Robert Sobot

Wireless Communication Electronics

Introduction to RF Circuits and Design Techniques

Springer

Wireless Communication Electronics

Robert Sobot

Wireless Communication Electronics

Introduction to RF Circuits and Design Techniques

Springer

Robert Sobot
Department of Electrical and Computer Engineering
The University of Western Ontario
Richmond Street 1151
N6A 5B8 London, ON
Canada

ISBN 978-1-4614-1116-1 e-ISBN 978-1-4614-1117-8
DOI 10.1007/978-1-4614-1117-8
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2012930048

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Allen

Preface

This textbook originated in my lecture notes for the "Communication Electronics I" undergraduate course that I have offered over the last six years to the students at The University of Western Ontario in London Ontario, Canada. The book covers the transitional area between low frequency and high frequency wireless circuits. Specifically, it introduces the fundamental physical principles related to the operation of a typical wireless radio communication system.

By no means have I attempted to touch upon all the possible topics related to wireless transmission systems. Most modern textbooks cover a large number of topics with relatively low level of details, which are usually left as an "exercise to the reader". In this textbook I have chosen to discuss the subject in more depth, and thus provide detailed mathematical derivations, applied approximations, and analogies. The chosen topics are, in my experience, suitable for a one semester, four hours per week, senior undergraduate engineering course. My intent was to tell a logical story that flows smoothly from one chapter to the next, hoping that the reader will find it easy to follow.

My main inspiration in writing this book came from my students, who at the beginning of the semester would always ask: "What do I need to study for this course?". Having a choice between writing a textbook that covers many topics at a high level, or the one that covers fewer fundamental principles but in more detail, I choose the latter. All of the material in this book is considered the basic knowledge that is expected to have been acquired by aspiring engineers entering the field of wireless communication electronics.

Therefore, the intended audience for this book are, primarily, senior undergraduate engineering students preparing for their carriers in communication electronics. At the same time, my hope is that graduate engineering students will find this book a useful reference for some of the topics that have been only touched upon in the previous stages of their education, or are explained from a different point of view. Finally, the practicing junior RF engineers may find this book a handy source for the quick answers that are routinely omitted from most textbooks.

London ON, Canada — Robert Sobot

Acknowledgements

I would like to acknowledge all those wonderful books that I used as references and the source of my knowledge, and to say thank you to their authors for providing me with the insights that otherwise I would not have been able to acquire. Under their influence, I was able to expand my own picture of reality, which is what acquiring of the knowledge is all about. My hope is that their guidance and shaping of my own understanding of the topics in this book are clearly visible, hence I do want to acknowledge their contributions, which are now being passed on to my readers.

In professional life one learns both from written sources and from experience. The experience comes from the interaction with people that we meet and projects that we work on. I am grateful to my former colleagues who I was fortunate to have as my technical mentors on really inspirational projects, first at the Institute of Microelectronic Technologies and Single Crystals, University of Belgrade, former Yugoslavia, then at PMC–Sierra Burnaby BC, Canada, where I gained most of my experiences of the real engineering world.

I would like to acknowledge the contributions of Professor John MacDougall, who initialized and restructured the course into the form of "design and build", and of Professors Alan Webster, Zine Eddine Abid, and Serguei Primak who taught the course at various times.

I would like to thank all of my former and current students who relentlessly keep asking "Why?" and "How did you get this?". I hope that the material compiled in this book contains answers to at least some of those questions and that it will encourage them to keep asking questions with unconstrained curiosity about all the phenomena that surround us.

Sincere gratitude goes to my publisher and editors for their support and making this book possible.

Most of all, I want to thank my wife for being my loyal supporter, and to our son for always hanging around my desk and for making me laugh.

Contents

Abbreviations

AC	Alternating current
A/D	Analogue to digital
ADC	Analogue to digital converter
AF	Audio frequency
AFC	Automatic frequency control
AGC	Automatic gain control
AM	Amplitude modulation
BiCMOS	Bipolar-CMOS
BJT	Bipolar junction transistor
BW	Bandwidth
CMOS	Complementary metal-oxide semiconductor
CRTC	Canadian Radio-Television and Telecommunication Commission
CW	Continuous wave
D/A	Digital to analogue
DAC	Digital to analogue converter
dB	Decibel
dBm	Decibel with respect to 1 mW
DC	Direct current
ELF	Extremely low frequency
EM	Electromagnetic
eV	Electron volts
FCC	Federal communication commision
FET	Field effect transistor
FFT	Fast Fourier transform
FM	Frequency modulation
GaAs	Gallium arsenide
GHz	Gigahertz
HBT	Heterojunction bipolar transistor
HF	High frequency
Hz	Hertz
IC	Integrated circuit
IF	Intermediate frequency
InGaAs	Indium gallium arsenide
InP	Indium phosphide
I/O	Input–output
IR	Infrared

JFET	Junction field-effect transistor
KCL	Kirchhoff's current law
KVL	Kirchhoff's voltage law
LC	Inductive–capacitive
LF	Low frequency
LNA	Low-noise amplifier
LO	Local oscillator
MMIC	Monolithic microwave integrated circuit
MOS	Metal-oxide semiconductor
MOSFET	Metal-oxide semiconductor field-effect transistor
NF	Noise figure
PCB	Printed circuit board
PLL	Phase-locked loop
PM	Phase modulation
pp	Peak-to-peak
ppm	Parts per million
Q	Quality factor
RADAR	Radion detecting and ranging
RF	Radio frequency
RMS	Root mean square
SAW	Surface acoustic wave
SHF	Super high frequency
SINAD	Signal-to-noise plus distortion
S/N	Signal to noise
SNR	Signal-to-noise ratio
SPICE	Simulation program with integrated circuit emphasis
TC	Temperature coefficient
THD	Total harmonic distortion
UHF	Ultra high frequency
UV	Ultraviolet
VCO	Voltage-controlled oscillator
V/F	Voltage to frequency
VHF	Very high frequency
V/I	Voltage current
VLF	Very low frequency
VSWR	Voltage standing wave ratio

Chapter 1
Introduction

Abstract Wireless transmission of information over vast distances is one of the finest examples of Clarke's third law, which states that "any sufficiently advanced technology is indistinguishable from magic". Even though a radio represents one of the most ingenious achievements of humankind and is now taken for granted; for the majority of the modern human population (including some of its highly educated members), this phenomenon still appears to be magical. This chapter introduces fundamental concepts in physics and engineering with the intention of preparing you for the material that follows; in return, that material is expected to reduce, if not completely remove, the "magic" part of the subject.

1.1 Fundamental Concepts in Physics

Although the word *energy*, which derives from the Greek $\varepsilon\nu\varepsilon\rho\gamma\iota\alpha$ (energia), was used by Aristotle way back in the fourth century BC, it is still one of the most ambiguous concepts in science. The limited ability of humans to understand our own reality puts this fundamental idea at the edge of our complexity horizon.

Twenty-four centuries later, this topic was addressed by Feynman in his famous *Lectures on Physics*, where he said:

> There is a fact, or if you wish, a *law*, governing all natural phenomena that are known to date. There is no known exception to this law – it is exact so far as we know. The law is called "conservation of energy"; it states that there is a certain quantity, which we call energy that does not change in manifold changes which nature undergoes. That is a most abstract idea, because it is a mathematical principle; it says that there is a numerical quantity, which does not change when something happens. It is not a description of a mechanism, or anything concrete; it is just a strange fact that we can calculate some number, and when we finish watching nature go through her tricks and calculate the number again, it is the same.

The concept of energy was famously united with the concept of matter by Einstein through his $E = mc^2$ equation. The arena needed to describe the interactions of energy and matter is then set by introducing a medium called *space*. In order to keep these interactions "catalogued", i.e., to be able to tell them apart, the last fundamental concept, the concept of *time*, had to be introduced. With this set of fundamental physical concepts, science is able to develop detailed models that can correctly describe present state and predict the future behaviour of many of the phenomena in this world.

For the purpose of our discussion, we may accept a rather vague definition of energy as "the ability to do work", while the work itself is defined in terms of both time and space. Hence, the process of transmitting (i.e., doing the work of carrying) a bit of information is equivalent to the process of moving a packet of energy from point A to point B in space and time, which brings us back to the main topic of this book. We refer to these streams of energy as "messages" or "signals", originating at

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8_1,

the transmitting side and terminating at the receiving side, with variations that are observed in time. Note that this broad definition does not favour any particular physical form of the signal; it does not matter whether the signals take the form of smoke clouds rising in the sky, a message in a bottle, sound caused by a distant thunderstorm, digital bits of data travelling from one computer to another through the network, or the light arriving to Earth from a star faraway. As long as the message has any meaning to the receiver, we say that signal transmission has taken place.

1.2 Wireless Transmission of Signals

Strictly speaking, wireless transmission of signals, i.e., transmission of signals between two points in space without any visible physical connection between them, has been available to us since the dawn of humanity. Most of us communicate with other people using our voice without additional special equipment. Our vocal cords and hearing system create a wonderful magical communication system; engineering efforts merely represent attempts to increase its range.

In the most general sense, a transmission (communication) system consists of: (a) a transmitter; (b) a transmitting medium; and (c) a receiver (see Fig. 1.1), existing for the sole purpose of moving a message between the transmitter and the receiver. In technical jargon, the vocal cord–ear system is called a "transceiver" because it is capable of both receiving and transmitting a signal, in this case encoded in the form of sound, while the air between the transmitter and the receiver serves as the transmitting medium.

Our bodies are also capable of receiving signals encoded in the form of light, by means of our visual cortex. In this case, only the receiving channel is available to us; for a message encoded in light, the human body is only a receiver—it cannot produce "light rays". The infrared radiation (IR) generated by the body is not really an encoded message—it merely reveals the existence of the source.

Humans have always needed to extend the distance over which messages can travel, which has resulted in the development of various communication systems. For example, carrier pigeons, writing systems, telegraph, radio, television, satellite systems, and cellphones all serve the same purpose of extending the distance that a message created by a person can travel in time and space. The message contained in this book will be received by readers who are widely spread in both time and space.

1.2.1 A Short History of Wireless Technology

In modern terminology, it is assumed that the term "wireless communication" refers to an electronic system for transmitting messages that consists of an electronic transmitter, an electronic receiver, and

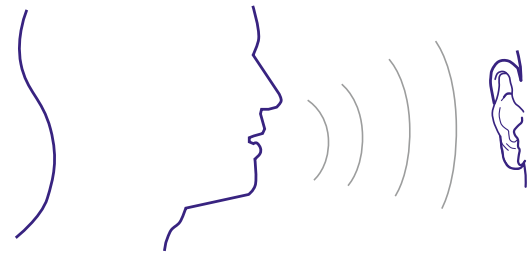


Fig. 1.1 A wireless system consisting of a transmitter (vocal cords), transmitting media (air, in this case), and the receiver (hearing system)

radio waves. While most of us have a vague idea what a radio wave is, it is not that simple to describe in plain words. We leave more detailed description of the waves for the following sections; for the time being, we accept that the term "wave" symbolizes the flow of energy.

In the nineteenth century, interest in the phenomenon of electricity, magnetism and light was at its height. A number of scientists worked on related problems and a long series of studies culminated in Maxwell's equations (Appendix B) of the electromagnetic (EM) field, first published in 1865, which describe electricity, magnetism and light in one uniform system. Consequently, all major laws in electrical engineering can be derived from his equations. In the May 24, 1940 issue of *Science*, Einstein said:

> The precise formulation of the time–space laws was the work of Maxwell. Imagine his feelings when the differential equations he had formulated proved to him that EM fields spread in the form of polarized waves and at the speed of light! To few men in the world has such an experience been vouchsafed. . . . it took physicists some decades to grasp the full significance of Maxwell's discovery, so bold was the leap that his genius forced upon the conceptions of his fellow-workers.

It goes without saying that studying Maxwell's equations and their derivatives is of the highest importance for electrical engineers.

In 1887, Hertz ventured to prove the theory of electromagnetism experimentally, eventually performing his famous "spark experiment" that proved the existence of radio waves, as predicted by Maxwell. His simple experimental setup consisted of a coil and two copper plates with spherical probes connected to a battery. Each time it was turned on and off, this structure would create a spark jumping across the small gap between the spherical probes. A short distance away, there was another copper ring with a short gap between two small spherical terminals. Each time the spark was created in the main apparatus, Hertz noticed a spark in the other copper ring. Wireless transmission was born. As often happens, Hertz did not realize the full practical implications of his discovery; he stated[1]:

> It's of no use whatsoever . . . this is just an experiment that proves Maestro Maxwell was right—we just have these mysterious EM waves that we cannot see with the naked eye. But they are there.

The same year, Tesla, who for most of his life was obsessed with the wireless transfer of energy, was granted a patent on a rotating magnetic field, originally conceived in 1882. By 1891, Tesla had invented the "Tesla Coil", a type of resonant circuit that can produce high-voltage, high-frequency alternating currents (AC), which he proposed could be used for "telecommunication of information".[2] In 1897, Tesla demonstrated the first radio communication system, which he used to control a model boat with his wireless transmitter and receiver (an inductively coupled system),[3] which started the era of practical wireless communications (see Fig. 1.2). On March 20, 1900 Tesla was issued a patent on the radio transmission of electrical energy.[4]

If Tesla is considered the father of practical wireless communications, Marconi should be considered the father of commercial radio communications. In 1901, he demonstrated the first wireless communication system for transmitting Morse-coded messages across the Atlantic. His demonstrations set in motion the wide use of radio for wireless communications, especially with ships (the Titanic disaster also helped the cause). He established the first transatlantic radio service and built the first commercial stations for the British short wave service. It is also recorded in history that Tesla was not pleased with the attention Marconi was getting while using Tesla's patented technology.

---

[1] Hertz, H. (1888) *Annalen der Physik* 270(7):551–569.

[2] Indeed, as late as the 1920s, Tesla coils were used in commercial radio transmitters.

[3] US Patent 613809, November 8, 1898.

[4] US Patent 645576, applied for on September 3, 1897.

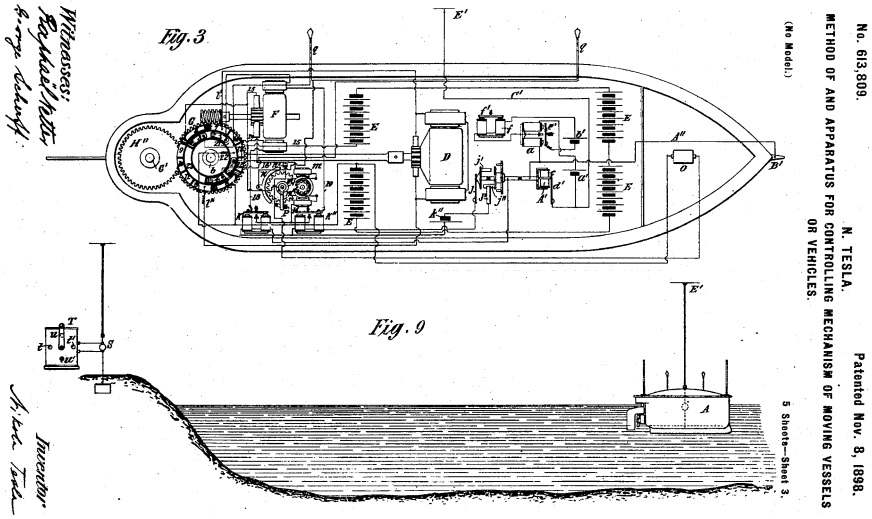


Fig. 1.2 Nikola Tesla's remotely controlled boat using radio waves, first demonstrated in 1897 in the Hudson river, New York

Nevertheless, it took until 1943 for the US Supreme Court to invalidate Marconi's patents in favour of Tesla, stating[5]:

> The Tesla patent No. 645,576, applied for 2 September 1897 and allowed 20 March 1900, disclosed a four–circuit system, having two circuits each at transmitter and receiver, and recommended that all four circuits be tuned to the same frequency. . . . [He] recognized that his apparatus could, without change, be used for wireless communication, which is dependent upon the transmission of electrical energy. [. . .]
>
> Marconi's reputation as the man who first achieved successful radio transmission rests on his original patent, which became reissue No. 11,913, and which is not here in question. That reputation, however well-deserved, does not entitle him to a patent for every later improvement which he claims in the radio field. Patent cases, like others, must be decided not by weighing the reputations of the litigants, but by careful study of the merits of their respective contentions and proofs.

Feud stories like this one repeatedly happen throughout history; bitter rivalry and disputes over important inventions are not exceptions, rather they are the rule. In another example, even though Bell was the first to receive a patent for the invention of the telephone in 1876, several other scientists demonstrated working prototypes as early as 1857 when Meucci developed a voice communication apparatus but, apparently, did not have enough money for the full patent fee. He was granted a *caveat* (i.e., a provisional patent) in 1871, which expired in 1874, leaving an opening for Bell's patent.

The most basic wireless data transmission is possible simply by repeating Hertz's experiment many times, i.e., switching on and off the transmitting coil. Morse was the first to formalize a "time sharing" scheme for encoding a message, famously known as "Morse code". Transmitting Morse code wirelessly requires only a simple tuned circuit being constantly turned on and off. It was not possible to transmit voice messages until 1904 when Fleming invented the thermionic valve (i.e., vacuum tube). This thermal device (which functions as a diode) was the key element needed for radio communication systems. Two years later, the addition of a third terminal was a natural development leading to the invention of a triode (a vacuum tube that functions as an amplifying element) (see Fig. 1.3). Again, Fleming argued bitterly with De Forest about ownership of these ideas. At the same

[5]US Supreme Court (1943) "Marconi Wireless Telegraph Co. of America v. United States". 320 US 1. Nos. 369, 373, April 9–12.

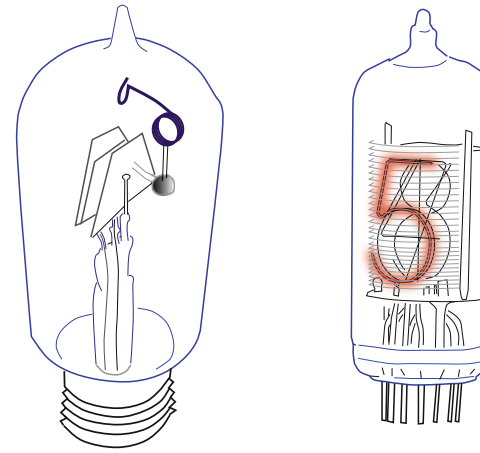


Fig. 1.3 The first electronic valve (*left*), designed by Fleming in 1904, and a modern alphanumeric valve used in electronic equipment to display numbers and letters (*right*)

time, Armstrong (still an undergraduate student) used the triode to create a "regenerative circuit" topology and patented it in 1914.[6] It should be remembered that virtually all modern radio equipment, including the radio receiver topology studied in this book traces its history back to this "heterodyne" topology (later expanded into "superheterodyne").

Although use of the term "radio" may imply the exclusion of television, that is not the case. The television should be looked at as no more than a sophisticated radio. To be historically correct, it has to be stated that television was invented by the many scientists and engineers who made incremental contributions while radio and television systems were being developed in parallel. It is worth mentioning that the first patent for an electro–mechanical television system was granted to Nipkow, a university student, back in 1884. In 1925, Baird demonstrated a system that paved the way to the first practical use of television in 1929, when regular television broadcasts started in Germany.

After the groundbreaking work on radio transmission in the early twentieth century, it is safe to say that there have been no new fundamental advances ever since. Incremental advances can be credited only to engineering ingenuity and technological improvements, most notably the invention of the transistor in 1948 (the three scientists who invented it received a Nobel Prize but never talked to each other again) and the integrated circuit (IC) in 1958 (Kilby, a newly employed engineer at Texas Instruments who did not yet have the right to a vacation, spent his summer working on this concept—it earned him a Nobel Prize and a place in history).

To conclude this short historical review, the importance of radio development is such that most engineers and scientists who have made major contributions also earned a Nobel Prize. They have also served as inspiration for the generations of engineers who have followed in their footsteps.

1.3 Nature of Waves

As hinted at in the previous sections, our understanding of "waves" is more intuitive than exact. Dropping a rock into a pond creates circular ripples that expand both in space and time (which we can visualize, as in Fig. 1.4); a "wave" of spectators can travel around a packed stadium at a soccer game (when each spectator stands up and sits down at the right moment). These familiar examples of waves are perceived by our vision.

[6] US Patent 1113149, October 6, 1914.

Fig. 1.4 Water ripples created in a pond by a small rock. Although the water particles move vertically, the wave expands horizontally while carrying away the kinetic energy of the falling rock. Eventually, the energy is dissipated in the pond and the wave dies out

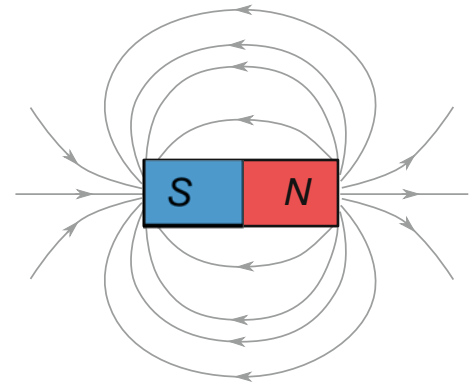


Fig. 1.5 The direction of magnetic field lines represented by the alignment of iron filings sprinkled on paper placed above a bar magnet. The mutual attraction of opposite poles of the iron filings results in the formation of elongated clusters of filings along "field lines"

We are also accustomed to talking about sound waves because we can detect them with our hearing system. It is a bit more complicated to envision sound waves in our mind because we would need to "see" air pressure regions that change from low pressure to high pressure and back. The situation becomes even more difficult if we try to envision light waves. Attempts to explain the nature of light waves led scientists into the development of theories of relativity and quantum mechanics, and touched the deepest questions of human existence.

At the fundamental level, we can accept that water in a pond carries the water waves; we can also accept that sound waves are carried by air; the natural question is to wonder what carries light waves. After all, light waves come from outer space. Is the space empty? What are the waves? When passing through airport security, how does the machine knows whether we are carrying metallic objects without touching us? How does MRI equipment, which always stays outside the body, make detailed pictures of the inside?

To answer these questions and create a meaningful model that correctly describes the observed reality, Faraday introduced the concept of a *field*. This abstract concept, expanded by Maxwell and many others, underlines many of the little mysteries we encounter in everyday life. Although it has proved very useful, being a very abstract concept the concept of field still does not answer the question of what waves are. Nevertheless, it does help us visualize something that otherwise would have been beyond the reach of our senses.

We all remember the magic of the elementary school experiment in which iron filings were sprinkled on paper above a bar magnet. When the paper was shaken, the iron filings aligned along the "field lines" of the magnetic field (see Fig. 1.5). That experiment makes it become obvious why a compass needle placed at a location close to a magnet always takes the direction that is tangential to the field lines. It should be noted, however, that the field lines do not really "exist". Instead, the whole volume of space surrounding the magnet is filled in by the magnetic field, while the strength of the magnetic force measured along the direction from one pole to the other changes in a way that it is

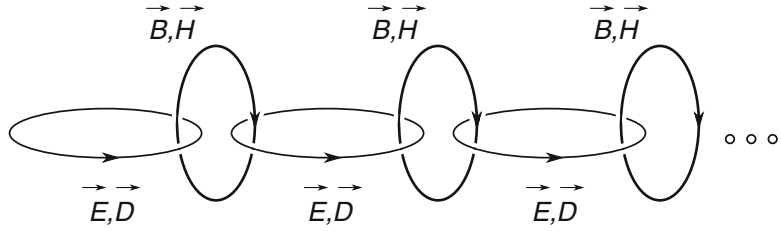


Fig. 1.6 An electromagnetic wave may be imagined as a self-propagating transverse oscillating wave of electric and magnetic fields. Starting, for instance, with a time-varying electric field, magnetic and electric fields are successively generated indefinitely

proportional to the density of the field lines. In other words, the magnetic field is strongest at locations where most of the field lines are crammed together, in this case close to the magnetic poles.

Visualizing the magnetic field certainly helps us to imagine other fields, especially the EM field that was introduced by Maxwell to explain the wave nature of the light. According to Maxwell's equations, a spatially varying electric field generates a time-varying magnetic field and vice versa. As an oscillating electric field generates an oscillating magnetic field, the magnetic field in turn generates an oscillating electric field, and so on. These time-varying fields together form an EM wave that propagates in space (see Fig. 1.6). A less obvious observation is that once the EM wave is established, its source can be removed without further influencing the wave. In free space, an EM wave propagating in the z direction is described as:

$$E_x = E_{0x} \sin(\omega t - \beta z), \tag{1.1}$$

$$H_y = H_{0y} \sin(\omega t - \beta z), \tag{1.2}$$

where E_x is the electric field vector in the x direction, H_{0x} is its maximum amplitude in V/m, H_y is the magnetic field vector in the y direction (which is orthogonal to both the electric field vector **x** and the wave propagation vector **z**), H_{0y} is its maximum amplitude in A/m, ω is the radial frequency in rad/s, and β is the propagation constant defined as

$$\beta = \frac{2\pi}{\lambda}, \tag{1.3}$$

where λ is the wavelength, which is defined in Sect. 1.4.5. Expanding on the propagation constant, we define "phase velocity" v_p as

$$v_\mathrm{p} = \frac{\omega}{\beta}, \tag{1.4}$$

which gives us information about how fast the wave phase propagates in space. A way to visualize phase velocity is by focusing on one single point on the wave (for example, on the crest) and follow it in space.

Motion of electric charges is, by definition, electric current. Electric current creates moving magnetic field, which in return creates moving electric field. Once the process is started, the initial source of this moving EM field (i.e., the moving electric charge) can be removed; the EM field keeps moving in space in self-perpetuating motion. By experiment,[7] the EM wave speed, i.e., its phase

[7] ε_0 is measured through capacitance and dimension of the capacitor (1.12).

velocity, c_0 was found to be the same as the speed of light in a vacuum:

$$c_0 = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} = 299,792,458\,\text{m/s} \approx 3 \times 10^8\,\text{m/s}. \tag{1.5}$$

Maxwell concluded that EM waves (i.e., radio waves) and light are fundamentally the same thing, hence Maxwell's equations deal with this moving EM wave and the relationship between electric and magnetic fields.

Example 1.1. Calculate the wavelengths of EM waves at the following frequencies: $f_1 = 3\,\text{kHz}$, $f_2 = 3\,\text{MHz}$, and $f_3 = 3\,\text{GHz}$.

Solution 1.1. EM waves have phase velocity of $c_0 \approx 3 \times 10^8$ m/s, hence, after substituting (1.4) into (1.3), we write

$$\lambda = \frac{2\pi}{\beta} = \frac{2\pi\, v_\text{p}}{\omega} = \frac{v_\text{p}}{f}, \tag{1.6}$$

which results in

$$\lambda_1 = \frac{3 \times 10^8\,\text{m/s}}{3\,\text{kHz}} = 100 \times 10^3\text{m}; \quad \lambda_2 = \frac{3 \times 10^8\,\text{m/s}}{3\,\text{MHz}} = 100\text{m};$$

$$\lambda_3 = \frac{3 \times 10^8\,\text{m/s}}{3\,\text{GHz}} = 100 \times 10^{-3}\text{m}.$$

An important observation to make is to realize what travels through space. Going back to the ripples in the pond (Fig. 1.4) and dropping a cork into the water, it is easy to see that the cork moves only in the vertical direction, indicating that the water particles move in the same way, i.e., they do not move away from the centre of the ripples. Similarly, the spectators do not run around the stadium—each person only moves up and down in their own seat. That is to say, it is not the particles of matter that propagate through space in the z direction but the wave carrying the energy of the disturbed particles while they vibrate around their nominal positions (in the x or y directions) in synchronicity with their neighbours. These repetitive "up" and "down" vibrations are usually referred to as "oscillations".

1.4 Wave Characteristics

Following the qualitative introduction of waves in the previous section, we now introduce a set of more specific characteristics to help us quantify general wave function properties. It will be shown many times in this book that, in its basic form, any general wave can be represented mathematically as the sum of one or more sinusoidal functions. A vertical cross-section of water ripples, an instantaneous picture of a piano string producing a single note, and the time-domain plot of a voltage signal recorded at the terminals of an electrical resonator all resemble the familiar shape of the sinusoid. In analogy with the sounds of a single note, these single sinusoidal functions are referred to as "single tones" or simply just "tones" (even though we cannot really hear them in their original form).

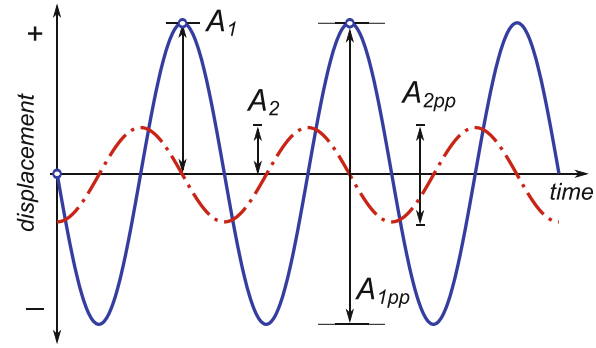


Fig. 1.7 A loud sound is symbolized by a large amplitude (A_1) while a relatively weak sound is quantified by a small amplitude (A_2). By definition, the numerical value of the peak-to-peak amplitude A_{pp} is double that of the single-sided (peak) value A

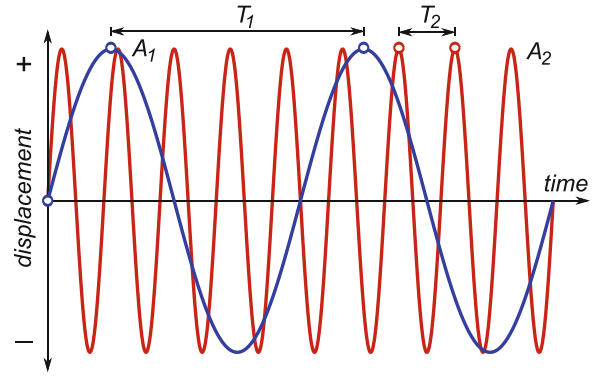


Fig. 1.8 Lower frequency tone (A_1) has longer period T_1 relative to the higher frequency tone (A_2). In this example, the A_2 waveform has a frequency about nine times higher than the A_1 waveform

1.4.1 Amplitude

Exploiting further the analogy of a sound wave created by a piano string playing a single note (for example, A), the wave amplitude is manifested by the volume of the tone. The harder the string is struck, the more violently it vibrates (i.e., the greater the displacement) or, to put it in technical terms, the greater is the amplitude of the sound wave. Figure 1.7 shows the amplitude change in time for two independent sinusoidal waves (A_1 and A_2).

Wave amplitude is quantified in two ways. It can be measured in the positive direction from the zero point (in this case, the average value) to the wave's maximum on the vertical scale (displacement), for example, amplitudes A_1 and A_2 in Fig. 1.7. It can also be measured by the distance between the wave's extreme vertical points, for example, amplitudes A_{1pp} and A_{2pp} in Fig. 1.7, where the index "pp" is pronounced "peak-to-peak". It should be noted that, by definition, the numerical value of the peak-to-peak (PP) amplitude A_{1pp} (or A_{2pp}) is double that of the single-sided (peak) value A_1 (or A_2).

1.4.2 Frequency

Various notes played on a piano, for example A and B, are perceived by our brains as different pitches. This quality of a sound wave is directly related to the amount of time required by the wave to complete one full pattern or, in technical terms, to complete one "period" (measured in seconds). In other words, this is the time required for the string to complete one full movement up, down and back again along the displacement axis in Fig. 1.8. This particular time is marked as T_1 for the A_1 waveform and as T_2 for the A_2 waveform.

A period T is measured between two adjacent extreme amplitude points or at any other two points on adjacent slopes of the same kind (i.e., either two up-slopes or two down-slopes) that have the same displacement value. A shorter period T implies a greater number of patterns being repeated in a given time—the waveform has higher *frequency*. Frequency is measured in *hertz* (abbreviated as Hz), where 1 Hz means that one full wave cycle took a second to complete; in other words, the associated period $T = 1$ s. In Fig. 1.8, the A_2 waveform has a frequency nine times higher than the A_1 waveform. For example, the middle-C tone played on a piano has a frequency of 261 Hz. The full frequency range[8] of piano tones is 27–3,516 Hz. Young people with normal hearing can perceive tones in the range of 20–20,000 Hz. Similarly, human eyes distinguish the various frequencies of light and our brain perceives them as various colours. The visible frequency band for most people is approximately 400–790 $\times 10^{12}$ Hz (i.e., 400–790 THz). This fascinating bandwidth represents almost unlimited resource for signal transmission. As defined, the period and frequency of a waveform are inversely proportional:

$$f = \frac{1}{T} \quad [\text{Hz}], \tag{1.7}$$

where f is the frequency in Hz and T is the period in seconds. A more practical representation of sinusoidal waveforms is based on a mathematical model known as a rotating *phasor*.[9] In a geometrical sense, the time to accomplish one pattern is easily mapped onto the time required to accomplish one full rotation around the circle. The usefulness of the model comes from equivalency between one full movement along the displacement axis and one full circle rotation of the phasor, which is expressed in angle units as 2π, i.e.

$$\omega = 2\pi f = \frac{2\pi}{T} \quad \left[\frac{\text{rad}}{\text{s}}\right], \tag{1.8}$$

where ω is called the "radian frequency". It is important to keep this distinction of radian frequency relative to "frequency" in mind because forgetting the 2π factor is one of the most common mistakes that students make. Also, it is common engineering practice to use the term "single-tone" (or just "tone") while referring to a wave that, mathematically speaking, consists of a single sinusoidal waveform, even in cases when the wave is not a sound wave. The term "wave" refers to the conceptual phenomenon; "waveform" refers to a graphical representation of a wave. These terms are often used interchangeably.

1.4.3 Envelope

Figure 1.9 shows an important case of a waveform. Similarly to an optical illusion drawing, where the presented image is perceived as either of two possible (and completely different) images imbedded in each other, it is valid to ask what you see in Fig. 1.9. Do you see a high-frequency tone whose amplitude varies or do you see a low-frequency tone and its mirrored image?

We should be able to recognize that, indeed, the waveform consists of a high-frequency tone whose amplitude is changing in accordance with a low-frequency sinusoidal function. The low-frequency waveform (not necessarily sinusoidal) that is "riding" on the high frequency peaks is very important in communications; it is referred to as the "envelope" of the high-frequency tone and the high-frequency waveform is referred to as the "carrier".

[8] In radio terminology, "range" means the distance that a waveform can travel. A range of frequencies is referred to as a "band" or "bandwidth".

[9] See Fig. 2.6.

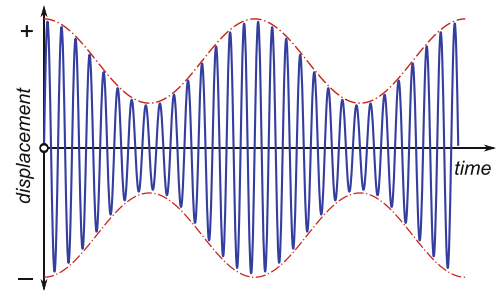


Fig. 1.9 A high-frequency waveform (*the solid line*) and its low-frequency embedded envelope (*the dashed line*)

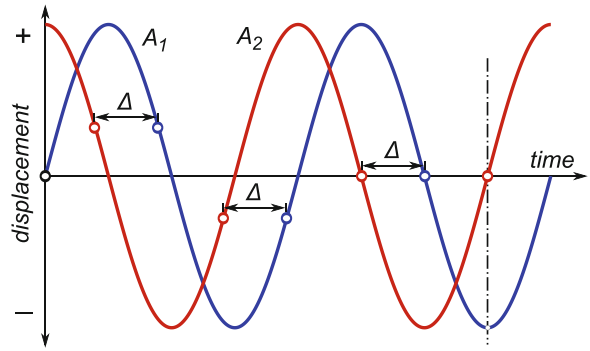


Fig. 1.10 Two single-tone waveforms with normalized amplitudes and the same frequency that have phase difference $\Delta = \pi/2$. The *dashed line* shows that the peak of the A_1 waveform coincides with the cross-over point of the A_2 waveform, i.e., the phase difference is $\Delta = 1/4$ of the period. When two signals are in this relationship, they are said to be "in quadrature"

Theoretical and practical techniques for imprinting an arbitrary envelope over a carrier (that originally had a constant amplitude) and for extracting the envelope and discarding the high-frequency carrier are the main subjects not only of this book, but also of the radio frequency (RF) circuit design field in general. The process of "imprinting" the envelope signal into the carrier is referred to as "modulation" and the process of envelope extraction is known as "demodulation". We will devote a large portion of this book to these two processes.

1.4.4 Phase, Group, and Signal Velocity

A stand-alone single-tone wave is fully described by its amplitude, frequency (or, equivalently, its period), and *phase*. The concept of a phase is derived from the rotating phasor model and it assumes sine (as opposed to cosine) as the default waveform function because at time $t = 0$ its phase is zero. Consequently, one period T in the time domain is mapped onto an angle of a circle, i.e., $T = 2\pi$ radians (or 360°). Note that numerical value for *time* T (measured in seconds) is scaled to a number 2π (with no unit); these two measuring units are used interchangeably. Usually (but it is not mandatory), the phase is measured at a point in time $t = 0$. Since the initial value of a sine function is zero ($\sin 0 = 0$), its initial angle (or phase) is $\phi = 0$. In Fig. 1.10, the A_1 waveform has phase $\phi = 0$ and the A_2 waveform has phase $\phi = \pi/2$ (or 90°) because $\sin \pi/2 = 1$, which is the initial value of its sine function.

A "phase" is a relative term, hence, it makes much more sense to define the term "phase difference", which implies existence of the second wave. This requires the introduction of a fictional arbitrary reference plane that serves as the "zero phase". With two sinusoidal waves, once the amplitudes are

normalized, it is important to compare their frequencies. If the frequencies are not the same than there is not much left to say, but to note existence of two relatively independent waves. However, if the two waves do have the same frequency (and not necessarily the same amplitude), than it makes sense to ask which wave arrives first.

To answer that question, let us set up a "race". The initial phase of one of the two waves is arbitrarily declared the zero-phase reference. The stopwatch starts when the first wave's amplitude crosses the zero-amplitude value, for instance, on its way from higher amplitudes to lower amplitudes (in technical terms, on its "falling edge"). The stopwatch stops when the second wave's amplitude crosses the zero value of its amplitude on its falling edge. Due to the fact that only either falling or rising edges are used as the start–stop triggers, the relative "timing difference" between the two waves must be within the range 0 to T (T is the common period of the two waveforms). It should be obvious that the result of the race depends on neither the absolute value of the amplitude cross-over point (i.e., it does not have to be zero) nor the choice of the rising or falling edge.

Therefore, under the condition of frequency equality, the phase difference is either "constant" (see Fig. 1.10) or, by definition, it does not exist. When it does exist, it is said that one wave either "leads" or "lags" the second one by Δ seconds (or, equivalently, by Y degrees). It is important to keep in mind that this measurement is "relative". In a practical sense, it is much easier to express the phase difference as a "fraction of the period", i.e., in degrees, instead of using the absolute time units. To illustrate the point, saying that the phase difference between two tones is, for instance, 5 ns still requires additional information about the frequency value. Even then, it is not easy to visualize the size of the 5 ns time difference relative to, for instance, a 100 Hz waveform or a 100 MHz waveform. However, saying that the phase difference is $\pi/2$ instantaneously brings into our mind a mental picture of two sinusoidal waves (see Fig. 1.10), where the peaks of one wave coincide with the zero-crossing points of the second, regardless of the wave frequency. In that case, the phase difference is one quarter of the cycle, i.e., 90°, and the two waves are said to be "in quadrature". The quadrature signals are very important and widely used in radio communication systems.

Example 1.2. Calculate differences in the times of arrival Δt for EM wave pairs with a phase difference of $\Delta\phi = \pi/2$ at each of the following three frequencies: $f_1 = 1\,\text{kHz}$, $f_2 = 1\,\text{MHz}$, and $f_3 = 1\,\text{GHz}$.

Solution 1.2. First, we convert the given frequencies into their equivalent periods as:

$$T_1 = \frac{1}{f_1} = \frac{1}{1\,\text{kHz}} = 1\,\text{ms}; \quad T_2 = \frac{1}{f_2} = \frac{1}{1\,\text{MHz}} = 1\,\mu\text{s}; \quad T_3 = \frac{1}{f_3} = \frac{1}{1\,\text{GHz}} = 1\,\text{ns}$$

then, by knowing that period $T \equiv 2\pi$ (i.e., one full cycle), we conclude that $\pi/2$ is equivalent to $T/4$. Therefore, in the time domain, the phase differences translate into

$$\Delta t_1 = \frac{1\,\text{ms}}{4} = 250\,\mu\text{s}; \quad \Delta t_2 = \frac{1\,\mu\text{s}}{4} = 250\,\text{ns}; \quad \Delta t_3 = \frac{1\,\text{ns}}{4} = 250\,\text{ps},$$

which illustrates how the phase difference translates into time-of-arrival differences at various frequencies.

1.4.5 Wavelength

An obvious, but often ignored, fact is that Fig. 1.4 shows a wave frozen in time: after the water wavefront has travelled outwards in space from the point where the rock hit the water to its last position. Again, keep in mind that (ideally) the water particles have only vertical movement, i.e., it

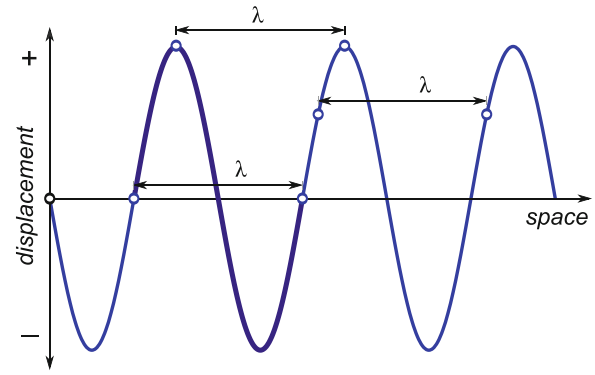


Fig. 1.11 By measuring the distance in space between the peaks (or any other pairs of equivalent points, as shown) the spatial dimension, wavelength λ, is established. One wave cycle is emphasized by a *bold line*

is only the displacement (energy) that travels horizontally. From Fig. 1.4, it is possible to measure the horizontal distance between any two wave peaks in *space*. This spatial dimension is denoted as wavelength λ. If, instead of a single frame, the full movie were available to us, then it would be possible to measure the same event in the time domain, namely, the period T for any given particle of water to complete the full up, down and back again vertical swing. By now, a careful reader should realize that the period T is the same time taken by the wavefront to travel distance λ. Figure 1.8 shows the vertical displacement of a single wave particle in time and Fig. 1.11 shows the vertical displacement of all wave particles in space (measured horizontally from the wave starting point to its end). It is important to keep a mental picture of these two views showing the same event .

As in any other case of linear motion in classical physics, knowing two of the three parameters (i.e., the distance travelled, the time taken for the trip, and the average speed) enables calculation of the third parameter. Using experimental methods, wave propagation speeds for sound and light waves through various materials were established. For instance, it was established that a sound wave travels at a speed of 343 m/s through dry air at 20°C, or approximately one kilometer in three seconds. Similarly, the speed of a light wave in a vacuum was established as 299,792,458 m/s, which is often rounded to 300,000 km/s. At this speed, it takes sunlight 8 min and 19 s to reach Earth.

Example 1.3. Estimate the distance of lightning if approximately nine seconds pass between the time you registered the lightning flash and the time you heard the thunder.

Solution 1.3. The speed of light, for purposes of this problem, is infinite (unless we use very precise measuring equipment); in 9 s, sound travels approximately 3 km. Therefore, we can ignore the delay in light travelling 3 km (which is about 10 µs) and just estimate that the lightning happened approximately 3 km away.

In mathematical terms, wavelength is expressed by the equation

$$\lambda = vT = \frac{v}{f} \quad [\text{m}], \tag{1.9}$$

where λ is the wavelength (i.e., the horizontal distance travelled by the disturbance while completing one full cycle of the vertical disturbance) in meters, T is the time in seconds needed by the waveform to travel the horizontal distance λ while completing one full vertical cycle, and v is the wave propagation speed[10] in meters per second (denoted by c in the special case of the speed of light).

It should be noted that it is the frequency of the wave that determines the pitch (or colour). The wavelength is a secondary phenomenon depending on the speed of the wave. To support this

[10]The correct term should be velocity, but most books (wrongly) use the speed instead.

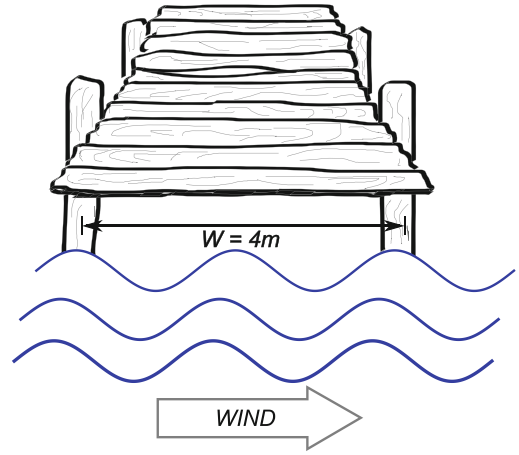


Fig. 1.12 Illustration for Example 1.5

observation, imagine sending a sound of the same frequency through two parallel channels, water and air. Even though the speed of sound in water is more than four times greater than the speed through air (and, therefore, there is more than four times the wavelength), the perceived tone at the receiving end remains the same in both cases, as confirmed by (1.9)—when $\lambda \to 4\lambda$ and $v \to 4v$, the frequency stays the same.

Example 1.4. For a voltage disturbance wave travelling at the speed of light and described as $v_1 = \sin(20\pi \times 10^6 t)$ find:

(a) Its maximum amplitude
(b) Its frequency
(c) Its period
(d) Its wavelength
(e) Its phase at time $t = 0\,\text{s}$

For a second wave v_2 with the same maximum amplitude and with a phase difference of $\Delta\phi = +45°$, find its amplitude at time $t = 0$ and the distance between one of its peaks and the following v_1 peak.

Solution 1.4. Inspecting the wave v_1 equation, we can write:

(a) Maximum value: $A_\text{m} = 1\,\text{V}$
(b) Radial frequency: $\omega = 2\pi f = 20\pi \times 10^6\,\text{Hz}$, therefore $f = 10\,\text{MHz}$
(c) Period: $T = 1/f = 100\,\text{ns}$
(d) Wavelength: $\lambda = cT \approx 30\,\text{m}$
(e) Phase at $t = 0$: $\phi = 0$

The second wave is leading with a phase difference of $\Delta\phi = \pi/4 = T/8$, whose amplitude at $t = 0\,\text{s}$ is $v_2 = \sin(\pi/4) = 1/\sqrt{2}\text{V} \approx 0.707\,\text{V}$. The phase difference is $T/8$ in the time domain, therefore in space domain it has to be $\lambda/8 = 3.75\,\text{m}$.

Example 1.5. Imagine standing at a lakeshore's boat dock that is $w = 4\,\text{m}$ wide, watching passing waves created by a wind (see Fig. 1.12). Assuming that you have only a stopwatch, explain the procedure to estimate the wavelength, frequency, period, and velocity of the passing waves.

Solution 1.5. First, observe and count how many wave crests fit along the dock's $w = 4\,\text{m}$ side. According to Fig. 1.12 there are approximately three crests from edge to edge. Compare the wave shape with Fig. 1.11 and conclude that the dock's width equals two wavelengths, i.e., $w = 2\lambda$,

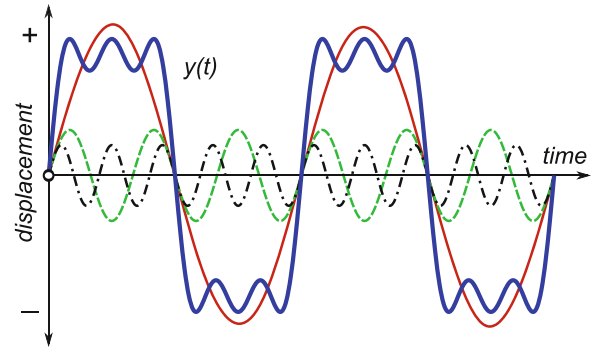


Fig. 1.13 A complicated waveform (*solid dark line*), created by linear addition of its first, third, and fifth harmonics. Increasing the number of harmonics to 15 or more would create an almost-perfect, square-pulse waveform

therefore $\lambda = 2\,\text{m}$. Second, using a stopwatch, measure the time taken for one crest to travel along the deck's edge. If the measured time is $t = 4\,\text{s}$, then the wave period is $T = t/2 = 2\,\text{s}$, because the total distance travelled by the wave is equal to two wavelengths. Third, it is now straightforward to calculate the frequency of the wave as $f = 1/T = 0.5\,\text{Hz}$. Finally, from (1.9) it follows that the wave propagation speed is $v = \lambda f = 1\,\text{m/s}$.

1.4.6 Multitone Waveform

By now we should be comfortable with using terminology related to a single-tone waveform and should be moving on to waveforms whose shapes do not have a simple sinusoidal form. For instance, take a look at the waveform in Fig. 1.13 shown by the solid dark line. It looks more like a square-pulse waveform than a single tone. A brilliant intuition led Fourier to speculate that an arbitrary waveform, which is a typical shape found in nature, is composed of more than single-tone waveforms. Eventually, he proved the idea and earned his space in history by developing the "Fourier transform", which is known to virtually every engineer and scientist in the world.

A very liberal interpretation of the Fourier transform is that any arbitrary waveform can be synthesized from an infinite number of harmonics added together in a certain proportion, as prescribed by a formula that was delivered for that particular waveform. We start with a single tone whose frequency is ω, referred to as the first harmonic, the second harmonic is a single-tone sinusoidal waveform whose frequency is 2ω, the third harmonic is a single-tone sinusoidal waveform whose frequency is 3ω, and so on. All single-tone terms in a Fourier transform (referred to as "harmonics") are then appropriately scaled in amplitude and added together (1.10).

Using a Fourier transform, the squarish looking waveform $y(t)$ in Fig. 1.13, is synthesized by using only the first three odd harmonics as

$$y(t) = \frac{4}{\pi}\left[\sin\omega t + \frac{1}{3}\sin 3\omega t + \frac{1}{5}\sin 5\omega t + \cdots\right], \tag{1.10}$$

where sin terms, together with their respective frequencies $n \times \omega$ ($n = 1, 3, 5, \ldots$) and amplitude scaling coefficients (1, 1/3, 1/5, ...) multiplied by $4/\pi$ represent harmonics of the waveform $y(t)$. One way to interpret (1.10) is to say that waveform $y(t)$ is constructed using the three single-tone signals as its basic building blocks. In a way, the Fourier transform serves a similar role for a waveform as an X-ray machine does for a body: it shows what a complicated waveform is made of. Or, to put it in technical terms, the frequency spectrum of the waveform $y(t)$ in (1.10) consists of the first, third, and fifth harmonics.

We refer back to this section a number of times in the rest of the book because it is one of the most important concepts in signal processing and *RF* circuit design.

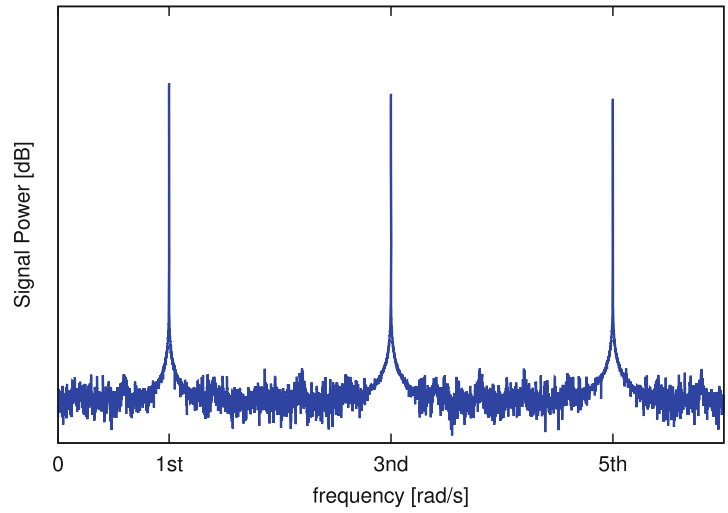


Fig. 1.14 The power spectrum plot of the $y(t)$ waveform in Fig. 1.13 and (1.10), showing the three single tones (harmonics) in the frequency domain and the noise floor. (Note that, due to the scale of the vertical axis, relatively small differences in powers of the three tones are not clearly visible.)

1.4.7 Frequency Spectrum

Once the familiar sinusoidal shape becomes a permanent image in our mind, there is no point in looking at its time domain plot. The shape of a sinusoid is always the same and all that we need to describe it are the three numbers representing its amplitude, frequency and phase. If the sine plot axes are labelled, the three numbers are found by inspection.

However, a more complicated waveform, such as $y(t)$ in Fig. 1.13, is not that easy to analyze by visual inspection only, because it is defined by its amplitude, frequency and phase parameters in the time domain plot. Instead, it is much more important to know its frequency spectrum, which may contain many tones in an infinite number of combinations, as implied by (1.10). It is very useful to create a plot that shows the relationship among all the harmonics in the frequency–power domain. To illustrate the point, a fast Fourier transform (FFT) numerical algorithm is applied to time domain waveform data, as calculated by (1.10), in order to transform it into its equivalent frequency spectrum function, Fig. 1.14. For the purpose of frequency domain plots, it is common practice to convert units of amplitude (for instance, volts or amperes) into *decibels* (dB) (a unit of relative *power*). In Sect. 2.7.2, we introduce definitions and units for power calculations in more detail.

The graph in Fig. 1.14 is interpreted as follows. Starting from the zero frequency point (i.e., DC) and moving along the horizontal axis (scaled in units of rad/s), each point of the graph symbolizes its respective $(x, y) = (\mathit{frequency}, \mathit{power})$ pair of numbers. In other words, each pixel of the curve shows individual power levels for each of the infinite possible single tones within this frequency band. The three distinct vertical lines represent the three waveform harmonics, each with its dB power level quantified by the highest vertical point. It should not be difficult to realize that there is a "sea of noise" (in technical terms, a "noise floor"), caused by various random sources that exist all around us, of which the level is relatively consistent. Since we started with three ideal single tones and nothing else, it is no surprise that they are far stronger than any other single tone at and below the noise floor. A more detailed introduction of the noise floor is left for Sect. 13.3.1.1.

It should be emphasized that detailed examination of a complicated waveform includes both time and frequency domain analysis. To help the process, an oscilloscope is a test instrument that serves as a time domain waveform plotter and a spectrum analyzer is a test instrument that performs real-time Fourier transformation of the given waveform and displays its "power spectrum plot". It is assumed

that all engineers and scientists are familiar with these two test instruments that enable us to see two distinct, yet complementary, perspectives of the same waveform, in the same way as we need to see all three projections of a solid before concluding its 3D shape.

A fine distinction needs to be noted. The noise floor in Fig. 1.14 is ideal, limited only by the numerical resolution of the algorithm and computer used for the calculations. Hence, the calculated noise floor is referred to as a "numerical noise". For the same waveform, a spectrum analyzer shows a similar plot, with the distinction that the measured noise floor is real. We keep in mind that the real, measured noise floor is expected to be much higher than the ideal, numerical noise floor. It is fair to say that a spectrum analyzer is among the most sophisticated and precise instruments ever invented.

1.5 Electromagnetic Waves

In our discussion so far, both sound waves and light waves have been used to introduce and define basic phenomena and definitions related to wave propagation.

Let us first elaborate a bit more on the process of creating sound waves. Our eardrums are extremely sensitive organs capable of distinguishing a huge range of sound intensities. For instance, the softest sound a healthy young human can detect is measured as a pressure of 20 μPa, which is so small that it causes the eardrum to move a distance of less than one tenth the diameter of a single hydrogen molecule. At the other extreme, the threshold of pain is measured as a pressure of 63.2 Pa, that is to say 3,160,000 times louder than the softest detectable sound (in technical terms, eardrums have a dynamic range of 130 dB; more details are given in Sects. 2.7.2 and 13.3.1).

Sound waves start with vibrations of our vocal cords that cause surrounding air molecules to start moving. First, the molecules in the air layer closest to the cords are pushed away; the molecules a bit further away (in the next layer of air) still have not detected any movement, which causes the neighbouring air layers to move closer to each other. This phenomenon is due to the fact that the mass of air molecules is still finite, although very small, which means that they exhibit some inertia and do not instantaneously change the direction of their movement. Keep in mind that shortening the distance between air molecules is also interpreted as an increase in pressure. By the time the first air layer starts its outward movement, the cords have changed their direction of movement by 180° and, instead of "pushing", they start "pulling" the surrounding air molecules, which must follow (nature does not like a vacuum very much). It is not difficult to envision this chain of push–pull actions spreading in all directions and affecting more and more air layers, causing alternate spherical layers of high and low air pressure to move outward (see Fig. 1.15). The imaginary expanding sphere that separates air still not affected from the sound wave inside is the wavefront and its speed is what we refer to as the "propagation speed"—the speed of sound, in this example. Eventually, the wavefront, followed by the wave of high- and low-pressure layers, reaches our eardrums enabling the cochlea (which is our own natural spectrum analyzer) to measure the wave's frequency, which is then perceived by our mind as a tone of a certain pitch. As a reminder, the sound wave's wavelength λ is the distance between two adjacent high-pressure or low-pressure layers (see Fig. 1.15).

To further exploit this analogy, imagine that you are trying to create sound waves by flapping your arms. We can flap our arms only a few times per second, which is too slow for the air molecules to start moving. Instead, the air molecules slip along the skin and allow the arm to pass easily. Obviously, there are two options for creating air waves: (a) start flapping much faster, so that the air molecules are "hit" in the same way as flying insects do with their wings; and/or (b) drastically increase the size of your hands, say up to hundreds of square meters, so that a large volume of air (hence, large inertia) is moved back and forth by your paddling. In the former case, we would be able to detect the waves with our ears because the flapping frequency would be within our hearing range (we can hear buzzing

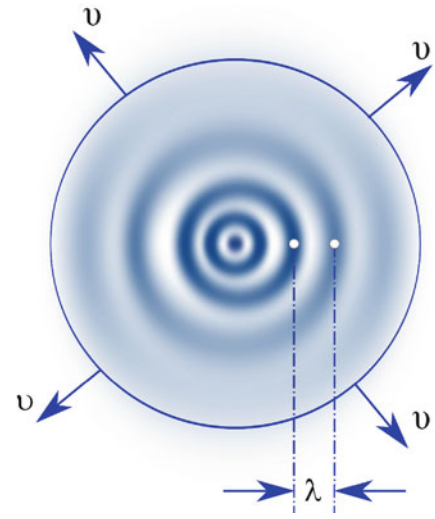


Fig. 1.15 A spherical wave spreading in all directions. The imaginary expanding sphere (represented by the *blue circle*) that separates space still not affected from the wave inside is the wavefront. Its velocity determines the wave propagation speed v. The wavelength λ is the spatial distance between two subsequent crests

Table 1.1 Classification of radio frequency bands

Frequency band	Abbreviation	Frequency range	Typical application
Extremely low	ELF*	3.0–30 Hz	Military underwater communications
Super low	SLF*	30.0–300 Hz	Military underwater communications
Ultra low	ULF*	0.3–3 kHz	Military underground communications
Very low	VLF	3.0–30 kHz	Submarine navigation
Low	LF	30.0–300 kHz	LORAN, time signals
Medium	MF	0.3–3 MHz	AM broadcasting, radio beacons
High	HF	3.0–30 MHz	Amateur radio
Very high	VHF	30.0–300 MHz	Short-distance terrestrial communication
Ultra high	UHF	0.3–3 GHz	TV broadcasting, cell phones
Super high	SHF	3.0–30 GHz	Wireless LAN, satellite links
Extremely high	EHF	30.0–300 GHz	Radio astronomy, research, military

* The whole earth may serve as an antenna

of a flying insect). In the latter case, however, even though the waves would be created, we would not be able to hear them because their frequency would be below our hearing range. On the other hand, they would cause neighbouring doors and windows to rattle and rumble.

In a very similar way, electric waves also create disturbance. Alas, because of the great speed of electric waves, the antenna (which is equivalent to your waving hand or an insect's flapping wings) must "flap" proportionally faster. Unlike with sound waves, however, air molecules cannot respond that fast. To radiate radio waves at audio frequencies, an antenna would have to be of the order of kilometers, so one might just as well (and more conveniently) use it at ground level as a telephone line. Fortunately, it turns out that RF waves do not need air to propagate. If anything, they travel much easier through a vacuum. To summarize, the fundamental difference between sound waves and RF waves is that sound waves propagate by *mechanical* vibration while RF waves radiate from antennas in the form of EM waves, i.e., light.

Now we have learned how to create RF waves at any frequency, and considering that they need less than a tenth of a second to circle the globe, they definitely qualify for the job of our chief message carrier. With Maxwell's help, we have learned that RF waves are of the same nature as light, the only difference being their frequencies (i.e., wavelengths). Because of the very wide range of useful frequencies used by radio communication systems, the most common frequency bands are categorized into sub-bands as shown in Table 1.1.

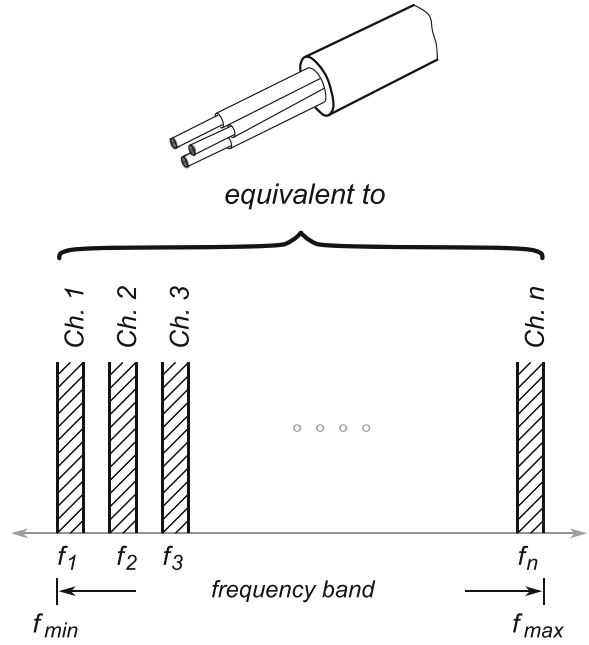


Fig. 1.16 Frequency band with multiple channels compared to a multi-wire communication cable

Relative to the complete frequency spectrum known to exist in nature, our natural wave receptors (ears and eyes) cover only two minor frequency ranges. In addition, there is a relatively big gap between the two frequency bands. It is no surprise that most of our engineering efforts go into building artificial wave receptors that operate in our "blind spots" and enable us to "see" the full EM spectrum.

1.5.1 Tuning

Even if it were practical to build wireless systems that operate in the audio frequency band for communications over longer distances than those achieved by natural speech, it is easy to see the immediate practical problem. We would create a world crammed with very loud giants all talking at the same time, all the time. Moreover, we would be able to hear them without the assistance of any artificial equipment.

Human speech, including music, requires a very narrow frequency band (only about 20 kHz wide, known as the "audio band"). In comparison, the EM spectrum is immense. Splitting that enormous frequency space into abutting "strips" 20 kHz wide would create many parallel "pipes" each of them wide enough to conduct the full audio spectrum. It is easy to show that the number of possible pipes (i.e., audio communication channels) is more than sufficient for human needs. It is important to note that these communication channels are strictly separated only in the frequency domain; in real space they co-exist at all times and everywhere (Fig. 1.16). Having the ability to visualize the same signal in all three of these domains, i.e., frequency, time and space, is essential to understanding wireless communication systems. Only then is it possible to understand how each of these communication channels could be made to connect a receiver with a specific transmitter that is located somewhere in space while, at the same time, an arbitrary number of other transmitter–receiver pairs also maintain their connections. With that in mind, it is not difficult to imagine a wireless communication system where each transmitter–receiver pair is assigned its own frequency channel for the duration of the communication.

However, in order for this multiple frequency band approach to be practical and for us to make use of it, we need to resolve the following design issues: (a) how to shift the frequency of each individual audio signal up and align it exactly with its assigned channel; (b) how to force the receiving equipment to listen only to that particular channel, while ignoring communications in all other parallel channels; and (c) how to shift the frequency of the received signal back to the audio range and decode the original message. Solutions to these three fundamental steps are required for virtually all communication systems invented so far.

A practical solution to the problem was enabled by the invention of *tuning*. In the case of mechanical sound waves, if one object of a certain size is made to vibrate and produce a tone, bringing a second object of similar size close to the first one causes the second object to vibrate with the same frequency: consider tuning forks used by elementary school music teachers. This phenomenon is known as "resonance" and the key point is that the two forks are of "similar mechanical size". It is possible to create equivalent conditions using EM waves; if one electronic circuit is made to vibrate (we say, "to oscillate"), then a second circuit of similar "electrical size", and within a certain distance, oscillates with the same frequency. If the two "electrical sizes" do not match, then the second circuit does not oscillate. Tuning (i.e., resonance) is one of the most important and most fundamental phenomena in nature. It is quite possible that without it our universe would not have existed, let alone practical wireless communication.

Engineering creativity supported by mathematical analysis is needed to remove the last obstacle in our quest for practical RF communication—the invention of a practical device capable of precisely shifting audio information up and down the frequency domain. The rest of this book deals with both the theoretical background and the practical implementation of electronic circuits for modulation, tuning, and frequency shifting.

1.5.2 Maxwell's Equations

In this section, we review the basic definitions and terminology associated with EM waves. Existence of the interleaved self-perpetuating magnetic and electric fields (Fig. 1.6) is fundamental to the propagation of EM waves and, therefore, to wireless communication systems. Their relationship is described by the set of Maxwell's equations.

The main goal of this book is to provide a first introduction to the fundamental principles of RF circuit design to students who have taken only introductory courses in electronics, without all the complications associated with extremely high-frequency field (EHF) theory and circuit design specifics. Because there is a big gap in the required theoretical background, complexity and design methods between linear low-frequency circuits and, for instance, the millimeter wave RF circuits (see Fig. 1.17), we focus only on relatively low-frequency non-linear RF circuits. The analysis of EHFs is difficult and not always necessary. By focusing on low-frequency RF circuits, we are able to use most of the methods acquired from previous courses and to apply approximate methods derived from Maxwell's equations under condition of low frequencies (see Fig. 1.18).

For the sake of completeness, let us review the basic definitions from EM theory related to the EM field and Maxwell's equations.

1.5.2.1 Magnetic Field

We determine the existence of a *magnetic field* in the space where a magnet is acted upon by a magnetic force. The magnetic field is quantified by two properties: "magnetic intensity" or "field strength" **H**, which is measured by the force acting on the magnet (in units of ampere-turn per meter,

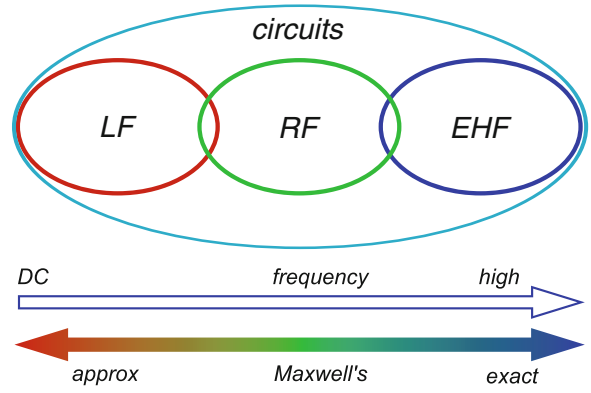


Fig. 1.17 The relationship between EHF, RF, and LF design methods with respect to exact and approximated Maxwell's equations

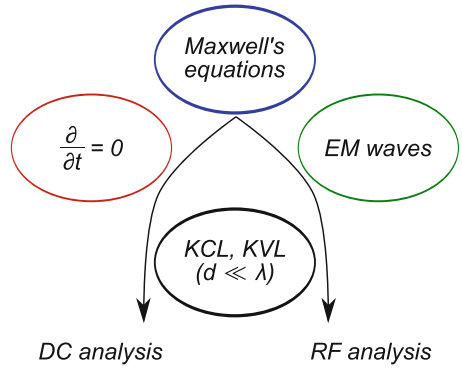


Fig. 1.18 The role of Maxwell's equations relative to electrical circuit analysis. After the low-frequency approximation is applied, i.e., $d \ll \lambda$ (that is, the transmission distance is much smaller than the signal wavelength), Kirchhoff's current law (KCL) and Kirchhoff's voltage law (KVL) equations are used instead of the full set of Maxwell's equations

A/m); and "magnetic induction" or "magnetic flux density" **B**, which is measured by a force acting upon moving electrical charges (in units of Tesla, T). Permeability of the medium where the magnetic field exists is defined as the ratio $\mu = \mathbf{B}/\mathbf{H}$. That is to say, the material dependent multiplication constant μ is referred to as "magnetic permeability" and shows how much **B** is modified by the material. In a vacuum, the two vectors **B** and **H** are identical, except for the vacuum permeability constant $\mu_0 = 4\,\pi \times 10^{-7}$ H/m (which is fixed through the definition of an ampere). In conclusion, a magnetic field is a field of force produced in two ways: by moving electric charges, i.e., by electric fields that vary in time, or by the "intrinsic" magnetic field of elementary particles associated with the spin of the particle.

1.5.2.2 Electric Field

We determine the existence of an electric field in the space where an electric charge is acted upon by an electric force. Similarly to a magnetic field, an electric field is quantified by two properties: "electric intensity" or "electric field strength" **E**, which is measured in units of volts per meter (V/m); and "electric flux density" or "induction" **D**, which is used to account for free charges within materials, also referred to as an "electric displacement field", measured in units of coulombs per square meter (C/m^2). The two electric fields **E** and **B** are connected through the permittivity constant ε as $\mathbf{D} = \varepsilon\mathbf{E}$. In a vacuum, the vectors **E** and **B** are identical, except for the vacuum permittivity constant ε_0.

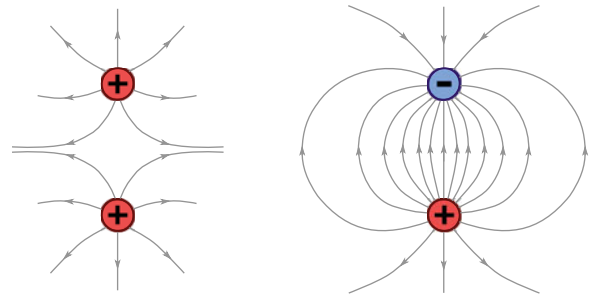


Fig. 1.19 An electric field and the induction lines between two point charges

As opposed to a magnetic field, which originates and ends at the same magnetic dipole (i.e., it is not possible to separate the "north" and "south" sections of a magnet), an electric field originates at positive charges and sources at negative charges, regardless of the separation distance (see Fig. 1.19).

Example 1.6. Derive an expression for the capacitance of a parallel-plate capacitor with a plate area of S and distance d between them, where the plate separation d is much smaller than the plate side $d \ll \sqrt{S}$, i.e., the fringe electric field is ignored.

Solution 1.6. If the two plates carry charges of $+q$ C and $-q$ C respectively, the flux density at any point between the plates is

$$D \equiv \frac{q}{S} \quad \left[\frac{\mathrm{C}}{\mathrm{m}^2}\right] \quad \therefore \quad E = \frac{D}{\varepsilon} = \frac{q}{\varepsilon S} \quad \left[\frac{\mathrm{V}}{\mathrm{m}}\right] \tag{1.11}$$

because the electric field between the two plates is constant and homogeneous, the potential difference between the plates is

$$V = E\,d = \frac{q}{\varepsilon S} d \quad [\mathrm{V}] \quad \therefore \quad C \equiv \frac{q}{V} = \varepsilon \frac{S}{d} \quad [\mathrm{F}]. \tag{1.12}$$

Example 1.7. Derive an expression for the capacitance of a co-axial capacitor, where the inner cylinder has radius a m and the outer cylinder has radius b m. The capacitor is l m long. Again, ignore the fringe electric field at the ends of the capacitor.

Solution 1.7. If the two cylinders carry charges of $+q$ C and $-q$ C respectively, the induction field is radial, hence at point r between the plates, i.e., $(a < r < b)$ the area of interest is $S = 2\pi rl$, leading to

$$D \equiv \frac{q}{S} = \frac{q}{2\pi r l} \quad \left[\frac{\mathrm{C}}{\mathrm{m}^2}\right], \tag{1.13}$$

$$\therefore$$

$$E = \frac{D}{\varepsilon} = \frac{q}{2\pi r l \varepsilon} \quad \left[\frac{\mathrm{V}}{\mathrm{m}}\right]. \tag{1.14}$$

Because the electric field is not constant (it is radial and varies with r), we write an expression for the potential difference between the plates as

$$\mathrm{d}V = E\,\mathrm{d}r \ \ \mathrm{V} \quad \therefore \quad V = \int_a^b E\,\mathrm{d}r = \int_a^b \frac{q}{2\pi l \varepsilon} \frac{\mathrm{d}r}{r} = \frac{q}{2\pi l \varepsilon} \ln \frac{b}{a} \quad [\mathrm{V}], \tag{1.15}$$

$$\therefore$$

$$C \equiv \frac{q}{V} = \varepsilon \frac{2\pi l}{\ln \frac{b}{a}} \quad [\mathrm{F}]. \tag{1.16}$$

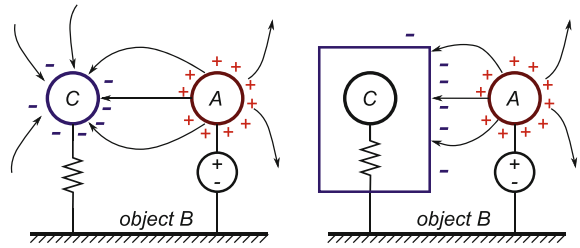


Fig. 1.20 Electrical shielding

It is important to note that if there is a potential difference between any two conductors, or two points on the same conductor, then an electric field exists between them. The electric field can be visualized as induction lines that terminate on induced charges in the conductors, therefore a "parasitic capacitance" must exist between two points at different potentials. If the points are on the same conductor, than the conductor has "self-capacitance".

1.5.2.3 Electrical Shielding

In most cases, the unavoidable parasitic capacitance is not desired, which calls for some form of isolation between two objects at different potentials. Consider the rather realistic situation in Fig. 1.20 (left), where an object A is at a higher potential than an object C, both referenced to an object B that serves as the local reference, i.e., "ground". Induction lines, therefore, originate at A and terminate at B and C, which is to say that there is parasitic capacitance created by these three objects. If the goal is to isolate object C from the influence of A, an additional "shield" must be added around C (as in Fig. 1.20 (right)). Ideally, the shield must be at the same potential as the ground, so that the induction lines are given an opportunity to terminate other than at C. In practice, sensitive electronics are literally encased in a metal box (sometimes a partial metal screen can be used) that is connected electrically to the local ground potential.

1.5.2.4 Magnetic Shielding

Magnetic shielding is a bit more complicated than electrostatic shielding and is never perfect. Magnetic induction lines cannot be terminated, only diverted, hence a thick material of high permeability is used to redirect magnetic flux away from the object that needs to be shielded.

1.5.2.5 Displacement Current

A fine point that has been ignored so far was, in fact, one of the most important issues that Maxwell had to work around. Let us consider the case of a plate capacitor with a *vacuum dielectric*. That is, there are no electrons nor any other charges to carry the current through the capacitor. Nevertheless, it was proven by experiment that the charge–discharge current does flow through, which means that KCL is not valid! In order to make his equations work, in a stroke of genius, Maxwell added a new term called "displacement current" into Ampère's current law equation. Displacement current is equal to the charging and discharging currents in the external circuit. By doing this, Maxwell was able to derive the EM wave equation and to prove theoretically the existence of EM waves and the speed of light. At the time, the additional term in the equations could not be experimentally confirmed and the concept

of a "field" was still some time ahead. As we now know, EM waves do exist, otherwise we would not, which means that, although, we still do not know the real nature of the displacement current, we have accepted it as part of our reality. If you think that Maxwell was the only one who artificially introduced a new term into his equations (a new term that initially seemed to be an arbitrary addition and was subsequently proven by experiment), just remember Einstein and Planck. If anything, these examples only prove the power of applied mathematics. More recently, the development of string theory is taking the same approach, although it may take a while (if ever) before it is proven experimentally.

For slow-changing fields, it is common practice to use quasi-static approximation of Maxwell's equations, which is what we will be using in this book.

Example 1.8. Derive an expression for the resistance R of a conductive piece of material having cross-sectional area S and length l.

Solution 1.8. Starting from Maxwell's equations, the definitions for potential difference ΔV along a one-dimensional electric field and current density J give

$$E = \rho J = \frac{J}{\rho} \qquad \Delta V = -\int E \mathrm{d}l \qquad J = \frac{I}{S}, \tag{1.17}$$

where E is magnitude of the electric field along the conductor, $\sigma = 1/\rho$ is the conductivity of the material, ρ is the resistivity of the material, J is the current density, S is the cross-sectional area of the conductor, and l is the length of the conductor. Assuming a uniform electric field and homogenous material, it follows that

$$E = \frac{V}{l} \qquad J = \frac{I}{S} \qquad \therefore \qquad \frac{V}{l} = \rho \frac{I}{S} \qquad \therefore \qquad R \equiv \frac{V}{I} = \rho \frac{l}{S} \qquad [\Omega]. \tag{1.18}$$

1.5.3 The Concept of High Frequency

We very often use the term "high frequency" (HF) and it is valid to ask how the term *high frequency* is defined. Is there any particular number, for example 1 kHz or 1 GHz, that is accepted as defining "high frequency" or is there something else that is important to notice?

In order to answer this question, let us take a look at a simple, one-dimensional wave of an electric field travelling the z direction along the conductive wire's length, whose length is l, as

$$E_x = E_{0x} \cos(\omega t - kz), \tag{1.19}$$

where E_x is the electric wave field component along the x coordinate, E_{0x} is its maximum amplitude, ω is the angular frequency, $k = |\mathbf{k}| = \omega/c = 2\pi/\lambda$ is the wave vector value, z is the space coordinate showing the direction of wave propagation (which is perpendicular to the electric field vector), and the initial phase ϕ_0 is assumed to be zero.

