction Effects

Refraction occurs in the ionosphere and the troposphere. Ionospheric refraction is depending on the (total) electron content (measured in so-called TEC units, TECU), which in turn depends on the position, the season, and the time of day. An example of the differences can be seen in Fig. 9.36. Often, the knowledge of this distribution is transferred via geostationary satellites (e.g. EGNOS) to the user and can thus be compensated for.

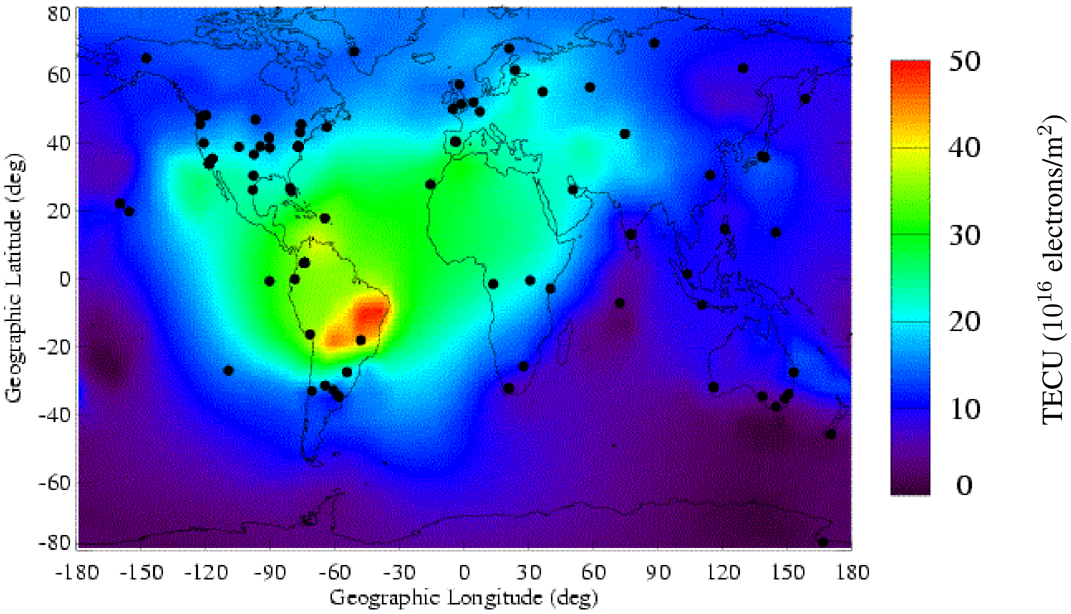

Geographic Latitude (deg)
80
60
40
20
0
-20
-40
-60
-80
-180 -150 -120 -90 -60 -30 0 30 60 90 120 150 180
Geographic Longitude (deg)
50
40
30
20
10
0
TECU (10^{16} electrons/m^2)

Figure 9.36 Distribution of total electron content (TEC) as of May 18, 2011. Source: `http://wondering-star.blogspot.ch/2011/05/jpl-global-maps-of-real-time.html`.

Refraction in the troposphere, which is dependent both on the weather and on the elevation angle of the satellite as can be seen in Fig. 9.37.

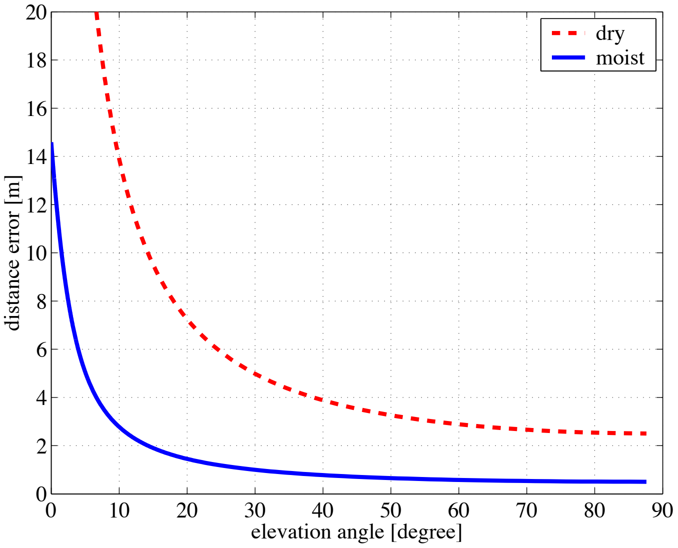


Figure 9.37 Positioning error due to refraction in the troposphere.

9.9.3 Accuracy

Accuracy is a big issue for GPS devices. Manufacturer often misuse it for marketing's sake. In order to compare accuracy of two different devices, we must always make sure we compare the same parameter, see also [80]. A popular figure of accuracy is the CEP (circle error probable). It describes a horizontal circle around the true location that contains half of the estimated locations. Similarly, we have another parameter, R95, containing 95% of the points. Furthermore,

$$\text{R95} = 2.08 \cdot \text{CEP}. \tag{9.52}$$

Yet another parameter is the RMS. The RMS (root mean square) can be derived as follows. The distance squared is a chi-square distribution, whose mean is 2 (for degree 2). If the underlying Gaussian processes have a variance of σ, the RMS is $\sqrt{2}\sigma$. Since for the Rayleigh distribution the median (CEP) is equal to $\sqrt{\ln 4}\sigma$, we have

$$\text{RMS} = \sqrt{\frac{2}{\ln 4}}\,\text{CEP} = 1.2 \cdot \text{CEP}. \tag{9.53}$$

Of course, the switching-off of SA in the year 2000 has helped a great deal in improving the accuracy of GPS. See Fig. 9.38 for the effect.

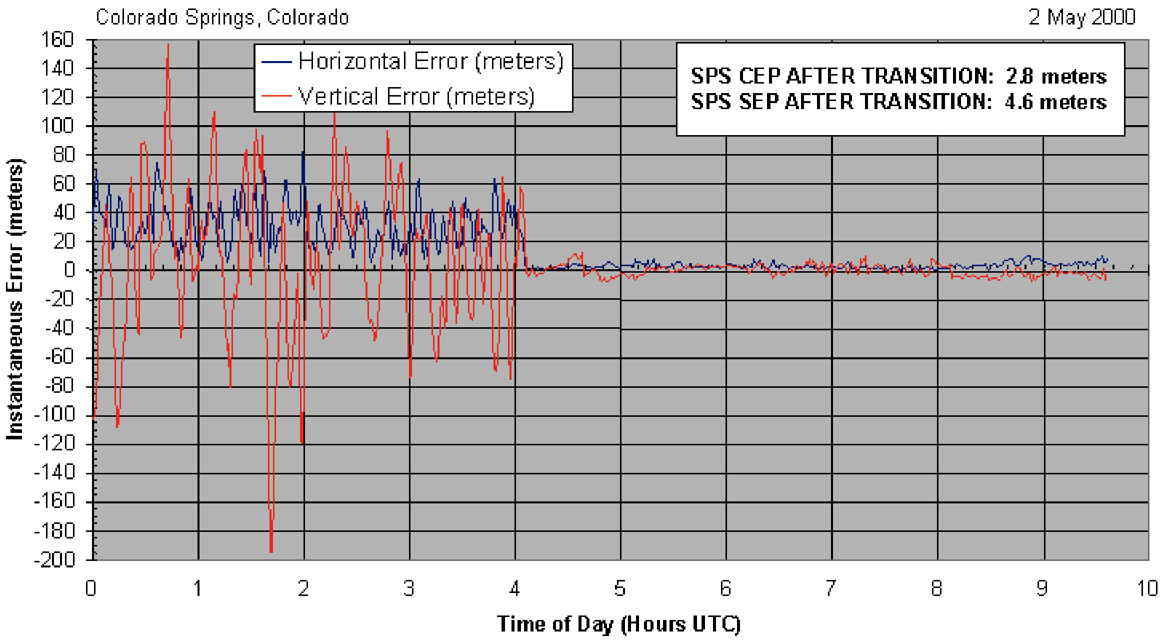


Figure 9.38 Instantaneous Error of GPS vs. TOD.

Besides precision and accuracy, there are two other quality parameters, namely availability and integrity. Availability describes the probability that the systems works at a given place. The integrity of GPS refers to the trustworthiness, in other word, can some imposter of GPS mislead someone to believe a wrong position by generating signals. The integrity of a system is very important in the area of aviation.

9.9.4 Nature of Jamming Noise

Different types of jamming are known. The different types can roughly be divided into in-band jammers and out-of-band jammers. The first type is obviously more damaging, since it passes through all filters in the receiver chain of a GPS receiver. In-band jammers might be further divided into narrowband jammers, usually referred to as cw[3] jammers and wideband jammers, which may occupy a large proportion of the GPS band.

Out-of-band jammers are usually less harmful, since they are considerably filtered once they reach the baseband processing. If they are very strong, however, they might block the AGC, amplifiers, and the ADC. In particular, out-of-band jammers may only block the ADC if the digital bandwidth is considerably different (smaller) from the analog IF filter, otherwise they would fall in-band or be filtered by the IF filter.

9.9.5 Sources of Jamming Noise

The sources of artificial jamming noise falling close to a GPS signal are multiple. Apart from deliberate jamming, which is not of serious concern to civil applications, but to military usage, there are two main concerns relating to jamming noise. First, the advent of *ultrawideband systems* (UWB) is thought to produce interference to GPS systems, although it is currently difficult either to confirm or to disprove any such effects. Second, the increasing integration of GPS into wireless communication equipment bears the difficulty of coping with intermodulation products.

For future wireless handheld transmitters, the frequency occupation will be important to any jamming effects in a GPS receiver integrated in the same product. The relevant unlicensed bands for mobile communication are allocated around 433, 866, 2400, and 5800 MHz. These bands are also referred to as ISM bands (industrial, scientific, medical) and are used for commercial applications such as WLAN (802.11b and 802.11a) and PAN (Bluetooth). Unlicensed transmissions are usually low in power density but rather wideband. The power output allowed is regulated by the FCC or the corresponding body from the country of operation.

The jamming effect can now occur in two different forms:

- out-of-band jamming noise by directly leaking through via wireless transmitting antenna and GPS receiving antenna
- in-band jamming noise created by intermodulation products of the transmitter power amplifier that fall into the GPS band

The second effect is widely independent of the IF filter used, if the bandwidth of the signal is considerably smaller than the IF filter.

In the absence of AGC, ADC, and any other possibly nonlinearly behaving component, the effect of jamming noise is the same as that of any noise. The carrier-to-noise-density ratio C/N_0 is reduced by the amount of the total additional noise. Hence, e.g., if a thermal noise power of -110 dBm results from a given choice of the IF filter, additional -110 dBm of in-band jamming power will result in a decrease of C/N_0 of 3 dB.

9.9.6 Doppler and Doppler change

The satellites (SV for space vehicle) revolve around the earth on a MEO (medium earth orbit) every twelve hours. In the following, we assume the orbit to be circular, an assumption which is roughly true. Fig. 9.39 displays the constellation between a user on the earth and one SV in the orbit. In order to evaluate the Doppler speed as a function of the elevation angle β, let us denote R_E the radius of the earth and R_S the orbit radius. Note that by elevation we mean the angle as seen from the center of the earth, not from the user.

[3] continuous wave, modeled as one discrete frequency

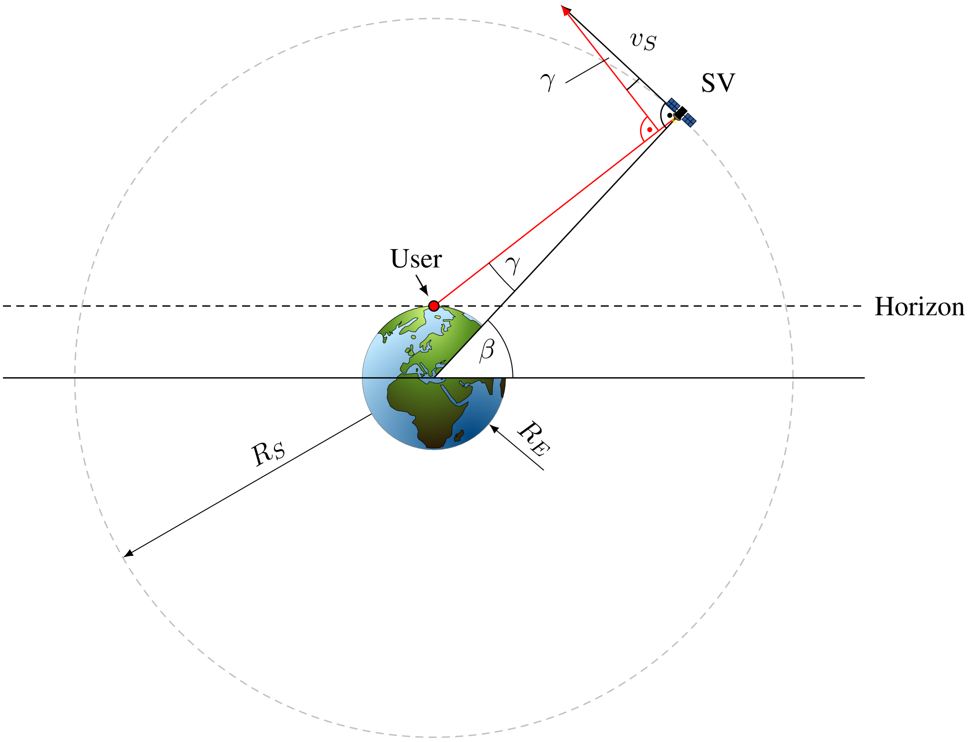


Figure 9.39 Geometrical setup of SV and user to evaluate Doppler speed and Doppler change.

Now we introduce the angle γ as seen in Fig. 9.39. Furthermore, r denotes the range (distance) user-SV. The law of sines states that

$$\frac{\sin\gamma}{R_E} = \frac{\sin(90^\circ - \beta)}{r} = \frac{\cos\beta}{r}. \tag{9.54}$$

On the other hand, the law of cosines states that

$$\begin{aligned} r^2 &= R_E^2 + R_S^2 - 2R_E R_S \cos(90^\circ - \beta) \\ &= R_E^2 + R_S^2 - 2R_E R_S \sin\beta. \end{aligned} \tag{9.55}$$

Once γ and r are known, the Doppler speed calculates as

$$\begin{aligned} v_{\text{Doppler}} &= v_{\text{SV}} \sin\gamma \\ &= v_{\text{SV}} R_E \frac{\cos\beta}{r} \\ &= v_{\text{SV}} R_E \frac{\cos\beta}{\sqrt{R_E^2 + R_S^2 - 2R_E R_S \sin\beta}}, \end{aligned} \tag{9.56}$$

where v_{SV} is the speed of the SV. For GPS satellites, we have v_{SV}=3.8704 km/s. Since the speed of the earth at the equator is only about 10 % of that value, we neglect its effect. The angle β is a linear function of the time. It can now be used to evaluate the Doppler speed, which is depicted in Fig. 9.40. As the radii of earth and SV track we have taken $R_E = 6378.1363$ km and $R_S = 26561.75$ km, respectively. The visibility of the SV is determined by $\gamma = \beta$ or the relation $\sin\beta = R_E/R_S$.

Although the Doppler speed is not a monotonous function of the elevation angle, the maxima still occur at the limits of visibility. The visibility of the SVs is determined by the horizon, which is around 14° above

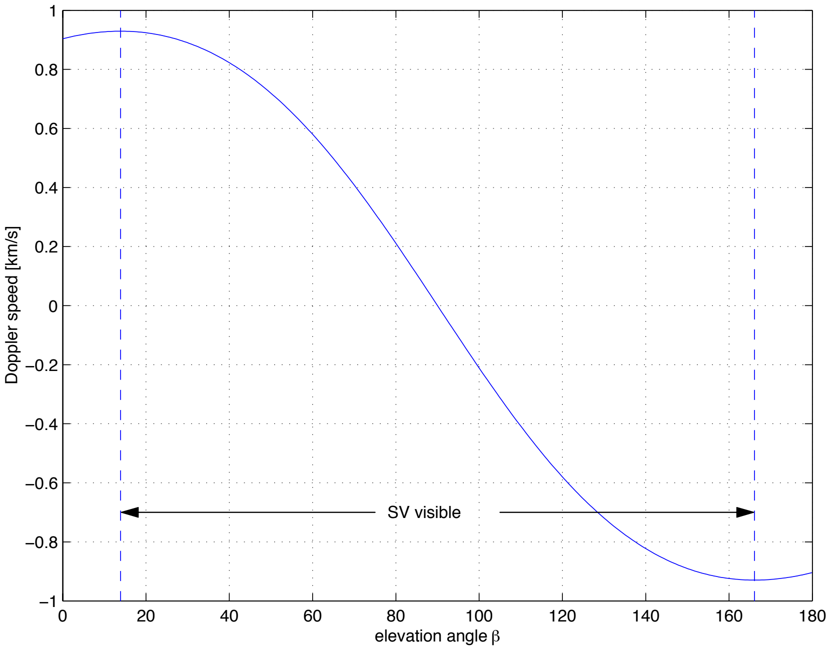


Figure 9.40 Doppler speed vs. elevation angle.

the zero-elevation point of the SV. The Doppler frequency (Doppler shift) itself is defined as

$$f = \frac{\frac{v}{c} f_{\mathrm{RF}}}{\sqrt{1-\frac{v^2}{c^2}}} \approx \frac{v}{c} f_{\mathrm{RF}} \quad (\text{if } v \ll c) \tag{9.57}$$

and becomes proportional to the Doppler speed when $v \ll c$. If the Doppler speed is 0 (at an elevation angle of 90°), the Doppler frequency vanishes. At the limits of visibility (earth curvature considered only), we have

$$f_{\max} = \frac{v_{\max}}{c} f_{\mathrm{RF}} = \frac{9.3 \cdot 10^2\,\text{m/s}}{3 \cdot 10^8\,\text{m/s}} 1.57542\,\text{GHz} \approx 4.88\,\text{kHz}. \tag{9.58}$$

Taking the difference of the Doppler speed, we can arrive at the Doppler change,

$$\Delta f = \frac{\Delta v}{c} f_{\mathrm{RF}}. \tag{9.59}$$

The maximum Doppler change occurs around 90° elevation angle. From Fig. 9.40 we can see that this is

$$\left|\frac{\Delta v}{\Delta \beta}\right|_{\max} = 0.0213\,\text{km/(°s)}. \tag{9.60}$$

With a change of

$$\Delta\beta = \frac{360°}{12 \cdot 60 \cdot 60} = 0.0083\,°\text{/s}, \tag{9.61}$$

since the SVs revolve around the earth every 12 hours, the maximum Doppler speed change is

$$\Delta v_{\max} = \Delta\beta \left|\frac{\Delta v}{\Delta \beta}\right|_{\max} = 1.7679 \cdot 10^{-4}\,\text{km/s}^2, \tag{9.62}$$

or expressed as the maximum Doppler frequency change

$$\Delta f_{\max} = \frac{\Delta v_{\max}}{c} f_{\mathrm{RF}} = \frac{1.7679 \cdot 10^{-1}\,\mathrm{m/s^2}}{3 \cdot 10^8\,\mathrm{m/s}} 1.57542\,\mathrm{GHz} \approx 1\,\mathrm{Hz/s}, \tag{9.63}$$

a value that has to be compared to that originating from normal user acceleration. To this end, let us imagine two scenarios: First, a vehicle acceleration from 0 to 100 km/s in 10 s is equivalent to an acceleration of 2.8 m/s^2, which, in turn, gives us a maximum Doppler frequency change of 15 Hz/s. Second, assume a traffic roundabout of diameter 20 m. If a car can manage to take it with a speed of 40 km/h, it will experience an acceleration of $a = v^2/r = 12.3\,\mathrm{m/s^2}$, which equates to a maximum Doppler frequency change of 65 Hz/s. In general, every g of acceleration in the direction of the transmission equates to a Doppler frequency change of

$$\Delta f = \frac{9.81\,\mathrm{m/s^2}}{3 \cdot 10^8\,\mathrm{m/s}} 1.57542\,\mathrm{GHz} = 51\,\mathrm{Hz/s}. \tag{9.64}$$

9.10 Navigation

After the correlation of the incoming signal with PRNs from at least four different satellites, the user can determine the time shift of the correlation peak with respect to a time reference in the receiver. This time reference is usually incorrect with respect to the time references of the SV, which are all synchronized to each other due to the atomic standards on board. The timing offset τ of the receiver reference can be computed using the fourth SV range. The principle mode of operation can be seen in Fig. 9.41. The time shifts mentioned give the so-called pseudorange ρ to the individual satellites. The name pseudorange indicates that they are not true positions but the positions acquired for the assumed time reference. Using the three-

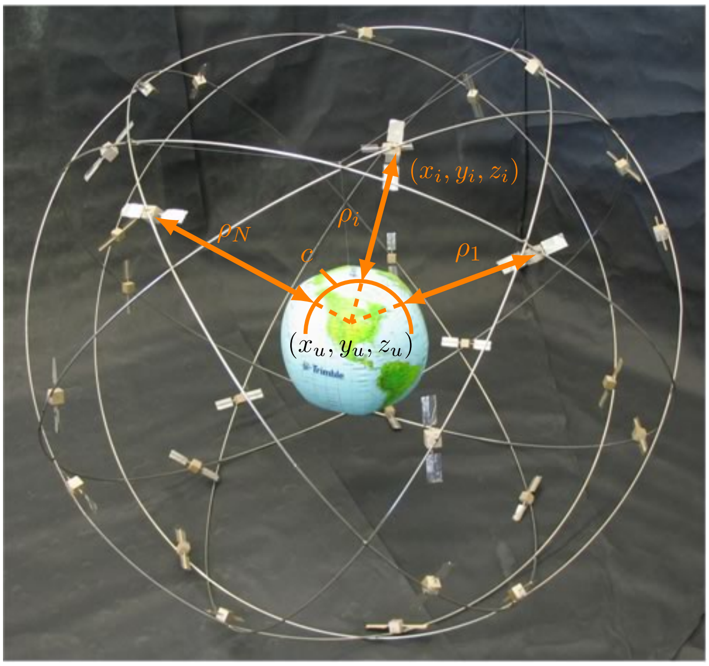


Figure 9.41 Principle of position estimation from pseudoranges.

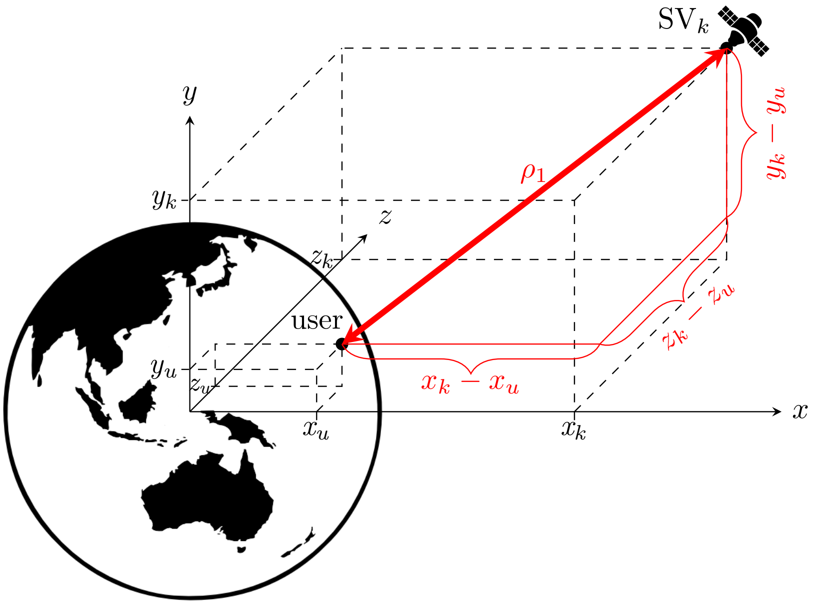


Figure 9.42 Coordinate system illustrating the three-dimensional Pythagoras theorem to build the system of equations (9.65a) to (9.65d).

dimensional coordinates of the kth SV, see also Fig. 9.42, (x_k,y_k,z_k), and coordinates of the user, (x_u, y_u, z_u), in an earth-centered cartesic coordinate system, a system of equations can be set using the coordinates of four different SVs

$$\rho_1 = \sqrt{(x_1 - x_\mathrm{u})^2 + (y_1 - y_\mathrm{u})^2 + (z_1 - z_\mathrm{u})^2} + c\tau, \tag{9.65a}$$

$$\rho_2 = \sqrt{(x_2 - x_\mathrm{u})^2 + (y_2 - y_\mathrm{u})^2 + (z_2 - z_\mathrm{u})^2} + c\tau, \tag{9.65b}$$

$$\rho_3 = \sqrt{(x_3 - x_\mathrm{u})^2 + (y_3 - y_\mathrm{u})^2 + (z_3 - z_\mathrm{u})^2} + c\tau, \tag{9.65c}$$

$$\rho_4 = \sqrt{(x_4 - x_\mathrm{u})^2 + (y_4 - y_\mathrm{u})^2 + (z_4 - z_\mathrm{u})^2} + c\tau, \tag{9.65d}$$

where c is the speed of light. The satellite positions are known from the navigation message. The system of equation now has to be solved for the unknowns x_u, y_u, z_u, and τ. If we have pseudoranges of more than four satellites, a more accurate solution can be computed solving the system of equations in the least-squares sense. On the other hand, if three pseudoranges are available only, a so-called 2D fix can be computed based on the assumption that the user 'sits' on the earth surface and his/her altitude is therefore known. A 2D fix can also be thought of as the solution achieved by a satellite being in the center of the earth whose pseudorange corresponds to the radius of the earth.

The solution of Eqs. (9.65a) to (9.65d) being a set of nonlinear equations is more involved than a set of linear equations. The most prominent approaches to the solution of the above system employ iterative techniques based on a linearization of the equations [32]. To this end, the pseudoranges

$$\rho_k = \sqrt{(x_k - x_\mathrm{u})^2 + (y_k - y_\mathrm{u})^2 + (z_k - z_\mathrm{u})^2} + c\tau = f_k(x_\mathrm{u}, y_\mathrm{u}, z_\mathrm{u}) \tag{9.66}$$

are approximated at first order using a Taylor series

$$\begin{aligned}f_k(\hat{x}_\mathrm{u}+\Delta x_\mathrm{u},\hat{y}_\mathrm{u}+\Delta y_\mathrm{u},\hat{z}_\mathrm{u}+\Delta z_\mathrm{u},\hat{\tau}+\Delta\tau) &= f_k(\hat{x}_\mathrm{u},\hat{y}_\mathrm{u},\hat{z}_\mathrm{u},\hat{\tau})\\ &+\frac{\partial f_k(\hat{x}_\mathrm{u},\hat{y}_\mathrm{u},\hat{z}_\mathrm{u})}{\partial\hat{x}_\mathrm{u}}\Delta x_\mathrm{u}+\frac{\partial f_k(\hat{x}_\mathrm{u},\hat{y}_\mathrm{u},\hat{z}_\mathrm{u})}{\partial\hat{y}_\mathrm{u}}\Delta y_\mathrm{u}+\frac{\partial f_k(\hat{x}_\mathrm{u},\hat{y}_\mathrm{u},\hat{z}_\mathrm{u})}{\partial\hat{z}_\mathrm{u}}\Delta z_\mathrm{u}+\frac{\partial f_k(\hat{x}_\mathrm{u},\hat{y}_\mathrm{u},\hat{z}_\mathrm{u})}{\partial\hat{\tau}}\Delta\tau.\end{aligned} \tag{9.67}$$

where $(\hat{x}_\mathrm{u},\hat{y}_\mathrm{u},\hat{z}_\mathrm{u})$ is an initial guess of the user position and $(\Delta x_\mathrm{u},\Delta y_\mathrm{u},\Delta z_\mathrm{u})$ is the unknown deviation of this estimation from the true position. Likewise, $\hat{\tau}$ and $\Delta\tau$ are the estimated time offset and the difference to the true time offset, respectively. The partial derivatives can now be evaluated as

$$\frac{\partial f_k(\hat{x}_\mathrm{u},\hat{y}_\mathrm{u},\hat{z}_\mathrm{u})}{\partial\hat{x}_\mathrm{u}} = -\frac{x_k-\hat{x}_\mathrm{u}}{\hat{r}_k}, \tag{9.68a}$$

$$\frac{\partial f_k(\hat{x}_\mathrm{u},\hat{y}_\mathrm{u},\hat{z}_\mathrm{u})}{\partial\hat{y}_\mathrm{u}} = -\frac{y_k-\hat{y}_\mathrm{u}}{\hat{r}_k}, \tag{9.68b}$$

$$\frac{\partial f_k(\hat{x}_\mathrm{u},\hat{y}_\mathrm{u},\hat{z}_\mathrm{u})}{\partial\hat{z}_\mathrm{u}} = -\frac{z_k-\hat{z}_\mathrm{u}}{\hat{r}_k}, \tag{9.68c}$$

$$\frac{\partial f_k(\hat{x}_\mathrm{u},\hat{y}_\mathrm{u},\hat{z}_\mathrm{u})}{\partial\hat{\tau}} = c, \tag{9.68d}$$

where

$$\hat{r}_k = \sqrt{(x_k-\hat{x}_\mathrm{u})^2+(y_k-\hat{y}_\mathrm{u})^2+(z_k-\hat{z}_\mathrm{u})^2} \tag{9.69}$$

is the estimated range and

$$\hat{\rho}_k = \hat{r}_k + c\hat{\tau} \tag{9.70}$$

is the estimated pseudorange. Note, that k is used here as the index to the satellite, not as an iteration index. Using Eqs. (9.68a) to (9.68d) and Eq. (9.66) in Eq. (9.67) results in

$$\rho_k = \hat{\rho}_k - \frac{x_k-\hat{x}_\mathrm{u}}{\hat{r}_k}\Delta x_\mathrm{u} - \frac{y_k-\hat{y}_\mathrm{u}}{\hat{r}_k}\Delta y_\mathrm{u} - \frac{z_k-\hat{z}_\mathrm{u}}{\hat{r}_k}\Delta z_\mathrm{u} + c\Delta\tau. \tag{9.71}$$

We have now transformed the set of nonlinear equations into a set of linear equations for the unknowns Δx_u, Δy_u, Δz_u, and $\Delta\tau$. We introduce new variables

$$\Delta\rho_k \overset{\triangle}{=} \hat{\rho}_k - \rho_k, \tag{9.72a}$$

$$\tilde{x}_k \overset{\triangle}{=} \frac{x_k-\hat{x}_\mathrm{u}}{\hat{r}_k}, \tag{9.72b}$$

$$\tilde{y}_k \overset{\triangle}{=} \frac{y_k-\hat{y}_\mathrm{u}}{\hat{r}_k}, \tag{9.72c}$$

$$\tilde{z}_k \overset{\triangle}{=} \frac{z_k-\hat{z}_\mathrm{u}}{\hat{r}_k}. \tag{9.72d}$$

The $\tilde{x}_k,\tilde{y}_k,\tilde{z}_k$ are the components of a unit vector pointing from the estimated user position to the kth SV.

Using the vector/matrix definitions

$$\boldsymbol{\Delta\rho} \triangleq \begin{bmatrix} \Delta\rho_1 \\ \Delta\rho_2 \\ \Delta\rho_3 \\ \Delta\rho_4 \end{bmatrix}, \tag{9.73a}$$

$$\boldsymbol{H} \triangleq \begin{bmatrix} \tilde{x}_1 & \tilde{y}_1 & \tilde{z}_1 & 1 \\ \tilde{x}_2 & \tilde{y}_2 & \tilde{z}_2 & 1 \\ \tilde{x}_3 & \tilde{y}_3 & \tilde{z}_3 & 1 \\ \tilde{x}_4 & \tilde{y}_4 & \tilde{z}_4 & 1 \end{bmatrix}, \tag{9.73b}$$

$$\boldsymbol{\Delta x} \triangleq \begin{bmatrix} \Delta x_\mathrm{u} \\ \Delta y_\mathrm{u} \\ \Delta z_\mathrm{u} \\ -c\Delta\tau \end{bmatrix}, \tag{9.73c}$$

we can rearrange Eq. (9.71) for $k = 1 \ldots 4$ into the matrix equation

$$\boldsymbol{\Delta\rho} = \boldsymbol{H}\boldsymbol{\Delta x}, \tag{9.74}$$

which has the solution

$$\boldsymbol{\Delta x} = \boldsymbol{H}^{-1}\boldsymbol{\Delta\rho}. \tag{9.75}$$

For more than four satellites, where $\boldsymbol{H}$ is no longer a square matrix, the solution can be found in the LS sense

$$\boldsymbol{\Delta x} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{\Delta\rho}. \tag{9.76}$$

The solution can now be found in an iterative manner going through Eqs. (9.69) to (9.75). For ease of reference, the algorithm is formulated again in the following compact form:

Iterative solution to the TOA system of equations

- We have measured some pseudoranges ρ_k.
- Initialization: assume some start values for $\hat{x}_\mathrm{u}, \hat{y}_\mathrm{u}, \hat{z}_\mathrm{u}, \hat{\tau}$.
- Iteration: compute

$$\hat{r}_k = \sqrt{(x_k - \hat{x}_\mathrm{u})^2 + (y_k - \hat{y}_\mathrm{u})^2 + (z_k - \hat{z}_\mathrm{u})^2}, \tag{9.77}$$

$$\hat{\rho}_k = \hat{r}_k + c\hat{\tau}. \tag{9.78}$$

- Update matrices: recompute $\boldsymbol{\Delta\rho}, \boldsymbol{H}$.
- Solve for $\boldsymbol{\Delta x} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{\Delta\rho}$.
- Update the estimated user positions

$$\hat{x}_\mathrm{u} = \hat{x}_\mathrm{u} + \Delta x_\mathrm{u}, \tag{9.79a}$$

$$\hat{y}_\mathrm{u} = \hat{y}_\mathrm{u} + \Delta y_\mathrm{u}, \tag{9.79b}$$

$$\hat{z}_\mathrm{u} = \hat{z}_\mathrm{u} + \Delta z_\mathrm{u}, \tag{9.79c}$$

$$\hat{\tau} = \hat{\tau} + \Delta\tau. \tag{9.79d}$$

- Go back to Eq. (9.77)

It is interesting to note that the matrix $\boldsymbol{H}$ containing information on the directions of the SVs used is also helpful in determining the positioning error. Similarly to Eq. (9.76) we can express the error on the three coordinates and time

$$d\boldsymbol{x} \triangleq \begin{bmatrix} dx_\mathrm{u} \\ dy_\mathrm{u} \\ dz_\mathrm{u} \\ cd\tau \end{bmatrix} \tag{9.80}$$

as a function of the pseudorange errors

$$d\boldsymbol{\rho} \triangleq \begin{bmatrix} d\rho_1 \\ d\rho_2 \\ d\rho_3 \\ d\rho_4 \\ \vdots \\ d\rho_n \end{bmatrix}, \tag{9.81}$$

where n is the number of SVs in view. A thorough derivation is provided by Kaplan [32], the result is

$$d\boldsymbol{x} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T d\boldsymbol{\rho}. \tag{9.82}$$

We are interested in the 2nd-order statistics and look for the covariance of $d\boldsymbol{x}$

$$\mathrm{cov}(d\boldsymbol{x}) = E\left\{d\boldsymbol{x}\, d\boldsymbol{x}^T\right\}. \tag{9.83}$$

Using Eq. (9.82) we get

$$\begin{aligned}\text{cov}(d\boldsymbol{x}) &= E\left\{(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T d\boldsymbol{\rho}\, d\boldsymbol{\rho}^T\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\right\}\\ &= (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\text{cov}(d\boldsymbol{\rho})\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}.\end{aligned} \tag{9.84}$$

Assuming independence of the pseudorange errors and equal size of the component errors, hence

$$\text{cov}(d\boldsymbol{\rho}) = \boldsymbol{I}_{n\times n}\sigma^2_{\text{UERE}}, \tag{9.85}$$

where σ^2_{UERE} is the pseudorange error variance and UERE stands for user equivalent range error, we may therefore write

$$\begin{aligned}\text{cov}(d\boldsymbol{x}) &= (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\sigma^2_{\text{UERE}}\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\\ &= (\boldsymbol{H}^T\boldsymbol{H})^{-1}\sigma^2_{\text{UERE}}\boldsymbol{H}^T\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\\ &= (\boldsymbol{H}^T\boldsymbol{H})^{-1}\sigma^2_{\text{UERE}}\\ &= \begin{bmatrix} D_{11} & D_{12} & D_{13} & D_{14}\\ D_{21} & D_{22} & D_{33} & D_{24}\\ D_{31} & D_{32} & D_{33} & D_{34}\\ D_{41} & D_{42} & D_{43} & D_{44}\end{bmatrix}\sigma^2_{\text{UERE}}.\end{aligned} \tag{9.86}$$

We are interested mainly in the diagonal components of the covariance matrix

$$\text{cov}(d\boldsymbol{x}) = \begin{bmatrix} \sigma^2_{x_\text{u}} & \sigma^2_{x_\text{u}y_\text{u}} & \sigma^2_{x_\text{u}z_\text{u}} & \sigma^2_{x_\text{u}c\tau}\\ \sigma^2_{y_\text{u}x_\text{u}} & \sigma^2_{y_\text{u}} & \sigma^2_{y_\text{u}z_\text{u}} & \sigma^2_{y_\text{u}c\tau}\\ \sigma^2_{z_\text{u}x_\text{u}} & \sigma^2_{z_\text{u}y_\text{u}} & \sigma^2_{z_\text{u}} & \sigma^2_{z_\text{u}c\tau}\\ \sigma^2_{c\tau x_\text{u}} & \sigma^2_{c\tau y_\text{u}} & \sigma^2_{c\tau z_\text{u}} & \sigma^2_{c\tau}\end{bmatrix}. \tag{9.87}$$

DOP (dilution of precision) parameters are defined in terms of the ratio of different combinations of these diagonal components and σ^2_{UERE}. The most important of these parameters is the geometric dilution of precision (GDOP), which is defined by

$$\text{GDOP} \triangleq \frac{\sqrt{\sigma^2_{x_\text{u}} + \sigma^2_{y_\text{u}} + \sigma^2_{z_\text{u}} + \sigma^2_{c\tau}}}{\sigma_{\text{UERE}}} = \sqrt{D_{11} + D_{22} + D_{33} + D_{44}}. \tag{9.88}$$

Other DOP parameters are the position (3D) dilution of precision

$$\text{PDOP} \triangleq \frac{\sqrt{\sigma^2_{x_\text{u}} + \sigma^2_{y_\text{u}} + \sigma^2_{z_\text{u}}}}{\sigma_{\text{UERE}}} = \sqrt{D_{11} + D_{22} + D_{33}}, \tag{9.89}$$

the horizontal dilution of precision

$$\text{HDOP} \triangleq \frac{\sqrt{\sigma^2_{x_\text{u}} + \sigma^2_{y_\text{u}}}}{\sigma_{\text{UERE}}} = \sqrt{D_{11} + D_{22}}, \tag{9.90}$$

the vertical dilution of precision

$$\text{VDOP} \triangleq \frac{\sqrt{\sigma^2_{z_\text{u}}}}{\sigma_{\text{UERE}}} = \sqrt{D_{33}}, \tag{9.91}$$

and the time dilution of precision

$$\text{TDOP} \triangleq \frac{\sqrt{\sigma_{c\tau}^2}}{\sigma_{\text{UERE}}} = \sqrt{D_{44}}. \tag{9.92}$$

Since the DOP values are derived from matrix $\boldsymbol{H}$ containing normalized vectors to the satellites used in the navigation solution, these values are heavily depending on the SVs that can be viewed at any point in time. A situation where the SVs are spread in all directions is preferrable to one where all the satellites are in the same sector. This can be easily seen by considering a two-dimensional example, where the user is situated in the origin of a (x, y)-coordinate system. Let us consider two satellites, one sitting at $(0, 1)$, the other at $(1, 0)$. We assume that we are close to the real positions with our estimations, hence $\hat{x}_{\text{u}} = x_{\text{u}}$, $\hat{y}_{\text{u}} = y_{\text{u}}$ and $\hat{r}_k = r_k$. We then have a 2 × 2 matrix

$$\boldsymbol{H} = \begin{bmatrix} \frac{x_1 - x_{\text{u}}}{r_1} & \frac{y_1 - y_{\text{u}}}{r_1} \\ \frac{x_2 - x_{\text{u}}}{r_2} & \frac{y_2 - y_{\text{u}}}{r_2} \end{bmatrix}. \tag{9.93}$$

Since both satellites sit on the unit circle, we have $r_1 = r_2 = 1$ and with $x_{\text{u}} = y_{\text{u}} = 0$ we have

$$\boldsymbol{H} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{9.94}$$

Furthermore

$$(\boldsymbol{H}^T\boldsymbol{H})^{-1} = \left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \right)^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{9.95}$$

and

$$\text{HDOP} = \sqrt{1+1} = \sqrt{2}. \tag{9.96}$$

This is the best case, since the two satellites are seen by the user under a right angle. If the second satellite is close to the first one (still on the unit circle but only separated by an angle shift of α) we have

$$\boldsymbol{H} = \begin{bmatrix} 0 & 1 \\ \sin\alpha & \cos\alpha \end{bmatrix}. \tag{9.97}$$

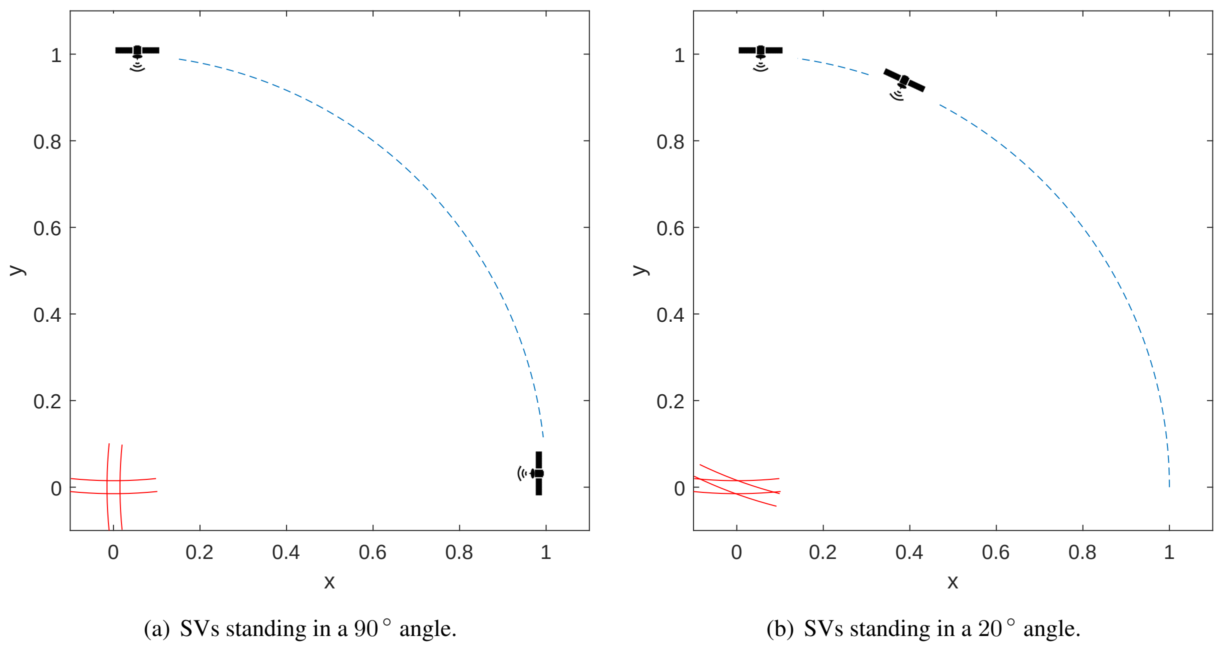
Furthermore

$$\begin{aligned}(\boldsymbol{H}^T\boldsymbol{H})^{-1} &= \left(\begin{bmatrix} 0 & \sin\alpha \\ 1 & \cos\alpha \end{bmatrix} \begin{bmatrix} 0 & 1 \\ \sin\alpha & \cos\alpha \end{bmatrix} \right)^{-1} = \begin{bmatrix} \sin^2\alpha & \sin\alpha\cos\alpha \\ \sin\alpha\cos\alpha & 1+\cos^2\alpha \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \frac{1+\cos^2\alpha}{\sin^2\alpha} & -\frac{\cos\alpha}{\sin\alpha} \\ -\frac{\cos\alpha}{\sin\alpha} & 1 \end{bmatrix} \end{aligned} \tag{9.98}$$

and

$$\text{HDOP} = \sqrt{\frac{1+\cos^2\alpha}{\sin^2\alpha} + 1} = \frac{\sqrt{2}}{|\sin\alpha|}. \tag{9.99}$$

For small values of α, i.e., when the two satellites are close together, we get quickly large values of HDOP. Figs. 9.43(a) and 9.43(b) show in a two-dimensional example with two SVs standing in a 90° and 20° angle, respectively, how the HDOP values increases from $\sqrt{2}$ to $\sqrt{2}/\sin 20^\circ = 4.13$. Fig. 9.43 also shows



(a) SVs standing in a 90 ° angle.

(b) SVs standing in a 20 ° angle.

Figure 9.43 Setup for a simple situation with two SVs and a user in the origin.

nicely that the tolerance area of the estimated user position around the origin is proportional to the respective HDOP value.

It is further illustrative to check if a third satellite would help in this situation. If the third SV has an angle of β relative to the first SV sitting at the coordinates (0,1) we have

$$\boldsymbol{H} = \begin{bmatrix} 0 & 1 \\ \sin\alpha & \cos\alpha \\ \sin\beta & \cos\beta \end{bmatrix} . \tag{9.100}$$

After some matrix operations we find

$$\text{HDOP} = \sqrt{\frac{3}{\sin^2\alpha + \sin^2\beta + \sin^2(\beta-\alpha)}} . \tag{9.101}$$

Numerically, if we insert a third SV between the two SVs of Fig. 9.43(b), hence $\beta = 10\,°$, we get an HDOP value of 4.11, only a marginal improvement of the previous value of 4.13. If, however, the third SV is outside the other two, e.g., at $\beta = 45\,°$, the HDOP value drops to 1.94, still worse than the situation of Fig. 9.43(a). The conclusion here is that two SVs covering a larger angle are better than three SVs covering only half of the angle.

9.11 Extensions

There are many extensions to the basic functions of GPS. Some are listed below:

- Dual-frequency receiver (or even triple-frequency receiver)
- DGPS (differential GPS)

- SBAS (satellite based augmentation systems)
- weak-signal reception
- carrier-phase estimation
- dead reckoning
- location-based services (LBS)
- emergency calls

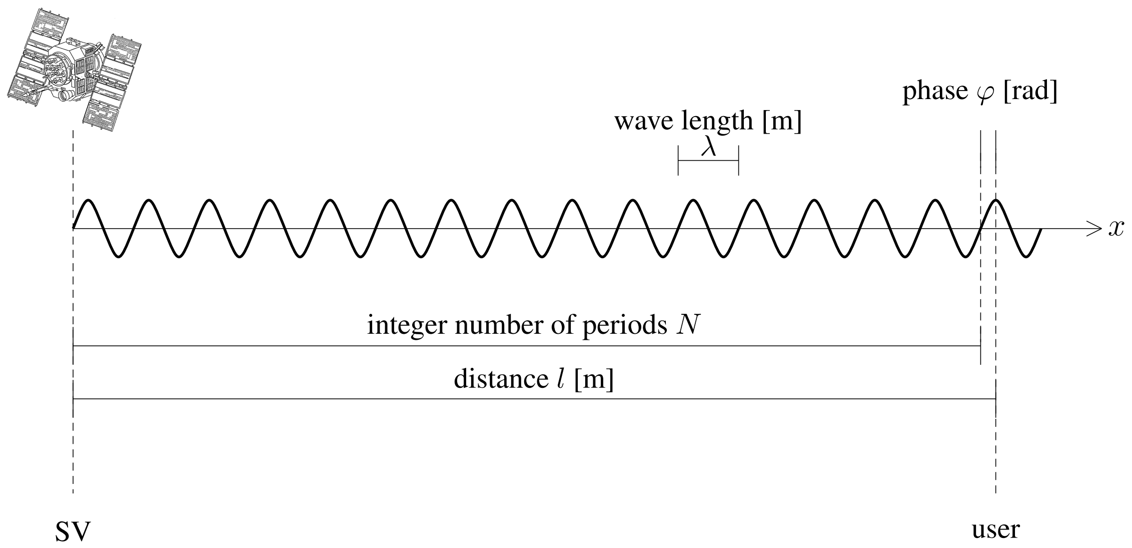


Figure 9.44 Carrier-phase estimation of a GNSS signal.

As an example, carrier-phase estimation is illustrated in Fig. 9.44.

9.11.1 Weak-signal reception

Conventional GPS receivers are designed for the reception of SV signals that are around -130 dBm, which corresponds to an SNR of around -20 dB, a level achieved in the line-of-sight (LOS) of a satellite. Locations under a rooftop or even modest foliage coverage attenuate the signal considerably. The GPS receiver is based on the extraction of the signal buried in the noise using a correlation with a pseudo-random noise (PRN) sequence. The signal can be amplified through longer correlation, so that, theoretically, a signal at any level could be detected by simply correlating long enough. In practice, there are limits to how long a signal can be coherently integrated (correlated). These limits are given by the data bit duration, frequency offset, speed and acceleration of the user.

9.12 Some applications of GNSS

Whereas originally GPS technology was deployed mainly by military and geosurveying related systems, today the applications in the commercial market are numerous.

There are dedicated handsets that give positions only, used mainly for hikers and other outdoor-attracted people. Besides position and time, they usually display speed and allow the setting and tracking of way points to follow a predefined route.

Navigation systems are standard in many middle-class and upper-class cars. A combination of GNSS and other sensors, e.g., incremental pulses make positioning in covered park decks possible. These aiding techniques to GNSS are known as *dead reckoning*. Since they are incremental measuring devices, such as gyros, they experience a lot of drift. In combination with GNSS, they can be recalibrated whenever satellite coverage is available.

In some places, the position is not important or might be known (permanent locations), but an accurate time reference would be important. GNSS might still be used in such places, providing a time reference as accurate as a few ns using C/A code tracking only.

Today, GNSS is included in most smartphones. This may, on the one hand, give the user position information, but also enables additional revenue for network providers due to so-called location-based services (LBS), such as information on restaurants in the neighborhood, current discounts in shops nearby and the like.

9.13 Other Positioning Standards

9.13.1 GLONASS

The Russian equivalent to GPS is GLONASS (*Global Navigation Satellite System*). Rather than different spreading sequences, GLONASS uses the same spreading sequence but different frequencies for each SV. These frequencies are

$$f = 1602 + 0.5625\,n \quad \text{[MHz]}, \tag{9.102}$$

where n represents the channel number. In essence, the access scheme is thus rather FDMA than CDMA, although spreading is still used, but not to distinguish channels. The spreading code used is a 9-bit maximal-length code with a rate of 0.511 Mchips/s. With a code length of $2^9 - 1 = 511$ chips, the repeat rate becomes 1 ms. The navigation message and many other aspects are very similar to GPS.

The full GLONASS constellation of 24 operational satellites was established in 1995, two years after GPS. Unfortunately, it could not be maintained in the following years due to lacking resources and because the first generation of satellites had a lifespan of only three years. The constellation was completed again in 2011 with satellites of the second and third generation that have lifespans of seven and ten years.

9.13.2 Galileo

In spring 2002, the European Parliament has agreed to finance an independent European satellite-based navigation system, called *Galileo*. The plan so far is to have an operating system by 2008. Currently, even the Chinese are interested in financing and developing such a system. Together with the Americans, who are naturally interested in an undegraded GPS performance, a lot of effort has been put into coordinating the use of spectrum and mutual interference of the systems. As can be seen from Table 9.4, many of the frequencies used must be shared with GPS.

Carrier	Center freq. [MHz]
E5a	1176.45
E5b	1196.91–1207.14
E6	1278.75
E2-L1-E1	1575.42

Table 9.4 Frequencies used by the Galileo system.

From the Americans' point of view, it is not so much the performance degradation that a possible overlay of the Galileo signal causes to the new military M-signals of GPS, but their inability to jam non-GPS navigation signals without jamming their own signals, which is of concern.

Many different services and precisions are provided by different data and code formats. The ultimate spreading sequences have not been finalized, but the lowest chip rate is expected to be 2.046 Mchips/s. Some of the services that are planned for Galileo are:

- GPS compatible mode (free of charge)
- service with enhanced security
- rescue service for naval and aviatic purpose
- service for rescue organizations
- encrypted service for police and military

9.13.3 BeiDou

Similarly to the respective global navigation satellite systems operated by the US, by Russia, and by Europe, there is a system run by China. For an intermediate generation, the system was also named COMPASS. The first satellite was launched in 2000. BeiDou's CDMA-based signal is similar to the other GNSS systems and the frequencies occupied are also in the L band.

9.14 Wireless-Based Geolocation Techniques

Wireless-based geolocation[4] techniques are subject of current investigations and research. It is, of course, much cheaper if positioning is possible without relying on GPS.

All systems can be grouped into one of two approaches: self-positioning systems (the handset computes its position without help from the system) and remote-positioning systems (where, e.g., only pseudoranges are computed in the handset and handed back to the system). The former systems have the advantage that they are user-friendly, since a position fix will be granted at a user's will. The equipment, however, might be more expensive than that of a remote-positioning systems. The latter system is interesting for cellphone operators since it adds revenue to the system and allows the operator to know where the users are.

In principle, the following parameters might be measured to estimate a user position:

- cell information only

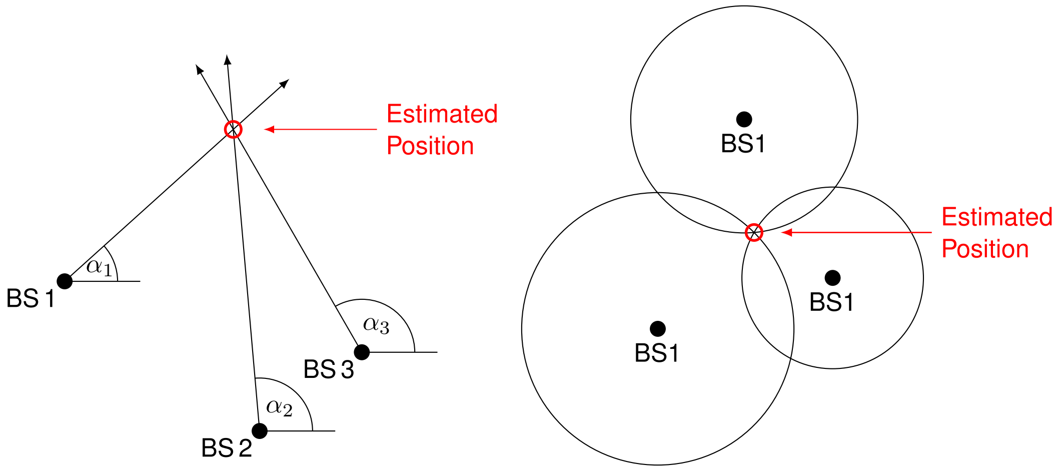
[4] Locationing is often referred to for a two-dimensional solution, wheras positioning usually refers to the full three-dimensional determination.

- signal strength (RSSI)
- angle of arrival (AOA)
- time of arrival or time difference of arrival (TOA and TDOA)

Strictly speaking, using cell information only does not require any measuring of signals. The information broadcast or any specific information to the mobile station is simply interpreted. The location accuracy is, of course, entirely dependent on the size of the cell and my be up to many kilometers in rural areas. If picocells are used in shopping malls, such information might be accurate enough for location-based services such as advertising.

Using a propagation model allows one to infer a distance of a mobile user from the base station. However, methods employing signal-strength measurements are not very reliable due to the mobile channel being governed by different fading mechanisms.

Angle-of-arrival techniques, see Fig. 9.45(a), are based on the principle of human hearing. Similar to our ears detecting slight phase differences in the arriving sound in the left and the right ear, an antenna array can estimate the direction of the incoming wave front. Such approaches are usually impractical due to the necessity of several antennas.



(a) angle-or-arrival (AOA)

(b) time-of-arriveal (TOA)

Figure 9.45 Principles of angle-of-arrival (AOA) and time-of-arrival (TOA) location determination.

Measuring the arrival time, shown in Fig. 9.45(b), or difference of the arrival times (see Fig. 9.46) of several signals from different places, it is possible to estimate the user position, much in the same way as it is done in GPS.

The standardization body of GSM (ETSI) has foreseen the possibility to broadcast information for A-GPS. The SMS cell broadcast information relevant to positioning is divided into four main categories:

- E-OTD data
- DGPS
- Ephemeris and clock correction
- Almanac

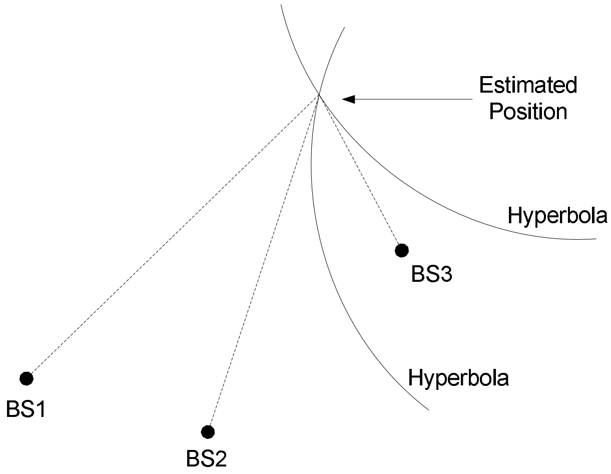


Figure 9.46 Principle of time-difference-of-arrival location determination.

The “enhanced observed time differences” method measures the time-of-arrival differences of a signal sent by the BTS between the MS and another reference station, the so-called Location Measurement Unit (LMU). Strictly speaking, at least three such measurement sets are needed to compute a position.

DGPS data (differential GPS) has lost some of its meaning on the grounds of the SA (selective availability) being switched off in the year 2000.

The last two points concern the payload of the GPS signal. A receiver may, instead of having to download the navigation message from the SVs, receive the same information from a cellular broadcast service.

Future wireless standards further support positioning techniques through increased broadcast of positioning information and the synchronization of the time bases of the different base stations, which is, for example, not done for GSM.