The electric wave travels inside a "long" conductive wire aligned along the z coordinate, where the wave equation (1.19) explicitly shows the time t and space z arguments of the electric field. It is very important to note that, in this case, the term "long" implies that the wire length d is measured in units of the wavelength λ. In other words, this is a relative measurement where the wire length d is measured by the number of wavelengths λ. For example, Fig. 1.21 shows the wire length to be $d \approx 2.25\lambda$. Therefore, it should be obvious that, for a given physical wire length d, whether the wire

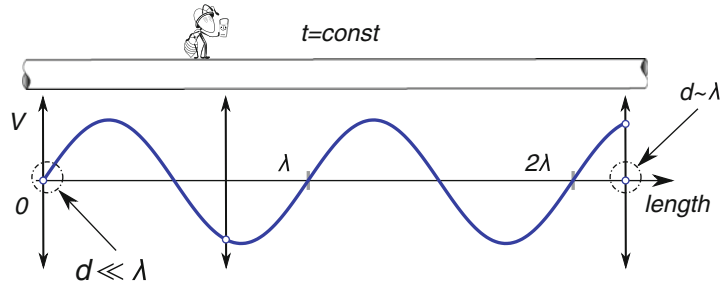


Fig. 1.21 A unidirectional wave front inside a conductive wire, with a single time frame shown in space. The measured voltage amplitude along the wire drastically depends upon the current location in space along the z axis. Try to compare the current that would flow into a branch connected at a point corresponding to $z = \lambda/4$ against the currents flowing into branches connected at points where $z = 2\lambda$ or $z = 3\lambda/4$

is quantified as "long" or "short" depends strictly on the signal frequency. Hence, a "short" wire is one where the wire length is much shorter than the wavelength, i.e., $d \ll \lambda$, while a "long" wire implies that the wire length d is either comparable to or longer than the waveform λ, regardless of whether the signal frequency is 60 Hz, 1 kHz, 1 GHz or any other number. The engineering rule of thumb is to estimate the wire length as "short" if $d \leqslant \lambda/10$; keep an open mind for the grey area between "short" and "long".

As a thought experiment, let us imagine that the time for this wave field has stopped (except for the little ant), so that the ant can observe and closely examine a "single frame" of this movie, i.e., $t = const$, while walking along a long conductive wire (see Fig. 1.21). Because this is a long wire, the waveform goes through more than a full cycle in space, which is to say that the measured potential along the wire varies between its minimum and maximum amplitude values in accordance with (1.19). A direct consequence of this situation is that, if a network branch contains portions of this wire, Kirchhoff's voltage law (KVL) is not valid, i.e.

$$\sum_{i=1}^{n} V_i \neq 0. \tag{1.20}$$

That consequence arises because KVL derived from Maxwell's equations assumes that the wire length is $d = 0$ (or, equivalently, $\lambda = \infty$), which in general is not the case, except for a DC signal. The spatial behaviour of the voltage (and its corresponding current) must be taken into account in cases when the signal wavelength is comparable to the conductor length (i.e., in the case of a "high-frequency signal") and Kirchhoff's circuit laws cannot be directly applied in their approximated form. The realization of this relationship led to the development of a mathematical model known as the *transmission line* model.

In order to circumvent the above problem, a long conductor carrying a high-frequency signal is split into a number of short-length sections Δz (mathematically $\Delta z \to 0$), which is to say that KVL is valid when applied to each section Δz separately (see Fig. 1.22). The physical properties of the wire section are then modelled using distributed electrical parameters R, L, C and G, where the corresponding electrical units are expressed in terms of the unit length, i.e., Ω/m, H/m, F/m, and S/m respectively. Analysis of each section is reduced to the analysis of a traditional circuit with lumped parameters.

Electric circuit representation of the line sections is a very useful modelling tool because it:

- Is a very intuitive model that is consistent with the two-port network methodology.
- Permits analysis using KVL and KCL.

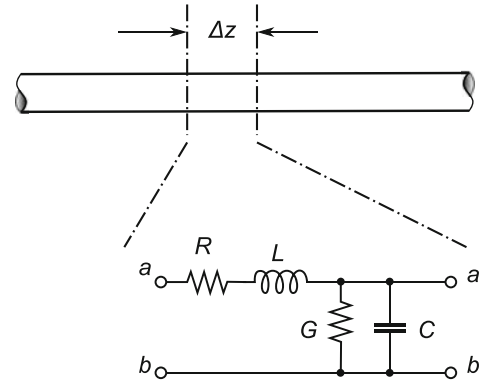


Fig. 1.22 A long conductor (relative to λ) is divided into infinitesimally short sections $\Delta z \ll \lambda$ and each section is modelled using distributed circuit elements R, L, C, and G

It has the following limitations:

- It is a one-dimensional model that does not include leaking fields and interference with other components.
- Material nonlinearities are mostly ignored.

In conclusion, KVL and KCL models are definitely applicable at DC and for "low-frequency" signals. For example, a 60 Hz signal ($\lambda \approx 5{,}000$ km) can be analyzed using Kirchhoff's laws if the signal is measured on a small PCB (with a wire length of, say, $d = 10$ cm). However, if the 60 Hz signal is carried across a continent, i.e. $\lambda \sim d$, than a more accurate transmission line model must be used. Similarly, a 1 GHz signal ($\lambda \approx 300$ mm) must be treated with the transmission line model if used on a 10 cm long PCB but the KVL model would result in a close-enough solution if the 1 GHz signal is carried by a 100 µm long wire inside an IC. Finally, an analysis of antennas and EM wave propagation through space must include full Maxwell's equations.

A solid understanding of these two extreme approximations, i.e., low frequency (LF) and high frequency (HF), is important for mastering RF circuit design. In this introductory book, however, we employ only low-frequency, quasi-static RF circuit design techniques for purposes of mastering the basic RF design principles without too much emphasis on specific properties of high and ultra-high frequency systems, which are the subject of more advanced courses.

1.6 RF Communication Systems

By now, we should be ready to carry out, at least in principle, a feasibility study of an RF communication system using the current technology. The goal is to transmit an audio signal using RF waves and faithfully reproduce it at the receiving end. Based on the principles introduced in this chapter, a rough block diagram of one possible system architecture is shown in Fig. 1.23.

At the beginning of the transmission chain, the mechanical sound wave produced by vocal cords (1) must be converted into its equivalent electrical signal (2) by a microphone. This electrical signal contains the complete information that needs to be transmitted and it is now ready to take a ride on its assigned carrier. It is the job of the modulator (3) to accept the signal and mix it with the carrier, which is enabled by the upconverter (4), so that the signal is imprinted as the carrier's envelope. The modulated carrier (5), with the information "riding" as its envelope, is now pushed into the transmitting antenna (6) and radiated into open space in the form of an EM wave (7). At this point, the information is available to anyone who is within the receiving range and whose "electrical length" matches that of the carrier. It should be noted that, at the same time, space is very busy and filled with

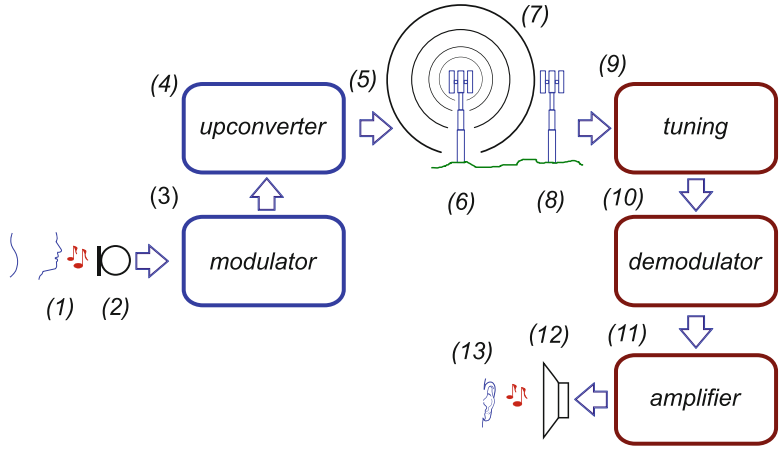


Fig. 1.23 A basic block diagram of a wireless communication system and its required components

many other carriers trying to reach their respective destinations. For the time being, the most important condition is that within the given space there must be no more than one carrier of any given frequency. Otherwise, two information packages travelling on separate carriers with indistinguishable carrier frequencies would unintentionally mix with each other and would be lost forever. In Sect. 9.6, we expand this condition to include one more frequency that, for now, we refer to as a "ghost frequency". There is no restriction on the number of receivers within the receiving range; in fact, radio and TV broadcasting companies spend vast amounts of money and resources to keep increasing the number of receivers for their broadcasting signals within the receiving distance range. This receiving distance range is limited by the power of the transmitted signal (its "loudness") and the sensitivity of the receiver (the quality of its "hearing system").

A receiver expecting a message must first adjust its "electric length" to match the frequency of the carrier. Under that condition, the receiving antenna (8) and the tuning section (9) start to oscillate in synchronicity with the incoming carrier, while (ideally) ignoring all other carriers. Using a simple analogy, we can visualize the receiver and the tuning section as a wall with a number of doors in various colours. At any given time, only one door is open (i.e., tuned) and only carriers of the matching colour get through. All other carriers face the wall with their matching colour doors closed.

Now there is no need for the carrier signal itself—it is the job of the demodulator (10) to extract the envelope and discard the carrier. After travelling a long way, the incoming wave is very weak (it is not economical to place receivers closer to the transmitter than needed), hence there is significant amplification (11) present in the receiving path. Stated differently, it is beneficial to design receivers with high sensitivity and maximize the distance between the transmitter and the receiver.

At the end of the receiving chain, the signal carrying the information is ready to be converted from electrical to mechanical form by a speaker (12) and finish the last leg of the journey the way it started—as a sound wave understandable by humans (13). The magic is done and the virtual distance between two humans becomes independent of the physical distance.

1.7 Summary

In this chapter, we have surveyed the fundamental principles required to understand the transmission of information over long distances using waves. The philosophy of building communication systems intended for this purpose is driven by the main constraint that, at the end, the information must be detectable by the human senses. The two most important are sight and hearing, both sensitive to wave

stimulus. Our hearing is capable of distinguishing sound waves in the range of 20–20,000 Hz, thanks to our internal spectrum analyzer, the cochlea. Each frequency from this band is perceived by our brain as a tone of a certain pitch, which means that we are capable of processing complicated, fast-changing sound signals and interpreting them, for instance, as speech, bird song, noise, or music. Our sight is sensitive to EM waves within a narrow band of frequencies centred around 5×10^{14} Hz, which we refer to as visible light. Unlike sound, light and other EM radiation does not need matter to travel through. Although it is still debatable exactly how EM radiation propagates through the space, nothing stops us from exploiting the fact that it does. Not only that, it travels at a speed of close to 300,000,000 m/s and is proven to be electrical in nature. An important parameter of any wave is its wavelength λ, which is calculated by dividing the wave's propagation speed by its frequency. Knowing a wave's wavelength helps us to compare waves to physical objects that need to interact with it. By doing so, we have learned how to design antennas suitable for interaction with waves of a particular frequency.

Discovery of this vast frequency band gave us an extremely valuable resource capable of carrying huge amounts of information simultaneously. It is comparable to a super highway with many parallel lanes. In order to make practical use of it, we had to invent precise and controllable methods for bidirectional translation of a given frequency to any other frequency and back, while preserving the original information. Thus, we also had to establish "rules of the road" for the frequency shifting to make sure that information travelling through very busy space does not collide. In order to do that, the whole EM frequency band is split into a large number of narrow bands following a very strict set of rules. We can visualize this collection of narrow frequency bands as a humongous set of parallel pipes, where each pipe originates at the information transmitting point and is just wide enough to carry that particular information. We use the terms "transmission channel" for each of these pipes and *bandwidth* to denote the "diameter" of the pipe. Once the information is radiated into space by a single transmitter and enters its assigned transmission channel, the number of receivers tapping into the channel is unlimited. Specifications for the transmitting equipment and the rules of transmission are strictly regulated by independent government agencies, for example, FCC in the USA and CRTC in Canada.

The conversion of the original sound wave to a radio wave at a particular frequency, and back, is done by electronic circuitry designed specifically to implement the mathematical operations of upconversion and downconversion. Details of these two primary steps in wireless communications were worked out using Maxwell's equations and basic trigonometry. Since the communication system is based on an application of electricity, the rest of this book is devoted to a detailed study of the general principles of electricity and time-varying electrical signals, the mathematical principles behind radio, and the design of practical electronic circuits that are synthesized to implement the required mathematical equations.

Problems

1.1. Calculate the intrinsic wave impedance, phase velocity, and wavelength of an EM wave in free space for the following frequencies: $f_1 = 10$ MHz, $f_2 = 100$ MHz, $f_3 = 10$ GHz.

1.2. Starting from the electrical line section model, Fig. 1.22 derive: (a) an expression for the general characteristic line impedance; (b) an expression for the lossless characteristic line impedance Z_0.

1.3. Plot the graph of the radial magnetic field $H(r)$ inside and outside an infinitely long wire in air of radius $a = 5$ mm, aligned along the z axis and carrying a DC current of $I = 5$A.

1.4. Find the induced voltage of a thin wire loop of radius $a = 5$ mm in air subject to a time-varying magnetic field $H = H_0 \cos \omega t$, where $H_0 = 5$ A/m and the operating frequency is $f = 100$ MHz.

1.5. The instantaneous voltage of a waveform is described as $v(t) = V_m \cos(2\pi f t + \phi)$ where $\omega = (2\pi\ 100)$ rad/s and $\phi = \pi/4$. Calculate the phase at $t = 15$ ns.

1.6. The instantaneous voltage of a waveform is described as

$$v(t) = \cos(2\pi \times 1 \times 10^3 t) + \frac{1}{3}\cos(2\pi \times 2 \times 10^3 t) + \frac{1}{5}\cos(2\pi \times 3 \times 10^3 t) \tag{1.21}$$

Using any plotting software, plot $v(t)$ on the same graph as the three single-tone terms over at least two periods of the slowest tone.

1.7. A sinusoidal wave is defined as $v(t) = 10V \sin(100\,t + 45^\circ)$. Determine: (a) the amplitude; (b) the v_{RMS} value; (c) the wave frequency; (d) the wave period; and (e) the phase at time $t = 1$ s. Convert $v(t)$ into its equivalent cosine function.

1.8. An arbitrary waveform $v(t)$ consists of $DC = 1$V, the fundamental tone $v_0 = 2\sin\omega t$, and the second harmonic $v_2 = 3/2 \sin 2\,\omega t$. Without using any plotting software, sketch the $v(t)$ waveform to scale.

Chapter 2
Basic Terminology

Abstract Our material world is reasonably well described by the set of working models that have been systematically derived in classical physics to map our perception of reality into compressed mathematical descriptions. These models are valid for most medium-intensity external conditions. For the purpose of our discussion, we accept them with their limitations and approximations as if they were the complete truth. That is, in this book we are not concerned with explaining the exact nature of all things. Instead, we merrily move forward and learn how to use the various phenomena and design new effects. In this section, we present the formulations of basic variables that are needed for RF circuit design and define the basic terminology.

2.1 Matter and Electricity

Conclusions that all matter consists mostly of electricity and that *electrons* are responsible for all chemical reactions in the known universe have been among the top intellectual achievements in human history. A positively charged nucleus contains almost all the mass of the atom and determines which element the atom is, e.g. silicon, oxygen, or any other element. The nucleus is generally very stable and it takes high amounts of energy to take it apart, i.e., to convert an atom into another element. In contrast, the fast-moving, negatively charged electrons contribute only a small percentage to the total atomic mass, however they determine the types of chemical reaction into which the atom enters. We can visualize fast-moving electrons as creating "shells" around the nucleus that are not easily penetrable, similar to the barrier created by a fast-spinning airplane propeller. However, a behaviour that is of much more relevance to our subject is that, in some cases, an electron may easily leave the atom and, in other cases, an external electron may join an atom by finding a place in the outermost shell. These "free-moving electrons" are responsible for a myriad of phenomena, only a small part of which are studied in this book (Fig. 2.1).

2.2 Electromotive Force

Once an electron leaves its native atom or joins some other hosting atom, a number of interesting things start happening. To start with, an atom that has lost one or more of its electrons is not electrically neutral—it now has an overall surplus of positive charges and is referred to as a "positive ion". Similarly, an atom that receives one or more external electrons in its outermost shell gains a surplus of negative charges and is referred to as a "negative ion". It is important to notice that ions do not

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8_2,

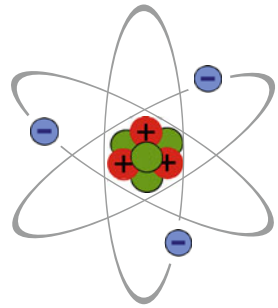


Fig. 2.1 The classical model of an atom consists of the nucleus (a massive central matter) and shells created by fast-moving electrons. The number of positive charges (protons) is equal to the number of negative charges (electrons) making the atom overall electrically neutral, even though an atom consists of charged particles. (Drawing not to scale.)

change the material itself, i.e., silicon is still silicon, oxygen is still oxygen, etc. However, charged particles interact by means of the *electric field* between them (see Fig. 1.19), which is a source of stress in the material that, theoretically, extends into the space infinitely far away. This stress due to unequal distribution of charges is manifested by an attractive or repelling force between the charged particles. Furthermore, the electric force is directly related to "potential energy", which defines the "potential" at a point inside the electric field, so that the potential energy of the charged particle at that point is measured relative to the reference point in infinity. A relative, and more practical, measure of potential is the "potential difference" (also known as the voltage) between two particles (or charged objects), where one of the objects serves as the reference point. In other words, the voltage V between objects A and B, is measured as the difference between their potentials, i.e., $V = V_A - V_B$. Most solid materials have their ions fixed (liquids and gases do not), while free electrons are pushed by electric field forces. Therefore, inside an electric field, the negatively charged electrons keep moving until, eventually, the overall electrical balance is restored. It is important to note that, by being negatively charged, the natural direction of the electron flow is towards the positive ions.

An interesting situation arises when, for example, a metallic wire (that happens to have a large number of free electrons) is connected to a device called a "battery". Then, the imbalance of the charges is maintained because the battery serves as an infinite source of free electrons because it provides "electromotive force" that moves the electrons from higher to lower potentials (which is opposite to their natural flow). The assembly of the metallic wire and the battery is referred to as a "closed circuit", where the battery enables the constant flow of electrons within the loop, as long as the path stays closed. This flow of free charged particles from higher potential to lower potential[1] caused by an electromotive force is referred to as "electric current". A battery serves a similar role to a water pump, which constantly pumps water to the top of a hill (i.e., increases its potential energy) and the water is pulled back to the bottom of the hill by the gravitational force (i.e., the potential energy is converted into kinetic energy). On its way down, the water flow may be used to do some extra work, for example, to spin a watermill.

[1]Keep in mind that, for historical reasons, the definition of the positive electric current direction is opposite to the direction of the moving electrons (a surplus of electrons means more negative charge).

2.3 Electric Current Effects

Now that we have established why and how the electric current came to be, it is natural to ask what it can do. At the very fundamental level, an electric current:

- *Generates heat*: The interaction of flowing electrons with the atomic lattice of a metallic material causes the atoms to increase the amplitude of their vibrations, which manifests as an increase in the material's temperature. Sometimes this heat generation is desirable, e.g. in an electric heater, and sometimes it is not, e.g. in a light bulb. Regardless of its desirability, it is very important to quantify the rate of heat generation (more details are presented in Sect. 4.1.4).
- *Generates a magnetic field around itself*: This property of an electric current is fundamental for wireless communication systems and is studied in great detail through the rest of the book.
- *Causes a chemical change in some materials*: This property of an electric current is exploited in chemistry, especially when a current passes through a liquid and enables the process of charging chemical batteries.

It should be noted that all of the above phenomena are bidirectional. Although of minor importance, the production of electricity in thermocouples is widely used for manufacturing thermal sensors. By far, the most exploited mechanism of electricity production is based on moving magnetic fields. Chemical batteries are still the most commonly used source of electric current for our mobile electronic devices.

2.4 Conductors, Semiconductors, and Insulators

In general, electrons do not have enough energy to leave a material. Instead, they keep exchanging their position by jumping from one atom to another. Every electron jump leaves a vacant spot, referred to as a positively charged "hole", behind in the positively charged ion, which in turn attracts some other electron and becomes neutral again. Due to large number of joggling electrons at any given time, a useful model is to treat them as an "electron cloud". Electron movements are induced in many ways, e.g. heat, and they happen randomly in time, which means that the average direction of the moving electron cloud is zero, similar to a swarm of bees that stays at one spot even though all the bees are very busy buzzing around; i.e., there is no spontaneous current flow in any particular direction. In order to force the electron cloud inside the material to have a non-zero average movement, an external electric field must be applied, for example by means of a battery. The external battery that is connected to a conductor serves as a "pump" that forces flow of the electrons from the battery's negative terminal through the conductor to its positive terminal. Materials that easily allow this directional drift of their electron cloud are called *conductors*. Most metallic materials are good conductors of electric current. A common model of a conductor assumes an ideal metallic wire, which allows an infinite electric current flow, even if an infinitely small voltage is applied at its ends. In other words, the ideal conductor is capable of dissipating an infinite amount of heat, which is to say that it can handle infinite power. Although real conductors do not have these properties, this idealization is very useful and commonly used every time you draw an electric schematic diagram. The connecting lines between the circuit components are assumed to be ideal wire conductors. This approximation is mostly valid, especially for circuits using low levels of current and operating at low frequencies.

Materials that do not have enough free charge–carriers to form the electron cloud are called *insulators*. Most plastic and glass-based materials are good insulators. That means that even if an internal electric field is created by an external potential difference across the isolating material and electric stress is induced in the material, (to the first approximation) there is no free current flow

through the insulator. A common model for the ideal insulator assumes that no single electron can leave the insulating material if a constant electric field is applied. Moreover, the ideal insulator is capable of handling infinite voltage across its terminals without allowing any current to flow. That is, because no current flow is allowed, the ideal insulator does not dissipate any amount of heat. This approximation is very useful because it enables us to model ideal discrete circuit components. In reality, there is always a small "current leakage" flowing through an insulator, however in applications with moderate requirements the leakage of current is safely ignored.

A third, and equally important, category of materials is known as "semiconductors". In general, semiconductive materials are neither good conductors nor good insulators. However, they do have a sufficient number of freely moving electric charges for a given volume of the material, which is strictly controlled so that the population of free charges is in the minority from the macro perspective. Under specially orchestrated conditions, for some types of semiconductor structures, it is possible to temporarily collect these free charges and to turn the non-conductive and localized volume of the semiconductor material into a very good conductor, i.e., to locally "invert" its conductive property. Once the controlling conditions are removed, everything reverts to the initial non-inverted state. This process is non-destructive, repeatable, and under full control of the circuit designer. Important variants of controlling current flow in semiconductor devices are outlined in Sect. 4.3.

2.5 Basic Electrical Variables

In this chapter, we have intuitively introduced the concept of "matter", which is a form of energy, and its basic property of "electric charge". In that model, the fine point is that electrons and protons are assumed to be merely material "carriers" of their respective charges.

2.5.1 Voltage

The concept of a particle charge, q (a scalar variable), leads to the concept of an electric field $\mathbf{E}$ (a vector variable), and to the electrical potential V_X (a scalar variable that is relative to a point infinitely far away) of a charged particle that occupies a point in space at the coordinate X. Two particle charges, occupying different points X and Y in space in the electric field, are therefore at different potentials. Thus they are said to have potential difference (p.d.) between them, which is referred to as "voltage" V and calculated as $\Delta V = V_X - V_Y$. Note that voltage is a relative measure that can be either positive or negative, depending upon which of the two charge potentials is assumed to be the local reference.

In a broad sense, electric fields are classified as either "static" or "dynamic". In the case of a static electric field, the electric potential created by a point charge q is described by Coulomb's law

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\,\mathbf{r}, \tag{2.1}$$

where $\mathbf{E}$ is the electric field vector, ε_0 is the vacuum permittivity (or electric constant), q is a single particle charge, r is the distance from the charge, and $\mathbf{r}$ is the unit vector pointing from the particle charge to the evaluation point in space.

By definition, the electric potential at a point r in a static electric field $\mathbf{E}$ is given by the line integral

$$\Delta V_{\mathbf{E}} = -\int_0^L \mathbf{E}\cdot \mathrm{d}\mathbf{l}, \tag{2.2}$$

where L is an arbitrary path connecting the point in infinity (i.e., with zero potential) to point r and $d\mathbf{l}$ is the unity path element. Note the dot product of the two vectors in the integral. In physical terms, (2.2) represents the electric work W (scalar variable) of the electric field along the integral path

$$W = q \int_0^l \mathbf{E} \cdot \mathrm{d}\mathbf{l} = q \Delta V_{\mathbf{E}}, \tag{2.3}$$

$$\therefore$$

$$V = \frac{\mathrm{d}W}{\mathrm{d}q} \quad [\mathrm{V}], \tag{2.4}$$

where voltage V is measured in volts [V]. Note that work represents the energy that is needed to move a particle over a certain distance.[2] In the case of a single particle charge q inside an electric field, (2.2) yields its potential as

$$V_{\mathbf{E}} = \frac{1}{4\pi\varepsilon_0} \frac{q}{r}. \tag{2.5}$$

A time-varying electric field, which is relevant to our subject, is always linked to a time-varying magnetic field (and vice versa). Consequently, it is not possible to describe the electric field in terms of a scalar potential V (because the integral (2.2) is now path dependent). Instead, one must use Maxwell's fundamental equations. For the sake of argument, one possible solution for the scalar potential is

$$-\nabla^2 V = \frac{\rho}{\varepsilon_0}, \tag{2.6}$$

where ρ is the charge density. For more details on Maxwell's equations, the reader is advised to consult more advanced textbooks on electromagnetism, some of which are listed in the references section.

2.5.2 *Current*

We have established the concept of an electric charge and concluded that a charge moves in space if a force $\mathbf{F}$ is applied, in this case, in the form of an electric field.[3] Observing this flow of charged particles, we define an electric current I (a scalar variable) as the net transfer of particle charges across a surface per unit of time. As a simple analogy, imagine standing on a sidewalk while a parade is marching by. Each person marching in the parade represents one unit of charge and the street width determines how many persons can fit in parallel. Start a stopwatch and count the people who pass over certain period of time, say, one second. Obviously, the wider the street, the more people pass through the street in a given time, i.e., the higher the "current" of people. Strictly, an electric current I is defined either as the rate of change of charge in time or the current density within the total conducting surface

$$I = \frac{\mathrm{d}Q}{\mathrm{d}t} = \int_S \mathbf{J} \cdot \mathrm{d}\mathbf{s} \quad [\mathrm{A}], \tag{2.7}$$

[2] By implication, one could sweat for whole day while trying to push a wall but no work would be done if the wall did not move, i.e., $d\mathbf{l} = 0$.

[3] The electric field $\mathbf{E}$ is defined as the force $\mathbf{F}$ per positive charge q that would be experienced by a stationary point charge at a given location in the field, i.e., $\mathbf{E} = \mathbf{F}/q$.

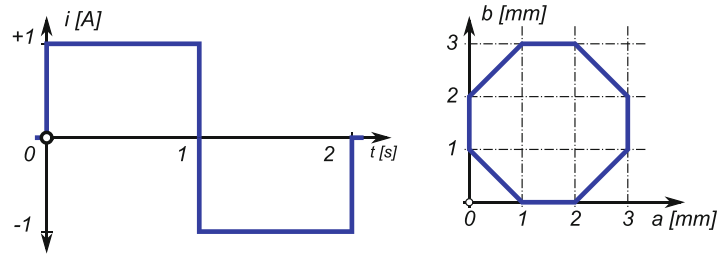


Fig. 2.2 Time diagram of charge flow for Example 2.1

where current I is measured in amperes [A], Q is the total amount of charge through the cross-sectional area S (not to be confused with the quality factor notification Q used in this book), dt is the differential unit of time, $\mathbf{J}$ is the current density vector, and $\mathbf{s}$ is the vector of the conducting surface element oriented in space. Note the dot product of the two vectors in the integral. Thus, for the known current, the total amount of transferred charge is

$$Q = \int_0^t i(t)\,\mathrm{d}t. \tag{2.8}$$

Example 2.1. A function that represents an instantaneous current amplitude is shown in Fig. 2.2 (left). The current flows through a conductor whose cross-section is shown in Fig. 2.2 (right). Determine the total amount of charge passing through: (a) from time zero to $t = 1$ s; (b) in the time period from $t_1 = 1$ s= to $t_2 = 2$ s; and (c) from time zero to $t = 2$ s. In addition, find the value of the current density $\mathbf{J}$.

Solution 2.1. By definition, the amount of electric charge is calculated using the integral (2.8), which in this case becomes trivial, because the flow of current is constant within each of the given time frames, hence

$$Q_1 = \int_0^{1s} i(t)\,\mathrm{d}t = 1\,\mathrm{A}\int_0^{1s} \mathrm{d}t = 1\,\mathrm{C},$$

$$Q_2 = \int_{1s}^{2s} i(t)\,\mathrm{d}t = -1\,\mathrm{A}\int_{1s}^{2s} \mathrm{d}t = -1\,\mathrm{C},$$

$$\therefore$$

$$Q = \int_0^{2s} i(t)\,\mathrm{d}t = 0, \tag{2.9}$$

that is, the net charge flow is zero. The current density is calculated by definition (2.7), which is also trivial because the current is constant from zero to $t = 1$s and, therefore the current density is constant, i.e.,

$$I = J\int_S \mathbf{ds} = J \times S \quad \therefore \quad J = \frac{I}{S} = \frac{1\,\mathrm{A}}{7\,\mathrm{mm}^2} = \frac{1}{7}\,\frac{\mathrm{A}}{\mathrm{mm}^2} \approx 142.86 \times 10^3\,\frac{\mathrm{A}}{\mathrm{m}^2},$$

where the cross-sectional area S is found by inspection of the plot Fig. 2.2 (right) to be $S = 7\,\mathrm{mm}^2$. Over the time period from $t = 1$ s to $t = 2$ s, the current density is $J = -142.86 \times 10^3\,\mathrm{A/m}^2$ because the current vector $\mathbf{I}$ points to the negative side.

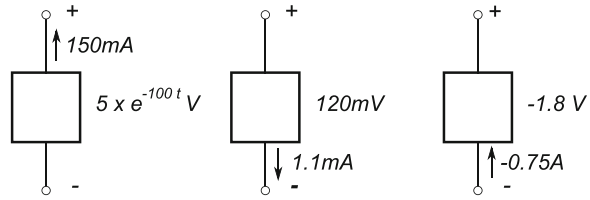


Fig. 2.3 Block diagrams for Example 2.2

2.5.3 Power

From the engineering perspective, it is important to establish not only the amount of energy needed to perform a work, but also the rate of energy exchange, i.e., the rate of either generation or absorption of energy. That brings us to the concept of *power P* (a scalar variable), which quantifies how fast, for a given amount of energy, the work is finished. Or, in a strictly mathematical sense, after substituting (2.4) and (2.7), the electrical power P is

$$P = \frac{\mathrm{d}W}{\mathrm{d}t} = \frac{\mathrm{d}W}{\mathrm{d}Q}\,\frac{\mathrm{d}Q}{\mathrm{d}t} = VI \quad [\mathrm{W}], \tag{2.10}$$

where power P is measured in watts [W]. To conclude, we keep in mind that all definitions introduced in this section assume either a static or a quasi-static (i.e., steady state) electric field.

Example 2.2. Find the power being delivered to or absorbed by the three elements in Fig. 2.3 at time instance $t = 5$ ms.

Solution 2.2. By definition, we write, -454.9 mW, 132 µW and 1.35 W.

2.5.4 Impedance

It is accepted convention to reserve the term *resistance R* for real resistive, i.e., frequency independent, components and to use the term *reactance* for the equivalent resistance of an inductor X_L or a capacitor X_C at a given frequency. By definition, a reactance is described only by the imaginary term $\mathrm{j}\Im$ (including the j part, which takes care of the phase) of a complex number $Z = \Re + \mathrm{j}\Im$. By the same convention, a serial combination of a resistance and either capacitance or inductance is referred to as *impedance Z*. For example, a serial connection of resistance R and inductance L is said to have impedance of $Z_\mathrm{L} = R + \mathrm{j}\omega L$ and ω is the radial frequency where the inductor reactance is calculated. Similarly, a serial connection of resistance R and capacitance C is said to have impedance of $Z_\mathrm{C} = R + 1/\mathrm{j}\omega C = R - \mathrm{j}/\omega C$. We note that inductive phase is positive and capacitive phase is negative.[4]

Two important parameters of any impedance are its absolute value $|Z|$ and argument ϕ (also referred to as a *phase*). In the complex plane, the real and imaginary axes are set at $\pi/2$ angle relative to each other, thus the absolute value and argument of a complex number are calculated using the Pythagorean theorem and trigonometric identities. For example, absolute values of *RL* and *RC* impedances are

$$|Z_\mathrm{RL}| = \sqrt{R^2 + (\omega L)^2} \qquad |Z_\mathrm{RC}| = \sqrt{R^2 + \left(\frac{1}{\omega C}\right)^2}, \tag{2.11}$$

[4] See the definition of the phase of a complex number in Appendix D.

while, by applying the same right-angle triangle rules, the phase is calculated as the ratio of the reactance and resistance (i.e., the imaginary and real) parts of the impedance, while paying attention to the sign of the reactance, i.e.

$$\tan\phi_{\mathrm{RL}} = \frac{\Im}{\Re} = \left|\frac{\mathrm{j}\omega L}{R}\right| = \frac{\omega L}{R}, \tag{2.12}$$

$$\tan\phi_{\mathrm{RC}} = \frac{1}{\mathrm{j}\omega CR} = -\left|\mathrm{j}\frac{1}{\omega CR}\right| = -\frac{1}{\omega CR}. \tag{2.13}$$

We use all of these relationships extensively in the rest of the book. For the time being, just note that both L/R and RC have dimensions of *time*.

Example 2.3. For an ideal capacitor $C = 100\,\mathrm{nF}$ and an ideal inductor $L = 100\,\mathrm{nH}$, calculate the following values at $f = 100\,\mathrm{MHz}$:

1. Find the impedance of a serial connection of the capacitor with a resistor $R = 6\,\mathrm{m\Omega}$.
2. Find the phase of a serial connection of the capacitor with a resistor $R = 6\,\mathrm{m\Omega}$.
3. Find the phase of a serial connection of the capacitor with a resistor $R = 0\,\Omega$.
4. Find the impedance of a serial connection of the inductor with a resistor $R = 4.6\,\Omega$.
5. Find the phase of a serial connection of the inductor with a resistor $R = 4.6\,\Omega$.

Solution 2.3. It is handy to first convert the frequency into its equivalent radial frequency, i.e., $\omega = 2\pi \times 100\,\mathrm{MHz} = 628.319\,\mathrm{Mrad/s}$, and then by direct implementation of (2.11) to (2.13), we write:

1.

$$|Z_{\mathrm{RC}}| = \sqrt{(6\,\mathrm{m\Omega})^2 + \left(\frac{1}{628.319\frac{\mathrm{Mrad}}{\mathrm{s}} \times 100\,\mathrm{nF}}\right)^2} \approx 17\,\mathrm{m\Omega}.$$

2. For the phase calculation, we must pay attention to the sign of reactance, i.e.

$$\tan\phi_{\mathrm{RC}} = -\frac{1}{\omega CR} = -\frac{1}{628.319\frac{\mathrm{Mrad}}{\mathrm{s}} \times 100\,\mathrm{nF} \times 6\,\mathrm{m\Omega}} = -2.65258,$$

$$\therefore$$

$$\phi_{\mathrm{RC}} = -69.344^\circ \approx -70^\circ.$$

3. When resistance $R = 0$, then $1/\omega RC \to \infty$, hence we must take a look at the limit of the tan function. However, this time we pay attention to the sign of the reactance, and we look only at the continuous range of angles[5] within the range -90° to 90°,

$$\tan\phi_{\mathrm{RC}} = -\lim_{R\to 0}\frac{1}{\omega CR} = -\infty \quad \therefore \quad \phi_{\mathrm{RC}} = -90^\circ.$$

4.

$$|Z_{\mathrm{RL}}| = \sqrt{(4.6\,\Omega)^2 + \left(628.319\frac{\mathrm{Mrad}}{\mathrm{s}} \times 100\,\mathrm{nH}\right)^2} = 63\,\Omega.$$

[5]Remember that the tan function is periodic and its values tend to $+\infty$ on one side and $-\infty$ on the other.

5.

$$\tan\phi_{\mathrm{RL}} = \frac{628.319\frac{\mathrm{Mrad}}{\mathrm{s}} \times 100\,\mathrm{nH}}{4.6\,\Omega} \quad \therefore \quad \phi_{\mathrm{RL}} = 85.813^\circ \approx 86^\circ.$$

6.

$$\tan\phi_{\mathrm{RL}} = \lim_{R\to 0}\frac{\omega L}{R} = +\infty \quad \therefore \quad \phi_{\mathrm{RL}} = 90^\circ.$$

Therefore, we note that even a small resistance in series with a reactance introduces a visible phase shift relative to the case of $R = 0\Omega$, i.e., when the phase equals $\pm 90^\circ$.

2.6 Electronic Signals

In electronic communication systems, the useful information, i.e., the signal, is embedded and carried in the form of voltage or current, or both. Time domain variations of either of these two variables are then modelled using appropriate mathematical functions. For example, digital information is transmitted by switching between two fixed voltage levels, which is modelled by using the pulse function. In wireless radio communications, at least the ones that are the subject of this book, the transmitted signal is embedded into a sinusoidal function. Therefore, we focus on properties of sine waves.

2.6.1 Properties of a Sine Wave

The basic characteristics of a travelling EM wave are based on a sinusoidal function (see Sect. 1.4). Hence, in this section, we focus on several properties of a sine functions that are relevant to RF signal analysis.

It is not difficult to prove that the average value of a sine wave over any integer number of cycles nT is zero, where n is the number of cycles and T is the sine wave period. A geometrical interpretation is that each period consists of one negative and one positive half-cycle, both having the same area. Since the cosine and sine functions are related, $\cos\omega = \sin(\omega - \pi/2)$, for the purposes of this discussion, it does not matter whether the sine or the cosine function is used in the analysis.

A very important case in engineering is the product of two sine waves. Let us consider the following two sine wave functions, with frequencies ω_1 and ω_2 and an initial phase difference θ at $t = 0$,

$$A = a\sin(\omega_1 t), \tag{2.14}$$

$$B = b\sin(\omega_2 t - \theta), \tag{2.15}$$

so that their product $x = AB$ is simply written as[6]

[6]Use the trigonometric identity $\sin(\alpha)\sin(\beta) = 1/2[\cos(\alpha-\beta) - \cos(\alpha+\beta)]$.

$$\begin{aligned} x &= ab\sin(\omega_1 t)\sin(\omega_2 t-\theta) \\ &= \frac{ab}{2}\{\cos[(\omega_1-\omega_2)t+\theta]-\cos[(\omega_1+\omega_2)t-\theta]\} \\ &= \frac{ab}{2}(x_1-x_2) \end{aligned} \tag{2.16}$$

and the average value x_{avg} is then calculated as the sum of the averages of the two terms x_1 and x_2. When $\omega_1 \neq \omega_2$, the average of the first term $x_{1\text{avg}}$ is

$$x_{1\text{avg}} = \cos[(\omega_1-\omega_2)t+\theta]_{\text{avg}} = 0 \tag{2.17}$$

for an integer number of cycles nT. Note, from (1.7), that the first term has a period of $T = 1/(f_1-f_2)$. Following the same argument, the same result is obtained for the second term,

$$x_{2\text{avg}} = \cos[(\omega_1+\omega_2)t-\theta]_{\text{avg}} = 0, \tag{2.18}$$

which is to say that, for the case of $\omega_1 \neq \omega_2$, the average value over the integer number of cycles of the product of two sine waves is zero.

However, for the case of identical frequencies $\omega_1 = \omega_2 = \omega$, (2.16) becomes

$$x = \frac{ab}{2}\cos\theta - \frac{ab}{2}\cos(2\omega t-\theta), \tag{2.19}$$

where the average of the second term $\cos(2\omega t-\theta)$ is zero, which leads to

$$x_{\text{avg}} = \frac{ab}{2}\cos\theta. \tag{2.20}$$

In this case, the average value depends upon the phase difference (the two frequencies are identical) and can, therefore, be adjusted to zero or anywhere between $(\pm ab/2)$. As will be demonstrated many times in this book, this observation is very important for RF design because the operation of RF circuits for wireless communication is based on perfect frequency relationships among multiple sinusoidal signals.

2.6.1.1 Root Mean Square

One possible view of a resistor is that it is a device that converts electrical energy into heat energy, which is then dissipated either intentionally (as in a stove heater, for example) or as wasted energy (as in a bulb, for example). Hence, it is important to know how much power is dissipated by the resistor in case of both DC and AC over an integer number of cycles. To do so, let us first consider the simple problem of calculating electrical power P dissipated by an ideal resistor R while conducting direct (i.e., constant in time) current I. Electric power was defined in (2.10) and additional forms are

$$P = VI = I^2R = \frac{V^2}{R}, \tag{2.21}$$

which, for a given resistance R, is dependent upon the current's (or the voltage's) squared value.

To find the answer for the case of periodic alternating current (e.g. $i = I_{\text{m}}\sin\omega t$), the calculation of the constant current term I^2 has to be replaced with the average value of time-varying quadratic

current, i.e i^2_{avg}, which, by definition, represents a "quadratic mean" or *root mean square (RMS)* of the current. Hence, calculation of the equivalent dissipated power is as follows,[7]

$$\begin{aligned} P_{\text{avg}} &= i\, v_{\text{avg}} = i^2_{\text{avg}} R \equiv i_{\text{rms}}\, R \\ &= \sqrt{\frac{1}{T}\int_0^T |i(t)|^2 \,\mathrm{d}t}\, R = \sqrt{\frac{1}{T}\int_0^T (I_{\text{m}} \sin \omega t)^2 \,\mathrm{d}t}\, R \\ &= \frac{I_{\text{m}} R}{T}\sqrt{\int_0^T \sin^2 \omega t \,\mathrm{d}t} = \frac{I_{\text{m}} R}{T}\sqrt{\left[\frac{t}{2} - \frac{\sin 2\omega t}{4\omega}\right]_0^T} \\ &= \frac{I_{\text{m}}}{\sqrt{2}} R. \end{aligned} \tag{2.22}$$

The equivalent effective direct current (DC) of a sinusoidal alternating current is the AC peak divided by the square root of two (2.22). In the case of a square wave, $i_{\text{rms}} = I_{\text{m}}$, while for a sawtooth wave, $i_{\text{rms}} = I_{\text{m}}/\sqrt{3}$.

It should be noted that most handheld multimeters assume a sine waveform. They filter the measured signal into an average value and then apply the $1/\sqrt{2}$ correction RMS factor. Therefore, the measured RMS voltage or current value is correct only if the input signal is sinusoidal. This is because the true RMS value is proportional to the area under the waveform, not to the average value of the waveform itself. For a sinusoidal waveform, the ratio of the average value to the area under the curve is constant, so most of the time the measured result is correct but any distortion or offset leads to errors.

If several sinusoidal functions with various frequencies are added together, for example

$$i = a\sin(\omega_1 t + \alpha) + b\sin(\omega_2 t + \beta) + c\sin(\omega_3 t + \gamma) + \cdots \tag{2.23}$$

then the RMS value of the sum (2.23) must be squared, however, in this case all inter–products between terms at different frequencies may be ignored because the average values of those products are zero, which leads to

$$i_{\text{rms}} = \sqrt{\left(\frac{a^2}{2} + \frac{b^2}{2} + \frac{c^2}{2} + \cdots\right)}. \tag{2.24}$$

This result shows that, when calculating the power of a multi-tone signal, each tone's power can be calculated separately, which is the property exploited in Fourier's analysis.

Finding the RMS value of a random signal is a bit more complicated and is left for a more advanced course.

2.6.1.2 Common Mode of a Signal

A periodic function that fluctuates around an average value other than zero may be thought of as being composed of a DC component I_{CM} and an AC component added together (see Fig. 2.4), i.e.

[7] Use the trigonometric identity $\sin^2\alpha = 1/2(1-\cos(2\alpha))$.

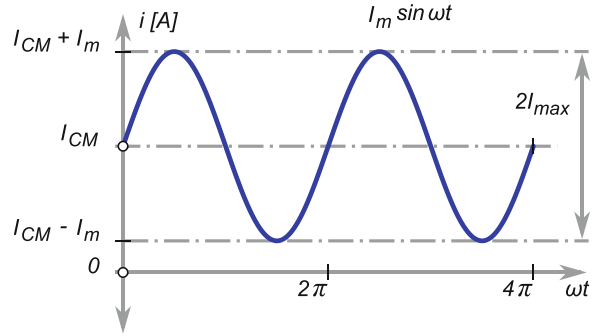


Fig. 2.4 A sinusoidal current signal whose common mode, i.e., average, level is I_{CM}

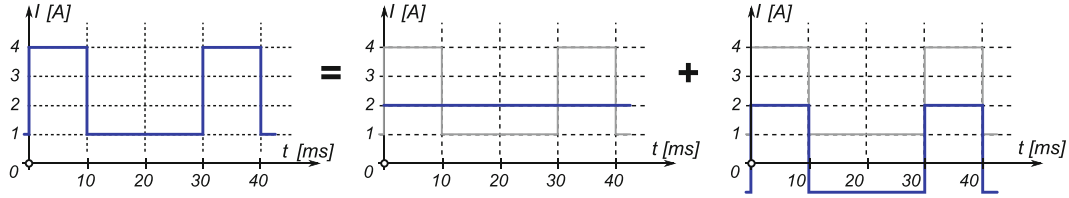


Fig. 2.5 Graphs of a square signal for Example 2.4

$$A = I_{CM} + I_m \sin \omega t, \tag{2.25}$$

where I_{CM} is the constant value and I_m is the maximum sine amplitude. Usually, $I_{CM} > I_m$ (often $I_{CM} \gg I_m$).

Example 2.4. Calculate the common mode level I_{CM}, AC amplitude I, RMS value of the AC component, and RMS value of the square signal in Fig. 2.5 (left). The value of current I is measured in [A] and the time is measured in [ms].

Solution 2.4. A square function consists only of linear sections, hence the integration is simplified to a simple addition over the period T. By inspection, the function period is $T = 30\,\text{ms}$; write,

- The common mode of Fig. 2.5 (left), i.e., the DC level, is

$$I_{avg} = \frac{4\,\text{A} \times 10\,\text{ms} + 1\,\text{A} \times 20\,\text{ms}}{30\,\text{ms}} = 2\,\text{A}. \tag{2.26}$$

- The AC component is found by realizing that the square waveform is the sum of its DC and AC components. By inspection, it is straightforward to recognize that the AC waveform must have $I_{AC} = 2\text{A}$ during the first 10 ms and $I_{AC} = -1\,\text{A}$ from 10 to 30 ms.
- The RMS value can be calculated, by definition, first for the AC component as

$$I_{rms}(AC) = \sqrt{\frac{(2\,\text{A})^2 \times 10\,\text{ms} + (-1\text{A})^2 + 20\,\text{ms}}{30\,\text{ms}}} = 1.414\,\text{A} \tag{2.27}$$

then, for the complete square waveform as

$$I_{rms} = \sqrt{\frac{(4\,\text{A})^2 \times 10\,\text{ms} + (1\text{A})^2 + 20\,\text{ms}}{30\,\text{ms}}} = 2.45\,\text{A} \tag{2.28}$$

or, alternatively, the total RMS value could be calculated as the sum of the RMS squares of the DC and AC components, as

$$I_{\mathrm{rms}} = \sqrt{I_{DC}^2 + I_{\mathrm{rms}}^2(AC)} = \sqrt{(2\,\mathrm{A})^2 + (1.414)^2} = 2.45\,\mathrm{A}, \tag{2.29}$$

which gives the same result because the RMS value of a DC level does not change.

2.6.2 DC and AC Signals

A *signal* is loosely defined as any time–varying event being observed. In electronic communications, signals are processed in form of either *current* or *voltage*; signal transmission can be either wired or wireless.

Two general categories of electronic signals are DC signals that have constant amplitude in time (for example, a battery voltage) and AC signals that have varying amplitude in time (for example, voltage amplitude measured at the wall power outlet). Further, an AC signal can be either periodic or aperiodic. Examples of periodic AC signal shapes are sinusoidal, square and saw waveforms, i.e., signals consisting of fixed, time-repetitive patterns. An example of an aperiodic electronic AC signal waveform is thermal noise. Naturally, DC signals have a simpler mathematical representation and treatment than periodic AC signals. On the other hand, aperiodic, or random, signals are more complicated than periodic signals and they are treated using mathematical tools from statistical analysis.

In this section, we review terminology related only to the most important form of AC signals, sinusoidal signals. Without being concerned about the nature of the signal, how it was generated, or what physical quantity it represents, a general sine-wave function is represented by

$$a = A_{\mathrm{p}} \sin(\omega t + \phi) \tag{2.30}$$

where:

a is the instantaneous value of time-varying quantity (voltage, current, power,...).
A_{p} is the maximum or peak amplitude.
ω is the angular frequency (related to frequency as $\omega = 2\pi f$).
ϕ is the initial phase (often assumed to be zero).
t is the time variable.

Figure 2.6 shows two common representations of AC signal (2.30), namely a phasor (or rotating vector) and its equivalent time-domain graph, where:

T is the period, i.e., the time interval required by the rotating vector to finish one full 2π cycle ($T = 1/f$).
θ is the instantaneous angle (not to be confused with the initial phase ϕ).

And, as we explained in Sect. 1.4.4, if two sinusoidal waveforms have the same frequency, then they are also related by phase difference $\Delta = \phi_1 - \phi_b$ (see Fig. 1.10). Further, depending upon the relative values of the instantaneous phases of the two, it is said that one waveform is either "lagging" or "leading" the other. For example, waveform A_1 in Fig. 1.10 is leading waveform A_2 by Δ. Of course, we keep in mind that, because the two waveforms have the same frequency, the phase difference is constant, otherwise it would not have been defined at all.

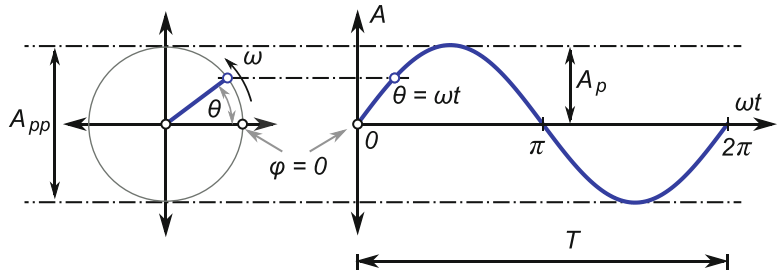


Fig. 2.6 Sine-wave representations: a phasor, i.e., a rotating vector, (*left*) and its equivalent time-domain sinusoidal function (*right*)

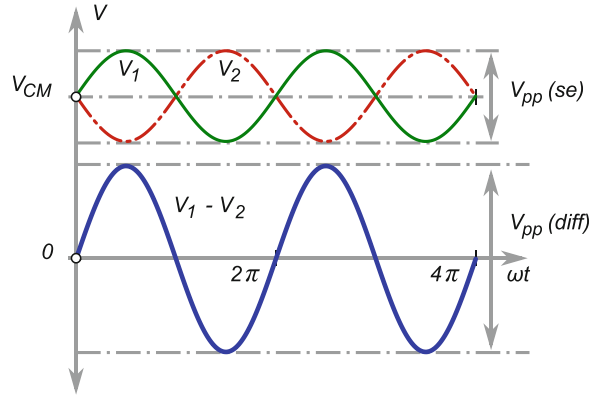


Fig. 2.7 A differential signal, $V_1 - V_2$, is constructed using two single-ended signals, V_1 and V_2. The two-single ended signals have equal frequency, amplitude and common-mode values, while their phases are inverted, i.e., they are at 180° relative to each other. The amplitude of the differential signal is twice the amplitude of the single-ended signal, i.e., $V_{\mathrm{pp}}(diff) = 2V_{\mathrm{pp}}(se)$

2.6.3 Single-Ended and Differential Signals

Typical signals, such as the sinusoids in Figs. 2.4 and 2.6, are also known as "single-ended" signals because they consist of only one waveform that is referenced to the local ground. In this section, we introduce a signal form that is very important to engineering, known as a *differential signal*, which is created by using two single-ended sinusoidal waveforms in the following relationship. They have:

- Equal amplitudes
- Equal frequencies
- Equal common mode
- Opposite phases, i.e., the phase difference is π

Let us consider two sinusoidal signals v_1 and v_2 as

$$v_1 = V_{\mathrm{CM}} + V_{\mathrm{m}} \sin \omega t, \tag{2.31}$$

$$v_2 = V_{\mathrm{CM}} - V_{\mathrm{m}} \sin \omega t, \tag{2.32}$$

where (2.31) and (2.32) formalize the required relationship between the two waveforms, shown in Fig. 2.7.

If these two signals are added, then the sum is, obviously, $v_1 + v_2 = 2V_{CM}$, which is a DC signal and the v_1 and v_2 waveforms are lost. However, if they are subtracted, then we write

$$v_{diff}(t) = v_1 - v_2 = 2V_m \sin \omega t, \tag{2.33}$$

which is a very interesting result because the original waveform[8] is still preserved, amplified by a factor of two and shifted down by the common mode value. The interesting part is that the amplification was achieved by the addition of two signals (one of which was negative) instead of multiplication. We note that the gain of factor two is significant, especially when we have weak signals to start with.

Let us explore this idea a bit further and assume that two conductive wires carrying the v_1 and v_2 signals are located physically close to each other. With that assumption, any interference signal $n(t)$ is added equally to both v_1 and v_2, i.e.

$$v_1 = V_{CM} + n(t) + V_m \sin \omega t, \tag{2.34}$$

$$v_2 = V_{CM} + n(t) - V_m \sin \omega t, \tag{2.35}$$

which, after subtraction again, results in (2.33). In other words, the common interfering signal is removed from the differential signal. These two properties of differential signals, namely the gain and the immunity to common noise, are beneficial and important enough that most modern, high-performance, signal-processing circuits are designed to process differential signals. However, for the sake of simplicity and accepted educational methodology, all circuits in this textbook are assumed to be single-ended, leaving differential architectures for more advanced courses.

2.6.4 Constructive and Destructive Signal Interactions

The relative phase between two periodic signals is very important from the perspective of their sum. In a circuit network, two currents entering the same node add up in accordance with KCL, while two voltages within the same branch add in accordance with KVL. In a realistic circuit implementation, it is almost inevitable to have two or more conductive wires in close proximity to each other. Unless they have exactly the same potential along their full respective lengths at all times, there is always capacitive cross-coupling between the two. Consequently, the two signals do interact, i.e., add, with each other.

In Sect. 2.6.3, we encountered the intentional subtraction of two signals with opposite phases for the purpose of creating a differential signal and exploiting its benefits, which is an example of constructive signal addition. However, in general, the amplitudes, phases, and frequencies of two adjacent signals are not equal (see Fig. 2.8). A special case of interest is when the two interacting signals are opposite in phase and have equal frequency and equal (or almost equal) amplitude (see Fig. 2.9). Under these special conditions, the two signals *cancel*, i.e., their sum is zero, and we refer to this interaction as *destructive* addition. We keep in mind that the concept of signal addition applies to all signals, not only to single tones. It is not difficult to see, for example, how one harmonic within the complicated signal spectrum is easily removed from the spectrum with destructive addition of the appropriate single tone, i.e., the one with the same frequency and amplitude and opposite phase.

[8] Remember, except for the phase difference, the two initial waveforms are identical.

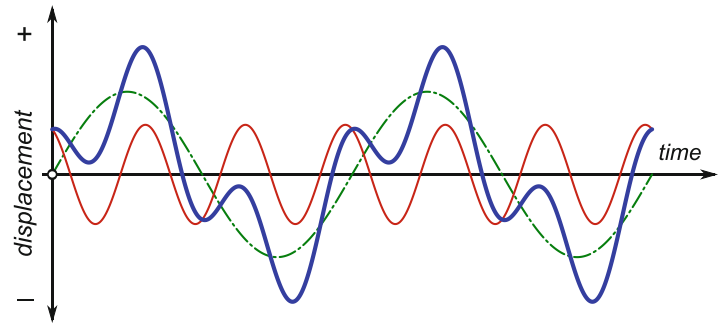


Fig. 2.8 A time domain graph of two arbitrary sinusoidal signals (*the dashed and thin lines*) and their sum (*the solid thick line*)

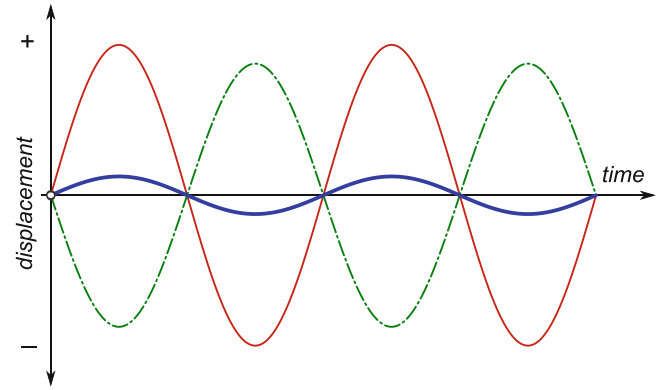


Fig. 2.9 A time domain graph of two sinusoidal signals with equal frequencies, inverted phase, and almost equal amplitudes (*the dashed and thin lines*) and their sum (*the solid thick line*). When the amplitudes are equal, the sum is exactly zero, i.e., DC

It is important to realize that both constructive and destructive signal additions are used intentionally in signal processing, as we will see later in this book.

2.7 Signal Quantification

Periodic signals are arguably the most important category of signals in the design of RF communication systems. Therefore, it is important that we become familiar with the metrics used to quantify periodic RF signals. Specifically, we are more interested in RF signal *power* levels than in the instantaneous values of individual voltages and currents. The level of an RF signal's RMS power is traditionally expressed in *dB*.

2.7.1 AC Signal Power

So far, we have introduced AC through a pure resistive network. In general, we need to expand our analysis to include inductive and capacitive elements as well. Being energy storage components, these *reactive elements* may cause reversal of energy flow (i.e., power flow) within the network. Consequently, it is common in the engineering community to define three "types" of power: real power P (i.e., power delivered to a pure resistive network); reactive power Q (i.e., power delivered to reactive components L and C); and complex power S (i.e., power delivered to a general RLC network); where the modulus of complex power $|S|$ is referred to as *apparent power*. At any given moment, the

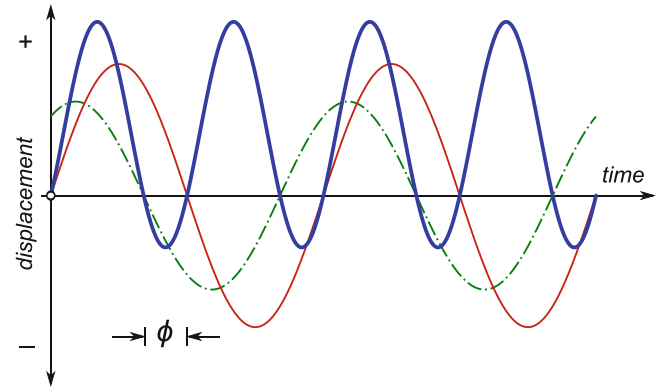


Fig. 2.10 The instantaneous voltage (*thin solid line*), current (*dashed line*), and power (*thick solid line*) in an AC circuit branch showing phase difference ϕ

instantaneous power delivered to any circuit element or network is given by product $p = vi$, where p is the instantaneous power, v is the instantaneous voltage and i is the instantaneous current. However, in the case of alternating currents and voltages, there is a very important consequence to notice, which we show here.

Let us assume that the instantaneous values of current and voltage in one branch of a circuit are given as follows:

$$i = I_\mathrm{p} \sin \omega t, \tag{2.36}$$

$$v = V_\mathrm{p} \sin(\omega t + \phi). \tag{2.37}$$

In other words, there is a phase difference of ϕ between the current and the voltage of that particular branch. Then, the instantaneous power is calculated as

$$p = vi = V_\mathrm{p} I_\mathrm{p} \sin \omega t \sin(\omega t + \phi). \tag{2.38}$$

Surprisingly, (2.38) suggests that at some instances in time the power is positive and at other instances the power is negative (see Fig. 2.10). In order to correctly interpret the above statement, keep in mind that the sign of power indicates only the direction of energy flow. Simply put, "positive power" indicates that external world is supplying power to the circuit, while "negative power" indicates that the circuit is delivering power to the world. This is possible only if some devices capable of storing energy are present in the circuit, i.e., inductors or capacitors.

Using the same method to calculate the average power of this circuit branch as for obtaining (2.22), we write

$$\begin{aligned} P &= \frac{1}{T} \int_0^T vi \, \mathrm{d}t = \frac{V_\mathrm{p} I_\mathrm{p}}{T} \int_0^T \sin \omega t \sin(\omega t + \phi) \, \mathrm{d}t \\ &= \frac{V_\mathrm{p} I_\mathrm{p}}{T} \left[\cos \phi \int_0^T \sin^2 \omega t \, \mathrm{d}t + \sin \phi \int_0^T \cos \omega t \sin \omega t \, \mathrm{d}t \right], \\ &\therefore \\ P &= \frac{V_\mathrm{p} I_\mathrm{p}}{2} \cos \phi = V_\mathrm{rms} I_\mathrm{rms} \cos \phi = \frac{V_\mathrm{rms}^2}{R} \cos \phi. \end{aligned} \tag{2.39}$$

An important observation regarding this result is that AC power depends upon the cosine of the phase difference between the corresponding current and voltage. A direct consequence of this relationship

is that in special cases when the phase difference $\phi = \pm 90^\circ$ (i.e., in a purely reactive circuit), the AC power factor $\cos\phi$ is zero. When the power factor $\cos\phi = 1$ (i.e., in a purely resistive circuit) the power is at maximum. Therefore, a power factor less than one always indicates the presence of reactive (i.e., L and C) components in the circuit. Keep this important observation in mind until we reach the discussion on capacitors and inductors in Chap. 4.

2.7.2 The Decibel Scale

In wireless communication systems, it is common to have an RF transmitter delivering signals at power levels of the order of watts, kilowatts or even megawatts. As a comparison, the signal power level at the receiving antenna can be only a few picowatts. That is, the power ratio of the transmitted and received signals may be as large as $1,000,000,000,000,000 : 1$. Clearly, using absolute numbers is not the most convenient way of presenting RF signal relations.

By definition, the dB is a logarithmic unit of measurement that expresses the magnitude of a physical quantity (usually power) relative to a specified or implied reference level. Its logarithmic nature allows very large and very small ratios to be represented by a convenient number. Being a simple ratio of two quantities, the dB is a *dimensionless* unit.

The *Bel* scale is defined as the logarithm of the base 10 of the power ratio. One Bel is a factor of 10, two Bels is a factor of 100, and so on. It is common, however, to use a more practical dB unit, so that 10 dB is a power ratio of 10, 20 dB is a ratio of 100, and so on. It is useful to remember that 3 dB is a power ratio of ≈ 2, 6 dB is a power ratio of ≈ 4, and so on.

Thus, power ratio (i.e., power gain G) is expressed in dB as

$$G_{\text{dB}} = 10 \log \frac{P_2}{P_1}, \tag{2.40}$$

where P_1 and P_2 are the two signal powers being compared, for example, the input and output powers of an amplifier. Keep in mind that when G_{dB} is a positive number, it indicates that $P_2 > P_1$ (often referred to as "gain"), while a negative G_{dB} number indicates that $P_1 > P_2$ (often referred to as "loss").

If we want to express a voltage (or current) ratio (i.e., a voltage or current gain A) of two signals v_2 and v_1 in the dB scale, and assuming that both signals are measured at the same impedance Z, then the gain is expressed in dB as:

$$A_{\text{dB}} = 10 \log \frac{P_2}{P_1} = 10 \log \frac{v_2^2/Z}{v_1^2/Z} = 20 \log \frac{v_2}{v_1}, \tag{2.41}$$

which is to say that a voltage (or current) ratio of 10 equals 20 dB gain, a ratio of 0.1 equals -20 dB gain, a ratio of 100 is equal to 40 dB gain, etc. It is handy to practice mental conversion between ratios and dB units by taking the number of zeros in the ratio exponent and multiplying it by 10 for power or by 20 for voltage or current; the final number is in dB units.

Because dB numbers are dimensionless they do not say anything about the absolute power levels being compared. Hence, from a specified gain, we can only conclude whether there was power amplification or power loss. From such a statement, however, we cannot conclude either what kind of gain it is (i.e., power, voltage, or current) or which two absolute signal values are being compared.

Therefore, for low-power applications, the standard reference value for power specification is defined in the form of the dBm scale, which is set to compare a given power level relative to the absolute power level of $P_1 = 1$ mW. After substituting the 1 mW level in (2.40), we write

$$G_{\text{dBm}} = 10 \log \frac{P_2}{1\,\text{mW}} \tag{2.42}$$

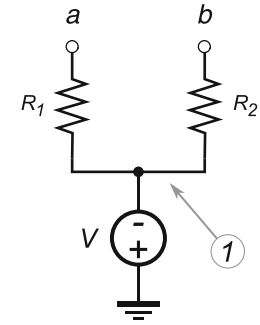


Fig. 2.11 A circuit with ground potential (for Example 2.6)

indicating that 1 mW of power is equivalent to 0 dBm. Similarly, if an amplifier delivers 10 mW of power, it is usually expressed as 10 dBm gain, 100 mW as 20 dBm, etc. Note that, due to the same scale, in power calculations the units of dB and dBm are added to or subtracted from each other, i.e they are interchangeable as long as we keep the 1 mW absolute reference in mind.

Example 2.5. A cell phone transmits $P_1 = +30\,\text{dBm}$ of signal power from its antenna. At the receiving side, the signal power is $P_2 = 5\,\text{pW}$. Calculate the propagation loss of the transmitting medium.

Solution 2.5. We convert the received power into dBm units as

$$P_2 = 10\log\frac{P_2}{1\,\text{mW}} = 10\log\frac{5\,\text{pW}}{1\,\text{mW}} = -83\,\text{dBm}. \tag{2.43}$$

Therefore the signal experienced attenuation A of

$$A = P_2 - P_1 = 30\,\text{dBm} - (-83\,\text{dBm}) = -113\,\text{dBm}. \tag{2.44}$$

2.7.3 The Meaning of "Ground"

In our discussions, we routinely assume that the concept of "ground" is clear to everyone and we simply assume that the ground is at zero potential. Often, we forget that the zero level was set as a relative point, not the absolute. Let us be reminded that any measured voltage value is implicitly assumed to be the potential difference between two points, one of which is arbitrarily declared the "ground", i.e., the zero reference. The absolute potentials are, by definition, measured relative to some point at infinity. Because of that, it is more practical to arbitrarily pick one of the two points and declare it to be the "local ground". When it is necessary to emphasize that a voltage is measured between two specific points in a circuit, the notation V_{AB} is used, where A is the node with higher potential and B is the node with lower potential (keep in mind that $V_{\text{AB}} = -V_{\text{BA}}$). It is especially important to have a clear understanding of the concept of "ground" when dealing with differential signal circuits because a differential signal is always measured as a difference between the two signals and its value is independent of the ground level.

Example 2.6. What is the value of resistance ($R_{a,b}$) between points a and b, in Fig. 2.11 if: (a) $V = 1\,\text{V}$; (b) $V = 0\,\text{V}$; (c) $V = -1\,\text{V}$; (d) $V = -1\,\text{MV}$?

Solution 2.6. If, for the moment, we completely ignore the existence of the voltage source V, then between points a and b there is a serial connection of two resistors, so we normally write $R_{a,b} = R_1 + R_2$. Did you notice that in order to calculate the equivalent serial resistance, we did not

need to find the potential at the joining node ① between the two resistors? That is, the potential at node ① is not part of the equation. Hence, the serial resistance stays the same whatever the potential at node ①. The voltage V is referenced to an arbitrary point in space that we temporarily declared the ground; it could have been node ① with no difference whatsoever.

2.8 Summary

Each profession has its own technical language and fluency in the language is critical for one's professional career. Similar to native speakers who immediately pick up even the smallest mistake by a non-native speaker, experienced professionals in a field are able to estimate the competence of the other person simply by picking up on incorrect use of terminology. In this chapter, we reviewed some of the very basic definitions that are considered fundamental knowledge in the field and are found in the vocabulary of all engineers and scientists.

Problems

2.1. Using a graphing tool of your choice, create overlapping plots of the following single-tone signals at $f = 10\,\text{MHz}$:

$$S_1 = 2.0\sin(\omega t),\quad S_2 = 2.0\sin(\omega t + \pi/3),\quad S_3 = 2.0\sin(\omega t + \pi/2),$$
$$S_4 = 2.0\sin(\omega t + 3\pi/4),\quad S_5 = 2.0\sin(\omega t + 2\pi),\quad S_6 = 2.0\sin(\omega t + 4\pi/3). \tag{2.45}$$

Observe the relationships between various signals in terms of their phase differences (hint: to start, begin at the zero time) and how the amplitudes are related to each other at any given point in time. Practice calculating the signal amplitudes at various time points by knowing their phase. For a given frequency, practice expressing various phase differences in the units of time.

2.2. Using a graphing tool of your choice, create plots of the following signals at $f = 10\,\text{MHz}$:

$$S_1 = 2.0\sin(\omega t),$$
$$S_2 = 2.0\sin(\omega t + \theta). \tag{2.46}$$

Plot $S_3 = S_1 + S_2$ for the following phase differences: $\theta = 0, \pi/3, \pi/2, \pi, 3\pi/2, 2\pi, 3\pi, 4\pi, \ldots$ Observe how the amplitude of S_3 changes relative to the phase differences between S_1 and S_2. In particular, pay attention to what happens to the amplitude of S_3 when $\theta = k\pi$ and $k = 0, 1, 2, 3, \ldots$.

2.3. Overlap plots of the following single-tone signals (assume $f = 10\,\text{MHz}$):

$$S_1 = 2\sin(\omega t),\quad S_2 = -\sin(2\omega t),\quad S_3 = \frac{2}{3}\sin(3\omega t),$$
$$S_4 = -\frac{1}{2}\sin(4\omega t),\quad S_5 = \frac{2}{5}\sin(5\omega t),\quad S = \sum_{k=1}^{5} S_k. \tag{2.47}$$

To what waveform shape is S converging, starting with $S = S_1 + S_2$, then $S = S_1 + S_2 + S_3$, etc., assuming that more S_k terms are added to the sum? Now, plot the sum without, for example, the S_2 term and observe what the S waveform looks like. What about without the S_3 term? Try dropping other terms or combinations of terms from the sum and observe the outcome.

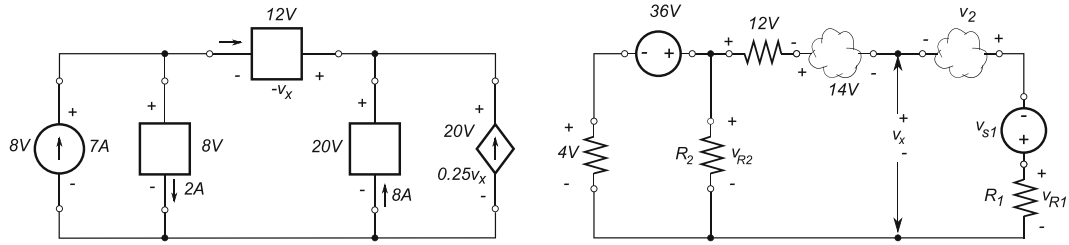


Fig. 2.12 Schematic diagram for Problem 2.7 (*left*) and Problem 2.9 (*right*)

This exercise demonstrates the frequency spectrum components of a complicated signal and how the signal becomes unrecognizable (i.e., distorted) to various extents if some of its terms are filtered out.

Note that this particular waveform consists (aside from the fundamental tone ω) of both even and odd harmonics, i.e., 2ω, 3ω, 4ω, ...

2.4. Overlap plots of the following single-tone signals (assume $f = 10\,\text{MHz}$):

$$S_1 = \frac{4}{\pi}\sin(\omega t), \quad S_2 = \frac{4}{3\pi}\sin(3\omega t), \quad S_3 = \frac{4}{5\pi}\sin(5\omega t),$$
$$S_4 = \frac{4}{7\pi}\sin(7\omega t), \quad S_5 = \frac{4}{9\pi}\sin(9\omega t), \quad S = \sum_{k=1}^{5} S_k. \tag{2.48}$$

Note that the frequency spectrum of this particular S signal comprises (aside from the fundamental tone ω) only odd harmonics, i.e., 3ω, 5ω, 7ω, ...

2.5. Calculate the average energy in a rectangular pulse whose amplitude is $v = 2\text{V}$ and width is $t = 1\,\text{ms}$. The energy is dissipated in a resistor $R = 100\,\Omega$.

2.6. A current flowing in a positive direction through a wire is defined as:

$$i(t) = \begin{cases} -2t, & \text{if } t < 0 \\ +3t, & \text{if } t \geqslant 0 \end{cases}. \tag{2.49}$$

Find the following values:

(a) $i(-2.2\text{s})$.
(b) $i(+2.2\text{s})$.
(c) The total charge g that has flowed through the wire within the time interval $-2s \leq t \leq 3s$.
(d) The average value of $i(t)$ within the same time interval.

2.7. Find the power absorbed by each element in the circuit shown in Fig. 2.12 (left).

2.8. For a resistor R with a current i entering its more positive terminal and voltage v across its terminals, find:

(a) The resistance R if $i = -1.6\,\text{mA}$ and $v = -6.3\,\text{V}$.
(b) The absorbed power P if $v = -6.3\text{V}$ and $R = 21\,\Omega$.
(c) The current i if the voltage is $v = 8\,\text{V}$ and R absorbs power $P = 0.24\,\text{W}$.
(d) The conductance G if the voltage is $v = -8\,\text{V}$ and R absorbs power $P = 3\,\text{mW}$.

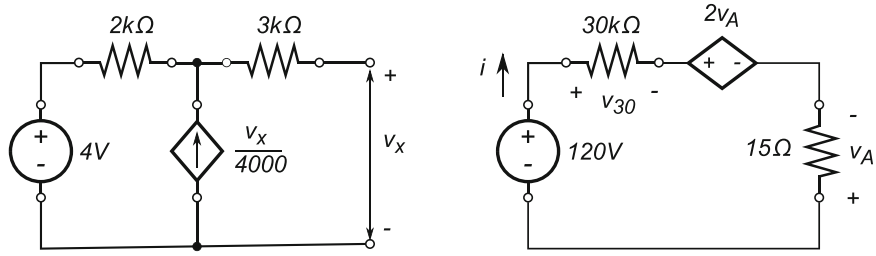
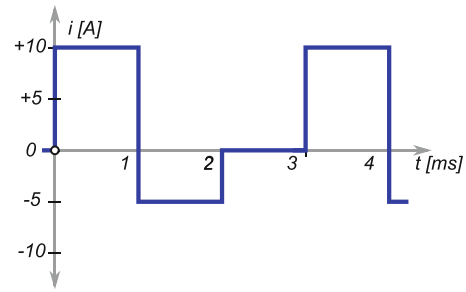


Fig. 2.13 Schematic diagrams for Problem 2.10 (*right*) and Problem 2.11 (*left*)

Fig. 2.14 Schematic diagram for Problem 2.13



2.9. Find v_{R_2} and v_x, shown in Fig. 2.12 (right).

2.10. Find the power absorbed by each component in Fig. 2.13 (right).

2.11. Find the equivalent Thévenin (see Sect. 4.2.3) circuit in Fig. 2.13 (left).

2.12. Find the RMS for the following waveforms where t is time, f is frequency, a is the peak amplitude, and T is the function period:

(a) Square wave

$$y = \begin{cases} a & t < 0.5T \\ -a & t \geqslant 0.5T \end{cases}. \tag{2.50}$$

(b) Modified square wave

$$y = \begin{cases} 0 & t < 0.25T \\ a & 0.25 \leqslant t < 0.5T \\ 0 & 0.5 \leqslant t < 0.75T \\ -a & t \geqslant 0.75T \end{cases}. \tag{2.51}$$

(c) Sawtooth wave

$$y = 2at - a \tag{2.52}$$

2.13. Calculate the average power delivered to an $R = 5\,\Omega$ resistor and $i_{\mathtt{rms}}$ for the current waveform in Fig. 2.14.

2.14. Calculate the average power delivered to an $R = 4\,\Omega$ resistor if the instantaneous current is: (a) $i = (2\cos 10t - 3\cos 20t)$ A; and (b) $i = (2\cos 10t - 3\cos 10t)$ A.

Chapter 3
Electrical Noise

Abstract Any electrical signal that makes recovery of the information signal more difficult is considered *noise*. For example, "white snow" on a TV picture and "hum" in an audio signal are typical electrical noise manifestations. Noise mainly affects receiving systems, where it sets the minimum signal level that it is possible to recover before it becomes swamped by the noise. It is important to note that amplifying a signal already mixed with noise does not help the signal recovery process at all. Once it enters the amplifier, noise is also amplified, which is to say that the ratio of S/N power does not improve and that is what matters. When the power of the noise signal becomes too large relative to the power of the information signal, information content may be irreversibly lost. In this chapter, we study the basic classification of noise sources and methods for evaluation of noise effects.