9.15 Further Literature

Good tutorial-style books are the ones by Kaplan [32], Parkinson and Spilker [49], Tsui [79]. Lakafosis and Tentzeris cover wireless localization in general in [36]. Furthermore, the IEEE has devoted one of its Proceedings [18] entirely to articles concerning GPS.

10 Mobile Communication Systems

10.1 Introduction

Thanks to the maturity of integrated circuit technology, mobile communication equipment has become cheap and small consumer products. Cellular telephone handsets are ubiquitous these days. Another factor that played a major role in the success story of cellular phones is the standardization effort that makes GSM compatible almost throughout the world. What once was planned as a European standard has long become accepted in major cities throughout the United States and Asia. In fact, the success of the digital standard GSM has created such high hopes in future revenue that some countries (e.g., U.K. and Germany) have raised ridiculous amounts of money in the frequency band auctions for UMTS, the next digital standard, such that many analysts fear the investment may never pay back.

Despite the fact that standards are ever changing, we shall try to throw some light on important standards in this chapter. Some fundamentals are valid throughout future standards and some standards may live longer than it was once anticipated.

The evolution of mobile communication systems in Switzerland, Europe and the USA is illustrated in Fig. 10.1. In the following, the generations and underlying motivations of these systems are discussed briefly, before a few standards are covered in more detail.

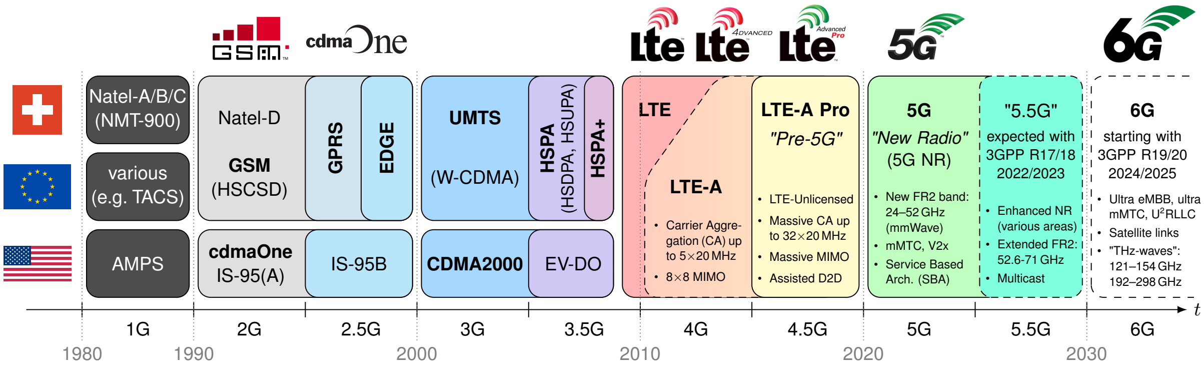


Figure 10.1 Evolution paths of the cellular mobile communication technologies[1] (generations).

[1] eCPRI stands for *Evolved Common Public Radio Interface* and denotes the highly desired standard interface to all communication systems allowing direct interconnections between systems of different manufacturers (e.g. Ericsson, Nokia, Huawei, and NEC), which is a fundamental requirement for future broadband communication technologies. The terms NaaS and XaaS stand for *Network as a Service* and *Everything as a Service*, respectively, and referring to high-level service models for general connectivity. NaaS involves the optimization of resource allocations by considering network and computing resources as a unified whole; XaaS goes one step further and also incorporates infrastructure and software (thus, *everything*) as parts of the unified system. What these ideas actually entail is still quite fuzzy and will have to be defined as these evolved communication standards are developed.

10.2 Overview of System Generations

10.2.1 1st-Generation Systems

1st-generation mobile telephony systems were not standardized beyond country borders. It is therefore difficult to give an exhaustive list of all systems ever in use. In Europe alone, there were around six systems in use. The map in Fig. 10.2 illustrates the situation at that time. We will therefore concentrate on the most important systems and on the systems in use in Switzerland.

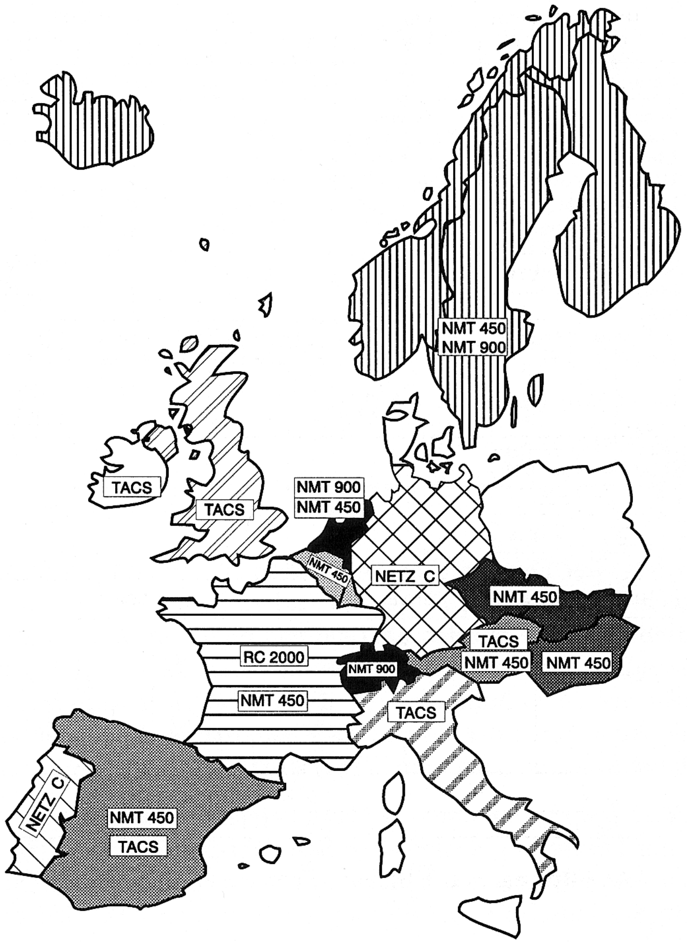


Figure 10.2 European map of 1G cellular standards.

1G systems were analog, meaning that the modulation format was some kind of FM. Although systems existed that were not cellular in nature, e.g., the Mobile Telephone System (MTS) in the 150 MHz band as early as 1946, the cellular concept was conceived early by Bell Labs in 1949. Using the same name convention, the system was named Improved Mobile Telephone System (IMTS). The main motivation for this move was the increase in capacity. No handover processes were possible with IMTS. However, it was not until late in the seventies, when the first trials with the *Advanced Mobile Phone Service* (AMPS) began. The commercial service of AMPS started 1983.

In Switzerland the cellular systems belonging to Swisscom (former PTT) are called *Natel*, which originally stood for *Nationales Autotelefon*, but is commonly associated with just *Nationales Telefon*. In the first-generation round, there were three different systems:

- Natel-A, introduced in 1978, fully operational by 1980 and eventually switched off in 1995. The capacity was limited to a few thousand customers, phone calls were restricted to a maximum duration of three minutes and only 32 calls could be carried simultaneously over the entire network. The air interface was on 150 MHz using analog phase modulation (PM). Even though the required equipment was bulky, heavy (about 12 kg) and expensive (beyond 10 000 CHF for the phone and monthly costs of 130 CHF), the PTT had greatly underestimated the demand and the limitation on the number of customers proved too restrictive, particularly in the Zurich area. Eventually, the network would be so overloaded, that PTT had to stop handing out new numbers all together, in 1980.

- Natel-B was launched as second network to extend the capacity of the overloaded Natel-A in 1980, using a similar air interface. It was fully operational by 1984 and eventually turned off in 1997.

- Natel-C was launched as the last of the analog generation networks in Switzerland in 1987. The main goals were to increase capacity from about 10k to over 250k subscribers, decrease monthly fees to below 100 CHF, and make user equipment available at below 5000 CHF. For the third network, PTT adopted the international standard NMT-900, providing 600 FDD channels in the frequency band of 890–960 MHz and using narrowband FM (2 × 25 kHz). By 1990, it surpassed 100k subscribers. Roaming was possible with Denmark, Sweden, Norway, Finland, and the Netherlands. It remained a long-lasting mobile phone solution and was eventually turned of in 1999.

None of these 1G systems continued operation in Switzerland beyond the year 2000.

10.2.2 2nd-Generation Systems

Usually, systems belonging to the first generation of mobile communication systems that employed digital modulation schemes are called 2G systems. In Switzerland, the 2G system *Natel-D* started operation in 1993. Although its name was chosen as a logical progression following its three analog predecessors Natel-A, -B, and -C, the standard behind it had always been GSM, *Global System for Mobile communications*, the pan-European cellular system, established by ETSI, the European Telecommunications Standards Institute, covered in detail in Section 10.4.

Extensions to GSM and other 2G systems are often called 2.5G systems, as they bridge from 2G to 3G systems, see Fig. 10.1 The main difference from standard 2G to 2.5G and beyond is that, before 2.5G, all traffic (mostly voice at first, but eventually more importantly data) was basically routed through wireless telephone lines. However, occupying an entire phone line to exchange small amounts of data is very inefficient (and hence expensive for the customer). Thus, all extensions to the 2G systems would move from so-called *circuit-switched data* (a circuit being a telephone line) to *packet-switched data*, which just exchanges data packets when needed and releases link for other customers otherwise. For GSM, these packet-based extensions are GPRS, *General Packet Radio Service*, and from 2003 on EDGE, *Enhanced Data Rates for GSM Evolution*, which is also sometimes called 2.75G. EDGE allowed data rates of 59.2 kb/s per time slot, theoretically allowing 473.6 kb/s when allocating 8 time slots for a single user. It was the first truly fast cellular data service; however, due to the technical difficulties it was rarely implemented to the full extent and the still very large end-to-end latency of 100 ms and beyond, often dampened the user experience considerably.

GSM networks eventually existed in more than 200 countries — but 2G systems were not all based on GSM, the most important exceptions are the North American continent (US and Canada) and Asian countries such as Japan and South Korea. The most important of other second-generation standards were:

- IS-54 and IS-136, together better known as *Digital AMPS* (D-AMPS), were the 2G successor to the 1G system AMPS, and started to be incorporated into the preexisting network in 1990. In essence, it built on the AMPS channels, which were based on FDMA, and added time-slotting, thereby making it into a TDMA standard similar to GSM. Further, it added digital compression and encryption (IS-54) as well as text messaging and circuit-switched data (IS-136). D-AMPS was mainly used in North America, but has been faded out by 2007-2009.
- IS-95, standing for *Interim Standard 95* (numbered according to its release year 1995), but better known by its proprietary name *cdmaOne* (see Fig. 10.1), was the first ever CDMA-based digital cellular technology and the 2G mobile communication system in the USA, Canada, Japan and a few other countries. It was developed by Qualcomm and later adopted as a standard by the Telecommunications Industry Association (TIA) in 1995. Similar to GSM in Europe, it was used as the the default wireless system in the 800 and 900 MHz bands, with 1.25 MHz wide subbands. To obtain this bandwidth, a 307.2 kb/s bitstream is exored with a PRN code running at 1.228 Mb/s and a period length of $2^{42} - 1$.
- Personal Digital Cellular (PDC), originally known as Japanese Digital Cellular (JDC), is a 2G standard that basically implements a physical layer quite similar to D-AMPS (with 6 TDMA time-slots) but its protocol stack resembles GSM. Defined in April 1991, PDC had quickly found widespread use in in Japan, peaking at 80 M subscribers (in a country of about 127 M people) of a single provider (NTT DoCoMo) by 2003. However, by 2005 the number had already dropped by almost half, with people adopting newer 3G standards very rapidly.

Some key parameters of some standards are collected in Table 10.1.

Standard	Access methods	Symbol rate	Modulation scheme	Bandwidth	Tx filter
GSM	TDMA/FDMA	270.833 kS/s	GMSK	200 kHz	Gauss. pulse, BT $= 0.3$
EDGE	TDMA/FDMA	270.833 kS/s	$3\pi/8$ 8-PSK	200 kHz	Gauss. pulse, BT $= 0.3$
D-AMPS	TDMA/FDMA	24.4 kS/s	$\pi/4$ DQPSK	30 kHz	RRC, $\rho = 0.5$
IS-95	CDMA/FDMA	0.6144 MS/s	QPSK	1.25 MHz	RRC, $\rho = 0.35$
PDC	TDMA/FDMA	21 kS/s	$\pi/4$ DQPSK	25 kHz	RRC, $\rho = 0.5$

Table 10.1 Modulation parameters of important 2G standards.

10.2.3 3rd-Generation Systems

Some evolution paths from 2G into 3G standards are shown in Fig. 10.1. The most prominent example of 3rd-generation systems in Europe is UMTS, an idea launched by the European Commission in 1989. UMTS stands for *Universal Mobile Telecommunications System* and is covered in detail in Section 10.5. Originally, the aim of the International Telecommunications Union (ITU) was to develop one single standard within the IMT-2000 project. This was not possible for political and technical reasons. Since different 2G systems existed, the actual migration scenarios were different depending on the region of the world.

In a parallel evolution path to UMTS in Europe, IS-2000, better known as *CDMA2000*, eventually took over from IS-95 as the 3G standard in North America and South Korea. CDMA2000 combined CDMA and TDMA techniques for enhanced data rates (then called *Evolution-Data Optimized*, EV-DO) and has been adopted with similar if not greater acceptance as UMTS (or HSPA, its evolved packet access for high-speed data rates) in Europe.

10.2.4 4th-Generation Systems

Motivated by the fast acceptance of the high-performance 3rd-generation systems, the next generations were planned to encompass even larger user bases. One of the most important steps was when the planned fourth-generation (4G) successor UMB (Ultra Mobile Broadband) for North America was abandoned by Qualcom in 2008. This opened the way for LTE, *Long-Term Evolution*[2], covered in more detail in Section 10.6, to become the first global communication standard.

Thus, one of the main goals of the 4G standard was the complete 'integration of the wireless world', a target enabled by software radio concepts. The trends followed by this technology are:

- higher data rates (of course)
- integration of different standards, always best connected (ABC)
- ad-hoc networks
- multihop networks
- MIMO (NLOS)
- OFDM

Initially, the plans to have such a revolutionizing and powerful network ready for launch around 2010 seemed rather far-fetched — but it all worked out rather smoothly in the end.

10.2.5 5th-Generation Systems

While it is currently unclear as to how fifth-generation mobile networks will be different from 4G networks such as LTE, it is strongly believed that new spectrum resources will be incorporate. The evergrowing demand for capacity, can no longer be satisfied by just closing the gap to the Shannon limit with sophisticated signal processing; instead, it pushes into exploring new regions in the spectrum with higher bandwidths, such as bands at 28 GHz, 38 GHz, and also as high as 60 GHz and 73 GHz, previously thought unfit for wireless transmission. Short latency is believed to be a key issue, e.g., vehicle-to-vehicle communication for self-driving cars. Together with high data volumes in other applications it will be important to get servers close to the user.

Non-orthogonal multiple access (NOMA) is believed to play a certain role, which allows to exploit bandwidth even more by allowing several users with different power on the same channel. The stronger signals can easily be detected in a conventional way (the weaker signal is treated like noise). Successive subtraction of stronger signals from the composite signals allows the detection of increasingly weaker signals.

Another key property of systems of the fifth generation and beyond is the fully compatible interoperability of the network hardware of different vendors. An *Evolved Common Public Radio Interface*, or *eCPRI* for short, is defined, denoting the highly desired standard interface to all communication systems which allows direct interconnections between systems of different manufacturers and opens the way for a single wireless network for all users.

[2] Sometimes LTE is also referred to as a 3.9G system, rather than a proper 4G standard, as it does not meet all the requirements originally required for a 4G declaration by the IMT.

From the feature point of view, 1G and 2G were designed with voice as being the main application. 2.5G was designed for packet-switched data and 3G supports multimedia applications. 4G is a combination of many services. For 5G, certainly also communication between vehicles will play an important role. Where as the main goal of 4G was on providing mobile internet, the focus of 5G is said to be interconnectivity, connecting different technologies and devices/users (while still providing even higher data rates and lower latencies).

10.3 The Cellular Concept

10.3.1 Introduction

Unlike, for example, wide-area paging systems, which were popular through the 80s and early 90s, modern communication systems are cell based, an idea that was already around in 1949. The signal is no longer distributed to the end user in a country-wide manner, but only to the cell the user is currently registered in. This way the capacity of a network can be immensely enlarged through so-called frequency reuse.

The reason for the hexagonal form of the cells can be found by looking into different grids on which the base-station antennas are placed. Rousseau and Saint-Aubin [62] argue that from all the choices of placing antennas on a regular triangular grid, a square grid, and a hexagonal grid, respectively, the first one results in the lowest antenna density per area covered, given a maximum distance of a mobile from the nearest base-station. Hence, if we draw the triangular grid, see Fig. 10.3 (a), we can divide the plane into regions of points that have the shortest distance to a certain base station. The well-known hexagonal regions for the cells can now be seen in Fig. 10.3 (b).

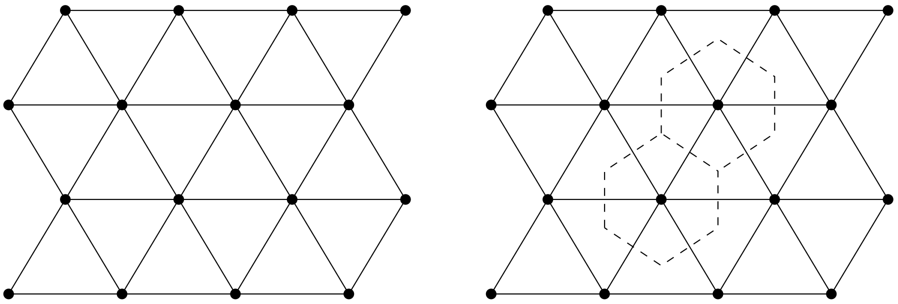


Figure 10.3 Triangular grid (left) and resulting regions for points closest to a certain base station (right). The black dots represent the base-station antennas.

Although the situation in reality is often determined by the actual landscape and therefore possibly complicated, a commonly established method is to divide a plane into hexagonal cells as can be seen in Fig. 10.4.

10.3.2 Cluster Size

In a cellular system, the neighbor cells have to use different frequencies, otherwise a mobile station sitting on the cell border receives two different base stations with the same power on a given channel. Still worse, base stations transmitting on the same frequency need to be at a certain distance in order not to interfere too much. This interference is called *cochannel interference*. The modulation scheme essentially determines

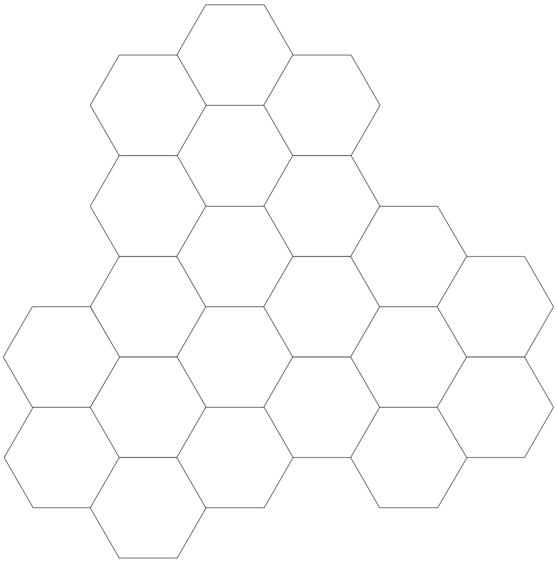


Figure 10.4 Hexagonal cells comprising a cellular wireless system.

the normalized distance q, which is also referred to as the reuse factor. The reuse factor, in turn, determines the cluster size, i.e., how many cells of different frequencies build the next largest object, the cluster. Only cluster sizes of the form

$$N = i^2 + ij + j^2, \tag{10.1}$$

where i and j are arbitrary positive integers, result in usable cluster patterns, i.e., clusters that can be tiled. To see this, we make the following lines of argument. Assume a cell of radius R, see Fig. 10.5. The cell-to-cell distance for a cluster of hexagons is given by

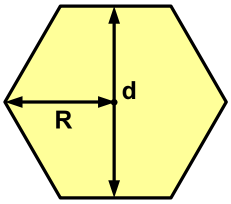


Figure 10.5 Single cell and size parameters.

$$d = \sqrt{3}R. \tag{10.2}$$

The area of a hexagon is thus given by

$$A = \frac{d}{2} \cdot R \cdot \frac{1}{2} \cdot 6 = \sqrt{3}\frac{3}{2}R^2. \tag{10.3}$$

In order to count the number of cells, we introduce a new coordinate system, as shown in Fig. 10.6, whose axes i and j are not perpendicular, but in an angle of 60°. The position of every cell can now be expressed in the coordinate pair (i, j). The distance D between two cells using the same channel in the new coordinate system, also called re-use distance, is

$$\begin{aligned} D^2 &= (id)^2 + (jd)^2 + 2(id)(jd)\cos 60^\circ \\ &= (i^2 + j^2 + ij)d^2. \end{aligned} \tag{10.4}$$

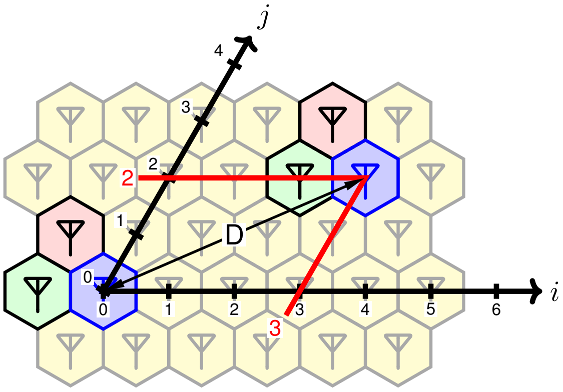


Figure 10.6 Coordinate system to arrange cells.

The condition for a group of cells to be clusterable (in an infinite tile carpet) is that it can be placed six times around one cluster (like a hexagon can be placed six times around another hexagon). In other words, the six nearest cochannel neighbors have all equal distance

$$D = \sqrt{(i^2 + ij + j^2)}\, d. \tag{10.5}$$

A new (virtual) hexagon, see Fig. 10.7, whose corners consist of the cochannel neighbors, and hence whose

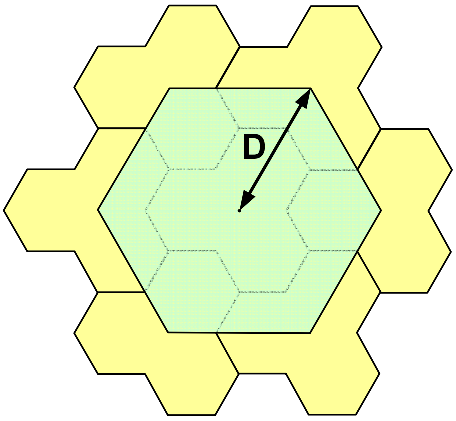


Figure 10.7 Example of virtual hexagon (dark area) to evaluate cluster area ($N = 3$).

radius is D, will contain a cluster of N cells plus one third of the cells corresponding to the six neighbor clusters, contains the area

$$A_{\text{virtual hexagon}} = \left(N + 6\,\frac{1}{3}N\right) A = 3NA. \tag{10.6}$$

Similar to the area calculation of one hexagon Eq. (10.3), we have

$$A_{\text{virtual hexagon}} = \sqrt{3}\,\frac{3}{2}D^2 = 3N\sqrt{3}\,\frac{3}{2}R^2. \tag{10.7}$$

With Eq. (10.7) it becomes clear that the *reuse factor* does not depend on the radius of the cell,

$$q = \frac{D}{R} = \sqrt{3N}. \tag{10.8}$$

10.3.3 Cochannel Interference

In a cellular system, very often not the signal-to-noise ratio is the limiting factor, but the interference from neighboring cells. If all the power of the transmitters is increased, the SNR is no longer a problem, but the cochannel interference remains the same. Fig. 10.8 illustrates this problem. Hence, since all the power levels can simply be scaled, it does not matter from an interference point of view how large the cells are (see Fig. 10.9). The interference is given by the distance of cells using the same frequencies. Thus, the cochannel interference can be estimated from the cell cluster parameters. If only the first tier of interferers is considered, the interference stems from 6 cells. For a path loss model with a power of 4, we can write for the cochannel interference (compared to a mobile station sitting at radius R from the base station)

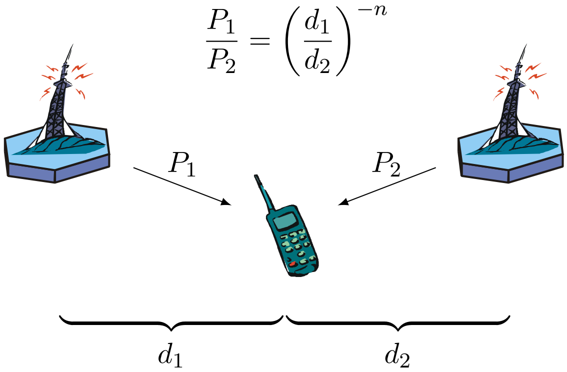


Figure 10.8 Cochannel problem. Source: [46].

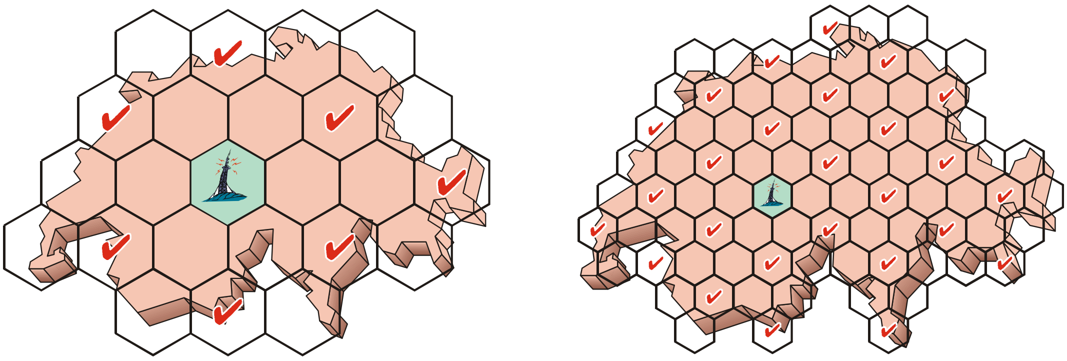


Figure 10.9 Switzerland divided up using large cells (left) and small cells (right). Source: [46].

$$C/I_{\text{lin}} = \frac{R^{-4}}{6D^{-4}} = \frac{3}{2}N^2, \tag{10.9}$$

or in terms of dB

$$C/I_{\text{dB}} = 1.76 + 20\log_{10} N. \tag{10.10}$$

Some results are listed in Table 10.2.

i	j	N	q	C/I_{pess} [dB]	C/I [dB]	C/I_{opt} [dB]
1	0	1	1.73	-13.20	1.76	9.68
1	1	3	3.00	4.26	11.30	16.30
2	0	4	3.46	7.89	13.80	18.21
2	1	7	4.58	14.39	18.66	22.09
3	0	9	5.20	17.13	20.84	23.90
2	2	12	6.00	20.18	23.34	26.02
3	1	13	6.24	21.01	24.04	26.62
4	0	16	6.93	23.14	25.84	28.19
3	2	19	7.55	24.87	27.34	29.50
4	1	21	7.94	25.87	28.20	30.27
5	0	25	8.66	27.59	29.72	31.62
3	3	27	9.00	28.34	30.39	32.22
4	2	28	9.17	28.70	30.70	32.50
5	1	31	9.64	29.69	31.59	33.30
4	3	37	10.54	31.39	33.12	34.70

Table 10.2 Estimation of cochannel interference.

Sectorization

One way to improve the C/I situation is to introduce sectorization. Sectorization means that the base station being located in the center of a cell has a directional preference towards one third of the hexagonal cell. This way, the C/I is improved by a factor of 3. The advantages of sectorization are

- C/I is improved
- cluster size may be reduced
- better use of base station equipment serving 'three' cells

The drawbacks of sectorization are

- additional hardware needed in base station (directional antenna etc.)
- increased signaling burden due to more frequent channel changes
- more frequencies needed

The logical extension to sectorization is to have narrow-beam antennas, which track the position of the mobile station as it moves through a cell.

10.3.4 Handover

When the mobile moves across cell boundaries, it has to change channels. This process is called *handover* or *handoff* and it must be performed quickly. Fast frequency synthesizers guarantee rapid change of operating frequencies, so that a user does not notice such a handover process.

The handover process is, however, a critical process. If handover happens whenever a concurrent base station appears temporarily stronger than the last serving, handover actions might occur too often, particularly when moving along cell boundaries. Some kind of hysteresis usually prevents this from happening.

10.3.5 Trunking

Obviously, the fixed assignment of a channel to every user in the system would quickly exhaust the number of available channels. Rather, a pool of common channels is shared among all users taking advantage of the fact that usually only a small percentage of all users is actively utilizing channels. This concept is called *trunking*. Since the resources are limited, there is a chance that a user requiring a channel cannot be served: the call is *blocked*. The probability of blocking determines the *grade-of-service* (GOS).