3.1 Thermal Noise

At the fundamental level, the numerical value of electrical current is just an average number of electrons coming out of the conductor per unit of time. Keep in mind that, even without any external electric field, an electron cloud moves inside a material and interacts with the vibrating ions, each electron moving in Brownian motion (i.e., similar to a pinball). The random motion of each individual electron makes a micro current that, together with all the other micro currents in the given volume, adds up to a macro current with zero average value. Due to its random nature, this current does not contain information, therefore we consider it "noise". This motion is responsible for the conductor's temperature, hence it is known as "thermal" noise; in real conductors, it is what constitutes the conductor's resistance. Given that the movement of electrons produces current, and current through a resistor creates voltage across its terminals, we also consider a resistor as a random noise generator. Both experiments and theory have found that the power spectrum of thermal noise is flat, which (loosely) means that each frequency component in the noise spectrum has the same power level, as shown in Fig. 3.1 (right). This conclusion is valid over a very wide range of frequencies (up to approximately 10^{13} Hz). Similarly to white light, which contains all colours (i.e., light frequencies), a noise signal that contains single tones at all possible frequencies is called, appropriately, *white noise*. Of course, it is only a very good approximation, because the implication is that, if measured over all possible frequencies, the total noise energy would be infinite. To accurately address this issue, we would need to delve into quantum mechanics theory, which is not the subject of this book.

Variables that have zero average, which is the case with thermal noise, are much better evaluated by measuring their RMS value as in Fig. 3.1 (left). Using methods from statistical thermodynamics

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8_3,

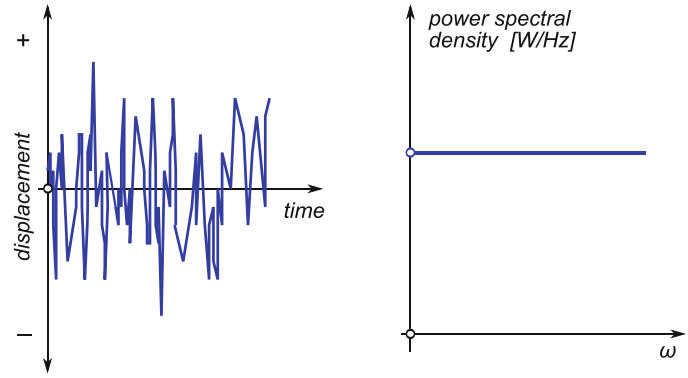


Fig. 3.1 Thermal noise in the time domain (*left*) and the noise power spectrum density (*right*)

and quantum mechanics, it has been shown that the noise spectrum density S_n (sometimes referred to as the available noise power) within a 1 Hz bandwidth, is

$$S_n(f) = kT \quad \left[\frac{W}{Hz}\right], \tag{3.1}$$

which is not a function of frequency, i.e., it is constant, Fig. 3.1 (right). Therefore, the noise power generated within frequency bandwidth Δf is, by definition

$$P_n = \int_{f_1}^{f_1+\Delta f} S_n(f)\,\mathrm{d}f = S_n(f)\int_{f_1}^{f_1+\Delta f} \mathrm{d}f = kT\Delta f \quad [\mathrm{W}], \tag{3.2}$$

where

k is Boltzmann's constant (1.38×10^{-23} J/K).
T is the absolute temperature of the conductor (in K).
Δf is the frequency bandwidth in which the noise measured (in Hz).

It is interesting to note that, even though it is modelled with a resistor, the noise power does not depend on the resistance of the conductor. Equation (3.2), also known as Johnson's law, implies that it is desirable to reduce the bandwidth of the receiver to a minimum since the noise power is proportional to the system bandwidth.

Example 3.1. Find:

(a) The spectrum density for thermal noise at room temperature ($T = 300$K).
(b) The available noise power within a bandwidth of 1 MHz.
(c) The available signal power for a 1 μV signal from a 50 Ω source delivered to the matched load.
(d) The SNR for the noise in part (b) and the signal in part (c).

Solution 3.1.

(a) $$S_n = 1.38 \times 10^{-23} \times 300 = 4.14 \times 10^{-21}\,\mathrm{W/Hz}$$

(b) $$P_n = 4.14 \times 10^{-21} \times 10^6 = 4.14 \times 10^{-15}\,\mathrm{W}$$

(c) $$P_s = \frac{(1 \times 10^{-6}/2)^2}{50} = 5 \times 10^{-15}\,\mathrm{W}$$

(d) $$SNR = \frac{P_s}{P_n} = \frac{5 \times 10^{-15}}{4.14 \times 10^{-15}} = 0.82\,\mathrm{dB}$$

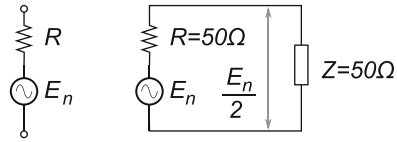


Fig. 3.2 Equivalent noise generator E_n with internal resistance R (*left*) and noise power delivered to a system whose input impedance is $R_\mathrm{L} = Z$ (*right*)

Because the real conductor with resistance R generates the electrical noise power, it is modelled as the equivalent voltage (or the equivalent current) generator circuit consisting of an ideal voltage source E_n and an ideal resistor R, Fig. 3.2. The average power delivered by a voltage generator of internal RMS voltage E_s and internal resistance R_s to a load R_L is at maximum, assuming matched impedances (i.e., $R_\mathrm{s} = R_\mathrm{L}$),[1] as shown in Fig. 3.2 (right). After substituting (3.2), we have

$$P_{\mathrm{L}\max} = \frac{(E_\mathrm{n}/2)^2}{R} \quad \therefore \quad kT\Delta f = \frac{E_\mathrm{n}^2}{4R} \quad \therefore \quad E_\mathrm{n} = \sqrt{4RkT\Delta f}. \tag{3.3}$$

Equation (3.3) is one of the most often used representations of electrical noise and is therefore widely used in calculating the system noise performance. Sometimes, because of the square root, it is more convenient to work with E_n^2 instead of E_n.

The equivalent noise voltage of combinations of resistors in series and in parallel is calculated after finding the equivalent resistance R or conductance G as

$$E_\mathrm{n}^2 = 4(R_1 + R_2 + \cdots)kT\Delta f = E_{\mathrm{n}1}^2 + E_{\mathrm{n}2}^2 + \cdots, \tag{3.4}$$

$$I_\mathrm{n}^2 = 4(G_1 + G_2 + \cdots)kT\Delta f = I_{\mathrm{n}1}^2 + I_{\mathrm{n}2}^2 + \cdots, \tag{3.5}$$

where R is the equivalent noise resistance, $G = 1/R$ is the equivalent noise conductance, E_n is the equivalent noise voltage, I_n is the equivalent noise current.

Example 3.2. Resistors $R_1 = 20\,\mathrm{k\Omega}$ and $R_2 = 50\,\mathrm{k\Omega}$ are at room temperature $T = 290\,\mathrm{K}$. For a given bandwidth of $BW = 100\,\mathrm{kHz}$, find: (a) the thermal noise voltage for each resistor; (b) for the resistors combined in series; (c) for the resistors combined in parallel.

Solution 3.2.

(a) From (3.3), it follows that

$$E_\mathrm{n}^2(R_1) = 4 \times 20\,\mathrm{k\Omega} \times 1.38 \times 10^{-23} \times 290\,\mathrm{K} \times 100\,\mathrm{kHz} = 32 \times 10^{-12}\,\mathrm{V}^2,$$

$$E_\mathrm{n}^2(R_2) = 4 \times 50\,\mathrm{k\Omega} \times 1.38 \times 10^{-23} \times 290\,\mathrm{K} \times 100\,\mathrm{kHz} = 80 \times 10^{-12}\,\mathrm{V}^2,$$

$$\therefore$$

$$E_\mathrm{n}(R_1) = 5.658\,\mu\mathrm{V},$$

$$E_\mathrm{n}(R_2) = 8.946\,\mu\mathrm{V}.$$

(b) Serial resistance is $R_\mathrm{s} = 70\,\mathrm{k\Omega} \quad \therefore \quad E_\mathrm{n}(R_\mathrm{s}) = 10.59\,\mu\mathrm{V}$.
(c) Parallel resistance is $R_\mathrm{p} = 14.286\,\mathrm{k\Omega} \quad \therefore \quad E_\mathrm{n}(R_\mathrm{s}) = 4.78\,\mu\mathrm{V}$.

[1] This statement is elaborated in more detail in Chap. 6.

3.2 Equivalent Noise Bandwidth

Although reactive components do not generate thermal noise because they do not dissipate thermal power, it is important to estimate the noise power of networks that contain inductive and capacitive reactances. This is because both capacitive and inductive components do influence the frequency bandwidth, hence, the effect of reactances on the noise spectrum must be taken into account. We consider two important network cases for thermal noise calculations: resistor–capacitor (RC) networks and resistor–inductor–capacitor (RLC) networks.

3.2.1 Noise Bandwidth in an RC Network

It can be shown that, when noise passes through a passive filter which has a complex transfer function $H(\omega)$, the noise output spectrum density S_{no} for the input spectrum density (3.1) is (in general) calculated as

$$S_{\text{no}} = |H(\omega)|^2 \, kT. \tag{3.6}$$

In the case of capacitive load (see Fig. 3.3 (left)), the LP filter with noise generator E_{n}^2 and output voltage V_{n}^2 taken across the capacitor has the transfer function $H(s)$ as

$$|H(\omega)| = \frac{1}{\sqrt{1+(\omega RC)^2}} \quad \therefore \quad S_{\text{no}} = \frac{kT}{1+(\omega RC)^2}, \tag{3.7}$$

$$P_{\text{no}} = \int_0^\infty S_{\text{no}} \, \mathrm{d}f = \int_0^\infty \frac{kT}{1+(2\pi RCf)^2} \, \mathrm{d}f = \frac{kT}{2\pi RC} \int_0^\infty \frac{1}{1+x^2} \, \mathrm{d}x. \tag{3.8}$$

The output spectrum, therefore, decreases as the frequency increases due to the bandwidth limiting of the LP filter. The total noise power available at the output is obtained by integrating (3.8) from zero to infinity. The integral is not difficult[2] and the result is

$$P_{\text{no}} = \frac{kT}{2\pi RC} \arctan x \Big|_0^\infty = \frac{kT}{4RC} = kT\Delta f_{\text{eff}} \quad \therefore \quad \Delta f_{\text{eff}} = \frac{1}{4RC}, \tag{3.9}$$

where we introduced Δf_{eff} as "effective noise bandwidth". This definition allows introduction of the equivalent circuit in Fig. 3.3 (right). Thus, the noise spectrum density within the effective bandwidth Δf_{eff} is considered to be equal to kT and zero everywhere else.

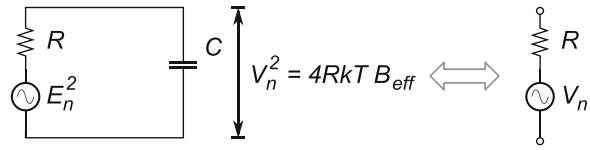


Fig. 3.3 Equivalent noise voltage in an RC circuit

[2] Use of the substitution $(2\pi RCf = x)$ leads to the tabular integral $\int \frac{1}{1+x^2} = \arctan x$.

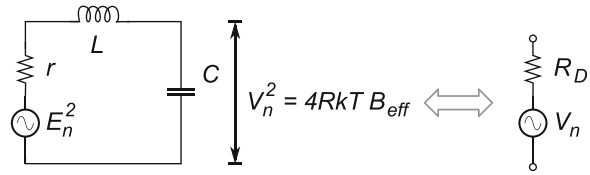


Fig. 3.4 Equivalent noise voltage in an RLC tuned circuit

Furthermore, the equivalent noise voltage V_n can be written as

$$V_\mathrm{n}^2 = 4RkT\frac{1}{4RC} = \frac{kT}{C}, \tag{3.10}$$

which shows that even though the noise was generated by the resistor R, the output noise voltage is not a function of the resistor R. Instead, it is determined by the capacitor C, which does not generate the thermal noise by itself.

Example 3.3. Calculate the equivalent noise voltage V_n generated by a resistor R in series with a $C = 100\,\mathrm{pF}$ capacitor at room temperature $T = 300\,\mathrm{K}$.

Solution 3.3. A straightforward implementation of (3.10) yields

$$V_\mathrm{n}^2 = \frac{kT}{C} = \frac{1.38\times10^{-23}\times300\,\mathrm{K}}{100\,\mathrm{pF}} = 4.14\times10^{-11}\,\mathrm{V}^2 \quad \therefore \quad V_\mathrm{n} = 6.434\,\mu\mathrm{V}.$$

3.2.2 Noise Bandwidth in an RLC Network

The RLC tuned circuit in Fig. 3.4 consists of the ideal lossless capacitor C, and a real inductor L whose resistance r generates noise. We consider the noise voltage E_n as the input to the network and V_n as the output from the network, thus, the modulus of the transfer $H(\omega)$ function is found (using the voltage divider rule) as

$$|H(\omega)| = \frac{|X_\mathrm{c}|}{|Z_\mathrm{s}|}, \tag{3.11}$$

where Z_s is the series impedance of the resonant circuit (5.38)[3] and X_c is the reactance of capacitor C.

If the noise calculation is limited to a narrow bandwidth $\Delta f \ll f_0$ around the resonant frequency f_0 then the transfer function $H(\omega_0)$ (3.11) is approximated as $H(\omega_0) \approx Q$, i.e., the Q factor of the RLC network. Solving an integral similar to (3.6), the mean square output noise voltage is found to be

$$V_\mathrm{n}^2 = Q^2E_\mathrm{n}^2 = Q^2\,4rkT\Delta f = 4R_\mathrm{D}kT\Delta f, \tag{3.12}$$

where $R_\mathrm{D} = Q^2r$ is the "dynamic impedance" of the RLC circuit at resonance. This result is very important for practical calculations because the noise bandwidth in RLC tuned networks is indeed limited to a narrow bandwidth around the resonant frequency.

[3]For more details see Sect. 5.2.1.

If the noise calculation is performed for unrestricted frequency bandwidth, the total noise spectrum must be taken into account by repeating a procedure similar to that in Sect. 3.2.1, while assuming the Q factor to be independent of frequency. After solving a rather more complicated integral, the total noise power is found as

$$P_{\text{no}} = \frac{kT}{4R_{\text{D}}C} = kT\Delta f_{\text{eff}}, \tag{3.13}$$

where $\Delta f_{\text{eff}} = 1/4R_{\text{D}}C$ is the effective noise bandwidth of the RLC network at resonance. It is practical to find the relation between the effective noise bandwidth Δ_{eff} and the bandwidth $B_{3\,\text{dB}}$ of the resonant circuit. It was shown that

$$R_{\text{D}} = \frac{1}{2\pi B_{3\,\text{dB}}C} \quad \therefore \quad \Delta f_{\text{eff}} = \frac{\pi}{2}B_{3\,\text{dB}}. \tag{3.14}$$

Even though it was assumed that the Q factor was constant over the full frequency range, which simplified the analysis, (3.14) is a good indicator of the expected noise. The idea of an equivalent noise bandwidth can be extended to amplifiers and receivers as well.

Example 3.4. A tuned parallel LC tank has the following data: $f_0 = 120\,\text{MHz}$, $C = 25\,\text{pF}$, $Q = 30$, bandwidth $\Delta f = 10\,\text{kHz}$. Find the effective noise voltage of the LC tank at room temperature within the given bandwidth.

Solution 3.4. From (5.81), the dynamic resistance of an LC resonator at resonance is calculated as

$$R_{\text{D}} = \frac{Q}{\omega_0 C} = \frac{30}{2\pi \cdot 120\,\text{MHz} \cdot 25\text{pF}} = 1.59\,\text{k}\Omega$$

then from (3.12)

$$V_{\text{n}}^2 = 4Q^2 R_{\text{L}} kT\Delta f = 4R_{\text{D}} kT\Delta f = 0.254 \times 10^{-12}\,\text{V}^2,$$

$$\therefore$$

$$V_{\text{n}} = 0.50\,\mu\text{V}.$$

3.3 Signal to Noise Ratio

One of the most (arguably, the most) important quantitative measures of a signal's "noisiness" is the signal-to-noise ratio (SNR), which is defined as the ratio of the signal and noise powers,

$$SNR = \frac{P_{\text{s}}}{P_{\text{n}}}, \tag{3.15}$$

where P_{s} is the signal power and P_{n} the noise power. As defined, it shows how many times more powerful is the signal than the noise; it is a relative measure of the two powers. Note that SNR is a unit-less number that merely shows the value of the signal-to-noise power ratio.

It is customary (and also very practical) to express the power ratios in units of *dB*, defined as follows:

$$SNR = 10\log\frac{P_s}{P_n} \quad \text{dB} \tag{3.16}$$

$$= 10\log\frac{V_s^2/R}{V_n^2/R} = 20\log\frac{V_s}{V_n} \quad \text{dB}, \tag{3.17}$$

where V_s the signal voltage and V_n is the noise voltage, measured across the resistive load R terminals. Note the multiplication constants in the expressions for power (3.16) and voltage (3.17). It is trivial to derive an expression similar to (3.17) for currents instead of voltages. Although it may appear a bit counterintuitive to introduce the cumbersome logarithmic function to replace the clean ratio, it turns out that calculations in units of dB are much simpler because the ratios become differences,[4] which is a much simpler arithmetic operation.

The relative measure of power (3.16) tells us only that, for example, P_1 is double P_2 in the case of $SNR = 3$ dB. It does not tell us whether we compared 6 mW to 3 mW or 6 kW to 3 kW, i.e. it does not tell us anything about the absolute power levels. In order to convey that information as well, we need to define the absolute unit of power P_{dBm} as

$$P_{\text{dBm}} = 10\log\frac{P_1}{1\,\text{mW}} \quad \text{dBm}, \tag{3.18}$$

where P_1 power is normalized to 1 mW. In the following sections, we will show examples of how to use dBm units. Its unity step is identical to the dB unity step, which means that adding dB and dBm units is a perfectly legal mathematical operation.

Example 3.5. Convert signal power levels of: (a) $P_1 = 1$ mW, (b) $P_2 = 1$ W, and (c) $P_3 = 10$ W into dBm units. Then, find the SNR of the same three signals if the noise power is $P_n = 1$ mW.

Solution 3.5. A straightforward application of (3.18) leads to: (a) $P_1 = 0$ dBm, (b) $P_2 = 30$ dBm, and (c) $P_3 = 40$ dBm. A straightforward application of (3.16) leads to: (a) SNR = 0 dB, (b) SNR = 30 dB, and (c) SNR = 40 dB.

Note that in the first set of calculations, the signal power levels are still showing the absolute values in dBm, while the second set of calculations show the signal power levels relative to the noise power of 1 mW.

3.4 Noise Figure

Knowing $SNR(in)$ of the signal presented to the input terminals of a circuit network is only one step in the circuit design process. For the purposes of measuring the "noisiness" of the circuit itself, i.e., of finding out how much noise was generated by the circuit's internal components, *SNR* is measured both at the input and output terminals (see Fig. 3.5),

$$SNR(in) = \frac{P_s(in)}{P_n(in)}, \tag{3.19}$$

[4] Some of the basic logarithmic identities are: $\log(x/y) = \log(x) - \log(y)$; $\log(xy) = \log(x) + \log(y)$; $\log(x^n) = n\log(x)$.

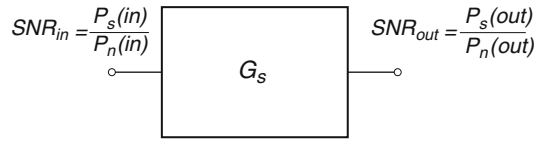


Fig. 3.5 A single stage with signal gain G_s, showing input and output SNRs

$$SNR(out) = \frac{P_s(out)}{P_n(out)}, \tag{3.20}$$

$$\therefore$$

$$F = \frac{SNR(in)}{SNR(out)} = \frac{P_s(in)\,P_n(out)}{P_n(in)\,P_s(out)} = \frac{P_n(out)}{A_P\,P_n(in)}, \tag{3.21}$$

where noise factor F is the ratio of the output and input SNRs and $A_P = P_s(out)/P_s(in)$ is the signal power gain. In practice, any of the three forms in (3.21) can be used to calculate F. If, for example, $SNR(out) = SNR(in)$ then $F = 1$, which is to say that there was no additional noise contribution between the input and output terminals, hence the circuit is noiseless. Note that, as defined, noise factor F is a unit-less number. It is practical to introduce the *noise figure (NF)* as

$$NF = 10\log F \text{ dB}, \tag{3.22}$$

where the noiseless circuit is said to have noise figure $NF = 0\,\text{dB}$, which is the ideal case, yet not achievable in real systems.

Example 3.6. An amplifier has SNR of $SNR(in) = 10$ at its input and $SNR(out) = 5$ at its output. Calculate its F and NF.

Solution 3.6. Using (3.19)–(3.21), simply write

$$F = \frac{SNR(in)}{SNR(out)} = \frac{10}{5} = 2,$$

$$\therefore$$

$$NF = 10\log 2 = 3\,\text{dB}.$$

3.5 Noise Temperature

Thermal noise power was defined earlier, in (3.2), as,

$$T_n = \frac{P_n}{k\Delta f}, \tag{3.23}$$

where index n is added to temperature T to indicate that the noise temperature T_n is referring to the noise power P_n.

For a given amplifier, however, its thermal noise is generated by the internal components and it can be measured at the output terminal. It is convenient in noise analysis to refer the noise back to the input terminal of the circuit and imagine that it is generated by the equivalent external noise source,

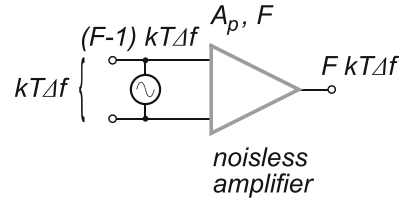


Fig. 3.6 Equivalent input referred noise power for a noiseless amplifier

while the circuit itself is assumed noiseless. If the circuit's power gain is A_P and if the equivalent noise power at the input is P_ni (see Fig. 3.6), then the output noise power P_no is calculated simply as

$$P_\mathrm{no} = A_\mathrm{P} P_\mathrm{ni} \;\therefore\; P_\mathrm{ni} = \frac{P_\mathrm{no}}{A_\mathrm{P}}. \tag{3.24}$$

On the other hand, if the input signal power is P_si and the input noise power is $P_\mathrm{ni} = kT\Delta f$, than from (3.19) the input side signal to noise ratio SNR_in is

$$SNR_\mathrm{in} = \frac{P_\mathrm{si}}{kT\Delta f} \tag{3.25}$$

then, while keeping in mind that both signal and noise are amplified with the same gain A_P, (3.21) can be formatted as

$$F = \frac{P_\mathrm{no}}{A_\mathrm{P} kT\Delta f}. \tag{3.26}$$

Substituting (3.26) back into (3.23), it follows that the total available noise at the input is

$$P_\mathrm{ni} = F\,kT\Delta f. \tag{3.27}$$

Therefore, the amplifier's noise contribution P_na is simply the difference between the output and input noise powers

$$P_\mathrm{na} = F\,kT\Delta f - kT\Delta f = (F-1)\,kT\Delta f. \tag{3.28}$$

Substituting (3.28) into (3.23) (in the case of an amplifier for which $P_\mathrm{n} = P_\mathrm{na}$), it is straightforward to write

$$T_\mathrm{n} = (F-1)\,T, \tag{3.29}$$

where T_n is the noise temperature and T is the ambient temperature. The significance of (3.29) is that it shows the equivalence between noise factor F and equivalent noise temperature T_n (which is not as same as the temperature of the noise source): if one is known, so is the other. In addition, in cases of low noise power levels, noise temperature turns out to be more sensitive than noise factor, which makes the measurements easier. Because of that, noise temperature is used mostly at higher frequencies and in radio astronomy.

Example 3.7. The equivalent noise temperature of an amplifier is $T_0 = 50\,\mathrm{K}$. Calculate the amplifier's noise factor F at room temperature $T = 300\,\mathrm{K}$.

Solution 3.7. Direct implementation of (3.29) leads to

$$T_{\rm n} = (F-1)\,T \quad \therefore \quad 50\,{\rm K} = (F-1)\times 300\,{\rm K} \quad \therefore \quad F = \frac{50\,{\rm K}}{300\,{\rm K}} + 1 = 1.167,$$

$$\therefore$$

$$NF = 10\log 1.167 = 0.669\,{\rm dB}.$$

3.6 Noise Figure of Cascaded Networks

Analysis in Sect. 3.4 and Sect. 3.5 demonstrated that any noise signal $P_{\rm ni}$ presented at the input terminals of an amplifier (or any general circuit, for that matter) is multiplied with its gain $A_{\rm P}$ and produces the output noise signal $P_{\rm no}$, as shown by (3.23). In addition, the amplifier itself generates internal noise $P_{\rm na}$, which is quantified by its noise factor F, as shown by (3.28). Therefore, a single-stage amplifier generates total output noise power P_1 as

$$P_1 = P_{\rm no} + P_{\rm na} = A_{\rm P}\,P_{\rm ni} + (F-1)\,kT\,\Delta f, \tag{3.30}$$

or, in general, rearranging (3.26) we can also write for the total output noise power

$$P_{({\rm no})({\rm tot})} = F_{({\rm tot})}\,A_{{\rm P}({\rm tot})}\,kT\,\Delta f, \tag{3.31}$$

where (tot) is added to indicate that the internal structure of the amplifier may consist of multiple stages.

Let us now evaluate the noise factor of a cascade of networks, each stage with its own noise factor F_i $(i = 1,\ldots,n)$. Considering that system-level analysis is almost always based on a cascade of driver–load pairs, it is important to find an expression for the total noise factor of the cascaded system (see Fig. 3.7).

In its simplest, very important case, the system consists of only two stages $(i = 1,2)$, so that the noise factor F_{12} of the combination is calculated as follows. The input to the first stage is connected to a resistor R_{eq}, which is used to model the thermal noise injected into the two-stage system. For the sake of simplicity, let us assume that the two noise bandwidths Δf of the stages are identical and equal to the noise bandwidth Δf of the cascaded combination.

The total gain $A_{\rm P12}$ of the two stages, obviously, must be

$$A_{\rm P12} = A_{\rm P1}\,A_{\rm P2} \tag{3.32}$$

and, according to (3.26), the noise output $P_{({\rm no})(1)}$ after the first stage is

$$P_{({\rm no})(1)} = F_1\,A_{\rm P1}\,kT\,\Delta f, \tag{3.33}$$

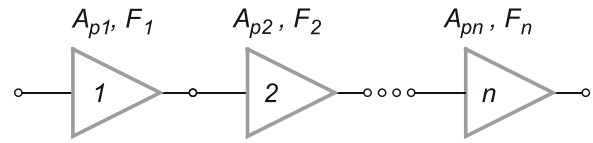


Fig. 3.7 Cascaded system of n stages, each with its own noise factor F_i $(i = 1,\ldots,n)$

which, after being multiplied by the second stage gain A_{P2} is

$$P_{(no)(2)} = A_{P2} P_{(no)(1)} = F_1 A_{P1} A_{P2} k T \Delta f \tag{3.34}$$

if the second stage were noiseless. However, it amplifies its own input referred thermal noise

$$P_{i2} = A_{P2} (F_2 - 1) k T \Delta f. \tag{3.35}$$

Therefore, the total noise output from the second stage is the sum of (3.35) and (3.34)

$$\begin{aligned} P_2 &= A_{P2} (F_2 - 1) k T \Delta f + F_1 A_{P1} A_{P2} k T \Delta f \\ &= \left(F_1 + \frac{F_2 - 1}{A_{P1}} \right) A_{P1} A_{P2} k T \Delta f \\ &= \left(F_1 + \frac{F_2 - 1}{A_{P1}} \right) A_{P12} k T \Delta f. \end{aligned} \tag{3.36}$$

By comparison of (3.31) and (3.36) we have

$$F_{(tot)} = F_1 + \frac{F_2 - 1}{A_{P1}}, \tag{3.37}$$

which is the noise factor expression for a two-stage cascaded network. It is not difficult to generalize (3.37) to a cascaded network of n stages, resulting in

$$F_{(tot)} = F_1 + \frac{F_2 - 1}{A_{P1}} + \frac{F_3 - 1}{A_{P1} A_{P2}} + \cdots + \frac{F_n - 1}{A_{P1} A_{P2} \cdots A_{P(n-1)}}. \tag{3.38}$$

Equation (3.38) is known as Friis's formula and is widely used for evaluating the NF of cascaded networks. Obviously, Friis's formula suggests that in a cascaded network, the noise factor of the very first stage, i.e., F_1, is the most critical because noise factors of the subsequent stages are divided by the combined gain of all previous stages.

Example 3.8. A three-stage amplifier has the following specifications: gain of the first stage is $A_{P1} = 14\,\text{dB}$ and its noise figure is $NF_1 = 3\,\text{dB}$; the second stage has $A_{P2} = 20\,\text{dB}$, and its noise figure is $NF_2 = 8\,\text{dB}$; and the third-stage amplifier is identical to the second stage. Calculate the overall noise figure NF of the system.

Solution 3.8. Using Friis's formula, we write:

$$\begin{aligned} &A_{P1} = 14\,\text{dB} = 25.1, \quad A_{P2} = A_{P3} = 20\,\text{dB} = 100, \\ &NF_1 = 3\,\text{dB}, \quad F_1 = 2, \quad NF_2 = NF_3 = 8\,\text{dB}, \quad F_2 = F_3 = 6.31, \end{aligned}$$

therefore,

$$F_{(tot)} = 2 + \frac{6.31 - 1}{25.1} + \frac{6.31 - 1}{25.1 \times 100} = 2.212 \quad \therefore \quad NF = 10 \log 2.212 = 3.448\,\text{dB}.$$

3.7 Noise in Active Devices

Semiconductor devices generate internal noise due to the discrete nature of electrons crossing p–n junctions. Similar to thermal noise, "shot noise" has a uniform spectrum density. Intuitively, it should not be difficult to visualize that the mean-square shot noise current depends upon the biasing point of the corresponding p–n junction. A number of models describe noise generated by the device, but the following two models are used most often.

First, a temperature-limited diode model (see Sect. 4.3.2) assumes that the emission from the cathode is limited only by temperature, and the mean-square shot noise current is given by

$$I_n^2 = 2q_e I_{DC} \Delta f \qquad \left[A^2\right], \tag{3.39}$$

where

I_n is the p–n junction noise current in A.
q_e is the electron charge (1.6×10^{-19} C).
I_{DC} is the biasing DC in A.
Δf is the effective noise bandwidth in Hz.

The simplicity of the above model is that the shot noise current is calculated based solely on the biasing current.

The second model applies to the semiconductor p–n junction diode and shows that

$$I_n^2 = 2q_e (I_{DC} + 2I_0) \Delta f \qquad \left[A^2\right], \tag{3.40}$$

where I_0 is the reverse saturation current. This model applies only at low frequencies and for low current injection.

More complex device behaviour is observed in form of flicker (or $1/f$) noise and burst (or $1/f^2$) noise. Both of them are very difficult to express analytically so we rely mostly on experimental results, which are usually published by device manufacturers. A qualitative function of noise against frequency dependence is shown in Fig. 3.8.

Example 3.9. For the amplifier in Fig. 3.9 (left), calculate the signal voltage V_s and the equivalent noise voltage V_n appearing at the input terminals. Data: bandwidth $\Delta f = 10\,\text{kHz}$, room temperature $T = 290\,\text{K}$, equivalent internal noise resistance $R_n = 400\,\Omega$, amplifier input resistance $R_i = 600\,\Omega$, source resistance $R_s = 50\,\Omega$, and source voltage $V_s = 1\,\mu\text{V}$.

Solution 3.9. Application of *Thévenin's theorem* on the E_s, R_s, and R_i network results in the following:

$$R_t = \frac{R_s R_i}{R_s + R_i} = 46.15\,\Omega,$$

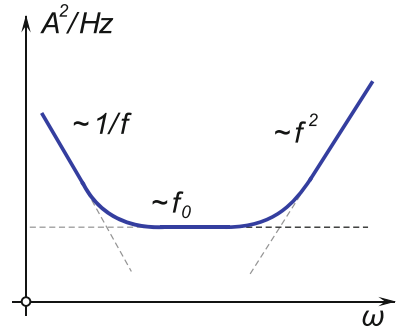


Fig. 3.8 Equivalent noise current spectral density for a bipolar transistor: $1/f$ dependence at low frequencies, approximately constant at medium frequencies (shot and thermal noise), and f^2 dependence at higher frequencies

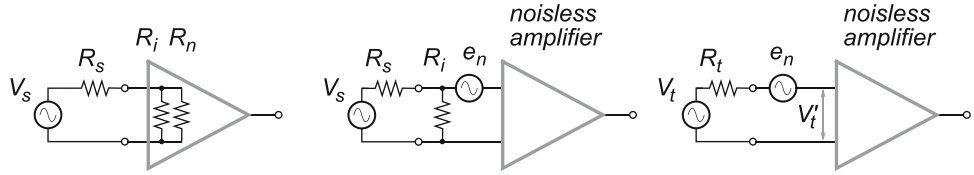


Fig. 3.9 An amplifier with internal noise and input resistances (*left*), the noise equivalent voltage generator (*centre*), and equivalent Thévenin representation (*right*)

$$V_t = V_\mathrm{s} \frac{R_\mathrm{i}}{R_\mathrm{s} + R_\mathrm{i}} = 0.923\,\mu\mathrm{V}.$$

The equivalent noise voltage at the amplifier input is calculated for the serial combination of $R_t + R_\mathrm{n} = 446.15\,\Omega$, which after applying (3.3), results in $V_\mathrm{n} = 0.267\mu\mathrm{V}$.

3.8 Summary

The topic of noise analysis is much broader than presented in this short chapter. A large number of research publications and textbooks are available for further study. In this chapter, we reviewed the most important basic definitions and applications that are considered essential for further discussion. The reader is encouraged to become fluent with the terminology and principles related to noise analysis. In Sect. 13.3.1.1, we expand on the role of noise within the context of signal dynamic range and system sensitivity.

Problems

3.1. Find:

(a) Spectrum density for thermal noise at room temperature ($T = 300\,\mathrm{K}$).
(b) Available noise power within a bandwidth of 1 MHz.
(c) Available signal power for a $1\mu\mathrm{V}$ signal from a $50\,\Omega$ source delivered to the matched load.
(d) SNR for the noise in part 2 and the signal in part 3.

3.2. Determine the noise voltage generated by $50\,\Omega$, $5\,\mathrm{k}\Omega$, and $5\,\mathrm{M}\Omega$ resistors at room temperature 300 K and within a 20 kHz bandwidth.

3.3. Resistors $R_1 = 20\,\mathrm{k}\Omega$ and $R_2 = 50\,\mathrm{k}\Omega$ are at room temperature $T = 290\,\mathrm{K}$. For a given bandwidth of $BW = 100\,\mathrm{kHz}$, find the thermal noise voltage for: (a) each resistor; (b) for their combination in series; (c) for their combination in parallel.

3.4. A tuned parallel LC tank has the following data: $f_0 = 120\,\mathrm{MHz}$, $C = 25\,\mathrm{pF}$, $Q = 30$, bandwidth $\Delta f = 10\,\mathrm{kHz}$. Find the effective noise voltage of the LC tank at room temperature within the given bandwidth.

3.5. For the amplifier in Fig. 3.9, calculate the signal voltage V_s and the equivalent noise voltage V_n appearing at the input terminals. Data: bandwidth $\Delta f = 10\,\mathrm{kHz}$, room temperature $T = 290\,\mathrm{K}$, equivalent noise resistance $R_\mathrm{n} = 400\,\Omega$, amplifier input resistance $R_\mathrm{i} = 600\,\Omega$, source resistance $R_\mathrm{s} = 50\,\Omega$, and source voltage $V_\mathrm{s} = 1\,\mu\mathrm{V}$.

3.6. An oscilloscope probe is specified as $R = 1\,\text{M}\Omega$ and $C = 20\,\text{pF}$, with bandwidth of $BW = 200\,\text{MHz}$. Determine the noise voltage generated due to the probe. If internal circuits of the oscilloscope add 20 dB of noise, determine the effective noise at the input of the oscilloscope.

3.7. A television set consists of the following chain of sub-blocks: two RF amplifiers with 20 dB gain and 3 dB NF each; a mixer with a gain of -6 dB and NF of 8 dB; two additional amplifiers with 20 dB gain and NF of 10 dB each. Calculate: (a) the system NF and (b) the system noise temperature.

3.8. An amplifier with input signal power of $5 \times 10^{-16}\,\text{W}$ and input noise power $1 \times 10^{-16}\,\text{W}$ has output signal power of $5 \times 10^{-12}\,\text{W}$ and output noise power of $4 \times 10^{-12}\,\text{W}$. Determine the noise factor F and the NF of this amplifier.

3.9. Calculate the noise current and equivalent noise voltage for a diode biased with $I_{\text{DC}} = 1\,\text{mA}$ at room temperature 300 K and within the bandwidth of 1 MHz.

3.10. The equivalent noise resistance of an amplifier is $R_{\text{in}} = 300\,\Omega$ and the equivalent shot noise current is $5\,\mu\text{A}$ at room temperature $T = 300\,\text{K}$. The signal generator has internal resistance $R_{\text{S}} = 150\,\Omega$ and provides a signal of $V_{\text{S}} = 10\,\mu V_{\texttt{rms}}$. Calculate the input $SNR(in)$, if the operational bandwidth is $\Delta f = 10\,\text{MHz}$.

3.11. A front-end RF amplifier whose gain is 50 dB and noise temperature is 90 K provides a signal to a receiver that has a NF of 12 dB. Calculate the noise temperature of the receiver by itself and the overall noise temperature of the amplifier plus the receiver system at room temperature $T = 300\,\text{K}$.

Chapter 4
Electronic Devices

Abstract Analysis and modelling of a general electrical network is based on four fundamental mathematical functions, associated with the behaviour of ideal devices, namely resistance (R), capacitance (C), inductance (L), and memristance (M). Each of these four fundamental elements is assumed to have one and one only property under all conditions and at any given time. For example, resistance R is always assumed to be the multiplying constant in the linear relationship between voltage and current at its terminals, i.e. $V = R \times I$. It is also assumed that both voltage and current can take any numerical value within the $[-\infty, +\infty]$ range. That is, the ideal elements have an infinite power-handling capability, either as the energy source or the energy sink.

In a very broad sense, all network elements that obey basic network laws can be classified as being either *passive* or *active*. In this chapter, we review the properties of the fundamental electronic devices and the basic laws.

4.1 Simple Circuit Elements

The main property of passive elements is that they absorb energy and subsequently convert it, for example, into heat. However, passive elements cannot generate energy, that is, passive elements are not capable of "power gain". In this section, we review the properties of the following basic passive elements: a simple conducting wire, ideal voltage and current sources, resistance, capacitance, inductance, a transformer, and memristance. In addition, we review arguably the simplest and one of the most important passive networks, the voltage divider.

4.1.1 Simple Conductive Wire

In circuit theory, an ideal conductive wire is defined as the most basic electrical element, possessing the following properties:

- It may be an arbitrary length from zero to infinity.
- Its surface is an ideal equipotential entity.
- It is a non-material entity capable of carrying an infinite amount of power.

A direct consequence of these assumptions is that there is no voltage difference between any two points on the wire, under all conditions. That is, an ideal conductive wire has zero resistance, zero

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8_4,

capacitance, and zero inductance. These simplifications greatly reduce the complexity of circuit analysis because, as a first approximation, the electrical influence of the wires is completely neglected.

A real wire, on the other hand, is made of physical matter (e.g. aluminum, gold, or copper), which means that there must be a physical limit to how much energy it can absorb before it overheats and melts due to its internal resistance. Hence, in reality, a wire does not behave as an equipotential entity. Consequently, there is always some voltage difference between the various points on its surface whose amplitude depends both on the external conditions and its internal structure. The operating range of a real wire is limited and set, on one side, by the inherent atomic thermal vibrations that generate electrical *noise*. The noise sets a limit on the minimum level of signal that the wire can carry before the signal is swamped by the noise. On the other side, too high a current forced through the wire causes the internal atomic vibrations to increase too much, which is perceived by the external world as an increase in temperature, which causes eventual physical destruction of the wire. As a side note, this heating phenomenon is not always necessarily a bad thing. For example, an electrical fuse is designed to exploit exactly that property of a real wire—if too much current is forced through it, the fuse wire splits, which breaks the current flow and protects other devices down the stream. In all other cases where the heat generation is not the primary goal, reduction of the wire's internal resistance is desired for a number of reasons, e.g. low resistance of the wire reduces the waste of energy due to thermal power dissipation and it enables the design of a coil with a high Q factor (see Sect. 4.1.6). In order to quantify its imperfection level relative to the ideal wire, a real wire is modelled as an RLC network where the internal RLC values are derived from the wire's geometry and its material properties. Because the wire is not intended to behave as a resistor, a capacitor or an inductor, those RLC values are referred to as "parasitic components" and they are, therefore, included in the realistic approximate wire model.

4.1.1.1 DC and RF Behaviours of a Simple Wire

Resistance to electric current flow is a fundamental property of all material conductors, including "simple" metal wires and printed circuit board (PCB) traces. For purposes of developing a good engineering feeling about a conductor's real behaviour across a range of frequencies, we need to consider currents starting from zero frequency $\omega = 0$, (i.e. *DC*) to the unachievable theoretical limit of infinite frequency ($\omega \to \infty$). It is important to learn how to quantify the change of the conductor's resistance as a function of the frequency so that we can determine its practical range of operation. Starting from the zero-frequency case, i.e. DC, it can be shown that the DC wire resistance is directly proportional to the conductor's length l and inversely proportional to the wire's cross-sectional area S and the material's conductivity σ. Intuitively, we visualize DC current flow as a river of electrons (i.e. the charge carriers) rushing through the conductor which, from the inside, looks more like a long prison hallway with bars every few steps because the metallic crystal lattice behaves as a rigid three-dimensional mesh. The longer the flow path, the higher the probability that more electrons collide with atoms in the lattice and pass on some of their kinetic energy, which increases the lattice vibrations, i.e. the conductor temperature. Continuing the prison hallway analogy, everything else being equal, the wider the diameter of the hallway (i.e. the larger the conductor's cross-sectional area), the more easily the current flows because there is more space for the electrons to spread (i.e. reduce the current density) and, therefore, it reduces the probability of hitting the lattice too often. In addition, not all materials have the same shape as a crystal lattice; some are more dense than others, therefore their resistivity constant ρ is widely different. The higher the resistivity, the higher the wire's resistance to the current flow. This reasoning is summarized in the well-known formula for wire resistance at DC

$$R_{\mathrm{DC}} = \rho_{\mathrm{cond}} \frac{l}{S} = \frac{l}{\sigma_{\mathrm{cond}} S} = \frac{l}{\sigma_{\mathrm{cond}} \pi a^2} = \frac{J}{\sigma_{\mathrm{cond}}}, \tag{4.1}$$

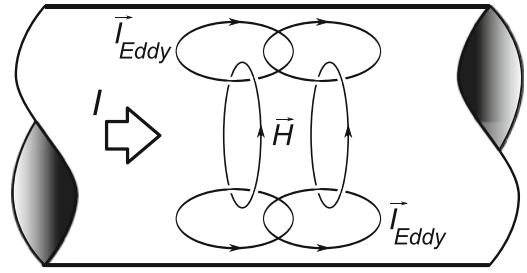


Fig. 4.1 Induced current causing the skin effect inside a cylindrical conducting wire

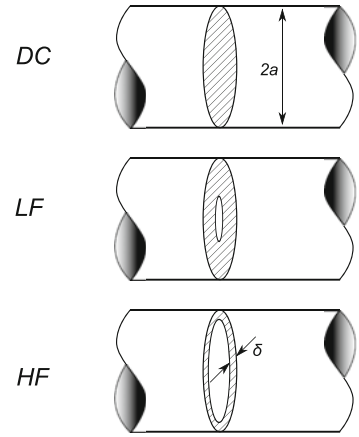


Fig. 4.2 Cross-sectional view of a round wire showing the skin effect. Note that the border between the conducting "skin" layer and the internal section of the wire is not as abrupt as shown—the change of resistance is gradual and the border line is calculated using the formula for skin depth

where R_{DC} is the DC wire resistance, l is the wire length, $\rho = 1/\sigma$ is the metal resistivity constant, σ is the metal conductivity constant, S is the wire's cross-sectional area ($S = \pi a^2$ in the case of a round wire whose radius is a), and $J = I/S$ is the current density across the cross-sectional area S.

A less obvious, and trivial at first glance, but very important point to note in (4.1) is that it assumes the DC I is uniformly distributed across the cross-sectional area $S = \pi a^2$, i.e. that the DC density is constant throughout the entire volume of the conducting wire. That assumption results in good accuracy of (4.1) for calculating a wire's DC resistance.

However, that conclusion becomes questionable in the more complicated case of alternating current (AC) flowing through the wire. In accordance with Faraday's law (Sect. 1.5.2), AC flow creates an alternating magnetic field, which further induces an alternating electric field. This induced electric field forces its own induced current (also known as "eddy currents") whose direction is such that it opposes the initial AC current flow (see Fig. 4.1). Moreover, this effect is not uniform; it is strongest at the wire's centre, i.e. for radius $r = 0 (r \leq a)$. This *skin effect* is perceived, in accordance with *Ohm's law*, as being a consequence of increased material resistance in inner regions of the wire, leaving only a thin layer close to the surface to carry all the current (that is where the "skin" part of the expression "skin effect" comes from), which is to say that the cross-sectional area of the wire is effectively reduced to a ring close to the wire perimeter (see Fig. 4.2). More importantly, the skin effect is more and more pronounced as the frequency increases, which is to say that at very high frequencies the cylindrical wire is reduced to a tube with a thin wall, as shown in Fig. 4.2 (bottom).

The observation that an increase of the AC frequency causes progressive increase in the wire resistance is very important because the wire's cross-sectional area (which, at DC, was a full circle) is reduced to a narrow ring and one could, rightly, conclude that at high frequencies there is no advantage to using a solid conductive wire. Instead, the same current carrying capability is achieved using hollow tubes, with the benefit of using less material, which results in much lighter wire transmission systems.

4.1.1.2 Skin Depth of a Simple Wire

In order to quantify the skin depth for a given frequency and material conductivity, one starts from Maxwell's equations (B.6) and (B.8). Eventually, it is shown that AC density J_z in the z direction is

$$J_z = \frac{p\,I}{2\pi a}\frac{J_0(pr)}{J_1(pa)}, \tag{4.2}$$

where J_z is the density of the total current I along the z axis, $p^2 = -j\omega\mu\sigma_{\text{cond}}$, J_0 and J_1 are Bessel functions of the zeroth and first orders, respectively, a is the wire diameter, and $r \leq a$ is a distance from the wire centre inside the wire. Hence, for a given wire diameter $2a$, J_z is a complex function of the radius $r, [0 \leqslant r \leqslant a]$. In the case of DC, (4.2) reduces to the known equation for current density J in a round cylindrical wire

$$J_{z0} = \frac{I}{\pi a^2}, \tag{4.3}$$

where J_{z0} is the DC density in the z direction.

The skin effect was noticed and studied relatively early in history. The studies of interaction between an EM wave and a conductive material uncovered a decline in AC density with the depth of the material, with the AC magnitude being greatest at the conductor's surface. Theoretical analysis of an infinitely thick slab of conductive material revealed that the current density decreases exponentially with depth d from the surface:

$$J(d) = J_{\text{S}}\,\text{e}^{-\left(\frac{d}{\delta}\right)}, \tag{4.4}$$

where J is the current density inside the conductive material at depth d, J_{S} is the current density at the conductor's surface, and δ is the skin depth. By convention, at skin depth $d = \delta$, the current density falls to $1/e$ of its value J_{S} at the surface (4.4). The widely cited formula for skin depth is

$$\delta = \sqrt{\frac{2\rho_{\text{cond}}}{\omega\mu_{\text{cond}}}} = \sqrt{\frac{1}{\pi f \mu_{\text{cond}}\,\sigma_{\text{cond}}}}, \tag{4.5}$$

where ρ is the conductor resistivity, $\omega = 2\pi f$ is the angular frequency of the current, $\mu_{\text{cond}} = \mu_0\mu_{\text{r}}$ is the magnetic permeability of the conductor, μ_0 is the magnetic permeability of the vacuum, μ_{r} is the relative magnetic permeability of the conductor, and σ_{cond} is the conductor's conductivity.

A detailed derivation of (4.5) from Maxwell's equations can be found in a number of textbooks. The calculated skin depth for three commonly used metals is shown in Fig. 4.3. Equation (4.5) shows that the skin depth is inversely proportional to the square root of the material conductivity σ_{cond} and drops to zero for a perfect conductor, i.e. the EM wave cannot penetrate a perfect conductor, it simply reflects back. Only lossy realistic materials suffer from the skin effect.

A long cylindrical conductive wire, whose diameter is $D \gg \delta$, has resistance approximately equal to that of a hollow tube with wall thickness δ carrying DC. That is, its AC resistance is approximately:

$$R = \rho\frac{l}{S} = \rho\frac{l}{\frac{\pi D^2}{4} - \frac{\pi(D-2\delta)^2}{4}} = \frac{\rho}{\delta}\frac{l}{\pi(D-\delta)} \approx \frac{\rho}{\delta}\frac{l}{\pi D}, \tag{4.6}$$

where l is the wire length and $D = 2a$ is its diameter. It should be noted that (4.6) is valid only for a single isolated wire. If there is a second wire nearby that also carries alternating current, then there is

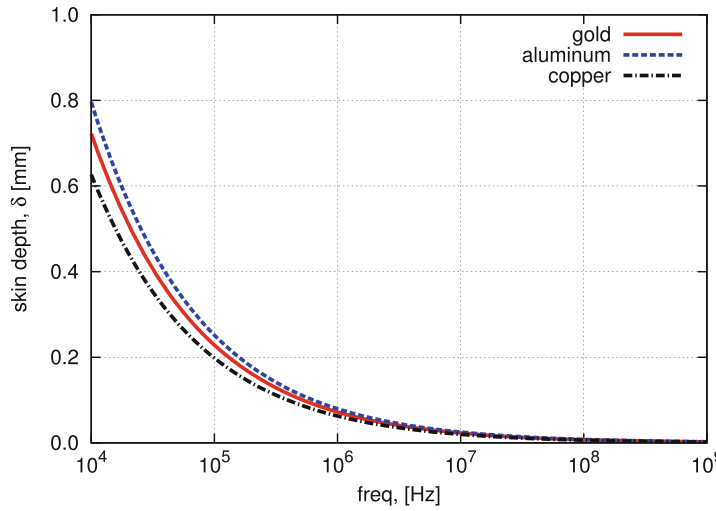


Fig. 4.3 Skin depth for three commonly used metals: gold, aluminum and copper

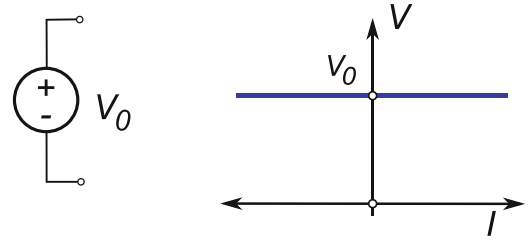


Fig. 4.4 An ideal voltage source circuit symbol and its V–I characteristics

additional eddy current induced in the first wire. This phenomenon is known as the "proximity effect". Analysis of the overall current density distribution becomes much more complicated in this case. For purposes of discussion in this book, the proximity effect is ignored.

4.1.2 Ideal Voltage Source

Circuit network analysis is based on the assumption that every ideal basic circuit element performs one and one only function under all (mathematically) possible conditions. Here, the term "function" refers to a relationship between voltage and its corresponding current at each of the device's nodes. The problem (or opportunity, depending on how you look at it) is that there is an infinite number of possible functions that can establish the voltage–current relationship, which leads to an equal number of possible ideal devices that can be defined. The simplest possible voltage function is that the voltage does not depend upon its corresponding current, i.e. $V \neq f(I)$, under all conditions. In other words, an ideal two-terminal element is capable of holding the preset voltage V_0 amplitude at its terminals, regardless of how much current flows through. The mathematical abstraction of such an element is known as an "ideal independent voltage source" (Fig. 4.4). In addition, as a general case, it is also possible to define a controlled ideal voltage source as a four-terminal device, two at the input and two at the output, where the output voltage is controlled by either voltage or current at its input terminals. Therefore, there are three possible flavours of ideal voltage source: an independent voltage source (a two-terminal device as shown in Fig. 4.4); a voltage-controlled voltage source (VCVS) (a four-terminal device); and a current-controlled voltage source (CCVS) (a four-terminal device).

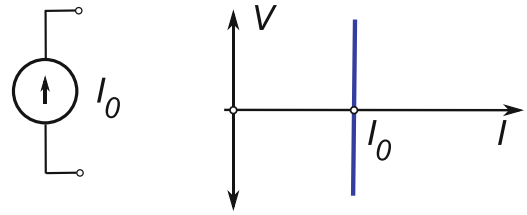


Fig. 4.5 An ideal current source symbol and its V–I characteristics

The theoretical ideal voltage source is capable of either delivering or absorbing an infinite amount of power because the current is allowed to take any value within the range $[-\infty, +\infty]$. The usual convention in circuit theory is that if current is entering a terminal at higher potential, the corresponding power is considered to be absorbed by the ideal element. The ideal element is said to deliver power to the outside circuit if the current is leaving the node at higher potential. Another important property of an ideal voltage source is its internal resistance, as seen through its output terminals. By definition, resistance is the change in voltage over the change in current, i.e. the derivative of voltage against current,

$$R \triangleq \lim_{\Delta \to 0} \frac{\Delta V}{\Delta I} = \frac{\mathrm{d}V}{\mathrm{d}I} = \frac{0}{\mathrm{d}I} = 0, \tag{4.7}$$

which leads to the conclusion that, inherently, the internal resistance of an ideal voltage source must equal zero. This property is very important, because any physical device or circuit that presents very low resistance at its terminals may be approximated and classified as a voltage source within a finite range of voltage and current values at its terminals. Of course, unlike their mathematical abstraction, a realistic device (e.g. a bipolar junction transistor) cannot hold voltage at its emitter terminal over an infinitely wide range of emitter currents. Hence, real devices can only approximate the ideal voltage source model over a very limited range of voltages and currents.

4.1.3 Ideal Current Source

A similar element is known as an "ideal current source" (see Fig. 4.5). By definition, an ideal current source is capable of holding the preset current amplitude regardless of the voltage at its terminals. Most of the comments made about ideal voltage sources apply also to ideal current sources, e.g. ideal current sources can also absorb or deliver an infinite amount of power. Again, as a general case, it is possible to define a controlled ideal current source as a four-terminal device, two at the input and two at the output, where the output current is controlled by either the voltage or current at its input terminals. Therefore, there are three possible flavours of ideal current source: an independent current source (a two-terminal device as shown in Fig. 4.5); a voltage-controlled current source (VCCS) (a four-terminal device); and a current-controlled current source (CCCS) (a four-terminal device).

A very important observation is that internal resistance of an ideal current source, as seen through its output terminals, is very different. Again, by definition, resistance is the change in voltage over the change in current, i.e. the derivative of voltage against current,

$$R \triangleq \lim_{\Delta \to 0} \frac{\Delta V}{\Delta I} = \frac{\mathrm{d}V}{\mathrm{d}I} = \frac{\mathrm{d}V}{0} = \infty, \tag{4.8}$$

which leads to the conclusion that, inherently, the internal resistance of an ideal current source must equal infinity. This property is very important because any physical device or circuit that presents very

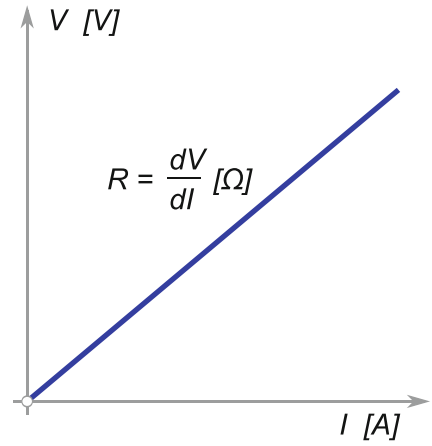


Fig. 4.6 Ideal V–I resistor characteristics

high resistance at its terminals may be approximated and classified as a current source within a finite range of voltage and current values at its terminals. Of course, unlike its mathematical abstraction, realistic devices cannot hold current at their terminals over an infinitely wide range of voltages.

4.1.4 Resistance

A resistor (a two-terminal device) is defined by a mathematical function that describes the relationship between voltage and current at its terminals. By definition, an ideal resistor is a linear device whose voltage and current at its terminals are proportional to each other in accordance with Ohm's law as

$$V = R\,I, \tag{4.9}$$

where resistance R is, in the case of a linear ideal device, the proportionality constant (see Fig. 4.6). In other words, the main purpose of a linear resistor is to create a voltage difference at its terminals that is proportional to the current flowing through. It achieves that task by converting some of the electrical energy into heat energy. An ideal resistive element is capable of absorbing an infinite amount of power.

Many types of material can be used to create a slab whose main property is to create a voltage difference at its terminals, within limited ranges of voltages and currents. Some of most commonly used materials for manufacturing realistic resistive components are:

- High-density carbon composites, usually in the shape of a cylinder.
- Metal film, either thick or thin.
- Metal wire wound around a nonconductive cylindrical core.
- Doped silicon or poly-silicon, used in IC technologies.

Traditional resistors are mostly through-hole components, i.e. cylindric devices that have wire leads as terminals, while modern RF and high-speed circuits employ much smaller surface-mounted devices (SMD) (see Fig. 4.7) that have much smaller parasitic components. The relationship between a resistor's geometry (its physical parameters) and its resistance (its electrical parameter) is given by

$$R = \rho\,\frac{l}{S}, \tag{4.10}$$

Fig. 4.7 State-of-the-art surface-mounted resistors are only a couple of millimetres long

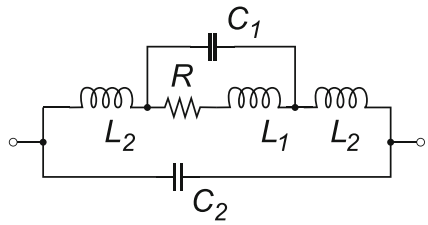


Fig. 4.8 Equivalent electrical circuit model of a high-frequency, wire-wound resistor. This model is based on lumped ideal RLC elements and is, therefore, suitable for a low to medium range of frequencies. In the case of very high frequencies, a more elaborate model, based on distributed elements, needs to be used

where R is the resistance, ρ is the material's resistive constant, l is the length, and S is the cross-sectional area.

The electrical behaviour of a real resistor is much more complicated than the basic ideal model (4.9) suggests. The main reason is that a real resistor is build using several materials, each of them with a different conductivity constant σ, a different temperature coefficient (TC), and so on. In Sect. 4.1.1.1 it was shown that even a simple wire on its own, which is commonly used to make the lead terminals for through-hole resistors, turns into a very complicated device once AC starts to flow through. To make things worse, a wire-wound resistor is designed to be the same shape as a regular inductor. The only difference is in the resistivity constant of the two wires: inductors are made of a very low-resistance metal wire and resistors are made of high-resistance alloys. In addition, there is a capacitive effect between the wire turns, between the resistor's body and the environment, and between the two wire terminals.

Consequently, the equivalent electrical circuit model needed to capture the behaviour of a real resistor over a range of frequencies includes, aside from the intended resistor device, the parasitic capacitances and inductances needed to model the behaviour of the real materials used to manufacture the resistor. In one commonly used model (suitable for a low to medium range of frequencies), component R represents the ideal intended value of the resistor, L_1 represents the parasitic inductance of the wire used to create the resistor, L_2 is the inductance of the wire used to crate the resistor's leads (ignored in the case of carbon or metal film resistors), C_1 is the parasitic capacitance associated with the wire coil used to create the resistor, and C_2 is the parasitic capacitance associated with the leads and the whole resistor itself, often referred to as the "feed-through" or "stray" capacitance (see Fig. 4.8).

A simple numerical analysis (see Fig. 4.9) of a resistor model whose DC resistance value was designed to be (for example) $R = 500\,\Omega$, shows that there are four distinctly different frequency regions where the resistor's behaviour is drastically different. Although the numbers shown in Fig. 4.9 are specific only for this particular example, the curve shape is similar for other examples.

- DC to 20 MHz: Inside this frequency region, the resistance is dominant. Note that the resistor's value does not change significantly with the frequency increase, i.e. $|Z| \approx R \neq f(\omega)$.

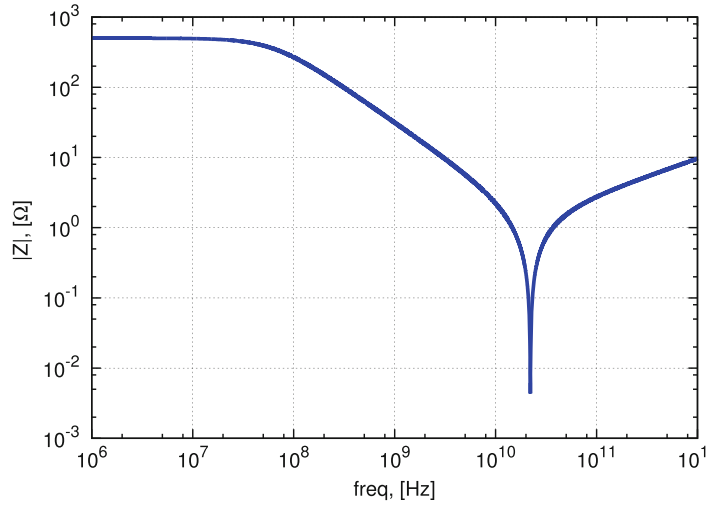


Fig. 4.9 Absolute impedance against frequency characteristics of a typical wire-wound resistor. A real resistor model (Fig. 4.8) used the following ideal component values: $R = 500\,\Omega$, $L_1 = 1.54/\sqrt{f}$, and $C_a = 5\,\text{pF}$

- 100 MHz to 10 GHz: Beyond the region in which $|Z|$ is approximately constant, the capacitive behaviour becomes dominant, which is illustrated by the drop in the impedance amplitude. It is consistent with capacitive impedance behaviour (capacitive impedance is inversely proportional to the frequency).
- 10 to 30 GHz: A sharp, pointy region is a very important property of any physical object. For the time being, let us only remember the frequency of the minimal point (in this case approximately 25 GHz) as the "self-resonating" frequency. More details of resonance in general are given in Chap. 5. In addition, resonance as a phenomenon is the fundamental principle behind wireless radio transmission and remains one of the main topics throughout this book.
- Above 30 GHz: At very high frequencies, inductive behaviour becomes most prominent, which is characterized by an increase in the impedance amplitude as the frequency increases, which is the typical behaviour of an ideal inductor.

This example illustrates the complexity of behaviour associated with real components due to frequency dependence that is caused by their internal parasitics, which directly limits the useful operating range of frequencies of real components.

4.1.4.1 Linear and Nonlinear Resistance

The definition of linear resistance in its basic form, (4.9), applies only to ideal linear resistor components whose voltage vs. current derivative is constant. In both cases, direct application of (4.9) is correct because the ratio of voltage to current is constant, i.e. R is constant. In a more realistic general case of resistance, which is inherently nonlinear in terms of the voltage–current relationship at the resistor's terminals, the change in voltage that corresponds to a change in current gives the correct value of point-by-point resistance.

One of the three possible forms of Ohm's law, which describes the relationship of time-varying voltage and current, is:

$$v = iR \Rightarrow R = \frac{v}{i}, \tag{4.11}$$

$$\therefore$$

$$R \triangleq \lim_{\Delta \to 0} \frac{\Delta v}{\Delta i} = \frac{\mathrm{d}v}{\mathrm{d}i}, \tag{4.12}$$

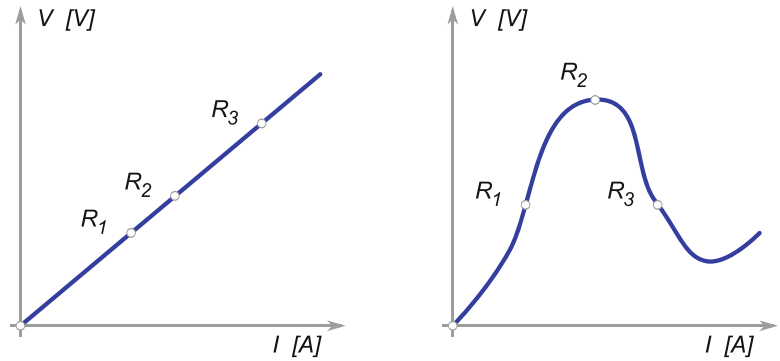


Fig. 4.10 Voltage–current characteristics of a linear element (*left*) and a nonlinear element (*right*)

which should be interpreted as being valid only at a particular voltage–current point. To clarify this statement, take a look at Fig. 4.10 which shows two VI characteristics: one for a linear element (e.g., an ideal resistor) and one for a nonlinear element (e.g., a diode). As shown in (4.12), strictly speaking, a resistance is defined as the first derivative of voltage over current, which implies the validity of (4.12) only at that particular VI point, which is a more natural interpretation of the concepts of *negative resistance* and the biasing point.

In the case of a linear resistor, the first derivative does not change from point to point in the VI plane, Fig. 4.10 (left); the resistance of linear devices is constant, i.e. $R_1 = R_2 = R_3$. For a nonlinear device, Fig. 4.10 (right), it should be obvious not only that at each point R_1, R_2, and R_3 the respective derivatives are different, but also that two possible values for resistance can be calculated at each point. For example, for point R_1, direct use of (4.9) results in the value of $R_1 = V_1/I_1$, while use of (4.12) generates a quite different value for R_1, which is due the definition of the first derivative. The situation at point R_2 is even more surprising: direct application of (4.9) gives $R_2 = V_2/I_2$ and application of (4.12) results in $R_2 = 0$. Finally, at point R_3, direct application of (4.9) gives $R_3 = V_3/I_3$, however, (4.12) gives $R_3 < 0$. What is more, two current values are associated with the same voltage value: note that the voltages across R_1 and R_3 are equal, while the currents are widely different.

The problem arises from the fact that application of (4.9) requires only a single measurement of the current and voltage. This is not sufficient to answer the question of what happens with the voltage–current relationship at other points. In contrast, application of (4.12) requires several readings, i.e. a priori knowledge of several voltage–current readings in close proximity to each of the points (V_k, I_k) before it is possible to apply (4.9). Usually, full nonlinear V–I characteristics are provided for practical nonlinear devices. Points defined by the (V_k, I_k) pairs are referred to as "DC biasing points", resistances calculated by (4.9) are "DC resistances" and resistances calculated using (4.12) are known as "AC resistances". Therefore, we have not one but two valid results for resistance at the same (V_k, I_k) pair, hence the need to specify the corresponding biasing points when calculating AC parameters of a circuit that contains nonlinear devices, such as diodes and transistors.

4.1.4.2 AC Signal Generator and Resistive Load

Parallel connection of a single-tone generator and a purely resistive load, as shown in Fig. 4.11 (left), forms the simplest AC circuit. In this section, we review the important characteristics of this class of circuits in terms of its voltage–current–power relationship. Because an AC signal, mathematically described by sinusoidal function, constantly changes in time, it is convenient to use its RMS value to quantify the energy transfer between the source and the load. We already know (Sect. 2.7.1) that electrical power is the product of voltage and current. In the same section, (2.39) showed that the phase

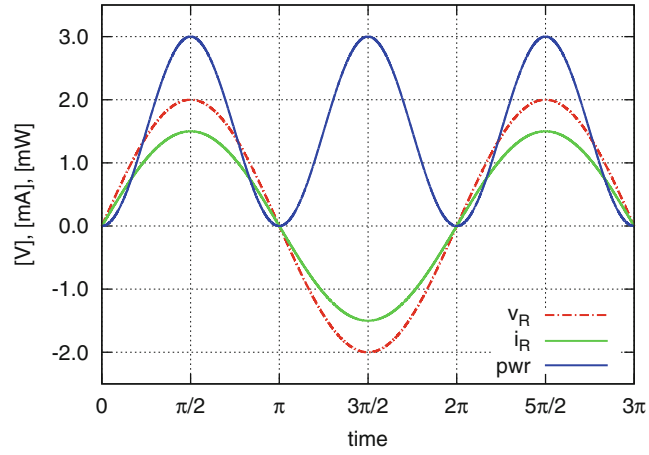


Fig. 4.11 A circuit with purely resistive load and AC voltage generator (*left*) and the corresponding voltage–current–power time domain plot (*right*). In this example, $E_0(t) = 2\sin(\omega t) = v_R$ V and $i_R = 1.5\sin(\omega t)$ mA

difference between voltage and current waveforms is an important factor. By inspection of (4.9), we note that resistance by itself does not have a frequency-dependent component, hence its voltage v_R and current i_R measured at the terminals must be in phase, as shown in Fig. 4.11 (right).

It is important to observe that:[1]

- Because the voltage and the current through a resistor are in phase, i.e. for half a cycle both are positive and for half a cycle both are negative, the power is always positive (it always flows out of the generator into the resistor and dissipates in heat).
- The power cycle is half the signal cycle.
- Even though average values for voltage and current are zero, the average power is halfway between its minimum and peak values; for Fig. 4.11 (right), it is calculated as $(3\,\text{mW} + 0\,\text{mW})/2 = 1.5\,\text{mW}$.
- At any moment, the resistor value is $R = v(t)/i(t)$; for Fig. 4.11 (right), $R = 2\,\text{V}/1.5\,\text{mA} = 1.333\,\text{k}\Omega$.
- The equivalent DC voltage E that is needed to generate a power level equal to the average power is $E = \sqrt{PR}$; for Fig. 4.11 (right), $E = \sqrt{1.5\,\text{mW} \times 1.333\,\text{k}\Omega} = 1.414\,\text{V}$.

By now, we should be quick to realize that the effective DC voltage E is what is commonly referred to as the "RMS" value of the AC peak voltage V_m. The same $1/\sqrt{2}$ factor applies to the current peak value, which leads to a relation for the RMS power as $P_\text{rms} = 1/\sqrt{2}V_\text{m} \times 1/\sqrt{2}I_\text{m} = {}^1\!/_2 V_\text{m} I_\text{m}$. To conclude, the average and RMS values are not equal, even if (for example) the voltage and current are calculated over the same half cycle.

4.1.5 Capacitance

The intuitive introduction of resistance as a material property in Sect. 4.1.1.1 is based on a physical argument that follows from the flow of electrons inside a slab of material and their interaction with the material's crystal lattice. We concluded that resistance is related to the ability of the material to allow the flow of free electrons that are influenced by a force due to the electrical field caused by potential difference at the material surfaces. We keep in mind though that the voltage difference was caused by an imbalance in the number of electrons at the two surfaces of the slab. The internal stress was created

[1] Reminder: $a\sin(x) \times b\sin x = \frac{ab}{2}\sin(2x)$.

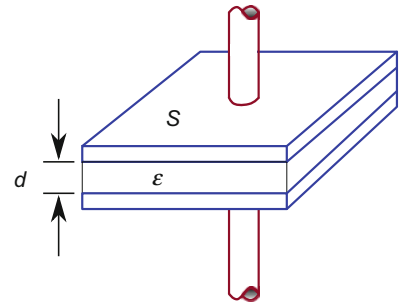


Fig. 4.12 Parallel plate capacitor structure

by the attracting electric force and it is reduced only by providing a low-resistance path between the two surfaces. However, if there is an external "pump" (i.e. an electron source—a battery) attached to the slab that is capable of reintroducing the electron imbalance, then the electric force is maintained and we perceive the steady flow of the electrons as the current. Based on this reasoning, we learned how to control the current by designing various conductive materials and voltage sources.

Following the same idea, we could ask how the charging process itself is controlled before the current is allowed to flow. Obviously, in order to introduce a potential difference between two objects separated in space (i.e. with no conductive path between them), we need to bring a certain number of electrons Q to one of the objects and store it there. We define *capacitance* C as the proportionality constant that defines the number of electrons Q that are required to create $V = 1$ V potential difference between the two objects, i.e.

$$Q \equiv C \times V, \tag{4.13}$$

where the charge Q and the potential V hold as long as the two objects are separated.

From a physical perspective, capacitance is calculated either for any two objects at different potentials separated in space by the high-resistance insulating layer or for any two points of the same object at different potentials (which is referred to as "self-capacitance" or parasitic capacitance). If resistance of the insulation material is assumed infinitely high, then equal and opposite charges placed on these two objects can never combine through the insulating layer; the charge recombination can happen only through the external path. We define a *capacitor* as a two-terminal device whose main role is to store and hold charges on its two separated surfaces. That is why a capacitor is also referred to as an energy storage device, where the energy is stored in its internal electric field.

From the implementation perspective, the most commonly used shape of capacitor is a parallel plate capacitor, which is made of two thin metal sheets with a thin insulating layer sandwiched between them. Similarly to (4.10), the relationship between the geometric properties of a parallel plate capacitor and its capacitance is given as

$$C = \varepsilon \frac{S}{d} = \varepsilon_0 \varepsilon_\mathrm{r} \frac{S}{d}, \tag{4.14}$$

where C is the capacitance that is directly proportional to the surface area S of the two conducting plates and to the permittivity of the insulating layer $\varepsilon = \varepsilon_0 \varepsilon_\mathrm{r}$ and inversely proportional to the separation distance d between the plates (see Fig. 4.12). The relative permittivity of the insulating material is ε_r and ε_0 is the vacuum permittivity. Even though (4.14), strictly speaking, applies only to a plate capacitor, the capacitance of many other shapes can be reasonably well approximated by the same formula. In order to reduce the component size, especially for larger values of capacitance C, a parallel plate capacitor is usually rolled into a cylindrical tube. To further increase the capacitance within a given volume, materials with higher permittivity values must be used, for example electrolytic

capacitors use an electrolyte (an ionic conducting liquid) as one of the plates. And, of course, making the insulation layer thinner increases the capacitance, as long as the insulating material's resistance is high enough to stop leakage of current caused by the strong electric field.

In the general case, when the voltage potential across the capacitor is not constant, we define the process of the capacitor charging by monitoring how fast the charges are brought to the capacitor plate. Mathematically, it is equivalent to finding the first derivative of (4.13), i.e.

$$\frac{\mathrm{d}}{\mathrm{d}t}Q \equiv i = C\,\frac{\mathrm{d}v}{\mathrm{d}t}, \tag{4.15}$$

where AC voltage v and current i satisfy the convention for passive elements, i.e. the current enters the terminal that is at higher potential. Therefore, the general definition of capacitance C is that it is the proportionality constant between the instantaneous current and the change of voltage over time.

4.1.5.1 Capacitive Reactance

Capacitive behaviour is in many ways parallel to pure resistive behaviour with a few important differences, as implied by (4.15). Capacitive current is proportional to the rate of change of its voltage. Therefore, if there is no voltage change, i.e. the capacitive voltage frequency is zero, then the term $\mathrm{d}v/\mathrm{d}t = 0$ in (4.15) becomes zero, with the direct consequence that the capacitive current must be zero by definition. We, therefore, conclude that, once the charging process is over, a capacitor does not let DC through, which is as same as saying that capacitive DC resistance is infinite.

An important case that is also of practical use is when the capacitive voltage changes periodically in time by following the sinusoidal function, which is also known as "steady-state analysis". Steady state signals change by the rate of radian frequency (1.8), $\omega = 2\pi/T$, therefore the maximum capacitor voltage V_m changes at the same rate. Mathematically, we define a periodic capacitor voltage as $v_\mathrm{C} = V_\mathrm{m}\cos\omega t$ and we are able to rewrite (4.15) as

$$\begin{aligned} i_\mathrm{C} &= C\frac{\mathrm{d}}{\mathrm{d}t}v_\mathrm{C} = C\,\frac{\mathrm{d}}{\mathrm{d}t}V_\mathrm{m}\cos\omega t = -\omega C\,V_\mathrm{m}\sin\omega t \\ &= \omega C\,V_\mathrm{m}\cos\left(\omega t + \frac{\pi}{2}\right) = \omega C\,\mathrm{j}\,v_\mathrm{C}, \end{aligned} \tag{4.16}$$

$$\therefore$$

$$Z_\mathrm{C} \equiv \frac{v_\mathrm{C}}{i_\mathrm{C}} = \frac{1}{\mathrm{j}\omega C}, \tag{4.17}$$

where the capital letters denote steady state variables. We have defined the steady-state capacitor impedance Z_C by forcing (4.15) into the shape of Ohm's law, which has revealed the expression $1/\mathrm{j}\omega C$ associated with capacitance that physically represents the capacitor resistance at frequency ω. Equation (4.17) shows the value of capacitive reactance and also shows the phase relationship between the capacitor voltage and current, through the complex variable j, which accounts for the 90° phase shift in (4.16) between the voltage and current.

The initial assumption of infinite resistance associated with the insulating layer is an abstraction that cannot be achieved in reality, of course. Good insulators, however, do block DC reasonably well, which makes the capacitor leakage current negligible. As a result, a well charged capacitor can hold its charge over very long periods of time where the leakage current is close to zero.

Example 4.1. To illustrate how capacitor impedance changes with frequency, calculate the impedance of a $C = 159\,\mathrm{pF}$ capacitor using (4.17) at the following frequencies: 1 Hz, 100 Hz, 10 kHz, 1 MHz, 100 MHz, and 1 GHz.

Table 4.1 Capacitor impedance for various frequencies, $C = 159$ pF

Frequency	Reactance
1 Hz	1 GΩ
100 Hz	10 MΩ
10 kHz	100 kΩ
1 MHz	1 kΩ
100 MHz	10 Ω
1 GHz	1 Ω

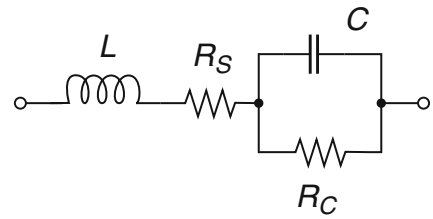


Fig. 4.13 Equivalent electrical circuit model of a high-frequency plate capacitor

Solution 4.1. Numerical results after substitution of the required frequency values in (4.17) are tabulated as shown in Table 4.1.

Because real dielectric materials are lossy (which is to say that there is a small current flow under all conditions)—in other words, a real capacitor insulator is leaky—the finite resistance of the insulating material is calculated in the same way as any other resistive material using (4.10). This parasitic resistance is perceived as being in parallel with the desired capacitance (effectively, it provides a DC path between the capacitor terminals), hence it is easier to take the inverse of (4.10) and calculate the conductance of the insulating dielectric as

$$G_\mathrm{C} = \frac{1}{R_\mathrm{C}} = \sigma_\mathrm{diel}\,\frac{S}{d}, \tag{4.18}$$

where G_C is the dielectric conductance, R_C dielectric resistance, σ_diel is the conductivity of the dielectric, S is the conductive cross-section, i.e. the surface of the plate, and d is the thickness of the dielectric material, i.e. the equivalent to resistive length l in (4.10). Engineering practice is to quantify the dielectric properties of material used as the insulating layer inside a capacitor by introducing the series loss tangent $\tan\Delta_\mathrm{s}$ as

$$\tan\Delta_\mathrm{s} = \varepsilon\,\frac{\omega}{\sigma_\mathrm{diel}}, \tag{4.19}$$

so that (4.18) is written as

$$G_\mathrm{C} = \sigma_\mathrm{diel}\,\frac{S}{d} = \omega\,\frac{\varepsilon S}{d}\,\frac{1}{\tan\Delta_\mathrm{s}} = \frac{\omega C}{\tan\Delta_\mathrm{s}}. \tag{4.20}$$

The equivalent electrical circuit for a real capacitor includes the desired capacitance C, the parasitic series resistance R_S, the inductance L of lead wires, and the dielectric loss resistance R_C (see Fig. 4.13). The overall impedance of a real capacitor Z_Cr is then calculated as

$$Z_\mathrm{Cr} = (R_\mathrm{S} + \mathrm{j}\omega L) + \frac{1}{G_\mathrm{C} + \mathrm{j}\omega C}. \tag{4.21}$$

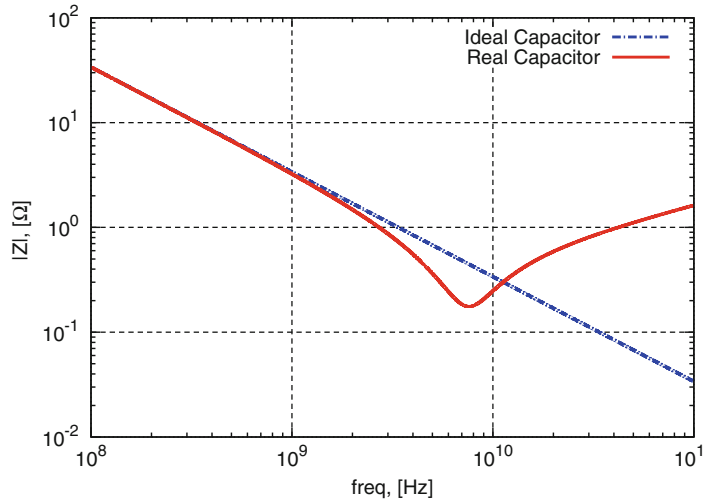


Fig. 4.14 Absolute impedance value against frequency of a typical capacitor. Ideal and realistic models are compared. The real capacitor model in Fig. 4.13 used the following ideal component values: $C = 47$ pF, $L = 771e-9/\sqrt{f}$, and $R = 4.8e-6\sqrt{f}$

A simple numerical analysis of the realistic capacitor model in Fig. 4.13, whose capacitance was designed to be, for example, $C = 47$ pF, shows that there are three distinctly different frequency regions where the capacitor's behaviour is drastically different (see Fig. 4.14). For purposes of comparison, an impedance amplitude of an ideal capacitor is shown in the same plot. Although the numbers used to create Fig. 4.14 are specific only for this particular numerical example, the curve shape is similar for other examples.

- Below 1 GHz: In this frequency region, the capacitance closely follows the one for the ideal capacitor, which is to say that the parasitic components are negligible.
- 1 to 10 GHz: The resonant behaviour of the real capacitor is clearly visible with a self-resonant frequency at approximately 8 GHz.
- Above 10 GHz: In this frequency region, the inductive parasitics are dominant, turning this capacitor into an inductor, and the desired capacitive function is completely suppressed.

This example illustrates the complexity of the behaviour associated with real capacitive components in the frequency domain due to their internal parasitics, which directly limits their useful frequency range of operation.

4.1.5.2 AC Steady State of a Circuit with Capacitor

A parallel connection of a single-tone generator and purely capacitive load is shown in Fig. 4.15 (left). In this section, we take a closer look at the important characteristics of this class of circuits in terms of the AC voltage–current relationship.

Although it may look trivial, (4.15) is very important for understanding the voltage–current relationship in Fig. 4.15, because it states that the AC through a capacitor depends on the rate of voltage change, i.e. on its first derivative in respect to time. At the beginning of the voltage waveform in Fig. 4.15 (right), i.e. at $t = 0$, the capacitor is discharged and, according to (4.15), the voltage v_C must also be zero. However, at that moment, the rate of voltage change is highest, which means that, according to (4.15), the corresponding current i_C must be at its maximum value. Moving along the voltage waveform, e.g. at the point $t = \pi/2$, the value of v_C is at its maximum but the rate of change is zero. The corresponding current i_C also must be zero, which is exactly what the plot shows. Once this analysis is done for all points in time, we reach the conclusion that the current waveform

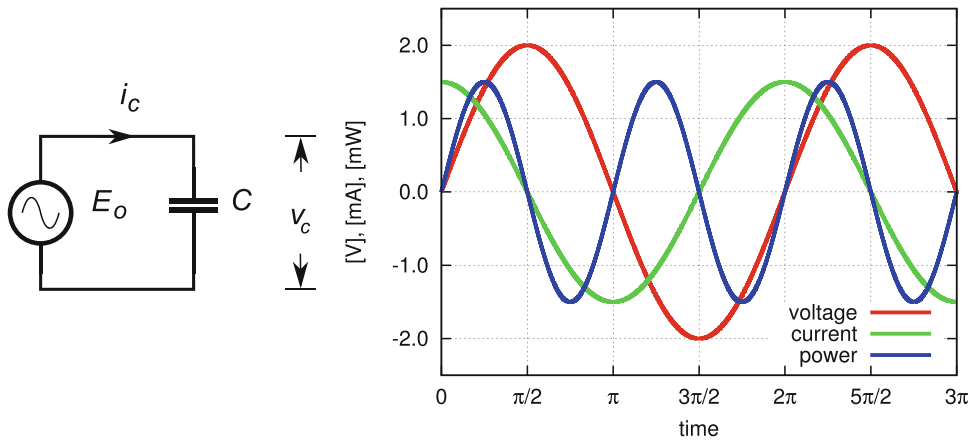


Fig. 4.15 A circuit with purely capacitive load and AC voltage generator (*left*) and the corresponding voltage–current–power time domain plot (*right*). In this particular example, $E_0(t) = 2\sin(\omega t) = v_C$ V and $i_R = 1.5\sin(\omega t)$ mA

also takes a sinusoidal shape, however it is "out of phase", i.e. it is 90° ahead of the voltage waveform. This "ahead" wording is a source of confusion for a number of students who ask the valid question "How the current could possibly know its value a quarter of the cycle ahead in time?" It does not know: at any given point in time the current value is proportional to the instantaneous rate of change of the corresponding voltage value, and the voltage value is a quarter of the cycle ahead in time. Indeed, this voltage–current relationship is conveniently described by saying that "the current leads the voltage by 90°".

For the specific numerical example in Fig. 4.15, the important points to observe are:

- Because the voltage and current through a capacitor are out of phase, the power changes from its most positive value through zero to its most negative value, and back. Its waveform also follows a sinusoidal shape. For half of its cycle, the power flows out of the generator into the capacitor where it is stored in the form of an electrical field. For the other half, the power flows out of the capacitor back into the voltage generator.
- The power cycle is half the signal cycle.
- The average power is zero, i.e. it keeps bouncing back and forth between the source and the capacitor. In short, in an ideal capacitor there is no thermal power dissipation.

4.1.5.3 Transient Capacitive Current

Under certain conditions, i.e. the capacitor C is not charged at the beginning and the abrupt voltage change, a.k.a. step function, is introduced by the pulse function (see Fig. 4.16), the behaviour of the capacitor in the time domain is not a steady state. It is derived as follows.

At any given moment in time, the source voltage E_0 is split between the voltages across the resistor v_R and across the capacitor v_C. At the beginning, the capacitor is not charged, which is to say that $v_C = 0$ (both plates are at the same potential). At the moment $t = t_0$, when the voltage E_0 becomes abruptly high, the source voltage is distributed only over the resistor (there is no voltage drop across the capacitor) and the current abruptly jumps to $i(0) = E_0/R$. However, as soon as the current starts to flow, the charges are "tending to" the capacitor plate, which is the equivalent of saying that the capacitor's voltage v_C starts to rise at a very high rate (limited only by the initial current $i(0)$). As a consequence, less voltage is left across the resistor, which further lowers the current, while the rate

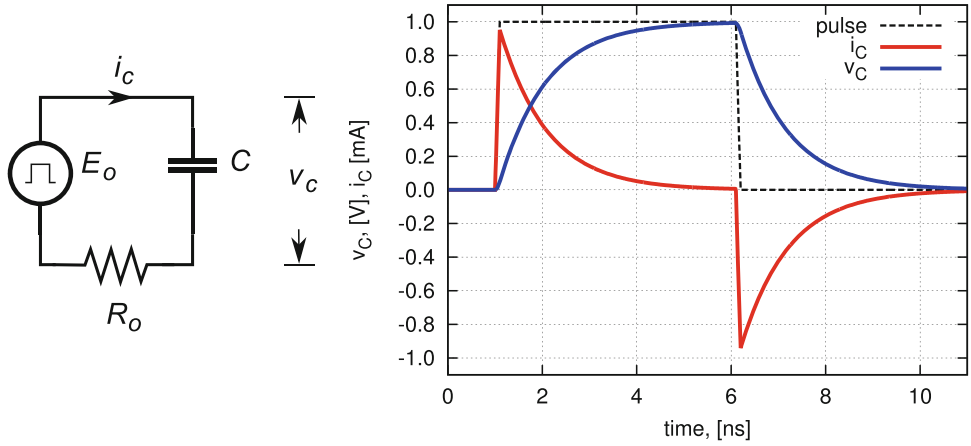


Fig. 4.16 A circuit with a capacitor C, limiting resistor R_0, and pulse voltage generator E_0 (*left*) and the corresponding voltage–current time domain plot (*right*). In this particular example, $E_0(t) = 1\,\mathrm{V} \cdot \texttt{pulse}(10\,\mathrm{ns})$. and $R_0 = 1\,\mathrm{k\Omega}$

of current change is constantly reduced. Theoretically, the process (which is very common in nature) keeps going forever. It is known as "exponential decay" or "natural growth".

Mathematically, the exponential decay process is modelled by a first-order differential equation. Keep in mind that the two constant voltage pulse levels by themselves are considered DC voltages, i.e. the same pulse shape could have been created with a DC source and a two-way switch. Hence, we write the Kirchhoff's current law (KCL) equation and take its derivative as

$$E_0 = v_\mathrm{R}(t) + v_\mathrm{C}(t) = i(t)R + \frac{q(t)}{C}, \tag{4.22}$$

therefore,

$$0 = R\frac{\mathrm{d}i(t)}{\mathrm{d}t} + \frac{1}{C}\frac{\mathrm{d}q(t)}{\mathrm{d}t}, \tag{4.23}$$

$$0 = \frac{\mathrm{d}i(t)}{\mathrm{d}t} + \frac{1}{RC}\,i(t). \tag{4.24}$$

The solution to the first-order differential equation for $i(t)$, with the initial condition $i(0) = E_0/R$, is

$$i(t) = i_\mathrm{C}(t) = \frac{E_0}{R}\,\mathrm{e}^{-t/\tau_0}, \tag{4.25}$$

where $\tau_0 = RC$ is the "time constant" of the system. After substituting (4.25) into (4.22), we find the voltage across the capacitor as

$$E_0 = i(t)R + v_\mathrm{C}(t) \qquad \therefore \qquad v_\mathrm{C}(t) = E_0\left(1 - \mathrm{e}^{-t/\tau_0}\right). \tag{4.26}$$

Equations (4.25) and (4.26) describe how the voltage and current across a capacitor follow abrupt changes in the DC voltage level across the capacitor terminals. The technical term for this type of change is that it is "transient" and it is, obviously, a very nonlinear process. Points to note are:

- A capacitor is very good at passing fast, abrupt voltage changes while presenting an open circuit for direct current.
- Theoretically, a capacitor never reaches the level of E_0 voltage, it only keeps tending towards it forever. Because of that, a practical decision is usually made that a capacitor is "fully charged" at

about $t = 5\tau_0$, because at that moment the capacitor voltage v_C is at over 99% of the maximum level set by E_0 (which is easily proved by (4.26)).

An important question regarding the charging and discharging process is "Where does the power go?" At the beginning, during the capacitor charging period, when the voltage across the capacitor abruptly jumps from low to high voltage, almost the whole power is dissipated in the resistor. As the transition current lowers due to the increase of the capacitor voltage, the portion of power being stored in the capacitor in form of an electrostatic field increases. When the polarity at the capacitor terminals is reversed (at the falling edge of the pulse), the capacitor serves as the source of energy, which now flows into the resistor where it is dissipated. At the end of the charge–discharge cycle, the total energy initially provided by the voltage source has been dissipated in the resistor. These two phases of the full cycle must contain exactly the same amounts of energy, which must add up to the total available energy, thus we can say that the total charge Q must result in voltage $V = Q/C$. That is to say, the average voltage must have been one half, i.e. $V_{\text{avg}} = Q/2C$. Therefore, the work (and the energy W) that was stored in the capacitor C is calculated as the charge Q times the average voltage V_{avg}, i.e.

$$W = QV_{\text{avg}} = Q\frac{Q}{2C} = \frac{Q^2}{2C} = \frac{V_{\text{avg}}^2 C^2}{2C} = \frac{V_{\text{avg}}^2}{2C}, \tag{4.27}$$

which is a commonly used expression for the amount of energy in a capacitor.

4.1.6 Inductance

Inductors are not often used in low-frequency electronic circuits. However, in wireless RF designs they are absolutely essential components. What is more, the frequency behaviour of an RF inductor arguably influences the final specifications of an RF circuit more than any other component.

Similar to a capacitor, an *inductor* is a two-terminal device that is capable of storing energy. This time the energy is stored in the form of an internal magnetic field. The voltage–current relationship at the inductor's terminals is described as

$$v = L\frac{\mathrm{d}i}{\mathrm{d}t}, \tag{4.28}$$

where voltage v and the change of current $\mathrm{d}i/\mathrm{d}t$ are connected by the proportionality constant L, which is defined as *inductance*. Hence, voltage generated at terminals of an inductor is proportional to *rate of change* of the current flowing through.

Using the same methodology as in Sect. 4.1.5.1, the periodic current $i_L = I_m \cos \omega t$. After expanding (4.28), we show that the inductive reactance Z_L is defined as

$$Z_L \equiv \frac{v_L}{i_L} = \mathrm{j}\omega L. \tag{4.29}$$

As a side note, there is an important version of an inductor known as "RF choke" (RFC) of which the main characteristic is that it is made intentionally large. Consequently, at higher frequencies it serves as an AC-blocking device while allowing DC to flow. It is used extensively in RF circuits for providing DC biasing to active devices without interfering with the AC signal. To conclude, we note that an RFC has the same relationship to AC signals as a capacitor to DC signals, and vice versa.

Example 4.2. To illustrate how inductor impedance changes with frequency, calculate the impedance of a, $L = 159\,\text{nH}$ inductor at the following frequencies: 100 Hz, 10 kHz, 1 MHz, 100 MHz, and 10 GHz.

Table 4.2 Inductor impedance for various frequencies, $L = 159\,\text{nH}$

Frequency		Reactance	
100	Hz	100	μΩ
10	kHz	10	mΩ
1	MHz	1	Ω
100	MHz	100	Ω
10	GHz	1	kΩ

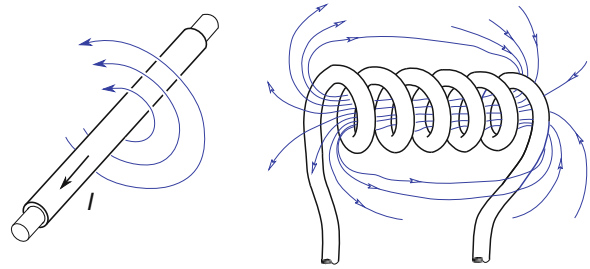


Fig. 4.17 Magnetic field of a straight wire (*left*) and a cylindrical air-core inductor (*right*)

Solution 4.2. The numerical results after substitution of the required frequency values in (4.29) are tabulated as shown in Table 4.2.

Typically, inductors are built by winding low-resistance wire around a cylindrical body (see Fig. 4.17). An approximate formula commonly used for short cylindrical air-core inductors is

$$L = \frac{\pi r^2 \mu_0 N^2}{l}, \tag{4.30}$$

where L is the desired inductance, r is the coil radius, l is the coil length, N is the number of turns, and μ_0 is permeability in a vacuum.

Example 4.3. Estimate the inductance L of a coil formed by $N = 50$ turns of a copper wire with radius $a = 80\,\mu\text{m}$, a radius of air core $r = 2\,\text{mm}$, and length of the coil $l = 10\,\text{mm}$. Note that the distance between two adjacent turns is $d = l/N = 100\,\mu\text{m}$.

Solution 4.3. A commonly used formula for estimating the inductance of an air-core solenoid for $r \ll l$ is

$$L = \frac{\pi r^2 \mu_0 N^2}{l} \approx 3.948\,\mu\text{H},$$

which is a close estimate of the coil inductance.

As simple as it sounds, inductor design is still considered as much an art as engineering. The problem is that, according to Faraday's law of EM induction, any current-carrying wire creates a magnetic field around it (see Fig. 4.17 (left)). Hence, it can be stated that every real wire also behaves as an inductor with finite internal resistance. Moreover, a current-carrying wire could always be rightly considered as one "plate" of a capacitor where the other "plate" could be the adjacent wire or any of the surrounding objects that happen to be at a different potential. Now, one could ask the following question: if a single wire exhibits behaviour typical of a resistor, a capacitor and an inductor, all at the same time, how do we use it so that (at least within some finite frequency–voltage–current range) it resembles a device known as an "ideal inductor"? That is, how close can we get to ideal inductor behaviour in reality?

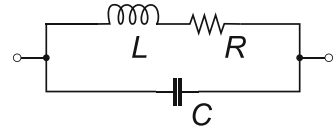


Fig. 4.18 Equivalent electrical circuit model for a high-frequency inductor

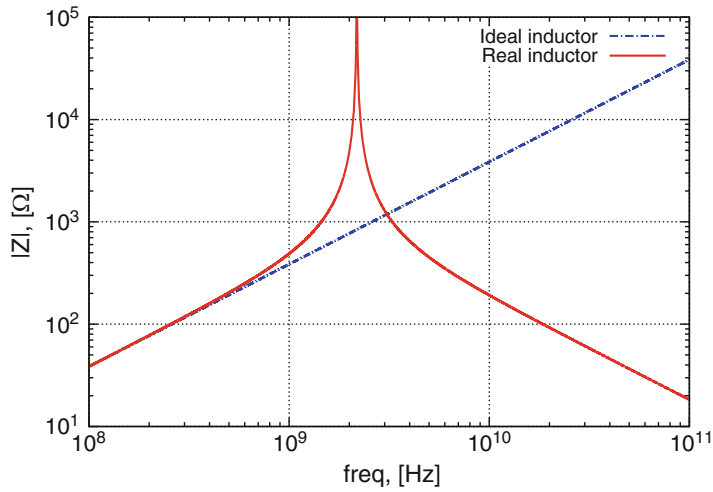


Fig. 4.19 Absolute impedance value against frequency of a typical wire-wound inductor. The inductor model in Fig. 4.18 used the following ideal component values: $R = 0.034\,\Omega$, $L = 61.4\,\text{nH}$, and $C = 0.087\,\text{pF}$

First, let us deal with the "resembling inductor behaviour" part of the question. The magnetic field around a straight wire, Fig. 4.17 (left), is relatively weak. Bending the wire into a circular shape forces all sections of the magnetic field, which are spread along the wire's length to fold over and "come close" to each other. Effectively, the magnetic energy density is increased within the encircled space, which is to say that the wire behaves more like an inductor. By increasing the number of turns and creating a cylindrical solenoid, Fig. 4.17 (right), the inductive behaviour is further emphasized. The magnetic field inside the solenoid is now much stronger (and more uniform) than the straight wire, implying that a considerable amount of energy is "stored" in the inductor's magnetic field.

Second, it should be noted that the internal resistance of the wire is always present with parasitic capacitances between the neighbouring turns and between the inductor and the surrounding environment, which means that it is possible to achieve a close approximation of ideal inductor behaviour, but never to become an ideal inductor. Therefore, a real inductor resembles the behaviour of a relatively complex RLC network with dominant inductive behaviour only within a limited range of operation, quickly losing its inductive property outside the optimal range. One of the limiting factors is a phenomenon known as "self-resonance", which is explained in more detail in the following paragraphs.

One of the possible equivalent circuits of a high-frequency inductor (see Fig. 4.18) includes the desired inductance L, the serial resistance of the wire R_s, and the parasitic shunt capacitance between the inductor's terminals C_s.

The typical realistic frequency behaviour of an inductor (Fig. 4.19) clearly shows three distinct regions. For purposes of comparison, an impedance amplitude of an ideal inductor is shown in the same plot. Although the numbers shown in Fig. 4.19 are specific only to this particular numerical example, the curve shape is similar for other values:

- Below 1 GHz: In this frequency region, the intended function, i.e. the inductance, closely follows that for the ideal inductor, which is to say that parasitic components are negligible.

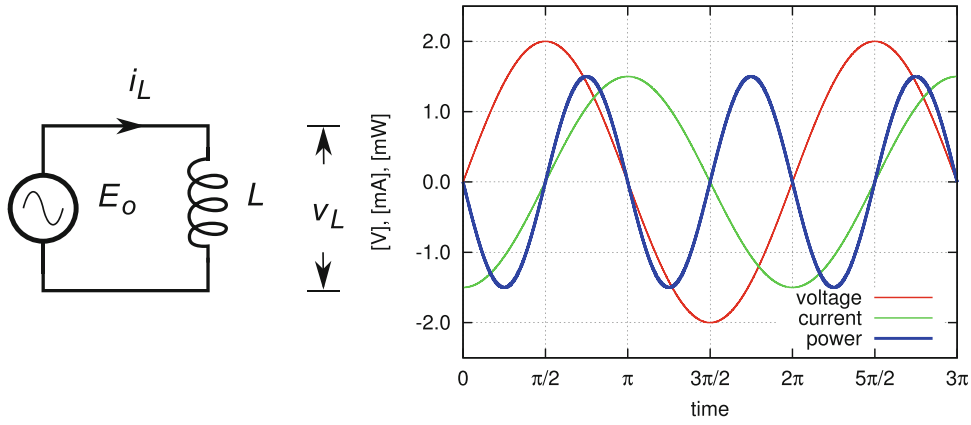


Fig. 4.20 A circuit with purely inductive load and AC voltage generator (*left*) and the corresponding voltage–current–power time domain plot (*right*). In this particular example, $E_0(t) = 2\sin(\omega t) = v_C\,\mathrm{V}$ and $i_R = 1.5\sin(\omega t)\,\mathrm{mA}$

- 1 to 10 GHz: A sharp, resonant behaviour of the real inductor is clearly visible with *self-resonant frequency* at approximately 2 GHz.
- Above 10 GHz: In this frequency region, capacitive parasitics are dominant (the desired function is almost completely suppressed), turning this inductor into a capacitor.

This example illustrates the complexity of behaviour associated with real components in the frequency domain due to their internal parasitics, which directly limits range of frequencies over which they are useful.

4.1.6.1 AC Steady State of a Circuit with Inductor

The parallel connection of a single-tone generator and purely inductive load is shown in Fig. 4.20 (left). In many ways, inductance (which is caused by magnetic phenomena) is equivalent to capacitance (which is caused by electrostatic phenomena). This means that they are like mirror images of each other, with some roles being swapped. In the case of inductance, the changing current forces its magnetic field to change, which in return forces the induced voltage to change.

Let us take a closer look at the voltage–current relationship in Fig. 4.20 (right). At the beginning of the voltage waveform cycle, i.e. at $t = 0$, it is at its maximum rate of change, which means that the rate of current change is highest, however it is negative. It is opposing the large rate of voltage change. On the other hand, when the current is at its maximum, with its first derivative equal to zero, the inductor voltage must be zero as well. This point in time is followed by a reduction in the current amplitude, which causes negative voltage, and the cycle keeps repeating. Once this analysis is done for all points in time, we reach the conclusion that the current waveform also takes a sinusoidal shape, however it is out of phase, 90° *behind* (i.e. it "lags") the voltage waveform. This voltage–current relationship is conveniently described by saying that "the current lags the voltage by 90°".

For the specific numerical example in Fig. 4.20, the important points to observe are:

- Because the voltage and current through an inductor are out of phase, the power changes from its most negative value through zero to its most positive value, and back. Its waveform also follows a sinusoidal shape. That means for half of its cycle the power flows out of the generator into the inductor where it is stored in the form of a magnetic field. For the other half, the power flows out of the inductor back into the voltage generator.
- The power cycle is half the signal cycle.

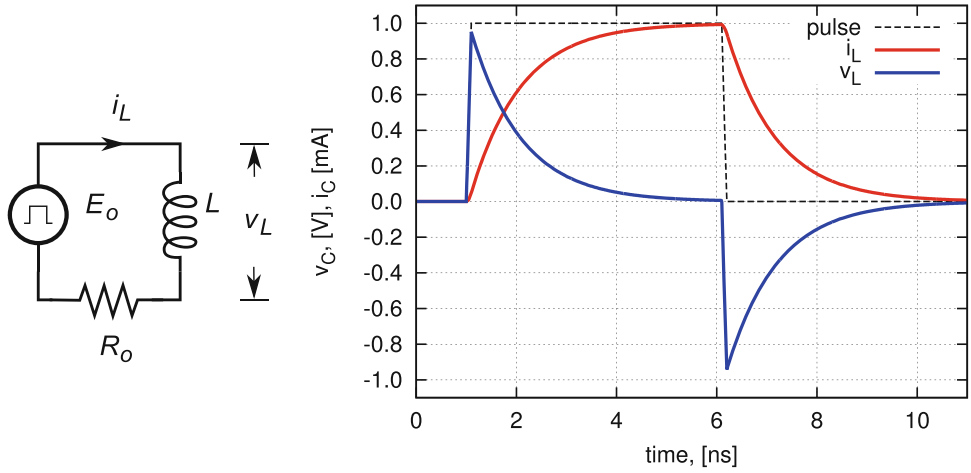


Fig. 4.21 A circuit with an inductor, limiting resistor R_0, and pulse voltage generator E_0 (*left*) and its corresponding voltage–current time domain plot (*right*). In this particular example, $E_0(t) = 1\,\mathrm{V} \cdot \mathtt{pulse}(10\,\mathrm{ns})$ and $R_0 = 1\,\mathrm{k\Omega}$

- The average power is zero, i.e. it keeps bouncing back and forth between the source and the inductor. In short, there is no thermal power dissipation in an ideal inductor.

4.1.6.2 Transient Inductive Current

For the most part, the story of the inductor–resistor network is similar to that of the capacitor–resistor network in Sect. 4.1.5.2, except that this time we start by looking at relationship (4.28). Again, at the time $t = 0$ the current is at its maximum negative value and its first derivative is zero. Hence, the induced voltage, according to (4.28), takes zero value. Following through the rest of the current waveform, we reach similar conclusions to those in Sect. 4.1.5.2 and we confirm the current–voltage waveform relationship, Fig. 4.20 (right). In conclusion, the current waveform lags the voltage waveform by a quarter of the cycle, i.e.

$$v_\mathrm{R} = iR, \tag{4.31}$$

$$v_\mathrm{L} = L\frac{\mathrm{d}i}{\mathrm{d}L}, \tag{4.32}$$

$$Ri + L\frac{\mathrm{d}i}{\mathrm{d}L} = V, \tag{4.33}$$

$$\therefore$$

$$i = \frac{V}{R}\left(1 - e^{-(R/L)\,t}\right), \tag{4.34}$$

where $\tau_0 = {}^{R}/_{L}$ is the timing constant of the LR circuit and the current follows the natural growth law. This example illustrates the duality of capacitor and inductor devices in terms of their time and frequency domain behaviour (Fig. 4.21).

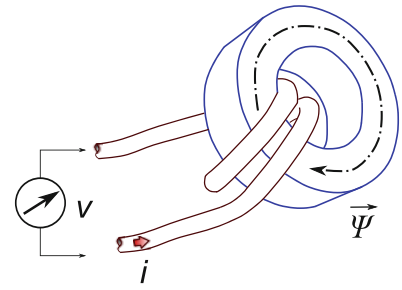


Fig. 4.22 Simple inductor with magnetic toroidal core

4.1.7 Transformer

Our definition of inductance (introduced in Sect. 4.1.6) was, strictly speaking, a definition of "self-inductance". We now generalize the definition by taking a closer look at how inductance works and then by introducing a second inductor in close proximity to the first. The presence of the second inductor creates a structure known as a "transformer".

Let us start with a single inductor connected to a voltage meter (Fig. 4.22), whose air core is replaced by a magnetic toroidal ring. The magnetic core serves as a container for magnetic flux Ψ and, for the moment, we assume the existence of a current I in the wire. To a first approximation, the magnetic flux Ψ is proportional to the current i and the number of inductor turns N as

$$\Psi = KNi, \tag{4.35}$$

where K is the proportionality constant that depends on the geometry and material properties of the toroidal ring. Magnetic flux is a vector variable, whose direction is determined using the right hand rule. In accordance with Faraday's law, a varying magnetic flux induces voltage in N turns of the inductor, after substituting (4.35), as

$$\frac{\mathrm{d}}{\mathrm{d}t}(N\Psi) = v \quad \therefore \quad v = N\frac{\mathrm{d}}{\mathrm{d}t}\Psi = N\frac{\mathrm{d}}{\mathrm{d}t}(KNi) = KN^2\frac{\mathrm{d}i}{\mathrm{d}t} \quad \therefore \quad v = L\frac{\mathrm{d}i}{\mathrm{d}t}, \tag{4.36}$$

where inductance is defined as

$$L \equiv KN^2, \tag{4.37}$$

which simply states that the inductance is a function of the square of the number of inductive turns and properties of the core. Alternatively, by using complex notation for the case of periodic magnetic flux, (4.36) is rewritten as

$$V = \mathrm{j}\omega LI, \tag{4.38}$$

which is the form commonly used in steady-state-circuit analysis. We recognize that magnetic flux is produced by a current, where the flux is proportional to the current, and voltage is produced by a time-varying magnetic flux, where the voltage is proportional to the rate of change of the magnetic flux. Therefore, voltage is proportional to the rate of change of the current.

Adding a second coil that shares the magnetic flux of the first coil (see Fig. 4.23) extends the single-coil argument and the structure is known as a "transformer". We note that a single coil is a two-terminal device, while a transformer is a four-terminal device. Although the roles of the two inductors are interchangeable, it is common to use the term "primary" when referring to the input side and

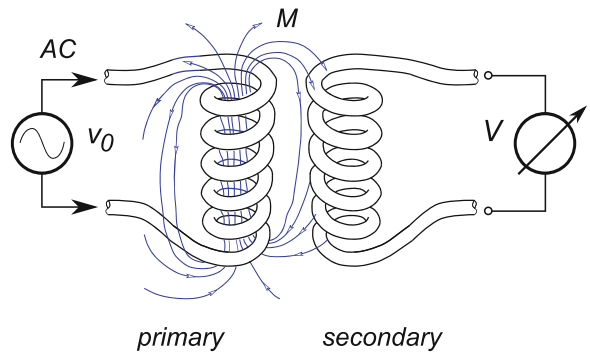


Fig. 4.23 Two coils forming a transformer. For simplicity, an air core is used

"secondary" when referring to the output side of the transformer. Intuitively, we expect that a voltage is induced in both coils if the flux is changed and it is interesting to find out how the two coils interact compared with how much of the flux is actually shared. By inspecting the transformer example in Fig. 4.22, we can deduce that if a second coil was added around the same magnetic coil then, in the ideal case, both coils would experience almost exactly the magnetic flux that is contained in the core. In contrast, if the two coils are separated as in Fig. 4.23 than only part of the flux going through the primary coil crosses to the secondary coil.

A current flowing through the primary coil produces magnetic flux not only in the primary but also in the secondary coil. Therefore, the time-varying flux in the secondary coil produces a voltage at its terminals that is proportional to the rate of change of current in the primary coil. Similar to (4.36), we write an expression for the secondary voltage v_S generated by secondary flux Ψ_S through N_S turns as

$$v_S = N_S \frac{d}{dt}\Psi_S, \tag{4.39}$$

where the secondary flux Ψ_S is a fraction of the primary flux Ψ_P. Hence we can establish their ratio as

$$k = \frac{\Psi_S}{\Psi_P}, \tag{4.40}$$

where k is a coupling coefficient that can take any value between zero and one. We can now connect the secondary voltage with the primary flux as

$$v_S = N_S k \frac{d}{dt}\Psi_P, \tag{4.41}$$

which, after substituting (4.35) for the primary side and (4.37) for both sides, yields

$$\begin{aligned} v_S &= N_S k \frac{d}{dt}(K N_P i_P) = k K N_P N_S \frac{di_P}{dt} = k K \sqrt{\left(\frac{L_P}{K}\right)} \sqrt{\left(\frac{L_S}{K}\right)} \frac{di_P}{dt} \\ &= k \sqrt{L_S L_P} \frac{di_P}{dt} = M \frac{di_P}{dt}, \\ &\therefore \\ V_S &= j\omega M I_P, \end{aligned} \tag{4.42}$$

where,

$$M \equiv k \sqrt{L_S L_P} \tag{4.43}$$

defines "mutual inductance", which is formally equivalent to the definition of inductance of a single coil in (4.36). Although it can be proven mathematically, we intuitively conclude that the mutual inductance M is equal in both directions, i.e. looking from the primary side as well as looking from the secondary side. Equivalently, following the same argument, we conclude that due to the mutual inductance there is additional voltage added in the primary side, i.e.

$$v_0 = L_\mathrm{P}\frac{\mathrm{d}i_\mathrm{P}}{\mathrm{d}t} + M\frac{\mathrm{d}i_\mathrm{S}}{\mathrm{d}t}, \tag{4.44}$$

$$\therefore$$

$$V_0 = \mathrm{j}\omega L_\mathrm{P} + \mathrm{j}\omega M I_\mathrm{S}, \tag{4.45}$$

which simply states that the input side voltage V_0 is the sum of the induced voltage in the primary coil (due to the change of the source current $\mathrm{j}\omega L_\mathrm{P}$) and the induced voltage due to change of the secondary current $\mathrm{j}\omega M I_\mathrm{S}$, where the capitalized notation of voltages and currents indicates complex numbers.

Example 4.4. Determine the voltage induced on the secondary side of a transformer whose primary side is driven by a sinusoidal current with maximum value 1 mA at 10 MHz. Primary inductance is $L_\mathrm{P} = 50\,\mathrm{nH}$, secondary inductance is $L_\mathrm{S} = 100\,\mu\mathrm{H}$, and the coupling factor is $k = 0.5$.

Solution 4.4. A straightforward implementation of (4.43) and (4.42) yields a result for mutual inductance and the induced secondary voltage as

$$M = k\sqrt{L_\mathrm{S}L_\mathrm{P}} = 0.5\sqrt{100\,\mu\mathrm{H}\times 50\,\mathrm{nH}} = 1.118\,\mu\mathrm{H},$$

$$\therefore$$

$$|V_\mathrm{S}| = \omega M I_\mathrm{P} = 2\pi\times 10\,\mathrm{MHz}\times 1.118\,\mu\mathrm{H}\times 1\,\mathrm{mA} = 70.248\,\mathrm{mV}.$$

4.1.7.1 Energy Stored in a Transformer

A pair of coupled inductors have energy stored in the form of a magnetic field that can be found starting from the zero energy initial condition and by looking at how the energy inside the transformer is built up step by step while using the following mathematical formalism. With no load on the secondary (i.e. $i_\mathrm{S} = 0$ or $v_\mathrm{S} i_\mathrm{S} = 0$), we increase the primary current from $i_\mathrm{P} = 0$ to an instantaneous value $i_\mathrm{P} = i_1(t_1)$ over the period of time from zero to $t = t_1$. Therefore, power and energy in the primary coil is found as

$$v_\mathrm{P} i_\mathrm{P} = L_\mathrm{P}\frac{\mathrm{d}i_\mathrm{P}}{\mathrm{d}t}\, i_1, \tag{4.46}$$

$$\therefore$$

$$W_\mathrm{P}(t_1) = \int_0^{t_1} v_\mathrm{P} i_\mathrm{P}\,\mathrm{d}t = \int_0^{t_1} L_\mathrm{P}\, i_\mathrm{P}\,\mathrm{d}i_\mathrm{P} = \frac{1}{2}L_\mathrm{P}\, i_\mathrm{P}^2(t_1), \tag{4.47}$$

which means that the total transformer energy is still stored in the primary. The change of current in the primary during the period of time from zero to t_1, is followed by the induced voltage v_P in the secondary even though the primary current $i_\mathrm{P}(t_1)$ is held constant during the period from t_1 to t_2,

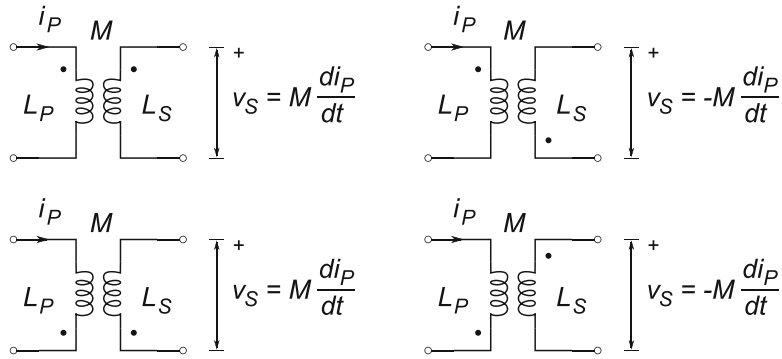


Fig. 4.24 Conventions regarding the direction of the input current and direction of the primary and secondary coil windings relative to the induced secondary voltage polarity

therefore the secondary induced current changed from zero to the instantaneous value of $i_S = i_s(t_2)$, hence the energy in the secondary is

$$W_S(t_2) = \int_{t_1}^{t_2} v_S i_S \, dt = \int_0^{t_2} L_S \, i_S \, di_S = \frac{1}{2} L_S \, i_S^2(t_2). \tag{4.48}$$

However, we should not forget that during the change of the secondary current there is an additional induced current in the primary, which contains energy $W_{Pi}(t_2)$ as

$$\begin{aligned} W_{Pi}(t_2) &= \int_{t_1}^{t_2} v_P i_P(t_1) \, dt = \int_{t_1}^{t_2} M \frac{di_S}{dt} \, i_P(t_1) \, dt \\ &= M \, i_P(t_1) \int_0^{t_2} di_S = M \, i_P(t_2) \, i_S(t_2), \end{aligned} \tag{4.49}$$

where $i_P(t_2) = i_P(t_1)$ is the current caused by the primary source that did not change during the period t_1 to t_2 and was already found in (4.47).

Adding all three energy components that are found at the moment in time $t = t_2$, or any other point in time t for that matter, gives the total energy W accumulated in the transformer as

$$W(t) = W_P(t) + W_S(t) + W_{Pi}(t) = \frac{1}{2} L_P \, i_P^2(t) + \frac{1}{2} L_S \, i_S^2(t) \pm M \, i_P(t) \, i_S(t). \tag{4.50}$$

So far, we have silently assumed positive values for the currents and voltages. In fact, the polarity of the primary and secondary voltages and currents depends upon the orientation of the coil windings, which results in four possible combinations (see Fig. 4.24) and two possible polarities of the induced secondary voltage, while we assume that the mutual inductance is always positive.

4.1.7.2 Transformer Loading

The analysis of a loaded transformer (Fig. 4.25) is based on the assumption that the transformer is linear. Strictly speaking that is not correct, because the magnetic core material used in a transformer almost always has a nonlinear magnetic flux–current characteristic. In the general case of a loaded transformer, we derive an expression for the input impedance, with reference to Fig. 4.25, as follows.

Total input impedance is the sum of the resistance and reactance associated with the input inductor L_P

$$Z_1 = R_P + j\omega L_P = \Re(Z_1) + j\Im(Z_1). \tag{4.51}$$

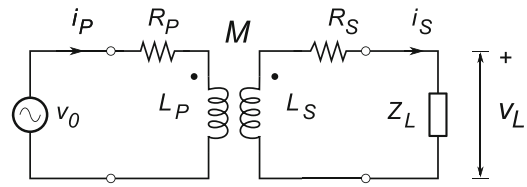


Fig. 4.25 A transformer with the signal source v_0 and loading impedance Z_L. Both primary and secondary resistances are included

The output impedance is the sum of the secondary impedance and the load

$$\begin{aligned} Z_2 &= R_S + j\omega L_S + (R_L + jX_L) = (R_S + R_L) + j(\omega L_S + X_L) \\ &= \Re(Z_2) + j\Im(Z_2), \end{aligned} \tag{4.52}$$

where the type of loading impedance Z_L is yet to be specified. Network equations in complex number notation for the input and output sides are as follows

$$V_0 = Z_1 I_P + j\omega M I_S, \tag{4.53}$$

$$0 = -j\omega M I_P + Z_2 I_S, \tag{4.54}$$

which, after substituting I_S from (4.54) into (4.53), leads to the following expression for the input impedance

$$Z_P = \frac{V_0}{I_1} = Z_1 - \frac{(j\omega M)^2}{Z_2}, \tag{4.55}$$

where, we note, the position of dots and polarities in Fig. 4.25 have no influence on the plus and minus signs. In addition, the input impedance equals the impedance of the primary coil Z_1, as expected, which is reduced by the term due to the presence of the secondary coil. This new term is referred to as "reflected impedance" and it plays a very important role in the behaviour of a transformer. Let us expand (4.55) and take a closer look at how the input impedance is changed.

$$\begin{aligned} Z_P &= [\Re(Z_1) + j\Im(Z_1)] + \frac{(\omega M)^2}{\Re(Z_2) + j\Im(Z_2)} \\ &= [\Re(Z_1) + j\Im(Z_1)] + \frac{(\omega M)^2}{\Re(Z_2) + j\Im(Z_2)} \frac{\Re(Z_2) - j\Im(Z_2)}{\Re(Z_2) - j\Im(Z_2)} \\ &= \left[\Re(Z_1) + \frac{\Re(Z_2)(\omega M)^2}{\Re^2(Z_2) + \Im^2(Z_2)}\right] + j\left[\Im(Z_1) - \frac{\Im(Z_2)(\omega M)^2}{\Re^2(Z_2) + \Im^2(Z_2)}\right] \qquad (4.56) \\ &= \Re(Z_P) + j\Im(Z_P), \qquad (4.57) \end{aligned}$$

which clearly states that the real part of the input impedance has increased due to the reflected impedance which, at the same time, has caused the reactance part to reduce. Equation (4.56) is a general result that applies to a realistic, loosely coupled transformer that is loaded with impedance Z_L. We simplify this result by assuming an ideal transformer that consists of two ideal inductors, that is, inductors whose Q factors and inductances tend to infinity, $L_P, L_S, Q \to \infty$. In addition, an ideal transformer is tightly coupled, hence its coupling index becomes $k = 1$, that is, the mutual inductance becomes $M^2 = L_P L_S$. Also, the two inductances may be expressed in the form of their ratio, hence, using (4.37) we write

$$
\begin{aligned}
L_S &= K N_S^2, \\
L_P &= K N_P^2, \\
&\therefore \\
\frac{L_S}{L_P} &= \frac{N_S^2}{N_P^2} = n^2 \quad \therefore \quad L_S = n^2 L_P,
\end{aligned} \tag{4.58}
$$

which further enables us to rewrite (4.55), after substituting $R_S = R_L = 0$, as

$$
\begin{aligned}
Z_P &= j\omega L_P + \frac{\omega^2 L_P L_S}{j\omega L_S + Z_L} = j\omega L_P + \frac{\omega^2 n^2 L_P^2}{j\omega n^2 L_P + Z_L} \\
&= \frac{j\omega L_P Z_L - \omega^2 n^2 L_P^2 + \omega^2 n^2 L_P^2}{j\omega n^2 L_P + Z_L} \\
&= \frac{j\omega L_P Z_L}{j\omega n^2 L_P + Z_L} \approx \frac{j\omega L_P Z_L}{j\omega n^2 L_P}, \\
&\therefore \\
Z_P &= \frac{Z_L}{n^2},
\end{aligned} \tag{4.59}
$$

where the assumption for the ideal inductor ($j\omega L_P \to \infty$) allowed for approximation $j\omega n^2 L_P + Z_L \approx j\omega n^2 L_P$. The last result is very important because it states that, for ideal transformers, the loading impedance Z_L is perceived at the side of the transformer as another impedance that is equal to R_L/n^2, which is under the control of the designer. This "impedance scaling" property of an ideal transformer is very useful in RF circuit design.

Example 4.5. If the number of turns of the ideal transformer primary coil is $N_P = 100$ and the number of turns of the secondary coil is $N_S = 10,000$, what is the perceived impedance at the input side Z_P if: (a) $Z_L = 20\,\text{k}\Omega$; (b) $Z_L = j\omega 200\,\text{mH}$; and (c) $Z_L = 1/j\omega 100\,\text{pF}$.

Solution 4.5. The turn ratio of this transformer is $n = N_S/N_P = 10,000/100 = 100$. A direct implementation of (4.59) yields the following results:

$$
\begin{aligned}
&\text{(a)}\ Z_P = \frac{Z_L}{n^2} = \frac{20\,\text{k}\Omega}{100^2} = 2\,\Omega \\
&\text{(b)}\ Z_P = \frac{j\omega\,(200\,\text{mH})}{100^2} = j\omega\,(20\,\mu\text{H}) \\
&\text{(c)}\ Z_P = \frac{1}{j\omega\,(100\,\text{pF}) \times 100^2} = \frac{1}{j\omega\,(1\,\mu\text{F})}
\end{aligned}
$$

Let us determine the relationship between the primary and secondary currents for an ideal transformer. From (4.54), we write

$$
\frac{I_S}{I_P} = \frac{j\omega M}{j\omega L_S + Z_L} \approx \frac{j\omega M}{j\omega L_S} = \frac{\sqrt{L_P L_S}}{L_S} = \sqrt{\frac{L_P}{L_S}} = \frac{1}{n} = \frac{N_P}{N_S}, \tag{4.60}
$$

hence, we conclude that

$$N_P I_P = N_S I_S. \tag{4.61}$$

Similarly, knowing that the power taken by the primary side must equal the power on the secondary side, we write

$$V_S = I_S Z_L; \qquad V_P = I_P Z_P = I_P \frac{Z_L}{n^2},$$

$$\therefore$$

$$\frac{V_S}{V_P} = n^2 \frac{I_S}{I_P} = n^2 \frac{1}{n} = \frac{N_S}{N_P},$$

$$\therefore$$

$$N_P V_S = N_S V_P. \tag{4.62}$$

This relationship for the primary and secondary voltages can be combined with (4.61) and we conclude that

$$V_P I_P = V_S I_S. \tag{4.63}$$

Equations (4.59), (4.61), and (4.63) are commonly used relations for an ideal, close-coupled transformer model.

From the perspective of RF circuit applications, the single-tuned transformer and the double-tuned transformer are important. Let us take a look at these two important cases.

1. *Single-Tuned Transformer:* in the case when the loading impedance is due to a pure capacitor, $Z_L = 1/\mathrm{j}\omega C_L$, and when the resonant frequency of the secondary circuit is set to

$$\omega_0 = \frac{1}{\sqrt{L_S C_L}} \tag{4.64}$$

then the output impedance becomes real, i.e. $X_S = 0$ and for a loose-coupled transformer (4.55) becomes

$$Z_P = Z_1 - \frac{(\mathrm{j}\omega_0 M)^2}{R_2} \approx \frac{(\omega_0 M)^2}{R_2} \tag{4.65}$$

because Z_1 is much smaller. This can be written symmetrically for the secondary impedance as

$$Z_S = R_2 - \frac{(\mathrm{j}\omega_0 M)^2}{Z_1} \approx \frac{(\omega_0 M)^2}{Z_1}. \tag{4.66}$$

2. *Double-Tuned Transformer:* This is a more complicated and interesting case, where both primary and secondary sides are tuned to the same frequency (see Fig. 4.26). We find the relationship between the voltage appearing across the secondary resonator v_L and the source signal v_0. By inspection of the circuit network in Fig. 4.26, we write

$$V_0 = \left[R_P + \mathrm{j}\left(\omega L_P - \frac{1}{\omega C_P}\right)\right] I_P + \mathrm{j}\omega M I_S, \tag{4.67}$$

$$0 = \mathrm{j}\omega M I_P + \left[R_S + \mathrm{j}\left(\omega L_S - \frac{1}{\omega C_S}\right)\right] I_S, \tag{4.68}$$

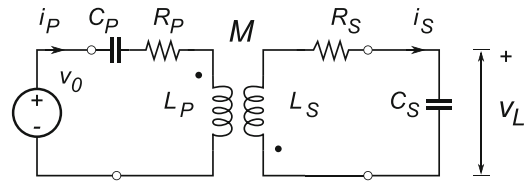


Fig. 4.26 A double-tuned transformer circuit with both primary and secondary sides tuned to the same frequency

which is straightforward to solve for current I_S as

$$I_S = -\frac{j\omega M V_0}{\left[R_P + j\left(\omega L_P - \frac{1}{\omega C_P}\right)\right]\left[R_S + j\left(\omega L_S - \frac{1}{\omega C_S}\right)\right] + (\omega M)^2}. \tag{4.69}$$

The secondary current is also found as

$$I_S = \frac{V_L}{\frac{1}{j\omega C_S}}, \tag{4.70}$$

therefore, the transformer voltage gain becomes

$$A_V = \frac{V_L}{V_0} = -\frac{\frac{j\omega M}{j\omega C_S}}{\left[R_P + j\left(\omega L_P - \frac{1}{\omega C_P}\right)\right]\left[R_S + j\left(\omega L_S - \frac{1}{\omega C_S}\right)\right] + (\omega M)^2}, \tag{4.71}$$

We keep in mind that both sides of the transformer are tuned to the same resonant frequency and, assuming $Q \geq 10$,

$$\omega_0 = \frac{1}{\sqrt{L_S C_L}} = \frac{1}{\sqrt{L_P C_P}}, \tag{4.72}$$

$$Q_P = \frac{\omega_0 L_P}{R_P} = \frac{1}{\omega_0 C_P R_P} = \frac{1}{R_P}\sqrt{\frac{L_P}{C_P}}, \tag{4.73}$$

$$Q_S = \frac{\omega_0 L_S}{R_S} = \frac{1}{\omega_0 C_S R_S} = \frac{1}{R_S}\sqrt{\frac{L_S}{C_S}}. \tag{4.74}$$

In addition, we introduce the useful substitution[2]

$$\delta = \frac{\omega}{\omega_0} - 1 \quad \therefore \quad \text{because } (\omega_0 \approx \omega) \quad \therefore \quad (\delta \ll 1). \tag{4.75}$$

[2] $\omega_0 \delta = BW/2$, which is used again shortly.

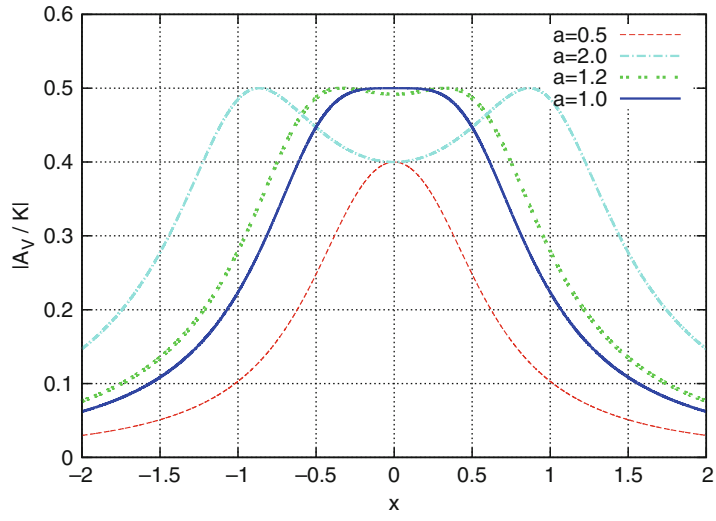


Fig. 4.27 Plot of $|A_V/K| = \sqrt{\Re^2(A_V/K). + \Im^2(A_V/K)}$, as a function of variable x and parameter a

Substitution of (4.75) helps to simplify the following expression

$$\begin{aligned} R + \mathrm{j}\left(\omega L - \frac{1}{\omega C}\right) &= R\left[1 + \mathrm{j}\frac{\omega L}{R}\left(1 - \frac{1}{\omega^2 LC}\right)\right] \\ &= R\left[1 + \mathrm{j}Q\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)\right] = R\left[1 + \mathrm{j}Q\left(\delta + 1 - \frac{1}{\delta + 1}\right)\right] \\ &= R\left[1 + \mathrm{j}Q\,\frac{\delta^2 + 2\delta}{\delta + 1}\right] \\ &\approx R(1 + \mathrm{j}2\delta Q) \end{aligned} \tag{4.76}$$

because $\delta^2 \approx 0$ and $\delta + 1 \approx 1$. For simplicity, we assume $Q_S = Q_P = Q$, which, after applying substitutions to (4.71), results in

$$\begin{aligned} A_V = \frac{V_L}{V_0} &= -\frac{\frac{\omega_0 M}{R_P R_S}\,\frac{1}{\omega_0 C_S}}{(1 + \mathrm{j}2\delta Q_P)(1 + \mathrm{j}2\delta Q_S) + \frac{(\omega_0 M)^2}{R_P R_S}} \\ &= -\frac{\frac{\omega_0 M}{\sqrt{R_P R_S}}\,\frac{1}{\omega_0 C_S \sqrt{R_P R_S}}}{1 + \mathrm{j}4Q\delta - 4\delta^2 Q^2 + \frac{(\omega_0 M)^2}{R_P R_S}} \\ &= -\frac{aK}{1 + \mathrm{j}4x - 4x^2 + a^2}, \end{aligned} \tag{4.77}$$

where we introduced substitutions

$$a = \frac{\omega_0 M}{\sqrt{R_P R_S}}; \quad K = \frac{1}{\omega_0 C_S \sqrt{R_P R_S}}; \quad x = Q\delta,$$

so that (4.77) can be simplified in its shape. A normalized plot of $|A_V/K|$ for a double-tuned transformer is shown in Fig. 4.27 as a function of variable x and parameter a.

If the two resonators have equal Q factor and the parameter $a = 1$ (also known as "critical coupling"), then the response has the maximum possible peak. By further increasing parameter a (i.e. over-coupling), two separate peaks start showing and the bandwidth is increased, which makes double-tuned transformers useful in AM radio receivers, because the bandwidth allows reception of sidebands as well. By differentiating (4.77) and finding the extreme points of the function, it is not difficult to derive analytical expressions for the locations of the two peaks. Under conditions of resonance, i.e. $\delta = 0$, (4.77) becomes

$$A_{\mathrm{V}}(\omega_0) = -\frac{aK}{1+a^2}, \tag{4.78}$$

therefore, the bandwidth (for the critical coupling $a = 1$) is found when the ratio of (4.77) and (4.78) is $1/\sqrt{2}$, as

$$\begin{aligned}
\frac{A_{\mathrm{V}}}{A_{\mathrm{V}}(\omega_0)} &= \frac{1+a^2}{1+\mathrm{j}4x-4x^2+a^2} \\
&= \frac{1+1}{1+\mathrm{j}4x-4x^2+1} = \frac{1}{1-2x^2+\mathrm{j}2x}, \\
\left|\frac{A_{\mathrm{V}}}{A_{\mathrm{V}}(\omega_0)}\right| &= \frac{1}{\sqrt{1+4x^4}} = \frac{1}{\sqrt{2}}, \\
4x^4 = 1 \quad \therefore \quad 4Q^4\delta^4 = 1 \quad \therefore \quad 2Q^2\delta^2 &= 1.
\end{aligned} \tag{4.79}$$

Because we already know that $BW = 2\delta\omega_0$, substitution of (4.79) yields

$$BW = 2\delta\omega_0 = 2\frac{1}{Q\sqrt{2}}\omega_0 = \sqrt{2}\frac{\omega_0}{Q}, \tag{4.80}$$

which shows a bandwidth that is a factor of $\sqrt{2}$ wider than a single-tuned LC resonator.

4.1.8 Memristance

For a long time, traditional engineering network theory was based on the three basic elements of the RLC. The capacitive element was discovered first, around 1745, by E. G. von Kleist who was experimenting with storing electric charges. He was followed by P. van Musschenbroek, who further refined the capacitive device by inventing the Leyden jar. Credit for the invention of the resistor goes to G. Ohm in 1827. M. Faraday and J. Henry are credited for inventing the inductor in 1831. For a long time, it seemed that those three elements were all there is in network theory.

However, in 1971, L. Chua hypothesized that there must be a fourth element. His symmetrical argument was based on the following reasoning:

1. The first fundamental physical property of matter is charge q, which is used to define electric current as

$$i = \frac{\mathrm{d}q}{\mathrm{d}t}. \tag{4.81}$$

2. Charged objects are associated with an EM field. The concept of magnetic flux ϕ is used to quantify the corresponding forces among charged objects, by introducing a potential v (voltage is just the difference between two potentials) as

$$v = \frac{\mathrm{d}\phi}{\mathrm{d}t}. \tag{4.82}$$

3. With those four variables (charge q, magnetic flux ϕ, current i and potential v) in place, it is possible to define, capacitance C as

$$C = \frac{\mathrm{d}q}{\mathrm{d}v} \tag{4.83}$$

resistance R as

$$R = \frac{\mathrm{d}v}{\mathrm{d}i} \tag{4.84}$$

and inductance L as

$$L = \frac{\mathrm{d}\phi}{\mathrm{d}i}. \tag{4.85}$$

That is where it all stopped, until Chua observed that it is possible to define one more relationship:

$$M = \frac{\mathrm{d}\phi}{\mathrm{d}q}, \tag{4.86}$$

which must represent a new, fourth, fundamental element because it cannot be derived by using any combination of the other three elements. He named it a "memristor" M and concluded that it must have a property of "remembering" its state after it was turned off. Unlike the other three, a memristor is a dynamic, nonlinear element. Even though scientists observed many phenomena that satisfy (4.86), it took more than 40 years before the proposed element was confirmed experimentally. In the meantime, Chua and his colleagues further developed the theory by discovering more elements with memristive properties.

The significance of the fourth element is still to be seen. A large number of research groups are racing to develop more practical devices and further refine the possible applications. Probably the most fascinating implication of the initial experiments involving memristive elements is that the learning mechanism used by single-cell living beings, amebas for example, appears to be similar to circuit models based on memristive elements. The following questions are yet to be answered: Are we on the verge of being able to design machines capable of learning, in a similar manner to human beings? Is the memristor the missing link needed to enable the design of intelligent machines, i.e. real artificial intelligence?

Although, not directly related to the subject of the book, the topic of memristors is introduced in this section for the sake of completeness and to raise awareness, especially because it is still not clear how large an impact this element is going to have on our traditional methods of circuit design.

4.1.9 Voltage Divider

Strictly speaking, a voltage divider in its basic form of two impedances connected in series is not a device. Nevertheless, it is the simplest circuit structure that is widely used in circuit analysis at the same level as the other basic circuit elements. In all its simplicity, like the other basic devices of this section, it has been an indispensable tool in every engineer's tool box. If only two of the three basic devices are

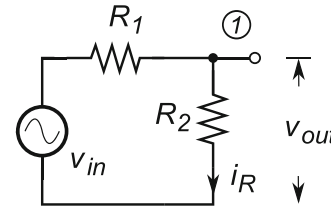


Fig. 4.28 Simple resistive voltage divider

used to build a voltage divider then there are nine possible serial configurations, each having slightly different behaviour. In this section, we review a few of most important configurations. The two main applications of a voltage divider can be categorized loosely as the literal and the conceptual.

When applied literally, a voltage divider is used simply to scale down the amplitude of a voltage (either DC or AC) applied at its terminals. In a more sophisticated version of the literal application, a single voltage divider is designed to impose different scaling factors on different tones in the signal frequency spectrum. In other words, the voltage divider is used to modify the "frequency profile" of the input signal in accordance with a Fourier transform (see Sect. 1.4.7), i.e. it serves as a filter.

When applied conceptually, a voltage divider is used to model the signal transfer process between any two system-level blocks. In Sect. 6.1, we introduce the concept of system partitioning based on voltage dividers in more detail. For the time being, we cannot emphasize enough that a clear understanding of simple voltage divider behaviour is of utmost importance to all electrical engineers. In the following sections, we review three of the most important voltage divider structures.

4.1.9.1 Resistive Voltage Divider

One of the most important simple networks consists of one ideal voltage source and two resistors connected in series. It is assumed that all the elements are ideal and, aside from the conversion of electrical energy into thermal energy, there is no energy loss. Analysis of this network structure is simple and the goal is to find the relationship between the source voltage V_{in} and the output voltage V_{out} at node ① (the connecting point between the two resistors) in Fig. 4.28. In the ideal case, there is no current flow in or out of node ① and the two resistors present serial resistance $R_1 + R_2$ to the voltage source. A straightforward application of Ohm's law (or Kirchhoff's laws, if you prefer) leads to an expression for the voltage gain A_{V} from the source V_{in} to the output voltage V_{out} at node ① as

$$i_{\text{R}} = \frac{V_{\text{in}}}{R_1 + R_2}; \quad v_{\text{out}} = i_{\text{R}} R_2 \Rightarrow v_{\text{out}} = \frac{v_{\text{in}}}{R_1 + R_2} R_2, \tag{4.87}$$

$$\therefore$$

$$A_{\text{V}} = \frac{v_{\text{out}}}{v_{\text{in}}} = \frac{R_2}{R_1 + R_2} = \frac{1}{1 + \frac{R_1}{R_2}}. \tag{4.88}$$

In other words, the ratio of the output voltage V_{out} and source voltage V_{in} is the same as the ratio of their respective resistances. The output voltage V_{out} is measured across R_2, while the source voltage V_{in} is distributed across $(R_1 + R_2)$ (see Fig. 4.28). When (4.88) is applied for a conceptual analysis, it is important to note that perfect signal transfer, i.e.equality of the voltage source V_{in} (i.e. the driver) and voltage at node ① (i.e. across the load), is possible only in two cases: $R_2 \to \infty$ and $R_1 = 0$, then $V_{\text{out}} = V_{\text{in}}$. Hence, (4.88) suggests that if a maximum voltage signal transfer efficiency is to be achieved, the loading resistance R_2 has to be infinite or the source resistance R_1 has to be zero. That is, in reality the input impedance of the loading stage (here symbolized by R_2) must be designed to

be much higher than the output impedance of the driver stage (here symbolized by R_1) so that the R_1/R_2 term in (4.88) is minimized. If that condition is not met, i.e. if the loading resistance is very low relative to the driver resistance, the real drivers are expected to deliver very high currents to keep the amplitude of the output voltage V_{out} away from zero value. This issue is of critical importance to all, not only RF circuit designs.

Example 4.6. Derive an expression for the maximum possible power $P_{\max}$ that can be delivered by a realistic voltage generator, i.e. with non-zero internal resistance, to a resistive load. This case is modelled with the circuit network in Fig. 4.28 where the ideal generator V_{in} and resistance R_1 represent the realistic voltage source, while resistance R_2 represents the load. That is, both the realistic voltage generator and the load are connected between the ground and node ①.

Solution 4.6. Using result (4.87) and the definition for power, it follows from Fig. 4.28 that

$$P \equiv I_1 V_{\text{out}} = I_1^2 R_2 = \left[\frac{V_{\text{in}}}{R_1 + R_2}\right]^2 R_2 = \frac{V_{\text{in}}^2 R_2}{(R_1 + R_2)^2}$$
$$= \frac{V_{\text{in}}^2}{R_1} \frac{\frac{R_2}{R_1}}{\left(1 + \frac{R_2}{R_1}\right)^2} = \frac{V_{\text{in}}^2}{R_1} \frac{x}{(1+x)^2} \tag{4.89}$$

after substitution of $R_2/R_1 = x$. The function $f(x)$

$$f(x) = \frac{x}{(1+x)^2} \tag{4.90}$$

has a maximum for $x = 1$, leading to $\max(f(x)) = 1/4$, hence

$$P_{\max} = \frac{V_{\text{in}}^2}{4R_1}. \tag{4.91}$$

The conclusion is that the maximum power (4.91) that can be generated by a voltage generator V_{in} whose internal resistance is R_1 is achieved for $R_1 = R_2$. This conclusion is used a number of times throughout the book.

4.1.9.2 RC Voltage Divider

A second and more elegant voltage divider structure consists of a serial RC network. Of the two possible serial RC networks, the one where resistor R_2 of Fig. 4.28 is replaced with a capacitor C is analyzed in this section and illustrated in Fig. 4.29 (left). Unlike the pure resistive voltage divider in Sect. 4.1.9.1, which was frequency independent because impedance of ideal resistors does not have any frequency-dependent term, the RC voltage divider includes a capacitive element C whose impedance (4.17) is frequency dependent. Therefore, this voltage divider structure is capable of altering the frequency spectrum of the output signal.

Steady-state analysis of an RC voltage divider routinely uses complex numbers. By doing so, all three variables (the amplitude, the phase, and the frequency of the output signal) are calculated using the same equation. Moreover, (4.88) still holds after resistance R_2 is replaced with the expression for impedance Z_C given by (4.17). Using complex algebra, and by inspection of the schematic diagram in Fig. 4.29 (left), it is straightforward to derive an expression for the impedance Z_{RC} that is connected to the voltage source V_{in} and its phase ϕ as

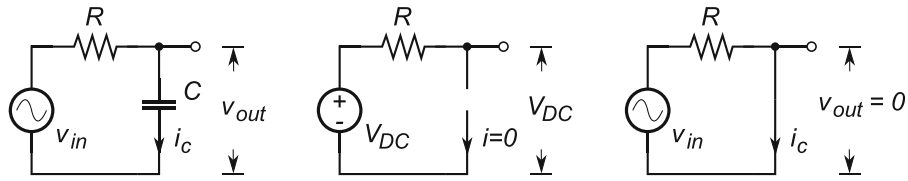


Fig. 4.29 RC voltage divider at AC (*left*), at DC (*middle*), and at $\omega = \infty$ (*right*)

$$Z_{\mathrm{RC}} = Z_{\mathrm{R}} + Z_{\mathrm{C}} = R - \frac{\mathrm{j}}{\omega C}, \tag{4.92}$$

$$\therefore$$

$$|Z_{\mathrm{RC}}| = \sqrt{R^2 + \frac{1}{\omega^2 C^2}} = R\sqrt{1 + \frac{1}{(\omega RC)^2}}, \tag{4.93}$$

$$\therefore$$

$$\phi = \arctan(-\omega RC). \tag{4.94}$$

With the help of simple algebra, (4.88) becomes

$$A_{\mathrm{V}} = \left|\frac{V_{\mathrm{out}}}{V_{\mathrm{in}}}\right| = \left|\frac{\frac{1}{\mathrm{j}\omega C}}{R + \frac{1}{\mathrm{j}\omega C}}\right| = \left|\frac{1}{1 + \mathrm{j}\omega RC}\right| = \frac{1}{\sqrt{1 + (\omega RC)^2}}. \tag{4.95}$$

Equation (4.95) includes the frequency dependence through the $\omega = 2\pi f$ term associated with the capacitor's impedance. A quick evaluation of (4.95) reveals that, for DC (i.e. $\omega = 0$), the capacitor has an infinite impedance, i.e. it becomes an open connection, as shown in Fig. 4.29 (middle), therefore $A_{\mathrm{V}} = 1$ or, to put it differently, $V_{\mathrm{out}} = V_{\mathrm{in}}$. At the opposite end of the spectrum, for $\omega = \infty$, the capacitor has zero impedance, i.e. it is a short connection, as shown in Fig. 4.29 (right), therefore $A_{\mathrm{V}} = 0$ or $V_{\mathrm{out}} = 0$. The output amplitude in the frequency domain between these two extremes is, therefore, described in accordance with (4.95). By definition, the frequency point where the power of the output signal equals half the input signal power is referred to as the "-3 dB" point, Fig. 4.30 (left); it is the frequency at which the real and imaginary parts of the voltage gain are equal, $\Re(A_{\mathrm{V}}) = \Im(A_{\mathrm{V}})$, which is to say that the ratio of the output voltage and the input voltage equals $1/\sqrt{2}$, i.e.[3]

$$\frac{1}{\sqrt{1 + (\omega_0 RC)^2}} = \frac{1}{\sqrt{2}} \quad \therefore \quad \omega_0 = \frac{1}{RC}. \tag{4.96}$$

The RC voltage divider is commonly referred to as a "low-pass filter" because it attenuates high-frequency components of the multi-tone signal while the DC tone passes unaffected. Frequency ω_0 that corresponds to the -3 dB amplitude point is the frequency parameter that determines its pass-band frequency and 45° phase shift, as in Fig. 4.30 (right). It should be noted that, for RC filters, the phase of the output voltage signal always lags the input voltage.

[3] Keep in mind Pythagoras' theorem in the complex domain.

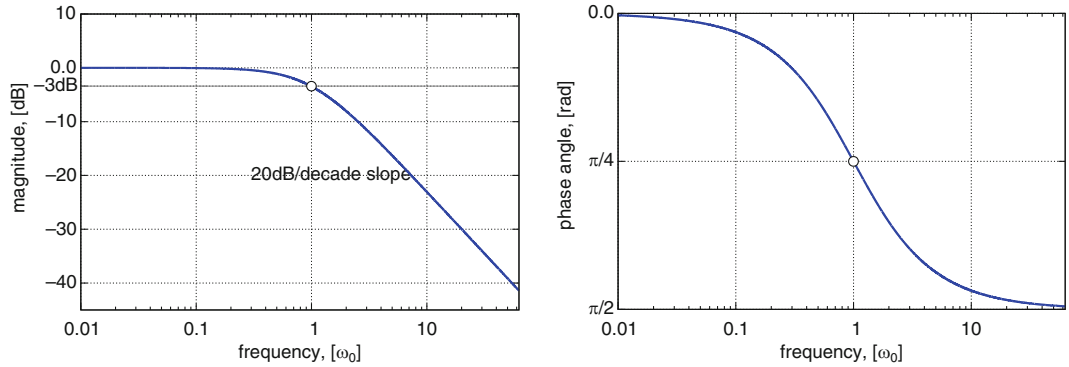


Fig. 4.30 Frequency domain plots of an LP RC filter: amplitude (*left*) and phase response (*right*)

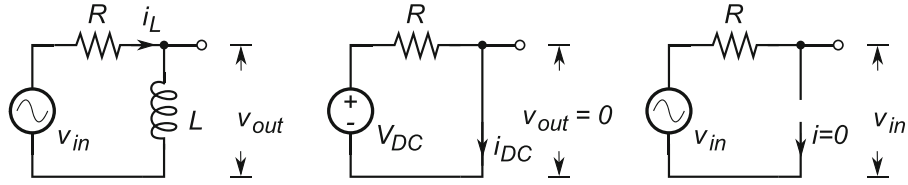


Fig. 4.31 Serial RL voltage divider at AC (*left*), at DC (*middle*), and at $\omega = \infty$ (*right*)

4.1.9.3 RL Voltage Divider

The third equally interesting and important type of voltage divider structure consists of a serial RL network. It is similar to the RC network of Sect. 4.1.9.2, with the capacitor being replaced by an inductor as shown in Fig. 4.31 (left). Due to the inductor's frequency dependence, this network also alters the frequency spectrum profile of the output signal. We find the frequency dependence of serial RL network as

$$Z_{\mathrm{RL}} = Z_{\mathrm{R}} + Z_{\mathrm{L}} = R + \mathrm{j}\omega L, \tag{4.97}$$

$$\therefore$$

$$|Z_{\mathrm{RL}}| = \sqrt{R^2 + (\omega L)^2} = R\sqrt{1 + \left(\frac{\omega L}{R}\right)^2}, \tag{4.98}$$

$$\therefore$$

$$\phi = \arctan\left(\frac{R}{\omega L}\right). \tag{4.99}$$

With the help of simple algebra, (4.88) then becomes

$$A_{\mathrm{V}} = \left|\frac{V_{\mathrm{out}}}{V_{\mathrm{in}}}\right| = \left|\frac{\mathrm{j}\omega L}{R + \mathrm{j}\omega L}\right| = \left|\frac{1}{1 - \mathrm{j}\frac{R}{\omega L}}\right| = \frac{1}{\sqrt{1 + \left(\frac{R}{\omega L}\right)^2}}. \tag{4.100}$$

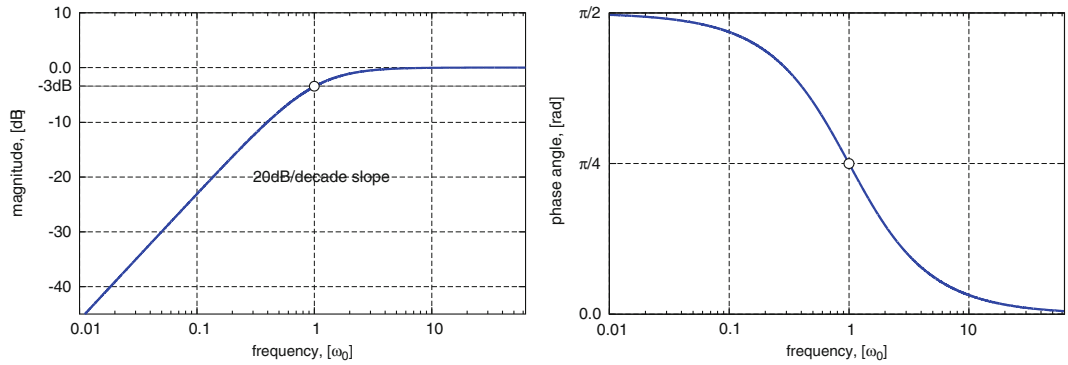


Fig. 4.32 Frequency domain plots of an RL HP filter: amplitude (*left*) and phase response (*right*)

This time, however, the inductor's reactance becomes zero at DC causing the output voltage to drop to zero as well, Fig. 4.31 (middle). At the other end of the spectrum, the inductor becomes an open connection when (because of infinite frequency) its reactance also becomes infinite, which effectively stops AC through its branch; stated differently, the output voltage becomes equal to the input voltage, Fig. 4.31 (right). The frequency domain amplitude plot in Fig. 4.32 (left) shows that the serial RL voltage divider behaves as a "high-pass filter". We find the $-3\,\text{dB}$ point by definition, as the frequency point where the power of the output signal equals half the input signal power, Fig. 4.32 (left); equivalently, it is the frequency amplitude where the real and imaginary parts of the voltage gain equation are equal, $\Re(A_\text{V}) = \Im(A_\text{V})$, which is to say that the ratio of the output voltage and the input voltage equals $1/\sqrt{2}$, i.e.[4]

$$\frac{1}{\sqrt{1+\left(\frac{R}{\omega_0 L}\right)^2}} = \frac{1}{\sqrt{2}} \quad \therefore \quad \omega_0 = \frac{R}{L} \tag{4.101}$$

and we note that the output voltage phase is "leading" the input phase by 90° at low frequencies, the phase lead reduces to 45° at the $-3\,\text{dB}$ point and, naturally, aligns its phase at high frequencies simply because the output signal becomes equal to the input signal.

4.2 Basic Network Laws

In addition to Ohm's law, the analysis of linear networks is based on writing a closed system of algebraic equations with the intent, for a given network, to solve for currents and voltages associated with all network branches. Kirchhoff's law and Thévenin's theorem are the two main procedures that have been developed specifically for the analysis of linear networks at low and medium frequencies.

[4] Keep in mind Pythagoras' theorem in the complex domain.

4.2.1 Ohm's Law

In Sect. 2.5, we introduced the two basic electrical variables: voltage and current. It was shown that electrical current I is defined by the number of electrons passing through a given cross-sectional area in a given amount of time. In the international system (SI) of units, electrical current, which is one of the seven basic units, is measured in amperes [A]. One ampere is defined as one *coulomb* (6.241×10^{18}) of electrons passing a given point per second. Further, we also stated that the number of the passing electrons in a closed circuit also depends on the value of the electromotive force V (measured in volts, V), which is not a basic SI unit; V is derived from energy and charge, J/C. Lastly, we also learned that the current depends on the type of material used to build the circuit, whether it was made of a good conductor or a good insulator, i.e. the material's *resistance*, R.

The relationship among these three variables, i.e. current, voltage, and the material's resistance, are summarized by Ohm in his elegant basic linear law

$$R = \frac{V}{I}. \tag{4.102}$$

Relationship (4.102) (and its variants in (4.11) and (4.12)) is considered fundamental engineering knowledge. A broader variant of (4.102), based on the connection with power P being delivered to a resistor, is

$$P = V\,I \qquad \therefore \qquad R = \frac{P}{I^2}. \tag{4.103}$$

This variant is more often used in radio electronics, because it covers all sorts of resistance that are used in wireless electronic systems, not only linear materials. The resistance associated with a general nonlinear material is measured directly by measuring the electric current flowing through the material and the total power in the system.