Traffic in a cellular network is measured in Erlang, in honor of a Danish mathematician, statistician, and engineer, Agner K. Erlang (1878–1929), who is renowned for "‘inventing"’ the fields of traffic engineering and queueing theory. One Erlang (1 E) is the measure of communication activity corresponding to one user making a call 100 % of the time. During busy hours, a typical cellular voice user generates around 2 mE of traffic, i.e., the user is active during 0.2 % of the time.

The Erlang-B formula determines the probability of blocking[3]. Its derivation [56] is rather lengthy and involves Markov chains, but the result is

$$\boxed{\text{Pr(blocking)} = \frac{\dfrac{A^C}{C!}}{\displaystyle\sum_{k=0}^{C} \frac{A^k}{k!}},} \tag{10.11}$$

where A is the traffic intensity in Erlang and C is the number of channels available. Fig. 10.10 shows the blocking probability as a function of traffic intensity. Note that the result corresponds in good correlation to the one for the slotted ALOHA collision probability, where we had earlier found that, taking the traffic

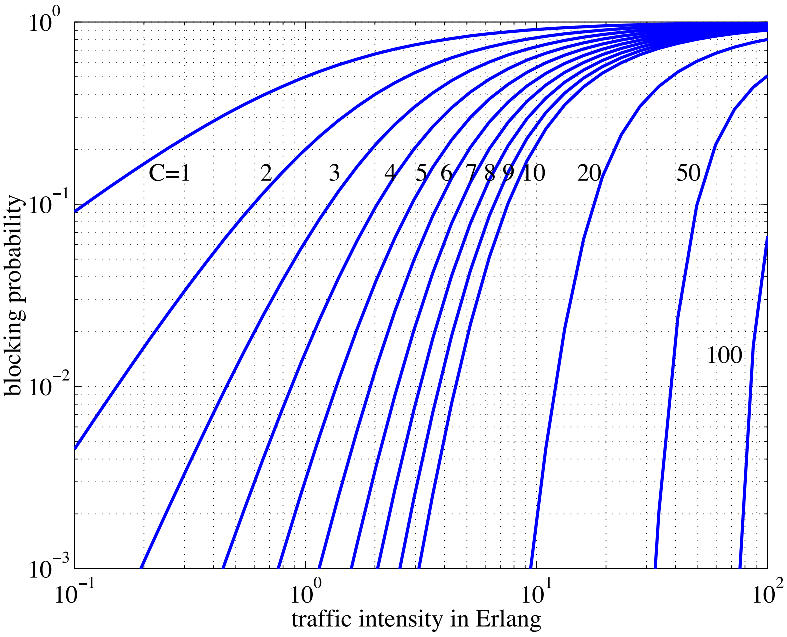


Figure 10.10 Probability of blocking as a function of traffic intensity (Erlang-B graph).

[3] Strictly speaking, it is assumed that the traffic is generated by an infinite number of users. This assumption is usually fullfilled by the fact that the number of users outnumbers the number of channels by far.

intensity as $A = N \cdot \tau/T$, where N is the number of users and the τ/T the usage ratio of one user, the probability of collisions can be expressed as the complementary probability of having no collisions

$$P_{\text{collision}} = 1 - P_{\text{success}} = 1 - \mathrm{e}^{-A}, \tag{10.12}$$

which for small A can be modelled as simply the linear term

$$P_{\text{collision}} \approx A. \tag{10.13}$$

Looking at above blocking probability we realize that in the case of one channel, i.e. $C = 1$, Eq. (10.11) can be written as

$$P_{\text{collision}} = \frac{A}{1+A} \approx A, \tag{10.14}$$

again for small A.

10.4 GSM

GSM has proved itself as a successful result of coordination and cooperation of companies and members of many countries. It has quickly found widespread acceptance, mostly due to the fact that the chip technology was at a stage where high enough integration was affordable for consumers. Whereas a few years back most wireless telephones were built into cars—at some point the word 'Natel' was clearly associated with a car-phone—today this is no longer the case. Long have cellular phones been reduced to a size that fits every pocket. Other categories of technology that had a great impact on the GSM development were RF, digital signal processing, software, control algorithms. In this chapter we shall restrict ourselves to the first two categories.

10.4.1 Short History

The abbreviation GSM originally stood for *Groupe Spéciale Mobile*, a group initiated by CEPT (Conférence Europeéne des Administrations des Postes et des Télécommunications) to address the design of a pan-European cellular standard. A short history of the GSM development is as follows:

- 1982: CEPT decides to set up pan-European standard.
- 1985: List of requirements and recommendations is adopted.
- 1987: Memorandum of Understanding (MoU) to develop and deploy a common cellular telephone system across Europe signed by telecommunication operators of 13 countries.
- 1989: ETSI (European Telecommunication Standardisation Institute) acquires GSM.
- 1990: GSM Phase 1 specifications released.
- 1991: First tests (world's first GSM call on July 1, in Finland), commercial launch.
- 1992: First SMS, first international roaming agreement (between UK and Finland).
- 1993: GSM coverage of main roads in Europe, first hand-held GSM mobile phone available.

10.4.2 Goals of GSM

Although not all goals of the pan-European cellular system to develop were clear at the beginning, some of the important targets proved to be:

- digital speech and data services
- high spectral efficiency
- compatibility to other networks (ISDN, PSTN)
- European roaming and handover
- independence of manufacturer
- small handset prices possible

10.4.3 System Parameters

The RF parameters of GSM and its derivatives are given in Table 10.3 and in Table 10.4. Fig. 10.11 shows the spectrum due to the GMSK modulation scheme. From the observation of the adjacent channel it becomes clear why the frequency allocation plan in GSM does not allow the use of adjacent channels in neighboring cells, let alone in the same cell.

	GSM850	**GSM900**	**GSM1800 (DCS)**	**GSM1900 (PCS)**
Uplink freq.[4]	824–849 MHz	890–915 MHz	1710–1785 MHz	1850–1910 MHz
Downlink freq.	869–894 MHz	935–960 MHz	1805–1880 MHz	1930–1990 MHz
Duplex separation	45 MHz	45 MHz	95 MHz	80 MHz
# RF channels	124	124	374	299
Avg. MS transmission power	250 mW	250 mW	125 mW	125 mW
Peak MS transmission power	2 W	2 W	1 W	1 W

Table 10.3 RF parameters of GSM and its derivatives.

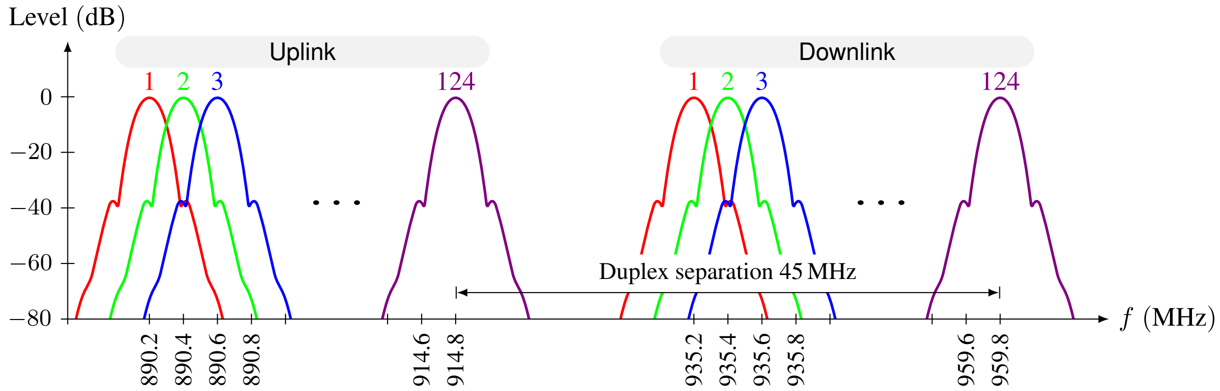


Figure 10.11 Spectra and frequency ranges of GSM 900.

For each uplink channel f_u there is a uniquely assigned downlink channel f_d using the relationship

$$f_\mathrm{u}(n) = 890\,\mathrm{MHz} + 200\,\mathrm{kHz} \cdot n, \qquad 1 \leq n \leq 124, \tag{10.15}$$

$$f_\mathrm{d}(n) = f_\mathrm{u}(n) + 45\,\mathrm{MHz}. \tag{10.16}$$

GSM is considered a TDMA system, see Fig. 10.12, despite using frequency division as a channel assignment method, too. Each RF channel is divided into 8 time slots, which together build one frame. One user normally only transmits once per frame and receives once per frame. The numbering of the downlink and uplink channels (in time) is shifted to prevent simultaneous transmission and reception in mobile stations, see Fig. 10.13. The length of slots and frames are given as

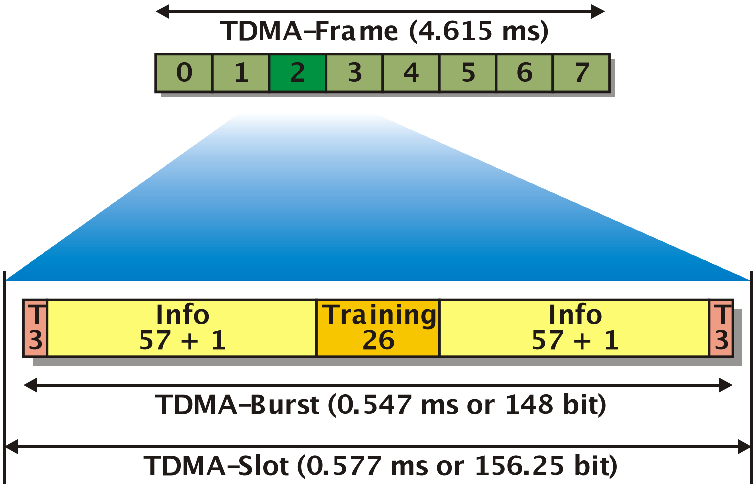
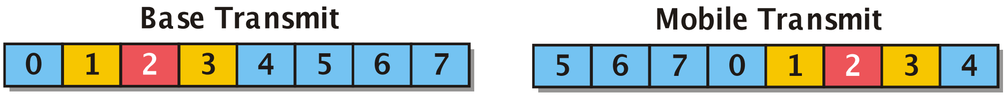

TDMA-Frame (4.615 ms)
0
1
2
3
4
5
6
7
T
3
Info
57 + 1
Training
26
Info
57 + 1
T
3
TDMA-Burst (0.547 ms or 148 bit)
TDMA-Slot (0.577 ms or 156.25 bit)

Figure 10.12 TDMA structure of GSM. Source: [46].

$$t_S = \frac{15}{26}\,\mathrm{ms} = 577\,\mu\mathrm{s}, \tag{10.17}$$

$$t_F = 8t_S = \frac{120}{26}\,\mathrm{ms} = 4.615\,\mathrm{ms}. \tag{10.18}$$



(a) Downlink

(b) Uplink

Figure 10.13 Signaling in different time slots for downlink and uplink. Source: [46].

10.4.4 Infrastructure

GSM is a cellular system. Even calls between handsets located in very close proximity take a detour to basestations and the corresponding infrastructure. An overview of this infrastructure is given in Fig. 10.14. The GSM infrastructure can be divided into three major parts: the mobile station (MS), the base station subsystem (BSS), and the network and switching subsystem (NSS). Two important elements of the MS are

Name	Value
Access method	TDMA-FDD
Multiplex number	8
TDMA frame length	4.615 ms
Channel spacing	200 kHz
Symbol rate	270.833 kbps
Symbol duration	3.7 µs
Modulation type	GMSK, BT= 0.3
User data transfer rate	2.4, 4.8, or 9.6 kbps
Channel equalization	up to 16 µs

Table 10.4 Common system parameters of GSM.

the mobile equipment or the handset, which comprises the hard- and software, and the subscriber identity module (SIM). In a GSM network, the calls are directed to SIMs rather than to a mobile equipment. Thus, somebody might use one number (defined by the SIM) with different mobile equipment or vice versa, e.g., a user has one handset but two SIMs, which he uses depending on business or private purpose. The SIM can also contain short messages and short-dial numbers.

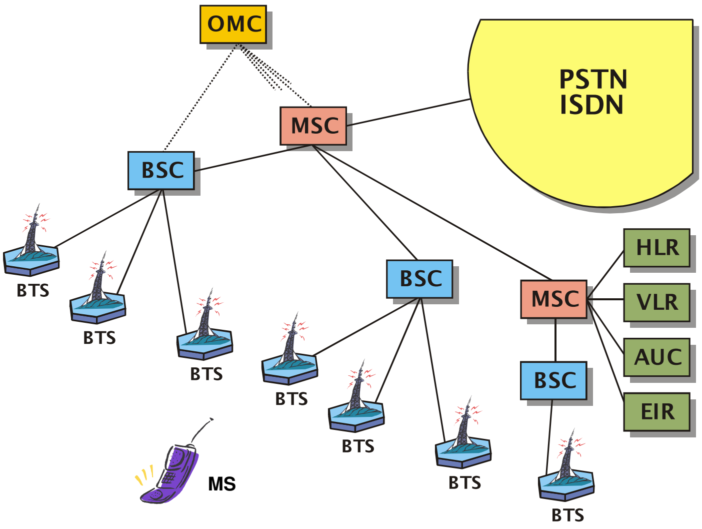


Figure 10.14 GSM infrastructure. Source: [46].

The BSS is the interface between the wireless world and the wired infrastructure. It provides all the translation between different protocols. The BSS consists of the base transceiver station (BTS) and the base station controller (BSC). Typically, several BTS are colocated at the same site (e.g., sectored cells) using a common antenna mast. BSC might be connected to BTS using either wiredline connections or microwave links. Usually, 20 to 30 BTS are controlled by a single BSC.

The NSS contains all the elements involved in the registration and authentication of users, in particular the operating and maintenance center (OMC). The mobile switching center (MSC) controls the traffic among a couple of different cells. It is connected to the home location register (HLR), the visitor location register (VLR), the authentication center (AUC), and the equipment identity register (EIR), so that the system can check if both user and equipment are legal subscribers. These elements help prevent the use of stolen or fraudulent equipment. The MSC also connects to other MSCs or the public switched telephone network (PSTN).

The HLR is a database software handling the mobile subscriber account. There is an HLR for every network, so for Switzerland there is a total of three HLRs. It stores address, service type, current location, forwarding address, authentication/ciphering keys, and billing information of a user. The address of a user is not only the ISDN telephone number, but also an international mobile subscriber identity (IMSI) number that is unique for every user.

The VLR is a temporary database holding the information of all subscribers currently being inside the coverage area of an MSC. The maintenance of these two location register databases is paramount to mobility as used in a cellular system. When switched on, every user is registered in his HLR (possibly far away from his present location) and in a VLR, which in turns informs the HLR about the location of the user, such that any incoming calls might be forwarded to the right MSC.

10.4.5 Burst and Channel Types

There are a lot of different burst types in GSM, some of which are displayed in Fig. 10.15. There are normal bursts both in uplink and downlink with a normal-length training sequence, which are used most of the time. For initial access, there are short access bursts for the uplink, since the mobile might not know the full time alignment at that point of time. Special synchronization bursts in the downlink allow the mobile station to adjust some of its parameters. Finally, the frequency correction burst in the downlink basically sends a continuous wave (CW), that allows the mobile to correct for its frequency offset compared to the base station reference.

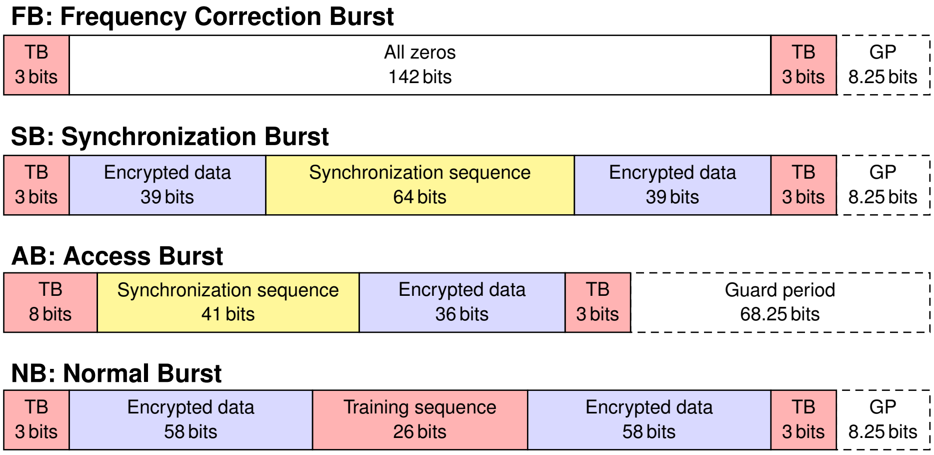


Figure 10.15 Some burst types in GSM.

It is interesting to note that GSM uses a technique called timing advance (TA) in order to ensure non-overlapping bursts arriving at the base station. The base station measure incoming access bursts of the individual mobile stations and asks them to shift the longer normal burst according to their respective distances to the base station (more distances mobiles need transmit earlier). The resolution is roughly 500 m and the assignment of the 6 bits used for this purpose is shown in Table 10.5.

TA	Distance between BS and MS
0	<550 m
1	550 – 1100 m
2	1100 – 1650 m
⋮	⋮
63	35 km

Table 10.5 Timing advance assignment used in GSM.

Many different logical channels are used in GSM. They are all mapped to one of these four burst types. There are two types of logical channels, namely traffic channels (TCH) and control channels (CCH). Let us start by briefly explaining the control channels.

Control Channels

There are three main types of control channels, the broadcast channels (BCH), the common control channels (CCCH), and the dedicated control channels (DCCH). These channel types are further divided into specific channels according to Table 10.6. To understand the signaling originating from a GSM base station, we need to have a look at the frame structure, which is displayed in Fig. 10.16. All BTS transmit broadcast channels (BCH) all the time. These BCHs are a bit like a beacon and allow the MS to orient itself. Moreover, all time slots on the channel of the BCH are transmitted with maximum power constantly. As can be seen from Fig. 10.17, there is a certain ordering of the different broadcast channels. On the downlink on certain frequencies, the time slot number 0 within each of the 51 frames inside a multiframe is alternately filled with FCCH (F), SCH (S), BCCH (B) and CCCH (C). The FCCH allows the synchronization of the MS with respect to any frequency uncertainty. The following SCH allows the MS to synchronize with respect to sampling time and frame time. It also identifies the serving base station. Cell and network identity are broadcast using the BCCH, which also transmits a list of channels currently in use within the cell. Besides these three types of broadcast channels, CCCHs are transmitted in the same time slot.

Control Channels	**BCH** Broadcast Channels	BCCH	↓	Broadcast Control CH	Normal Burst
		FCCH	↓	Freq. Correction CH	Freq. Correction Burst
		SCH	↓	Synchronization CH	Synch. Burst
	CCCH Common C. Channels	RACH	↑	Random Access CH	Access Burst
		AGCH	↓	Access Grant CH	Normal Burst
		PCH	↓	Paging CH	
	DCCH Dedicated C. Channels	SDCCH	↕	Stand-Alone Ded. CCH	
		SACCH	↕	Slow Associated CCH	
		FACCH	↕	Fast Associated CCH	

Table 10.6 Types of control channels in GSM.

At the same time, on the uplink, there is a possibility for an MS to address the BS using random access channels (RACH). The RACH (R) uses a slotted ALOHA access scheme, hence, collisions may occasionally occur. Since most traffic, however, occurs on assigned channels, the capacity is not noticeably influenced by these collisions.

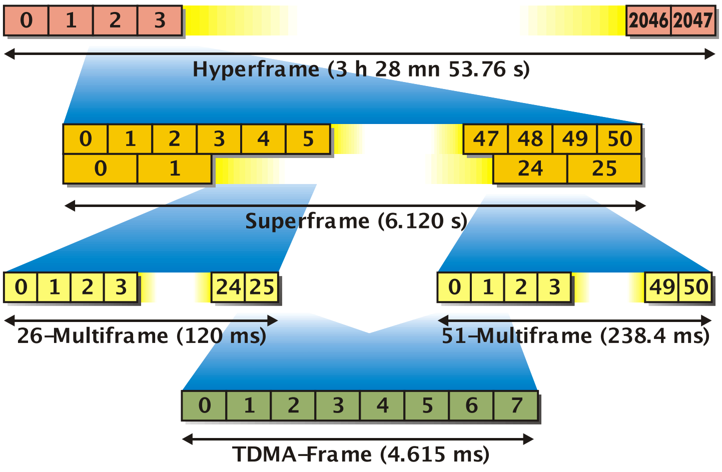
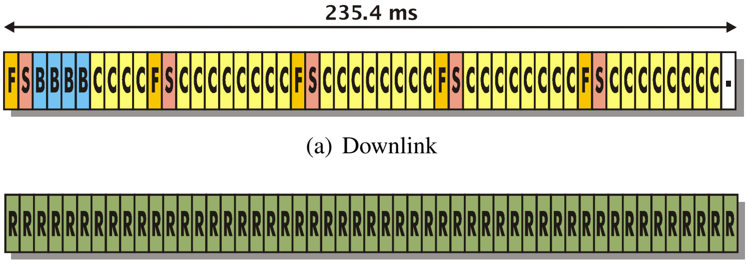


Figure 10.16 Frame structure of GSM. Source: [46].


235.4 ms
FSBBBBCCCCFSCCCCCCCCFSCCCCCCCCFSCCCCCCCCFSCCCCCCCC-
RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRR

(a) Downlink

(b) Uplink

Figure 10.17 Frame content of GSM. BCH signaling in different frames. Source: [46].

In contrast to the common control channels, the dedicated control channels occur on the same slots as the traffic channels and bear information in context with individual signaling services.

Traffic Channels

By far the most space in terms of frequencies and time slots are used for traffic channels. This is where all data and speech content is transmitted. In GSM there are a couple of different traffic channels taking care of different needs in data and speech transmission. Table 10.7 shows the traffic channels available in GSM. In normal speech mode, the data rate for delivering the voice of people through the air is thus 13 kb per second. The compression from a sampled speech stream is achieved using a so-called voice coder. Users that are

Traffic Channels					Coded bit rate	Raw bit rate
	Encoded Speech	TCH/FS	↕	Full Rate	13 kb/s	22.8 kb/s
		TCH/HS	↕	Half Rate	6.5 kb/s	11.4 kb/s
	Data	TCH/F14.1	↕	Full Rate	14.1 kb/s	22.8 kb/s
		TCH/F9.6	↕		9.6 kb/s	22.8 kb/s
		. . .				
		TCH/H2.4	↕	Half Rate	2.4 kb/s	11.4 kb/s

Table 10.7 Types of traffic channels in GSM.

given a full-rate TCH, transmit in their assigned time slot in every frame. Users that are assigned a half-rate TCH, transmit only in every second frame.

10.4.6 Speech Encoding

The following GSM speech coders are available (see also Fig. 10.7):

- **Full-Rate Codec** (GSM 06.10)
- **Half-Rate Codec** (GSM 06.20)
  - Audibly worse speech quality, at 3...4 times higher computational cost
  - Only used in emergencies
- **Enhanced Full-Rate (EFR) Codec** (GSM 06.60)
  - Uses *Adaptive Code Excited Linear Prediction* for enhanced sound quality (at 22.8 kb/s data rate)
  - Slightly worse error performance → only at sufficient SNR
- **Adaptive Multi-Rate (AMR) Codec** (GSM 06.90)
  - Variable coded data rate 4.75 to 12.2 kb/s
  - Rarely supported/used

Fig. 10.18 illustrates an example of error coded speech transmission: The full-rate codec produces a 13 kb/s data stream, of which 20 ms (equivalent to 260 bits) are buffered, partitioned and error coded with parity bits and convolutional coding.

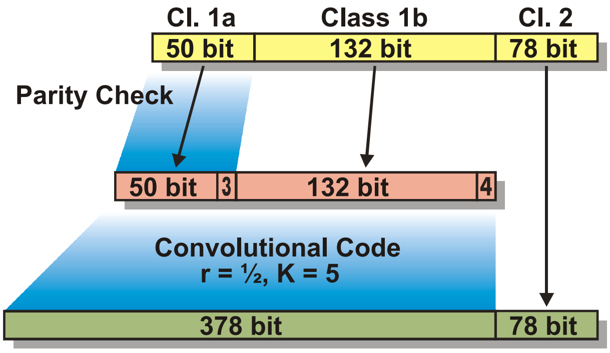
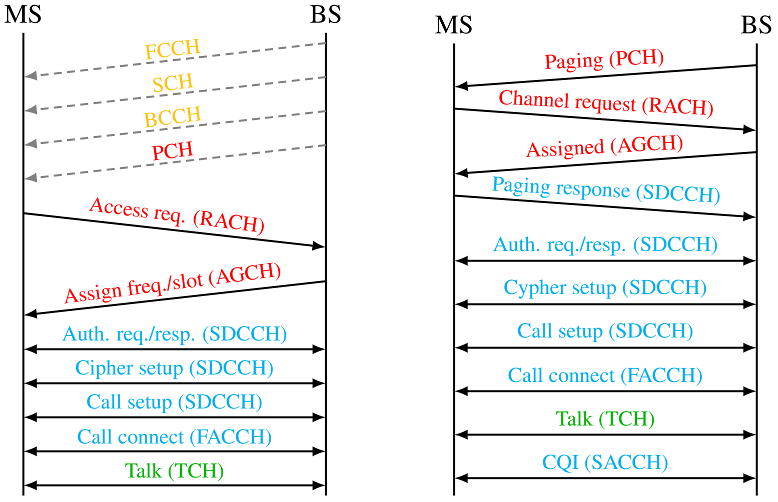


Figure 10.18 Error coding of speech within GSM.

10.4.7 Call Establishment

In the following, coarse scenarios are given of how GSM calls are established. Depending on whether a mobile station is calling or being called, we distinguish between two different cases as follows.

- **MS Originated Call:** The time line of an MS originated call is illustrated in Fig. 10.19(a) and proceeds as follows. After a user has been registered within a cell, his handset will continuously monitor the broadcast channels FCCH, SCH, PCH, and BCCH.In order to initiate a call, the MS transmits a burst of RACH (random access) on the same ARFCN (channel assignment) as the base station it has monitored. The BS then responds sending an AGCH (access grant) giving the mobile a new frequency (ARFCN) and a new time slot (TS). On this new channel, which the MS has to tune in, first authentication takes place, followed by the transmission of timing advance (TA) value and power control information. The mobile is then given yet another ARFCN and TS on which the traffic is being transmitted and received.
- **MS Terminated Call:** A call of which the MS is the responding party is essentially the same as an MS-originated call except that the BS first transmits a paging call (PCH) with the address of the MS to make it respond through an RACH, from which point the same procedure as above is started. This time line is illustrated in Fig. 10.19(b).



(a) MS originted call (outgoing) (b) MS terminated call (incoming)

Figure 10.19 Timeline diagrams of outgoing and incoming GSM calls.

10.4.8 System Extensions

Faster data transfer modes in existence are:

- **HSCSD** (High-Speed Circuit Switched Data): A single channel can now transfer 14.4 kbit/s, up to 4 channels can be combined for a total throughput of 57.6 kbit/s.
- **GPRS** (General Packet Radio Service): Single packets can be transferred at up to 14.4 kbit/s; a maximum of 8 channels can be combined, resulting in 115.2 kbit/s.
- **EDGE** (Enhanced Data rates for GSM Evolution): Single packets can be transferred at $>$ 48 kbit/s through the use of new modulation schemes. Maximally 8 channels can be combined resulting in $>$ 384 kbit/s.
- **EGPR** (Enhanced General Packet Radio Service, a subset of EDGE): Using $3\pi/8$ rotated 8-PSK, the bitrate is three times higher compared to GMSK. Up to 474 kbit/s is possible.

10.5 UMTS

The most prominent example of a CDMA-based cellular communication system in Europe is UMTS (Universal Mobile Telecommunications System), launched by the European Commission in 1989. Originally planned to cover the entire globe, it has mostly been adopted in Europe, Australia and Africa as well as in parts of Asia. Since different 2G systems existed, the migration scenarios to 3G were different depending on the region of the world. The possible evolutions into 3G standards are shown in Fig. 10.1.

The current 3rd-generation system in Europe is UMTS, which consists of two modes:

- W-CDMA, which originally occupied the frequency bands from 1920–1980 and 2110–2170 MHz in FDD (Frequency Division Duplex) for uplink and downlink.
- TD-CDMA, which was planned to occupy the frequency bands from 1900 MHz–1920 and 2010–2025 MHz in TDD (Time Division Duplex).

Some key parameters of UMTS are given in Table 10.8.

Standard	Chip rate	Modulation scheme	Bandwidth	Tx filter
WCDMA	3.84 Mchip/s	QPSK	5 MHz	RRC, $\rho = 0.22$

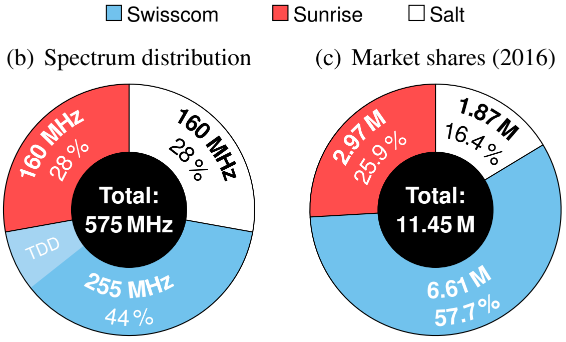
Table 10.8 Modulation parameters of UMTS.

10.5.1 Spectrum Allocation

Table 10.9(a) lists the spectrum allocations of all three cellular communication providers in Switzerland, which resulted from the spectrum auction in early 2012 and are valid up to 2028. It is interesting to note (albeit not entirely surprising) that the percentages of the allocated spectra do not directly correspond to the the market share percentages of each mobile network provider, as the diagrams in (b) and (c) reveal.

(a) Spectrum allocation

Band	Duplex	Swisscom	Sunrise	Salt
800	FDD	20 MHz	20 MHz	20 MHz
900	FDD	30 MHz	30 MHz	10 MHz
1800	FDD	60 MHz	40 MHz	50 MHz
2100	FDD	60 MHz	20 MHz	40 MHz
2100	TDD	–	–	–
2600	FDD	40 MHz	50 MHz	40 MHz
2600	TDD	45 MHz	–	–



(b) Spectrum distribution

(c) Market shares (2016)

Table 10.9 Spectrum allocation of the three mobile communication providers in Switzerland resulting from the auction of February 2012 (valid until 2028).

Unlike in previous years, the awarded licenses are now technology-neutral. Each provider can decide which communication technology to use within each band separately and dynamically. However, on the basis of international developments, it is expected that most bands will witness rapid deployment of the newest technology for high-speed data transfer, currently LTE (Long Term Evolution), the successor to UMTS. Table 10.10 shows which of the communication standards typically are used by each of the three network providers in Switzerland, within the specific bands. Many mobile network operators are abandoning 2G services within the next few years. For example, all of Japan's operators have done so already; Canadian operators have mostly phased out their 2G services as well and the US is expected to follow soon. Vodaphone switched off 2G in Australia but is waiting with similar steps in Europe. In most European countries, the operators have phase-out plans for the years 2020 to 2025. In Switzerland, Sunrise has announced phasing out 2G by end of 2018 (they previously announced to cease 2G service by 2017, but did not follow through). Swisscom and Salt have announced plans to turn off GSM by 2020.

Band	GSM/EDGE (2G)	UMTS/HSPA(+) (3G)	LTE/LTE-A (4G)
800	–	–	🟦🟥□
900	🟦🟥□	🟦🟥□	–
1800	□	–	🟦🟥□
2100	–	🟦🟥□	🟦
2600	–	–	🟦🟥□

🟦 Swisscom
🟥 Sunrise
□ Salt

Table 10.10 Typical usage of the spectrum for the different mobile communication technologies in Switzerland (as of 2012, the awarded licenses are technology-neutral).

Fig. 10.20 illustrates the current spectrum in and around the cellular communication bands in Switzerland. Additional resources in the 700 and 1400 MHz bands as well as in the range of 3.4 to 3.8 GHz are planned to become available and be auctioned off within the next few years. A few additional bands are currently being investigated for future use by CEPT (the European Conference of Postal and Telecommunications Administrations). However, as can be seen, the potential increase in spectrum is very small. Thus, expanding into frequencies above 6 GHz (most importantly in the 24 to 26 GHz range, but also in the 40 and 60 GHz bands) is being considered as well.

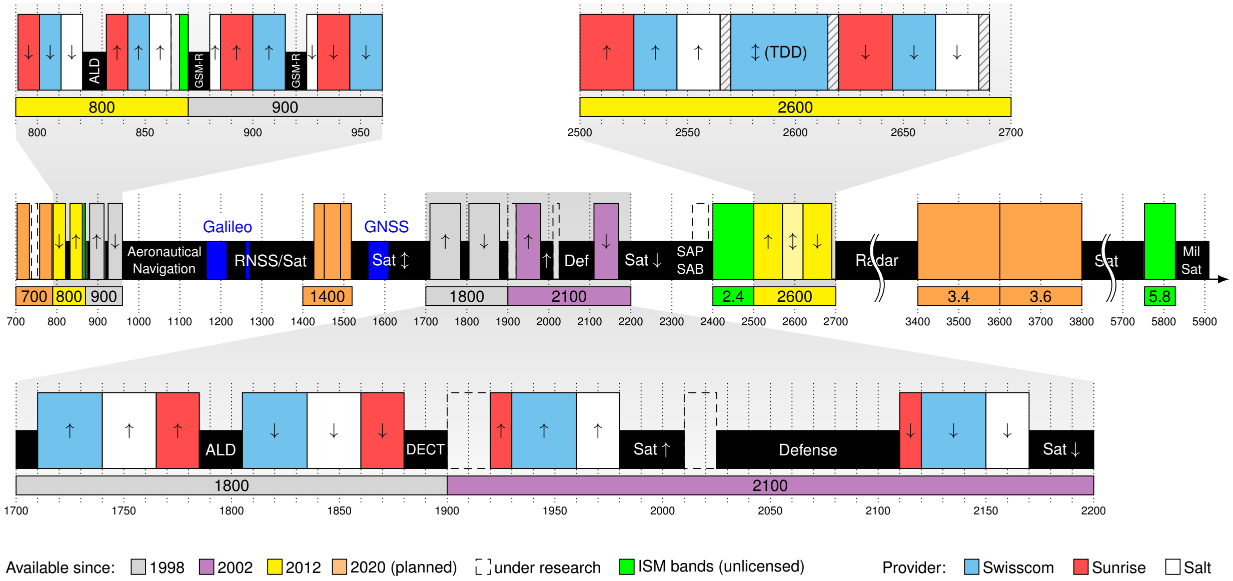


Figure 10.20 Illustration of the current frequency allocation plan for cellular communication systems and other systems in spectral proximity.

10.5.2 Soft Handover

As opposed to GSM and other TDMA-based cellular systems, where one user has a link to one cell only at one given time, a UTMS user can use signals from two base stations. They send with the same user code, such that the signal looks as if it was sent through a multipath environment. Thus, we do not have a hard handover when changing from one cell to another, but rather a smooth, so-called *soft handover*, where the user may have a simultaneous link to two base stations during a transition phase, see Fig. 10.21.

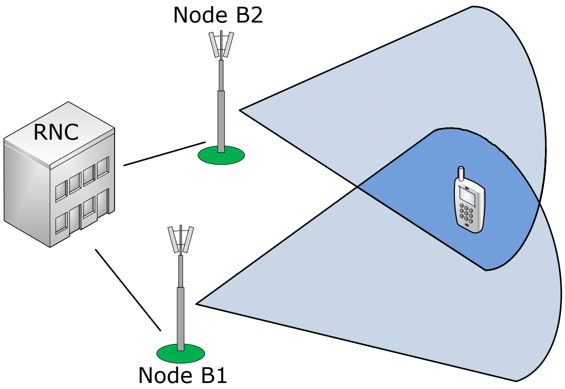


Figure 10.21 Soft handover as a special feature of UMTS.

10.5.3 Orthogonal Variable Spreading Factors

The spreading factor is an instrument to control the amount of capacity a user gets. If one user has a low data rate, but his signal is spread using a higher spreading factor to get the final bandwidth, we may accommodate many such users in one given channel. On the other hand, if one user's data rate is already quite high, only a small spreading factor is needed to spread the signal to the final bandwidth. In a UMTS cell, many users using different data rates are served simultaneously. Still, the codes need to be orthogonal. This can be achieved using *orthogonal variable spreading factors* (OVSF). For an illustration of how to choose the codes, see Fig. 10.22. Once a node is picked, all other codes later and earlier in the branch are no longer available. Usually, more than one code is applied, e.g., in order to distinguish the user (channelization) and the region (scrambling), see Fig. 10.23.

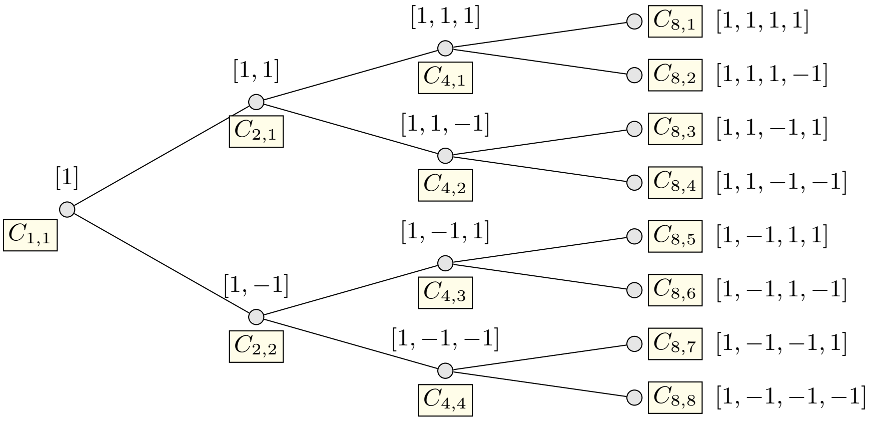


Figure 10.22 OSVF in UMTS to adapt the data rate to user need and capacity.

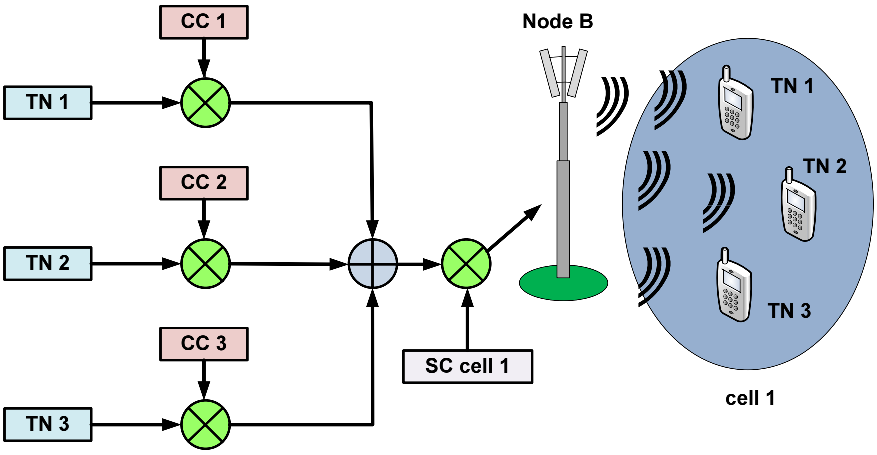


Figure 10.23 Messages from different users are spread using different spreading codes.

10.5.4 Cell Breathing

The cell planning of 2G systems is usually a static process in the sense that during operation, cell sizes are constant. For special events involving many people at a certain location, network planners usually extend the number of channels for a basestation by adding hardware modules. The mobile device adjusts its transmit power based on link measurements obtained from the BTS in order to minimize battery consumption. However, the dynamic range is rather limited, as shown in Table 10.11.

	GSM	UMTS
Dynamic	30 dB	80 dB
Measurement interval	480 ms (= 104 Frames)	0.667 ms
Adjustment step size	2 dB	1 – 3 dB
Max. adjustment	2 dB/60 ms	full range

Table 10.11 Power control in GSM and UMTS.

A 3G cell in UMTS is much more dynamic. The 'bandwidth'[5] may be shared among few users with high data rates or among many users with much lower data rates, depending on the user behavior. High data rates can only be provided at a short distance to the base station (due to the lower spreading gain), hence the radius of the cell changes to adapt to the situation. This process is called *cell breathing*. In order to solve the near-far problem, in UMTS, the transmit power is adjusted in a large range and often.

[5] We set quotes here since bandwidth is not meant in the literal sense, but more in the meaning of capacity.

10.6 LTE

The abbreviation LTE stands for Long-Term Evolution. LTE has the goal to cover the broadband internet connectivity in mobile communication environments, demanded by the users. To do so, the ITU-R (International Telecommunications Union Radiocommunication sector) specified the next-generation mobile communication in March 2008. It is called fourth-generation mobile telecommunication (4G) which includes the following peak-speed requirements on the downlink side:

- 100 Mbit/s in high mobility communication such as cars and trains
- 1 Gbit/s in low mobility communication such as stationary and pedestrians

LTE only covers up to 300 Mbit/s datarates, which strictly speaking does not reach the 4G specification in either case. Nevertheless, many providers call their LTE network a 4G network. In fact, only LTE-Advanced fulfills the requirement of 1 Gbit/s data rates on the downlink side.

10.6.1 Transmission Scheme

LTE uses an OFDM transmission scheme. The available spectrum is divided into subcarriers. Each subcarrier is modulated with a separate low-rate data stream. The main benefits of OFDM are efficient receiver architecture and robustness against multipath fading. To reduce inter-symbol-interference (ISI) additional guard intervals are inserted between the OFDM symbols to overcome the delay spread. In LTE, the guard intervals are created by adding a Cyclic Prefix (CP) in front of each OFDM symbol. Fig. 10.24 shows such an OFDM signal in a time-frequency diagram. In this example, the signal uses 5 MHz of bandwidth but other bandwidths are also possible in LTE, see Table 10.13.

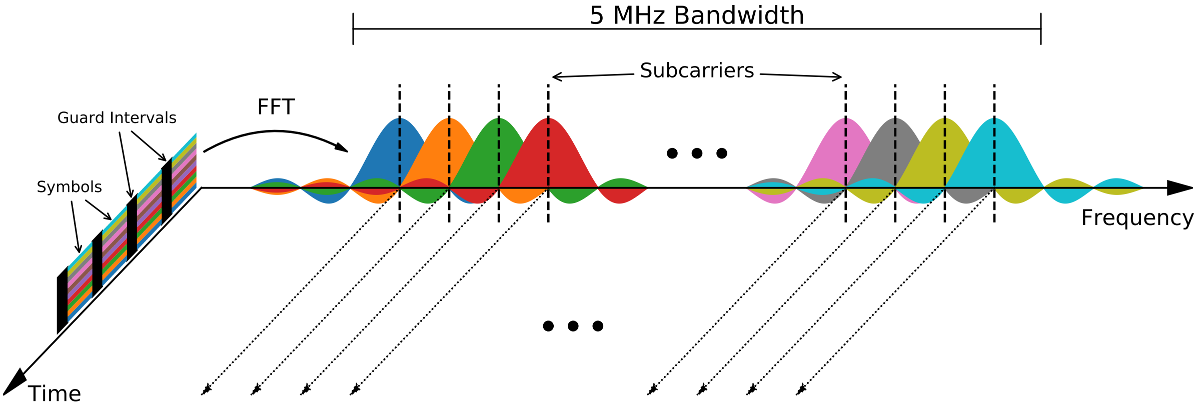


Figure 10.24 OFDM signal.

The LTE definition of subcarrier spacing is generally 15 kHz. There is also a 7.5 kHz spacing which may be available for Multimedia Broadcast Multicast Service (MBMS) transmission. The transmitted downlink signal is organized in a so-called resource grid, as shown in Fig. 10.25.

One cell in the grid is called a Resource Element (RE) and corresponds to one subcarrier in the frequency domain and one OFDM symbol in the time domain. The REs (k, l) are grouped to so-called Resource Blocks

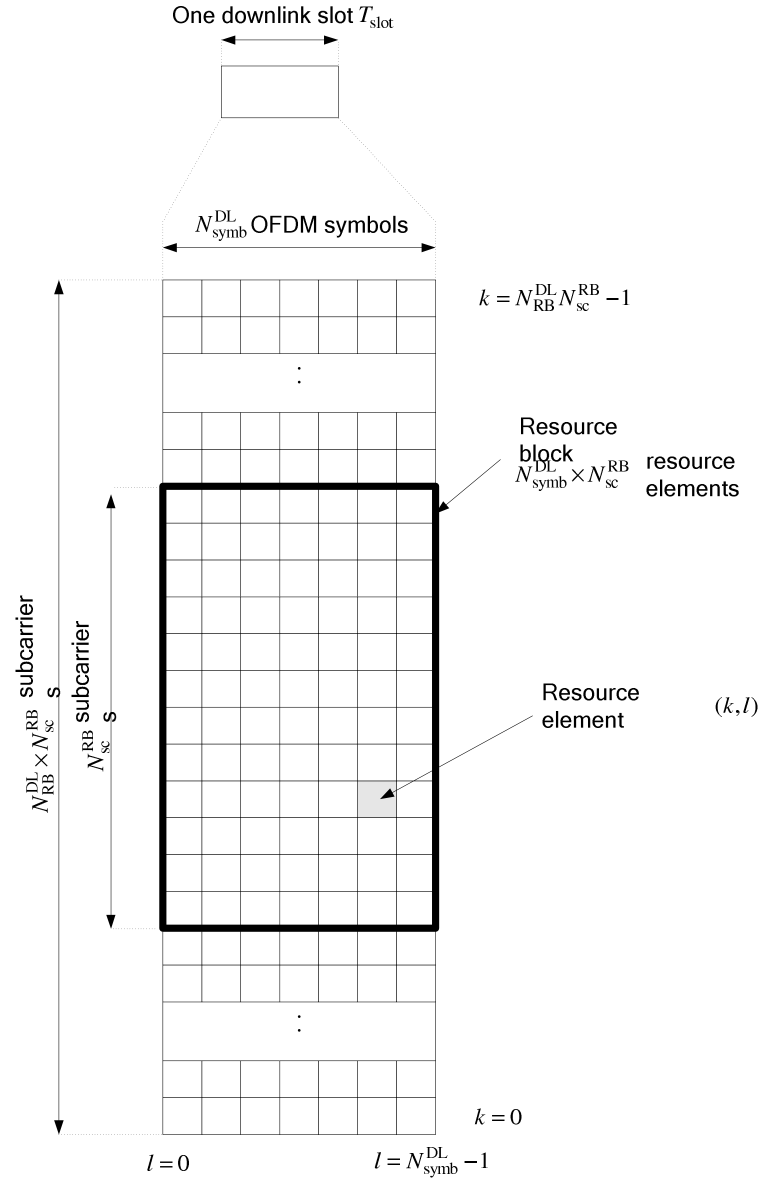

One downlink slot T_{slot}
$N_{\text{symb}}^{\text{DL}}$ OFDM symbols
$k = N_{\text{RB}}^{\text{DL}} N_{\text{sc}}^{\text{RB}} - 1$
Resource block
$N_{\text{symb}}^{\text{DL}} \times N_{\text{sc}}^{\text{RB}}$ resource elements
Resource element
(k,l)
$N_{\text{RB}}^{\text{DL}} \times N_{\text{sc}}^{\text{RB}}$ subcarriers
$N_{\text{sc}}^{\text{RB}}$ subcarriers
$k = 0$
$l = 0$
$l = N_{\text{symb}}^{\text{DL}} - 1$

Figure 10.25 Downlink resource grid.

(RB). A resource block is defined as $N_{\text{sc}}^{\text{RB}}$ contiguous subcarriers over $N_{\text{symb}}^{\text{DL}}$ contiguous OFDM symbols, where $N_{\text{sc}}^{\text{RB}}$ and $N_{\text{symb}}^{\text{DL}}$ are specified in Table 10.12.

Configuration		$N_{\textbf{sc}}^{\textbf{RB}}$	$N_{\textbf{symb}}^{\textbf{DL}}$
Normal cyclic prefix	$\Delta f = 15\,\text{kHz}$	12	7
Extended cyclic prefix	$\Delta f = 15\,\text{kHz}$	12	6
	$\Delta f = 7.5\,\text{kHz}$	24	3

Table 10.12 Resource block parameters

Typically, a RB consists of 12 subcarriers that occupy a bandwidth of 180 kHz (15 kHz spacing) and 7 OFDM symbols that corresponds to one slot (0.5 ms). In LTE, the system sample rate is defined as

$$f_\mathrm{s} = \frac{1}{T_\mathrm{s}} = 2048 \cdot 15\,\mathrm{kHz} = 30.72\,\mathrm{MHz} \tag{10.19}$$

which is eight times the sample rate of UMTS. This way, it is possible to perform a 2048-point FFT in order to extract the subcarrier amplitudes. The number of subcarriers actually used is only a fraction of the 2048 theoretically possible. The valid numbers of subcarriers and RBs used are listed in Table 10.13. Further, the center subcarrier, which corresponds to DC in baseband, is not transmitted in downlink, i.e., it is skipped and thus not used at all. The RBs are always equally distributed on both the negative and positive frequencies, as visualized in Fig. 10.26. The time-continuous signal in an OFDM symbol is defined in [20, Sec. 6.12] by

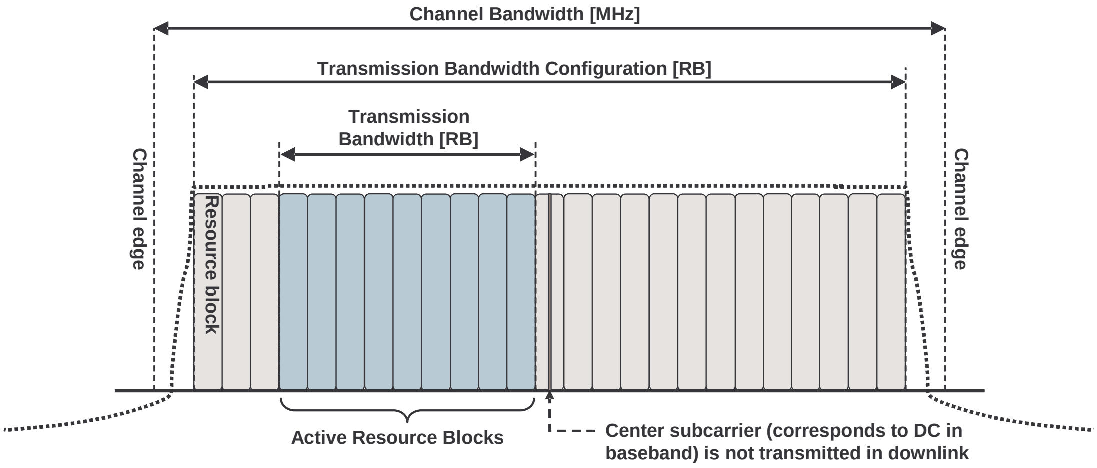

Channel Bandwidth [MHz]
Transmission Bandwidth Configuration [RB]
Transmission Bandwidth [RB]
Channel edge
Resource block
Channel edge
Active Resource Blocks
Center subcarrier (corresponds to DC in baseband) is not transmitted in downlink

Figure 10.26 Definition of channel and transmission bandwidth. Source: [19, Fig. 5.6-1]

Channel BW (MHz)	Transmission BW (MHz)	Resource Blocks ($N_\mathbf{RB}$)	Subcarriers used
1.4	1.08	6	72
3	2.7	15	180
5	4.5	25	300
10	9	50	600
15	13.5	75	900
20	18	100	1200

Table 10.13 Transmission bandwidth configuration

$$s_l^{(p)}(t) = \sum_{k=-\lfloor N_\mathrm{RB}^\mathrm{DL} N_\mathrm{sc}^\mathrm{RB}/2 \rfloor}^{-1} a_{k^{(-)}}^{(p)} \cdot \mathrm{e}^{j2\pi k \Delta f (t - N_{\mathrm{CP},l} \cdot T_\mathrm{s})} + \sum_{k=1}^{\lceil N_\mathrm{RB}^\mathrm{DL} N_\mathrm{sc}^\mathrm{RB}/2 \rceil} a_{k^{(+)}}^{(p)} \cdot \mathrm{e}^{j2\pi k \Delta f (t - N_{\mathrm{CP},l} \cdot T_\mathrm{s})} \tag{10.20}$$

for $0 \le t < (N_{\mathrm{CP},l} + N) \cdot T_\mathrm{s}$, where $k^{(-)} = k + \lfloor N_\mathrm{RB}^\mathrm{DL} N_\mathrm{sc}^\mathrm{RB}/2 \rfloor$, $k^{(+)} = k + \lfloor N_\mathrm{RB}^\mathrm{DL} N_\mathrm{sc}^\mathrm{RB}/2 \rfloor - 1$ and p denotes the antenna port number. In [20] an antenna port is explained as follows:

> *An antenna port is defined such that the channel over which a symbol on the antenna port is conveyed can be inferred from the channel over which another symbol on the same antenna port is conveyed. There is one resource grid per antenna port. The set of antenna ports supported depends on the reference signal configuration in the cell:*
>
> - *Cell-specific reference signals (CRS) support a configuration of one, two, or four antenna ports and are transmitted on antenna ports* $p = 0$, $p \in \{0, 1\}$, *and* $p \in \{0, 1, 2, 3\}$, *respectively.*
> - *Positioning reference signals (PRS) are transmitted on antenna port* $p = 6$.

In practice, this means that signals defined on these 'virtual' antenna ports may be either transmitted individually over one physical antenna per port or if convenient, combined to a smaller number of physical antennas. Actually, there is always a reference signal assigned to each antenna port in order to be able to estimate the channel that corresponds to that port. The reference signals that are relevant for the SoO receiver are described in detail in the next chapters.

10.6.2 Frame Structure

Two different frame structures are defined in LTE: Type 1 and Type 2. Type 1 is for frequency division duplex (FDD), i.e., it separates the base station (BS) downlink frequency from the mobile station (MS) uplink frequency. The second frame structure, Type 2, is managed in time division (TDD) duplex. Transmissions of BS and MS are done on the same frequency but in different time intervals. Fig. 10.27 shows the frame structure of Type 1.

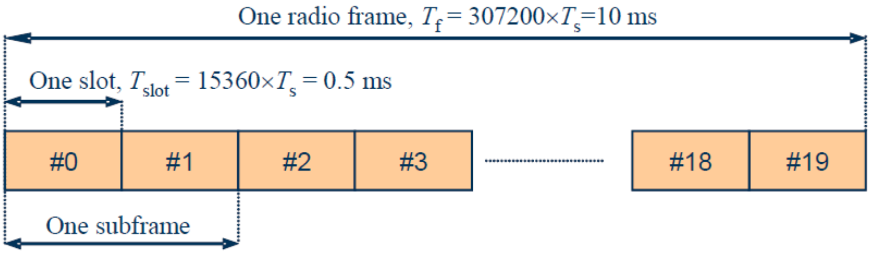

One radio frame, $T_f = 307200 \times T_s$=10 ms
One slot, $T_{slot} = 15360 \times T_s = 0.5$ ms
#0
#1
#2
#3
#18
#19
One subframe

Figure 10.27 Type 1 frame structure.

One main frame has a duration of 10 ms. The main frame is divided into 10 subframes of 1 ms each. Each subframe includes two slots, where a slot consists of either $N^{\mathrm{DL}}_{\mathrm{symb}} = 3, 6, 7$ OFDM symbols, depending on whether the cyclic prefix length and the subcarrier spacing, as listed in Table 10.14. As shown in Fig. 10.24, cyclic prefixes are inserted into the guard time between adjacent OFDM symbols to prevent ISI.

Configuration	Δf	$N^{\mathbf{DL}}_{\mathbf{symb}}$	$N_{\mathbf{CP},l}$	**CP in µs**
Normal CP	15 kHz	7	160 for $l = 0$ 144 for $l = 1 \dots 6$	5.2 µs for $l = 0$ 4.7 µs for $l = 1 \dots 6$
Extended CP	15 kHz 7.5 kHz	6 3	512 for $l = 0 \dots 5$ 1024 for $l = 0 \dots 2$	16.7 µs

Table 10.14 Cyclic prefix configuration

10.7 Handset Architectures

10.7.1 Functionality

A very coarse block diagram of the activities carried out between the microphone and the antenna in a wireless handset is given in Fig. 10.28. Depending on the number of standards a cellular handset complies to, we distinguish between different types:

- standard
- multiband
- multimode
- SDR (software-defined radio)

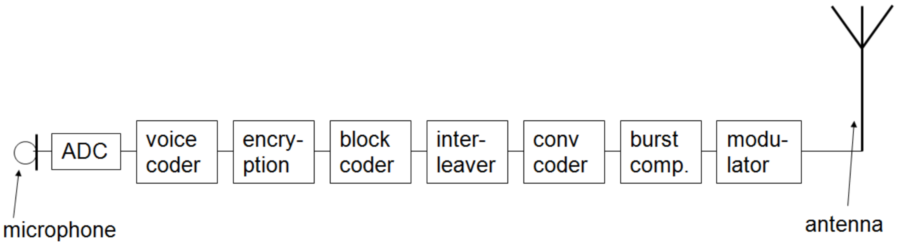


Figure 10.28 Block diagram of a wireless handset.

10.7.2 Architectures

Fig. 10.29 shows an example of a superhet architectures, for years the architecture of choice. These days, more and more wireless handsets emply direct-conversion or low-IF architectures. Also, modern hand-sets are much more than just a phone. The GSM system within an iPhone for example, see Fig. 10.30, is but one part in a multiplicity of subsystems containing digital camera, WLAN, sound processing, GPS and so on.

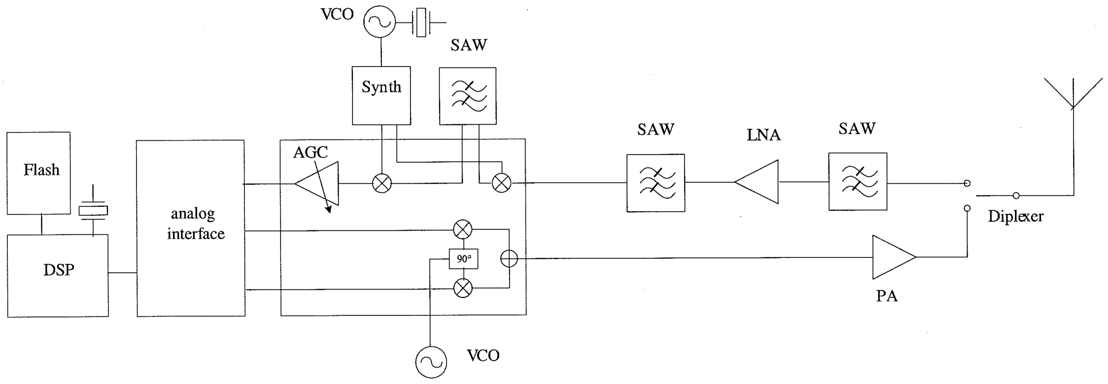


Figure 10.29 Architecture of the Siemens dual-band GSM cell phone S35, built in the year 2000.