4.2.2 Kirchhoff's Laws

The most important method for solving a circuit network is a consequence of the conservation of charge and energy in electrical circuits and, therefore, can be derived using Maxwell's equations. The first of the two laws is most often referred to as *KCL* and it states that, at any instant, the total current entering any point in a network is equal to the total current leaving the same point. The second of the two laws is most often referred to as *KVL* and it states that, at any instant, the algebraic sum of all electromotive forces and potential differences around a closed loop is zero. These two rules are sufficient to generate a closed set of algebraic equations that are needed to solve any given network.

To illustrate the methodology, we consider the circuit in Fig. 4.33 (left) where all component values are given. The goal is to find all the voltages and currents associated with each network branch.

There are total of six currents in the network, therefore there must be six independent equations in the algebraic system:

$$I_1 = I_2 + I_4, \tag{4.104}$$

$$I_4 = I_5 + I_6, \tag{4.105}$$

$$I_3 = I_2 + I_6, \tag{4.106}$$

$$V_1 = R_4 I_4 + R_5 I_5, \tag{4.107}$$

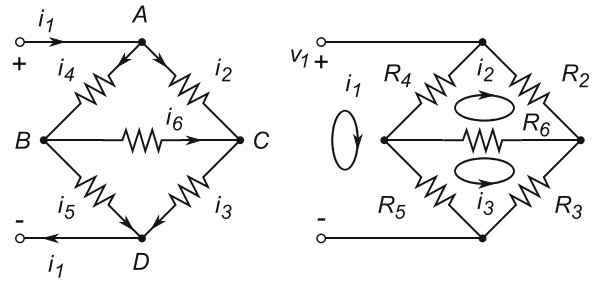


Fig. 4.33 Kirchhoff's laws

$$0 = R_6 I_6 - R_2 I_2 + R_4 I_4, \tag{4.108}$$

$$0 = R_6 I_6 + R_3 I_3 - R_5 I_5. \tag{4.109}$$

The first three equations are derived using KCL at nodes A, B, and C, while the second three equations are derived using KVL around loops L1, L2, and L3. It is important to note that it is possible to derive additional equations from the circuit, for example $I_1 = I_5 + I_3$ and $V_1 = R_2 I_2 + R_3 I_3$. However, they are not independent: they can be derived from (4.104) to (4.109).

The same circuit can be viewed as in Fig. 4.33 (right) where the concept of Maxwell's circulating currents is used, i.e. each loop is assigned its own independent "circulating current". The current from voltage source V_1 is labelled I_1, the current in branch AC is $I_{AC} = I_2$, in branch CD is $I_{\mathrm{CD}} = I_3$, in branch AB is $I_{\mathrm{AB}} = (I_1 - I_2)$, in branch BD is $I_{\mathrm{BD}} = (I_1 - I_3)$, and in branch BC is $I_{\mathrm{BC}} = (I_3 - I_2)$, leading to the following three voltage equations

$$V_1 = R_4(I_1 - I_2) + R_5(I_1 - I_3), \tag{4.110}$$

$$0 = R_2 I_2 + R_6(I_2 - I_3) + R_4(I_2 - I_1), \tag{4.111}$$

$$0 = R_3 I_3 + R_5(I_3 - I - 1) + R_6(I_3 - I_2), \tag{4.112}$$

which, after relatively simple algebra, are solved for I_1, I_2 and I_3. Knowing these three circulating currents, it is straightforward to resolve the branch currents and voltages. In this particular example, the calculation may be simplified if one node, for example node D is declared the local ground, i.e. $V_{\mathrm{D}} = 0$ and potentials at nodes B and C are labelled as independent voltages V_{B} and V_{C}. Because the voltage at node A is set by the voltage source, i.e. $V_{\mathrm{A}} = V_1$, there are only two independent equations for the two independent voltages. Once the potentials at the four nodes are known, it is trivial to resolve the branch currents and voltages.

As a last note regarding methodologies for solving circuit networks, be reminded that if multiple voltage or current sources are present in the network, each of the sources is considered to drive its independent current through the network. Hence, the superposition principle may be used to simplify the process.

4.2.3 Thévenin and Norton's Transformations

A very elegant, and intuitively very useful, approach to network analysis is to introduce concept of the Thévenin generator (first discovered by H. Helmholtz), which consists of V_{Th} and R_{Th} (see Fig. 4.34). It is used to replace any linear electrical network with its two-terminal "black box" model. One way of looking at the Thévenin generator is that it represents a non-ideal voltage source, where

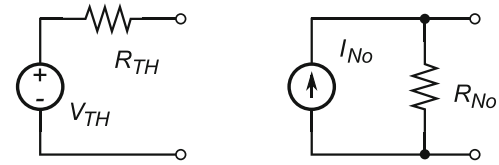


Fig. 4.34 Thévenin (*left*) and Norton (*right*) generator models

its internal resistance is equal to R_{Th} and electromotive force to V_{Th} assuming there is no external load. Consequently, as soon as some other resistance is connected to a Thévenin generator, then the resulting circuit diagram becomes identical to the one for the simple voltage divider shown in Fig. 4.28 (keep in mind, however, the Thévenin component values). Using Thévenin's theorem also leads to the concept of "input" and "output" impedances, i.e. the "looking into" approach is used extensively in circuit analysis. For example, looking into the two terminals of the Thévenin generator in Fig. 4.34 (left), it is straightforward to conclude that its output impedance is only R_{Th} because the ideal voltage source has zero resistance.

Thévenin's dual black-box model consists of an ideal current source I_{No} in parallel with a resistor R_{No}, as shown in Fig. 4.34 (right). It is also known as "Norton's generator". The two models are equivalent and it is only matter of convenience which one is used. Looking into the two terminals of Norton's generator, it is straightforward to conclude that its output impedance is only R_{No} because the ideal current source has infinite resistance.

4.3 Semiconductor Devices

The basic devices described in Sect. 4.1 by themselves are not capable of amplifying an electrical signal introduced at the input terminals of a network. Being *passive devices*, they can only provide a series of voltage dividers along the path, which can only progressively reduce the amplitude of the input signal. In order to enable an increase in the signal amplitude, i.e. to have the gain larger than one, the network must include *active devices*, namely diodes and transistors. It is important to clarify that the increased signal power showing at the output terminals of a transistor is provided by the external energy source, i.e. a battery, and is not magically created inside the transistor (the law of energy conservation still holds). That is, a transistor merely serves as a valve that controls the flow of a large current through the transistor's output terminals, where the current is drawn from the energy source, by means of low power signal at the input terminals. Simply put, we use the flow of a small amount of energy (the control signal) to control the flow of a large amount of energy (the external energy source, e.g. the battery), hence, the "amplification" effect. At the same time, we keep in mind that a transformer can increase the amplitude of either voltage or current at its output terminals but not both at the same time, which is the definition of power. Hence, a transformer is not a power-amplifying device.

In the rest of this section, the basic properties of p–n junctions, diodes and transistors are reviewed in order to consider their application to the amplification of weak RF signals.

4.3.1 Doped Semiconductor Material

In Sect. 2.4, we briefly introduced the basic terminology related to semiconductor materials. It is now time to take a closer look at the mechanism that enables the functionality of all semiconductor devices. As already implied, pure (intrinsic) semiconductor material, such as silicon, is electrically neutral and its atoms are arranged in a regular three-dimensional crystalline structure, the crystal lattice shown in

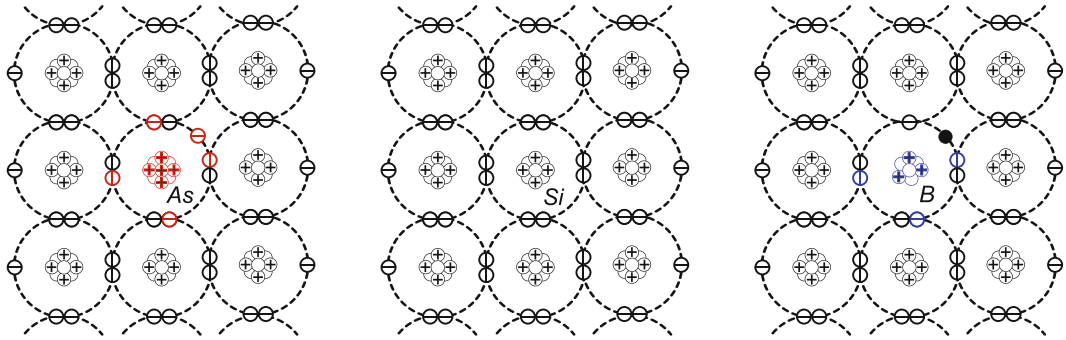


Fig. 4.35 Types of silicon: doped n type with added arsenic (As) that donated an electron (*left*); intrinsic (*centre*); and doped p type with added boron (B) that donated an "electron hole" (*right*)

Fig. 4.35 (centre). Using a special manufacturing process (known as "doping"), it is possible to replace some of the silicon atoms in the crystal lattice with either arsenic or boron atoms. These two elements are used because their atomic sizes are close to the size of a silicon atom and, hence, replacing the silicon atoms with either of the two "dopants" does not disrupt the silicon lattice too much. Because atoms of any element are electrically neutral, the overall "doped" slab of silicon is still electrically neutral.

Silicon belongs to the semiconductor group, IV in the periodic table, which is to say that its atoms have four electrons in the outermost shell (valence electrons) available for establishing bonding connections with other atoms, Fig. 4.35 (centre). Arsenic comes from group V (i.e. it has five valence electrons), while boron comes from the group III (i.e. it has three valence electrons). An interesting situation occurs when a relatively small number of dopants is added to intrinsic silicon. For example, for every boron atom added there are three bonding connections with the surrounding silicon atoms plus one "missing electron" (i.e. positive hole) connection, Fig. 4.35 (right), with the fourth silicon atom. Any electron generated elsewhere, due to thermal movement for example, is attracted to fill in the hole (i.e. "to recombine") and complete the bonding connection. However, that electron leaves behind an empty spot, which is equivalent to saying the positive hole moved into that spot. From a macro perspective, this manifests as a random movement of positive charge carriers. Keep in mind though that, overall, the boron-doped silicon is still electrically neutral, i.e. there is still an equal number of protons and electrons in the given volume. However, because a number of "holes" are created in the process, this kind of silicon is referred to as "p-type" silicon (a lack of electrons is equivalent to an excess of positive charges).

An equivalent situation arises if arsenic is added to intrinsic silicon as in Fig. 4.35 (left). In that case however, for each added arsenic atom, four valence electrons complete the four connections with the surrounding silicon atoms and the fifth arsenic atom is free to go. Therefore, silicon with added arsenic is referred to as "n-type" silicon, because a number of free negatively charged electrons are created in the process.

In both cases, the free charge carriers are referred to as "minority charge carriers". A side but very important observation for the operation of semiconductor devices is that it turns out that the average mobility of holes in p-type silicon is approximately half of the electron mobility in n-type silicon. That is, for a given cross-sectional area and timeframe, the p-type electric current is half of the n-type electric current. Remember, electric current is defined as the number of charges passing through a cross-sectional area in a given unit of time, thus "half mobility" means half the current for the given time.

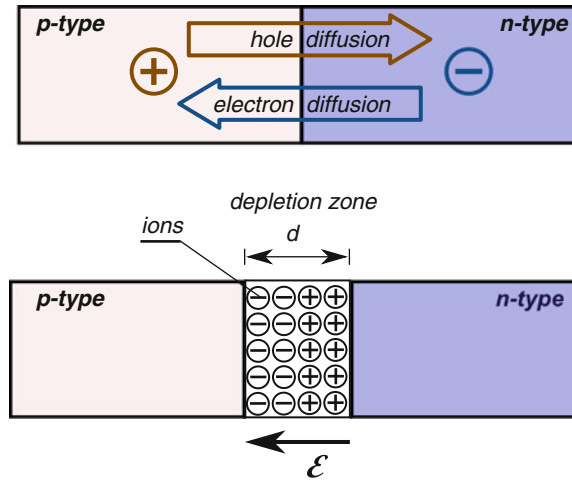


Fig. 4.36 P-n junction at the beginning of diffusion process before recombination started (*top*), and after depletion region has been established by the internal "built–in" electric field $\mathcal{E}$ (*bottom*)

4.3.2 P–N Junction

In order to understand what happens when pieces of p-type and n-type silicon material are brought into contact with each other, i.e. when a p–n junction is created, let us go through the following mental experiment.

To a first approximation, soon after the p–n junction is created inside a slab of silicon by the doping process, both p-type and n-type charge carriers, i.e. holes and electrons, are attracted by the opposite type of charge. Thus, electrons diffuse into the p-type region and fill in the electron holes, while at the same time (effectively) holes diffuse into the n-type region and recombine with the electrons, Fig. 4.36 (top). The primary effect of this diffusion process is that electrons leave positive ions[5] behind in the n-type silicon and create negative ions in the p-type region, Fig. 4.36 (bottom). In other words, it is equivalent to saying that holes moved from the p-type region into the n-type region, because the electrons that "crossed the border" recombined with the holes and completed the missing bonding connections. Consequently, by accepting the incoming electrons to recombine with the holes, the immobile hosting atoms are not electrically neutral any more; they are now negative ions fixed inside a p-type lattice. This process of charge carrier recombination and ion creation starts first in the region closest to the p–n junction boundary and expands deeper in both directions.

The region of space with ions only, i.e. without mobile carriers, expands on both sides of the original p–n junction plane and is referred to as a "depletion zone", Fig. 4.36 (bottom). The secondary effect of this forced diffusion is that the two types of fixed ion left in the depletion zone form the internal fixed electrical field $\mathscr{E}$ in the direction that opposes the free carrier diffusion. In other words, the newly created internal electrical field forms the electrical barrier that, as it increases, opposes the flow of the diffusion current. Eventually, the force of the built-in potential becomes high enough to stop the diffusion current and brings the p–n junction into its equilibrium state, Fig. 4.36 (bottom). The p-type–depletion region–n-type structure behaves as a charged capacitor: the p-type and n-type regions by themselves are not electrically neutral (they each have surplus of charges). For all practical purposes, the depletion region behaves as a non-conductive dielectric of thickness d (there are no free charge carriers inside). Typically, the built-in potential is controlled by the process and is designed to less than one volt.

[5]Ions are atoms with an imbalance of charge. Being part of the lattice, they are not mobile.

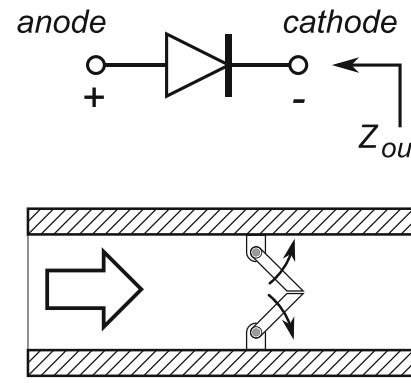


Fig. 4.37 Electrical symbol of a diode (*top*) and its equivalent mechanical function of a unidirectional water valve (*bottom*)

However, once established, the equilibrium state is easily affected by the external electric field caused by a properly connected external voltage source, e.g. a battery. In technical terms, the equilibrium is affected by the "external biasing". If the externally created electric field is in the same direction as the built-in field, it helps some of the charges to pass through and the depletion zone widens. Equally well, if the external electric field is in the opposite direction to the internal field, some of the charges are "pushed back" and the depletion zone becomes narrower and it may even disappear altogether. Hence, depending on the strength of the external field, the depletion region may be completely cancelled (which happens when $E_{\text{external}} \geq E_{\text{internal}}$) which enables free carrier current flow again (i.e. the p–n junction enters the forward-biased mode). Otherwise, the external bias may further reinforce opposition to the current flow, i.e. the p–n junction enters the reverse-biased mode. In other words, from the functional point of view, a p–n junction may be described as:

- A unidirectional, bias-dependent valve for electric current.
- A simple voltage-controlled capacitor whose capacitance is roughly calculated, using (4.14) for plate capacitance. The main difference is that the p–n junction capacitance is very nonlinear and dependent upon its biasing voltage. Nevertheless, it has important RF applications that are described in more detail in Sect. 8.7.

These two properties of a p–n junction are very important for the analysis of networks that include active components, as well as for analysis of the active components themselves. What is more, existence of the capacitive behaviour makes the p–n junction inherently sensitive to the frequency of externally induced electrical fields, which in return influences the behaviour of high-frequency electronic circuits.

4.3.3 Diode

The simplest electronic component that employs a single p–n junction is a diode. Its symbol, shown in Fig. 4.37 (top), implies that the forward-biasing mode of operation, i.e. "turning on" the diode, is achieved if the potential difference between the anode and cathode terminals is at least equal to or greater than the built-in p–n junction potential. Under that condition, looking into the cathode, one perceives an easy flow of current, hence it can be stated that the impedance Z_{out} of a forward-biased diode is very low (in the ideal case, zero). In contrast, when potential at the cathode node is greater than the anode potential, the diode is in reverse-biasing mode and, therefore, does not conduct appreciable current; in other words, the output impedance Z_{out} is very high (in the ideal case, infinite).

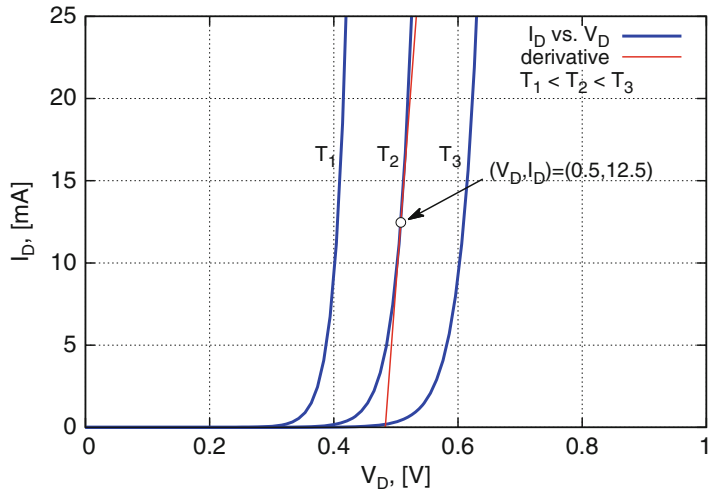


Fig. 4.38 A diode current plotted against voltage transfer characteristics (4.113) for three distinct temperatures. A derivative at the operating point illustrates the impedance of the forward-biased diode and its similarity to an ideal voltage source

The voltage–current characteristics of this two-terminal device obey the "exponential law", hence a diode mathematical model illustrates its "exponential nature" as

$$I_D = I_S \left[\exp\left(\frac{V_D}{n V_T} \right) - 1 \right] = I_S \left[\exp\left(\frac{q V_D}{n k T} \right) - 1 \right], \tag{4.113}$$

where,

I_D is current flowing through the diode,
I_S is the diode leakage current,
V_D is voltage across the diode, i.e. biasing voltage,
V_T is the thermal voltage ($V_T = kT/q$),
k is the Boltzmann constant ($k = 1.380650 \times 10^{-23}$ J/K),
T is the temperature in degrees Kelvin ($K = {}^\circ\text{C} + 273.15$),
q is the elementary charge, ($q = 1.602176487 \times 10^{-19}$ C),
n is the emission coefficient, usually between 1 and 2.

Equation (4.113) shows that the amplitude of the current flowing through the diode is controlled, in a fashion somewhat similar to a resistor (4.11), by voltage across its terminals. The main difference is that the relationship between I_D and V_D (4.113) is very nonlinear (the solid lines in Fig. 4.38). Instead of focusing on the absolute value of the diode voltage V_D, we note that actually it is the ratio of the biasing voltage V_D and the thermal voltage V_T that matters. A direct consequence is that a diode current is very temperature dependent. Therefore, calculation of the diode current I_D is valid only at one specific temperature and biasing point (V_D, I_D) in the V–I plane.

Realistic values of the diode leakage current I_S vary over a wide range of values even for the same diode type, sometimes as much as $\pm 50\%$. In addition, because of the exponential function (which is considered a "strong" function from a mathematical perspective), it is common practice to work with approximated V–I expressions after recognizing that there are two distinct regions of diode operation leading to the following important and useful approximations:

- $V_D \gg V_T$: when the diode biasing voltage is much larger (ten times or more, for example) than the thermal voltage V_T, the exponential term in (4.113) becomes much larger than the -1 term. In that

case, (4.113) degenerates into a plain exponential function that is much easier to handle from the mathematical perspective (think about the first, second, ... derivatives of an exponential function $\exp(x)$), i.e.

$$I_D = \left[\exp\left(\frac{V_D}{nV_T}\right) - 1\right] \approx I_S \exp\left(\frac{V_D}{nV_T}\right), \qquad (V_D \gg V_T). \tag{4.114}$$

For this extreme case, the diode is said to be fully "forward biased", i.e. it fully conducts current and its behaviour is similar to a plain wire or a closed switch in series with an ideal voltage source with V_D volts at its terminals, which is to say that its internal impedance is very low. It is easy to see how this conclusion is reached by looking at the first derivative of (4.113) in Fig. 4.38. At a given point the current–voltage characteristics are usually approximated with the first derivative at that point (the straight line in Fig. 4.38), which is very close to the V–I characteristics of an ideal voltage source and illustrates very low impedance of a forward-biased diode. The voltage across the diode is almost constant, hence, it is clear that the overall behavioural model of a forward-biased diode is very similar to an ideal DC voltage source when the built-in voltage is taken into account. For the example in Fig. 4.38, the forward-biased diode can be approximated as a voltage source of $V = 0.5$ V.

Moreover, in the first, very crude approximation, even the built-in voltage may be assumed equal to zero, and just state that the diode is "shorted", which is sometimes useful to quickly reach conclusions about the circuit operation. One needs to be comfortable using all of these levels of approximations appropriately while analyzing circuits that contain diodes.

- Constant voltage source: Once a diode is forward biased, for all practical purposes it could be replaced with an ideal voltage source. Direct application of (4.114) leads to the conclusion that the voltage across diode terminals V_D is set by current I_D forced through the diode by the external current source, and vice versa. In other words, from (4.114) it follows that setting the biasing voltage across the diode terminals ($V_D = 0.5$ V, for example) dictates that its current is set in accordance with (4.114), ($I_D = 12.5$ mA, for example; see Fig. 4.38).
- $V_D \ll V_T$: when the biasing voltage V_D is much smaller (ten times or more, for example) than the thermal voltage V_T, the exponential term in (4.113) becomes very close to one. In that case, (4.113) degenerates into the following expression:

$$I_D = I_S \left[\exp\left(\frac{V_D}{nV_T}\right) - 1\right] \approx 0, \qquad (V_D \ll V_T). \tag{4.115}$$

The diode is said to be "reverse biased", i.e. it is fully turned off and its behaviour is similar to that of an open switch – only a small portion of the leakage current I_D flows thtrouh the p–n junction boundary. It is important to note that if the anode and cathode terminals are shorted (or at the same potentials), in other words $V_D = 0$, then from (4.113) it follows that $I_D = 0$. That method is commonly used in circuits when there is a need to guarantee that a diode is turned off.

All three approximations are very useful and practical for a quick estimate of the behaviour of circuits that contain diodes. In practice, there are number of ways to design diodes optimized for a particular behaviour. For example, a Schottky diode is designed to have very fast switching times; a Zener diode (i.e. an avalanche diode) is designed for a specific reverse-bias *breakdown voltage* that is useful as a reference in voltage-stabilizing circuits; a varactor diode is designed specifically for its voltage-controlled capacitance (we meet it again in Sect. 8.7); a PiN diode is designed with a region of intrinsic silicon between p-type and n-type regions (hence the PiN name) to enable its linear voltage-controlled resistance behaviour, which is especially useful in microwave systems;

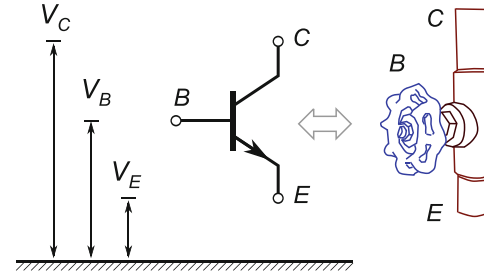


Fig. 4.39 Electrical symbol of a BJT (*left*) and its functional valve analogy (*right*) showing the relative potential levels of the three terminals in the case of an "open" valve

and a light-emitting diode (LED) is designed so that the recombination process results in the release of photons of light – by controlling the free carrier's energy levels, we control the frequency of the released photons, i.e. the emitted light colour.

4.3.4 Bipolar Junction Transistor

While a single p–n junction behaves as a unidirectional valve, the addition of a third semiconductor layer creates a second p–n junction and introduces a completely new dimension into the operation of this three-layer sandwich structure (a bipolar junction transistor). There are two possible ways to align the three layers, either as an NPN or a PNP structure, where the three layers are commonly referred to as collector (C), base (B), and emitter (E). It is important to know that the middle layer (the base) must be much thinner than the other two layers and that it is not possible to make a bipolar junction transistor (BJT) by mechanically placing in contact three layers that have been manufactured separately. Instead, the starting point is a single slab of either p-type or n-type semiconductor material (for example, silicon) whose side regions are then changed into the opposite type during the manufacturing process.

In its basic function, a transistor can be described as a simple valve that controls the flow in the CE branch. If the analogy for a diode was one of a unidirectional water pipe that is either fully open or fully closed, a BJT transistor is a unidirectional pipe with a third terminal (base) that controls the amount of current flow through the CE "pipe", from fully closed to fully open (see Fig. 4.39). The amount the transistor is open is controlled by the potential of the gate terminal relative to the other two terminals. If the gate potential V_B is below the emitter potential V_E, we see it as the transistor being "closed"—like the reverse-biased diode, it is fully turned off. When the gate potential V_B is equal to the collector potential V_C, the transistor is maximally open. Further increase of the gate voltage above the collector level does not change much—the transistor "pipe" has already reached its maximum diameter and the transistor is equivalent to a forward-biased diode, fully turned on. If you have ever used a slide potentiometer to control the volume of your audio equipment, you should not have a problem visualizing how the gate voltage moves up and down between the fixed emitter and collector "end points". It is important to keep in mind that it is the base–emitter diode (which is forward biased) that controls the transistor current, not the base–collector diode (which is reverse biased).

Of course, once the valve position is set, water flow through a pipe is controlled by pressure at its ends, where the pressure is generated by a water pump used to push the water through the pipe. In the case of a transistor, the "pressure" is provided by the external voltage source between the collector and the emitter terminals. In this analogy, a water pump is equivalent to a battery that forces electrons (i.e. charge carriers) to flow through wires and the flow of water drops is equivalent to the flow of electrons, i.e. electric current. However, a water valve is controlled manually, which obviously has a limitation in the number of open–close mechanical cycles that a human hand can achieve per second.

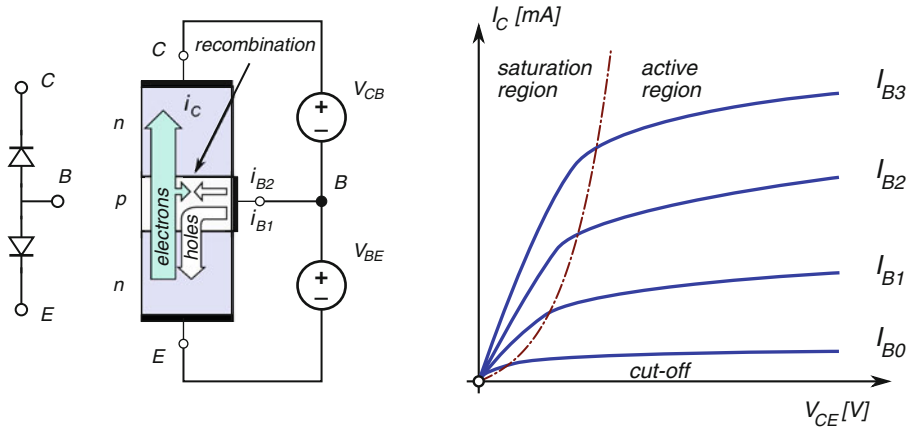


Fig. 4.40 NPN junctions in a BJT (*left*) and typical transfer characteristics showing the three main operating regions (*right*): cut-off, when there is no current flow; saturation region, when the increase of collector current is proportional to collector–emitter voltage (for a given base current); and active region, when the collector current is approximately independent of the collector–emitter voltage

An equally obvious limitation is that the controlling hand must be accompanied by the human owner. All in all, this is not a terribly efficient way of controlling the water flow. The open–close cycles of a transistor are controlled electronically, which enables a much higher number of switchings per second.

Indeed, current flow through transistor's CE branch is controlled by a small current injected into the base terminal, which is equivalent to saying that the transistor is controlled by a voltage difference V_{BE} between the base and emitter terminals across the input resistance between these two terminals R_{in}, Fig. 4.40. As a consequence, the small control current provided by the signal source connected to the gate terminal causes a large current variation in the CE branch, where the CE current is provided by the power supply source. Hence, it is said that a BJT enables power amplification—it is important to understand that the amplified power is generated not within the transistor itself, but by the external power supply source (e.g. a battery); the power is merely controlled by the transistor in a similar manner to the way in which a valve controls water flow through a pipe.

The electrical functionality of a BJT may also be described using a back-to-back diode analogy, Fig. 4.40 (left). If the BE diode is forward biased (i.e. $V_{\mathrm{B}} > V_{\mathrm{E}}$) *and* and the BC diode is reverse biased (i.e. $V_{\mathrm{C}} \geq V_{\mathrm{B}}$), the transistor is said to be in "active" or "constant current source" mode. This mode is of particular interest because the constant current at the collector node is set by the biasing point $(V_{\mathrm{BE}}, I_{\mathrm{C}})$ and therefore determines g_{m}, which subsequently controls the voltage gain A_{V} of a single stage amplifier. In addition, the two-diode model helps visualize how impedance looking into the collector node has to be "high" (keep in mind the reverse-biased diode BC) and impedance looking into the emitter node has to be "low" (keep in mind the forward-biased diode BE).

A detailed description of BJT operation, as presented in solid-state physics, is based on the mechanism known as the "injection of the minority carriers". It is a "bipolar" junction transistor because both minority and majority carriers are involved in the operation. In contrast, unipolar devices (e.g. *field-effect transistors*) rely on only one type of carrier for their operation (see Sect. 4.3.5). For the purposes of our discussion, we focus only on the descriptive explanation of BJT operation; interested readers are advised to take an introductory course in solid-state physics and semiconductor devices.

A mathematical description of the relationship between collector current I_C and base–emitter voltage V_{BE} in a BJT is similar to the voltage–current relationship of a diode, (4.113).[6] In addition, a transistor may be looked at as a three-terminal node, where the three currents must obey KCL. Finally, the collector current amplification factor β is introduced. Therefore, for a given V_{BE} voltage, it follows that

$$I_C = \beta I_B, \tag{4.116}$$

$$I_E = I_C + I_B, \tag{4.117}$$

$$I_C = I_S \left[\exp\left(\frac{V_{BE}}{n V_T} \right) - 1 \right], \tag{4.118}$$

where

I_C is the collector current,
I_B is the base current,
I_E is the emitter current,
I_S is the BJT leakage current,
V_{BE} is the base–emitter voltage,
β is the current gain factor,
n is the emission coefficient, usually between 1 and 2.

Equation (4.116) illustrates the basic current-amplifying property of a BJT—the collector current is β times larger than the base current. The current amplification factor β is usually of the order of 100 and it is controlled by the manufacturing process. However, β is not constant, indeed it is a strong function of the temperature, the transistor type, collector current, and the collector–emitter voltage V_{CE}. All in all, circuit designers try hard to design circuits with gains that are independent of β.

The relationship between the emitter and collector currents is derived by substituting (4.116) into (4.117) as follows

$$I_E = I_C + \frac{I_C}{\beta} \quad \therefore \quad I_C = \frac{\beta}{\beta + 1} I_E = \alpha I_E \approx I_E, \tag{4.119}$$

where α is the ratio of the collector and emitter currents and the last approximation is valid when $\beta \gg 1$ (which is valid for virtually all transistors).

Equation (4.117) illustrates that a BJT obeys KCL, while (4.118) emphasizes the fact that a BJT device is fundamentally a "transconductance" amplifier. The collector current I_C (the output variable) is controlled by the base–emitter voltage V_{BE} (the input variable). Similar to diode approximations (4.114) and (4.115), the BJT's base–emitter diode expression can be approximated (assuming $n = 1$) as

$$I_C \approx I_S \exp\left(\frac{V_{BE}}{V_T} \right) \tag{4.120}$$

under the condition that $\exp(V_{BE}/V_T) \gg 1$, i.e. when V_{BE} voltage is 2 to 3 times greater than the V_T voltage. By definition,

$$V_T \equiv \frac{kT}{q} \approx 25\,\text{mV} \quad \text{at room temperature } T = 290.22\,\text{K}, \tag{4.121}$$

[6] Remember the forward-biased diode BE.

which means that, at room temperature, as soon as V_{BE} voltage is greater than 50–75 mV or so, approximation (4.120) is valid. Just keep in mind that this base–emitter voltage V_{BE} is the one that controls the collector current. It is useful to note that, at room temperature ($T = 290.22$ K), (4.120) becomes

$$I_C \approx I_S\, e^{40 V_{BE}} \tag{4.122}$$

and we need to find an expression for the transconductance gain g_m of a BJT. By definition,[7] g_m is the derivative of the output current I_C against the input voltage V_{BE}, that is

$$g_m \equiv \frac{\partial I_C}{\partial V_{BE}} = \frac{\partial}{\partial V_{BE}}\left[I_S \exp\left(\frac{V_{BE}}{V_T}\right)\right] = \frac{I_S \exp\left(\frac{V_{BE}}{V_T}\right)}{V_T} = \frac{I_C}{V_T}, \tag{4.123}$$

which is a very important result. By inspection of (4.123), we conclude that the transconductance gain of a BJT is set, at the given temperature, strictly by its biasing point. This is a far-reaching conclusion that, basically, means as soon as the biasing point is provided to the BJT device by the external biasing circuitry, the details of the biasing can be completely ignored in the subsequent signal analysis, as long as the g_m value is used. Or, equivalently, an amplifying circuit is usually designed by specifying the required g_m value of the BJT device, which immediately translates into the collector current I_C ((4.123)) that is required to provide that particular g_m value. The external biasing circuit is designed to set the collector current by using (4.118).

Example 4.7. Estimate by how much the base–emitter voltage V_{BE} must be increased at room temperature so that the collector current is increased ten times.

Solution 4.7. A direct implementation of (4.120) for the two currents is written as

$$\begin{aligned}
I_C &= I_S \exp\left(\frac{V_{BE1}}{V_T}\right),\\
10 \times I_C &= I_S \exp\left(\frac{V_{BE2}}{V_T}\right),\\
&\therefore\\
\frac{10 \times I_C}{I_C} &= \frac{I_S \exp\left(\frac{V_{BE2}}{V_T}\right)}{I_S \exp\left(\frac{V_{BE1}}{V_T}\right)} = \exp\left(\frac{V_{BE2} - V_{BE1}}{V_T}\right) = \exp\left(\frac{\Delta V_{BE}}{V_T}\right),\\
&\therefore\\
\Delta V_{BE} &= V_T \ln 10 = 25\,\text{mV} \times 2.3026 = 57.567\,\text{mV} \approx 60\,\text{mV},
\end{aligned}$$

that is, if the gate voltage is increased by about 60 mV, the collector current is increased ten times.

An important BJT parameter is the small-signal impedance r_e looking into the emitter. Assuming that the collector and emitter currents are approximately the same, i.e. $I_C \approx I_E$, which is close enough if $\beta \gg 1$, or equivalently the base current is negligible relative to the collector current, the impedance is calculated by definition as

[7] See Sect. 7.1.4.

$$
\begin{aligned}
r_e &\equiv \frac{\partial V_{BE}}{\partial I_E} \approx \frac{\partial V_{BE}}{\partial I_C} \quad \therefore \quad \frac{1}{r_e} = \frac{\partial I_C}{\partial V_{BE}} = g_m, \\
r_e &= \frac{1}{g_m} \approx \frac{25\,\text{mV}}{I_C} \quad \text{at room temperature } T = 290.22\,\text{K} \\
&= \frac{25}{I_C\,[\text{mA}]},
\end{aligned} \tag{4.124}
$$

which is a very useful rule of thumb for estimating the emitter's output impedance of a BJT in the active mode of operation. For example, a typical biasing current $I_C = 1\,\text{mA}$ results in emitter output resistance of $r_e = 25\,\Omega$; $I_C = 0.5\,\text{mA}$ leads to $r_e = 50\,\Omega$, etc. This is an intrinsic emitter resistance that acts as if in series with the emitter node. It should be noted that this resistance is parasitic and caused by the silicon material that is used to manufacture the transistor. The small emitter resistance r_e is important because it limits the maximum possible transistor gain by preventing the emitter resistance becoming zero. In more detailed analysis, we find that the base–emitter voltage has a positive TC, contrary to what (4.118) would suggest, because of the very strong temperature dependence of the BJT leakage current I_S. It is useful to remember that the base–emitter voltage V_{BE} increases by approximately 2 mV/°C, which is an important data point for the design of temperature-independent voltage references, known as "bandgap references".

Example 4.8. To illustrate the point of BJT voltage gain A_V at the output terminal, let us assume that a 1 mV signal is applied across the input impedance of a BJT, Fig. 4.40 (left), and that the output load impedance connected between the collector and the external power supply is $R_L = 10\,\text{k}\Omega$. Calculate the voltage gain A_V and the power gain delivered to the loading resistor at room temperature if $\beta = 100$ is assumed.

Solution 4.8. The input current into the emitter is $I_E = V_{BE}/r_e = 1\,\text{mV}/25\,\Omega = 40\,\mu\text{A}$. Assuming the same current at the collector output, $I_C \approx I_E$, the voltage generated across the load resistance is $V_{out} = I_C R_L = 40\,\mu\text{A} \times 10\,\text{k}\Omega = 400\,\text{mV}$. Thus the voltage gain is $A_V = 400\,\text{mV}/1\,\text{mV} = 400 = 52\,\text{dB}$, which illustrates the high voltage gains achievable at a BJT output terminal. Note that the output current is converted into voltage by means of the loading resistance, otherwise there is no internal mechanism that would provide the voltage gain.

In addition, we find that power delivered into the load is

$$
\begin{aligned}
P_{in} &= I_B V_{BE} = \frac{I_C}{\beta} \times V_{BE} = \frac{40\,\mu\text{A}}{100} \times 1\,\text{mV} = 400\,\text{pW}, \\
P_{out} &= I_C V_{out} = 40\,\mu\text{A} \times 400\,\text{mV} = 16\,\mu\text{W}, \\
&\therefore \\
A_P &= \frac{16\,\mu\text{W}}{400\,\text{pW}} = 40{,}000 = 46\,\text{dB},
\end{aligned}
$$

which illustrates the point that a BJT device is capable of power gains, of course with the help of the external power supply from which the 40 μA current is drawn.

4.3.4.1 BJT Equivalent Circuits

Linear circuit analysis employs the traditional small-signal BJT AC model. In our discussion, however, we are going to use simplified models that are more appropriate for large-signal analysis, which

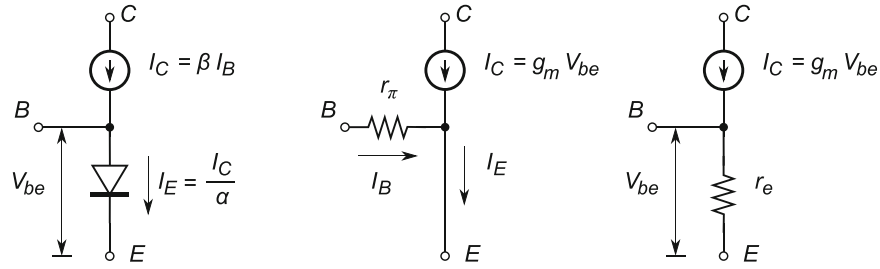


Fig. 4.41 Common large-signal model functional diagrams of NPN BJT in active mode

reflects the nature of many RF circuits and facilitates the approximate analysis approach that is used throughout of this book. In addition, all models presented in this section assume an active mode of operation, at room temperature, and large β values.

A forward-biased base–emitter diode controls the overall current in accordance with its exponential function (4.118); the β factor controls the ratio between the collector current and the base current, hence the straightforward circuit implementation of this model is shown in Fig. 4.41 (left). Equivalently, we define the base resistance r_π as

$$v_{\mathrm{be}} = r_\pi \, i_{\mathrm{B}}, \tag{4.125}$$

therefore,

$$i_{\mathrm{C}} = g_{\mathrm{m}} \left(r_\pi \, i_{\mathrm{B}} \right) = g_{\mathrm{m}} \left(\frac{\beta \, i_{\mathrm{B}}}{g_{\mathrm{m}}} \right) = \beta \, i_{\mathrm{B}}, \tag{4.126}$$

which is illustrated in Fig. 4.41 (centre). Another useful variant of the BJT large-signal model emphasizes the small emitter resistance r_{e}, Fig. 4.41 (right), as

$$v_{\mathrm{be}} = r_{\mathrm{e}} \, i_{\mathrm{e}}, \tag{4.127}$$

therefore,

$$v_{\mathrm{be}} = r_\pi \, i_{\mathrm{B}} = r_\pi \frac{i_{\mathrm{C}}}{\beta} = r_\pi \frac{\alpha \, i_{\mathrm{E}}}{\beta} = \frac{r_\pi}{\beta + 1} \, i_{\mathrm{e}} \qquad \therefore \qquad r_{\mathrm{e}} = \frac{r_\pi}{\beta + 1}, \tag{4.128}$$

where the last expression shows the relationship between the base and emitter resistances. It is very handy to interpret (4.128) as the "magnifying effect" between the two resistances: the resistance associated with the emitter node is perceived at the base node as being multiplied $\beta + 1$ times and the resistance associated with the base is perceived at the emitter node as being divided $\beta + 1$ times. In order to visualize this effect, just imagine that you are located at the base node and you are using binoculars to look at the emitter resistance. If the binocular has magnification of $\beta + 1$, than the size of the emitter resistor is enlarged by the same factor. Now, move to the emitter node and take a look at the base resistor. However, this time look through the binoculars from the wrong side (i.e. the picture is reduced in size instead of enlarged). That is, the size of base resistance is seen as being $\beta + 1$ times smaller than its true value. We use this trick very often to evaluate the input and output resistances associated with the base and emitter nodes of active BJT devices.

Example 4.9. If a BJT with current gain factor $\beta = 99$ has $R_{\mathrm{E}} = 1\,\mathrm{k\Omega}$ connected between its emitter node and the ground, estimate the input impedance perceived at the input node. For simplicity, ignore values of r_{e} and r_π resistances.

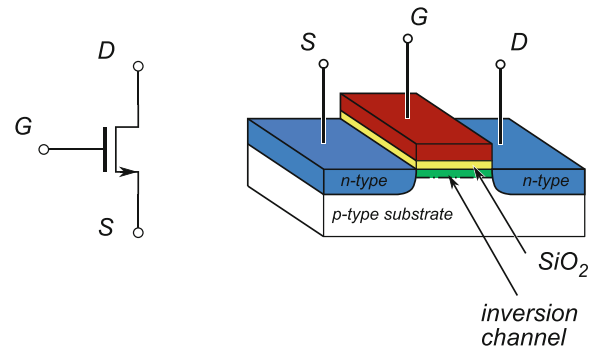


Fig. 4.42 Electrical symbol of an NMOS and its physical geometry

Solution 4.9. By using the magnification effect reasoning, the resistance associated with the emitter node, $R_E = 1\,k\Omega$, is seen from the base node as $R_{in} = (\beta + 1)\,R_E = 100\,k\Omega$, which illustrates how the emitter resistance influences the base resistance and makes it relatively high.

4.3.5 MOS Field-Effect Transistor

From the functional perspective, a field-effect transistor (FET) is equivalent to a BJT. It is a three-terminal device (see Fig. 4.42), the three terminals being drain (D), gate (G), and source (S), whose roles are equivalent to the collector, base, and emitter of a BJT.[8] Similarly to a BJT, its main role is to serve as a valve that controls the flow of current through its drain–source branch. Over time, several FET varieties have been developed and used. In this textbook, we review only the most common one, the enhanced NMOS and its symmetrical counterpart the PMOS.

Although they are made to serve essentially the same function, there are number of fundamental differences between MOS Field-Effect Transistor (MOSFET) and BJT devices in terms of how the current flow control function is implemented:

- Current flow in a BJT device is caused by the movement of electrons and holes at the same time (hence, it is "bipolar"). In a FET device, only one type of carrier makes the current, i.e. either electrons or holes, hence a FET device is "unipolar".
- A BJT device is asymmetrical by design, hence the current always flows from the collector to the emitter. FET devices are symmetrical and the roles of drain and source are determined only by their potentials within the specific circuit for each device separately.
- A FET device is constructed by placing a thin isolation layer of SiO_2 underneath the conducting gate layer and above the substrate, as shown in Fig. 4.42 (right). That is, for all practical purposes a FET gate represents one plate of a *capacitor* (the substrate itself being the second). In contrast to the base current of a BJT, under normal operational conditions there is no DC flow into the gate[9] and it is always assumed that the gate current is zero. Therefore, while a BJT is considered mostly as a current-amplifying device (i.e. base current in becomes collector current out), a FET device is a true *transconductance device* (i.e. voltage input controls current output).
- The current conduction mechanism in a BJT device is based on the principle of injection of minority carriers. The current conduction mechanism in a FET device is based on the "inversion channel" principle. In short, gate voltage creates a vertical electric field that attracts free carriers

[8] Strictly speaking, it is a four-terminal device, the fourth terminal being the body, i.e. the substrate.

[9] Modern FET devices have a very thin gate layer and, therefore, there is visible "current leakage" which is ignored in the first approximation.

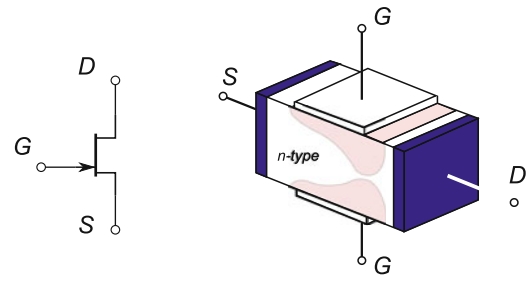


Fig. 4.43 Electrical symbol of a JFET and its physical geometry

from the substrate to amass underneath the SiO_2 isolation layer until, eventually, the gate voltage creates a strong enough field that the concentration of free carriers underneath the gate increases so much that the thin layer of semiconductor substrate turns into a conductive layer that is called the "inversion layer". As soon as the gate voltage is removed, the free electrons repel and return to the substrate. The minimum gate voltage that is needed to create the inversion layer is called the "threshold voltage" V_T. Once the inversion layer is established, even a small horizontal electric field caused by a voltage difference between the drain and source potentials causes current flow. Hence, a FET device relies on two electric fields for its operation.

- A BJT device has one forward-biased diode (the base–emitter diode). Inside a FET device, all internal diodes are reverse biased under normal working conditions. Note that each p–n junction inside a transistor is a diode in its own right.
- A BJT is a "vertical device"—it is manufactured so that the NPN (or PNP) sandwich is vertical relative to the substrate surface. A FET is considered to be a "lateral device"—its NPN (or PNP) sandwich is parallel to the substrate surface. The orientation of the sandwich determines the direction of the CE current: vertical in a BJT and lateral in a FET, as shown in Fig. 4.42 (left).
- We have already established that the collector current in a BJT is controlled by the exponential function (4.118). A FET device is controlled by the square function between its drain current I_D and its gate voltage V_{GS}, i.e. in saturation mode (which is equivalent to active mode in BJT), we have

$$I_D \approx K\,(V_{GS} - V_T)^2, \tag{4.129}$$

where K is a constant and V_T is the threshold voltage.

Once the above differences are accounted for, analysis of circuits using FET devices is, in the first approximation, almost identical to that of BJT circuits. For our purposes, we ignore the other differences that are normally important to the FET circuit designer.

4.3.6 Junction Field-Effect Transistor

Although the junction field-effect transistor (JFET) structure was conceived before the structures presented in the previous sections, technology limitations delayed its realization until years after the other structures. In principle, a JFET is the simplest structure and could be considered as a hybrid between BJT and FET devices. However, unlike the other two types of device, there is no p–n junction on the drain–source path (see Fig. 4.43). Instead, a JFET is built from a slab of, for example, n-type doped material with metallic contacts as shown in Fig. 4.43 (right). It is very similar to a resistor with the drain and source as its terminals, where the applied voltage V_{DS} causes the current flow. On two sides of this conductive slab, there are two gate p-type doped gates controlled by the same voltage V_G.

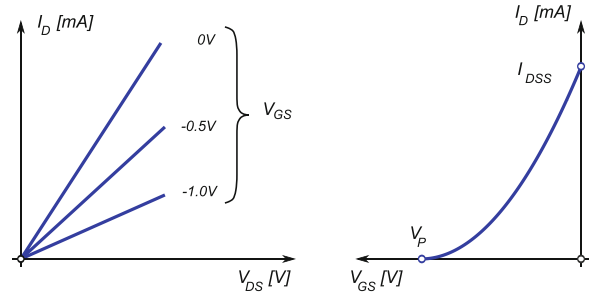


Fig. 4.44 Characteristics of the linear (*left*) and pinched-off (*right*) regions of a JFET

This gate voltage causes the depletion layer to extend deeper into the n-type slab, which effectively "squeezes" the current path and reduces the available cross-section, i.e. it increases the resistance of the current path. When the gate voltage is removed, the depletion layer is removed and the original resistance is restored.

The main difference between JFET and other FET devices is that, when the gate voltage $V_{GS} = 0$, it is said that JFET operates in the "ohmic region" and the depletion layer is very narrow, which means that the JFET behaves as a voltage-controlled resistor. In order to increase the depletion region, the gate potential must be negative relative to the source potential.

Let us take a closer look at the linear and pinched-off regions of a JFET. In the linear region, shown in Fig. 4.44 (left), JFET behaves as a voltage-controlled resistor, which is controlled by gate voltage V_{GS} as

$$I_D = \frac{2I_{DSS}}{|V_P|}\left(1 - \frac{V_{GS}}{V_P}\right) V_{DS}, \tag{4.130}$$

where V_P is the pinch-off voltage, i.e. the V_{DS} voltage when drain current $I_D = 0$, and I_{DSS} is the drain voltage when $V_{GS} = 0$. It is straightforward to find the transconductance of a JFET by finding the derivative of (4.130) as

$$g_m \equiv \frac{\partial I_D}{\partial V_{DS}} = \frac{2I_{DSS}}{|V_P|}\left(1 - \frac{V_{GS}}{V_P}\right), \tag{4.131}$$

which is the linear function for reverse-biasing values of V_{GS}. This linearity is useful in, for example, JFET-based multiplying circuits.

When a JFET operates in constant current mode, which is the most often used mode of operation, it is said to operate in the "pinch-off region", which is set when the drain–source voltage V_{DS} is above the linear region. In that mode, a JFET behaves as a VCCS. In pinch-off mode, the relationship of the gate voltage V_{GS} to the drain current I_D is given by

$$I_D = I_{DSS}\left[1 - \frac{V_{GS}}{V_P}\right]^2. \tag{4.132}$$

This parabolic relationship is shown in Fig. 4.44 (right). We note that the characteristic curves do not extend in positive direction much beyond $V_P = 0$. This is because a JFET must be operated with a reverse-biased gate, otherwise the gate diode is turned on and the pinch-off transfer function is not valid any more. For a specific JFET device, its values of I_{DSS} and V_P are determined by experimental measurement. It is then easy to apply the parabolic function (4.132) and calculate the drain current.

4.4 Summary

In this chapter, we have reviewed the basic devices that are used in RF circuit design. This review is by no means complete and thorough; it merely serves the purpose of being a reminder to the reader about the very basic and approximate facts describing the functionality of devices. Detailed treatment of each of the devices mentioned would cover a book similar to this one. The reader is advised to follow the literature and expand on the concepts learned in this chapter; without knowledge of fundamental device behaviour, any attempt to design an RF circuit is futile.

Problems

4.1. By definition, voltage across an inductor is related to current flow as

$$L \equiv \frac{\mathrm{d}i}{\mathrm{d}t}. \tag{4.133}$$

If the current plot in time is shown in Fig. 4.45, sketch the voltage graph across the inductor.

4.2. A capacitor $C = 1\,\mu\mathrm{F}$ and a resistor $R = 1\,\mathrm{k}\Omega$ are connected in parallel. At time $t_0 = 0\,\mathrm{s}$, the capacitor was charged up to voltage $V_0 = 10\,\mathrm{V}$. Sketch a graph of the voltage v_C across the capacitor over the next 5 ms. (Hint: the timing constant is $\tau = RC = 1\,\mathrm{ms}$.)

4.3. Starting from (4.5) and using mathematical software of your choice, recreate the plot in Fig. 4.3.

4.4. Using the model in Fig. 4.8, derive an expression for impedance amplitude $|Z|$ and, using mathematical software of your choice, pick component values and generate plots for (a) a through-hole resistor and (b) a surface-mounted resistor. Make recommendations for the useful frequency range of operation for these two resistors.

4.5. Assuming a 3.3 V power supply, design a 1 V voltage reference using a resistive divider. Assume that there is no branching current and that Thévenin resistance is $R_\mathrm{th} < 100\,\Omega$.

4.6. Using the model in Fig. 4.13, derive an expression for impedance amplitude $|Z|$ and, using mathematical software of your choice, generate plots for (a) a through-hole capacitor and (b) a surface-mounted capacitor. Make recommendations for the useful frequency range of operation for these two capacitors.

4.7. Using (4.28) and the plot of current waveform $i(t)$ shown in Fig. 4.45, for an inductor $L = 3\,\mathrm{H}$, sketch $v(t)$.

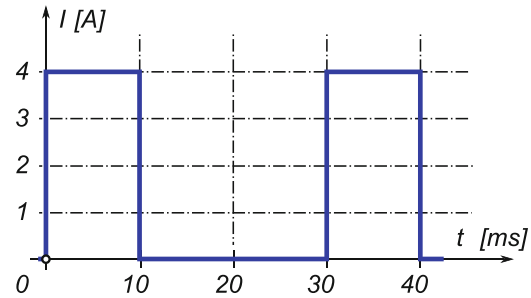


Fig. 4.45 Schematic diagram for Problem 4.1

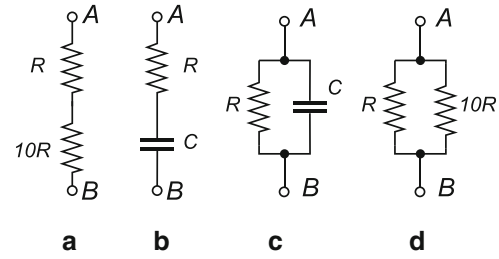


Fig. 4.46 Resistive networks for Problem 4.13

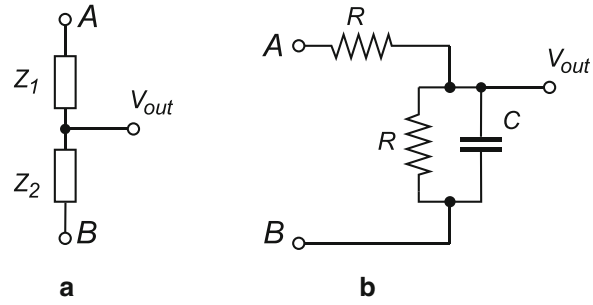


Fig. 4.47 Networks for Problem 4.14

4.8. A typical 1N4004/1A diode has the following parameters: $I_S = 76.9\,\text{nA}$ and $n = 1.45$. At a junction temperature of $T = 28°\text{C}$, calculate the diode current I_D if the forward-biasing voltage is (a) $V_D = 616.17\,\text{mV}$ and (b) $V_D = 50\,\text{mV}$. Find the calculation error (for each I_D) as a percentage, if (4.114) is used instead of (4.113)

4.9. The typical 1N4004/1A diode from Problem 4.8 is connected in series with an ideal current source $I = 1\,\text{A}$. At junction temperature $T = 28°\text{C}$, calculate the voltage V_D across the diode terminals as well as the voltage range if the current source varies by $\pm 10\%$.

4.10. For a JFET whose $V_P = 4.5\,\text{V}$ and $I_{DSS} = 7\,\text{A}$, plot the curve of I_D against V_{GS}.

4.11. For a BJT whose $I_S = 5 \times 10^{-15}\,\text{A}$ and, at room temperature, $V_T = 25\,\text{mV}$, the biasing current is $I_C = 1\,\text{mA}$. Calculate the base emitter voltage V_{BE}. Now, for $V_{BE} = 0.50\,\text{V}$, 0.55 V, 0.60 V, 0.65 V, 0.70 V, 0.75 V, 0.80 V, calculate the collector current I_C.

4.12. For the BJT from Problem 4.11, calculate g_m gain.

4.13. Calculate the equivalent resistance R_{AB} for the four resistive networks in Fig. 4.46. Find the equivalent resistances at the following frequencies: 1 Hz, 1 kHz, 100 kHz, 1 MHz, 100 MHz, ∞. Round the final result using reasonable engineering approximations.

4.14. For the networks given in Fig. 4.47, find the output voltage gain $V_{AB}/V_{out,B}$, assuming the frequencies from Problem 4.13 and for impedance ratios of $Z_1/Z_2 = R_1/R_2 = 10:1$, and $Z_1/Z_2 = R_1/R_2 = 1:10$. Assume $C = 15.915\,\text{pF}$.

4.15. Sketch approximate time domain plots of a signal that consists of:

(a) DC = 0 V and AC = 1 V
(b) DC = 5 V and AC = 1 V
(c) DC = 5 V and AC = 10 mV

4.16. For the network shown in Fig. 4.48 (left):

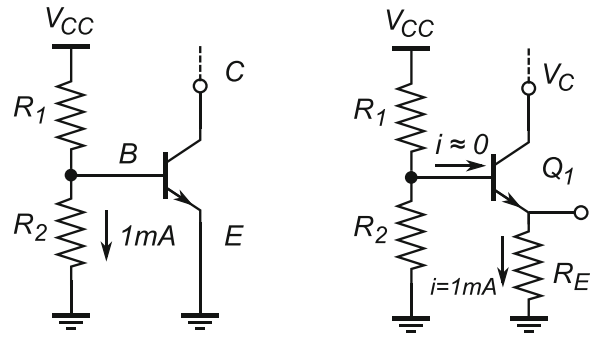


Fig. 4.48 Schematic of a network for Problem 4.16 (*left*) and for Problem 4.17 (*right*)

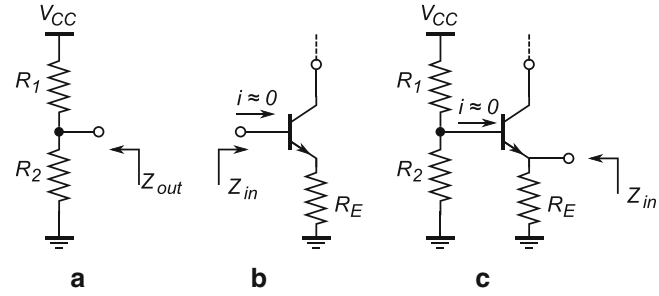


Fig. 4.49 Schematic of a network for Problem 4.18

(a) Assuming an ideal BE diode (i.e. the base–emitter diode threshold voltage is $V_{\text{th}}(BE) = 0\,\text{V}$), find values of R_2 so that the transistor Q_1 is turned on. What potential V_C is required at collector node C to maintain the saturation mode of operation?
(b) Assuming a realistic BE diode (i.e. the base–emitter diode threshold voltage is $V_{\text{th}}(BE) = 1\,\text{V}$), find values of R_2 so that the transistor Q_1 is turned on. What potential is required at collector node V_C to maintain the saturation mode of operation?

4.17. What is the required resistor ratio R_1/R_2 for the network in Fig. 4.48 (right), so that the transistor Q_1 is operating in saturation mode, if $V_{CC} = 10\,\text{V}$ and $R_E = 1\,\text{k}\Omega$:

(a) Assuming an ideal BE diode (i.e. the base–emitter diode threshold voltage is $V_{\text{th}}(BE) = 0\,V$), find values of R_2 so that the transistor Q_1 is turned on. What potential is required at collector node V_C to maintain the saturation mode of operation?
(b) Assuming a realistic BE diode (i.e. the base–emitter diode threshold voltage is $V_{\text{th}}(BE) = 1\,V$), find values of R_2 so that the transistor Q_1 is turned on. What potential is required at collector node V_C to maintain the saturation mode of operation?

4.18. Estimate the impedances looking into each of the networks in Fig. 4.49.

4.19. Calculate thermal voltage V_T under the following conditions: (a) $T = -55°\text{C}$, (b) $T = 25°\text{C}$, and (c) $T = 125°\text{C}$. Note that these three temperatures are commonly used to characterize military-grade electronic equipment.

4.20. For a typical 1N4004 diode, the specification sheet lists its saturation current as $I_S = 18.8\,\text{nA}$. Calculate the diode current I_D at the three temperatures listed in Problem 4.19, under the following conditions for the diode voltage V_D: (a) $V_D = 0.1V_T$, (b) $V_D = V_T$, and (c) $V_D = 10V_T$.

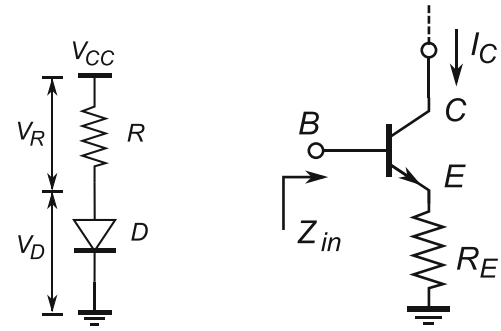


Fig. 4.50 Voltage reference networks for Problem 4.21 (*left*) and Problems 4.22–4.26 (*right*)

4.21. A simple voltage reference is built using a resistor and a 1N4004 diode as in Fig. 4.50 (left). Calculate the voltage across the diode at all three temperatures listed in Problem 4.19, under the following conditions: $V_{CC} = 9\,\text{V}$, $R = 1\,\text{k}\Omega$, $I_S = 18.8\,\text{nA}$. Express the result using scientific notation to three decimal places.

4.22. For the network in Fig. 4.50 (right), find the biasing voltage V_{BE} at all three temperatures listed in Problem 4.19 if BJT collector current is set to $I_C = 1\,\text{mA}$ and $I_S = 100\,\text{fA}$. Repeat the calculations for $I_S = 200\,\text{fA}$.

4.23. For the network in Fig. 4.50 (right), estimate the unknown collector current I_C that is required to force the biasing voltage $V_{BE} = 768.78\,\text{mV}$ if: (a) $I_S = 100\,\text{fA}$ and (b) $I_S = 200\,\text{fA}$. Do the calculations for all three temperatures listed in Problem 4.19.

4.24. For the BJT in Fig. 4.50 (right), estimate the biasing voltage V_B required at the base node so that the collector biasing current is set to $I_C = 1\,\text{mA} \approx I_E$. Data: $I_S = 100\,\text{fA}$, $R_E = 100\,\Omega$, $T = 25^\circ\text{C}$.

4.25. For the network in Fig. 4.50 (right), estimate the input impedance Z_{in} looking into the base node, if the emitter resistor is $R_E = 100\,\Omega$ and the forward gain β_F is assumed to be: (a) $\beta_F = 99$ and (b) $\beta_F \to \infty$.

4.26. For the circuit in Fig. 4.50 (right), design a preliminary resistive voltage divider to set the base biasing voltage. Use your engineering judgement for the design. Assume power supply voltage $V_{CC} = 9\,\text{V}$, emitter resistor $R_E = 100\,\Omega$, and: (a) $\beta_F = 99$; (b) $\beta_F \to \infty$. What is the percentage error between the solutions for the two values of β_F?

4.27. For a BJT with $I_S = 100\,\text{fA}$, $V_{BE} = 768.78\,\text{mV}$ at temperature $T = 25^\circ\text{C}$, calculate transconductance g_m. How large is the intrinsic emitter resistance r_E? How large is the collector current I_C? What happens to r_E if $I_C = 2\,\text{mA}$, $3\,\text{mA}$, ...? What if the temperature changes to, for example, $T = 30^\circ\text{C}$? Can you make any useful observations?

Chapter 5
Electrical Resonance

Abstract In the most familiar form of mechanical oscillations, the pendulum, the total system energy constantly bounces back and forth between the kinetic and potential forms. In the absence of friction (i.e., energy dissipation), a pendulum would oscillate forever. Similarly, after two ideal electrical elements capable of storing energy (a capacitor (which is initially charged) and an inductor) are connected in parallel then the total initial energy of the system bounces back and forth between the electric and magnetic energy forms. This process is perceived by the observer as electrical oscillations and the parallel LC circuit is said to be "in resonance". The phenomenon of electrical resonance is essential to wireless radio communications technology because without it, simply put, there would be no modern communications. In this chapter, we study behaviour and derive the main parameters of electrical resonant circuits.

5.1 The LC Circuit

The simplest electrical circuit that exhibits oscillatory behaviour consists of an inductor L and capacitor C connected in parallel (see Fig. 5.1). Let us assume the initial condition where the capacitor contains q amount of charge, hence the initial voltage V across the LC parallel network is related to the charges as $q = CV_C = Cv(max)$.

At time $t = 0$, the voltage across the capacitor is at its maximum $v(max)$, its associated electric field and stored energy are also at maximum, and the network current is still at zero value. That is, at time $t = 0$, the inductor is still "seen" by the capacitor charge as an ideal wire. Naturally, due to the electric field, the capacitive charge is forced to move through the only available path, the inductive wire. However, as soon as the first electron leaves the capacitor plate, this movement qualifies as a change of current in time and, according to (4.28), this "ideal wire" starts to show strong inductive properties accompanied by an appropriate magnetic field. Hence, while this current in creation flows through the inductor, it must obey Lenz's law and create the magnetic field that opposes the change that produced it. Eventually, the current reaches its maximum value $i(max)$ (at $t = T/4$) when the capacitor is fully discharged; the whole energy of the LC system is now stored in the inductor's magnetic field. It is now up to the inductor to serve as the energy source in the circuit and to push charges inside the wire while gradually passing the magnetic energy into the capacitive electrostatic energy. The uninterrupted flow of the current continues to cause the charges to keep accumulating at the other capacitor plate and along the way to create an electric field in the opposite direction relative to the initial state. This process continues until the capacitor is fully charged again (at $t = T/2$) (see Fig. 5.1), this time the voltage across the capacitor is at its minimum $\upsilon(min) = -\upsilon(max)$. Keep in mind that the system is assumed to be ideal, i.e., there is no thermal dissipation in wires, capacitor and

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8_5,

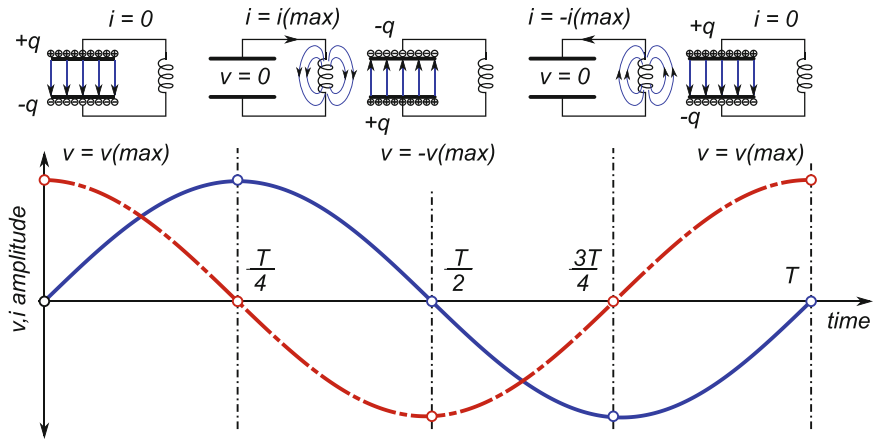


Fig. 5.1 Ideal LC resonance, the first cycle

inductor. Consequently, the energy conservation law must be maintained, which is the condition for a sustained repetitive exchange of energy between the inductor and capacitor.

It is straightforward to show that, in the time domain, the ideal LC circuit in Fig. 5.1 indeed creates electrical current that follows a sinusoidal waveform. We write the KVL equation around the loop as

$$v_C - v_L = 0 \quad \therefore \quad \frac{q}{C} + L\frac{di}{dt} = 0, \tag{5.1}$$

$$\text{(by definition)} \quad i = \frac{dq}{dt}, \tag{5.2}$$

therefore, after differentiating (5.2) we have

$$\frac{i}{C} + L\frac{d^2i}{dt^2} = 0 \quad \therefore \quad \frac{d^2i}{dt^2} + \frac{1}{LC}i = 0, \tag{5.3}$$

hence,[1]

$$i = I_0\cos(\omega_0 t + \phi) \quad \text{or} \quad i = I_0\sin(\omega_0 t + \theta), \tag{5.4}$$

where (5.4) is the standard solution of the second-order differential equation (5.3) and

$$\omega_0 = \frac{1}{\sqrt{LC}} \quad \Rightarrow \quad f_0 = \frac{1}{2\pi\sqrt{LC}}, \tag{5.5}$$

where ω_0 is the frequency of oscillation of a pure LC resonating circuit (i.e., with no thermal losses). We duly note that (5.4) is indeed a sinusoidal form that applies both to the resonating current and to the voltage (Fig. 5.1). Angular frequency ω_0, defined in (5.5), is the most important variable in RF design, so much so that it was given its own name, the *resonant frequency*. The resonant frequency is calculated either as ω_0 in rad/s or as f_0 in Hz, where $\omega_0 = 2\pi f_0$. The physical definition of resonance is the tendency of a system to oscillate at maximum amplitude at a certain frequency. This frequency is known as the system's resonant frequency. It is very important to distinguish the resonant frequency

[1] This is the second-order differential equation with a standard form of the solution.

from other modes of oscillations. While a system can oscillate at many frequencies, only the frequency associated with the maximal amplitude of oscillation is named the resonant, or natural, frequency.

For the sake of completeness, we repeat again the expression for the total energy $W = W_C + W_L$ contained in the LC resonator network, which is the sum of energies stored in the capacitor W_C and the inductor W_L, i.e.

$$W_C = \frac{1}{2}\frac{q^2}{C} = \frac{1}{2}v_C\, q = \frac{1}{2}C\, v_C^2, \tag{5.6}$$

$$W_L = \frac{1}{2}L i^2, \tag{5.7}$$

where, at time $t = 0$ there is no initial energy stored in the inductor $W_L(t = 0) = 0$; that is, the complete initial energy of the LC network is stored in the capacitor.

5.1.1 Damping and Maintaining Oscillations

The ideal resonating system introduced above demonstrated that, once started, the sinusoidal oscillations would maintain forever the amplitude of its waveform. Of course, that would have been a kind of perpetual motion machine, because we ignored the internal energy losses due to the system's internal resistance. The analogical mechanical system would be, for example, a swing that once pushed keeps swinging forever. In reality, this situation does not happen because of energy losses caused by internal friction in the swing's joints and air resistance to the body movement. As a result, the amplitude of oscillations becomes smaller with every passing cycle until, eventually, the movement completely stops.

It is in our interest to determine under what conditions oscillations of a realistic oscillator are maintained and at what rate energy is lost from the oscillator due to the internal and external imperfections in the system. A harmonic oscillator of which the oscillations lose amplitude over time is referred to as a "damped harmonic oscillator", which is a very general and common mode of behaviour in nature, exhibited by many seemingly unrelated systems: an imperfect LC resonator, a pendulum, a guitar string, or a bridge, to name a few.

The general mathematical treatment of a damped harmonic oscillator is found in many textbooks on mathematics and physics; for completeness of our topic, we repeat the basic definitions. The second-order linear differential equation is

$$a_2\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + a_1\frac{\mathrm{d}x}{\mathrm{d}t} + a_0 x = 0, \tag{5.8}$$

where a_2, a_1, a_0 are constants and x is the variable. It is more convenient to rewrite (5.8) in a form where the constant a_2 associated with the second derivative is normalized to one, hence we have

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \frac{a_1}{a_2}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{a_0}{a_2}x = 0 \quad \therefore \quad \frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \gamma\frac{\mathrm{d}x}{\mathrm{d}t} + \omega_0^2 x = 0, \tag{5.9}$$

where constant γ is the ratio a_1/a_2 and ω_0 is the natural frequency of the damped harmonic oscillator, defined as the ratio a_0/a_2. In general, (5.9) has three possible solutions depending on how much dampling is applied to the oscillator (see Fig. 5.2):

- Lightly damped oscillators have minimal energy loss (i.e., they are very close to the case of a non-damped oscillator). This class of oscillators, if left on its own, is able to sustain oscillations for an appreciable amount of time.

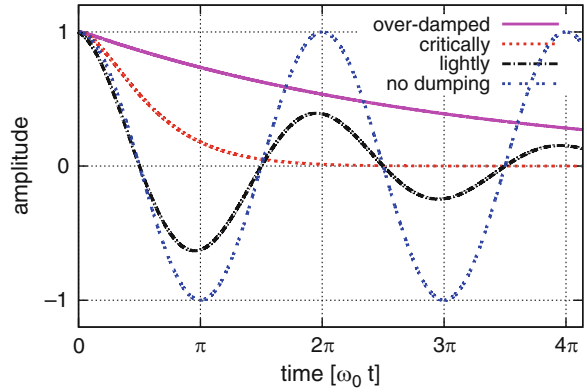


Fig. 5.2 Oscillation waveforms for various types of damping: ideal oscillations with no energy loss (no damping); oscillations with light energy loss (lightly); a system on the verge of starting oscillations (critically); and an over-damped system that cannot start oscillations (over-damped)

- Critically damped oscillators are on the verge of being able to start and sustain oscillations.
- Over-damped oscillators cannot start the oscillation process because of large energy losses.

Let us focus on a lightly damped oscillator because it is the only one of the three cases that can start oscillations. Intuitively, we conclude that the solution of (5.9) must include a term that reduces the initial amplitude over time; hence, we multiply the solution from (5.4) with an exponentially decaying function and adopt the solution of (5.9) in the following form

$$x = \exp\left(-\frac{t}{\tau}\right) A_0 \cos \omega t, \tag{5.10}$$

where t is the time variable, ω is the oscillation frequency, and τ is the timing constant that controls the rate of amplitude decay. For example, if $\tau = \infty$ than there is no reduction in the amplitude A_0 because the exponential term becomes equal to one at all times. At the other extreme, if $\tau = 0$ than the exponential term becomes zero, that is, the cosine function is completely suppressed. For any other value of τ, there will be natural decay in the initial amplitude A_0.

The first and second derivatives of (5.10) are

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -A_0 \exp\left(-\frac{t}{\tau}\right) \left(\omega \sin \omega t + \frac{1}{\tau} \cos \omega t\right), \tag{5.11}$$

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = A_0 \exp\left(-\frac{t}{\tau}\right) \left[\frac{2\omega}{\tau} \sin \omega t + \left(\frac{1}{\tau^2} - \omega^2\right) \cos \omega t\right]. \tag{5.12}$$

After substituting (5.10)–(5.12) into (5.9), we have

$$A_0 \exp\left(-\frac{t}{\tau}\right) \left[\left(\frac{2\omega}{\tau} - \gamma\omega\right) \sin \omega t + \left(\frac{1}{\tau^2} - \omega^2 - \frac{\gamma}{\tau} + \omega_0^2\right) \cos \omega t\right] = 0. \tag{5.13}$$

Equation (5.13) is possible at all times if the two multiplying constants of both sine and cosine terms are zero, i.e.

$$\left(\frac{2\omega}{\tau} - \gamma\omega\right) = 0 \quad \therefore \quad \tau = \frac{2}{\gamma}, \tag{5.14}$$

$$\therefore$$

$$\frac{1}{\tau^2} - \omega^2 - \frac{\gamma}{\tau} + \omega_0^2 = 0 \quad \Rightarrow \quad \omega = \sqrt{\omega_0^2 - \left(\frac{\gamma}{2}\right)^2} \tag{5.15}$$

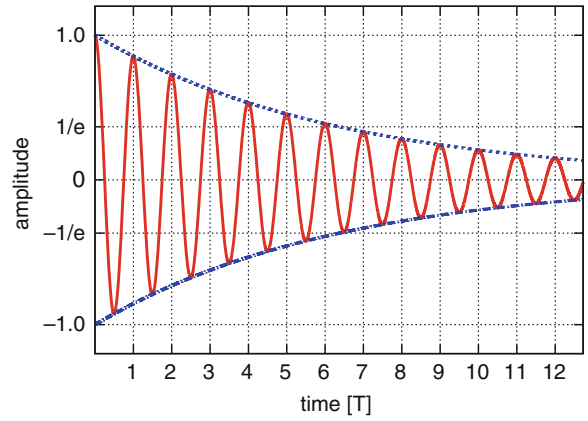


Fig. 5.3 Normalized amplitude and period of a decaying oscillator whose $Q = 20$. It can be found from the plot by counting the number of oscillation periods N until the amplitude drops to $1/e$, then $Q = N\pi$. In this case, $N \approx 6.5$, and $6.5 \times \pi \approx 20$. The *dotted lines* are the exponential envelope function $\exp(-\gamma t/2)$

after eliminating τ from the cosine coefficient. Now, we can rewrite solution (5.10) for the case of a lightly damped oscillator with zero initial phase as

$$x = A_0 \exp\left(-\frac{\gamma}{2}\, t\right) \cos \omega t. \tag{5.16}$$

The solution (5.16) is valid for ω as found in (5.15) and represents oscillatory motion if ω is real, i.e., if

$$\omega_0^2 > \frac{\gamma^2}{4}, \tag{5.17}$$

which is the condition for lightly damped harmonic oscillations. In addition, the frequency of a lightly damped oscillator is close to its natural resonant frequency if

$$\omega_0^2 \gg \frac{\gamma^2}{4} \qquad \therefore \qquad \omega = \sqrt{\omega_0^2 - \left(\frac{\gamma}{2}\right)^2} \approx \omega_0. \tag{5.18}$$

It is important to note that parameters γ and ω_0 are solely determined by the physical parameters of the circuit. An example of lightly damped harmonic oscillation is shown in Fig. 5.3. Let us now determine how the amplitude of a decaying cosine function changes along the envelope function by finding the ratio of the two maxima of the cosine function.