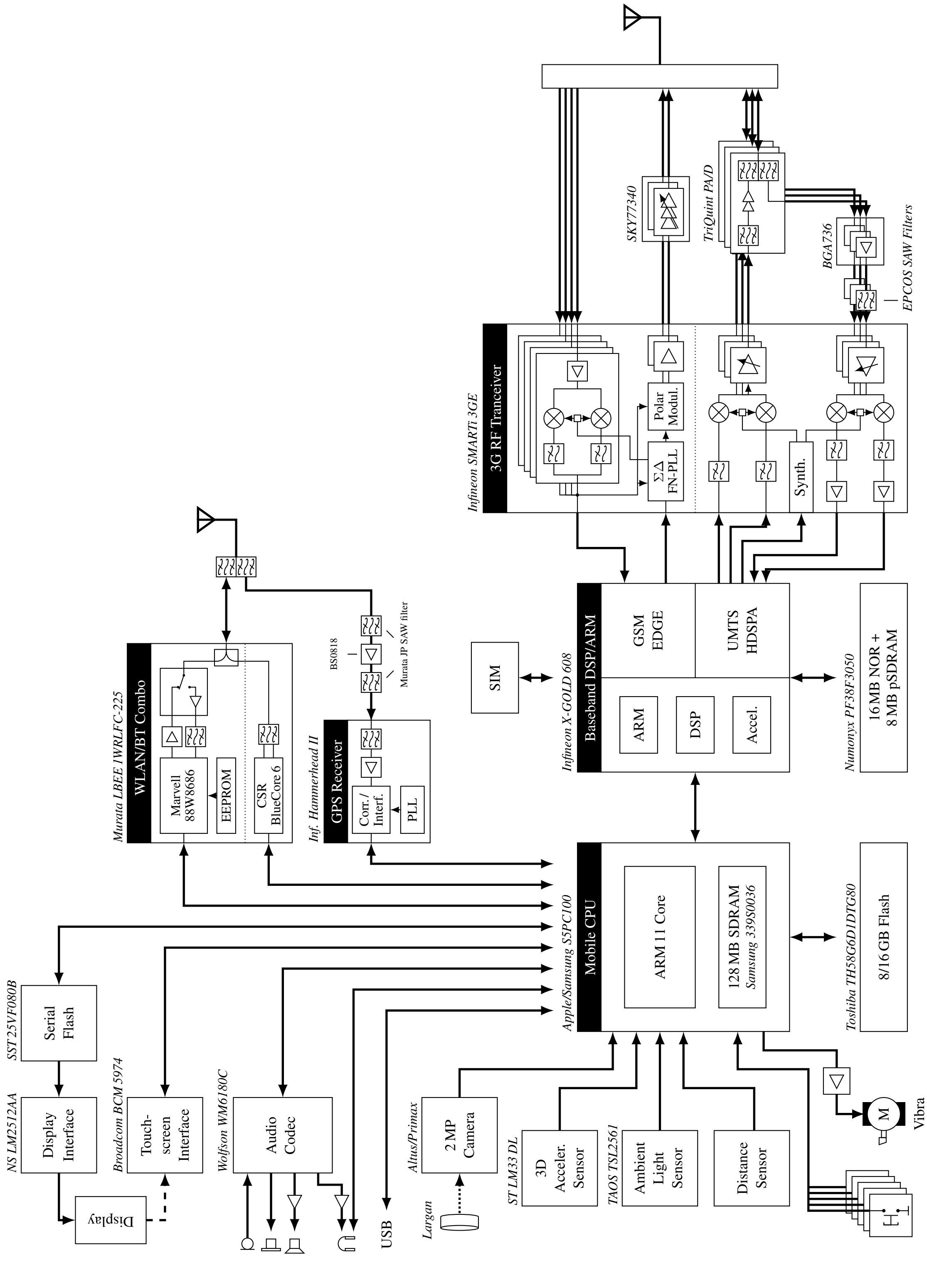


Figure 10.30 Hardware architecture of the Apple iPhone 3G.

10.8 The NIR Problem

NIR means *non-ionizing radiation* and implies that such radiation cannot change the structure of the molecules. The frequency range used in mobile communication systems produces NIR, hence, cannot cause mutations to the human genotype, unlike x-ray and radioactive radiation. Still, the issue of biological influence of electromagnetic radiation has attracted much interest over the last years. People are more and more worried about the increasing amount of radiation due to cordless telephones (inside the house) and wireless telephones (outside the house). It is part of the responsibility of an engineer to study this issue in an objective way and being able to judge between real threats and myths.

Das NIS Problem

NIS bedeutet[6] *Nichtionisierende Strahlung* und impliziert, dass solche Strahlung keine Moleküle in ihrer Struktur verändern kann. Der Frequenzbereich der für die Mobilkommunikation vorgesehen ist, erzeugt sogenannte NIS, kann also z.B. keine Mutation des menschlichen Erbguts vornehmen, wie das z.B. von Röntgenstrahlen und radioaktiver Strahlung bekannt ist. Trotzdem hat die Frage nach den biologischen Auswirkungen elektromagnetischer Strahlung in den letzten Jahren enorm an Bedeutung gewonnen. Viele Leute sind wegen der wachsenden Bestrahlungsmenge von drahtlosen Telefonen (sowohl im Hause wie ausserhalb) zunehmend verunsichert. Es liegt in der Verantwortung des Ingenieurs, sich mit dieser Problematik zu befassen um sich objektiv ein Bild davon machen zu können und beurteilen zu können, welche Gefahren möglich sind und wo Trugschlüsse herrschen.

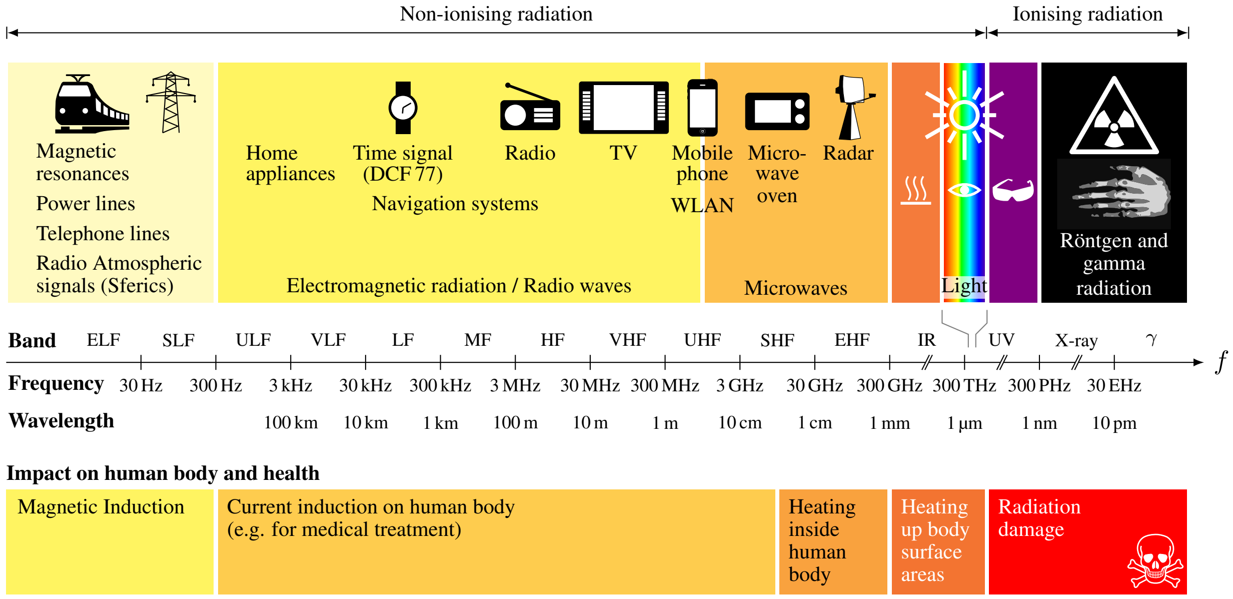


Figure 10.31 The electromagnetic spectrum and its effects on the human body.

[6] Dieses Kapitel ist zusätzlich in deutsch enthalten, um die spezifischen Begriffe, welche in der Schweizer Presse gebraucht werden, richtig zu formulieren.

10.8.1 The Effects of Electromagnetic Radiation

Fig. 10.31 shows the electromagnetic spectrum and its effects on the human body. In the range applicable to mobile communication systems there are two effects to name, so-called athermic and thermic effects. As a thermic effect, we consider the phenomenon that intense radiation can heat up human tissue, similar to a microwave oven heating up a water-containing meal. Thermal effects can potentially cause damage in the brain. However, such effects could only occur after hour-long phone calls with a wireless phone. Never will thermic effects be caused by basestation antennas. They are simply too far away for such effects to occur. The user can therefore decide for himself/herself (through his/her use of mobile technology) how far he/she wants to minimize thermal effects. Further measures that can be taken to reduce thermal effects are:

- use of a handsfree set
- put handset close to ear only after connection is established
- use of a handset with directive antenna
- use of an external antenna in the car
- generally use the phone outside buildings only (power of the handset is controlled according to the signal level received by the basestation)

With respect to athermic effects, little can be concluded at the current time. No study that has seemingly proven negative effects has been confirmed by a second, independent study. Due to the still early stage within the mobile communication age, long-term effects are currently unknown and are being further investigated. In the meantime, the World Health Organization (WHO) has developed a substantial programme for research into the as yet unknown effects of mobile communications to the health of human beings. The research programme allocates more than 150 Mio. US$ of money for international activities during the time between 1996 and 2006.

Die Wirkung elektromagnetischer Strahlung

Fig. 10.31 zeigt das elektromagnetische Spektrum und die Effekte, die dabei auf den menschlichen Körper einwirken. In dem Bereich, in dem die gesamte Mobilkommunikation abläuft, wirken zwei Effekte, sog. athermische und thermische Effekte. Als thermischen Effekt bezeichnet man das Phänomen, dass durch intensive Bestrahlung das menschliche Gewebe erwärmt wird, ähnlich wie ein Mikrowellenofen wasserhaltige Nahrung erwärmen kann. Thermische Effekte können vor allem in der Hirnregion zu Schäden führen. Allerdings sind solche Effekte erst nach stundenlangem Telefonieren mit dem Handy messbar und kommen immer vom Handset aus und nie von der Basisstation. Der Benützer hat es daher in der Hand (sprichwörtlich), durch sein Konsumverhalten etwelche thermischen Effekte zu minimieren. Weitere Massnahmen, die thermische Effekte reduzieren, sind:

- Freisprecheinrichtung benutzen
- Handy erst nach Verbindungsaufbau ans Ohr halten
- Handy mit direktiver Antenne benutzen
- im Auto Aussenantenne benutzen
- generell im Freien telefonieren (Leistung des Handys wird angepasst)

Bezüglich athermischen Effekten lässt sich zur Zeit wenig schlüssiges beweisen. Keine Studie, die vermeintlich negative Auswirkungen auf den menschlichen Organismen scheinbar nachweisen konnten, konnte durch eine unabhängige Zweitstudie bestätigt werden. Durch das junge Alter der Mobilkommunikation sind allerdings Langzeiteffekte noch unbekannt und es wird weiterhin auch in dieser Richtung geforscht. Mittlerweile hat sogar die Weltgesundheitsorganisation WHO ein umfangreiches Programm zur Erforschung noch unbekannter Auswirkungen des Mobilfunks auf die Gesundheit des Menschen ausgearbeitet. Ein Forschungsprogramm zwischen 1996 und 2006 alloziert mehr als 150 Mio. US$ an Geldern für die internationale Forschung in dieser Periode.

10.8.2 Legal Issues

In Switzerland, the permitted field strength human beings are exposed to is governed in a strict fashion by the so-called *Ordinance relating to Protection from Non-Ionising Radiation (ONIR)*, which was created in 1999 by the former BUWAL (today BAFU (in German) or the Federal Office for the Environment FOEN). The ordinance is based on the recommendations of the International Commission on Non-Ionizing Radiation Protection (ICNIRP), which, in turn, collaborates closely with the WHO.

The ONIR consisist of the following chapters:

- Chapter 1: **General provisions**
- Chapter 2: **Emissions**
- Chapter 3: **Exposure**
- Chapter 4: **Requirements for the designation of building zones**
- Chapter 5: **Final provisions**
- Annex 1: **Precautionary emission limitations**
- Annex 2: **Exposure limit values**

The field strengths relevant to mobile communication base stations can be found in Annex 1, Paragraph 6 of the ONIR. It is generally distinguished between two limit values, the exposure limit value and the installation limit value. The exposure limit value is the electric field strength in [V/m] that is not to be exceeded by the sum of all installations under any circumstances. It is applicable to all places accessible to persons. The limits are derived directly from the recommendations given in the ICNIRP.

Gesetzliche Grundlagen

In der Schweiz ist die maximal zulässige Feldstärke, der Menschen ausgesetzt sind, durch die sog. NIS-Verordnung (NISV), die 1999 ins Leben gerufen wurde, strikte geregelt. Auf Englisch nennt sich diese Verordnung *ordinance relating to protection from non-ionizing radiation (ONIR)*. Das dazu zuständige Bundesamt für Umwelt, Wald und Landschaft (BUWAL), heute Bundesamt für Umwelt BUFA, beruft sich dabei auf die Empfehlungen der Internationalen Kommission für den Schutz vor nichtionisierender Strahlung (ICNIRP), das seinerseits eng mit der WHO zusammenarbeitet.

Die NISV setzt sich aus folgenden Kapiteln zusammen:

- Kapitel 1: **Allgemeine Bestimmungen**
- Kapitel 2: **Emissionen**
- Kapitel 3: **Immissionen**
- Kapitel 4: **Anforderungen an die Ausscheidung von Bauzonen**
- Kapitel 5: **Schlussbestimmungen**
- Anhang 1: **Vorsorgliche Emissionsbegrenzungen**
- Anhang 2: **Immissionsgrenzwerte**

Die für Mobilfunk-Basisstationen relevanten Bestimmungen finden sich in Anhang 1 Ziffer 6 der NISV. Man unterscheidet grundsätzlich zwischen zwei einzuhaltenden Werten, dem Immissionsgrenzwert (IGW) und dem Anlagegrenzwert (AGW). Der Immissionsgrenzwert ist der Wert der Feldstärke in [V/m], die durch die Summe aller Anlagen auf keinen Fall überschritten werden darf und gilt für Orte kurzfristigen Aufenthalts (OKA). Hier bezog sich das BUWAL bei der Festsetzung auf die vom ICNIRP gegebene Empfehlung.

A much stricter (lower) limit value is the installation limit value, which limits the maximum radiation of one specific installation. It is also given in [V/m] and applies to places of sensitive use, i.e., places where people may stay for longer period of times. This limit is lower by a factor of 10 as compared to the recommendations given in the ICNIRP. Generally, adaptations of limit values are currently ongoing in many countries. At the current time, the limit values differ up to a factor of 1000 in different countries. Table 10.15 shows the limit values relevant to GSM in Switzerland.

Der viel strengere (kleinere) Wert ist der Anlagegrenzwert, der die Emission einer einzelnen Anlage limitiert. Er wird ebenfalls als Feldstärke in [V/m] angegeben und gilt für Orte mit empfindlicher Nutzung (OMEN), also Orte an denen sich Personen für längere Zeit aufhalten können. Dieser Grenzwert liegt um Faktor 10 unterhalb der vom ICNIRP gegebenen Empfehlung. Allgemein ist eine Anpassung der Grenzwerte in vielen Ländern im Gange. Zur Zeit variieren diese Grenzwerte bis zu einem Faktor 1000 in den verschiedenen Ländern. Tab. 10.15 zeigt die für GSM relevanten Grenzwerte in der Schweiz.

Frequency band	**Exposure limit value**	**Installation limit value**
900 MHz	42 V/m	4 V/m
1800 MHz and above	58 V/m	6 V/m
mixed use		5 V/m

Table 10.15 NIR limit values in Switzerland.

There is room for interpretation with respect to the question of what is considered as a place of sensitive use and what not. For example, the Federal Supreme Court of Switzerland has recently decided (Fall 2002) that terraces and balconies do **not** belong to places of sensitive uses.

Thermal effects can only be caused by the handset when holding it directly to the head. In order to compare the effects of handsets from different manufacturers, the so-called specific absorption rate (SAR) is used to indicate how much power is absorbed by the human tissue. Its unit is [W/kg]. Current handsets range from 0.1 W/kg to 3 W/kg. The antenna type used has most implications on the actual SAR value of the handset under consideration.

Natürlich gibt es bei diesen Definitionen noch Interpretationsspielraum, bezüglich was nun als OMEN oder was als OKA gilt. Beispielsweise hat das Bundesgericht kürzlich (Herbst 2002) entschieden, dass Terrassen und Balkone **nicht** zu den OMEN zu zählen sind und daher als OKA behandelt werden.

Thermische Effekte können nur durch das Handset, das unmittelbar an den Kopf gehalten wird, entstehen. Um die Auswirkungen der verschiedenen Handsetmodelle vergleichen zu können, bedient man sich einer dritten Grösse, der sog. spezifischen Absorptionsrate (SAR), die beschreibt wieviel Leistung pro Gewebemasse aufgenommen wird. Ihre Einheit ist daher [W/kg]. Die Spannweite bei heutigen GSM-Handsets reicht von 0.1 W/kg bis 3 W/kg. Grösste Auswirkungen auf diesen Wert hat natürlich der Antennentyp im betrachteten Handset.

10.8.3 Measurements

In addition to the ONIR, there is also a recommendation for the validation of the limit values, produced by the Federal Office of Metrology METAS. For a specific location, the limit values must not be exceeded anywhere within the location. A well-known method is therefore to swing a LogPer antenna and store the maximum measured value. If the assessment value (measurement and extrapolation) is within the installation limit value, the operator has fulfilled his duty.

Even if the assessment value exceeds the ONIR limits (based on wideband measurements), it is not necessarily the responsibility of the operator under consideration. It could potentially be a competitor's network, which further increases the assessment value. The measurement is carried out using isotropic, wideband probes, e.g., in the range $800 \dots 1000\,\mathrm{MHz}$ and $1800 \dots 1900\,\mathrm{MHz}$ and a power meter or a field-strength meter. The maximum value measured $E_{\max}$ is being multiplied by an extrapolation factor $K_{\max}$ and results in the assessment value

Messungen

Zusätzlich zur NISV hat das BUWAL zusammen mit der METAS (Metrologie und Akkreditierung Schweiz) eine Messempfehlung für die Ueberprüfung von IGW und AGW herausgegeben. Bei einer Messung innerhalb eines bestimmten Gebietes muss der Messwert überall eingehalten werden. Eine gängige Messmethode ist daher, mit einer LogPer-Antenne zu schwenken und den Maximalwert festzuhalten.

Eine Messung kann grundsätzlich breitbandig oder frequenzselektiv erfolgen. Eine breitbandige Messung dient zur Orientierung. Ist dabei der Beurteilungswert (Messung und Hochrechnung) unterhalb des AGW so hat der Netzbetreiber seine Pflicht getan. Bei Ueberschreitung desselben müssen Nachmessungen erfolgen. Eine Ueberschreitung des breitbandigen Orientierungswertes hat nicht notwendigerweise Massnahmen des Netzbetreibers zur Folge, es könnte sich ja auch um andere Netze handeln, die den Beurteilungswert in die Höhe treiben. Die Messung selbst findet mit isotropischen Breitbandsonden z.B. im Bereich $800 \dots 1000\,\mathrm{MHz}$ und $1800 \dots 1900\,\mathrm{MHz}$ und einem Leistungsmessgerät oder einem Feldstärkemessgerät statt. Der damit gemessene Höchstwert $E_{\max}$ wird mit dem Hochrechnungsfaktor $K_{\max}$ multipliziert und ergibt den Beurteilungswert

$$E_B = E_{\max} \cdot K_{\max}. \tag{10.21}$$

The extrapolation factor $K_{\max}$ considers the power ratio between the omnipresent BCCH and the TCHs, for which the installation is specified.

For the narrowband measurement, the single BCCH can be measured selectively. The BCCH is constantly transmitting with a constant power. Hence, the maximum power of a transmitting base station can be estimated be extrapolation from the power of the BCCH.

Der Hochrechnungsfaktor $K_{\max}$ berücksichtigt das Leistungsverhältnis zwischen dem omnipräsenten BCCH und den TCHs, für die eine Anlage spezifiziert ist.

Bei der schmalbandigen Messung werden die einzelnen BCCH selektiv ausgemessen. Weil der BCCH immer sendet und zwar konstant mit der gleichen Leistung (im Gegensatz zu den TCHs), wird dieser Kanal gemessen und die Gesamtleistung daraus extrapoliert.

This measurement is usually carried out using a spectrum analyzer. The assessment value can then be computed from the RMS values of the BCCH measurements

Dies geschieht meist mit einem Spektrumsanalysator. Der Beurteilungswert berechnet sich dann als RMS der einzelnen hochgerechneten BCCH-Messungen

$$E_B = \sqrt{\sum_k (E_{k,\max} \cdot K_k)^2}. \tag{10.22}$$

For either method, the measurement uncertainty is considerable. We have measurement inaccuracies in the order of ±45 % (standard deviation). For the estimation of the exposure limit value, a possible measurement error is considered, while for the estimation of the installation limit value, no such consideration is made. Residents who live within 10 % of the calculated installation limit value are legitimized to require measurements.

For UMTS, the situation is trickier. The so-called pilot channel cannot just be selected in the time or the frequency domain and separated from the traffic channels, owing to the CDMA system used. Thus, special measurement equipment is needed, which knows the spreading sequences and can thus separate the channels.

Bei beiden Methoden sind die Messunsicherheiten beträchtlich. So spricht man von einer totalen Messunsicherheit (Standardabweichung) von ±45 %. Bei der Ueberprüfung des IGW wird ein möglicher Messfehler mit eingerechnet, während für das Einhalten des AGW der Messfehler nicht berücksichtigt wird. Anwohner, die laut Rechnung innerhalb von 10 % des AGW wohnen, sind gegenüber der Baubehörde berechtigt, Messungen zu verlangen.

Bei UMTS ist die Problematik, dass der sog. Pilot-Channel nicht einfach durch Selektion im Zeit- oder Frequenzbereich von den Verkehrskanälen getrennt werden kann (CDMA). Es müssen deshalb Messgeräte verwendet werden, die die verwendeten *Spreading* Sequenzen kennen und so die Kanäle auseinander halten können.

10.8.4 Common Misconceptions

The WHO points out that the most common misconception in the public and the media is the fact that the terms *biological effects* and *adverse health effects* are used interchangeably. However, they are different. The latter is defined as a biological effect outside the body's normal range, detrimental to health and well-being. Even if a study shows that some biological effect might be due to some radiation, this might not necessarily cause an adverse health effect.

In the following, we list some other common misconceptions:

- pulsating radiation can easily penetrate walls
- UMTS has increased radiation power
- many antennas increase field strengths

10.9 Further literature

A book without any equations that gives a good overview of current systems and systems to come is the "Survival Guide for Business Managers" [69]. On the other hand, a very detailed book with lots of problems to solve is "Wireless Communications" by Rappaport [56]. Also, Juha Korhonen's "Introduction to 3G mobile communications" [34] is recommended for new standards. The GSM standard is covered in detail in the first volume of Walke's book [82]. German books include the ones written by Duque-Antón [15] and Schiller [67].

A Appendix

A.1 General

A.1.1 Greek Letters

A	α	alpha	I	ι	iota	R	ρ	rho
B	β	beta	K	κ	kappa	Σ	σ, ς	sigma
Γ	γ	gamma	Λ	λ	lambda	T	τ	tau
Δ	δ	delta	M	μ	mu	Υ	υ	upsilon
E	ϵ, ε	epsilon	N	ν	nu	Φ	ϕ, φ	phi
Z	ζ	zeta	Ξ	ξ	xi	X	χ	chi
G	η	eta	O	o	omikron	Ψ	ψ	psi
Θ	θ, ϑ	theta	Π	π	pi	Ω	ω	omega

A.1.2 Unit Conversions

1 inch	= 2.54 cm	Fahrenheit F	$= 9/5\,\mathrm{C} + 32$
1 foot	= 0.3048 m	Celsius C	$= 5/9\,(\mathrm{F} - 32)$
1 mile	= 1.609344 km	1 lb (US)	= 0.45359243 kg
		1 US gallon	= 3.785411 l

A.1.3 Dimensions

Y	Yotta	10^{24}	m	milli	10^{-3}	h	hecto	10^{2}
Z	Zetta	10^{21}	µ	micro	10^{-6}	da	deka	10^{1}
E	Exa	10^{18}	n	nano	10^{-9}	d	deci	10^{-1}
P	Penta	10^{15}	p	pico	10^{-12}	c	centi	10^{-2}
T	Tera	10^{12}	f	femto	10^{-15}			
G	Giga	10^{9}	a	atto	10^{-18}			
M	Mega	10^{6}	z	zepto	10^{-21}			
k	kilo	10^{3}	y	yocto	10^{-24}			

A.1.4 Resistivity of some Common Metals

Aluminum	$2.82 \cdot 10^{-8}\,\Omega\mathrm{m}$
Copper	$1.7 \cdot 10^{-8}\,\Omega\mathrm{m}$
Gold	$2.4 \cdot 10^{-8}\,\Omega\mathrm{m}$
Patinum	$12 \cdot 10^{-8}\,\Omega\mathrm{m}$
Silver	$1.59 \cdot 10^{-8}\,\Omega\mathrm{m}$

A.2 RF issues

A.2.1 Reactance of Inductors and Capacitors

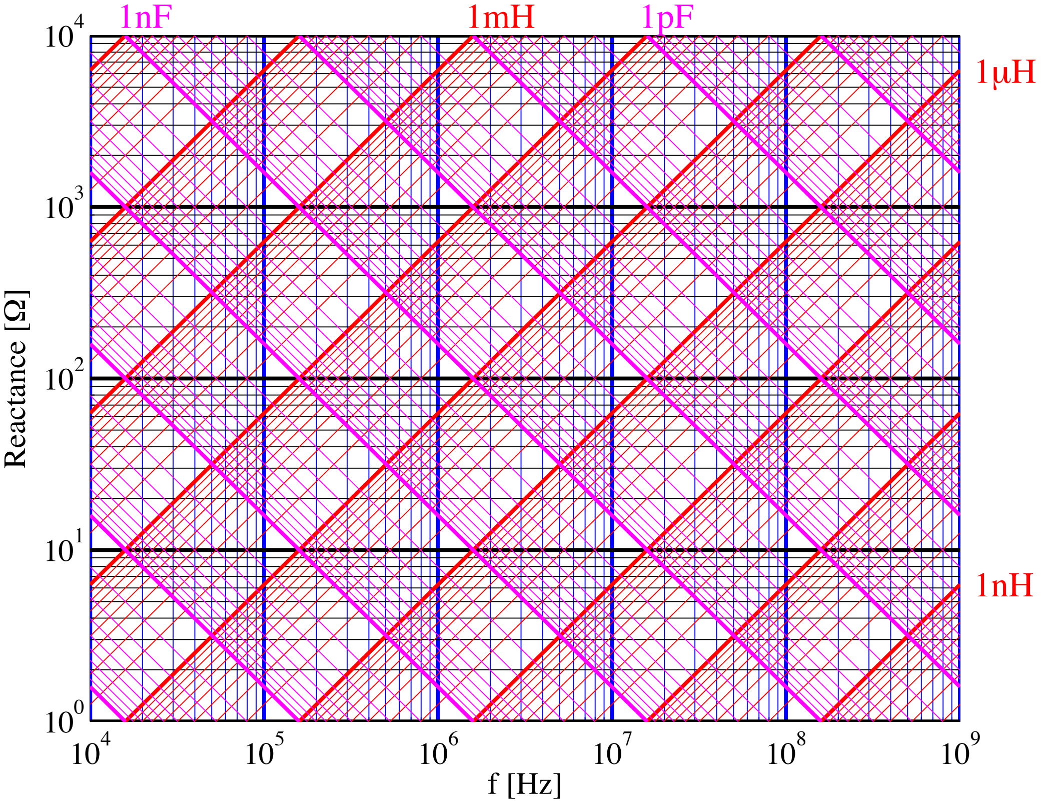


Figure A.1 Reactance of inductors and capacitors

A.2.2 dB Conversion

There is a variety of ratios that are expressed using the dB scale. The following table shall provide some of the most commonly used.

dB Ratio	Description	Value
dB	general ratio (power, voltage, etc.)	
dBc	value with respect to carrier	
dBd	power densitiy with respect to dipole antenna	0 dBd=2.15 dBi
dBi	power density with respect to isotropic radiator (antenna)	
dBm	absolute power with respect to 1 mW	0 dBm=1 mW
dBW	absolute power with respect to 1 W	0 dBW=1 W
dB/m	attenuation in dB per meter (e.g., cable loss)	
dBμV	absolute voltage with respect to 1 μV	0 dB μV=1 μV

Table A.1 Commonly used dB ratios.

Some conversions are shown in the following:

$$\text{dBm} = 10\log_{10}(P) + 30\,\text{dB} = \text{dBW} + 30\,\text{dB} = \text{dBµV} - 107\,\text{dB} \qquad (R_0 = 50\,\Omega) \tag{A.1}$$

$$\text{dBµV} = 20\log_{10}(\text{µV}_{\text{rms}}) \tag{A.2}$$

$$\text{V}_{\text{rms}} = \sqrt{R_0 \cdot 10^{\frac{\text{dBm}-30\,\text{dB}}{10}}} \tag{A.3}$$

$$\text{V}_{\text{peak}} = \sqrt{2} \cdot \text{V}_{\text{rms}} \tag{A.4}$$

dBm		**V$_{\text{rms}}$**		**V$_{\text{peak}}$**		**dBµV**		**P**	
-150	dBm	7.07	nV	10.00	nV	-43	dBµV	1.00	aW
-140	dBm	22.36	nV	31.62	nV	-33	dBµV	10.00	aW
-130	dBm	70.71	nV	100.00	nV	-23	dBµV	100.00	aW
-120	dBm	223.61	nV	316.23	nV	-13	dBµV	1.00	fW
-110	dBm	707.11	nV	1.00	µV	-3	dBµV	10.00	fW
-100	dBm	2.24	µV	3.16	µV	7	dBµV	100.00	fW
-95	dBm	3.98	µV	5.62	µV	12	dBµV	316.23	fW
-90	dBm	7.07	µV	10.00	µV	17	dBµV	1.00	pW
-85	dBm	12.57	µV	17.78	µV	22	dBµV	3.16	pW
-80	dBm	22.36	µV	31.62	µV	27	dBµV	10.00	pW
-75	dBm	39.76	µV	56.23	µV	32	dBµV	31.62	pW
-70	dBm	70.71	µV	100.00	µV	37	dBµV	100.00	pW
-65	dBm	125.74	µV	177.83	µV	42	dBµV	316.23	pW
-60	dBm	223.61	µV	316.23	µV	47	dBµV	1.00	nW
-55	dBm	397.64	µV	562.34	µV	52	dBµV	3.16	nW
-50	dBm	707.11	µV	1.00	mV	57	dBµV	10.00	nW
-45	dBm	1.26	mV	1.78	mV	62	dBµV	31.62	nW
-40	dBm	2.24	mV	3.16	mV	67	dBµV	100.00	nW
-35	dBm	3.98	mV	5.62	mV	72	dBµV	316.23	nW
-30	dBm	7.07	mV	10.00	mV	77	dBµV	1.00	µW
-25	dBm	12.57	mV	17.78	mV	82	dBµV	3.16	µW
-20	dBm	22.36	mV	31.62	mV	87	dBµV	10.00	µW
-15	dBm	39.76	mV	56.23	mV	92	dBµV	31.62	µW
-10	dBm	70.71	mV	100.00	mV	97	dBµV	100.00	µW
-5	dBm	125.74	mV	177.83	mV	102	dBµV	316.23	µW
0	dBm	223.61	mV	316.23	mV	107	dBµV	1.00	mW
5	dBm	397.64	mV	562.34	mV	112	dBµV	3.16	mW
10	dBm	707.11	mV	1.00	V	117	dBµV	10.00	mW
15	dBm	1.26	V	1.78	V	122	dBµV	31.62	mW
20	dBm	2.24	V	3.16	V	127	dBµV	100.00	mW
25	dBm	3.98	V	5.62	V	132	dBµV	316.23	mW
30	dBm	7.07	V	10.00	V	137	dBµV	1.00	W
35	dBm	12.57	V	17.78	V	142	dBµV	3.16	W
40	dBm	22.36	V	31.62	V	147	dBµV	10.00	W
45	dBm	39.76	V	56.23	V	152	dBµV	31.62	W
50	dBm	70.71	V	100.00	V	157	dBµV	100.00	W
55	dBm	125.74	V	177.83	V	162	dBµV	316.23	W
60	dBm	223.61	V	316.23	V	167	dBµV	1.00	kW
65	dBm	397.64	V	562.34	V	172	dBµV	3.16	kW
70	dBm	707.11	V	1.00	kV	177	dBµV	10.00	kW

Table A.2 Power conversion ($R_0 = 50\,\Omega$, V$_{\text{peak}}$ single tone).

A.2.3 Power of the Sum of Two Signals

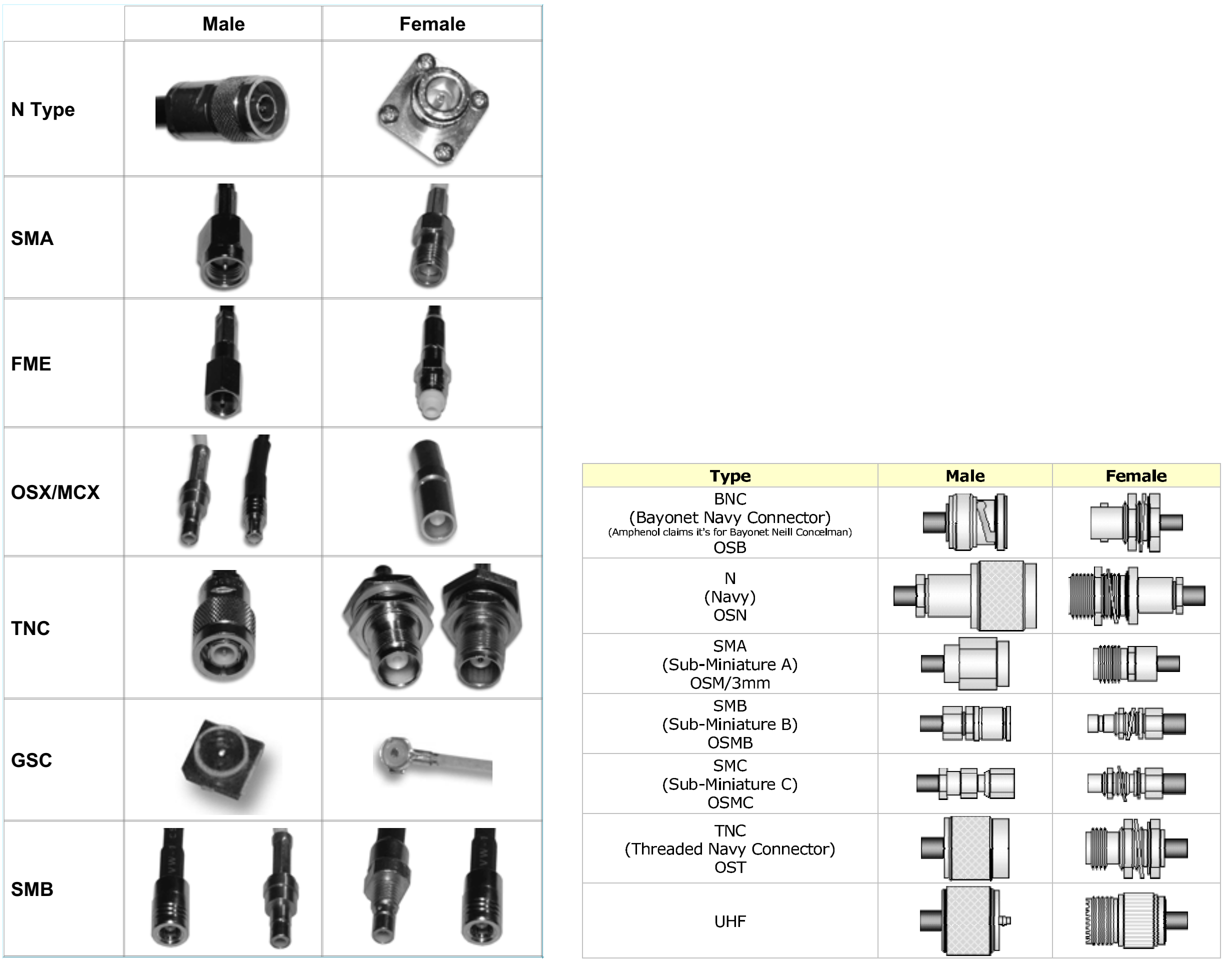
Very often, one finds oneself in the situation where a quick evaluation of the power of the sum of two signals whose individual powers are given in dB is required. Instead of painfully transforming each power to the linear domain using exponential functions, adding them and bringing the total power back into the dB domain using logarithms, a little table comes in handy.

Power difference [dB]	Power increase [dB]
0	3.0103
0.5	2.7675
1.0	2.5390
1.5	2.3247
2.0	2.1244
2.5	1.9378
3.0	1.7643
3.5	1.6037
4.0	1.4554
4.5	1.3188
5.0	1.1933
5.5	1.0783
6.0	0.9732
6.5	0.8774
7.0	0.7901
7.5	0.7108
8.0	0.6389
8.5	0.5738
9.0	0.5150
9.5	0.4618
10.0	0.4139

Power difference [dB]	Power increase [dB]
10.5	0.3708
11.0	0.3320
11.5	0.2971
12.0	0.2657
12.5	0.2376
13.0	0.2124
13.5	0.1898
14.0	0.1695
14.5	0.1514
15.0	0.1352
15.5	0.1207
16.0	0.1077
16.5	0.0962
17.0	0.0858
17.5	0.0766
18.0	0.0683
18.5	0.0609
19.0	0.0543
19.5	0.0485
20.0	0.0432

The table can be used in the following way. Consider the difference of the two signals in dB and take this value for the first column. Then look for the amount by which the total power is increased (over the higher power level) in the second column. An example: we have two signals of -110 dBm and -120 dBm power, respectively. According to our table, the increased power for a power difference of 10 dB is 0.4139 dB higher, hence, the total power is -109.5861 dBm.

A.2.4 RF Connectors



(a) (b)

Figure A.2 Connector types

A.2.5 Coaxial Cables

Dielectric Type	Time Delay (ns/ft)	Propagation Velocity (% of c)	ε_{eff}
Solid Polyethylene (PE)	1.54	65.9	2.3
Foam Polyethylene (FE)	1.27	80.0	1.56
Foam Polystyrene (FS)	1.12	91.0	1.21
Air Space Polyethylene (ASP)	1.15-1.21	84-88	1.35
Solid Teflon (ST)	1.46	69.4	2.08
Air Space Teflon (AST)	1.13-1.20	85-90	1.32

Table A.3 Coaxial cable parameters.

Type(/U)	MIL-W-17	Z0(W)	Dielect.Type	Cap.(pF/ft)	O.D.(in.)	dB/100 ft@400MHz	Vmax(rms)	Shield
RG-4		50.0	PE	30.8	0.226	11.7	1,900	Braid
RG-5		52.5	PE	28.5	0.332	7.0	3,000	Braid
RG-5A/B		50.0	PE	30.8	0.328	6.5	3,000	Braid
RG-6	/2-RG6	76.0	PE	20.0	0.332	7.4	2,700	Braid
RG-6A	/2-RG6	75.0	PE	20.6	0.332	6.5	2,700	Braid
RG-8		52.0	PE	29.6	0.405	6.0	4,000	Braid
RG-8A		52.0	PE	29.6	0.405	6.0	5,000	Braid
RG-9		51.0	PE	30.2	0.420	5.9	4,000	Braid
RG-9A		51.0	PE	30.2	0.420	6.1	4,000	Braid
RG-9B		50.0	PE	30.8	0.420	6.1	5,000	Braid
RG-10		52.0	PE	29.6	0.463	6.0	4,000	Braid
RG-10A		52.0	PE	29.6	0.463	6.0	5,000	Braid
RG-11	/6-RG11	75.0	PE	20.6	0.405	5.7	4,000	Braid
RG-11A	/6-RG11	75.0	PE	20.6	0.405	5.2	5,000	Braid
RG-12	/6-RG12	75.0	PE	20.6	0.463	5.7	4,000	Braid
RG-12A	/6-RG12	75.0	PE	20.6	0.463	5.2	5,000	Braid
RG-17A		52.0	PE	29.6	0.870	2.8	11,000	Braid
RG-22	/15-RG22	95.0	PE	16.3	0.405	10.5	1,000	Braid
RG-22A/B	/15-RG22	95.0	PE	16.3	0.420	10.5	1,000	Braid
RG-23/A	/16-RG23	125.0	PE	12.0	0.650	5.2	3,000	Braid
RG-24/A	/16-RG24	125.0	PE	12.0	0.708	5.2	3,000	Braid
RG-34	/24-RG34	71.0	PE	21.7	0.625	5.3	5,200	Braid
RG-34A	/24-RG34	75.0	PE	20.6	0.630	5.3	6,500	Braid
RG-35	/64-RG35	71.0	PE	21.7	0.928	2.8	10,000	Braid
RG-35A/B	/64-RG35	75.0	PE	20.6	0.928	2.8	10,000	Braid
RG-55B		53.5	PE	28.8	0.200	11.7	1,900	Braid
RG-58	/28-RG58	53.5	PE	28.8	0.195	11.7	1,900	Braid
RG-58A	/28-RG58	52.0	PE	29.6	0.195	13.2	1,900	Braid
RG-58B		53.5	PE	28.8	0.195	14.0	1,900	Braid
RG-58C	/28-RG58	50.0	PE	30.8	0.195	14.0	1,900	Braid
RG-59/A	/29-RG59	73.0	PE	21.1	0.242	10.5	2,300	Braid
RG-59B	/29-RG59	75.0	PE	20.6	0.242	9.0	2,300	Braid
RG-62/A/B	/30-RG62	93.0	ASP	13.5	0.242	8.0	750	Braid
RG-63/A/B	/31-RG63	125.0	ASP	10.0	0.405	5.5	1,000	Braid
RG-65/A	/34-RG65	950.0	ASP	44.0	0.405	16 @5MHz	1,000	Braid
RG-71/A/B	/90-RG71	93.0	ASP	13.5	0.245	8.0	750	Braid
RG-79/A/B	/31-RG79	125.0	ASP	10.0	0.436	5.5	1,000	Braid
RG-83		35.0	PE	44.0	0.405	9.0	2,000	Braid
RG-88		48.0		50.0	0.515	0.7 @1MHz	10,000	Braid
RG-108/A	/45-RG108	78.0	PE	19.7	0.235	2.8@10MHz	1,000	Braid
RG-111/A	/15-RG111	95.0	PE	16.3	0.478	10.5	1,000	Braid
RG-114/A	/47-RG114	185.0	ASP	6.5	0.405	8.5	1,000	Braid
RG-119	/52-RG119	50.0	ST	29.4	0.465	3.8	6,000	Braid
RG-120	/52-RG120	50.0	ST	29.4	0.523	3.8	6,000	Braid
RG-122	/54-RG122	50.0	PE	30.8	0.160	18.0	1,900	Braid
RG-130	/56-RG130	95.0	PE	17.0	0.625	8.8	3,000	Braid
RG-131	/56-RG131	95.0	PE	17.0	0.683	8.8	3,000	Braid
RG-133/A	/100-RG133	95.0	PE	16.3	0.405	5.7	4,000	Braid
RG-141/A		50.0	ST	29.4	0.190	9.0	1,900	Braid
RG-142/A/B	/60-RG142	50.0	ST	29.4	0.195	9.0	1,900	Braid
RG-144	/62-RG144	75.0	ST	19.5	0.410	4.5	5,000	Braid
RG-164	/64-RG164	75.0	PE	20.6	0.870	2.8	10,000	Braid
RG-165	/65-RG165	50.0	ST	29.4	0.410	5.0	5,000	Braid
RG-166	/65-RG166	50.0	ST	29.4	0.460	5.0	5,000	Braid
RG-177	/67-RG177	50.0	PE	30.8	0.895	2.8	11,000	Braid
RG-178/A/B	/93-RG178	50.0	ST	29.4	0.072	29.0	1,000	Braid
RG-179	/94-RG179	70.0	ST	20.9	0.100	21.0	1,200	Braid
RG-179A/B	/94-RG179	75.0	ST	19.5	0.100	21.0	1,200	Braid
RG-180	/95-RG180	93.0	ST	15.4	0.140	17.0	1,500	Braid
RG-180A/B	/95-RG180	95.0	ST	15.4	0.140	17.0	1,500	Braid
RG-210	/97-RG210	93.0	ASP	13.5	0.242	8.0	750	Braid
RG-211/A	/72-RG211	50.0	ST	29.4	0.730	2.3	7,000	Braid
RG-212	/73-RG212	50.0	PE	29.4	0.332	6.5	3,000	Braid

RG-213	/74-RG213	50.0	PE	30.8	0.405	5.5	5,000	Braid
RG-214	/75-RG214	50.0	PE	30.8	0.425	5.5	5,000	Braid
RG-215	/74-RG215	50.0	PE	30.8	0.463	5.5	5,000	Braid
RG-216	/77-RG216	75.0	PE	20.6	0.425	5.2	5,000	Braid
RG-217	/78-RG217	50.0	PE	30.8	0.545	4.3	7,000	Braid
RG-218	/79-RG218	50.0	PE	30.8	0.870	2.5	11,000	Braid
RG-219	/79-RG219	50.0	PE	30.8	0.928	2.5	11,000	Braid
RG-223	/84-RG223	50.0	PE	19.8	0.211	8.8	1,900	Dbl Braid
RG-302	/110-RG302	75.0	ST	19.5	0.201	8.0	2,300	Braid
RG-303	/111-RG303	50.0	ST	29.4	0.170	9.0	1,900	Braid
RG-304	/112-RG304	50.0	ST	29.4	0.280	6.0	3,000	Braid
RG-307/A	/116-RG307	75.0	80	16.9	0.270	7.5	1,000	Braid
RG-316	/113-RG316	50.0	ST	29.4	0.102	20.0	1,200	Braid
RG-391	/126-RG391	72.0		23.0	0.405	15.0	5,000	Braid
RG-392	/126-RG392	72.0		23.0	0.475	15.0	5,000	Braid
RG-393	/127-RG393	50.0	ST	29.4	0.390	5.0	5,000	Braid
RG-400	/128-RG400	50.0	ST	29.4	0.195	9.6	1,900	Braid
RG-401	/129-RG401	50.0	ST	29.4	0.250	4.6	3,000	Cu. S-R
RG-402	/130-RG402	50.0	ST	29.4	0.141	7.2	2,500	Cu. S-R
RG-403	/131-RG403	50.0	ST	29.4	0.116	29.0	2,500	Braid
RG-405	/133-RG405	50.0	ST	29.4	0.086	13.0	1,500	Cu. S-R
9914 (Belden)		50.0		26.0	0.405	10.0	——	

Table A.4 Coaxial cables

A.3 Frequency Usage

A.3.1 General Frequency Bands

Band	Wavelength	Frequency range	Propagation	Examples
ELF	10^5 km - 10^4 km	<30 Hz		
SLF	10^4 km - 10^3 km	30 - 300 Hz		
ULF	1000 km - 100 km	300 Hz - 3 kHz		
VLF	100 km - 10 km	3 - 30 kHz	ground and room wave	heart-beat monitor
LF	10 km - 1 km	30 - 300 kHz	ground wave $>$ 1000 km room wave at night	Long wave radio (150 - 300 kHz)
MF	1 km - 100 m	300 kHz - 3 MHz	ground wave $>$ 100 km room wave at night	Medium wave radio (500 - 1500 kHz)
HF	100 m - 10 m	3 MHz - 30 MHz	sky wave reflected by ionosphere	Short wave radio (5 - 30 MHz)
VHF	10 m - 1 m	30 MHz - 300 MHz	free-space openfield	FM radio (88.1 - 108.1 MHz)
UHF	1 m - 10 cm	300 MHz - 3 GHz	free-space openfield	Consumer Communications, GNSS
SHF	10 cm - 1 cm	3 GHz - 30 GHz	free-space openfield	Radar, Satellite
EHF	1 cm - 1 mm	30 GHz - 300 GHz	free-space	Radar, Satellite
THF	1 mm - 0.1 mm	300 GHz - 3 THz	free-space	Airport, body scanner ('T' for tremendously high ...)

Table A.5 Frequency bands.

SHF is also called microwaves and EHF is called millimeter waves.

Frequency	Wavelength		Band (old)
1 – 2 GHz	30 – 15	cm	L
2 – 4 GHz	15 – 7.5	cm	S
4 – 8 GHz	7.5 – 3.75	cm	C
8 – 12 GHz	3.75 – 2.5	cm	X
12 – 18 GHz	2.5 – 1.67	cm	Ku
18 – 27 GHz	1.67 – 1.11	cm	K
27 – 40 GHz	11.1 – 7.5	mm	Ka
40 – 75 GHz	7.5 – 4	mm	V
75 – 110 GHz	4 – 2.72	mm	W
110 – 300 GHz	2.72 – 1	mm	mm

Table A.6 IEEE frequency band designations.

A.3.2 RADAR Bands

Frequency	Band (old)
0.22 - 0.3 GHz	P
33 - 50 GHz	Q
40 - 60 GHz	U
60 - 90 GHz	E
90 - 140 GHz	F
110 - 170 GHz	D
140 - 220 GHz	G
170 - 260 GHz	Y
220 - 325 GHz	J

Table A.7 IEEE radar band designations.

There is yet another set of letters in use, which matches the alphabetical order to the increasing frequencies in succeeding bands. Although this set is newer and hence more modern, it has not found wide acceptance and is therefore left out here.

A.4 Modulation Schemes

First letter:	
A	Main carrier is amplitude modulated and uses double sideband.
B	Main carrier is amplitude modulated and uses independent sideband.
C	Main carrier is amplitude modulated and uses vestigial sideband
F	Main carrier is angle modulated and uses frequency modulation
G	Main carrier is angle modulated and uses phase modulation
H	Main carrier is amplitude modulated and uses single sideband, full carrier
J	Main carrier is amplitude modulated and uses a single sideband, suppressed carrier
R	Main carrier is amplitude modulated and uses a single sideband, reduced or variable-level carrier

First number:	
1	Signal modulating the main carrier is a single channel containing quantized or digital information without the use of a modulating subcarrier
2	Signal modulating the main carrier is a single channel containing quantized or digital information with the use of a modulating subcarrier
3	Signal modulating the main carrier is a single channel containing analog information
8	Signal modulating the main carrier is 2 or more channels containing analog information

Second letter:	
A	Telegraphy for aural reception
B	Telegraphy for automatic reception
C	Facsimile transmission
D	Data transmission, telemetry or telecommand
E	Telephony
F	Television (video)
W	A combination of any of the kinds of information described in the previous items

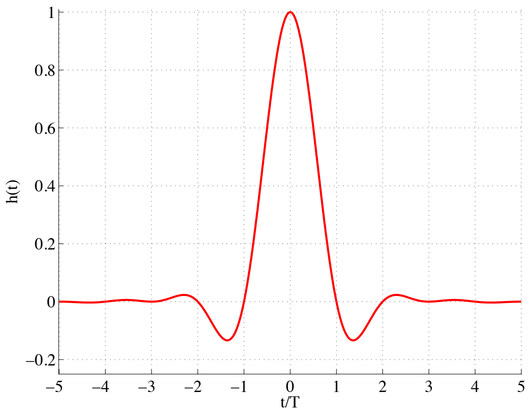
Table A.8 Modulation codes

Mode	Description	Emission Code
AM	amplitude modulation, double sideband, full carrier, voice	A3E
FM	frequency modulation, voice	F3E
PM	phase modulation, voice	G3E
SSB	amplitude modulation, single side band, suppressed carrier, voice	J3E
SSB-TV	amplitude modulation, single side band, suppressed carrier, image	J3F
TV	Television using FM	F3F
RTTY	telegraphy using FSK, no modulating audio tone	F1B
RTTY	telegraphy using AFSK, modulating audio tone	F2B
packet	FM data	F1D
CW	amplitude modulation, double sideband, full carrier, data	A1A

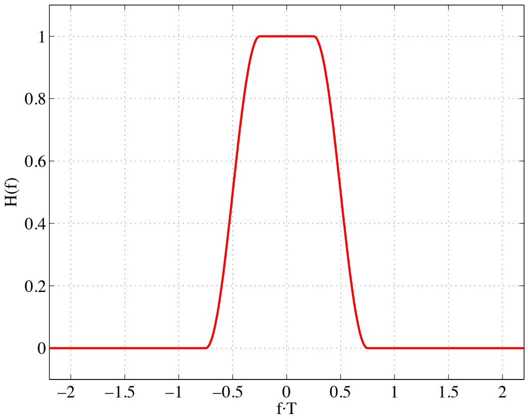
Table A.9 Examples of modulation codes

A.5 Pulses and their Fourier Transforms

A.5.1 Raised Cosine



○—●

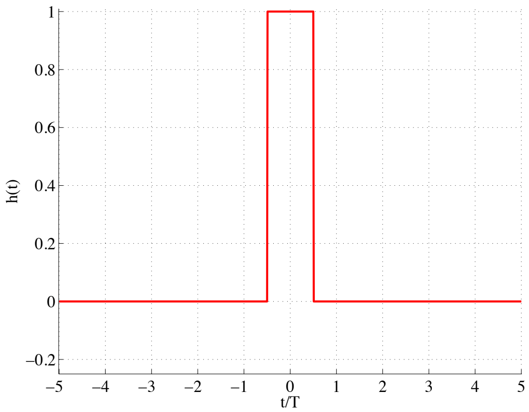


here: $\rho = 0.5$

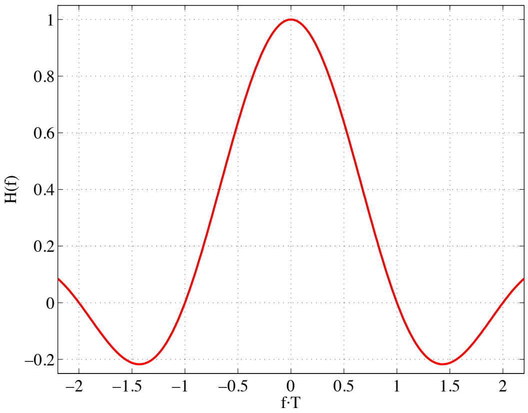
$$h(t) = \frac{\sin\left(\frac{t}{T}\pi\right)}{\frac{t}{T}\pi} \cdot \frac{\cos\left(\rho\frac{t}{T}\pi\right)}{1 - 4\rho^2(\frac{t}{T})^2}$$

$$H(f) = \begin{cases} 1, & 0 \leq |f| \leq \frac{1-\rho}{2T}, \\ \frac{T}{2}\left[1 - \sin\left(\frac{\pi(|f|T-\frac{1}{2})}{\rho}\right)\right], & \frac{1-\rho}{2T} \leq |f| \leq \frac{1-\rho}{2T}, \\ 0 & \frac{1-\rho}{2T} \leq |f|. \end{cases}$$

A.5.2 Rectangular



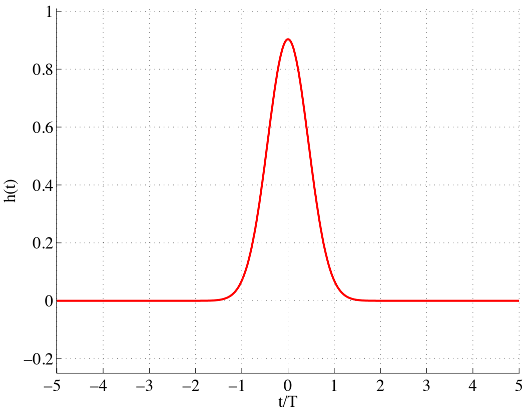
○—●



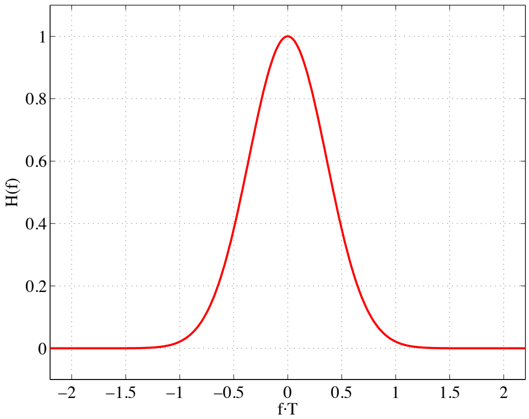
$$h(t) = \begin{cases} 1, & |t| \leq 0.5T, \\ 0, & |t| > 0.5T. \end{cases}$$

$$H(f) = \frac{\sin(\pi f T)}{\pi f T}$$

A.5.3 Gaussian



$\circ\!\!-\!\!\bullet$



here: $BT = 0.3$

$$h(t) = B\sqrt{\frac{2\pi}{\ln 2}}\mathrm{e}^{-\pi^2 B^2 \frac{2}{\ln 2} t^2} \qquad H(f) = \mathrm{e}^{-\left(\frac{f}{B}\right)^2 \frac{\ln 2}{2}}$$

A.5.4 M-Sequences

Stages	Possible feedback taps
2	[2, 1]
3	[3, 2]
4	[4, 3]
5	[5, 3]
	[5, 4, 3, 2]
	[5, 4, 3, 1]
6	[6, 5]
	[6, 5, 4, 1]
	[6, 5, 3, 2]
7	[7, 6]
	[7, 4]
	[7, 6, 5, 4]
	[7, 6, 5, 2]
	[7, 6, 4, 2]
	[7, 6, 4, 1]
	[7, 5, 4, 3]
	[7, 6, 5, 4, 3, 2]
	[7, 6, 5, 4, 2, 1]
8	[8, 7, 6, 1]
	[8, 7, 5, 3]
	[8, 7, 3, 2]
	[8, 6, 5, 4]
	[8, 6, 5, 3]
	[8, 6, 5, 2]
	[8, 7, 6, 5, 4, 2]
	[8, 7, 6, 5, 2, 1]
9	[9, 5]
	[9, 8, 7, 2]
	[9, 8, 6, 5]
	[9, 8, 5, 4]
	[9, 8, 5, 1]
	[9, 8, 4, 2]
	[9, 7, 6, 4]
	[9, 7, 5, 2]
	[9, 6, 5, 3]
	[9, 8, 7, 6, 5, 3]
	[9, 8, 7, 6, 5, 1]
	[9, 8, 7, 6, 4, 3]
	[9, 8, 7, 6, 4, 2]
	[9, 8, 7, 6, 3, 2]
	[9, 8, 7, 6, 3, 1]
	[9, 8, 7, 6, 2, 1]
	[9, 8, 7, 5, 4, 3]
	[9, 8, 7, 5, 4, 2]
	[9, 8, 6, 5, 4, 1]
	[9, 8, 6, 5, 3, 2]
	[9, 8, 6, 5, 3, 1]
	[9, 7, 6, 5, 4, 3]
	[9, 7, 6, 5, 4, 2]
	[9, 8, 7, 6, 5, 4, 3, 1]
10	[10, 7]
	[10, 9, 8, 5]
	[10, 9, 7, 6]
	[10, 9, 7, 3]
	[10, 9, 6, 1]
	[10, 9, 5, 2]
	[10, 9, 4, 2]
	[10, 8, 7, 5]
	[10, 8, 7, 2]
	[10, 8, 5, 4]
	[10, 8, 4, 3]
	[10, 9, 8, 7, 5, 4]
	[10, 9, 8, 7, 4, 1]
	[10, 9, 8, 7, 3, 2]
	[10, 9, 8, 6, 5, 1]
	[10, 9, 8, 6, 4, 3]
	[10, 9, 8, 6, 4, 2]
	[10, 9, 8, 6, 3, 2]
	[10, 9, 8, 6, 2, 1]
	[10, 9, 8, 5, 4, 3]
	[10, 9, 8, 4, 3, 2]
	[10, 9, 7, 6, 4, 1]
	[10, 9, 7, 5, 4, 2]
	[10, 9, 6, 5, 4, 3]
	[10, 8, 7, 6, 5, 2]
	[10, 9, 8, 7, 6, 5, 4, 3]
	[10, 9, 8, 7, 6, 5, 4, 1]
	[10, 9, 8, 7, 6, 4, 3, 1]
	[10, 9, 8, 6, 5, 4, 3, 2]
	[10, 9, 7, 6, 5, 4, 3, 2]

Table A.10 Feedback taps that result in m-sequences for up to ten stages. Source: `http://www.newwaveinstruments.com`

A.5.5 The Q-function

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.0+	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1+	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2+	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3+	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4+	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5+	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6+	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7+	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8+	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9+	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0+	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1+	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2+	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3+	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4+	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5+	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6+	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7+	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8+	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9+	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0+	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1+	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2+	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3+	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4+	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
3.5+	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
3.6+	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

Table A.11 Numerical evaluations of some values of $Q(x)$.