We write (5.16) at time $t = t_0$ and at $t = t_0 + nT$, where T is the cosine period, and n represents the index of the n-th maxima away from the one at t_0. Hence, we have expressions for the two amplitudes as

$$A_k = x(t_0) = A_0 \exp\left(-\frac{\gamma}{2}\, t_0\right) \cos \omega t_0, \tag{5.19}$$

$$A_{k+n} = x(t_0 + nT) = A_0 \exp\left(-\frac{\gamma}{2}\,(t_0 + nT)\right) \cos \omega (t_0 + nT) \tag{5.20}$$

$$= A_0 \exp\left(-\frac{\gamma}{2}\, t_0\right) \exp\left(-\frac{\gamma}{2}\, nT\right) \cos \omega t_0, \tag{5.21}$$

$$\therefore$$

$$\ln\left(\frac{A_k}{A_{k+n}}\right) = \frac{\gamma\, nT}{2} = \frac{\gamma\, n 2\pi}{2\omega} = \frac{\gamma\, n\pi}{\omega} \approx \frac{\gamma\, n\pi}{\omega_0} = \frac{n\pi}{Q} \tag{5.22}$$

because $\cos\omega(t_0+T)=\cos\omega t_0$. In (5.22), we introduced the ratio of the natural resonant frequency and γ as the figure of merit for the quality of oscillations, the Q factor,

$$Q=\frac{\omega_0}{\gamma}. \tag{5.23}$$

Obviously, for finite frequencies, $Q\to\infty$ implies that $\gamma\to 0$, which causes term $A_0\exp(-\gamma^t/2)\to A_0$, in other words, there is no damping.

Example 5.1. It was determined by measurement that the amplitude of a decaying cosine function at approximately 6.5 periods from $t=0$ is e times smaller than the initial amplitude value A_0. Estimate the Q factor of this resonator.

Solution 5.1. From (5.22), we write

$$\ln(e)=\frac{6.5\pi}{Q}\quad\therefore\quad Q=6.5\pi\approx 20, \tag{5.24}$$

which is shown in Fig. 5.3.

Now that we have determined the behaviour of a lightly damped harmonic oscillator, we conclude that condition (5.17) sets boundaries for the other two damping conditions as

$$\omega_0^2>\frac{\gamma^2}{4}\quad\text{lightly damped,} \tag{5.25}$$

$$\omega_0^2=\frac{\gamma^2}{4}\quad\text{critically damped,} \tag{5.26}$$

$$\omega_0^2<\frac{\gamma^2}{4}\quad\text{over-damped.} \tag{5.27}$$

The critically damped oscillator has the fastest time to equilibrium and still does not start oscillations. The over-damped system is dominated by the exponential decay function and slowly follows the envelope path (see Fig. 5.2).

Example 5.2. The amplitude of a decaying oscillation obeys $E(t)=E_0\exp(-t/\tau)$, where E_0 is the initial amplitude and τ is the decay time (Figs. 5.2 and 5.3). For a guitar string that produces a tone at 334 Hz, the sound decayed by factor 2 after 4 s. Estimate the decaying time τ and the quality factor Q.

Solution 5.2. We write

$$E(t)=E_0\exp(-t/\tau)\quad\therefore\quad\tau=\frac{t}{\ln\left(\dfrac{E_0}{E(t)}\right)},$$

$$\tau=\frac{4\,\text{s}}{\ln(2)}=5.77\,\text{s}$$

and, from (5.14) and (5.23) we write

$$Q=\frac{\omega_0}{\gamma}=\frac{\omega_0\tau}{2}=\frac{2\pi\times 334\,\text{Hz}\times 5.77\,\text{s}}{2}\approx 6\times 10^3.$$

A realistic electrical resonator consists of a serial RLC loop, similar to that shown in Fig. 5.6 (left) except that nodes a and b are connected. Under those conditions, we write KVL around the loop, as

$$L\frac{\mathrm{d}i}{\mathrm{d}t} + iR + \frac{q}{C} = 0, \tag{5.28}$$

$$\therefore$$

$$\frac{\mathrm{d}^2 q}{\mathrm{d}t^2} + \frac{R}{L}\frac{\mathrm{d}q}{\mathrm{d}t} + \frac{1}{LC}q = 0, \tag{5.29}$$

where (5.29) has an identical form to (5.9). Hence, we write by inspection that

$$\omega_0^2 = \frac{1}{LC}; \qquad \gamma = \frac{R}{L}, \tag{5.30}$$

$$\therefore$$

$$q(t) = q_0 \exp\left(-\frac{R}{2L}t\right)\cos\left(\sqrt{\frac{1}{LC} - \frac{R^2}{4L^2}}\,t\right), \tag{5.31}$$

therefore, the Q factor is found from (5.23) and (5.30) as

$$Q = \frac{1}{R}\sqrt{\frac{L}{C}}. \tag{5.32}$$

We observe that the presence of resistive element R is the cause of damping factor γ and a finite value of Q factor. In the case of $R = 0$, we have again the ideal resonator, i.e. $Q \to \infty$, with no damping and $\omega = \omega_0$.

5.1.2 *Forced Oscillations*

Understanding (5.16) for a free running resonator is important in order to be able to control conditions that are favourable for maintaining oscillations in a real system. Going back to the swing analogy, in order to maintain the oscillations, at the end of every cycle the swing needs to receive just the right amount of push in the right direction. This action causes just the right amount of energy to be regularly injected into the system so that the energy loss due to friction is compensated. The key points are that the compensating energy must be injected at the right moment in time and with the right phase, i.e., synchronized with the oscillations in the right direction.

As opposed to a simple mechanical system, such as a swing, it is not as practical to manually compensate for thermal losses in electronic systems, at least if we are to achieve any decent speed of operation. The good news, however, is that it is not difficult to synchronize electronic systems so that the losses are correctly compensated for and to maintain the oscillations. For example, a realistic LC resonator with internal losses, Fig. 5.4 (left), is connected to a transistor Q_1 that serves as a current source (the biasing details are omitted for simplicity). If at the end of each cycle we manually press the switch for a short period of time, just enough to inject the right amount of energy that is needed to compensate for the thermal losses per cycle, and if the phase of signal generator v_{err} is synchronized with the output oscillations v_{out}, the compensation signal is of the right amplitude, and the finger presses the switch fast enough, then the amplitude and frequency of the output voltage v_{out} is maintained. Although theoretically possible, it is not a terribly practical solution. Instead, we can

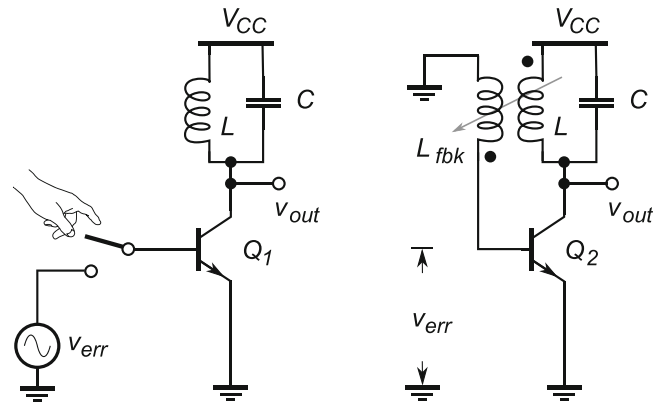


Fig. 5.4 LC resonating circuits with (*left*) a manual compensation mechanism for the internal thermal losses and (*right*) an automatic compensation mechanism

tap into the output oscillations v_{out} through inductive coupling with inductor L, Fig. 5.4 (right), and create a scaled copy of v_{out} in the coupled inductor L_{fbk} through a transformer effect with the phase controlled by the relative direction of the coil turns in the two inductors. The scaled, in phase signal is made equal to the required correction signal v_{err}, which controls the collector current of transistor Q_2 and closes the loop while always being in synchronicity with the output oscillations. With proper engineering, this mechanism provides the right amount of "push" at the right moment and in the right direction so that the injected energy precisely compensates for the thermal losses per cycle. We note that this is another example of intentional addition of two signals. Indeed, this principle is used in realistic oscillating circuits to maintain the amplitude of the sinusoidal signal. The main role of an LC resonator is to provide a physical realization of a sinusoidal function, which is fundamental for wireless radio communications, and therefore for radio transceiver circuit design.

The case of a forced RLC resonator is essential to RF communication systems and we are going to take a closer look by rewriting (5.29) as

$$\frac{\mathrm{d}^2 q}{\mathrm{d}t^2} + \frac{R}{L}\frac{\mathrm{d}q}{\mathrm{d}t} + \frac{1}{LC}\,q = V_0 \cos \omega t, \tag{5.33}$$

where $V_0 \cos \omega t$ represents a signal source that is connected, for example, between nodes a and b in Fig. 5.6. The mathematical procedure of a non-homogeneous, linear differential equation is a bit more involved, however it is easily found in calculus textbooks, hence we only write the solution for voltage across the capacitor $v_{\text{C}}(t)$ in the RLC forced resonator as

$$v_{\text{C}}(t) = V_{\text{C}}(\omega)\, \cos(\omega t - \delta), \tag{5.34}$$

$$V_{\text{C}}(\omega) = \frac{\frac{V_0}{LC}}{\sqrt{(\omega_0^2 - \omega^2)^2 + \left(\frac{\omega R}{L}\right)^2}}, \tag{5.35}$$

where δ is the phase difference between the voltage source and the oscillator's frequency. At resonance, when $\omega = \omega_0$, then (5.35) becomes

$$V_{\text{C}}(\omega_0) = \frac{V_0}{\omega_0 RC} = Q V_0. \tag{5.36}$$

We observe the very important fact, that when the local RLC resonator is designed for the resonant frequency ω_0 and the frequency of the external driving voltage source $V(\omega)$ coincides, i.e., $\omega = \omega_0$,

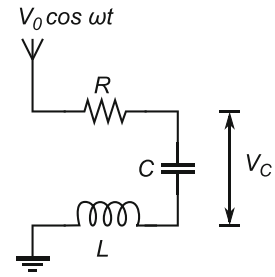


Fig. 5.5 RLC resonating circuit driven by RF voltage source $V_0 \cos \omega t$

there is significant amplification of the incoming tone (Q is usually very large). The simplified RLC circuit is driven by the incoming radio signal provided by the antenna shown in Fig. 5.5.

5.2 The RLC Circuit

In real systems, there is always a small resistance R associated with the connection wires and the inductor, as well as a small leakage current in the capacitor (due to the less than infinite resistance of the capacitor's dielectric material) which, all combined, cause a small amount of energy to be lost each cycle in the form of heat. As a result, if generated by a real RLC circuit with no external compensation, the waveform amplitude exponentially decays (Fig. 5.3). That is the main reason for having an external energy source that compensates for the internal thermal losses in real oscillating RLC circuits (Fig. 5.4). This statement is confirmed experimentally by LC resonators made of superconductive materials; once the internal current is induced, the superconductive resonator oscillates for a very long time (measured in days and months) without any external energy source (strictly speaking this is not correct—a large amount of external energy is spent on keeping the resonator cool, however that is not the point).

5.2.1 Serial RLC Network

In Sects. 4.1.5 and 4.1.6, we already learned that, in a capacitive network, the capacitive voltage V_C lags the current by 90° and that, in an inductive network, the inductive voltage leads the current by 90°. Intuitively, we conclude that if the two elements are put in the same network the two voltages must, therefore, have the phase difference of 180°, which leads to an interesting question: what happens if the two voltages are equal in amplitude? Obviously, one voltage must be subtracted (remember differential signals?) from the other, which leads to interesting conclusions. To illustrate the point, let us take a look at the following example.

Example 5.3. An AC voltage source V is connected across a serial LC connection. The data is: $V = 5\,\text{V}$, $f = 10\,\text{MHz}$, $C = 1\,\text{nF}$, and $L = 1\,\mu\text{H}$. Find the capacitive X_C and inductive X_L reactances and voltages V_C and V_L across their respective terminals.

Solution 5.3. The two reactances are calculated as $X_L = 2\pi fL = +62.832\,\Omega$ and $X_C = 1/2\pi f C = -15.915\,\Omega$.[2] Therefore, the total reactance must be $X_{LC} = X_L + X_C = +46.916\,\Omega$. That is, the total

[2]Remember, this is a relative comparison, thus it is agreed convention that the negative reactance is associated with a capacitance because $X_C = \frac{1}{j\omega C} = -j\frac{1}{\omega C}$.

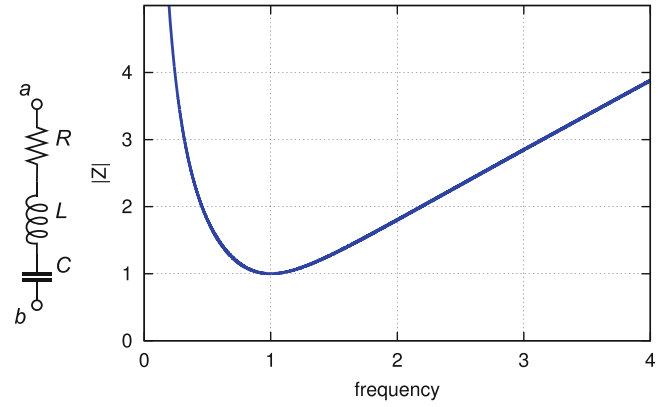


Fig. 5.6 Serial RLC circuit network with normalized resonant frequency $\omega_0 = 1$ (*left*) and its total impedance plot versus frequency (*right*)

reactance at this frequency is equivalent to the reactance of an inductor $L = X_{\mathrm{LC}}/2\pi f = 746.697\,\mathrm{nH}$, which further implies that, from the perspective of the voltage generator, at this particular frequency the serial LC connection could be replaced with a single 746.697 nH inductor without disturbing the rest of the circuit. The total branch current is, therefore, $I = V/X_{\mathrm{LC}}$=106.573 mA with -90° phase relative to the voltage.

While keeping in mind the phase relationships, the voltage across the inductor is calculated as $V_{\mathrm{L}} = I \times X_{\mathrm{L}} = 106.573\,\mathrm{mA} \times 62.832\,\Omega = 6.696\,\mathrm{V}$, while the voltage across the capacitor is $V_{\mathrm{C}} = I \times X_{\mathrm{C}} = 106.573\,\mathrm{mA} \times 15.915\,\Omega = 1.696\,\mathrm{V}$. Note that the inductor voltage is much higher than the one provided by the voltage source. However, the difference between the voltages is $V_{\mathrm{L}} - V_{\mathrm{C}} = 6.696\,\mathrm{V} - 1.696\,\mathrm{V} = 5\,\mathrm{V}$, as it should be in order to agree with the applied voltage. The conclusion is that we must be careful about the operational range of components used to build RLC resonators. If you try some other L and C values, then you find that it is also possible to set up larger voltages across the capacitor.

The addition of resistance R into a serial LC network, turning it into an RLC network (see Fig. 5.6), changes the overall circuit behaviour. The resistive component becomes responsible for thermal dissipation of the energy that was originally stored in the electrostatic and magnetic fields. As a consequence, the sinusoidal resonating current calculated in (5.4) cannot sustain its maximum value indefinitely. With each cycle, some of the electrical energy dissipates into heat, while the output voltage decays (Fig. 5.2). The absolute value of the total impedance is calculated as

$$Z = R + \mathrm{j}\omega L + \frac{1}{\mathrm{j}\omega C}, \tag{5.37}$$

$$|Z| = \sqrt{R^2 + (X_{\mathrm{L}} - X_{\mathrm{C}})^2} = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}, \tag{5.38}$$

where the two reactances, X_{L} and X_{C}, determine the equivalent total reactance X_{RLC} of the RLC network, which becomes zero if the two reactances are equal in their absolute values. The equality $X_{\mathrm{L}} = X_{\mathrm{C}} \;\; \therefore \;\; X_{\mathrm{RLC}} = 0$ is considered the required condition for resonance.

Under the condition of resonance, the absolute value of serial impedance is at its minimum, i.e., it is equal to R (Fig. 5.6), and it is therefore real, which means that the total phase angle equals zero (Fig. 5.7). By forcing the total reactance term in (5.38) to zero, we have

$$\omega L - \frac{1}{\omega C} = 0 \Rightarrow \omega L = \frac{1}{\omega C} \Rightarrow \omega_0 = \frac{1}{\sqrt{LC}}, \tag{5.39}$$

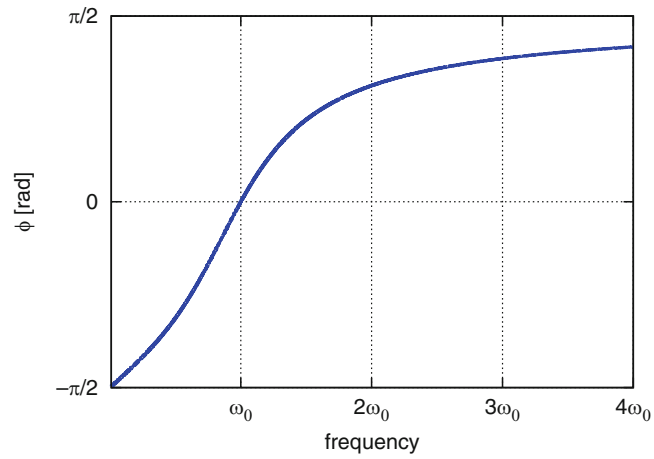


Fig. 5.7 Phase plot of serial RLC network, normalized to the resonant frequency ω_0

where ω_0 indicates the resonant frequency of the LC network under the $X_L = X_C$ condition, which is the same conclusion as the result (5.5) from the differential equation. At other frequencies, the total reactance is greater than zero, (5.38), which implies that R is the minimum value of the absolute impedance Z. The phase plot is easily obtained, after using definitions (2.12) and (2.13) and (5.38) as

$$\phi = \tan^{-1} \frac{\left(\omega L - \frac{1}{\omega C}\right)}{R}, \tag{5.40}$$

which is given in Fig. 5.7 after normalizing to the ω_0 resonant frequency.

Example 5.4. For the circuit in Fig. 5.6 and $f = 10\,\text{MHz}$, $C = 1\text{nF}$, $L = 1\mu\text{H}$, $R = 1\text{m}\Omega$, source resistance $r_i = 50\,\Omega$ and input voltage source V_{ab} is connected between terminals a and b, i.e. $V_{ab} = V_{in} = 1\text{mV}$, find: (a) output voltage $V_{out} = V_R$ as measured across the resistor R, at resonant frequency f_0; (b) output voltage $V_{out} = V_C$ as measured across the resistor C at 1 kHz.

Solution 5.4. After applying the voltage-divider rule, it follows:
(a) For resonant frequency f_0, $Z_L = Z_C \Rightarrow Z = R$, therefore

$$\frac{V_{out}}{V_{in}} = \frac{R}{R + r_i} = \frac{1\,\text{m}\Omega}{1\,\text{m}\Omega + 50\,\Omega} \approx 20 \times 10^{-6} \quad \therefore \quad V_{out} = 20 \times 10^{-6} \times 1\,\text{mV} = 20\,\text{nV}.$$

(b) For frequency $f = 1\,\text{kHz}$, total impedance seen by the voltage source is

$$\begin{aligned} Z_{in} &= \sqrt{(r_i + R)^2 + (X_L - X_C)^2} \\ &= \sqrt{(50\,\Omega + 1\,\text{m}\Omega)^2 + (29.581\,\mu\Omega - 1.592\,\text{k}\Omega)^2} \\ &= 1.593\,\text{k}\Omega \end{aligned}$$

and output impedance (as "seen" by V_{out}) is $Z_{out} = 1.592\,k\Omega$, hence

$$\frac{V_{out}}{V_{in}} = \frac{Z_{out}}{Z_{in}} = \frac{1.593\,\text{k}\Omega}{1.592\,\text{k}\Omega} \approx 1 \qquad \therefore \qquad V_{out} = 1 \times 1\,\text{mV} = 1\,\text{mV}.$$

Example 5.5. Find the resonant frequency f_0 for the circuit in Fig. 5.6 with the following data: $r_i = 0$, $R = 1\,\text{m}\Omega$, $L = 4.708\,\text{nH}$ and $C = 100\,\text{nF}$. Calculate impedance Z at: (a) 1 kHz; (b) 7.335 MHz; and (c) 1 GHz.

Solution 5.5. The resonant frequency is

$$f_0 = \frac{1}{2\pi\sqrt{LC}} = \frac{1}{2\pi\sqrt{4.708\,\text{nH}\,100\,\text{nF}}} \approx 7.335\,\text{MHz}.$$

(a) At 1 kHz:

$$X_\text{L} = 2\pi f L = 2\pi\,1\,\text{kHz}\,4.708\,\text{nH} = 29.581\,\mu\Omega,$$
$$X_\text{C} = \frac{1}{2\pi f C} = \frac{1}{2\pi\,1\,\text{kHz}\,100\,\text{nF}} = 1.592\,\text{k}\Omega,$$

therefore,

$$Z = \sqrt{R^2 + (X_\text{L} - X_\text{C})^2} = \sqrt{1\,\text{m}\Omega^2 + (29.581\,\mu\Omega - 1.592\,\text{k}\Omega)^2} = 1.592\,\text{k}\Omega.$$

(b) At 7.335 MHz: this is the resonant frequency, hence $Z = R = 1\,\text{m}\Omega$.
(c) At 1 GHz: $Z \approx X_\text{L} = 29.581\,\Omega$.

5.2.2 *Parallel RLC Network*

A parallel RLC network has a few subtle differences from the serial version which was discussed in the previous paragraphs. It also represents a frequency controlled impedance which has the same expression for the resonant frequency. However, the impedance behaves slightly differently.

In a parallel RLC circuit (see Fig. 5.8), the voltage is equal across all three components, where each component defines the branch whose current is described by the Ohm's law and each component keeps its own voltage, current, and phase relationship. For a parallel RLC network, the total current I_tot is written as

$$I_\text{tot} = \sqrt{I_\text{R}^2 + (I_\text{L} - I_\text{C})^2}. \tag{5.41}$$

By inspection of Fig. 5.8 (left) and examining the two extreme cases, at DC and very high frequency, we easily conclude the following. At DC, the inductor has zero impedance (i.e., it becomes a short connection), the resistor holds the R value, and the capacitor has infinite impedance (i.e., it becomes an open connection). The three components are in parallel, hence the equivalent impedance is zero. At very high frequencies (i.e., $\omega \to \infty$), the inductor has infinite impedance (i.e., it becomes an open connection), the resistor still holds R, and the capacitor has zero impedance (i.e., it becomes a short connection). Again, the equivalent impedance is zero. This behaviour implies the existence of at least one maxima between the two extreme points (Fig. 5.8). For the time being, let us stay with this conclusion and only roughly plot the frequency dependence of the impedance, as shown in Fig. 5.8 (right).

The following example illustrates how the three components in parallel share the current branches.

Example 5.6. For the circuit in Fig. 5.8 (left) and component values of $r_i = 0$, $V_\text{in} = 12\,\text{V}$, $R = 400\,\Omega$, $X_\text{L} = 500\,\Omega$ and $X_\text{C} = 200\,\Omega$, estimate the total current I_tot supplied by the source and the circuit impedance Z_tot.

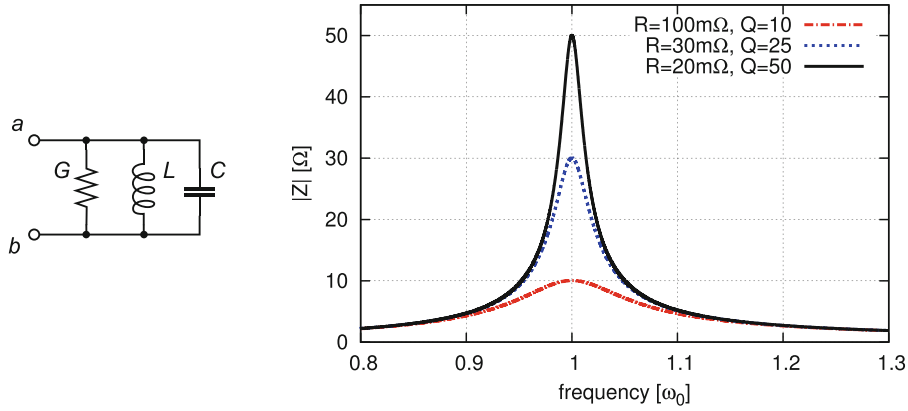


Fig. 5.8 Parallel RLC circuit network, $G = 1/R$, (*left*) and the plot of impedance $|Z_{ab}|$ against frequency, normalized at $\omega_0 = 1\,Hz$, (*right*). The maximum impedance value is $Z_{\omega_0} = Q^2 R$

Solution 5.6. The three branch currents are:

$$I_R = \frac{V_{in}}{R} = \frac{12\,V}{100\,\Omega} = 30\,mA; \quad I_L = \frac{V_{in}}{X_L} = \frac{12\,V}{500\,\Omega} = 24\,mA,$$
$$I_C = \frac{V_{in}}{X_C} = \frac{12\,V}{200\,\Omega} = 60\,mA.$$

Then, from (5.83) it follows that

$$I_{tot} = \sqrt{I_R^2 + (I_L - I_C)^2} = \sqrt{(30\,mA)^2 + (24\,mA - 60\,mA)^2} = 46.862\,mA,$$
$$\therefore$$
$$Z_{tot} = \frac{V_{in}}{I_{tot}} = \frac{12\,V}{46.862\,mA} = 256.07412\,\Omega.$$

We note that the current through the capacitive branch is larger than the total current I_{tot} provided by the signal generator. In the following sections, we explore this phenomenon in more detail.

5.3 Q Factor

It is now logical to ask questions about the behaviour of LC oscillating circuits. What can they be used for? When is the voltage amplitude of the sinusoidal waveform at the maximum? When is it at the minimum? To answer these questions and gain more understanding of the phenomenon, we study two simple networks: serial and parallel RLC circuits. Both of them exhibit the same oscillatory behaviour with slight differences, and it is important to study these two types of circuits in order to understand how can we use them efficiently. Before we proceed, let us first introduce one of the most important parameters in RF circuit design, the Q factor, through a more general definition than the one used in Sect. 5.1.1.

The Q factor is the ratio of the energy stored in the resonator and the energy supplied by a generator, i.e., it is evaluated each cycle as

$$Q = 2\pi \times \frac{\text{Energy Stored}}{\text{Energy dissipated per cycle}} = \omega_0 \times \frac{\text{Energy Stored}}{\text{Power Loss}}, \tag{5.42}$$

where, in electrical systems, the stored energy is the sum of energies initially stored in lossless inductors and capacitors and the lost energy is the sum of the energies dissipated in resistors per cycle.

In the ideal case, energy stored in the magnetic field of the inductor is eventually converted without loss into energy of the electrostatic field of the capacitor. At the resonant frequency, the maximum energy stored in the network keeps bouncing back and forth between the inductor and the capacitor without loss and, therefore, is calculated either at the moment when the capacitor is fully discharged (and therefore the inductor holds the full amount of energy W_L) or when the capacitor is fully charged (and therefore temporarily holds the full amount of the energy W_C), i.e.

$$W_\mathrm{L} = \int_0^T v(t)i(t)\,\mathrm{d}t = \int_0^T i(t)L\frac{\mathrm{d}i(t)}{\mathrm{d}t}\,\mathrm{d}t = L\int_0^{I_\mathrm{p}} i\,\mathrm{d}i = \frac{1}{2}LI_\mathrm{p}^2 = LI_{RMS}^2, \tag{5.43}$$

or, similarly,

$$W_\mathrm{C} = \int_0^T v(t)i(t)\,\mathrm{d}t = \int_0^T v(t)C\frac{\mathrm{d}v(t)}{\mathrm{d}t}\,\mathrm{d}t = C\int_0^{V_\mathrm{p}} v\,\mathrm{d}v = \frac{1}{2}CV_\mathrm{p}^2 = CV_{RMS}^2, \tag{5.44}$$

where $I_\mathrm{p} = \sqrt{2}I_\mathrm{max}$ is the peak current through the inductor, and $V_\mathrm{p} = \sqrt{2}V_\mathrm{max}$ is the peak voltage across the capacitor.

The energy dissipated in the resistor W_R during one full resonant cycle

$$T_0 = \frac{1}{f_0} = \frac{2\pi}{\omega_0} \tag{5.45}$$

is simply, by definition, power multiplied by the time, i.e.

$$W_\mathrm{R} = P_\mathrm{R} \times T_0 = RI_{RMS}^2 \times T_0 = \frac{2\pi}{\omega_0}RI_{RMS}^2, \tag{5.46}$$

which means that (5.42) becomes (using either W_L or W_C) for serial RLC

$$Q_\mathrm{s} = 2\pi\frac{W_\mathrm{L}}{W_\mathrm{R}} = 2\pi\frac{LI_{RMS}^2}{2\pi/\omega_0 RI_{RMS}^2} = \frac{\omega_0 L}{R}. \tag{5.47}$$

At resonance, the resonant frequency ω_0, inductance L, and capacitance C are related as in (5.39), therefore the three equivalent formulations of Q_s are

$$\omega_0 = \frac{1}{\sqrt{LC}} \quad \therefore \quad Q_\mathrm{s} = \frac{\omega_0 L}{R} = \frac{1}{\omega_0 RC} = \frac{1}{R}\sqrt{\frac{L}{C}}, \tag{5.48}$$

where (5.48) shows all three variants of the expression for the Q factor of an RLC network in series. Expressions (5.47) and (5.48) for the quality factor Q are very important and are used to quantify a number of specifications in radio design.

It is important to note that for the ideal inductor, i.e., $R = 0$, the Q factor becomes $Q = \infty$. It is desirable to keep control over the Q factor for many reasons that are mentioned throughout this book. In addition, it should be noted that in serial configurations, the Q factor is inversely proportional to the resistance R.

Example 5.7. For a typical serial RLC network, $L = 1\,\text{mH}$, $C = 25.33\,\text{pF}$, and the total resistance at the resonant frequency in the loop is $R = 15\,\Omega$. Estimate the resonant frequency and the Q factor.

Solution 5.7. The resonant frequency is

$$f_0 = \frac{1}{2\pi\sqrt{(LC)}} \approx 1\,\text{MHz}$$

and the Q factor is

$$Q = \frac{1}{R}\sqrt{\frac{L}{C}} = \frac{1}{15\,\Omega}\sqrt{\frac{1\,\text{mH}}{25.33\,\text{pF}}} \approx 420,$$

which are typical numbers in the current state of the art.

5.3.1 Q Factor of a Serial RLC Network

In a serial RLC circuit (see Example 5.3), we found that voltage across the inductor was magnified a number of times relative to the voltage provided by the source (i.e., the total voltage at the network terminals). It should be noted that this effect is similar to what happens in a transformer, i.e., only the voltage is amplified, but not the current at the same time. That is, passive networks are not capable of power amplification.

This phenomenon of amplification of either the voltage or the current is very important for the resonant behaviour, so we should take a closer look and try to quantify it. Let us take, for instance, the serial RLC network in Fig. 5.6. Under the condition of resonance, the circuit reactance equals zero, which means that the input voltage v_{ab} must be equal to the voltage v_R across the resistor. The voltage amplification is then calculated as the "output" voltage, in this case v_L, divided by the "input" voltage, in this case v_{ab}. Voltage is calculated as a product of current and impedance; in a serial circuit there is only one branch current, therefore the voltage gain Q_s is

$$Q_\text{s} = \frac{v_\text{L}}{v_\text{in}} = \frac{v_\text{L}}{v_\text{R}} = \frac{iX_\text{L}}{iR} = \frac{\omega L}{R} = \frac{1}{\omega RC} = \frac{1}{R}\sqrt{\frac{L}{C}}. \tag{5.49}$$

The variable Q_s is commonly used as a symbol for the kind of amplification that happens inside a resonating serial RLC network. The last two terms in (5.49) are derived after substituting $X_\text{L} = X_\text{C}$ (which is true at resonance, see (5.5)). In (5.49), it is assumed that the Q factor is larger than approximately ten. In the following sections, we justify this assumption.

A couple of observations are in order. First, in an ideal network where the serial resistance $R = 0$ the implication is that $Q \to \infty$. In other words, in order to increase the Q_s factor of a serial RLC network, we need to reduce the total internal resistance or increase inductive reactance. Second, real resistance R causes thermal dissipation (i.e., loss) of energy, while the total energy contained in the circuit initially was stored in the inductor (or the capacitor, for that matter). Thus, we also say that the Q factor represents the ratio of the total energy and the dissipated (i.e., lost) energy in the circuit.

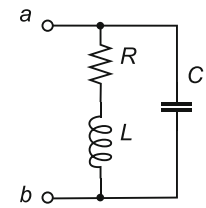


Fig. 5.9 Realistic parallel LC network

5.3.2 Q Factor of a Parallel RLC Network

We now find the resonant frequency ω_{p0} of a realistic parallel LC network, Fig. 5.9, where the resistance R accounts for all thermal losses, i.e., the combined resistance of the inductor and wires and the effective series resistance (ESR) of the capacitor.

$$Y(\omega) = \frac{1}{R+\mathrm{j}\omega L} + \mathrm{j}\omega C = \frac{R-\mathrm{j}\omega L}{R^2+(\omega L)^2} + \mathrm{j}\omega C$$
$$= \frac{R}{R^2+(\omega L)^2} + \mathrm{j}\left(\omega C - \frac{\omega L}{R^2+(\omega L)^2}\right). \tag{5.50}$$

At resonance (i.e., $\omega = \omega_{p0}$), the two reactances are equal $|Z_L| = |Z_C|$, therefore the imaginary part is $\Im(Y) = 0$, hence we write

$$\omega_{p0}C = \frac{\omega_{p0}L}{R^2+(\omega_{p0}L)^2} \quad \therefore \quad R^2+(\omega_{p0}L)^2 = \frac{L}{C}, \tag{5.51}$$

which leads to the conclusion,

$$\omega_{p0} = \sqrt{\frac{1}{LC} - \frac{R^2}{L^2}}. \tag{5.52}$$

We conclude that the resonant frequency ω_{p0} of a parallel LC network that includes realistic inductance has the additional term $(R/L)^2$ due to the finite wire resistance, which slightly reduces the resonant frequency relative to the case of ideal LC resonator. When $R \to 0$, (5.52) becomes the same as (5.39) for the ideal LC resonator, i.e., $\omega_{p0} \to \omega_0$.

Example 5.8. For typical RLC components, $L = 1\,\mathrm{mH}$, $C = 25.33\,\mathrm{pF}$, and $R = 15\,\Omega$, find by how much the resonant frequency of the realistic resonator is off relative to the ideal resonator.

Solution 5.8. The ideal resonant frequency is simply

$$\omega_0 = \frac{1}{\sqrt{LC}} = \frac{1}{\sqrt{1\,\mathrm{mH}\times 25.33\,\mathrm{pF}}} = 1.00000584\,\mathrm{MHz},$$

while the realistic resonant frequency is calculated as

$$\omega_0 = \sqrt{\frac{1}{LC} - \frac{R^2}{L^2}} = \sqrt{\frac{1}{1\,\mathrm{mH}\times 25\,\mathrm{pF}} - \frac{(15\,\Omega)^2}{(1\,\mathrm{mH})^2}} = 1.00000299\,\mathrm{MHz}, \tag{5.53}$$

therefore, the difference of 2.85 Hz relative to 1×10^6 Hz is negligible for most practical purposes.

For parallel RLC configuration, as shown in Fig. 5.8 (left), however, it is desirable to have R as high as possible in order to reduce the power dissipation (i.e., to reduce the current through the R branch of the RLC network), which is to say that, using the principle of duality, the three equivalent quality factor Q_{p} formulations for a parallel RLC network are

$$Q_{\mathrm{p}} = \frac{R}{\omega_0 L} = \omega_0 RC = R\sqrt{\frac{C}{L}}. \tag{5.54}$$

To elaborate the point, it is useful to evaluate the size of the difference between resonant frequencies of series and parallel RLC networks. We already derived expression (5.52) for the resonant frequency ω_{p0} of a parallel RLC network, which can be reformulated as

$$\omega_{\mathrm{p0}} = \sqrt{\frac{1}{LC} - \frac{R^2}{L^2}} = \sqrt{\omega_{\mathrm{s0}}^2 - \frac{R^2}{L^2}} = \omega_{\mathrm{s0}}\sqrt{1 - \frac{R^2}{\omega_{\mathrm{s0}}^2 L^2}} = \omega_{\mathrm{s0}}\sqrt{1 - \frac{1}{Q_s^2}}, \tag{5.55}$$

$$\therefore$$

$$\omega_{\mathrm{p0}} \approx \omega_{\mathrm{s0}} \quad \text{for} \quad (Q_{\mathrm{s}} > 10), \tag{5.56}$$

where we appropriately introduced series resonant frequency ω_{s0} through the serial Q_{s} factor. Equation (5.55) shows that for ideal or high Q networks (i.e., $Q > 10$) there is a very small error in calculating resonating frequencies ω_{s0} and ω_{p0} of the series and parallel circuits. Hence, they can be used interchangeably, as long as the Q factor is high.

To simplify the calculation, for high Q values, we assume that $(\omega L)^2 \gg R^2$ (this is justified because an inductor's wire resistance is relatively small) or, equivalently, the same condition is written as $R^2 + (\omega L)^2 \cong (\omega L)^2$. Therefore, at resonance the admittance is resistive, which after applying condition (5.56) to the real part of (5.50) yields

$$Y_0 = \frac{R}{R^2 + (\omega_0 L)^2} \approx \frac{R}{(\omega_0 L)^2} = \frac{CR}{L} \quad \therefore \quad R_{\mathrm{D}} = \frac{1}{Y_0} = \frac{L}{CR}, \tag{5.57}$$

where R_{D} now represents the dynamic resistance of the LC tank at resonance.

After having delivered expressions for both non-resonant admittance Y and admittance at resonance Y_0, it becomes straightforward to find out how LC tank admittance changes with frequency, relative to its resonant value. Then, from (5.50) and (5.57) we write

$$\frac{Y}{Y_0} \cong \frac{L}{RC}\left[\frac{R}{(\omega L)^2} + \mathrm{j}\left(\omega C - \frac{1}{\omega L}\right)\right] \tag{5.58}$$

$$= \frac{1}{\omega^2 LC} + \mathrm{j}\left(\frac{\omega L}{R} - \frac{1}{\omega CR}\right) \tag{5.59}$$

$$= \frac{\omega_0^2}{\omega^2} + \mathrm{j}\,\delta\, Q, \qquad \text{where,} \qquad \delta = \frac{\omega}{\omega_0} - \frac{\omega_0}{\omega} \tag{5.60}$$

after substituting (5.49) and rearranging (for high Q factor, the serial resonance ω_{s0} and parallel resonance ω_{p0} are equal). Therefore,[3]

$$|Y| = Y_0\sqrt{\left(\frac{\omega_0}{\omega}\right)^4 + (\delta\, Q)^2}. \tag{5.61}$$

[3] That is, $|Z| = \sqrt{\Re(Z)^2 + \Im(Z)^2}$.

Result (5.61) is an important relation that is used to estimate the amplitude of a signal located not exactly at the resonant frequency. More applications of this formula are shown in Sect. 9.6.1.

5.4 Self-resonance of an Inductor

As implied in the previous paragraphs, real inductors show characteristic properties of a complex RLC circuit. Based on the analysis of resonance, and knowledge that non-ideal inductors have parasitic capacitances related to the wire, it should be intuitively obvious that the circuit diagram shown in Fig. 5.9, could also be used to represent a real inductor by itself. Of course, it is only one of the possible ways to create a model of real inductor; it is also one of the most often used models. Following the same procedure as in the previous sections, and after applying the low wire resistance approximation, an expression for the admittance of a non-ideal inductor by itself (see Fig. 5.9) is found as

$$\begin{aligned} Y_{\mathrm{L}} &= \frac{1}{R_{\mathrm{L}}+\mathrm{j}\omega L}+\mathrm{j}\omega C_{\mathrm{L}} \cong \frac{R_{\mathrm{L}}}{(\omega L)^2}+\mathrm{j}\left(\omega C_{\mathrm{L}}-\frac{1}{\omega L}\right) \\ &= \frac{R_{\mathrm{L}}}{(\omega L)^2}-\mathrm{j}\left(\frac{1-\omega^2 LC_{\mathrm{L}}}{\omega L}\right) = \Re(Y_{\mathrm{L}})+\mathrm{j}\,\Im(Y_{\mathrm{L}}). \end{aligned} \tag{5.62}$$

Development of the non-ideal inductor model makes the concept of a *self-resonant frequency* easier to accept. Because of the $R_{\mathrm{L}}C_{\mathrm{L}}L$ component values associated with physical realization of the inductor, it becomes obvious that, based on the knowledge of resonance, the non-ideal inductor does have resonant frequency $\omega_{0\mathrm{L}}$ of its own. It is very important for a designer to have at least some estimate of where this self-resonant frequency might be. Typically, the wire resistance is $R_{\mathrm{L}} \leq 1\,\Omega$ and associated parasitic capacitance C_{L} is in the order of pF, which means that the self-resonant frequency is, typically, in the order of megahertz to a few hundreds of megahertz. That is, if a non-ideal inductor is to be used in an LC tank, the designer is forced to limit the intended signal frequency to no more than one decade (i.e., ten times) below the inductor's self-resonant frequency. This rule of thumb is most often used as a measure of how good an inductor is needed for the intended design.

A natural question one could ask would be why an external capacitor C is needed in parallel with a non-ideal inductor to set the resonant frequency. Why not just use the non-ideal inductor alone? This approach would simplify things, at least in terms of the component count. Indeed, that approach is used for the design of circuits working at very high frequencies, for example in satellite communication systems. However, detailed analysis of such components and circuits is beyond the scope of this book. For purposes of designing circuits with discrete components working at frequencies in the order of up to a few 100 MHz, one should keep in mind that controlling the inductor's parasitic capacitance is not practical. Therefore, a more practical LC model is used (Fig. 5.9).

The first point to note is that all wire resistances r_{W} associated with the practical LC model are now merged with inductive resistance R_{L}. The effective parallel resistance R_{p} is represented by the real part of (5.62), i.e.

$$\frac{1}{R_{\mathrm{p}}} = \frac{R_{\mathrm{L}}}{(\omega L)^2} \qquad \therefore \qquad R_{\mathrm{p}} = \frac{(\omega L)^2}{R_{\mathrm{L}}}, \tag{5.63}$$

while the effective parallel inductance of the coil L_{eff} is represented by the imaginary part of (5.62), i.e.

$$\frac{1}{\omega L_{\mathrm{eff}}} = \frac{1-\omega^2 LC_{\mathrm{L}}}{\omega L},$$

$$\therefore$$

$$L_{\text{eff}} = \frac{L}{1-\omega^2 LC_{\text{L}}} = \frac{L}{1-(\omega/\omega_{0\text{L}})^2}. \tag{5.64}$$

Next, we estimate the deviation from the ideal LC tank model at resonance ω_0 (which has to be at least one decade below the self-resonance $\omega_{0\text{L}}$ of the coil). Another way of stating this condition is that the resonator's external capacitor C_{T} has to be much larger than the parasitic capacitance C_{L}, i.e., $C_{\text{T}} \gg C_{\text{L}}$, (remember, the ideal inductance L is always the same). At circuit resonance ω_0, the dynamic resistance R_{D} of the LC tank in Fig. 5.9, as defined in (5.81) and (5.82), is

$$R_{\text{D}} = Q\,\omega_0 L. \tag{5.65}$$

At the same time, the dynamic resistance R_{D} of the equivalent circuit is described using effective values

$$R_{\text{D}} = Q_{\text{eff}}\,\omega_0 L_{\text{eff}}. \tag{5.66}$$

Because of the equivalence of these two circuits and from (5.65) and (5.66) it follows that

$$Q\,\omega_0 L = Q_{\text{eff}}\,\omega_0 L_{\text{eff}}, \tag{5.67}$$

$$\therefore$$

$$Q_{\text{eff}} = Q\left(1-\omega^2 LC_{\text{L}}\right) = Q\left[1-\left(\frac{\omega}{\omega_{0\text{L}}}\right)^2\right], \quad (\omega \ll \omega_{0\text{L}}) \tag{5.68}$$

where it is assumed $(\omega \ll \omega_{0\text{L}})$, and $Q = \omega_0 L/R$, as defined in (5.49). This result shows that the effective Q factor Q_{eff} of a realistic LC tank decreases as the operating frequency ω approaches the self–resonating frequency $\omega_{0\text{L}}$.

How do we use the above results? Well, it depends. For the parallel circuit case in Fig. 5.9, the dynamic resistance can be calculated using either (5.65) or (5.66). The general definitions from (5.81) and (5.82) can be used as well, as long as either C is replaced with $(C+C_{\text{L}})$ or Q is replaced with Q_{eff}. The bandwidth Δf from (5.69), however, must be calculated using Q and not Q_{eff} because, for a given resonant frequency, capacitance C is adjusted to absorb C_{L}, because they are in parallel connection.

Note that a serial tuned circuit is different, i.e., capacitor C is in serial connection with the inductive capacitance C_{L}. Effectively, C resonates with L_{eff} which means that instead of using Q, Q_{eff} is used, so that

$$Q_{\text{eff}} = \frac{f_0}{\Delta f}. \tag{5.69}$$

Those little differences should be accounted for when analyzing serial and parallel RLC resonant circuits.

5.5 Serial to Parallel Impedance Transformations

Often, it is useful to transform a serial RLC network into its equivalent parallel configuration or vice versa. This transformation must be done only at a single frequency, which does not affect the serial and parallel Q factor of the networks,

$$Q_{\text{s}} = \frac{X_{\text{s}}}{R_{\text{s}}}, \tag{5.70}$$

$$Q_{\text{p}} = \frac{R_{\text{p}}}{X_{\text{p}}}, \tag{5.71}$$

so that, assuming $Q_s = Q_p = Q$ at the given frequency

$$Z_s = R_s + jX_s = R_s + jQ_sR_s = R_s(1 + jQ_s), \tag{5.72}$$

$$Y_p = \frac{1}{Z_s} = \frac{1}{R_s(1+jQ)} = \frac{1}{R_s(1+jQ)}\frac{1-jQ}{1-jQ} \tag{5.73}$$

$$= \frac{1}{R_s(1+Q^2)} - j\frac{Q}{R_s(1+Q^2)} \tag{5.74}$$

$$= \frac{1}{R_s(1+Q^2)} - j\frac{Q}{\frac{X_s}{Q}(1+Q^2)} \tag{5.75}$$

$$= \frac{1}{R_p} - j\frac{1}{X_p}, \tag{5.76}$$

$$\therefore$$

$$R_p = R_s(1+Q^2), \tag{5.77}$$

$$X_p = X_s\left(1 + \frac{1}{Q^2}\right), \tag{5.78}$$

after replacing (5.70) and (5.71) in (5.72)–(5.75). Again, for large Q, i.e., $Q > 10$,

$$R_p \approx Q^2 R_s, \tag{5.79}$$

$$X_p \approx X_s. \tag{5.80}$$

The last two expressions are often-used approximations in resonant circuit network analysis.

5.6 Dynamic Resistance

The imaginary part $\Im(Y)$ determined the resonant frequency, while the real part $\Re(Y)$, from (5.50), determines the dynamic resistance R_D, i.e., real resistance of the LC resonator at the resonant frequency, as,

$$\Re(Y(\omega_{p0})) = \frac{R}{R^2 + (\omega_{p0}L)^2}$$

$$= \frac{R}{R^2 + \left[\sqrt{\frac{1}{LC} - \frac{R^2}{L^2}}\right]^2 L^2}$$

$$= \frac{R}{R^2 + \left(\frac{L}{C} - R^2\right)}$$

$$= \frac{RC}{L},$$

$$\therefore$$

$$R_D = \frac{1}{\Re(Y(\omega_{p0}))} = \frac{L}{RC}. \tag{5.81}$$

In the ideal case, i.e., $R = 0$, the dynamic resistance of the LC resonator (see Fig. 5.9) becomes $R_D = \infty$. It should be noted that, from the perspective of the resonant current, which circulates inside the RLC loop, the three elements are in series. Hence, reducing the resistance associated with the inductive branch is desirable in order to increase the dynamic impedance perceived by the network external to the RLC resonator.

Finally, the expression for dynamic resistance (5.81) can also be reformulated in terms of the Q factor ($Q > 10$), after using (5.48), as:

$$R_D = \frac{L}{RC} = \omega_0 LQ = \frac{Q}{\omega_0 C} = Q^2 R, \tag{5.82}$$

which is, again, the resistance of a realistic RLC tank in resonance, as perceived by the external network. An important distinction to make is that the resistance R is a physical entity: in a serial RLC network, it needs to be as small as possible; in a parallel configuration, it needs to be as large as possible. However, at resonance, this small resistance is perceived by the external network as if magnified by the Q^2 factor. In the ideal case, i.e., when serial the resistance $R = 0$, value of the Q factor becomes infinity. Hence, expression (5.82) is only a mathematical approximation, see Fig. 5.8.

The maximum impedance value happens at the resonant frequency ω_0, as shown in Fig. 5.8 (right), however this time it is calculated as

$$Z_{max} = Z(\omega_0) = Q^2 R. \tag{5.83}$$

This is a very important property of parallel RLC networks, which indicates that, at resonance, an ideal LC parallel network (i.e., $R = \infty$) would have infinite Q factor and, therefore, infinite voltage output. One should note the obvious risk associated with the possible destruction of components used in high Q resonators. An ideal parallel RLC network is used to suppress all frequencies except the resonant frequency, Fig. 5.8 (right).

5.7 General RLC Networks

A truly realistic model of an LC resonator must include losses of both the inductor and capacitor, modelled by effective series resistance ESR and resistor r_1 (see Fig. 5.10). In this section, we derive expressions for the resonant frequency ω_0 and dynamic resistance R_D of a general LC circuit as the function of Q_1 and Q_2 factors of the inductor and capacitor respectively. Finally, we show how to transform the resonator in Fig. 5.10 into its equivalent parallel RLC network, assuming that the capacitor is lossless, i.e., $ESR = 0$.

By definition, Q factors of inductive and capacitive branches in the LC network (after substitution $ESR = r_2$) at resonant frequency ω_0 are:

$$Q_1 \equiv \frac{X_L}{r_1} = \frac{\omega_0 L}{r_1} = \tan\theta_1 \quad \therefore \quad \theta_1 = \arctan Q_1, \tag{5.84}$$

$$Q_2 \equiv \frac{X_C}{r_2} = \frac{1}{\omega_0 C r_2} = \tan\theta_2 \quad \therefore \quad \theta_2 = \arctan Q_2, \tag{5.85}$$

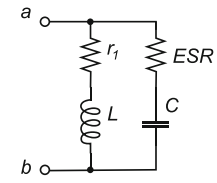


Fig. 5.10 Realistic parallel LC network

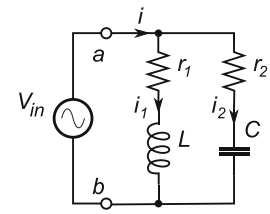


Fig. 5.11 Realistic LC resonator driven by the external signal source V_{in}

where θ_1 and θ_2 are the respective phase angles in the inductor and capacitor due to the thermal losses (resistances $r_{1,2}$ denote the internal resistances of the coil and the ESR of the capacitor, respectively). We also define, after including (5.48),

$$Z_1 = r_1 + \mathrm{j}\omega_0 L = \frac{\omega_0 L}{Q_1} + \mathrm{j}\omega_0 L,$$

$$\therefore$$

$$|Z_1| = \sqrt{\left(\frac{\omega_0 L}{Q_1}\right)^2 + (\omega_0 L)^2} = \omega_0 L\sqrt{1 + \frac{1}{Q_1^2}}, \tag{5.86}$$

as well as,

$$Z_2 = r_2 + \frac{1}{\mathrm{j}\omega_0 C} = \frac{1}{Q_2\omega_0 C} + \frac{1}{\mathrm{j}\omega_0 C}, \tag{5.87}$$

therefore,

$$|Z_2| = \sqrt{\frac{1}{(Q_2\omega_0 C)^2} + \frac{1}{(\omega_0 C)^2}} = \frac{1}{\omega_0 C}\sqrt{1 + \frac{1}{Q_2^2}}. \tag{5.88}$$

From (5.84) and (5.85), in addition to straightforward application of trigonometric identities,[4,5] we write

$$\sin\theta_1 = \frac{Q_1}{\sqrt{1+Q_1^2}}\ ;\ \cos\theta_1 = \frac{1}{\sqrt{1+Q_1^2}}, \tag{5.89}$$

$$\sin\theta_2 = \frac{Q_2}{\sqrt{1+Q_2^2}}\ ;\ \cos\theta_2 = \frac{1}{\sqrt{1+Q_2^2}}. \tag{5.90}$$

5.7.1 *Derivation for the Resonant Frequency* ω_0

If an AC voltage source $V_{\text{in}} = V\cos\theta_1$ is connected to the resonator (see Fig. 5.11), the total current needed to compensate for the thermal losses $i = i_1 + i_2$ is split between the two branches.

[4] $\cos[\arctan x] = 1/\sqrt{1+x^2}$.
[5] $\sin[\arctan x] = x/\sqrt{1+x^2}$.

The inductive branch current i_1 has two components: one that is in phase with the source voltage V_{in}, i.e., $(V\cos\theta_1)/Z_1$, and one that is lagging by 90°, $(V\sin\theta_1)/Z_1$. At the same time, the capacitive branch current i_2 also has two components: one that is in phase with the source voltage V_{in}, i.e., $(V\cos\theta_2)/Z_2$, and one that is leading the source voltage V_{in} by 90°, i.e., $(V\sin\theta_2)/Z_2$.

At resonance, the two quadrature current components must be opposite and equal (so that the vector sum is zero), which leads to the following expressions (after using results (5.86) to (5.90))

$$(V\sin\theta_1)\frac{1}{Z_1} = (V\sin\theta_2)\frac{1}{Z_2}, \tag{5.91}$$

therefore,

$$\begin{aligned}
&\frac{Q_1}{\sqrt{1+Q_1^2}}\frac{1}{\omega_0 L\sqrt{1+\frac{1}{Q_1^2}}} = \frac{Q_2}{\sqrt{1+Q_2^2}}\frac{\omega_0 C}{\sqrt{1+\frac{1}{Q_2^2}}},\\
&\frac{Q_1}{\sqrt{\left(1+Q_1^2\right)\left(1+\frac{1}{Q_1^2}\right)}} = \frac{Q_2}{\sqrt{\left(1+Q_2^2\right)\left(1+\frac{1}{Q_2^2}\right)}}\omega_0^2 LC,
\end{aligned} \tag{5.92}$$

where both the left and right side of (5.92) contain algebraic terms that can be simplified as follows:

$$\begin{aligned}
\frac{x}{\sqrt{(1+x^2)(1+\frac{1}{x^2})}} &= \sqrt{\frac{x^2}{(1+x^2)(1+\frac{1}{x^2})}} = \sqrt{\frac{x^2}{x^2+2+\frac{1}{x^2}}}\\
&= \sqrt{\frac{x^2}{\left(x+\frac{1}{x}\right)^2}} = \frac{x}{x+\frac{1}{x}} = \frac{1}{1+\frac{1}{x^2}}.
\end{aligned} \tag{5.93}$$

Using (5.93) it is straightforward to rewrite (5.92) as:

$$\begin{aligned}
&\frac{1}{1+\frac{1}{Q_1^2}} = \frac{1}{1+\frac{1}{Q_2^2}}\omega_0^2 LC,\\
&\therefore\\
&\omega_0 = \frac{1}{\sqrt{LC}}\sqrt{\frac{1+\frac{1}{Q_2^2}}{1+\frac{1}{Q_1^2}}} \approx \frac{1}{\sqrt{LC}} \qquad (Q_{1,2} \gg 1),
\end{aligned} \tag{5.94}$$

which is the solution for the resonant frequency of an LC resonator with a non-ideal inductor and a non-ideal capacitor. Naturally, for very good L and C components the thermal losses are negligible, in other words $Q_{1,2} \gg 1$, hence (5.94) can be approximated with the expression for the resonant frequency ω_0 that was defined earlier for the case of ideal LC resonator. However, note that the assumption of high Q is not always valid. A very dramatic example is the case of the on-chip inductors manufactured in the standard CMOS process that are used in modern wireless devices; their Q is in the order of five. Consequently, additional design and technological techniques must be employed to improve the performance of integrated LC resonators.

5.7.2 Derivation for the Dynamic Resistance R_{D}

At resonance, the sum of complex quadrature components of the two branch currents is zero, which leaves only the two in-phase current components. Similarly to the previous derivation, we write,

$$\begin{aligned} i &= V\left(\frac{\cos\theta_1}{Z_1}+\frac{\cos\theta_2}{Z_2}\right) \\ &= V\left[\frac{1}{\sqrt{1+Q_1^2}}\frac{1}{\omega_0 L\sqrt{1+\frac{1}{Q_1^2}}}+\frac{1}{\sqrt{1+Q_2^2}}\frac{\omega_0 C}{\sqrt{1+\frac{1}{Q_2^2}}}\right] \\ &\therefore \\ &= V\left[\frac{Q_1}{\omega_0 L(1+Q_1^2)}+\frac{Q_2\,\omega_0 C}{1+Q_2^2}\right]. \end{aligned} \tag{5.95}$$

It is now convenient to introduce substitution for the $\omega_0 C$ term in (5.95) by rewriting (5.94) as follows:

$$\omega_0^2 LC = \frac{1+\frac{1}{Q_2^2}}{1+\frac{1}{Q_1^2}} \quad \therefore \quad \frac{Q_2^2\,\omega_0 C}{1+Q_2^2} = \frac{Q_1^2}{(1+Q_1^2)\omega_0 L} \quad \therefore \quad \omega_0 C = \frac{1+Q_2^2}{Q_2^2}\frac{Q_1^2}{(1+Q_1^2)\omega_0 L}. \tag{5.96}$$

After these substitutions (5.95) becomes,

$$i = V\left[\frac{Q_1}{\omega_0 L(1+Q_1^2)}+\frac{Q_2}{1+Q_2^2}\frac{1+Q_2^2}{Q_2^2}\frac{Q_1^2}{(1+Q_1^2)\omega_0 L}\right] = V\frac{Q_1}{\omega_0 L(1+Q_1^2)}\left[1+\frac{Q_1}{Q_2}\right], \tag{5.97}$$

which now leads straight into the expression for dynamic resistance R_{D} as

$$R_{\mathrm{D}} \equiv \frac{V}{i} = \omega_0 L\frac{Q_1+\frac{1}{Q_1}}{1+\frac{Q_1}{Q_2}} \tag{5.98}$$

for the case of a non-ideal inductor and a non-ideal capacitor. As is, (5.98) shows the dependence of dynamic resistance versus the Q factors of L and C components. In case of very good (but still not perfect) inductors, i.e., $Q_1 \gg 1$ or in other words $(1/Q_1) \approx 0$, (5.98) can be written as the very close approximation,

$$R_{\mathrm{D}} = \omega_0 L\frac{Q_1}{1+\frac{Q_1}{Q_2}} = \omega_0 L\frac{Q_1 Q_2}{Q_1+Q_2}. \tag{5.99}$$

Modern capacitors are made using very good dielectrics, which is to say that Q_2 is not only large but could be approximated as $Q_2 \to \infty$, in other words $Q_2 \gg Q_1$, in other words $Q_1/Q_2 \approx 0$. Therefore, in case of a lossless capacitor, (5.99) is further approximated as

$$R_{\mathrm{D}} = \omega_0 L Q_1, \tag{5.100}$$

which is commonly used in practice because, in comparison with capacitors, inductors are much harder components to build.

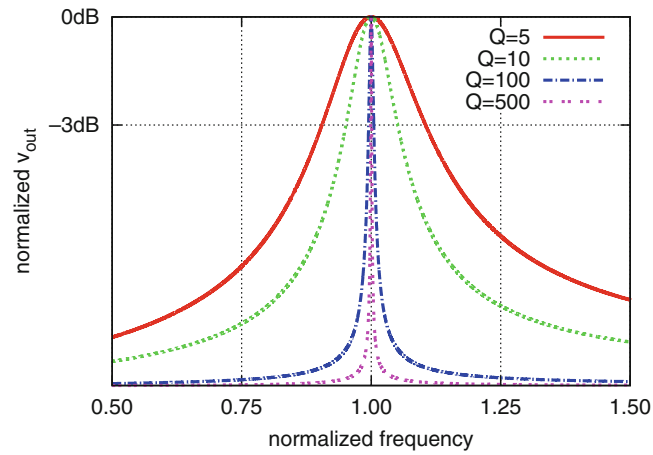


Fig. 5.12 Normalized output voltage across the inductor at normalized resonant frequency $\omega_0 = 1$ for various Q factors

Finally, in the extreme approximation that is good only for fast "back-of-an-envelope" analysis, even the inductor is assumed to be perfectly lossless, i.e., $Q_1 \to \infty$, which means that (5.100) becomes simply

$$R_D \to \infty, \tag{5.101}$$

which is what was concluded earlier, in (5.82).

Expressions (5.98)–(5.101), in addition to (5.109), for dynamic resistance R_D are useful, as long as the applied assumptions are kept in mind.

5.8 Selectivity

The ability of a resonating circuit to select and amplify a weak voltage signal at one specific frequency ω_0 is its core quality used in RF circuits, it is referred to as "selectivity". In the ideal case of $Q \to \infty$, the resonating circuit would pick one and only one frequency, ω_0, while all other tones would be completely suppressed. However, in realistic circuits there is always some finite resistance causing the thermal loss, which is measured by the circuit's finite Q factor. A plot of selectivity curves as a function of Q factor is shown in Fig. 5.12. The plot indicates that, for good selectivity, we need high Q factor resonating circuits.

An interesting question to answer is: In the case of resonators with finite Q, what is the range of frequencies that passes through the resonator without being significantly suppressed? In the following paragraphs, we examine the realistic case of RLC network behaviour regarding this important filtering property.

5.9 Bandpass Filters

Let us consider a serial RLC network from the perspective of voltage source with resistance R driving impedance $Z = \mathrm{j}\,(\omega L - 1/\omega C)$ (see Fig. 5.13). Maximum power transfer, therefore, happens when the source is matched to the load, i.e., $R = |Z|$. Otherwise, at DC the capacitor becomes open while the

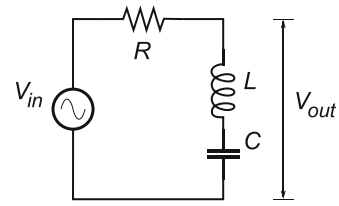


Fig. 5.13 LC network in series

inductor becomes a short connection; and at the other side of the frequency spectrum, at very high frequencies, the capacitor becomes short while the inductor becomes an open connection. In both extreme cases, there is no power transfer because the loop current must drop to zero.

Hence, the condition for maximum power transfer $R = |Z|$, leads to

$$V_{\text{out}} = \frac{V_{\text{in}}}{|R+Z|}|Z| = \frac{V_{\text{in}}}{|R \pm \text{j}R|}R = \frac{V_{\text{in}}}{|1 \pm \text{j}|} = \frac{V_{\text{in}}}{\sqrt{2}}, \tag{5.102}$$

which happens at two frequency points. Let us label them (for the time being) as ω_{U} and ω_{L} (for the "upper" and "lower" frequency, respectively), so that $R = |Z|$ is written as

$$R = \omega_{\text{U}}L - \frac{1}{\omega_{\text{U}}C}, \tag{5.103}$$

$$-R = \omega_{\text{L}}L - \frac{1}{\omega_{\text{L}}C}, \tag{5.104}$$

which, at resonance leads to

$$\begin{aligned} R = \omega_{\text{U}}L - \frac{1}{\omega_{\text{U}}\frac{1}{\omega_0^2 L}} \therefore R = \omega_{\text{U}}L - \frac{\omega_0^2 L}{\omega_{\text{U}}} \therefore \frac{R}{\omega_0 L} = \frac{\omega_{\text{U}}}{\omega_0} - \frac{\omega_0}{\omega_{\text{U}}}, \\ -R = \omega_{\text{L}}L - \frac{1}{\omega_{\text{L}}\frac{1}{\omega_0^2 L}} \therefore -R = \omega_{\text{L}}L - \frac{\omega_0^2 L}{\omega_{\text{L}}} \therefore -\frac{R}{\omega_0 L} = \frac{\omega_{\text{L}}}{\omega_0} - \frac{\omega_0}{\omega_{\text{L}}}. \end{aligned} \tag{5.105}$$

We substitute $Q = \omega_0 L/R$:

$$\frac{1}{Q} = \frac{\omega_{\text{U}}}{\omega_0} - \frac{\omega_0}{\omega_{\text{U}}}, \tag{5.106}$$

$$-\frac{1}{Q} = \frac{\omega_{\text{L}}}{\omega_0} - \frac{\omega_0}{\omega_{\text{L}}}. \tag{5.107}$$

After adding (5.106) and (5.107), it follows that

$$\frac{\omega_{\text{U}}}{\omega_0} + \frac{\omega_{\text{L}}}{\omega_0} = \frac{\omega_0}{\omega_{\text{U}}} + \frac{\omega_0}{\omega_{\text{L}}} \quad \therefore \quad \omega_0^2 = \omega_{\text{U}}\omega_{\text{L}} \tag{5.108}$$

and now, using (5.106) and (5.108), we write

$$\frac{1}{Q} = \frac{\omega_{\text{U}}}{\omega_0} - \frac{\omega_0}{\omega_{\text{U}}} = \frac{\omega_{\text{U}}^2 - \omega_0^2}{\omega_{\text{U}}\omega_0},$$

$$\therefore$$

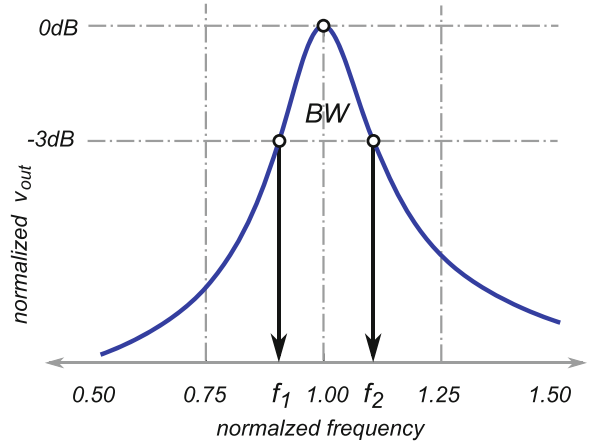


Fig. 5.14 Bandwidth definition plot, where f_1 corresponds to ω_L, and f_2 corresponds to ω_U

$$Q = \frac{\omega_U \omega_0}{\omega_U^2 - \omega_0^2} = \frac{\omega_U \omega_0}{\omega_U^2 - \omega_U \omega_L} = \frac{\omega_0}{\omega_U - \omega_L} = \frac{\omega_0}{\Delta\omega}. \tag{5.109}$$

The last expression is very important, because the two frequencies ω_U and ω_L are used to define the resonator's *bandwidth* BW (see Fig. 5.14). The two frequencies are at -3 dB points relative to the maximum amplitude of the resonator (which is at ω_0). Also, (5.109) shows that a narrow band is achieved by using high Q components.

In serial RLC configuration, high Q also means very low resistance R and high inductance L, which implies that it is good for matching with a low impedance source, such as an antenna, for example, which usually has impedance in the order of 50 Ω. Otherwise, if the source impedance is very high then a parallel RLC configuration must be used, where high Q means high resistance and very low inductance.

Example 5.9. A parallel LC tank consists of: $L = 2.533$ nH with internal wire resistance of $R_L = 1$ mΩ, and $C = 100$ nF. Calculate: (a) the resonant frequency f_0; (b) the Q factor at resonance; (c) the impedance at resonance Z_{max}; and (d) the bandwidth BW.

Solution 5.9.

(a) $$f_0 = \frac{1}{2\pi\sqrt{LC}} = \frac{1}{2\pi\sqrt{2.533\,\text{nH}100\,\text{nF}}} = 10\,\text{MHz}.$$

(b) $$Q = \frac{X_L}{R_L} = \frac{2\pi f_0 L}{R} = \frac{2\pi 10\,\text{MHz}2.533\,\text{nH}}{1\,\text{m}\Omega} = 159.153.$$

(c) $$Z_{max} = Q^2 R_L = (159.153)^2 \cdot 1\,\text{m}\Omega = 29.330\,\Omega.$$

(d) $$BW = \frac{f_0}{Q} = \frac{R_L}{2\pi L} = \frac{1\,\text{m}\Omega}{2\pi 2.533\,\text{nH}} = 62.833\,\text{kHz}.$$

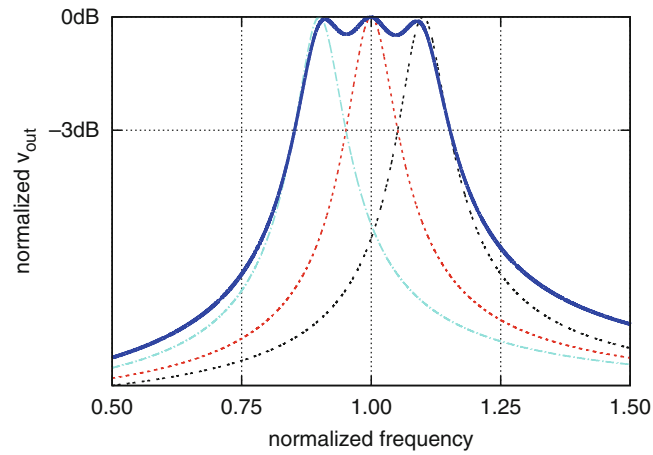


Fig. 5.15 Normalized output voltage across three staggered inductors normalized around the resonant frequency $\omega_0 = 1$ for the same Q factors

5.10 Coupled Tuned Circuit

Although it may seem that increasing the Q factor of a resonating circuit is always desirable, that is not the case. In addition to improving the selectivity of a receiver, an increased Q factor helps with amplification of weak RF signals arriving to the antenna (i.e., with the sensitivity). However, higher Q also reduces the bandwidth, which may start cutting into the frequency content of the signal and, therefore, start introducing distortions.

For example, if a receiver is meant to receive the complete voice frequency spectrum, i.e., 20 Hz–20 kHz, using a 10 MHz carrier signal, then the minimum required bandwidth, calculated according to (5.109), is $Q = f_0/\Delta f = 100\,\text{MHz}/20\,\text{kHz} = 500$. Using a wider bandwidth would not benefit the quality of the received signal; instead, it would allow more noise into the system. If, for whatever reason, a resonator with higher Q is being used, in practical systems it is always possible to widen the overall bandwidth by staggering more than one resonator while maintaining the required sensitivity (see Fig. 5.15). Each of the resonators is tuned to a slightly different resonant frequency and the overall frequency response becomes equal to the sum of the individual responses.

5.11 Summary

In this section, we introduced serial and parallel resonant LC circuits. The LC resonant behaviour is very important for generating voltage and current variables that follow the sinusoidal shape in the time domain. We explored both ideal and realistic cases of LC resonators, and introduced the Q factor as a commonly used measure of internal thermal losses. In the second important use of LC resonators, by controlling the Q factor, we are able to determine the bandwidth of the bandpass LC resonating filter and, therefore, limit the frequency range of single-tone signals that pass through the LC resonator. These two functions are fundamental for RF circuit design and we use both of them.

Problems

5.1. For a given coil, $L = 2\,\mu\text{H}$, $Q = 200$, $f_0 = 10\,\text{MHz}$, calculate:

(a) Its equivalent series resistance.
(b) Its parallel resistance.
(c) The value of the resonating capacitor.
(d) Parallel resistance which, when added, provides bandwidth of 200 MHz.

5.2. A single-tone signal $f_0 = 8\,\text{MHz}$ is passed through a LP RC filter followed by a high-pass RC filter.

(a) Choose R and C values such that the bandwidth around f_0 is $BW = 10\,\text{kHz}$.
(b) What would you choose for $BW = 5\,\text{kHz}$?
(c) Design RLC filters with the same characteristics.

Note: pick component values at your will. They do not have to be the standard values.

5.3. For a given inductor $L = 2.533\,\text{nH}$ and trimming capacitor C whose range is 80 nF to 120 nF, calculate the tuning range ($\Delta f = f_{\max} - f_{\min}$) of this LC resonator.

5.4. Design an LC resonator whose resonant frequency is $f_0 = 10\,\text{MHz}$ and only the following components are available:

(a) $L = 2.533\,\text{nH}$, $C_1 = 10\,\text{nF}$, $C_2 = 40\,\text{nF}$, and $C_3 = 50\,\text{nF}$
(b) $L = 2.533\,\text{nH}$, $C_1 = 20\,\text{nF}$, $C_2 = 30\,\text{nF}$, and $C_3 = 60\,\text{nF}$
(c) $L = 2.533\,\text{nH}$, $C_1 = 70\,\text{nF}$, $C_2 = 60\,\text{nF}$, and $C_3 = 60\,\text{nF}$

5.5. Calculate the Q factor of a serial RLC network if inductor $L = 2.533\,\text{nH}$ and the lumped wire resistance $r = (\pi)\text{m}\Omega$, at: (a) $f_1 = 10\,\text{MHz}$; and (b) $f_2 = 100\,\text{MHz}$.

5.6. For a serial RLC network, derive an expression for bandwidth BW at the resonant frequency ω_0 as a function of Q. What is the conclusion?

5.7. A $1\,\mu\text{H}$ inductive coil has wire resistance of $R = 5\,\Omega$ and self-capacitance of 5 pF. The inductor is used to create an LC resonator with $f_0 = 25\,\text{MHz}$. Calculate the effective inductance and effective Q factor.

5.8. Calculate the resonant frequency of a serial RLC network with $R = 30\,\Omega$, $L = 3\,\text{mH}$, and $C = 100\,\text{nF}$. Calculate its impedance at $f = 10\,\text{kHz}$ and at $f = 5\,\text{kHz}$.

5.9. A frequency response curve of an LC resonator looks as in Fig. 5.14. Assume $f_1 = 450\,\text{kHz}$, $f_2 = 460\,\text{kHz}$, and $f_0 = 455\,\text{kHz}$. Determine the resonator bandwidth, the Q factor, the inductance L if the capacitance is $C = 100\,\text{nF}$, and the total internal circuit resistance R.

5.10. A parallel LC tank consists of $L = 1\,\text{mH}$ whose wire resistance is $R = 1\,\Omega$, and a capacitor $C = 100\,\text{nF}$. Determine the resonant frequency, the Q factor, the dynamic resistance R_D, and the bandwidth of this resonator.

5.11. A serial RC branch consists of $R_\text{S} = 10\,\Omega$ and $C_\text{S} = 7.95\,\text{pF}$. Convert it into its equivalent parallel RC network form at $f = 1\,\text{GHz}$.

Chapter 6
Matching Networks

Abstract The main purpose of an electronic circuit is to process an electronic signal that has arrived at its input terminals. In other words, an abstract mathematical operation that was envisioned during the initial phases of the design is materialized in the form of a physical electronic circuit and its transfer function. The circuit is then expected to modify (i.e., to process) the input signal in accordance with the intended mathematical function and to pass the result to the next stage. In addition, a real, well-designed system should perform the signal-processing operations efficiently with minimal waste of time and energy. Hence, a number of interesting questions arise: What is the most optimal strategy for the energy transfer and signal processing? How should the interface between two subsequent stages be modelled and designed? Is it more beneficial to pass the signal from one stage to another in the form of voltage or in the form of current? Based on what criteria should the decision be made? What happens if it is not possible to achieve the optimal goal and what kind of compromises are appropriate to make? How should we deal with general, more complicated networks? In this chapter, we study a simple basic methodology for interfacing two stages in the signal processing chain that is commonly used in the design of RF electronic systems, with the main criterion being maximum power transfer between the stages. This approach is justified by the argument that wireless RF signals that have arrived at the system input terminals are very weak, thus subsequent power loss would have broad consequences for the overall system performance.

6.1 System Partitioning Concept

It is not difficult to extrapolate analytical methodologies based on Ohm and Kirchhoff's laws to large networks and realize that manual analysis of large systems is not practical because modern electronic systems are synthesized using billions of RLC-equivalent components. Considering that humans need quite a bit of time to manually solve even a relatively simple system of four equations, it becomes obvious that some levels of abstraction must be introduced into the design process. The trick being applied is to "divide and conquer", i.e., to partition the system along the signal flow path in order to create a chain of system-level blocks, and then consider each of the subsystems (blocks) as a stand-alone unit. The same methodology is then applied to each of the stand-alone units until we reach the very end of the chain, which consists only of basic devices.

In accordance with this methodology, a complicated system is virtually always designed hierarchically; most practical systems are split into fewer than ten levels of hierarchy. Once the hierarchy chain is established and each of the stages is replaced by its equivalent *Thévenin* or *Norton model*, at the conceptual level each of the blocks is considered to be a "black box" with input and output terminals (see Fig. 6.1).

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8_6,

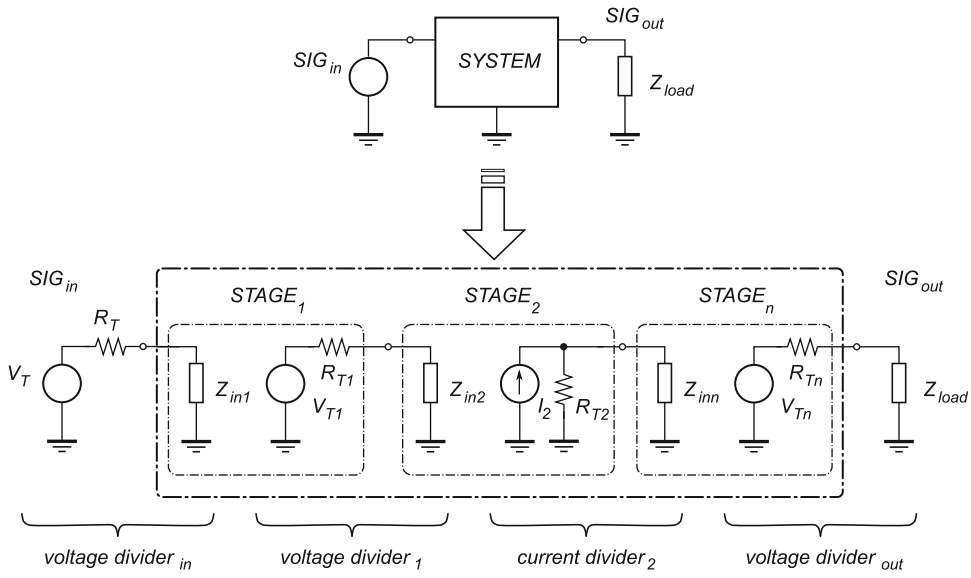


Fig. 6.1 General system partitioning based on Thévenin voltage divider and/or Norton current divider model at interface planes between adjacent stages

We appreciate the elegance of the approach once we realize that each stage that generates a signal at its output terminals serves as "driving stage" (or simply "the driver") while each stage that receives the signal at its input nodes is the "loading stage" (or simply "the load"). It is important to understand that, by itself, each stage is both driver (relative to its neighbour down the signal path) and load (relative to its neighbour up the signal path). From the signal transmission perspective, the internal structure of each stage is not relevant; indeed, it is only important to know the following:

- The amplitude of voltage (or current) generated by the driver V_{Ti}.
- The output impedance of the driver R_{Ti}.
- The input impedance of the load Z_{in}.

Conceptually, these three ideal elements (i.e., they do not really exist as such inside real circuits, they are only symbolic representations of the circuit behaviour) are used to model the signal transfer at each of the interface nodes. By doing so, the analysis of a complicated system is reduced to repeated calculations of a simple *voltage/current divider* at each of the interface planes.

6.2 Maximum Power Transfer

Routinely, analysis of a low-frequency system focuses on calculation of the voltage and current levels within the circuit, which results in "good enough" answers. This is true because the parasitic elements, which inherently have small RLC values, usually do not have a significant impact on circuit performance at low frequencies. Therefore, for low-frequency designs, approximation of parasitic reactances with either short or open connections, and dealing only with the "real" resistors and thermal energy losses is the generally accepted methodology.

At RF frequencies, however, voltage and current levels internal to the active circuit elements are, in general, not equal to the ones at the circuit terminals. Consequently, there is a non-negligible amount of wasted energy that is caused by the parasitic components associated with the circuit elements.

Because of that hidden waste of energy, instead of evaluating the internal voltages and currents separately, it is much more important to evaluate how the "instantaneous signal power" ($p = vi$) is transferred from one stage to another, with the implication that all internal impedances need to be accounted for. That is, the question becomes how the power is transferred between any two stages influenced by the values of the impedance divider at the interface.

Intuitively, it should be relatively easy to reach useful conclusions about the interface structure in the two extreme cases of the loading impedance, i.e., when the load impedance is either $Z_\mathrm{L} = 0$ or $Z_\mathrm{L} = \infty$. In the case of $Z_\mathrm{L} = 0$, the voltage level at the load terminals is zero, which means that the power delivered to the load must be $P_\mathrm{L} = V\,I = 0 \times I = 0$. In the case of $Z_\mathrm{L} = \infty$, however, the current delivered to the load is zero, hence the delivered power must be zero again. Considering that electronic circuits transfer signal power for all other cases that fall between these two extreme zero-power cases, the conclusion is that there must be at least one non-zero maximum power transfer point somewhere between them.

Although, strictly speaking, proof of the maximum power transfer condition requires calculus, it is possible to derive the same condition using a less rigorous approach based on complex algebra.

Let us assume complex source impedance $Z_0 = R_0 + \mathrm{j}X_0$ and complex load impedance $Z_\mathrm{L} = R_\mathrm{L} + \mathrm{j}X_\mathrm{L}$ driven by ideal voltage source V_0. The average power P_L is dissipated in the resistive part of the load, while the current is complex, i.e.

$$P_\mathrm{L} = I_{RMS}^2 R_\mathrm{L} = \frac{1}{2}|I|^2 R_\mathrm{L} = \frac{1}{2}\left(\frac{|V_0|}{|Z_0 + Z_\mathrm{L}|}\right)^2 R_\mathrm{L} = \frac{1}{2}\frac{|V_0|^2}{(R_0 + R_\mathrm{L})^2 + (X_0 + X_\mathrm{L})^2} R_\mathrm{L}. \tag{6.1}$$

By inspection of (6.1), we note that the power P_L increases when the reactive term of the denominator is at a minimum. The minimum of a square function (which always has a non-negative value) is, of course, zero. That is, the minimum value of $(X_0 + X_\mathrm{L})^2$ is achieved when the source and load reactances are equal and with opposite sign, i.e., $X_0 = -X_\mathrm{L}$.

That leaves (6.1) with the resistive terms only, hence

$$P_\mathrm{L} = \frac{1}{2}\frac{|V_0|^2}{(R_0 + R_\mathrm{L})^2} R_\mathrm{L} = \frac{1}{2}\frac{|V_0|^2}{\dfrac{R_0^2}{R_\mathrm{L}} + 2R_0 + R_\mathrm{L}}. \tag{6.2}$$

Therefore, the problem of finding the maximum value of P_L is reduced to the problem of finding the minimum value of the denominator in (6.2) in respect to the load resistance R_L, i.e.

$$\frac{\mathrm{d}}{\mathrm{d}R_\mathrm{L}}\left(\frac{R_0^2}{R_\mathrm{L}} + 2R_0 + R_\mathrm{L}\right) = -\frac{R_0^2}{R_\mathrm{L}^2} + 1 = 0, \quad \therefore \quad R_0 = R_\mathrm{L} \tag{6.3}$$

because resistive values are always positive. The two derived conditions for reactances, $X_0 = -X_\mathrm{L}$, and resistances, $R_0 = R_\mathrm{L}$, are combined and written as

$$Z_0 = Z_\mathrm{L}^*, \tag{6.4}$$

which is called "conjugate matching". This condition guarantees the most efficient power transfer; however, only at the one frequency for which source and load reactances are conjugate, i.e., $\mathrm{j}X_\mathrm{L} = -\mathrm{j}X_0$.

By substituting $R_\mathrm{L} = xR_0$ in (6.2), we have

$$P_\mathrm{L} = \frac{1}{2}\frac{xR_0}{(R_0 + xR_0)^2}\,|V_0|^2 = \frac{1}{2}\frac{x}{(1+x)^2}\frac{|V_0|^2}{R_0} = \frac{1}{2}\frac{x}{(1+x)^2}P_0, \tag{6.5}$$

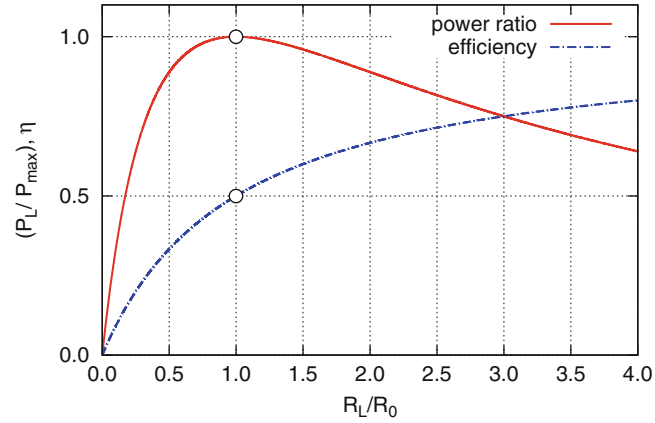


Fig. 6.2 Power transfer ratio P_L/P_0 (6.5), and maximum power efficiency (6.8), normalized to $P_L(\max) = 1/8$, versus $x = R_L/R_0$

therefore, for $x = 1$

$$P_L(max) = \frac{1}{2} \cdot \frac{1}{4}, \tag{6.6}$$

where, $P_L(max)$ is the maximum power dissipated in the load, i.e., under the condition $R_L = R_0$ (i.e., $x = 1$), and after normalizing. The normalized plot in Fig. 6.2 shows the delivered power at its maximum at the load ratio $(R_L/R_L(max))$ when $R_L = R_0$. In addition, if we define power transfer efficiency as

$$\eta = \frac{R_L}{R_L + R_0} = \frac{1}{1 + \frac{R_0}{R_L}} = \frac{1}{1 + \frac{1}{x}}. \tag{6.7}$$

Figure 6.2 shows that when the maximum power is transferred to the load, efficiency is only 50%, which is intuitively correct for the case of matched impedance. We note that efficiency tends to 100% while power transfer ratio tends to zero.

Alternatively, a non-conjugate matching or broadband matching condition,

$$Z_0 = Z_L \tag{6.8}$$

called *reflectionless match*, is used. It is not as efficient as conjugate matching but it does offers broader maxima. The matching condition used in practice depends upon the application.

6.3 Measuring Power Loss Due to Mismatch

In cases when the maximum power transfer is not achieved, it is useful to quantify the source–load offset. For example, in the case of two arbitrary matching impedances Z_1 and Z_2, the amount of the mismatch is quantified by the "reflection coefficient", which is defined as

$$\Gamma = \frac{Z_2 - Z_1}{Z_2 + Z_1}, \tag{6.9}$$

where, $0 \leq |\Gamma| \leq 1$. In mathematical terms, the power transfer is represented as a sum of two power waves: the incident power originating from the source and reflected power that was not delivered to the load. For good power transfer, the reflection should be as small as possible. In the case of perfect matching, the impedances Z_1 and Z_2 are equal, i.e., $\Gamma = 0$. In most communication systems, the associated standards specify the maximum allowed value of the reflection coefficient. More often,

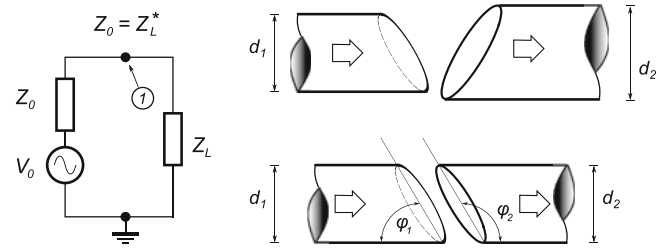


Fig. 6.3 Complex impedance matching network (*left*) and the pipe analogy showing the non-matched and matched cases (*right*)

the reflection coefficient, which is a unitless number, is converted to dBs and referred to as the "return loss"

$$RL_{\text{dB}} = 10\log(|\Gamma|^2) = 20\log|\Gamma|, \tag{6.10}$$

where, $0\,\text{dB} \leq RL \leq \infty$. Loosely stated, the return loss quantifies the difference in power delivered to the two interfacing impedances. To find out how much power is wasted at the interface, the "mismatch loss" (ML) is defined as

$$ML = \frac{1}{1-|\Gamma|^2}, \tag{6.11}$$

or, after conversion to dBs

$$ML_{\text{dB}} = -10\log(1-|\Gamma|^2), \tag{6.12}$$

where it is assumed that the signal source itself is ideal. In the case of arbitrary impedances at the network ports, calculation of the mismatch loss is a bit more complicated and we leave that for another occasion. To summarize this section, the return loss represents the difference between the reflected and the incident powers and a good match is indicated by a low return loss value. At the same time, mismatch loss represents the maximum possible power gain improvement relative to the case of a perfect match. Therefore, close to unity value of mismatch loss ML indicates a good match.

In order to visually illustrate the concept of complex power matching, it helps to introduce an analogy to the power flow in the form of a water flow through two pipes with diameters d_1 and d_2 (see Fig. 6.3). Pipe diameters and water flow have a relationship similar to that between resistance and current flow. We intuitively know that the most efficient water flow (i.e., no spills) happens when the two pipes have the same diameters, $d_1 = d_2$. To make the analogy even closer, let us note that if the two connecting pipes are cut at a right angle $\varphi = 90°$, then it does not matter how the pipes are rotated along their axes; they always make a good connection, i.e., they "match". However, if the pipes are cut at some other angle $\varphi \neq 90°$, the most efficient water flow is when the two angles are complementary, i.e., $\varphi + \varphi_2 = \pi$. Non-perpendicular angles are equivalent to complex impedances Z_0 and Z_L in the voltage divider, where the positive slope is equivalent to "inductive" and the negative slope to "capacitive" impedance, while the right angle is the special (and simpler) resistive case.

6.4 Matching Networks

Now that we intuitively understand the consequences of impedance mismatch, it is natural to ask what we should do in a more general (and realistic) case when the two matching impedances are not the same. Going back to the pipe analogy, when two pipes with unequal diameters need to be connected, we add a third pipe to serve as an adapter. Similarly, in order to enable efficient power transfer between two stages with non-matching impedances, an additional circuit network has to be

designed and inserted at the interface to serve as an "impedance converter". Detailed coverage of the art of matching network design is beyond the scope of this book; nevertheless, in the following sections, some of the basic concepts of matching network design is introduced by means of examples.

For sake of clarifying the terminology, it is important that we distinguish between two similar, and therefore often confused, circuit design activities: impedance transformation and impedance matching.

Impedance transformation is used to transform one impedance to a different value. At the output node of the transformation network, the new impedance is visible and it effectively masks the impedance connected to the input node of the transformation network. This interface is always unidirectional and is intended to interface only one impedance with the rest of the system.

Impedance matching is always performed between two impedances. The interface is always bidirectional and intended to maximize power transfer between the two impedances. In this book, unless otherwise specified, we design the inserted matching network with the goal of maximizing signal power transfer.

6.5 Impedance Transformation

In Sect. 4.1.7.2 and (4.59), we introduced loaded transformers. For convenience, we repeat here the important voltage–current relationship between the primary and secondary coils,

$$v_s i_s = v_p i_p, \tag{6.13}$$

which implies that in a transformer coil, increase in the coil voltage is accompanied by decrease in the coil current. Effectively, a transformer presents an impedance at its primary side that is different from the impedance of the load.

$$Z_p = \left(\frac{N_p}{N_s}\right)^2 Z_s, \tag{6.14}$$

which states that ratio of the primary and secondary impedances is equal to the square of the primary and secondary coil turns ratio. In other words, impedance Z_L at the secondary coil is "seen" at the primary coil as $(N_p/N_s)^2 R_L$. For this property alone, transformers are often used as impedance converters in RF circuits.

6.6 The Q Matching Technique

This impedance matching technique is based on the idea that a single L-shaped (X_S,X_P) branch is sufficient to provide impedance transition between two real, unequal resistances R_0 and R_L. When the two resistances are already equal, i.e., $R_0 = R_L$, there is no need for additional matching. We observe that by looking into the connecting node ① in Fig. 6.3, towards the source we see the same impedance as when looking at the load.

When the two matching resistances are not equal $R_0 \neq R_L$, we intuitively try to equalize the two sides at node ① by adding serial reactance X_S to the side with the lower initial resistance and, at the same time, by adding parallel reactance X_P to the side with the higher initial resistance. Of course, we are exploiting the fact that addition of a serial resistance increases the overall branch resistance, while addition of a parallel resistance reduces the overall branch resistance.

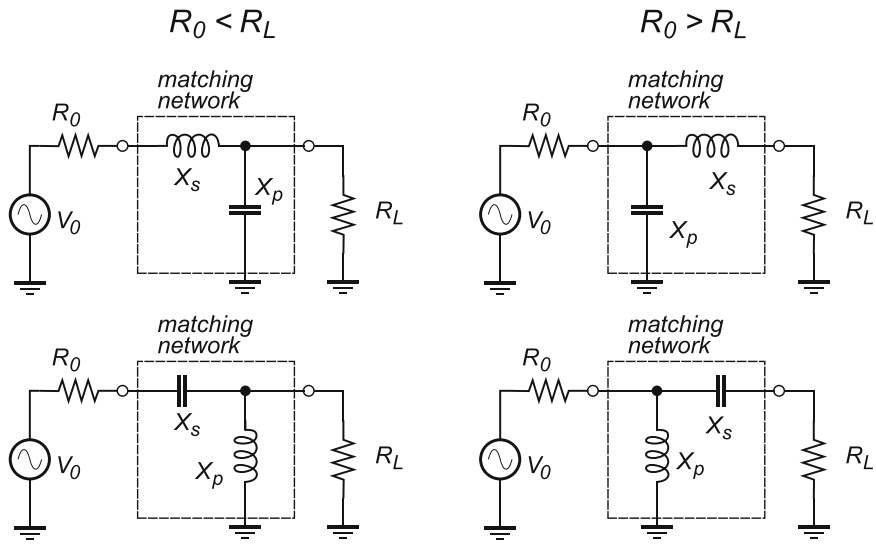


Fig. 6.4 Four ways to use a single X_0–X_P circuit branch as a matching network between the source resistance R_0 and the load resistance R_L. On the *left*, $R_0 < R_L$; on the *right*, $R_0 > R_L$

Because there are only two initial resistances to compare, R_0 and R_L, and two possible flavours of reactances that can be used $(j\omega L)$ and $(1/j\omega C)$, there are only four possible combinations that can be made. On the left of Fig. 6.4 either inductive or capacitive reactance X_S is used in series with $R_0 < R_L$. At the same time, either capacitive or inductive reactance X_P is used in parallel with load resistance $R_L > R_0$. The rule is that the two reactive components X_S and X_P must not be of the same type, i.e., one must be inductive and the other capacitive. Similarly, on the right of Fig. 6.4, either capacitive or inductive reactance is used in parallel with the source resistance $R_0 > R_L$ and either inductive or capacitive reactance is used in series with $R_L < R_0$.

At this point it is valid to ask why we use reactances when the same goal is achievable with a resistive network. It is possible to design a matching network using only resistors, however the power loss increases drastically and wideband networks are always much noisier, which reduces SNR. For each of the two relations between $R_0 \lessgtr R_L$, there two possible matching networks; we may ask if there is any difference between the two. If there are no additional constraints, either solution is valid. For example, either of the two matching networks on the left of Fig. 6.4 is valid when $R_0 < R_L$. In practice, we usually have additional constraints, for example, if a DC connection needs to be maintained between the source and load resistance then the serial reactance must be inductive, $X_S = j\omega L$ (the upper two cases in Fig. 6.4); if an AC connection is desired, the serial reactance must be capacitive, $X_S = 1/j\omega C$ (the lower two cases in Fig. 6.4).

6.6.1 Matching Real Impedances

A typical matching problem involves only real source and load resistances that are not equal, $R_0 \neq R_L$. For example, in Fig. 6.5, source resistance $R_0 = 5\,\Omega$ must drive a load of $R_L = 50\,\Omega$. After the matching network is designed and inserted, the source should "see" a load value of, in this case, $5\,\Omega$ and, at the same time, the load resistor should "feel" as if it was driven by a source resistance equal to its own, in this example $50\,\Omega$. Let us find out how a general problem such as this one is solved using the Q matching technique. As a side note, one of the drawbacks of this technique is its use of reactive components, meaning that the matching is possible at only one frequency.

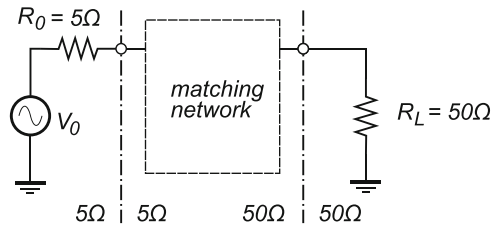


Fig. 6.5 A typical case of mismatched source and load resistances, $R_0 < R_L$

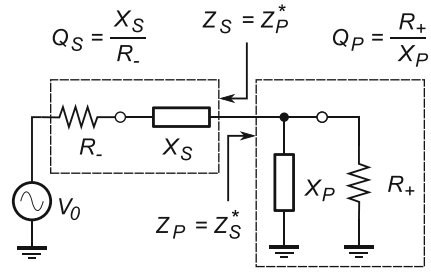


Fig. 6.6 An LC section placed between two resistive terminations creates a serial subnetwork and a parallel subnetwork. When the two subnetworks are conjugate matched to each other, their Q factors are equal

If we apply knowledge about serial–parallel transformations of resonant LC networks (see Sect. 5.5), the matching network design procedure is summarized by four simple steps[1]:

1. Add a series reactive element X_S next to R_- and increase the impedance of the serial subnetwork branch. Add a parallel reactive element X_P next to R_+ and reduce the impedance of the parallel subnetwork branch. We note that, if the serial element is an inductor, adding a parallel capacitor creates a LP topology (see Fig. 6.6); a serial capacitor in combination with a parallel inductor forms a high-pass section.
2. At the design frequency, the two newly created subnetworks, one in series and one in parallel (Fig. 6.6), must represent complex conjugate impedances to each other. Thus, the Q factors of these two subnetworks must be equal at the frequency where the match is computed. The serial Q factor Q_S and the parallel Q factor Q_P of the two subnetworks are

$$Q_S = \frac{X_S}{R_-} \quad \text{and} \quad Q_P = \frac{R_+}{X_P}. \tag{6.15}$$

3. Using (5.79) and (5.80) we calculate the serial and parallel Q factors of the two subnetworks as

$$Q_S = Q_P = \sqrt{\frac{R_+}{R_-} - 1}. \tag{6.16}$$

4. Once the Q factors are calculated, the next step is to calculate the series and parallel reactances from (6.15) and to compute the inductor and capacitor values by using their respective impedance definitions for the given design frequency.

In summary, the Q matching methodology for the case of signal source V_0 with real source resistance R_0 (either R_- or R_+) that drives a load with real resistance R_L (either R_+ or R_-) is a straightforward procedure because there are only four possible matching networks to consider. In order

[1] In order to better visualize the design steps, the lower of the two resistances is labelled R_- while the higher of the two is labelled R_+.

to make the solution unique, an additional constraint must be introduced to further determine the nature of serial and parallel impedances in the matching network. For example, if the matching network is to preserve a DC connection between the source and the load, then an inductor must be chosen as the serial element. Similarly, if an AC connection between the source and the load is to be preserved, a capacitor must be chosen as the serial element.

Example 6.1. Using the Q matching technique, design a single-section LC network to match a source resistance $R_0 = 5\,\Omega$ to a resistive load $R_L = 50\,\Omega$ at $f = 100\,\text{MHz}$ (see Fig. 6.5). Maintain a DC connection between the source and the load.

Solution 6.1. The source resistance is smaller than the load resistance, $R_S < R_L$, hence, $R_S = R_-$ and $R_L = R_+$ (Figs. 6.5 and 6.6). Therefore, serial reactance needs to be added to the source resistance R_- and parallel reactance to the load resistance R_+. Adding a serial inductor to the $5\,\Omega$ source side and a parallel capacitor to the $50\,\Omega$ load side keeps the DC connection and creates the LP matching configuration.

From (6.16), the required Q factors are calculated as

$$Q_S = Q_P = \sqrt{\frac{R_+}{R_-} - 1} = \sqrt{\frac{50}{5} - 1} = 3.$$

From (6.15) it follows, first for the serial component,

$$X_S = Q_S R_- = 3 \cdot 5\,\Omega = 15\,\Omega,$$

$$\therefore$$

$$L = \frac{15\,\Omega}{2\pi\,100\,\text{MHz}} = 23.873\text{nH}$$

and then for the parallel component,

$$X_P = \frac{R_+}{Q_P} = \frac{50\,\Omega}{3} = 16.667\,\Omega,$$

$$\therefore$$

$$C = \frac{1}{2\pi\,100\,\text{MHz}\,16.667\,\Omega} = 95.491\,\text{nF}.$$

Let us verify the above result. After inserting the matching network and looking into the source side (Fig. 6.6), there is a serial connection of the source resistance R_0 and the matching network's inductor X_S. Therefore, the total serial source side impedance is $|Z_0| = \sqrt{R_0^2 + X_S^2} = \sqrt{5^2 + 15^2}\,\Omega = 15.811\,\Omega$. At the same time, looking into the load side, there is a parallel connection of R_L and X_P. Therefore, the parallel impedance at the load side is $|Z_L| = 1/\sqrt{1/R_L^2 + 1/X_P^2} = 1/\sqrt{1/50^2 + 1/16.667^2}\,\Omega = 15.811\,\Omega$. Thus, the source side impedance has increased and the load side impedance has decreased, with the apparent matching of the two sides at $15.811\,\Omega$.

Example 6.2. Match a $50\,\Omega$ resistive source at 100 MHz to a load that is a serial connection of a $50\,\Omega$ resistor and a 95.491 nF capacitance.

Solution 6.2. This is a trivial case because the real parts of the source and load impedances are equal. Because the load side has an additional $X_S = -15\,\Omega$, the required matching circuit is simply

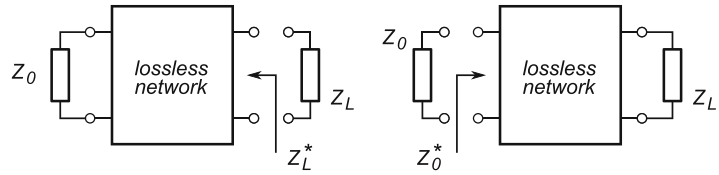


Fig. 6.7 Matching complex impedances by design of the lossless matching network simultaneously provides bidirectional complex conjugate matching

an inductor in series with the source resistance R_s, i.e., $X_L = +15\,\Omega$, therefore $L = 23.873\,\text{nH}$. At $f = 100\,\text{MHz}$, the serial connection of the inductor and the capacitor results in zero impedance, which leaves the two real resistances matched.

6.6.2 Matching Complex Impedances

A general case of matching complex impedances follows the same design methodology presented in the previous sections, i.e., properly designed matching network must provide correct complex conjugate matching both at the input terminal plane and at the output terminal plane. When looking into the output terminals of the matching networks we need to see the complex conjugate value of the output impedance and when looking into the input terminals of the matching network we need to see the complex conjugate value of the input impedance (Fig. 6.7). Under those conditions, all of the signal power is delivered to the load without any reflection at the output port.

The reactances associated with source and load impedances are referred to as "parasitics". If any of the two matching impedances Z_0 and Z_L already contains parasitics, the matching network design problem can be approached in two possible ways that may lead to the desired solution: we could try to "absorb the parasitics" into the matching network or to eliminate the parasitics by resonance, i.e to "resonate them out". In both of these methods, the parasitic reactances may be eliminated either completely or partially. A design procedure for matching complex impedances starts by solving the matching network for the real parts and then proceeds by absorbing the parasitics or resonating them out, either completely and partially.

6.6.2.1 Absorbing the Parasitics

Let us consider a case where source or load impedances include parasitic reactances. In addition, let us assume that values of the parasitic reactances are lower than the component values of the matching network that is required to match only the real parts of the two impedances. If that is the case, there is an opportunity to "absorb" i.e., to combine, these source or load parasitics with the matching network components. Let us take a look at the following example.

Example 6.3. Design a single-stage LC matching network at $f = 100\,\text{MHz}$ for the case of a source V_0 whose impedance consists of a resistor $R_0 = 5\,\Omega$ connected in series with an $L_S = 13.873\,\text{nH}$ inductor, which has to drive an $R_L = 50\,\Omega$ load resistance in parallel with $C_L = 45.491\,\text{nF}$, Fig. 6.8 (left). The matching network is expected to maintain a DC path between the source and the load.

Solution 6.3. In the case of a source or load with complex impedances, a good starting point is to first resolve the matching network only for the real parts of the two impedances. In Example 6.1,

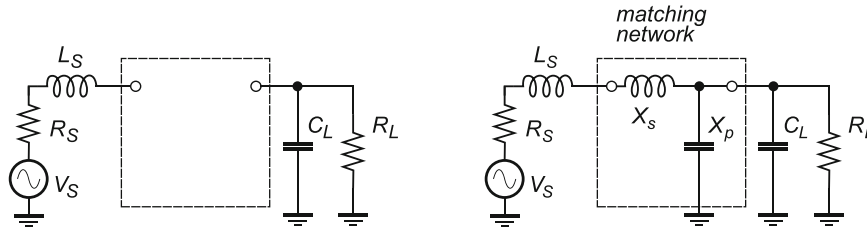


Fig. 6.8 Complex source with impedance $Z_0 = R_0 + j\omega L_S$ is required to drive a load with impedance $Z_L = R_L||1/j\omega C_L$ (*left*). Both the source and the load reactive parasitics are absorbed by the matching network (*right*)

we designed a matching network for the case of real $R_0 = 5\,\Omega$ source and $R_L = 50\,\Omega$ load at 100 MHz, which happen to be numerically equal to the real parts of the impedances in this example. Hence, we reuse the results and treat those calculations as the first phase of this example.

As we already found in Example 6.1, to match $R_0 = 5\,\Omega$ to $R_L = 50\,\Omega$ we need an $X'_S = 23.873\,\text{nH}$ inductor and an $X'_P = 95.491\,\text{nF}$ capacitor. However, the source impedance in this example already contains $L_S = 13.873\,\text{nH}$ inductance, which means that only the additional $X_S = 23.873\,\text{nH} - 13.873\,\text{nH} = 10\,\text{nH}$ inductor in series is needed, as shown in Fig. 6.8 (right). By doing this, we "absorb" the existing parasitic inductance into the value of the inductance required by the matching network. At the same time, the loading impedance needs a total of $X'_P = 95.491\,\text{nF}$ capacitance, which means that only the additional $X_P = 95.491\,\text{nF} - 45.491 = 50\,\text{nF}$ capacitor is needed in parallel with the existing $C_L = 45.491\,\text{nF}$ parasitic capacitance. By doing this, we "absorb" the existing parasitic capacitance into the value of the capacitance required by the matching network. Therefore, the required matching network consists of an $X_S = 10\,\text{nH}$ inductor and an $X_P = 50\,\text{nF}$ capacitor, as shown in Fig. 6.8 (right), i.e., $L_S + X_S = 23.873\,\text{nH}$ and $C_L + X_P = 95.491\,\text{nF}$.

6.6.2.2 Resonating out Excessive Parasitics

Let us consider a case where, for example, the load impedance includes parasitic reactance. In addition, let us assume that value of the parasitic reactance is greater than the value of the component of the matching network designed to match only the real parts of the two impedances. If that is the case, there is an opportunity to "resonate out", either fully or partially, the load's parasitic reactance with the matching network's components. To illustrate the point, let us reuse the results of the previous examples.

Example 6.4. Design a single-stage LC matching network at $f = 100\,\text{MHz}$ for the case of a source V_0 with a $R_0 = 5\,\Omega$ output resistance, which has to drive a $R_L = 50\,\Omega$ load resistance in parallel with $C_L = 105.491\,\text{nF}$. The matching network is expected to maintain a DC path between the source and the load.

Solution 6.4. As we already found in Example 6.1, to match $R_0 = 5\,\Omega$ to $R_L = 50\,\Omega$ we need $X'_S = 23.873\,\text{nH}$ inductor and $X'_P = 95.491\,\text{nF}$ capacitor. However, the source impedance already includes the parasitic capacitance of $C_L = 105.491\,\text{nF}$, which means that somehow we need to reach the required $X'_P = 95.491\,\text{nF}$. In general, there are two possible ways to approach this kind of problem.

- *Total resonating out*: Let us first create an LC resonator that consists of the existing parasitic capacitance C_L in parallel with an inductor L_L (a new component). If we set the resonant frequency of this LC resonator to $f_0 = 100\,\text{MHz}$, we create (in the ideal case) dynamic impedance $R_D = \infty$ in parallel to the load resistance R_L. Consequently, the total loading impedance becomes

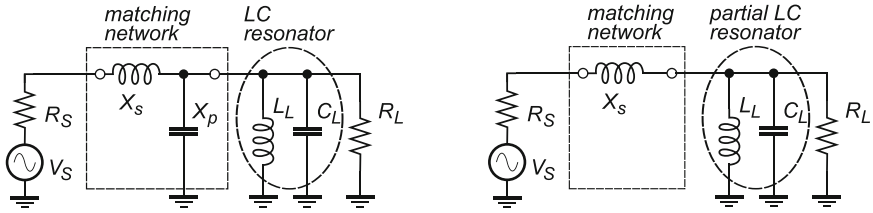


Fig. 6.9 Parasitic load capacitance C_L is completely "resonated out" by adding parallel inductor L_L (*left*). Only the excess amount of parasitic capacitance C_L is "partially resonated out" (i.e. its effective size is reduced) by adding parallel inductance L_L (*right*)

$Z_L = R_L || R_D = R_L$, i.e., the parasitic capacitance is "resonated out" and, effectively, disappears. In this case, we would need to use an $L_L = 24.024\,\text{pH}$ inductor. Although it is very difficult to create such a small inductor, for sake of argument let us keep the numbers. Once the parasitic capacitance is fully resonated out, we are back to the problem of Example 6.1, which means that, in order to finalize the matching network, we need to add an $X_S = 23.873\,\text{nH}$ inductor and an $X_P = 95.491\,\text{nF}$ capacitor, as shown in Fig. 6.9 (left). This solution requires three new components: X_S, X_P, and L_L.

- *Partial resonating out*: Let us try to resonate out only the excess part of the parasitic capacitance, i.e., $C'_L = C_L - X'_P = 105.491\,\text{nF} - 95.491\,\text{nF} = 10\,\text{nF}$. That is, let us create an LC resonator that consists of the existing part of the parasitic capacitance C'_L in parallel with an inductor L_L (a new component), so that they resonate at $f_0 = 100\,\text{MHz}$. To do so, we need $L_L = 1/(2\pi f_0)^2 C'_L = 252.3\,\text{pH}$. By adding the new component, the L_L inductor, in parallel with the parasitic capacitance, we effectively reduce the size of the capacitor. One way to visualize this situation is to imagine that the load capacitance C_L consists of two capacitors in parallel, i.e., $C_L = 95.491\,\text{nF} + 10\,\text{nF}$. The newly created LC resonator resonates out the 10 nF part and becomes effectively infinite dynamic impedance $R_D = \infty$. Hence, the 95.491 nF capacitance required by the matching network is still available – all we need to do is to add the series inductance $X'_S = 23.873\,\text{nH}$, as shown in Fig. 6.9 (right). Therefore this solution requires only the two new components, L_L and X_0. And, as a side note, in this solution the resonating inductor is a bit larger.

6.7 Bandwidth of a Single-Stage LC Matching Network

So far in our discussion of single-stage LC matching networks, we have only focused on the main goal of matching the source side impedance to the load side impedance. We had no freedom to control the bandwidth of the overall network. We have learned by now that a general RLC network always behaves as a bandpass filter centred around the resonant frequency ω_0, which is determined by the LC components. We also have learned that, as a good approximation (assuming Q factor larger than ten or so), the serial and parallel RLC networks resonate in the same way. Therefore, it is very important to estimate the bandwidth of matched networks, because we may reach a solution that offers too wide a bandwidth (and allows too much noise into the system) or too narrow a bandwidth (and alters the frequency content of the passing signals, i.e., the matching network distorts the signal).

A more detailed network analysis, which is beyond the scope of this book, would have revealed that determining the network bandwidth using the standard definition based on the 3 dB points turns out to be problematic, to say the least. There are at least two good reasons for this statement. One has to do with the fact that some resonant networks may never reach the 3 dB attenuation points. For example, a low Q resonant curve is almost flat—it may not even have 3 dB difference between its maximum amplitude at the point of the resonant frequency and the side points. A second, and less obvious, reason for our difficulties in determining the 3 dB bandwidth of LC matching networks

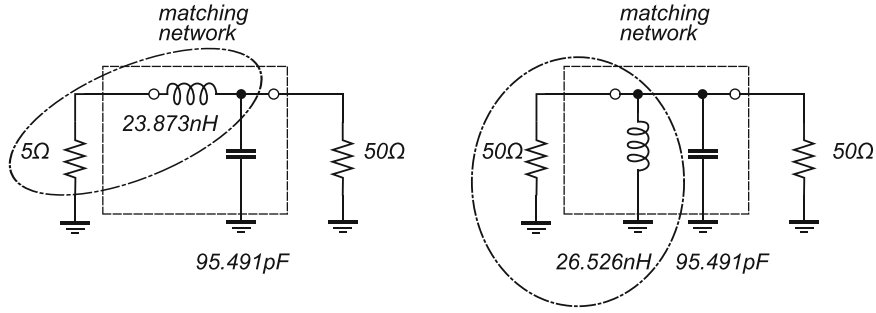


Fig. 6.10 Converting the serial RL subnetwork portion of the matching network into its equivalent parallel RL subnetwork, for purposes of calculating the overall bandwidth

unambiguously is that, in general, resonant curves are not symmetrical around the resonant frequency. Nevertheless, in this book we assume high Q resonant networks (which is a reasonable assumption in the case of wireless radio) and we assume symmetrical resonant curves (which is a valid assumption in the narrow region around the resonant frequency point).

With these assumptions in mind, let us now determine the bandwidth for the matching network of Example 6.1, as shown in Fig. 6.10 (left). It is important to recognize that, for a narrow bandwidth, the LC circuit is treated as a "resistively loaded resonator". To show how this works, we need to convert the serial source impedance $Z_0 = (5 + \mathrm{j}15)\,\Omega$ subnetwork into its equivalent parallel subnetwork (at $f = 100\,\mathrm{MHz}$). The serial RL branch is treated as a non-ideal inductor whose Q_S factor is $Q_S = {}^{15}/_{5} = 3$. Conversion of the serial RL branch is easily done by using (5.79) and (5.80), which results in a parallel resistor of $R_p = R_s(1 + Q_S^2) = 5\,\Omega(1 + 3^2) = 50\,\Omega$ and the equivalent parallel inductor is $L_p = L_s({}^{1}/_{Q_S^2} + 1) = 23.873\,\mathrm{nH}({}^{1}/_{3^2} + 1) = 26.526\,\mathrm{nH}$, shown in Fig. 6.10 (right). Again, note that we used the Q_S factor for the stand-alone RL branch by itself. After this conversion, it is easy to see a parallel resonant circuit loaded by two $50\,\Omega$ resistors in parallel, i.e., which is effectively $R_{\mathrm{loaded}} = 25\,\Omega$, as in Fig. 6.10 (right).

In order to determine the 3 dB bandwidth of the matching circuit we need to calculate the Q factor of the "loaded" network at resonance. By using either the capacitive or inductive reactance at resonance, we write as usual

$$Q_{\mathrm{loaded}} = \frac{R_{\mathrm{loaded}}}{X_P} = \frac{25\,\Omega}{16.667\,\Omega} = 1.5, \tag{6.17}$$

$$\therefore$$

$$\Delta f = \frac{f_0}{Q_{\mathrm{loaded}}} = \frac{100\,\mathrm{MHz}}{1.5} = 66.667\,\mathrm{MHz}, \tag{6.18}$$

which, although not completely accurate still provides a reasonable estimate of the bandwidth. Numerical simulations show that result (6.18) underestimates the actual bandwidth by approximately 20–30%, which is what expected due to the assumptions being made.

6.7.1 Increasing Bandwidth with Multisection Impedance Matching

The single-stage matching network does not have enough steps of freedom to allow for setting the network bandwidth. In order to gain control over that parameter, we need to expand our single-stage,

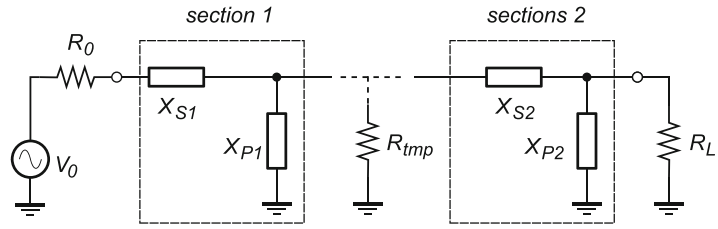


Fig. 6.11 Two-stage matching network with an intermediate temporary resistance R_{tmp} for the purpose of intermediate calculations

L-shaped matching network and add a second section. In principle, the two-section network is solved by repeating two times the methodology that we used for single-section networks. The additional step of freedom is achieved by introducing a temporary loading resistance R_{tmp} (see Fig. 6.11). This allows us to split the two-section matching network problem into two single-section matching networks, where R_{tmp} serves as the temporary load for the first section and as the source resistance for the second section.

We set the value of the temporary resistance R_{tmp} to

$$\frac{R_0}{R_{\text{tmp}}} = \frac{R_{\text{tmp}}}{R_{\text{L}}} \qquad \therefore \qquad R_{\text{tmp}} = \sqrt{R_0 R_{\text{L}}}, \tag{6.19}$$

which is the geometrical mean between the source R_0 and load R_{L} resistances. The addition of the second section using condition (6.19) provides an optimal compromise in increasing the bandwidth.

Once again, the temporary resistance R_{tmp} is a "ghost" value, not a real physical component; it is merely a number that would have been seen by looking into the matching network if it were split at the middle.

6.7.2 *Decreasing Bandwidth with Multisection Impedance Matching*

There are circumstances when we want to design a narrow bandwidth matching network. For example, input stages of RF amplifiers should be limited to only the minimum necessary bandwidth. As you already expect, bandwidth reduction also requires a two-section matching network (Fig. 6.12), except that this time the value of the temporary resistance R_{tmp} is chosen outside the range set by the source R_0 and load R_{L} resistance values. Consequently, there are two possible choices for its value, one where $R_{\text{tmp}} < R_-$ and one where $R_{\text{tmp}} > R_+$.[2]

As long as the condition $R_{\text{tmp}} \lesseqgtr [R_0, R_{\text{L}}]$ is satisfied, we have almost arbitrary freedom to pick the value of the temporary resistance. However, in practice the decision about where to place R_{tmp} depends on the impedance levels of the terminations and practically realizable component values. For example, if the existing values of $[R_0, R_{\text{L}}]$ are already low, then it is more practical to select a temporary resistance on the high side, $R_{\text{tmp}} > R_+$, while if the existing values of $[R_0, R_{\text{L}}]$ are high, then we pick a low value for the temporary resistance so that $R_{\text{tmp}} < R_-$. Aside from these notes, there is nothing special about the two-stage (or even multi-stage) matching networks. It is good engineering practice to design circuits with a minimum number of components, hence, as the last step in the design of two-section matching networks, multiple serial inductances should be replaced by a single component and multiple parallel capacitances should be replaced by a single capacitor.

[2] We use the same notation: $R_- = min[R_0, R_{\text{L}}]$ and $R_+ = max[R_0, R_{\text{L}}]$.

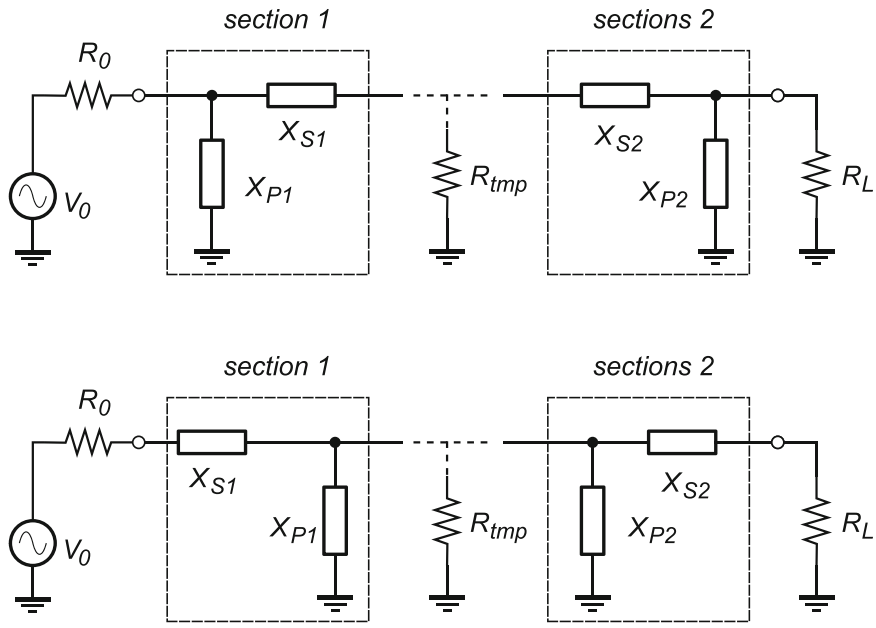


Fig. 6.12 Selection of an independent R_{tmp}

6.8 Summary

In this chapter, we familiarized ourselves with the basic concepts of power transfer between stages of a general multi-stage system. Having the main motivation of designing networks for use in wireless radio systems, it became our main priority to maximize power transfer of the RF signals. In order to achieve that goal, we introduced the concept of matching networks that serve as the gradual impedance converter between the source impedance and the load impedance. Their first application is between the antenna and the RF amplifier. The reader should keep in mind that Q matching is just one of several key techniques for matching network design. The next logical step is to start using Smith Charts, which offer a somewhat more elegant way of designing RF matching networks, especially at higher frequencies.

Problems

6.1. Using the Q matching technique, find an equivalent parallel network to the serial connection of $R_S = 5\,\Omega$ and $L_S = 2.8\,\text{nH}$ at $f = 100\,\text{MHz}$.

6.2. Design a single-stage LC matching network between a source with $R_S = 5\,\Omega$ and load $R_L = 50\,\Omega$ termination at $f = 100\,\text{MHz}$. An additional condition is to maintain a DC connection between the source and load sides.

6.3. Using the results from Problem 6.2, find reflection coefficient Γ and mismatch loss ML at the interface between the serial and parallel parts of the matching network.

6.4. Using the results from Problem 6.2, estimate the 3 dB bandwidth, assuming a symmetrical network.

6.5. Using the results from Problem 6.3, if the input signal changes to $f = 80\,\text{MHz}$, recalculate Γ and ML. Can you comment on the results?

6.6. Using the results from Problem 6.3, If the input signal changes to $0.2\,\text{GHz}$, recalculate Γ and ML. Can you comment on the results?

6.7. Design a single-stage LC matching network when parasitic inductance L_S exists in series with the source resistance R_S, where: $R_S = 5\,\Omega$, $R_L = 50\,\Omega$, $L_S = 0.93\,\text{nH}$, and $f = 0.85\,\text{GHz}$.

6.8. Design a single-stage LC matching network when parasitic capacitance C_L exists in parallel with the load resistance R_L, where $R_S = 5\,\Omega$, $R_L = 50\,\Omega$, $C_L = 20\,\text{pF}$, and $f = 0.1\,\text{GHz}$.

6.9. Match $5\,\Omega$ source resistance to $50\,\Omega$ load resistance at $200\,\text{MHz}$. Use a two-stage LC matching network with low-pass–high-pass filter combination. The goal is to increase the bandwidth relative to the single-stage LC matching network solution.

6.10. Match source resistance of $5\,\Omega$ to load resistance of $50\,\Omega$ at $200\,\text{MHz}$. Use a two-stage LC matching network. The goal is to decrease the bandwidth relative to the single-stage LC matching network solution. Make your own choice, and justify it, of temporary resistance R_{tmp}.

6.11. Antenna impedance is assumed to be resistive $50\,\Omega$. An RF amplifier is tuned at $665\,\text{kHz}$ and has input impedance of $Z_{\text{in}} = 2\,\text{k}\Omega$. Design two possible matching networks using the Q matching technique and comment on differences between the two solutions.

6.12. Using the parasitic absorption method, match a source impedance of $Z_S = (50 + \text{j}100)\,\Omega$ to a load impedance of $Z_L = (1{,}000 - \text{j}750)\,\Omega$ (the capacitor is in parallel with R_L) at $100\,\text{MHz}$.

6.13. Using the parasitic resonance method, match a source resistance of $R_S = 50\,\Omega$ to a load impedance that consists of $C_L = 10\,\text{pF}$ in parallel with $R_L = 500\,\Omega$ at $100\,\text{MHz}$. The matching circuit should maintain a DC connection from the input to the output.

Chapter 7
RF and IF Amplifiers

Abstract After a weak radio frequency (RF) signal has arrived at the antenna, it is channeled to the input terminals of the RF amplifier through a passive matching network. As we learned in Chap. 6, the matching network enables maximum power transfer of the receiving signal by equalizing the antenna impedance with the RF amplifier input impedance. After that, it is job of the RF amplifier to increase the power of the received signal and prepare it for further processing. In the first part of this chapter, we review the basic principles of linear baseband amplifiers and common circuit topologies. In the second part of the chapter, we introduce RF and IF amplifiers. In order to clarify the difference between RF and IF amplifiers, we need to know that in most radio receiver topologies the incoming high-frequency signal is not shifted down to the baseband in a single step. Instead, for reasons that we discuss in detail in Chap. 9, frequency down-shifting inside radio receivers is usually done in one or more intermediate steps. RF amplifiers used at those lower frequencies are referred to as intermediate frequency (IF) amplifiers. Aside from their operating frequency, for all practical purposes, there is not much difference between the schematic diagrams of RF and IF amplifiers. In this book, unless we need to specifically separate the two functions, we refer to all tuned amplifiers as RF amplifiers.

7.1 General Amplifiers

The topic of linear baseband amplifiers is usually covered in introductory undergraduate courses in electronics, thus there is a large number of excellent textbooks available with thorough treatments of the subject, some of them listed in the reference section. Assuming that the reader is familiar with basic concepts in electronics, in this book, we introduce the "back-of-the-envelope" approach to circuit analysis with the intent of encouraging the reader to start developing intuition for circuits and to start developing mental skills of circuit analysis. Indeed, even though the "back-of-the-envelope" approach, which is often based on very rough approximations, leads to conclusions that are sometimes an order of magnitude from the "correct" numerical solution, its usefulness is in enabling the circuit designer to focus on the underling principles of circuit operation instead of on the fine and tedious numerical details. As a result, the amount of time spent reaching correct conclusions is often measured in seconds. By practicing mental analysis of circuits, designers eventually develop their intuition for the underlying principles and an ability to immediately spot possible problems and circuit limitations, which is the bedrock of innovation and creativity. Indeed, the machines and simulators still cannot solve problems and improve existing solutions: they merely produce numbers that may or may not have anything to do with the problem at hand. Until we reach the age of intelligent machines, our brain is still the only tool we have that is capable of creative reasoning.

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8_7,

The other point in support of mental analysis approach is that the very notion of a "correct" answer is often fuzzy. The point is that a circuit's internal states keep changing in both the time and frequency domains. That is, before the signal processing operation is finished, the circuit's internal voltage and current levels have changed many times. Therefore, it is valid to ask which of the states is the "correct" one. The answer is "all of them" and that is why numerical simulators are useful. They enable designers to observe the ever-changing internal states of the circuit, which would have been too much numerical information for our brains to handle. Hopefully, I have succeeded in convincing you, the reader, that in order to be practical engineers, we have to be fluent both in intuitive reasoning and in the use of numerical methods that help us to quantify our intuitive conclusions.

7.1.1 Amplifier Classification

A simple, general classification of amplifiers is done in respect of the nature of their input and output signals. In the world of electronic circuits, the signals are in the form of either voltage or current. Keep in mind though that voltage and current are not two independent variables that you could separate at your will. Instead, they are two representations of the same phenomenon, i.e., the position of the charge carriers in time and space. At the highest level of abstraction, the relationship between voltage and current is described by Maxwell's equations, which also include media where the charges are located. Kirchhoff's and Ohm's laws are simply the low-frequency approximations of Maxwell's equations derived under the assumption that wavelength λ of the signal being observed is much longer than the distance d it needs to travel, i.e., $\lambda \gg d$.

Students often ask how they should decide whether to use the voltage or current signal. In fact, deciding whether to process a signal in the form of voltage or current is a venture by itself and, except in the purely abstract mathematical world, there is no such thing as a pure voltage amplifier or a pure current amplifier, or any other "pure" signal-processing circuit for that matter. Instead, we approximate a circuit function as "a voltage amplifier" or "a current amplifier" based on the circuit characteristics, e.g. input and output impedances.

From the purely mathematical perspective, the function of an ideal linear amplifier is written as

$$y(x(t)) = K\,x(t), \tag{7.1}$$

where $x(t)$ is the time-dependent signal variable that is presented at the amplifier's input terminals, $y(x(t))$ is the time-dependent variable at the amplifier's output terminals, and K is the multiplication factor between the $x(t)$ and $y(t)$ variables, which is called *gain*. Strictly speaking, although the word "gain" implies a number larger than one, gain K can take any value, i.e., $-\infty \leq K \leq \infty$. Negative gain indicates that variables x and y have opposite phase while their amplitude relation is still controlled by the absolute value of K. Although gain less than one, i.e., $y < x$, is sometimes referred to as *loss*, the term gain assumes both "gain" and "loss". Additionally, in (7.1) it is assumed that K is constant, which translates into $y(x)$ being a linear function. We will review this particular assumption a number of times.

In the material world, the two abstract variables (x, y) are given physical meaning. Electronic amplifiers have two sets of terminals, i.e., the input and the output, that are capable of accepting two forms of signal, i.e., voltage and current, hence there are only four possible amplifier variants:

- *Voltage amplifier*: A circuit is classified as a "voltage amplifier" if the voltage signal v_{in} at its input causes a proportional voltage signal v_{out} at its output, with the circuit's input–output (I/O) transfer function as $v_{\text{out}} = A_{\text{v}}\,v_{\text{in}}$, where the multiplication constant A_{v} is referred to as the "voltage gain" in units of $[V/V]$(or in dB).

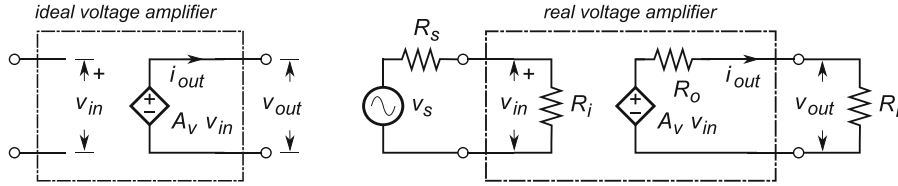


Fig. 7.1 An ideal voltage amplifier (*left*) and a realistic voltage amplifier connected with the input signal source and the output load (*right*). In order to distinguish a voltage source, which is a two-terminal device, from a voltage-controlled voltage source, which is a four-terminal device, it is a convention to use the diamond-shaped symbol instead of the circular shape

- *Current amplifier*: A circuit is classified as a "current amplifier" if the current signal i_{in} at its input causes a proportional current signal i_{out} at its output, with the circuit's I/O transfer function as $i_{\text{out}} = A_{\text{i}}\, i_{\text{in}}$, where the multiplication constant A_{i} is referred to as the "current gain" in units of $[A/A]$ (or in dB).
- *Transconductance amplifier*: A circuit is classified as a "transconductance (G_{m}) amplifier" if the voltage signal v_{in} at its input causes a proportional current signal i_{out} at its output, with the circuit's I/O transfer function as $i_{\text{out}} = G_{\text{m}}\, v_{\text{in}}$, where the multiplication constant G_{m} is referred to as the "voltage to current gain" in units of siemens ($\text{S} = \text{A}/\text{V} = 1/\Omega$). The accepted convention is to use G_{m} to indicate transconductance of a circuit, e.g. an amplifier, and g_{m} to indicate transconductance of a single device, e.g. a BJT.
- *Transresistance amplifier*: A circuit is classified as a "transresistance amplifier" if the current signal i_{in} at its input causes a proportional voltage signal v_{out} at its output, with the circuit's I/O transfer function as $v_{\text{out}} = A_{\text{R}}\, i_{\text{in}}$, where the multiplication constant A_{R} is referred to as the "current to voltage gain" in units of ohms ($\Omega = \text{V}/\text{A}$). A trivial example of a transresistance amplifier is a linear resistor, $v_{\text{R}} = R\, i_{\text{R}}$.

By combining these four possible ideal amplifying functions, we are able to both synthesize and analyze any complicated multi-stage amplifying circuit that may be optimized to process either the voltage or the current form of signals, or even to keep switching the signal form along the way. As a first step in moving from the ideal mathematical concept of amplifiers into the real world, we need to take a closer look at the characteristics of each of the four ideal amplifiers and the consequences of interfacing them with the signal source and the subsequent loading stages. For the sake of simplicity, in the following discussion we assume that amplifiers operate properly regardless of the signal frequency, hence we use the terms "resistance" and "impedance" interchangeably.

7.1.2 Voltage Amplifier

A functional symbol of an ideal voltage amplifier, Fig. 7.1 (left), shows the literal implementation of (7.1), which is based on an ideal voltage-controlled *voltage source* (VCVS) element whose voltage gain is A_{v}. Voltage gain A_{v} is, by definition, measured in $[V/V]$ or, more often, in dB.

Important characteristics of an ideal voltage amplifier to observe are as follows: Beginning from the left of the ideal voltage amplifier symbol, any voltage v_{in} presented at the input terminals outside the amplifier is immediately transferred inside the amplifier without any loss or change. The lack of any components on the input terminals indicates that the input impedance Z_{i} of the ideal voltage amplifier is equivalent to an open connection. To put it in technical terms, the input impedance of the ideal voltage amplifier is $Z_{\text{i}} = \infty$, which is another way of saying that the input current is $i_{\text{in}} = v_{\text{in}}/Z_{\text{i}} = v_{\text{in}}/\infty = 0$.

On the right of the ideal voltage amplifier, the output voltage v_{out} is generated by the internal VCVS that simply takes the input voltage value v_{in} as seen at the internal nodes of the amplifier and multiplies it by the multiplication constant $A_v = v_{out}/v_{in}$. Thus, as we learned in Sect. 4.1.2, because the internal impedance of an ideal voltage source is zero the output impedance of the ideal voltage amplifier in Fig. 7.1 (left) must also be zero, i.e., $Z_o = 0$ (looking into the output terminals of the ideal voltage amplifier, VCVS is the only element connected).

An element or circuit aspiring to be classified as a "voltage amplifier" must be as close as possible to the ideal model (Fig. 7.1 (left)). The criteria for quantifying the success of this aspiration are:

- The input impedance Z_i has to be very high, ideally infinite.
- The output impedance Z_o has to be very low, ideally zero.
- The voltage gain A_v has to be uniquely defined and constant.

To develop a sense of how closely and under what conditions any circuit could reach the ideal voltage amplifier model, let us take a look at the realistic amplifier model in Fig. 7.1 (right). The amplifier's input resistance is explicitly modelled by resistor R_i and the output VCVS is connected in series with resistor R_o. A careful reader should easily recognize that, under conditions of $R_i \to \infty$ and $R_o \to 0$ the real voltage amplifier model, Fig. 7.1 (right), degenerates into the ideal voltage amplifier model, Fig. 7.1 (left). Following the same idea, the input signal source (i.e., the driver) is modelled as an ideal voltage source v_S in series with a non-zero resistance $R_S > 0$. The loading circuit (i.e., the load) that receives the amplifier's output signal v_{out} is modelled only by its input impedance, a simple load resistor R_L.

Our main concern is to quantify the relationship between the output voltage v_{out} and the source voltage v_S, i.e., to find out how the voltage gain $A_v = v_{out}/v_S$ is influenced by the combination of the non-ideal driver, the non-ideal amplifier, and the load. The following analysis is very general and should be used as the foundation for establishing circuit analysis skills for any case of interface between a voltage source and a load. Even a casual reader should immediately recognize the application of the voltage divider concept that was introduced in Sect. 4.1.9.1.

At the input side of the amplifier, there is a voltage divider created by the voltage source v_S, source resistance R_S, and the amplifier's input resistance R_i. Therefore, the portion of the source signal level v_S that is transferred to the amplifier's internal nodes v_{in} (and subsequently multiplied by gain A_v) is calculated as

$$v_{in} = i_{in} R_i = \frac{v_S}{R_S + R_i} R_i, \qquad \therefore \qquad A'_v = \frac{v_{in}}{v_S} = \frac{R_i}{R_S + R_i} = \frac{1}{\frac{R_S}{R_i} + 1}, \tag{7.2}$$

where A'_v is the voltage gain of the input voltage divider itself. While keeping in mind our main goal with this circuit, i.e., to efficiently transfer the source voltage signal into the amplifier with no attenuation, let us for a moment take a closer look at what happened at this interface. Non-zero impedances at the source side created a resistive voltage divider (R_S, R_i) that caused a proportional reduction in the source signal v_S on its way to the amplifier's internal nodes. We need to determine the severity of this attenuation and conclude under what conditions the voltage signal transfer is ideal, i.e., lossless.

With passive components, the best that we can hope for is to transfer the full source signal level v_S to the inside of the amplifier, i.e., to achieve $v_{in} = v_S$ (or equivalently, $A'_v = 1$). By inspection of (7.2), we easily conclude that there are two conditions that would lead to $v_{in} = v_S$, the first one being

$$\lim_{R_S \to 0} A'_v = \frac{1}{\frac{0}{R_i} + 1} = 1, \tag{7.3}$$

that is, if a real voltage amplifier (the one whose input impedance is $0 < R_\mathrm{i} < +\infty$) is driven by an ideal source signal generator (the one whose source impedance is $R_\mathrm{S} = 0$) then the input voltage divider degenerates into a single resistor R_i driven by the ideal voltage source v_S, hence there is no voltage attenuation. The second limiting case is,

$$\lim_{R_\mathrm{i}\to\infty} A'_\mathrm{v} = \frac{1}{\frac{R_\mathrm{S}}{\infty}+1} = 1, \tag{7.4}$$

which means that if an ideal voltage amplifier (the one whose input impedance is $R_\mathrm{i} = \infty$) is driven by a real voltage source (the one whose source impedance is $0 < R_\mathrm{S} < +\infty$), we again achieve lossless voltage transfer across the input voltage interface.

The two conditions for perfect voltage transfer at the source side (7.3) and (7.4) can be combined by stating that a voltage amplifier must have large input impedance relative to the signal source impedance, i.e.,

$$R_\mathrm{i} \gg R_\mathrm{S}. \tag{7.5}$$

As a side note, while we are on this topic, note that under the condition of matched impedances $R_\mathrm{i} = R_\mathrm{S}$, the voltage gain is $A'_\mathrm{v} = 1/2$, which means that only half of the input voltage is transferred through matched networks. Remember that it means only a quarter of the signal power is transferred due to the $P = f(V^2)$ relationship.

We can now move our focus to the output terminals of the amplifier, Fig. 7.1 (right). The output voltage divider consists of the amplifier's output impedance R_o and the load impedance R_L, and it is driven by an ideal voltage source that generates $v_\mathrm{o} = A_\mathrm{v} v_\mathrm{in}$ output signal. Note that the signal v_o is internal to the amplifier, hence our main concern is to determine what percentage of it reaches the terminals of the load resistor R_L. It is straightforward to write

$$v_\mathrm{out} = i_\mathrm{out} R_\mathrm{L} = \frac{A_\mathrm{v} v_\mathrm{in}}{R_\mathrm{o}+R_\mathrm{L}} R_\mathrm{L}, \quad \therefore \quad A''_\mathrm{v} = \frac{v_\mathrm{out}}{v_\mathrm{in}} = A_\mathrm{v} \frac{R_\mathrm{L}}{R_\mathrm{o}+R_\mathrm{L}} = A_\mathrm{v} \frac{1}{\frac{R_\mathrm{o}}{R_\mathrm{L}}+1}, \tag{7.6}$$

where A''_v is the voltage gain of the output voltage divider. Again, non-zero impedances at the load side create a resistive voltage divider which causes proportional reduction of the output signal v_o on its way to the load terminals.

By inspection of (7.6), we easily conclude that there are two conditions that would lead to $v_\mathrm{out} = A_\mathrm{v} v_\mathrm{in}$, the first being

$$\lim_{R_\mathrm{o}\to 0} A''_\mathrm{v} = A_\mathrm{v} \frac{1}{\frac{0}{R_\mathrm{L}}+1} = A_\mathrm{v}, \tag{7.7}$$

that is, zero output impedance R_o is required if an amplifier is to maximize voltage transfer at its output side. The second limiting case is,

$$\lim_{R_\mathrm{L}\to\infty} A''_\mathrm{v} = A_\mathrm{v} \frac{1}{\frac{R_\mathrm{o}}{\infty}+1} = A_\mathrm{v}, \tag{7.8}$$

which means that if an infinitely large loading impedance is attached to a real amplifier, we have again achieved lossless voltage transfer across the output voltage interface. The physical interpretation of this condition is that the amplifier is disconnected from the load. Surprisingly, this mistake is often made by junior designers while trying to maximize the gain of their new amplifiers—watch out for it.

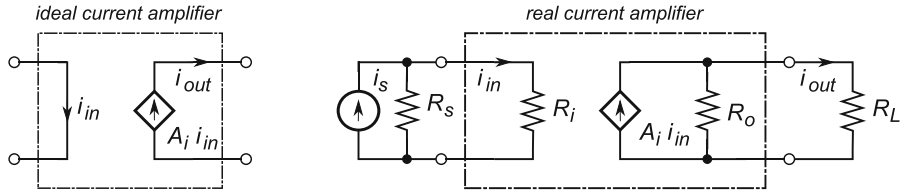


Fig. 7.2 An ideal current amplifier (*left*) and realistic current amplifier connected with the input signal source and the output load (*right*). In order to distinguish a current source, which is a two-terminal device, from a current-controlled current source, which is a four-terminal device, it is a convention to use the diamond-shaped symbol instead of the circular shape

The two conditions for perfect voltage transfer at the load side (7.7) and (7.8) can be combined by stating that a voltage amplifier must have much smaller output impedance than the load impedance, i.e.,

$$R_\mathrm{L} \gg R_\mathrm{o}. \tag{7.9}$$

It is educational to put together conditions (7.2) and (7.6) and take a look at the total gain that can be achieved with a real source, a real amplifier, and a real load chain. The two gains A'_v and A''_v follow each other, thus the total gain A_v from the source to the load terminals is their product

$$A_\mathrm{v} = \frac{v_\mathrm{out}}{v_\mathrm{s}} = A'_\mathrm{v} A''_\mathrm{v} = \frac{R_\mathrm{i}}{R_\mathrm{S} + R_\mathrm{i}}\, A_\mathrm{v}\, \frac{R_\mathrm{L}}{R_\mathrm{o} + R_\mathrm{L}}, \tag{7.10}$$

which clearly shows the three terms contributing to the total gain of a real amplifier. The first and third terms show attenuation caused by the input and output voltage dividers, and each is less than one. The second term is the only term possibly larger than one, and it represents the maximum possible voltage gain that could be achieved under the ideal condition of zero loss at the input and output terminals. The last statement, which in plain language summarizes (7.10), is the first key to our ability to mentally evaluate the effectiveness of voltage amplifiers and the appropriateness of processing the signal in voltage form. Without knowing the actual values of resistances R_i, R_S, and R_L, while analyzing voltage amplifiers, we can still make the ideal voltage amplifier approximation and evaluate the best possible case, by assuming high input impedance and low output impedance of the voltage amplifier.

In our quest to work out what types of circuit qualify as a voltage amplifier, we have concluded that low signal source impedance, high amplifier input impedance, low amplifier output impedance, and high load impedance together define a voltage amplifier. In short, (7.5) and (7.9) state that a voltage amplifier must have its input impedance much larger than the source impedance and at the same time it must have its output impedance much lower than the load impedance. Depending how closely a circuit reaches these conditions, we quantify how good a voltage amplifier it is. As good as these conditions are for enabling efficient voltage transfer and amplification, note that we simply ignored the fact that power transfer, under the ideal voltage transfer conditions, is zero. Our crudest approximations help us perform circuit analysis in our heads but we fall for the ideal voltage amplifier model; we must keep in mind the total picture regarding power transfer.

7.1.3 Current Amplifier

A functional symbol of an ideal current amplifier, Fig. 7.2 (left), shows the literal implementation of (7.1), which is based on an ideal current-controlled *current source* (CCCS) element whose current gain is A_i, which is, by definition, measured in $[A/A]$ or, more often, in dB.

Ii is important to observe the following characteristics of an ideal current amplifier. Beginning at the left of the ideal current amplifier symbol, any input current i_{in} presented at the input terminals outside the amplifier is immediately transferred to inside the amplifier without any loss or change. A short connection at the input terminals indicates that the input impedance Z_i of an ideal current amplifier is equivalent to a short connection. To put it in technical terms, the input impedance of the ideal current amplifier is $Z_i = 0$, which is another way of saying that the input voltage is $v_{in} = i_{in} Z_i = i_{in} \times 0 = 0$.

At the right of the ideal current amplifier symbol, the output current i_{out} is generated by the CCCS, which simply takes the value of the input current i_{in} as seen at the internal branch of the amplifier and multiplies it by the multiplication constant $A_i = i_{out}/i_{in}$. Thus, as we learned in Sect. 4.1.3, because the internal resistance of an ideal current source is infinite, the output resistance of an ideal current amplifier is also infinite, i.e., $R_o = \infty$ (looking into the output terminals, CCCS is the only element connected and its impedance is infinite).

Overall, an element or circuit aspiring to be called a "current amplifier" must be as close as possible to the ideal current amplifier model, Fig. 7.2 (left). The criteria for quantifying the success of this aspiration are:

- The input impedance R_i has to be zero.
- The output impedance R_o has to be infinite.
- The current gain A_i has to be uniquely defined and constant.

Following the same analytical steps as in Sect. 7.1.2, we derive conditions required for the efficient transfer of the source current signal i_S to the output current i_{out} entering the load. A realistic current source is modelled using its equivalent Norton model, hence we include resistances R_S and R_o, as shown in Fig. 7.2 (right). Again, we easily recognize that, under conditions of $R_i \to 0$ and $R_o \to \infty$, the real current amplifier model, Fig. 7.2 (right), degenerates into the ideal current amplifier model, Fig. 7.2 (left).

Similarly to the voltage amplifier circuit network, which was analyzed by using a voltage divider model, we recognize that there is current divider created at the input terminals of a realistic current amplifier that is driven by a realistic current source, Fig. 7.2 (right), where a non-zero input voltage v_{in} develops across the input resistance R_i (which is also across $R_S||R_i$). Therefore, by inspection we write

$$i_s = \frac{v_{in}}{R_i||R_S} = \frac{R_S + R_i}{R_S}\frac{v_{in}}{R_i} = \frac{R_S + R_i}{R_S} i_{in}, \quad \therefore \quad A_i' = \frac{i_{in}}{i_S} = \frac{R_S}{R_S + R_i} = \frac{1}{1 + \frac{R_i}{R_S}}, \tag{7.11}$$

where A_i' is the current gain of the input current divider. We keep in mind that our main goal is to efficiently transfer the source current signal into the amplifier with no attenuation. Non-zero impedances at the source side create a resistive current divider, which causes a proportional reduction of the current signal i_S on its way to the amplifier's internal nodes. Again, we need to determine the severity of this attenuation and conclude under what conditions the current signal transfer is ideal.

The minimum current signal loss condition translates into $i_{in} = i_S$ (or equivalently, $A_i' = 1$). By inspection of (7.11), we easily conclude that there are two conditions that lead to lossless current transmission through the input side amplifier terminals, the first being

$$\lim_{R_i \to 0} A_i' = \frac{1}{1 + \frac{0}{R_S}} = 1, \tag{7.12}$$

that is, if an ideal current amplifier (one whose input impedance is $R_i = 0$) is driven by a real current source signal generator (one whose source impedance is $0 < R_S < \infty$) it means that there is no current attenuation, i.e., the complete source current i_S flows into the amplifier. The second limiting case is,

$$\lim_{R_S \to \infty} A_i' = \frac{1}{1 + \frac{R_i}{\infty}} = 1, \tag{7.13}$$

which means that if a real current amplifier (one whose input impedance is $0 < R_i < \infty$) is driven by an ideal current source (one whose source impedance is $R_S = \infty$), we again achieve lossless current transfer through the input current interface. The two conditions for perfect current transfer at the source side, (7.12) and (7.13), can be combined by stating that a current amplifier must have small input impedance relative to the signal source impedance, i.e.,

$$R_i \ll R_S. \tag{7.14}$$

Similarly, by inspection of the current amplifier's output terminals, Fig. 7.2 (right), it is straightforward to write

$$i_{out} = \frac{R_o}{R_o + R_L} A_i \, i_{in}, \quad \therefore \quad A_i'' = \frac{i_{out}}{i_{in}} = A_i \frac{R_o}{R_o + R_L} = A_i \frac{1}{1 + \frac{R_L}{R_o}}, \tag{7.15}$$

where A_i'' is the current gain of the output current divider. Non-zero impedances at the load side creates a resistive current divider that causes proportional reduction of the output current i_{out} on its way to the load terminals.

By inspection of (7.15), we can easily write the two conditions that would lead to $i_{in} = A_i \, i_{out}$,

$$\lim_{R_L \to 0} A_i'' = A_i \frac{1}{1 + \frac{0}{R_o}} = A_i, \tag{7.16}$$

$$\lim_{R_o \to \infty} A_i'' = A_i \frac{1}{1 + \frac{R_L}{\infty}} = A_i, \tag{7.17}$$

that is, either infinite output impedance R_o or zero load impedance is required if an amplifier is to maximize the current transfer at its output side.

The two conditions for perfect current transfer at the load side, (7.16) and (7.17), can be combined by stating that a current amplifier must drive a small load impedance relative to its own output impedance, i.e.,

$$R_L \ll R_o. \tag{7.18}$$

We now put together the conditions (7.11) and (7.15) and take a look at the total current gain that can be achieved with a real source, a real amplifier, and a real load chain. The total gain A_i from the source to the load terminals is

$$A_i = \frac{i_{out}}{i_s} = A_i' A_i'' = \frac{R_S}{R_S + R_i} A_i \frac{R_o}{R_o + R_L}, \tag{7.19}$$

which clearly shows how the three terms contribute to the total gain of a real amplifier. The first and third terms show attenuation caused by the input and the output current divider, and each is less than one. The second term is the only term possibly larger than one, and it represents the maximum possible current gain that could be achieved under the ideal condition of zero loss at the input and output terminals. The last statement, which in plain language summarizes (7.19), is the second key to our ability to evaluate the effectiveness of current amplifiers and the appropriateness of processing the signal in current form.

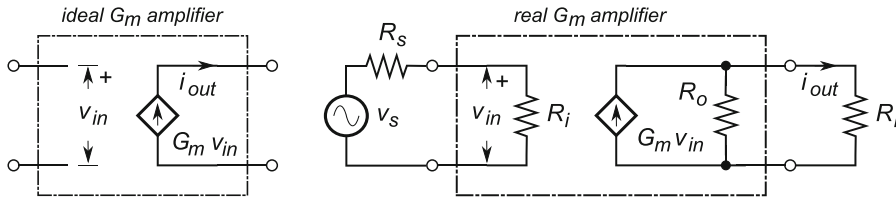


Fig. 7.3 An ideal transconductance (G_m) amplifier (*left*) and a realistic transconductance amplifier connected with input signal source and output load (*right*)

To summarize this section, we have concluded that high source impedance, low amplifier input impedance, high amplifier output impedance, and low load impedance together define a current amplifier. In short, (7.14) and (7.18) state that a current amplifier must have its input impedance much smaller than the source impedance and at the same time it must have its output impedance much larger than the load impedance. Depending how closely a circuit reaches these conditions, we quantify how good a current amplifier it is. We should note that we simply ignored the fact that power transfer, under the ideal current transfer conditions, is also zero. As with the ideal voltage amplifier, the crude approximation of the ideal current source helps us to perform mental circuit analysis.

7.1.4 Transconductance Amplifier

By definition, a transconductance (G_m) amplifier converts the input voltage signal v_in into the output current signal i_out, and it looks as if we took the input stage of a voltage amplifier and merged it with the output stage of a current amplifier. A functional symbol of an ideal G_m amplifier, Fig. 7.3 (left), shows the literal implementation of (7.1), which is based on an ideal voltage-controlled current source (VCCS) element whose current gain is G_m, measured in siemens (S), or electrical conductance that is derived as $^\mathrm{A}/_\mathrm{V} = {}^1/_\Omega$.

All the comments and conclusions that we have made about the input side of a voltage amplifier and about the output side of a current amplifier in the previous sections of this chapter still apply, which simplifies our analysis of this kind of amplifier.

By combining results in (7.2)–(7.19), we can write directly an expression for a real G_m amplifier gain as

$$G_\mathrm{m} = \frac{i_\mathrm{out}}{v_\mathrm{s}} = \frac{R_\mathrm{i}}{R_\mathrm{i} + R_\mathrm{S}}\, G_\mathrm{m}\, \frac{R_\mathrm{o}}{R_\mathrm{o} + R_\mathrm{L}} \tag{7.20}$$

and state that, in order to make an amplifier that would efficiently control the output current signal by the means of the input voltage signal, we need to make its input impedance much larger than the source impedance, i.e., $R_\mathrm{i} \gg R_\mathrm{S}$ and, at the same time, make its output impedance much larger than the load impedance, i.e., $R_\mathrm{o} \gg R_\mathrm{L}$. That is, an element or circuit aspiring to be called a "G_m amplifier" must be as close as possible to the ideal model, Fig. 7.3 (left). The criteria for quantifying the success of this aspiration are:

- The input impedance R_i has to be infinite.
- The output impedance R_o has to be infinite.
- The transconductance gain G_m has to be uniquely defined and constant.

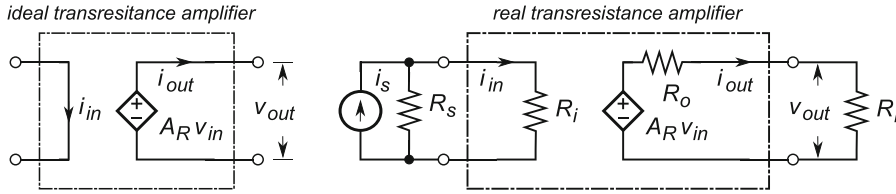


Fig. 7.4 An ideal transresistance (A_R) amplifier (*left*) and a realistic transresistance amplifier connected with input signal source and output load (*right*)

To illustrate how the important G_m stage is for the design of electronic circuits, let us just note that a single MOSFET transistor device is as close to the ideal g_m[1] stage as our technology allows. We remember that FET devices have extremely high input impedance (of the order of MΩ) and that, in active mode, the drain current i_D is controlled by the overdrive voltage at the input side $v_{OV} = (v_{GS} - V_{th})$, where V_{GS} is the gate–source voltage and V_{th} is the threshold voltage. That is, a FET device is a literal implementation of the transconductance gain definition, $i_D = g_m v_{OV}$.

7.1.5 Transresistance Amplifier

The fourth kind of amplifier is a transresistance (A_R) amplifier that, by definition, converts the input current signal i_{in} into the output voltage signal v_{out}. It looks as if we took the input stage of a current amplifier and merged it with the output stage of a voltage amplifier. A functional symbol of an ideal A_R amplifier, Fig. 7.4 (left), shows the literal implementation of (7.1), which is based on an ideal CCVS element whose current gain is A_R. The transresistance gain A_R is, by definition, measured in [V/A], i.e., in Ω.