Table A.11 has been produced by Matlab. In Matlab, the Q-function as given by

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}} dy \tag{A.5}$$

is not directly accessible. It can, however, be easily computed using the erfc command, since

$$Q(x) = \frac{1}{2} \operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right). \tag{A.6}$$

B Abbreviations

ABC	always best connected (G4 slogan)
AC	alternating current
ACK	acknowledgment
ADC, A/D	analog-to-digital converter
adf	amplitude density function
ADSL	asymmetric digital subscriber line
AF	audio frequency
AGC	automatic gain control
AGCH	access grant channel
A-GPS	aided GPS
AGW	Anlagegrenzwert
ALOHA	random access system
AM	amplitude modulation
AMPS	Advanced Mobile Phone Service
AOA	angle of arrival
ARFCN	absolute radio frequency channel number
ARQ	automatic repeat request
ARRL	American Radio Relay League
AS	anti-spoofing
ASK	amplitude shift keying
ASP	air space polyethylene
AST	air space teflon
AUC	authentication center
AWGN	additive white Gaussian noise
BAFU	Bundesamt für Umwelt
BAKOM	Bundesamt für Kommunikation
BB	baseband
BCCH	broadcast control channel
BCH	Bose-Chaudhuri-Hocquenghem Code
BCH	broadcast channel
BER	bit error rate
BLE	Bluetooth Low-Energy (standard)
BNC	Bayonet Neill-Concelman (connector)
BP	band pass (filter)
BPSK	binary phase shift keying
BS	base station
BSC	base station controller
BSS	base station subsystem
BT	bandwidth-time product
BTS	base transceiver station
BUWAL	Bundesamt für Umwelt, Wald und Landschaft
BW	bandwidth
C/A	coarse acquisition
CATV	cable television (formerly: community antenna television)
CCCH	common control channel
CCH	control channel
cdf	cumulative density function
CD	compact disc
CDMA	code division multiple access
CEM	comutational electromagnetics
CEP	circle error probable
CEPT	Conférence Europeéne des Administrations des Postes et des Télécommunications
CIR	channel impulse response
CM	constant modulus
CMA	constant-modulus algorithm
CORDIC	COrdinate Rotation DIgital Computer (algorithm)
CPM	continuous-phase modulation
CRT	cathode ray tube
Cs	caesium, chemical element nr. 55
CTF	channel transfer function
CW	continuous wave
DAB	digital audio broadcasting
DAC, D/A	digital-to-analog converter
DBM	double-balanced mixer
DC	direct current
DCCH	dedicated control channel
DCF-77	**D**eutschland (Germany), **C**ontinuous Wave (CW), **F**rankfurt, 77.5 kHz
DCO	digitally controlled oscillator
DCS	digital cellular system
dd	decision directed
DDS	direct digital synthesizer
DECT	digital enhanced cordless telecommuniations
DFE	decision-feedback equalizer
DGPS	differential GPS
DLL	delay-locked loop
C/N_0	carrier-to-noise-density ratio
DFT	discrete Fourier transform
DMT	discrete multitone
DOD	Department of Defense (USA)
DOP	dilution of precision
DPSK	differential phase shift keying

DQPSK	differential quaternary phase shift keying
DS	direct sequence
DSB	double sideband
DSP	digital signal processor
DSSS	direct sequence spread spectrum
DTTB	Digital Terrestrial Television Broadcasting
DTV	digital television
DUT	device under test
DVB	digital video broadcasting
DVB-C	digital video broadcasting, over cable
DVB-H	digital video broadcasting for handhelds
DVB-S	digital video broadcasting over satellite
DVB-T	digital video broadcasting, terrestrial system
E_b/N_0	bit energy over noise power density
E-911	extended emergency location
ECC	error control code
EDGE	enhanced data rates for GSM evolution
EGPRS	enhanced general packet radio service (subset of EDGE)
EGNOS	European Geostationary Navigation Overlay Service
EHF	extremely high frequency (30–300 GHz)
EIR	equipment identity register
EIRP	equivalent isotropic radiated power
EM	electro-magnetic
ENR	excess noise ratio
E-OTD	enhanced observed time difference
ERP	equivalent radiated power (with respect to dipole)
ESA	European Space Agency
ETSI	European Telecommunications Standards Institute
EVM	error vector magnitude
FAA	Federal Aviation Agency
FACCH	fast associated control channel
FAQ	frequently asked questions
FCC	Federal Communications Commission (USA)
FCCH	frequency correction channel
FDD	frequency division duplex
FDMA	frequency division multiple access
FE	foam polyethylene
FEC	forward error correction
FFSK	fast frequency shift keying
FFT	fast Fourier transform
FH	frequency hopping
FIR	finite impulse response
FOEN	Federal Office for the Environment
FM	frequency modulation
FMCW	frequency-modulation continuous wave (radar)
FOEN	Federal Office for the Environment
FS	full slot
FS	fractionally spaced
FS	foam polystyrene
FSE	fractionally spaced equalizer
FSK	frequency shift keying
FSM	finite state machine
GaAs	Gallium-Arsenide
GDOP	geometric dilution of precision
GEO	geostationary earth orbit
GMSK	Gaussian minimum shift keying
GNSS	global navigation satellite system
GOS	grade of service
GPR	ground-penetrating radar
GPRS	general packet radio service
GPS	Global Positioning System
GSM	global system for mobile communications
HDTV	high-definition television
HDOP	horizontal dilution of precision
HF	high frequency (3–30 MHz)
HLR	home location register
HPBW	half-power beamwidth
HS	half slot
HSCSD	high-speed circuit switched data (GSM)
HSDPA	high-speed downlink packet access (UMTS)
HSI	high side injection
HTx	hilly terrain (speed x km/h)
IC	integrated circuit
ICI	interchannel interference
ICNIRP	international commission on non-ionizing radiation protection
IEEE	Institute of Electrical and Electronics Engineers
IF	intermediate frequency
IFF	information friend or foe (secondary radar)
IFFT	inverse fast Fourier transform
IGW	Immissionsgrenzwert
i.i.d.	independent and identically distributed
IIP	input-refered intercept point
IL	insertion loss
IM	intermodulation
IMSI	international mobile subscriber identity

IMT	international mobile telecommunications
IMTS	improved mobile telephone systems
IQ	in-phase/quadrature
IS	intermediate standard (e.g., IS-95)
ISB	independent sideband
ISDN	integrated services digital network
ISI	intersymbol interference
ISM	industrial, scientific, medical
ITU	International Telecommunications Union
JPO	Joint Program Office
LBS	location-based service
LCP	left-hand circularly polarized
LEO	low earth orbit
LF	low frequency (30–300 kHz)
LFM	linear frequency modulation
LHCP	left-hand circularly polarized
LHS	left-hand side
LMS	least mean square
LMU	location measurement unit
LNA	low-noise amplifiers
LO	local oscillator
LOS	line of sight
LS	least square
LSB	lower sideband
LSI	low side injection
LTE	Long Term Evolution
LUT	look-up table
MAC	medium-access control
MAC	multiply and accumulate
MC	Monte-Carlo
MCS	master control station
MDR	multipath-to-direct-path ratio
MDS	minimum discernible signal
MEO	medium earth orbit
MER	message error rate
METAS	Metrologie und Akkreditierung Schweiz
MF	medium frequency (300 kHz–3 MHz)
MIMO	multiple in—multiple out
ML	maximum likelihood
MMSE	minimum mean square error
MS	mobile station
MSAS	Multi-Functional Satellite Augmentation System
MSC	mobile switching center
MSE	mean square error
MSK	minimum shift keying
MSS	mobile satellite service
MTS	mobile telephone systems
NADC	North American Digital Cellular
NAK	negative acknowledgment
NATEL	Nationales Autotelefonnetz
NAVSTAR	Navigation System with Timing and Ranging
NCO	numerically controlled oscillator
NEP	noise equivalent power
NF	noise figure
NIR	non-ionizing radiation
NIS	nichtionisierende Strahlung
NISV	Verordnung über den Schutz vor nichtionisierender Strahlung
NLOS	non-line of sight
NMT	nordic mobile telephone
NOMA	non-orthogonal multiple access
NSS	network and switching subsystem
OCS	operational control segment
OCXO	oven-controlled crystal oscillator
OEM	original equipment manufacturer
OFDM	orthogonal frequency division multiplex
OIP	output-refered intercept point
OKA	Orte für kurzfristigen Aufenthalt
OMC	operating and maintenance center
OMEN	Orte mit empfindlicher Nutzung
ONIR	ordinance relating to protection from non-ionizing radiation
OSI	open system interconnection
OSL	open, short, load (calibration method)
OSM	open, short, match (calibration method)
PA	power amplifier
PAM	pulse amplitude modulation (also called ASK)
PAN	personal area network
PCB	printed circuit board
PCH	paging channel
PCM	pulse code modulation
PCS	Personal Communication System
PDC	Personal Digital Cellular
PDP	power-delay profile
PDOP	position dilution of precision
PE	polyethylene
pdf	probability density function
PDMA	polarization division multiple access
PHY	physical layer
PIFA	printed inverted-F antenna
PLL	phase-locked loop
PM	phase modulation
PMR	Private Mobile Radio
ppm	parts per million
PPS	precise positioning service
PR	pseudo random
PRN	pseudo random noise

PRS	(Galileo) publicly regulated service
P/S	parallel to serial
PSK	phase shift keying
PSTN	public switched telephone network
PVT	position, velocity, time (GPS)
QAM	quadrature amplitude modulation
QOS	quality of service
QPSK	quaternary phase shift keying
RACH	random access channel
RADAR	radio detection and ranging
RAx	rural area (speed x km/h)
Rb	rubidium, chemical element nr. 37
RBW	resolution bandwidth
RC	raised cosine
RCP	right-hand circularly polarized
RCS	radar cross section
RF	radio frequency
RFID	radio frequency identification
RHCP	right-hand circularly polarized
RHS	right-hand side
RLS	recursive least square
RMS	root mean square
rpm	revolutions per minute
RRC	root raised cosine
RS	Reed-Solomon Code
RSSI	receiver signal strength indicator
RTOF	roundtrip time of flight
RTTY	radio transmission technology
SA	selective availability
SACCH	slow associated control channel
SAR	specific absorption rate
SAR	synthetic aperture radar
SAW	surface acoustic wave
SBAS	satellite based augmentation systems
SBM	single-balanced mixer
SC	stress compensated (crystal cut)
SCH	synchronization channel
SDCCH	stand-alone dedicated control channel
SDMA	space division multiple access
SDR	software-defined radio
SER	symbol error rate
SFDR	spurious-free dynamic range
SHF	super high frequency (3–30 GHz)
SIM	subscriber identity module
SMS	short message service
SNR	signal-to-noise ratio
SOC	system-on-a-chip
S/P	serial to parallel
SPS	standard positioning service
SRD	short-range device
SS	spread spectrum
SSB	single sideband
ST	solid teflon
SV	space vehicle
SWR	standing-wave ratio
SWT	sweep time
TA	timing advance
TCH	traffic channel
TCXO	temperature-controlled crystal oscillator
TDD	time division duplex
TDMA	time division multiple access
TDOA	time difference of arrival
TDOP	time dilution of precision
TDR	time-domain reflectometry
TETRA	Terrestrial Trunked Radio
TOA	time of arrival
TRF	tuned RF set
TS	time slot
TS	training sequence
TUx	typical urban (speed x km/h)
TV	television
TWT	travelling wave tubes
UE	unit element
UHF	ultra high frequency (300 MHz–3 GHz)
UIM	user identification module
UMTS	Universal Mobile Telecommunication System
US	United States
USB	upper sideband
UTRAN	UMTS terrestrial radio access network
UWB	ultrawideband
VA	Viterbi algorithm
VBW	video bandwidth
VCO	voltage controlled oscillator
VDOP	vertical dilution of precision
VHF	very high frequency (30–300 MHz)
VLF	very low frequency (3–30 kHz)
VLR	visitor location register
VNA	vector network analzyer
VoWLAN	voice over WLAN
VSB	vestigial sideband
VSWR	voltage standing wave ratio
WAAS	wide-area augmentation systems
W-CDMA	wideband CDMA
WHO	World Health Organization
WLAN	wireless local area network
w.r.t.	with respect to
XO	crystal oscillator

C Important Constants

Standard acceleration	$g = 9.80665\,\mathrm{m/s^2}$
Gravitational constant	$G = 6.6742 \cdot 10^{-11}\,\mathrm{Nm^2\,kg^{-2}}$
Mean radius of earth	$6.37103 \cdot 10^{6}\,\mathrm{m}$
Electron’s charge	$q = 1.6 \cdot 10^{-19}\,\mathrm{C}$
Permittivity of free space (dt. *Permittivität, früher Dielektrizitätskonstante*)	$\varepsilon_0 = 8.85418 \cdot 10^{-12}\,\mathrm{As/Vm}$ $\approx 8.85\,\mathrm{pF/m} \quad \approx \frac{10^{-9}}{36\pi}\,\mathrm{F/m}$
Permeability of free space	$\mu_0 = 4\pi \cdot 10^{-7}\,\mathrm{Vs/Am} \;\approx 1.26\,\mathrm{\mu H/m}$
Speed of light	$c_0 = 1/\sqrt{\varepsilon_0\mu_0} = 2.99792 \cdot 10^{8}\,\mathrm{m/s} \;\approx 3 \cdot 10^{8}\,\mathrm{m/s}$
Free-space impedance	$Z = \sqrt{\frac{\mu_0}{\varepsilon_0}} = 120\pi\,\Omega = 377\,\Omega$
Boltzmann constant	$k = 1.38 \cdot 10^{-23}$
@ $T_0 = 290\mathrm{K}$	$kT_0 = 4.004 \cdot 10^{-21} = -174\,\mathrm{dBm/Hz}$

D Bibliography

[1] American Radio Relay League, *The ARRL handbook.* ARRL, 1994.

[2] American Radio Relay League, *The ARRL antenna book.* ARRL, 1996.

[3] W. Bächtold, *Höchstfrequenz- und Mikrowellenelektronik (different lecture notes).* ETH Zurich, 1997.

[4] I. J. Bahl and P. Bhartia, *Microstrip Antennas.* Artech House, 1980.

[5] C. A. Balanis, *Antenna Theory.* John Wiley & Sons, 1997.

[6] C. A. Balanis, *Modern Antenna Handbook.* John Wiley & Sons, 2008.

[7] P. Beckmann and A. Spizzichino, *The scattering of Electromagnetic Waves from Rough Surfaces.* Macmillan, New York, 1963.

[8] H. W. Bode, *Network Analysis and Feedback Amplifier Design.* Van Nostrand, New York, 1945.

[9] C. Bowick, *RF Circuit Design.* SAMS, 1995.

[10] I. N. Bronstein and K. A. Semendyayev, *Handbook of Mathematics.* Springer Verlag, 3rd ed., 1997.

[11] R. W. Chang, "Orthogonal frequency division multiplexing," *U.S. Patent 3'488'445*, 1970.

[12] B. G. Colpitts and G. Boiteau, "Harmonic radar transceiver design: Miniature tags for insect tracking," *IEEE Trans. Antennas Propag*, 2004, pp. 2825–2832.

[13] D. Daly and A. C. Carusone, "A sigma-delta based open-loop frequency modulator," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, Bangkok, Thailand, Cambridge University Press, May 25–28, 2003, pp. 929–932.

[14] G. Delisle, J.-P. L. evre, M. Lecours, and J.-Y. Chouinard, "Propagation loss prediction: a comparative study with application to the mobile radio channel," *IEEE Trans. on Vehicular Technology*, vol. 34, no. 2, May 1985, pp. 86–96.

[15] M. Duque-Antón, *Mobilfunknetze.* Vieweg & Sohn Verlagsgesellschaft mbH, 1st ed., 2002.

[16] M. L. Edwards and J. H. Sinsky, "A new criterion for linear 2-port stability using a single geometrically derived parameter," *IEEE Transactions on Microwave Theory and Techniques*, vol. 40, no. 12, December 1992, pp. 2303–2311.

[17] J. J. Egli, "Radio propagation above 40 mc over irregular terrain," *Proceedings of the IRE*, vol. 45, no. 10, Oct. 1957, pp. 1383–1391.

[18] P. Enge and P. Misra, "Special issue on: GPS—the global positioning system," *Proceedings of the IEEE*, vol. 87, no. 1, January 1999.

[19] ETSI, TS, "136 104 v11. 4.0."

[20] ETSI, TS, "136 211 v11. 2.0."

[21] R. M. Fano, “Theoretical limitations on the broadband matching of arbitrary impedances,” *MIT technical report no. 41*, January 1950.

[22] D. Fleisch, *A Student’s Guide to Maxwell’s Equations*. Cambridge University Press, 2008.

[23] R. Garg, P. Bhartia, I. Bahl, and A. Ittipiboon, *Microstrip Antenna Design*. Artech House, 2000.

[24] J. Gibson, ed., *The Mobile Communications Handbook*. CRC Press, 1996.

[25] R. Gold, “Optimal binary sequences for spread spectrum multiplexing,” *IEEE Transactions on Information Theory*, vol. 13, no. 4, October 1967, pp. 619–621.

[26] R. Gold, “Maximal recursive sequences with 3-valued recursive cross-correlation functions,” *IEEE Transactions on Information Theory*, vol. 14, no. 1, January 1968, pp. 154–156.

[27] D. D. Grieg and H. F. Engelmann, “Microstrip – a new transmission technique for the klilomegacycle range,” *Proceedings of the IRE*, vol. 40, no. 12, Dec 1952, pp. 1644–1650.

[28] F. W. Grover, *Inductance Calculations*. Dover, 2009.

[29] J. B. Hagen, *Radio-Frequency electronics*. Cambridge University Press, 1996.

[30] E. O. Hammerstad, “Equations for microstrip circuit design,” in *5th European Microwave Conference*, 1975.

[31] M. Hata, “Empirical formula for propagation loss in land mobile radio services,” *IEEE Trans. on Vehicular Technology*, vol. 29, no. 3, Aug. 1980, pp. 317–325.

[32] E. D. Kaplan, ed., *Understanding GPS: Principles and Applications*. Artech House, 1996.

[33] K. Kark, *Antennen und Strahlungsfelder*. Vieweg & Sohn Verlagsgesellschaft mbH, 1st ed., 2004.

[34] J. Korhonen, *Introduction to 3G Mobile Communications*. Artech House, 2001.

[35] J. Kraus and D. Fleisch, *Electromagnetics with Applications*. McGraw-Hill, 1999.

[36] V. Lakafosis and M. M. Tentzeris, “From single- to multihop: The status of wireless localization,” *IEEE Microwave Magazine*, vol. 10, no. 7, December 2009, pp. S34–S41.

[37] T. H. Lee, “A nonlinear history of radio,” in *The design of CMOS Radio-Frequency Integrated Circuits*, Cambridge University Press, 1998.

[38] F. Losee, *RF systems, components, and circuits handbook*. Artech House, 1997.

[39] W. Mansfeld, ed., *Satellitenortung und Navigation*. Vieweg & Sohn Verlagsgesellschaft mbH, 1998.

[40] H. Mathis and P. O. Vontobel, “Shape optimization for a rectangularly constrained small loop antenna,” in *International Zurich Seminar on Broadband Communications*, Zurich, Switzerland, February 15–17, 2000, pp. 73–76.

[41] H. Meinke and F. W. Gundlach, *Taschenbuch der Hochfrequenztechnik*. Springer Verlag, 4th ed., 1985.

[42] R. B. Messaros and F. J. Seifert, “Performance of GMSK transmission under typical indoor and outdoor channel conditions using coherent and noncoherent reception,” in *Proc. VTC ’90*, vol. 40, May 1990, pp. 492–495.

[43] Mini-Circuits, *RF/IF Designer’s Handbook*. Mini-Circuits, 1993.

[44] Mini-Circuits, “Übertrager für HF- und Mikrowelle,” *HF-Praxis*, vol. 14, no. 12, December 2009, pp. 24–31.

[45] M. Nyffenegger, *GNSS Multipath Mitigation Using Antenna Arrays.* MSE Project Thesis OST University of Applied Science Eastern Switzerland, Jan 29, 2021.

[46] H. Ochsner, *Mobilkommunikation.* ETH Zurich, 2001.

[47] Y. Okumura, E. Ohmori, T. Kawano, and K. Fukuda, “Field strength and its variability in vhf and uhf land mobile radio service,” *Review of the Elec. Comm. Lab*, vol. 16, no. 9–10, May 1968, pp. 825–873.

[48] M. E. O’Neal, D. A. Landis, E. Rothwell, L. Kempel, and D. Reinhard, “Tracking insects with harmonic radar: a case study,” *American Entomologist*, Winter 2004, pp. 212–218.

[49] B. W. Parkinson and J. J. Spilker, Jr., eds., *Global Positioning System: Theory and Applications, Vol. I.* American Institute of Aeronautics and Astronautics, 1996.

[50] P. Pauli, *Moderne Antennentechnik.* Weiterbildungszentrum Sarnen, 2003.

[51] D. M. Pozar and D. H. Schaubert, eds., *Microstrip Antennas.* IEEE Press, 1995.

[52] J. G. Proakis, *Digital Communications.* McGraw-Hill, 3rd ed., 1995.

[53] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing.* Prentice Hall, 3rd ed., 1996.

[54] J. G. Proakis and M. Salehi, *Grundlagen der Kommunikationstechnik.* Pearson, 1st ed., 2004.

[55] D. Psychoudakis, W. Moulder, C. Chi-Chih, Z. Heping, and J. L. Volakis, “A portable low-power harmonic radar system and conformal tag for insect tracking,” *IEEE Antennas and Wireless Propagation letters*, 2008, pp. 444–447.

[56] T. S. Rappaport, *Wireless Communications.* Prentice Hall, 2nd ed., 2002.

[57] C. Rauscher, *Fundamentals of Spectrum Analysis.* Rohde & Schwarz, 5th ed., 2007.

[58] F. Riehle, *Frequency Standards – Basics and Applications.* John Wiley & Sons, 2004.

[59] J. R. Riley and A. D. Smith, “Design considerations for an harmonic radar to investigate the flight of insects at low altitude,” *Compute. Electron. Agricult., Elsevier*, vol. 35, Aug. 2002, pp. 151–169.

[60] J. Rollett, “Stability and power-gain invariants of linear twoports,” *IRE Transactions on Circuit Theory*, vol. 9, no. 1, March 1962, pp. 29–32.

[61] J. Rollett, “Correction to stability and power-gain invariants of linear twoports,” *IEEE Transactions on Circuit Theory*, vol. 10, no. 1, March 1963, pp. 107–107.

[62] C. Rousseau and Y. Saint-Aubin, *Mathematics and Technology.* Springer Verlag, 2008.

[63] A. W. Rudge, K. Milne, A. D. Olver, and P. Knight, *The Handbook of Antenna Design.* IEE Electromagnetic Waves Series, 2nd ed., 1968.

[64] C. L. Ruthroff, “Some broad-band transformers,” *Proceedings of the IRE*, vol. 47, no. 8, Aug 1959, pp. 1337–1342.

[65] W. E. Sabin and E. O. Schoenike, *Single Sideband Systems & Circuits.* McGraw-Hill, 1995.

[66] S. R. Saunders, *Antennas and Propagation for Wireless Communication Systems.* John Wiley & Sons, 1999.

[67] J. Schiller, *Mobilkommunikation*. Pearson, 2003.

[68] M. V. Schneider, "Microstrip lines for microwave integrated circuits," *Bell System Technology Journal*, May 1969.

[69] R. Schneiderman, *The Mobile Technology: Question and Answer Book*. American Management Association (Amacom), 2002.

[70] R. N. Simons, *Coplanar Waveguide – Circuits, Components, and Systems*. John Wiley & Sons, 2001.

[71] B. Sklar, *Digital Communications*. Prentice Hall, 1988.

[72] P. H. Smith, "Transmission line calculator," *Electronics*, no. 1, January 1939, pp. 29–31.

[73] P. H. Smith, "An improved transmission line calculator," *Electronics*, no. 1, January 1944, pp. 130–133, 318–325.

[74] P. H. Smith, *Electronic applications of the Smith Chart: in waveguide, circuit, and component analysis*. Atlanta : Noble, 1995.

[75] M. I. Solnik, *Introduction to RADAR systems*. McGraw-Hill, 3rd ed., 2001.

[76] E. J. Sterba and C. B. Feldman, "Transmission lines for short-wave radio systems," *Proceedings of the IRE*, vol. 20, no. 7, July 1932, pp. 1163–1202.

[77] W. L. Stutzman and G. A. Thiele, *Antenna Theory and Design*. Springer Verlag, 2nd ed., 1998.

[78] N. Tahir and G. Brooker, "Recent developments and recommendations for improving harmonic radar tracking systems," in *Proc. 5th European Conf. on AandP (EUCAP)*, Apr. 2011, pp. 1531–1535.

[79] J. B.-Y. Tsui, *Fundamentals of Global Positioning System Receivers*. John Wiley & Sons, 2000.

[80] F. van Diggelen, "Lies, damn lies, and statistics," *GPS World*, January 2007, pp. 26–32.

[81] B. C. Wadell, *Transmission Line Design Handbook*. Artech House, 1991.

[82] B. Walke, *Mobilfunknetze und ihre Protokolle 1: Grundlagen, GSM, UMTS und andere zellulare Mobilfunknetze*. B. G. Teubner Verlagsgesellschaft, 3rd ed., 2001.

[83] D. K. Weaver, "A third method of generation and detection of single-sideband signals," *Proceedings of the IRE*, vol. 44, no. 12, December 1956, pp. 1703–1705.

[84] W. Webb and L. Hanzo, *Modern Quadrature Amplitude Modulation*. Pentech Press, 1994.

[85] S. Wyrsch, *Nachrichtentechnik 1*. atis Ingenieurschule HTL, 1998.