As you have already guessed, the comments and conclusions that we have made about the input side of a current amplifier and the output side of a voltage amplifier in the previous sections of this chapter still apply.

By combining the results in (7.2)–(7.19), we can directly write an expression for a real A_R amplifier gain as

$$A_R = \frac{v_{out}}{i_s} = \frac{R_S}{R_i + R_S} A_R \frac{R_L}{R_o + R_L} \tag{7.21}$$

and state that, in order to make an amplifier that efficiently controls the output voltage signal by means of the input current signal, we need to make its input impedance much lower than the source impedance, i.e., $R_i \ll R_S$ and, at the same time, to make its output impedance much smaller than the load impedance, i.e., $R_o \ll R_L$. That is, an element or circuit aspiring to be called a "transresistance amplifier" must be as close as possible to the ideal model, Fig. 7.4 (left). The criteria for quantifying the success of this aspiration are:

- The input impedance R_i has to be zero.
- The output impedance R_o has to be zero.
- The transresistance gain A_R has to be uniquely defined and constant.

[1] We use the lower case "g" in g_m to indicate transconductance of a single device, as opposed to G_m for the full circuit.

The last circuit concludes our review of general amplifier types and completes the set of equations, (7.10), (7.19), (7.20), and (7.21) that we are going to use in the rest of the book along with our knowledge of resistive voltage dividers as a foundation for our mental circuit analysis skills.

Our first decision on how to classify a circuit as one of the four general amplifier types is based solely on evaluation by inspection and considering an ideal case of signal transfer. In the second pass, we use simple manual calculations. Finally, in the third pass, we use numerical simulators to quantify and confirm our initial estimates. A note of caution is in order: models used in this section still do not include frequency-dependent behaviour. However, if we keep in mind that in most cases resistances could be interpreted as impedances at a single frequency point, then the usefulness of the simplified models is preserved.

7.2 Single-Stage Amplifiers

In Sect. 7.1, we considered idealized amplifying functions from the most abstract perspective. Our goal was to determine how the external parameters influence the overall amplifier behaviour and to derive general rules of circuit interaction with the external world. It turns out that knowing the input and output impedances, the source and load impedances, and the internal gain factors is sufficient to specify conditions for four possible ways of amplifying the input signal. The exact details of the circuit's internal structure and the ways it may be implemented did not play any role in the analysis.

In this section, we continue to search for efficient ways of analyzing commonly used realistic amplifying circuits at the transistor level and of establishing practical rules and procedures for the mental analysis of amplifying circuits. Hence, we analyze the three main single-transistor amplifier topologies, i.e., common base (or common gate), common emitter (or common source), and common collector (or common drain), shown in Fig. 7.5, using both BJT and FET devices. In order to further develop our intuitive understanding of circuit operation, for each of the three single-stage amplifiers, we first derive the same sets of parameters as in Sect. 7.1, then we apply approximations and simplify the derived results so that we can easily apply them while doing circuit analysis "by inspection".

7.2.1 Common-Base Amplifier

We first analyze the common-base (CB) amplifier configuration, Fig. 7.6, in terms of its input resistance R_i, output resistance R_o, and voltage or current gain. The approach is based on treating a BJT as a "black box" and describing its properties strictly by observing the voltage and current relationships at each of its terminals. For the sake of simplicity, we do not show details of the amplifier biasing network: the transistor is assumed to be biased in the forward gain mode. A three terminal T-model is assumed (see Sect. 4.3.4.1).

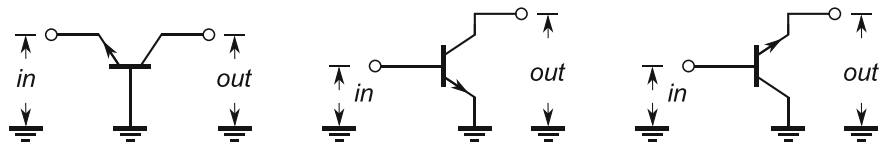


Fig. 7.5 Basic single-stage amplifiers: common gate (*left*), common emitter (*centre*), and common collector (*right*). Details of biasing are not shown, i.e., the ground symbols are *small signal* grounds

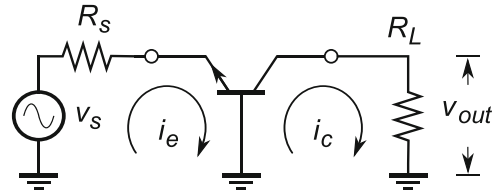


Fig. 7.6 Basic single-stage CB amplifier: It is driven by a voltage source v_S, whose internal resistance is R_S, and it drives resistive load R_L

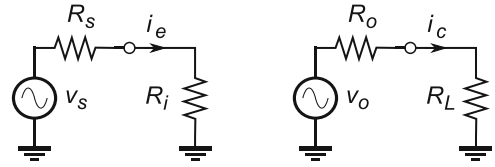


Fig. 7.7 Equivalent voltage divider schematics for a CB amplifier, looking into the input resistance (*left*) and the output resistance (*right*)

7.2.1.1 Input Resistance

By definition, the input current i_e, shown in Fig. 7.7 (left) is calculated as

$$i_e = \frac{v_S}{R_S + R_i}, \tag{7.22}$$

where R_i is the input resistance looking into the emitter node.

Assume that by looking into the emitter terminal, we see resistance R_e, by looking into the base terminal, we see resistance R_b, and by looking into the collector terminal, we see resistance R_c, then there are two voltage loops, on the input side and on the output side, whose KVL equations are

$$v_S = i_e(R_S + R_e + R_b) - i_c R_b, \tag{7.23}$$

$$0 = i_c(R_B + R_C + R_L) - i_e R_B - \alpha i_e R_C, \tag{7.24}$$

where i_e is the current entering the emitter terminal in the input branch and i_c is the current leaving the collector terminal in the output branch. Also, we use the simple relationships for BJT terminal currents:

$$i_e = i_b + i_c, \tag{7.25}$$

$$i_c = \alpha i_e, \tag{7.26}$$

$$i_c = \beta i_b. \tag{7.27}$$

Calculating current i_c from (7.24) and its substitution in (7.23) results in

$$i_e = \frac{v_S}{R_S + R_e + R_b - \frac{R_b(R_b + \alpha R_c)}{R_b + R_c + R_L}}. \tag{7.28}$$

From (7.22) and (7.28), it follows that

$$R_i = R_e + R_b - \frac{R_b(R_b + \alpha R_c)}{R_b + R_c + R_L}. \tag{7.29}$$

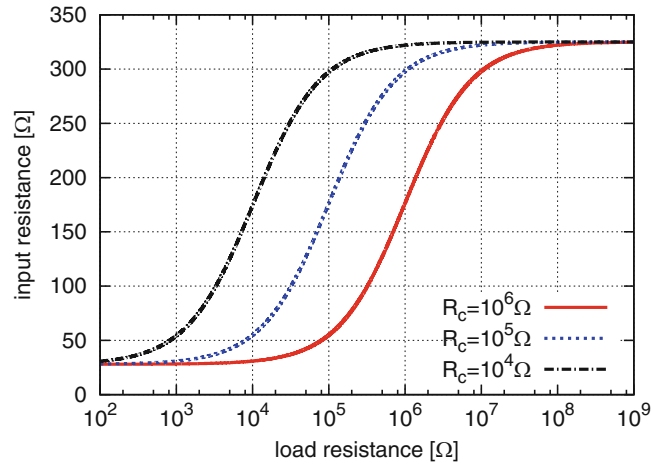


Fig. 7.8 Input resistance of a common-base amplifier as a function of collector resistance R_c for a typical case of $R_e = 25\,\Omega$, $R_b = 300\,\Omega$, and $\beta = 100$

Expression (7.29) shows that the input resistance R_i, under constant DC conditions, depends on the loading resistance R_L (see Fig. 7.8). To find out how large the spread of the R_i value is, we push (7.29) to two extremes: zero loading resistance, i.e., shorted output; and infinite loading resistance, i.e., removed load.

For $R_L = 0$, (7.29) becomes

$$R_i = R_e + R_b - \frac{R_b(R_b + \alpha R_c)}{R_b + R_c}, \tag{7.30}$$

which can be evaluated as follows. In BJT forward-bias mode, collector resistance R_c is much higher than the base resistance, i.e., $R_c \gg R_b$ and, because $\alpha \approx 1$, also $\alpha R_c \gg R_b$, which further simplifies (7.30) into

$$R_i = R_e + R_b - \frac{R_b(\alpha R_c)}{R_c} = R_e + R_b(1-\alpha) = R_e + \frac{R_b}{\beta} \approx R_e \tag{7.31}$$

for medium values of R_b and large values of β. Result (7.31) is very useful for our mental circuit analysis, because it is safe to say that, for low resistance loads, the input resistance of a CB amplifier is simply the emitter resistance, which is very low, e.g. if BJT is biased at $i_c = 1$ mA at room temperature, then $R_e \approx 25\,\Omega$.

For the case of open load, i.e., $R_L \to \infty$, (7.29) becomes simply

$$R_i = R_e + R_b \approx R_b, \tag{7.32}$$

that is, the input resistance is increased a bit, relative to the case of shorted output. At this stage, we can only take typical values for R_e and R_b and quantify the input resistance. For example, if BJT is biased at $i_c = 1$ mA at room temperature, then $R_e \approx 25\,\Omega$, while for small transistors the base resistance is of the order of a few hundred ohms, say $R_i \approx R_b = 300\,\Omega$ (Fig. 7.8).

7.2.1.2 Output Resistance

Referring to Fig. 7.7 (right), the output resistance R_o appears in the expression for the output current, which is the collector current i_C as

$$i_C = \frac{v_o}{R_L + R_o}, \tag{7.33}$$

where v_o is the equivalent output voltage generated by the CB stage at the collector terminal. We start again with (7.24) and write

$$i_e = \frac{R_C + R_B + R_L}{R_B + R_C} i_c, \tag{7.34}$$

which after substitution into (7.23) and a bit of rearranging of the terms, yields

$$i_c = \frac{\frac{R_b + \alpha R_c}{R_s + R_e + R_b} v_S}{R_L + R_b + R_c - \frac{R_b(R_b + \alpha R_c)}{R_S + R_e + R_b}} \tag{7.35}$$

then, by comparison of (7.33) and (7.35), it follows that

$$R_o = R_c + R_b - \frac{R_b(R_b + \alpha R_c)}{R_S + R_e + R_b}, \tag{7.36}$$

which shows that the output resistance depends on the source resistance. We find out how severe is this dependence by pushing (7.36) to two extremes: one for the ideal voltage source, i.e., $R_S = 0$, and one for the ideal current source, i.e., $R_S = \infty$. For the zero source resistance, (7.36) becomes

$$\begin{aligned} R_o &= R_c + R_b - \frac{R_b(R_b + \alpha R_c)}{R_e + R_b} = R_c + R_b \frac{(R_e + R_b) - (R_b + \alpha R_c)}{R_e + R_b} \\ &= R_c + R_b \frac{R_e - \alpha R_c}{R_e + R_b} = R_c + \frac{R_b R_e}{R_e + R_b} - \frac{R_b \alpha R_c}{R_e + R_b} \\ &\approx R_c - \frac{R_b \alpha R_c}{R_e + R_b} = R_c \frac{(R_e + R_b) - R_b \alpha}{R_e + R_b} \\ &= R_c \frac{R_e + R_b(1 - \alpha)}{R_e + R_b} \end{aligned} \tag{7.37}$$

because $(R_b R_e)/(R_e + R_b) = R_b || R_e < R_e \ll R_c$, and hence can be neglected relative to the collector resistance R_c. The ratio in the last term of (7.37) depends on the small percentage of the base resistance R_b, which is comparable with the emitter resistance R_e, i.e., it stands for a relatively small number divided by a number that is a bit larger, hence it should stay. As a quick estimate, let us use $R_e = 25\,\Omega$, $R_b = 300\,\Omega$, $\beta = 100$, and $R_c = 1\,\text{M}\Omega$; then, from (7.37), we calculate $R_o \approx 100\,\text{k}\Omega$ (Fig. 7.9).

At the other extreme, $R_S \to \infty$, expression (7.36) is simply reduced to

$$R_o = R_c + R_b \approx R_c, \tag{7.38}$$

which, for our numerical example gives $R_o \approx 1\,\text{M}\Omega$. That is, depending on the source resistance or the output resistance may change by a factor of ten (Fig. 7.9).

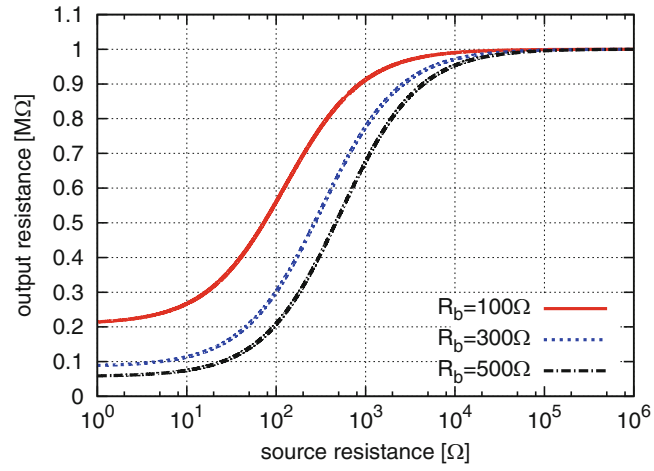


Fig. 7.9 Output resistance of a common-base amplifier as a function of collector resistance R_c for a typical case of $R_e = 25\,\Omega$, $R_c = 1\,\mathrm{M}\Omega$, and $\beta = 100$

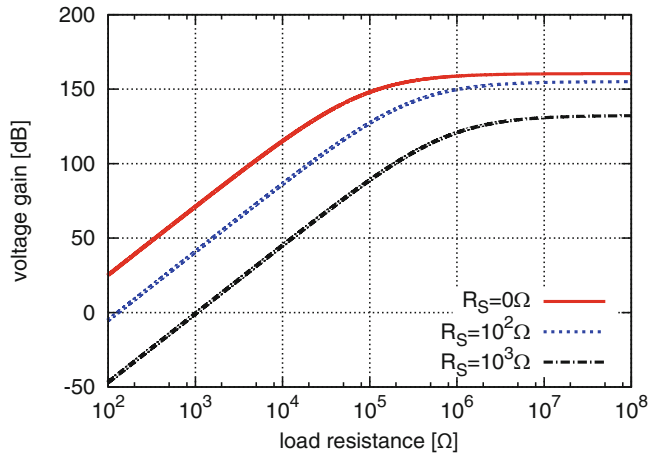


Fig. 7.10 Voltage gain of a common-base amplifier as a function of load resistance R_L for a typical case of $R_e = 25\,\Omega$, $R_c = 1\,\mathrm{M}\Omega$, and $\beta = 100$

7.2.1.3 Voltage Gain

The voltage gain of a CB amplifier, Fig. 7.6, is written as

$$A_v = \frac{v_{out}}{v_S} = \frac{i_c R_L}{v_S} = \frac{\frac{R_b+\alpha R_c}{R_s+R_e+R_b} R_L}{R_L + R_b + R_c - \frac{R_b(R_b+\alpha R_c)}{R_S+R_e+R_b}}$$
$$= \frac{(R_b+\alpha R_c) R_L}{(R_s+R_e+R_b)(R_L+R_b+R_c) - R_b(R_b+\alpha R_c)}, \tag{7.39}$$

where we substituted the expression for the collector current i_C from (7.35); a plot of the voltage gain is shown in Fig. 7.10.

A useful approximation is to assume large collector resistance, R_C, i.e., $\alpha R_C \gg R_b$ and $R_C \gg R_L$, so that (7.39) becomes

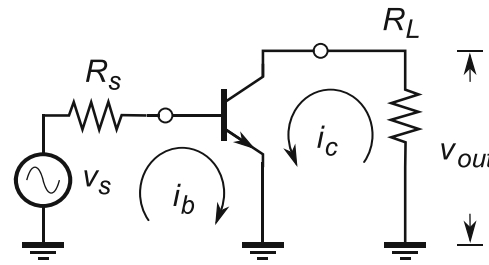


Fig. 7.11 A common-emit -ter amplifier. For the sake of simplicity the biasing network is not shown, therefore the ground terminals are small signal grounds

$$A_v \approx \frac{\alpha R_c R_L}{(R_s + R_e + R_b) R_c - R_b \alpha R_c}$$
$$\approx \frac{R_L}{R_s + R_e + \dfrac{R_b}{\beta}}$$
$$= \frac{R_L}{R_S + R_i}, \tag{7.40}$$

after substituting the second-last form from (7.31). It is also very useful to note that for $R_S = 0$, the source voltage is $V_S = V_{BE}$ for a BJT (Fig. 7.6), hence we have another useful expression for the voltage gain as

$$A_v = \frac{v_{out}}{v_{BE}} = \frac{R_L}{R_i}, \tag{7.41}$$

which indicates that, because the input resistance of a CB amplifier is typically very low, we need to keep the source resistance high and drive the CB stage with a current signal. The loading resistance also needs to be high.

7.2.2 Common-Emitter Amplifier

A basic common-emitter (CE) amplifier (no biasing details are shown, i.e., it is assumed to be in the forward-gain active mode) is driven by a realistic voltage source v_S with internal resistance R_S (see Fig. 7.11). Loading resistance is connected to the collector node, with the input and output branch currents as indicated. Hence, the three current relationships are $i_e = i_c + i_b$.

We will work out simplified expressions for the input resistance, output resistance, and voltage gain.

7.2.2.1 Input Resistance

Using the same approach as in Sect. 7.2.1.1, for the input branch we can write that

$$i_b = \frac{v_S}{R_S + r_i}. \tag{7.42}$$

For the input and output branch voltage loops (Fig. 7.11), the KVL equations are

$$v_S = i_b(R_S + r_b + r_e) + i_c r_e, \tag{7.43}$$
$$0 = i_c(R_L + r_c + r_e) + i_b r_b - \alpha i_e r_c. \tag{7.44}$$

From (7.44), we can find an expression for the current gain β, after substituting $i_e = i_c + i_b$, as

$$\beta = \frac{i_c}{i_b} = -\frac{r_e - \alpha r_c}{R_L + r_e + r_c(1-\alpha)}. \tag{7.45}$$

The current gain β can be evaluated for the extreme values of the load R_L. First, for $R_L = 0$

$$\beta = -\frac{r_e - \alpha r_c}{r_e + r_c(1-\alpha)} \approx -\frac{\alpha}{1-\alpha} \approx -\frac{1}{1-\alpha}, \tag{7.46}$$

after assuming that $\alpha r_c \gg r_e$ and $\alpha \approx 1$. This result shows the gain of the unloaded BJT; the negative sign is because of the i_c and i_b current directions. At the other extreme, $R_L \to \infty$, the current gain, as expected, drops to $\beta = 0$ (because infinite load means that $i_c = 0$).

After substituting (7.46) into (7.43) and a bit of rearranging of the terms, we come to the following expression

$$i_b = \frac{v_S}{R_S + r_b + r_e + \frac{r_e(\alpha r_c - r_e)}{R_L + r_e + (1-\alpha)r_c}}. \tag{7.47}$$

After comparison with (7.42), it follows that

$$r_i = r_b + r_e + \frac{r_e(\alpha r_c - r_e)}{R_L + r_e + (1-\alpha)r_c}, \tag{7.48}$$

which, again, shows that the input impedance does depends on the load. We estimate how much the input resistance changes by checking the two extreme cases of the load. In the case of shorted load, i.e., $R_L = 0$, (7.48) becomes

$$r_i = r_b + r_e + \frac{r_e(\alpha r_c - r_e)}{r_e + (1-\alpha)r_c} = r_b + r_e \frac{1}{(1-\alpha)} = r_b + \beta r_e, \tag{7.49}$$

which is an important result for us because it shows that the emitter branch resistance is mapped to the input resistance after being multiplied by a factor of β. This is often known as the "emitter resistor magnification factor" and it may easily become the dominant term (especially if there is an external large emitter resistor R_E in the emitter branch). Another important conclusion is that, for light loads, the input resistance strongly depends on β (Fig. 7.12).

At the second extreme, $R_L \to \infty$, i.e., an unloaded CE amplifier, we have

$$r_i = r_b + r_e \approx r_b, \tag{7.50}$$

which is constant. For a typical numerical example of $r_e = 25\,\Omega$, $r_b = 300\,\Omega$, $r_c = 1\,\mathrm{M}\Omega$, and $\beta = 100$, dependence of the input resistance versus the load is shown in Fig. 7.12.

7.2.2.2 Output Resistance

The collector current of a CE amplifier (Fig. 7.11) is

$$i_c = \frac{v_o}{r_o + R_L}, \tag{7.51}$$

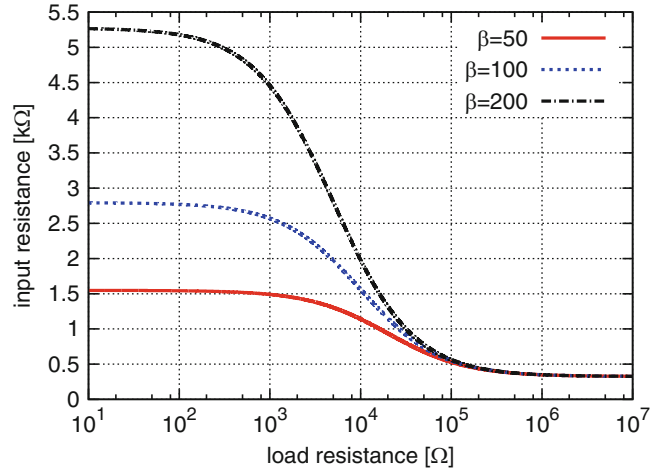


Fig. 7.12 Input resistance of a common-emitter amplifier as a function of load resistance R_L for a typical case of $R_\mathrm{e} = 25\,\Omega$, $R_\mathrm{c} = 1\,\mathrm{M}\Omega$

where v_o is the internal BJT voltage source driving the BJT output resistance r_o and the load R_L. From (7.43) and (7.44), after a bit of rearranging of the terms, we write the expression for collector current as

$$i_\mathrm{c} = \frac{\frac{\alpha r_\mathrm{c} - r_\mathrm{e}}{R_\mathrm{S} + r_\mathrm{b} + r_\mathrm{e}} v_\mathrm{S}}{R_\mathrm{L} + r_\mathrm{e} + r_\mathrm{c}(1-\alpha) + \frac{r_\mathrm{e}(\alpha r_\mathrm{c} - r_\mathrm{e})}{R_\mathrm{S} + r_\mathrm{b} + r_\mathrm{e}}}, \tag{7.52}$$

which, after comparing (7.51) and (7.52) yields

$$r_\mathrm{o} = r_\mathrm{e} + r_\mathrm{c}(1-\alpha) + \frac{r_\mathrm{e}(\alpha r_\mathrm{c} - r_\mathrm{e})}{R_\mathrm{S} + r_\mathrm{b} + r_\mathrm{e}}. \tag{7.53}$$

Therefore, the output resistance decreases as the source resistance R_S increases. In order to estimate the boundary values for the output resistance, we take a look at the two extremes of the source resistance, $R_\mathrm{S} = 0$ and $R_\mathrm{S} \to \infty$.

For $R_\mathrm{S} = 0$, (7.53) becomes

$$\begin{aligned} r_\mathrm{o} &= r_\mathrm{e} + r_\mathrm{c}(1-\alpha) + \frac{r_\mathrm{e}(\alpha r_\mathrm{c} - r_\mathrm{e})}{r_\mathrm{b} + r_\mathrm{e}} \\ &= r_\mathrm{c}(1-\alpha) + r_\mathrm{e}\,\frac{(r_\mathrm{b} + r_\mathrm{e}) + (\alpha r_\mathrm{c} - r_\mathrm{e})}{r_\mathrm{b} + r_\mathrm{e}} \\ &= r_\mathrm{c}\frac{r_\mathrm{e} + (1-\alpha) r_\mathrm{b}}{r_\mathrm{b} + r_\mathrm{e}} = r_\mathrm{c}\frac{r_\mathrm{e} + \dfrac{r_\mathrm{b}}{\beta}}{r_\mathrm{b} + r_\mathrm{e}} \approx r_\mathrm{c}, \end{aligned} \tag{7.54}$$

which shows mild dependence on the β factor.

For $R_\mathrm{S} = \infty$, (7.53) becomes simply

$$r_\mathrm{o} = r_\mathrm{e} + r_\mathrm{c}(1-\alpha) = r_\mathrm{e} + \frac{r_\mathrm{c}}{\beta}. \tag{7.55}$$

For typical values of a BJT, we create the plot in Fig. 7.13, which shows that larger gain factors of β mildly reduce the output impedance.

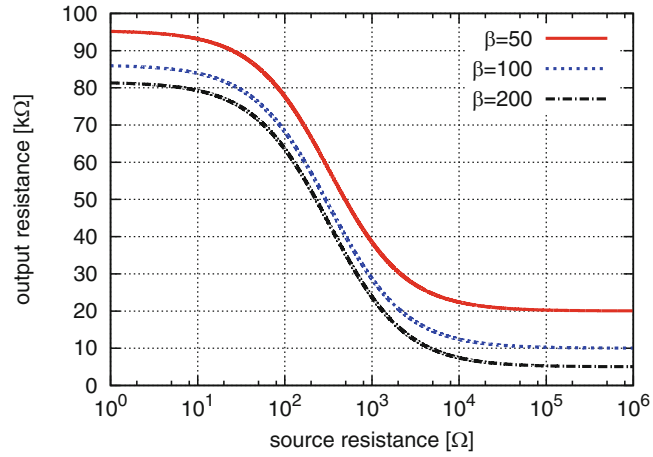


Fig. 7.13 Output resistance of a common-emitter amplifier as a function of load resistance R_L for a typical case of $R_e = 25\,\Omega$, $R_c = 1\,\mathrm{M}\Omega$

7.2.2.3 Voltage Gain

The voltage gain of a CE amplifier is, by definition

$$A_v = \frac{v_{out}}{v_S} = \frac{i_c R_L}{v_S}, \tag{7.56}$$

hence, after substituting (7.52) into (7.56) we have

$$A_v = \frac{\frac{\alpha r_c - r_e}{R_S + r_b + r_e} R_L}{R_L + r_e + r_c(1-\alpha) + \frac{r_e(\alpha r_c - r_e)}{R_S + r_b + r_e}}. \tag{7.57}$$

Let us now take a look at a CE amplifier driven by an ideal voltage source $R_S = 0$ and large loads $R_L \to \infty$, i.e., $R_L \gg r_c(1-\alpha)$. That simplifies (7.52) as

$$\begin{aligned} A_v &= \frac{\frac{\alpha r_c - r_e}{r_b + r_e} R_L}{R_L + r_e + r_c(1-\alpha) + \frac{r_e(\alpha r_c - r_e)}{r_b + r_e}} \\ &= \frac{(\alpha r_c - r_e) R_L}{(r_b + r_e)(R_L + r_e + r_c(1-\alpha)) + r_e(\alpha r_c - r_e)} \\ &\approx \frac{\alpha r_c R_L}{(r_b + r_e)(R_L) + r_e \alpha r_c} = \frac{\alpha r_c R_L}{r_b R_L + r_e(R_L + \alpha r_c)} \\ &\approx \frac{\alpha r_c R_L}{r_b R_L + r_e R_L} = \frac{\alpha r_c}{r_b + r_e} \approx \frac{r_c}{r_b + r_e}, \end{aligned} \tag{7.58}$$

where $\alpha \approx 1$ in the last approximation. Expression (7.58) is important to us because we see that, for the ideal voltage source drive and large loads, it is safe to say that the voltage gain of the CE amplifier is bounded by (i.e., is less than) the ratio of the collector resistance and emitter resistance r_c/r_e, which is a very useful result for quick estimates. For a set of typical device values, voltage gain versus load is shown in Fig. 7.14.

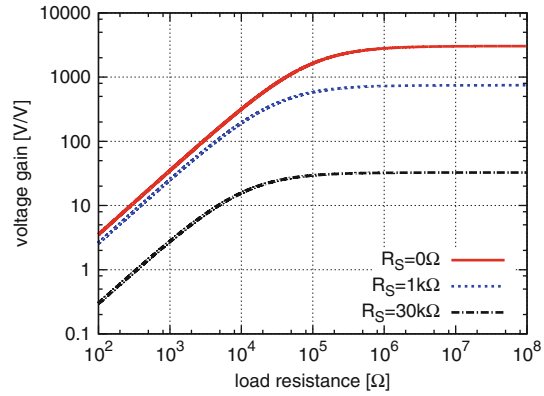


Fig. 7.14 Voltage gain of a common-emitter amplifier as a function of load resistance R_L for a typical case of $R_e = 25\,\Omega$, $R_c = 1\,M\Omega$, and $\beta = 100$

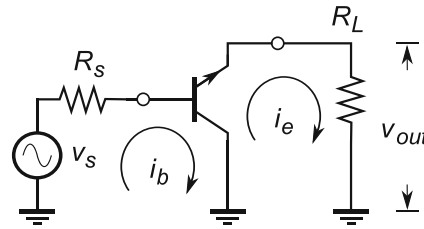


Fig. 7.15 A common-coll -ector amplifier stage. For simplicity, the biasing network is not shown, therefore the ground terminals are small signal grounds

7.2.3 Common-Collector Amplifier

Using current notation in Fig. 7.15, the three terminal currents are related as:

$$i_c = i_e - i_b. \tag{7.59}$$

Accordingly, the KVL equations for the common-collector (CC) amplifier circuit in Fig. 7.15 are:

$$v_S = i_b(r_b + r_c + R_S) + \alpha i_e r_c - i_e r_e, \tag{7.60}$$

$$0 = i_e(r_c + r_e + R_L) - \alpha i_e r_c - i_b r_c. \tag{7.61}$$

Hence, from (7.61), we write an expression for the current gain of a CC amplifier

$$\frac{i_e}{i_b} = \frac{r_c}{r_c + r_e + R_L - \alpha r_c} = \frac{r_c}{r_c(1-\alpha) + r_e + R_L} \approx \frac{1}{(1-\alpha)}. \tag{7.62}$$

For current signal transfer, the load needs to be small, that is, approximation $r_c(1-\alpha) \gg R_L, r_e$ is appropriate for this case.

7.2.3.1 Input Resistance

Applying the same approach as in the previous sections, we write an expression for the base current as

$$i_b = \frac{v_S}{R_S + r_i}, \tag{7.63}$$

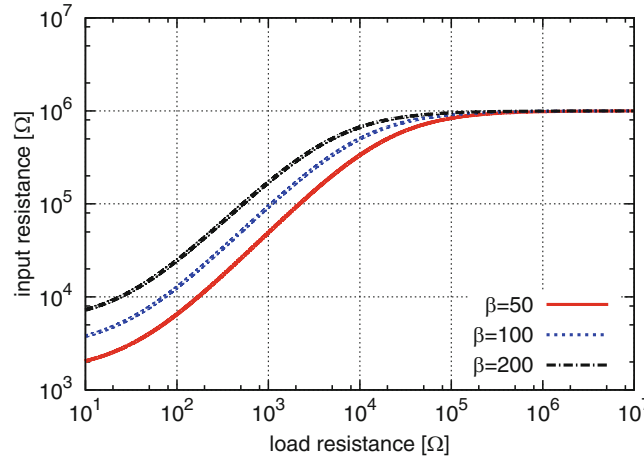


Fig. 7.16 Input resistance of a common-collector amplifier as a function of load resistance R_L for a typical case of $R_e = 25\,\Omega$ and $R_c = 1\,\mathrm{M}\Omega$

where r_i is the circuit input resistance as seen by the source side. From (7.60) and (7.61), after rearranging the terms, we write an expression for the base current as

$$i_b = \frac{v_S}{R_S + r_b + r_c - \frac{r_c^2(1-\alpha)}{r_c(1-\alpha)+r_e+R_L}}. \tag{7.64}$$

By comparison of (7.63) and (7.64), we write the expression for the input resistance of a CC amplifier as

$$r_i = r_b + r_c - \frac{r_c^2(1-\alpha)}{r_c(1-\alpha)+r_e+R_L}, \tag{7.65}$$

which shows strong dependence on the collector resistance r_c, in addition to its dependence on the loading resistance.

Let us find out how much the input resistance changes under extreme conditions of the loading resistance. In the case of the ideal source, $R_L = 0$, i.e., shorted output, after applying approximation $r_e \ll r_c(1-\alpha)$, relation (7.65) becomes

$$\begin{aligned} r_i &= r_b + r_c - \frac{r_c^2(1-\alpha)}{r_c(1-\alpha)+r_e} = r_b + \frac{r_c^2(1-\alpha)+r_c r_e - r_c^2(1-\alpha)}{r_c(1-\alpha)+r_e} \\ &\approx r_b + \frac{r_e}{1-\alpha} = r_b + \beta r_e, \end{aligned} \tag{7.66}$$

which is, as expected, the same result as we found for the CE amplifier (it is the same circuit from the source side).

At the other extreme, for the case of a disconnected load, i.e., $R_L = \infty$, the input resistance (7.65) is simply

$$r_i = r_b + r_c \approx r_c. \tag{7.67}$$

For a typical BJT device, the input resistance of a CC amplifier changes as in Fig. 7.16.

7.2.3.2 Output Resistance

The output current i_e is generated by the internal device voltage v_o that is driving its internal output resistance r_o and the load resistance R_L, i.e.

$$i_e = \frac{v_o}{R_L + r_o}, \tag{7.68}$$

where an expression for the emitter current is derived from (7.60) and (7.61) and arranged as

$$i_e = \frac{\frac{r_c}{r_b + r_c + R_S} v_S}{R_L + r_e + r_c(1-\alpha) - \frac{r_c^2(1-\alpha)}{r_b + r_c + R_S}}. \tag{7.69}$$

By comparison of (7.68) and (7.69), we conclude that the output resistance is

$$r_o = r_e + r_c(1-\alpha) - \frac{r_c^2(1-\alpha)}{r_b + r_c + R_S}, \tag{7.70}$$

which shows dependence on the source resistance. We will evaluate the influence of the source resistance by considering the two extreme cases. For ideal source, $R_S = 0$, we further write

$$\begin{aligned} r_o &= r_e + \frac{r_c(1-\alpha)(r_b + r_c) - r_c^2(1-\alpha)}{r_b + r_c} \\ &\approx r_e + \frac{r_c(1-\alpha)r_b}{r_c} = r_e + \frac{r_b}{\beta} \approx r_e \approx \frac{1}{g_m}, \end{aligned} \tag{7.71}$$

which is very important approximation, because we conclude that the output resistance of a CC amplifier is very low, i.e., it may serve as a good voltage driver.

For the case of $R_S = \infty$, expression (7.70) becomes

$$r_o = r_e + r_c(1-\alpha) \approx \frac{r_c}{\beta}. \tag{7.72}$$

For typical numerical example, the variation of the output resistance with the source resistance is shown in Fig. 7.17.

7.2.3.3 Voltage Gain

By definition, the voltage gain of a CC amplifier, Fig. 7.15, is

$$A_v = \frac{v_{out}}{v_S} = \frac{i_e R_L}{v_S}, \tag{7.73}$$

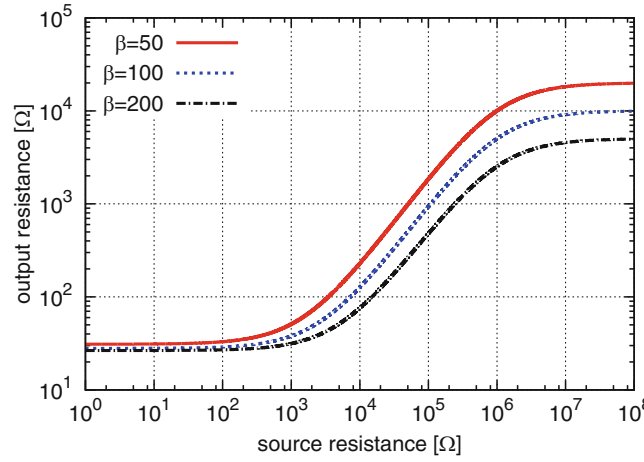


Fig. 7.17 Output resistance of a common-collector amplifier as a function of load resistance R_L for a typical case of $R_\mathrm{e} = 25\,\Omega$ and $R_\mathrm{c} = 1\,\mathrm{M}\Omega$

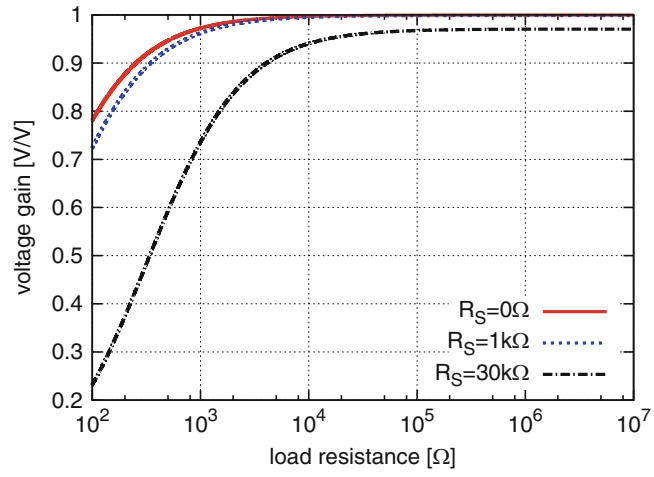


Fig. 7.18 Voltage gain of a common-collector amplifier as a function of load resistance R_L for a typical case of $R_\mathrm{e} = 25\,\Omega$, $R_\mathrm{c} = 1\,\mathrm{M}\Omega$, and $\beta = 100$

which can be further derived by substituting (7.69) into (7.73), hence

$$\begin{aligned} A_\mathrm{v} &= \frac{\frac{r_\mathrm{c}}{r_\mathrm{b}+r_\mathrm{c}+R_\mathrm{S}}\,R_\mathrm{L}}{R_\mathrm{L}+r_\mathrm{e}+r_\mathrm{c}(1-\alpha)-\frac{r_\mathrm{c}^2(1-\alpha)}{r_\mathrm{b}+r_\mathrm{c}+R_\mathrm{S}}} \\ &= \frac{r_\mathrm{c}\,R_\mathrm{L}}{(r_\mathrm{b}+r_\mathrm{c}+R_\mathrm{S})\,[R_\mathrm{L}+r_\mathrm{e}+r_\mathrm{c}(1-\alpha)]-r_\mathrm{c}^2(1-\alpha)} \\ &\approx \frac{r_\mathrm{c}\,R_\mathrm{L}}{(r_\mathrm{c}+R_\mathrm{S})\,[R_\mathrm{L}+r_\mathrm{c}(1-\alpha)]-r_\mathrm{c}^2(1-\alpha)} \\ &= \frac{R_\mathrm{L}}{R_\mathrm{L}\left(1+\dfrac{R_\mathrm{S}}{r_\mathrm{c}}\right)+R_\mathrm{S}(1-\alpha)} \approx \frac{R_\mathrm{L}}{R_\mathrm{L}+\dfrac{R_\mathrm{S}}{\beta}} \approx 1. \end{aligned} \tag{7.74}$$

This approximation is also important to us. Under normal operation, a CC amplifier has a voltage gain of one, which makes it suitable to serve as a "voltage buffer", or impedance converter (high input impedance and low output impedance). For our numerical example, the voltage gain plot is shown in Fig. 7.18.

7.3 Cascode Amplifier

The "cascode" amplifier, so important that it has its own name, is a combination of two single-stage amplifiers, the common emitter followed by the common base (in Fig. 7.19, the single-stage amplifiers are connected at the ① node). Let us intuitively analyze this useful structure and conclude its main characteristics.

- *Input resistance*: The first stage of a cascode amplifier is a standard CE stage. We need to notice that the Q_1 collector, i.e., the CE output, is connected to Q_2 emitter, i.e., the CB input. We have already learned that the input resistance of a CB amplifier is low for loading resistors $R_\mathrm{L} \ll r_{\mathrm{c}2}$. According to Fig. 7.12, that means the input resistance of a CE amplifier is on the high side, in the order of 3 kΩ for $\beta = 100$.
- *Output resistance*: The common base stage is driven by the Q_1 collector, which represents high resistance, therefore the overall output resistance is high, i.e., the Q_2 collector also mimics a current source. More detailed analysis shows that the cascode output resistance is $r_{\mathrm{o}2} = \beta r_{\mathrm{o}1} = g_{\mathrm{m}2} r_{\mathrm{o}1} r_{\mathrm{o}2}$, which is a significant increase relative to the output resistance of either a CE or a CB amplifier separately.
- *Voltage gain*: The second, CB, stage acts as a load to the first stage. The voltage gain of a CE stage followed by a CB gain, after using an ideal voltage source with resistance $R_\mathrm{S} = 0$, is

$$A_\mathrm{v} = \frac{v_\mathrm{out}}{v_\mathrm{S}} = \frac{v_\mathrm{out}}{v_{o1}} \frac{v_{o1}}{v_\mathrm{S}} = \frac{R_\mathrm{L}}{r_{\mathrm{e}1}} \frac{-r_{\mathrm{e}2}}{r_{\mathrm{e}1}} = -\frac{R_L}{r_{\mathrm{e}1}} = -g_\mathrm{m} R_\mathrm{L}, \tag{7.75}$$

  which is the same as for the single-stage CE amplifier. If this conclusion comes as a surprise to you, just remember that although the voltage gain of a CE amplifier with low resistance load is low (around one or so, as shown in Fig. 7.14), its collector current is passed through the CB stage with approximately unity gain, hence the CE voltage gain is realized after the Q_2 collector current is delivered to the load resistor R_L (which serves as a current-to-voltage amplifier). Adding the CB stage on the path between the load and the Q_1 collector did not change anything in terms of the voltage gain; remember that a CB amplifier serves as a current buffer, i.e., the input current at the emitter shows up as the output current at the Q_2 collector (i.e., $i_\mathrm{C}(Q_1) \approx i_\mathrm{C}(Q_2)$). Therefore, from the perspective of the loading resistor R_L nothing has changed, the same current is converted into a voltage.

To summarize, although the voltage gain of a cascode amplifier is the same as for a single-stage CE, its output resistance is increased. The increased resistance is interpreted as a current source that is much closer to the ideal one (which has the output resistance equal to infinity). Therefore, almost by default, realistic current sources are made out of cascode stages. Another important application of the cascode amplifier is related to RF applications and the "Miller effect" (described in Sect. 7.6).

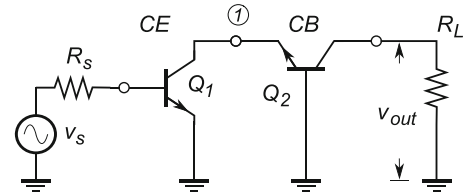


Fig. 7.19 A cascode amplifier stage. For simplicity, the biasing network is not shown, therefore the ground terminals are small signal grounds

7.4 The Biasing Problem

So far in our circuit analysis, we have ignored all details related to the device biasing and simply assumed that the device is somehow set to a stable DC operating point, and the details of how exactly that was achieved and maintained were just left out. That approach makes perfect sense because once the biasing point of a transistor is set, i.e., its g_m is set, then only its small signal behaviour is relevant—not the biasing details. However, we should not conclude that details of the biasing network are simple and unimportant. If anything, we keep in mind that the design of a biasing network is a fundamental and non-trivial issue that needs to be taken seriously. In this section, we go over the evolutional development of a typical BJT biasing setup, and eventually reach conclusions about what constitutes a good biasing setup and why.

Let us again take a look at a single BJT device and the external setup required to keep it operational. As we learned in Sect. 4.3.4, a BJT device requires two independent voltage sources for its operation. One voltage source V_{BE} is required to set up the current through the forward-biased, base–emitter diode and the second voltage source is required to keep the reverse-biasing voltage across the collector–base diode by making sure that $V_C \geq V_B$, i.e., the two voltage sources must be related $V_{CE} \geq V_{BE}$, as shown in Fig. 7.20 (left).

The sole purpose of this arrangement is to set up a stable (V_C, I_C) operating point of the BJT device, and therefore its g_m gain, for a given base–emitter voltage V_{BE} (see graph in Fig. 4.40).

This is where our biasing problems start. Fundamentally, a BJT device behaves as a "current-controlled" current source, where the output, i.e., the collector, current is controlled by the input, i.e., the base, current through the $I_C = \beta I_B$ relationship, where β is the current gain of the device. Unfortunately, our manufacturing technology is not ideal and, therefore, there are at least three main problems relevant to electronic circuit design:

- A BJT current gain factor β depends both on device geometry and on the manufacturing parameters. Unavoidably, the two have certain processing variations, which leads to β variations as large as $\pm 50\%$ around the nominal value that it was designed for. That is, if the original design was targeting, for example, $\beta = 100$, it is realistic to expect any value between $\beta = 50$ and $\beta = 150$ for a large number of tested devices. The final consequence is that the overall circuit gain, that was expected to be *fixed*, would have its minimum value three times smaller than its maximum value, i.e., the realistic gain variation would be in the order of 300%, which renders the device practically useless.
- A little less obviously, g_m also depends on temperature variations. Detailed device analysis shows that the base–emitter voltage V_{BE} changes at the rate of 2.5 mV/°C. On the other hand, high-reliability electronic devices must satisfy, for example, a military and airspace standard

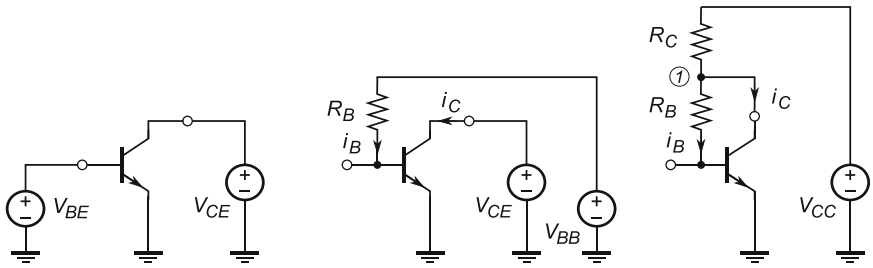


Fig. 7.20 Evolution of the biasing network required to set up a BJT device: a literal biasing network implementation with two voltage sources, V_{BE} and V_{CE} (*left*); a base current source (V_{BB}, R_B) implementation (*centre*); and a single voltage source V_{CC} implementation (*right*)

which is specified within an environmental temperature range of $T_1 = -55°C$ to $T_2 = +125°C$ (sometimes $-65°C$ to $+175°C$). Just to find out how drastic this requirement really is, let us take, for example, a temperature variation of $\Delta T = 180°C$, which causes the base–emitter V_{BE} voltage to change by $\Delta V_{BE} = 180°C \times 2.5\,mV/°C = 450\,mV/$. We know that collector current and the base–emitter voltage V_{BE} are related through the exponential function. Therefore, we can estimate using typical transistor parameters for two V_{BE} voltages as $V_{BE}(T_2) = 0.925\,mV$ and $V_{BE}(T_1) = 0.475\,mV$ with, for the given temperature range, $V_T(T_2) = kT_2/q = 34.31\,mV$ and $V_T(T_1) = 18.8\,mV$ (assuming constant saturation current I_S because the temperature variation is already accounted for in the ΔV_{BE}), hence we write

$$I_{C1}(T_1) \approx I_S \exp \frac{V_{BE1}}{V_{T1}} \quad \text{and} \quad I_{C2}(T_2) \approx I_S \exp \frac{V_{BE2}}{V_{T2}},$$

$$\therefore$$

$$\frac{I_{C2}(T_2)}{I_{C1}(T_1)} = \exp\left(\frac{V_{BE2}}{V_{T2}} - \frac{V_{BE1}}{V_{T1}}\right) \approx 5.4, \tag{7.76}$$

which directly translates into the overall circuit gain variation. Again, the device is not that useful as an amplifier without some form of external mechanism for stabilizing the DC biasing point.

- The combination of component aging, leakage currents in active devices, and other secondary effects of the IC technology amounts to an inconsistent and unpredictable variation of the current gain, which must also be evaluated and compensated for by some external active mechanism.

For a given design, one could always find (at least temporarily) the right combination of voltage source V_{BB} and base resistor R_B and adjust the base current i_B to generate the desired collector current i_C, Fig. 7.20 (centre). However, this is a tedious manual procedure that would have to be repeated for every single transistor in every single design and would not last long; this is neither efficient nor elegant engineering. In addition, we note that there are at least two fundamental problems with the approach: it fixes the base current instead of the collector current and it requires two independent power supplies (which are both bulky and expensive).

The biasing problem is a typical example of a whole category of problems in nature that require some sort of continuous monitoring and error-correcting mechanism. Using engineering terminology, the solution requires the design of a "negative feedback control system".

In principle, if we manage to inject a small sample of the collector current into the base in such a way that any increase of the collector current causes the base to oppose the change and to force reduction of the collector current, and hence to hold the collector current's mean value constant, then we have realized the control system. Practical realization of this principle looks like the schematic arrangement in Fig. 7.20 (right). Let us take a closer look at how exactly it works. First, any increase of the collector current i_C causes an increase in the voltage across the R_C resistor. Therefore, the potential at node ① is reduced because $V(1) = V_{CC} - i_C R_C$. In other words, the base current is reduced because $i_B = (v(1) - V_B)/R_B$. Now, it becomes obvious, the reduced base current forces reduction of the collector current because $i_C = \beta i_B$, opposing its initial increase. With the right combination of R_C, R_B, and β the control loop maintains the collector current's mean value permanently.

As a side benefit, the basic biasing control circuit in Fig. 7.20 (right) works with only one power supply source (for the price of one additional resistor). A truly simple and elegant solution. Alternatively, if the R_C resistor is moved along its branch through the V_{CC} source into the emitter branch, it becomes the emitter resistor R_E that keeps supporting the feedback mechanism. That variant of stabilizing the DC biasing point with emitter resistor R_E is known as "emitter degeneration" and is used almost all the time.

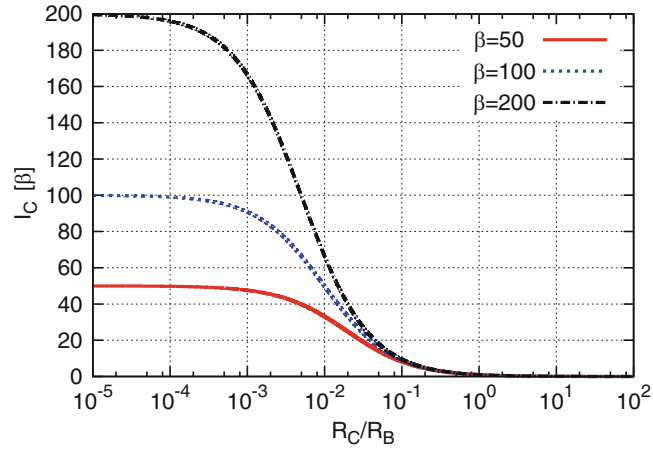


Fig. 7.21 Sensitivity of collector current I_C normalized to $(V_{CC} - V_{BE})$ and R_C/R_B; the current is in the units of β

Let us evaluate the effectiveness of the feedback mechanism and find out how the collector current in Fig. 7.20 (right) becomes approximately independent of the β factor. First, we state the current and voltage relations as

$$I_C = \beta I_B, \tag{7.77}$$

$$V_{CC} = R_C(I_C + I_B) + R_B I_B + V_{BE}, \tag{7.78}$$

where solving these two equations for i_C leads to

$$I_C = \frac{\frac{\beta(V_{CC}-V_{BE})}{R_B}}{1+(\beta+1)\frac{R_C}{R_B}}. \tag{7.79}$$

To find out how the ratio of the base and collector resistors R_B/R_C influences the feedback mechanism, let us look at the extreme cases of (7.79): $R_C/R_B \to 0$, i.e., either R_C is small or R_B is large (in other words, $R_C \ll R_B$); and $R_C/R_B \to \infty$, i.e., either R_C is large or R_B is small (in other words, $R_C \gg R_B$).

For $R_C \to 0$, we have

$$\lim_{R_C \to 0} I_C = \lim_{R_C \to 0} \frac{\frac{\beta(V_{CC}-V_{BE})}{R_B}}{1+(\beta+1)\frac{R_C}{R_B}} = \frac{\beta(V_{CC}-V_{BE})}{R_B}, \tag{7.80}$$

which shows that for small values of the collector resistor, i.e., $R_C/R_B \to 0$, the collector current is *directly proportional* to β (see Fig. 7.21). That is, if $R_C \ll R_B$ there is no feedback stabilization at all.

For $R_C \gg R_B$, we have

$$I_C = \frac{\beta(V_{CC}-V_{BE})}{R_B+(\beta+1)R_C} \approx \frac{\beta(V_{CC}-V_{BE})}{(\beta+1)R_C} \approx \frac{\beta(V_{CC}-V_{BE})}{\beta R_C} = \frac{V_{CC}-V_{BE}}{R_C}, \tag{7.81}$$

where the approximation $(\beta+1) \approx \beta$ is valid for all practical values. Result (7.81) is very important to us because we conclude that if $R_C \gg R_B$ then the feedback control is perfect and the collector current is not dependent upon the β factor any more (Fig. 7.21). Instead, only the external components are used to determine the collector current I_C value. We do realize, however, that this control is achieved by sacrificing the circuit gain, which is what we need to keep the DC voltage level.

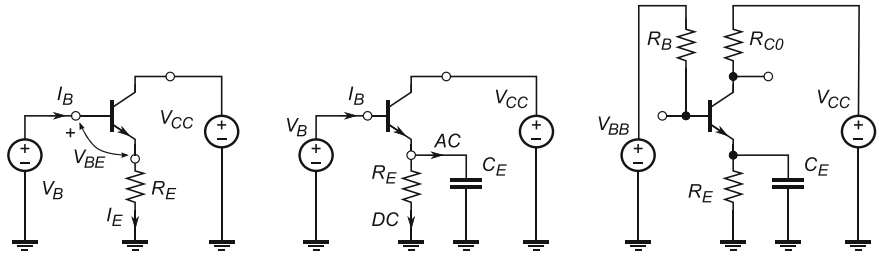


Fig. 7.22 Emitter-degenerated version of the CE amplifier (*left*); decoupling of the emitter's DC and AC currents by adding a C_E capacitor path (*centre*); and separation of R_B from R_C, which requires the use of two batteries (*right*)

As a side note, the base–emitter voltage still depends on temperature, i.e., $V_\mathrm{BE} = f(T)$. However, temperature compensation techniques are beyond the scope of this book and we assume that the environment temperature is constant, i.e., that the circuit operates at room temperature.

7.4.1 Emitter-Degenerated CE Amplifier

An alternative implementation of the negative feedback loop for the purpose of stabilizing the DC operating point is known as the "degenerated emitter" variant of the CE amplifier, Fig. 7.22 (left). We can show relatively easily how adding an emitter resistor R_E helps stabilize the collector current I_C. Starting with the base terminal, we first write (assuming $I_\mathrm{C} \approx I_\mathrm{E}$)

$$V_\mathrm{B} = I_\mathrm{B}R_\mathrm{B} + V_\mathrm{BE} + I_\mathrm{E}R_\mathrm{E} \approx I_\mathrm{B}R_\mathrm{B} + V_\mathrm{BE} + I_\mathrm{C}R_\mathrm{E} \tag{7.82}$$

and then we consider two cases. In the first case, $R_\mathrm{E} = 0$, the base voltage becomes

$$V_\mathrm{B} \approx I_\mathrm{B}R_\mathrm{B} + V_\mathrm{BE} = \frac{I_\mathrm{C}}{\beta}R_\mathrm{B} + V_\mathrm{BE}, \tag{7.83}$$

which shows that the base voltage strongly depends on the β factor. In the second case, $R_\mathrm{E} \neq 0$, the term $I_\mathrm{B}R_\mathrm{B}$ is much smaller than the other two, hence (7.82) becomes

$$V_\mathrm{B} \approx V_\mathrm{BE} + I_\mathrm{C}R_\mathrm{E}, \tag{7.84}$$

which is not dependent on the β factor and we have achieved the goal of fixing the collector current.

A silent feature of the bias stabilization mechanism is that it does not discriminate between the DC biasing signal (the one that we want to keep stable) and the AC signal (the one that we want to amplify). In other words, the gain is reduced for both DC and AC signals, which was not the intention. Therefore, we must modify the stabilization method so that only DC signals experience a low gain path and the gain of AC signals is only minimally affected. In other words, we have to "decouple" the AC signals from the DC signals, which implies that we need to use reactive components in order to provide two separate signal paths. In the case of a degenerated emitter, the emitter resistor R_E may be bypassed with a capacitor C_E, Fig. 7.22 (centre), which effectively reduces AC resistance $r_\mathrm{e} \approx 0$ in the emitter path. Hence, the AC signal is maximally amplified, (7.58). At the same time, the emitter resistor R_E provides the DC feedback control, (7.82). The emitter degeneration method is more often used than the literal implementation in Fig. 7.20.

7.4.2 Voltage Divider for Biasing Control

There are a number of different ways of implementing biasing schemes that can be reduced to the schematic Fig. 7.20 (right), regardless of whether we use the collector or emitter resistor. A resistive network associated with the base node can always be reduced to a single equivalent resistor R_B, hence (7.79) can be used in general to describe the relationship between the collector current and β.

It is very useful to develop a method of evaluating the effectiveness of the bias regulation method by inspecting component values in Fig. 7.20 (right). We can develop such a method by introducing a figure of merit called a *SF*, which enables us to compare the effectiveness of different biasing stabilization structures. One way of looking at (7.79) is to realize that its numerator represents the case of no stabilization. That is, if $R_B \to \infty$, which is equivalent to disconnecting the base from the collector, then the denominator of (7.79) becomes one. Hence, we can generalize and say that whatever multiplies the expression to achieve the no stabilization case is the multiplication factor and (7.79) may be rewritten as

$$I_C = \frac{\beta(V_{CC} - V_{BE})}{R_B} \frac{1}{1 + (\beta + 1)\dfrac{R_C}{R_B}} = \frac{\beta(V_{CC} - V_{BE})}{R_B} \times SF, \tag{7.85}$$

where the SF is $0 \geq SF \leq 1$ and, in general, is written in the following form

$$SF = \frac{1}{1 + \beta F}, \tag{7.86}$$

where F is the fraction of collector current used as the feedback. The two extremes are, of course, with no feedback current (i.e., $F = 0$) and with the whole collector current used as the feedback current (i.e., $F = 1$). By inspection of Fig. 7.20 (right) and by applying the current division rule at node ①, where the collector current I_C is split between the R_C and R_B branches, we write an expression for the base current I_B as

$$I_B = I_C \frac{R_C}{R_C + R_B}, \qquad F = \frac{R_C}{R_C + R_B}. \tag{7.87}$$

In other words, we can estimate the fraction of collector current that is redirected into the feedback path by knowing the values of the two resistors, and subsequently we can calculate the stability factor *SF* of the network. This approach is very useful and, in the following pages, we demonstrate its application.

We already concluded in (7.81) that in order to maximize the SF we need to reduce base resistance, i.e., $R_B \approx 0$. However, that choice also reduces the gain of the amplifier because the total effective collector resistance is also reduced (for small signals, R_B is in parallel with R_C), Fig. 7.20 (right). If we are to keep R_B, one possible solution is to connect the base terminal directly to the V_{BB} voltage that is provided by the V_{BB} battery, Fig. 7.22 (right).

Let us estimate the SF of this circuit by using the following numerical example.

Example 7.1. For a given transistor, in order to set the average collector current of the circuit in Fig. 7.22 to $I_C = 5\,\text{mA}$, the base voltage has to be set at $V_{BB} = 3\,\text{V}$. That collector current forces the base–emitter voltage to $V_{BE} = 0.7\,\text{V}$ and the emitter voltage to $V_E = 2.3\,\text{V}$. The base resistance is $R_{s2} = 100\,\Omega$, and the transistor current gain is $\beta = 100$. Estimate the value of the emitter resistor R_E and the SF of this circuit.

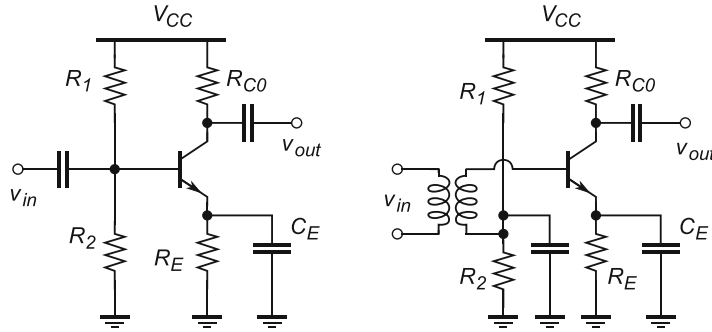


Fig. 7.23 Voltage divider biasing network for a CE amplifier: capacitive input signal coupling (*left*) and inductive input signal coupling (*right*). The large coupling capacitors are not labelled

Solution 7.1. As β is large, we have $I_E \approx I_C$ and it is straightforward to calculate the emitter resistor as $R_E = V_E/I_E \approx V_E/I_C = 2.3\,V/5\,mA = 460\,\Omega$. Keep in mind that in this configuration R_E is equivalent to R_C in Fig. 7.20 (right), while R_{C0} in this schematic only separates the output terminal and the V_{CC} battery and provides the voltage gain. notation in the general formula and escape this confusion.

Therefore, from (7.86) and (7.87), we write

$$SF = \frac{1}{1+\beta\,\frac{R_E}{R_{s2}+R_E}} = \frac{1}{1+100\,\frac{460\,\Omega}{100\,\Omega+460\,\Omega}} = 0.012, \tag{7.88}$$

which is a very good result not too far from the maximum theoretical value of 0.0099 (i.e., for $\beta = 100$ and $R_B = 0$). However, using that low value for R_B reduces the amplifier input impedance. For example, in a more realistic case of $R_B = 3.3\,k\Omega$, the stability factor becomes $SF = 0.076$, which is not a bad result.

The second pressing issue, the need for two batteries, may be solved by implementing a voltage divider instead of a single resistor R_B, Fig. 7.23. There are wo possible ways to inject an AC signal into an amplifier without disturbing its DC operating point: capacitive and inductive coupling.

By inspection of the circuits in Fig. 7.23, we recognize that the voltage divider R_1, R_2 effectively presents resistive load of $R_{1,2} = R_1||R_2$ at the base node.[2] That is, in order to keep the input resistance R_{in} of the amplifier high, these two resistances need to be rather large. Thus, by comparison with Fig. 7.22 (right), we conclude that in order to estimate SF we need to substitute the $R_{1,2}$ resistance in place of R_B in (7.87).

To illustrate the voltage divider biasing scheme, let us consider the following numerical example.

Example 7.2. For a CE amplifier (Fig. 7.23) intended to amplify a 50 Hz signal with $R_E = 1\,k\Omega$, $I_C = 1\,mA$, $\beta = 50$, and $V_{CC} = 6\,V$, design a voltage divider circuit.

Solution 7.2. Voltage at the emitter node is $V_E \approx I_C R_E = 1\,V$. Base current is $I_B = I_C/\beta = 1\,mA/50 = 20\,\mu A$. This current flows through R_1 in addition to R_1, R_2 current because of the V_{CC}. For steady base voltage, it is, therefore, required that the base current is much smaller than the DC current due to V_{CC}, which in engineering vocabulary means at least ten times smaller. Hence, the current caused by the battery is set to $I = 200\,\mu A$, which makes the total current through R_1, R_2 equal to $220\,\mu A$.

[2] Keep in mind that, because the internal resistance of the voltage source is zero, the two resistors are implicitly connected in parallel through the voltage source.

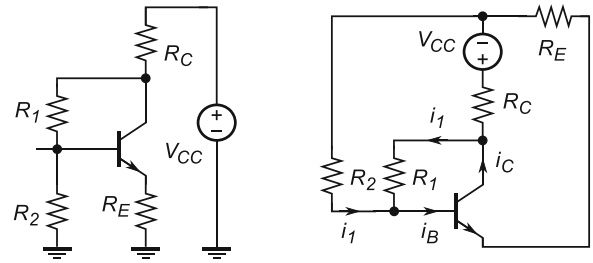


Fig. 7.24 Voltage divider biasing network version for CE amplifier (*left*); and its equivalent schematic (*right*)

Bypass capacitance must have a small reactance at the signal frequency relative to $r_e = 25\,\Omega$. So, at 50 Hz we can choose $C_1 = 500\,\mu\text{F}$, which translates to $Z_{C_1} = 6.5\,\Omega$. Voltage $V_{BE} = 0.7\,\text{V}$ hence $V_{R_1} = 5\,\text{V} - 0.7\,\text{V} = 4.3\,\text{V}$, therefore $R_1 = 4.3\,\text{V}/220\,\mu\text{A} \approx 20\,\text{k}\Omega$. Voltage $V_{R_2} = 1.7\,\text{V}$ hence $R_2 = 1.7\,\text{V}/200\,\mu\text{A} = 8.5\,\text{k}\Omega$ (the base current goes into the base).

Therefore, $R_{1,2} = R_1||R_2 \approx 6\,\text{k}\Omega$, which leads to

$$SF = \frac{1}{1+\beta\,\frac{R_\text{E}}{R_{1,2}+R_\text{E}}} = \frac{1}{1+50\,\frac{1\,\text{k}\Omega}{6\,\text{k}\Omega+1\,\text{k}\Omega}} = 0.12, \tag{7.89}$$

which means that the variations of the mean emitter current are about $1/8$ relative to the non-stabilized case. After carefully considering the influence of all parameters in our calculations, we can conclude that it is possible to further improve the SF at the expense of higher emitter current, i.e., higher power consumption, and it always helps to use better transistors with higher β. The example also illustrates that a good engineering solution results from carefully balanced compromises. We keep in mind that the voltage divider biasing scheme is one of the most commonly used methods of biasing LF amplifiers.

7.4.3 Two-Stage Biasing Control

For even better values of SF, we need to review the fundamental limitations of the voltage divider scheme. First, we need to recognize that reconnection of resistor R_1 (Fig. 7.23) back to the collector node enables R_C to again contribute to the feedback mechanism, Fig. 7.24 (left). Its equivalent schematic diagram is shown in Fig. 7.24 (right). In order to see how the improvement is achieved, let us consider the following example.

Example 7.3. The circuit in Fig. 7.24 has the following typical component values: $R_\text{E} = 1\,\text{k}\Omega$, $R_\text{C} = 5\,\text{k}\Omega$, $R_1 = 10\,\text{k}\Omega$, $R_2 = 5\,\text{k}\Omega$, and $\beta = 100$. Estimate the SF.

Solution 7.3. The collector current splits into two branches with R_C and R_1, where for the current through R_1, after applying the current divider rule, we write

$$I_1 \approx I_\text{C}\,\frac{R_\text{E}+R_\text{C}}{R_\text{E}+R_\text{C}+R_1} \tag{7.90}$$

(neglecting the parallel $R2||R_\text{E}$). Therefore, current through the resistor R_C is

$$I_\text{C}\,\frac{R_1}{R_\text{E}+R_\text{C}+R_1}, \tag{7.91}$$

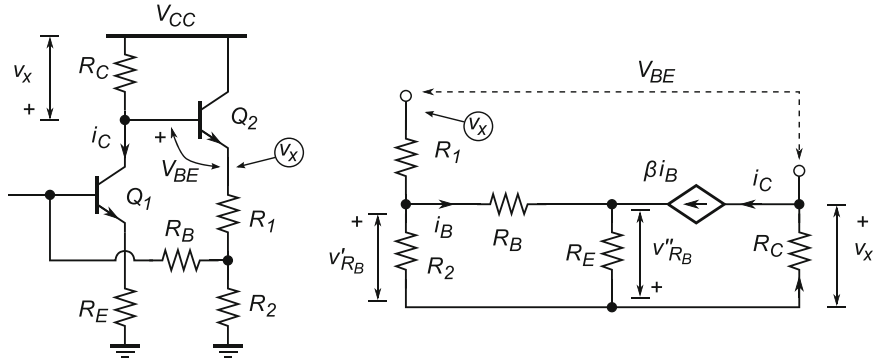


Fig. 7.25 A two-stage biasing network version for a CE amplifier (*left*) and its equivalent small circuits schematic diagram (*right*)

which splits at the junction of R_E and R_2. It follows that the current through R_2 is

$$I_2 = I_C \frac{R_1}{R_E + R_C + R_1} \frac{R_E}{R_E + R_2} \tag{7.92}$$

and the base current is then $I_B = I_1 + I_2$. Hence, we calculate the feedback factor as

$$F = \frac{I_B}{I_C} = \frac{R_E + R_C}{R_E + R_C + R_1} + \frac{R_1}{R_E + R_C + R_1} \frac{R_E}{R_E + R_2}. \tag{7.93}$$

Substituting (7.93) into (7.86) and using the numerical values of this example, we first find $F = 0.48$, which is to say that $SF = 0.02$. This variant of the feedback control represents a significant improvement relative to the version in Example 7.2.

In order to better evaluate the real gain achieved by the above idea, let us find out what price was paid to achieve the last result. The first consequence of the approach is that the gain in stability factor was achieved by placing $R_1||R_2$ across the collector resistance, effectively reducing the load resistance. The second consequence is that the AC feedback path of the signal is again closed, which, as we already discussed, reduces the overall signal AC gain. Neither of these consequences is a desirable property of a good voltage amplifier.

Therefore, we realize that we need to somehow increase resistance that is added in parallel to the collector resistor and we need to break the AC feedback path while keeping the feedback loop working. This sounds like a classical example of an interface that needs the addition of some sort of buffering stage. Hence, one possible way to resolve these two issues is to employ an emitter-follower stage in the feedback path, Fig. 7.25 (left) and its equivalent schematic diagram Fig. 7.25 (right). The large input resistance of the added CC stage does not affect the collector resistance of the first stage too much, while the low-resistance output of the CC stage, which looks more like an ideal voltage source, provides good voltage drive to the passive part of the feedback network R_1, R_2, and R_B.

In order to demonstrate this idea, let us evaluate the SF using the following numerical example.

Example 7.4. The circuit diagram in Fig. 7.25 (left) assumes typical component values of $R_C = 5\,\text{k}\Omega$, $R_E = 5\,\text{k}\Omega$, $R_B = 3.3\,\text{k}\Omega$, a half-voltage divider (i.e., $R_1 = R_2$), and $\beta = 100$. Estimate its SF.

Solution 7.4. Assuming high input resistance of the emitter follower at the base of Q_2, Fig. 7.25 (left), there is no current splitting at the Q_1 collector node. In addition, we observe that the small signal voltage across the Q_1 collector resistor R_C is $v_x = i_C R_C$, which is the same as the small signal voltage at the output of the emitter follower Q_2 that appears at the top of resistive divider R_1, R_2, (of course, with its common mode voltage shifted by v_{BE}).

After the circuit is simplified by using a BJT small signal T-model, Fig. 7.25 (right), we observe that the voltage across resistor R_B is generated by two currents, one through resistive voltage divider R_1, R_2 and one through the parallel connection of $R_E||R_B$ (after approximation $R_B + R_2 \approx R_B$). The former current generates voltage V'_{R_B} across resistor R_2 while the latter generates voltage V''_{R_B} across resistor R_E (in parallel with R_B). Hence, we write $V_{R_B} = V'_{R_B} + V''_{R_B}$. In this estimate, we also assume large β, i.e., $\beta \approx (\beta + 1)$.

With the above assumptions and approximations, current through R_B is approximately,

$$i_B = \frac{v_B}{R_B} = \frac{V'_{R_B} + V''_{R_B}}{R_B}, \tag{7.94}$$

where,

$$V'_{R_B} = R_2\, i_{R_2} \approx R_2\, \frac{v_x}{R_1 + R_2} = \frac{R_2}{R_1 + R_2}\, R_C\, i_C, \tag{7.95}$$

$$V''_{R_B} = i_C\, [R_E||(R_B + R_2)] \approx i_C\, [R_E||R_B] = \frac{R_E R_B}{R_E + R_B}\, i_C. \tag{7.96}$$

After substituting the above expressions into (7.94), we calculate the base current and the feedback factor as

$$i_B = \frac{V_{R_B}}{R_B} = \frac{V'_{R_B} + V''_{R_B}}{R_B} = \frac{\frac{R_2}{R_1+R_2}\, R_C\, i_C + \frac{R_E R_B}{R_E+R_B}\, i_C}{R_B},$$

$$\therefore$$

$$F = \frac{i_B}{i_C} = \frac{R_2}{R_1 + R_2}\frac{R_C}{R_B} + \frac{R_E}{R_E + R_B}. \tag{7.97}$$

Substituting (7.97) into (7.86), and after substituting the numerical values from this example, we find the SF as

$$SF = \frac{1}{1 + \beta\left(\frac{R_2}{R_1+R_2}\frac{R_C}{R_B} + \frac{R_E}{R_E+R_B}\right)} \tag{7.98}$$

$$= \frac{1}{1 + 100\left(0.5 \times \frac{5\,\text{k}\Omega}{3.3\,\text{k}\Omega} + \frac{5\,\text{k}\Omega}{8.3\,\text{k}\Omega}\right)} = 0.0073, \tag{7.99}$$

which indicates an order of magnitude improvement over the simple bias control schemes. Of course, the complexity of the circuit has increased, which is the price paid for the high DC biasing performance. A large number of variants of this two-stage DC control approach are commonly used in commercial amplifier designs.

Understanding the biasing principles presented in this section, from now on we accept that the biasing point is somehow established and we can focus on amplifier analysis while ignoring details of the DC bias as long as the g_m values of the active devices are set.

7.5 AC Analysis of Voltage Amplifiers

In our analysis, so far, we have made an important assumption that the functionality of circuits is completely independent of the signal frequency. In other words, we made a low-frequency approximation and simply assumed that the circuit behaves the same regardless of whether the frequency of the input signal is DC or infinite or anywhere in between for that matter, i.e., we assumed a frequency-independent amplifier gain. Of course, by now we know that, as tiny as they are, even electrons have both inertia and finite velocity associated with their movement. Consequently, even purely resistive networks do have measurable propagation delays, which is another way of saying that a signal applied at the input terminals of a network does not show up instantaneously at the output terminals. Indeed, common engineering practice is to calculate the propagation delay, for the given material properties and physical size of the network components, and confirm it by measurement. The addition of components capable of storing energy, i.e., capacitors and inductors, further undermines the low-frequency approximation.

A number of, often surprising, phenomena exist in our physical world because of the finite time that is required to, for example, deliver a certain number of electrons onto one plate of a capacitor and then to take them off the plate. To illustrate the point, let us do the following mental experiment.

Let us imagine a black box with only two wires coming out of it that are available to us. Let us further assume that we have an instrument that measures the amount of electrical charge passing through the wires. Initial measurement shows that there is no measurable potential difference between the two terminals. Then, let us connect an ideal $V = 1\,\mathrm{V}$ voltage source to the two terminals, Fig. 7.26 (left). Eventually, the voltage at the terminals becomes steady and we see that, relative to its initial value, the input terminal voltage changed by $\Delta V = 1\,\mathrm{V}$ (i.e., close enough). At the same time, the flow of electrons practically stopped and we measure (again, close enough, see Sect. 4.1.5.3), the amount of charge that moved through the instrument as $\Delta q = 1\,\mathrm{pC}$. The total amount of charge and the associated voltage are related as described by (4.13), hence through calculation

$$C = \frac{\Delta q}{\Delta V} = \frac{1\,\mathrm{pC}}{1\,\mathrm{V}} = 1\,\mathrm{pF}, \tag{7.100}$$

we rightfully conclude that the black box contains (or at least, it behaves as) a 1 pF capacitor.

Now, let us imagine that some aliens with a good sense of humour add, in series with the capacitor, a 99 V battery whose polarity is aligned with the 1 V external source, Fig. 7.26 (right), at the exact moment when the external 1 V voltage source is connected. Without knowing anything about the prank, we now measure a flow of 100 pC and, being aware of only the external 1 V source, we conclude that the black box behaves as a 100 pF capacitor

$$C = \frac{\Delta q}{\Delta V} = \frac{100\,\mathrm{pC}}{1\,\mathrm{V}} = 100\,\mathrm{pF}, \tag{7.101}$$

which, of course is not what the real network in the box is. This apparent *capacitive magnification* is known as the Miller effect, and it is a consequence of the internal voltages acting upon the real capacitor.

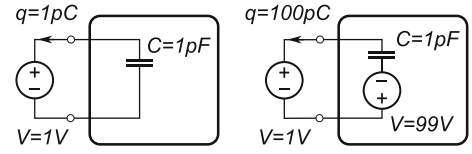


Fig. 7.26 A black box that contains: only a capacitor (*left*); a capacitor with an internal voltage source (*right*)

7.6 Miller Capacitance

An amplifying network configuration that satisfies the following conditions is quite common in electronics and in nature:

- The amplifier is an inverting voltage amplifier.
- The amplifier's voltage gain is larger than one, i.e., $|A_v| \gg 1$.
- A capacitor C is connected between its input and output terminals.

The general network that illustrates the three conditions listed above (Fig. 7.27) is analyzed as follows. Assuming an inverting voltage amplifier with infinite input impedance, the output and input voltages are related as $v_{out} = -A_v v_{in}$. Under the condition of no current flowing into the amplifier's input terminal, the input impedance Z_{in} of the network is calculated as

$$i_{in} = \frac{v_{in} - v_{out}}{Z_C} = \frac{v_{in} + A_v v_{in}}{Z_C} = \frac{v_{in}(1 + A_v)}{Z_C}, \tag{7.102}$$

$$\therefore$$

$$Z_{in} = \frac{v_{in}}{i_{in}} = \frac{Z_C}{1 + A_v}, \tag{7.103}$$

which, in the case of capacitive bridging impedance $Z_C = 1/sC$, takes the form of

$$Z_{in} = \frac{1}{j\omega C\,(A_v + 1)} = \frac{1}{j\omega C_M}, \tag{7.104}$$

where, the effective Miller capacitance is defined as

$$C_M = C\,(A_v + 1). \tag{7.105}$$

Result (7.105) is very important for high-frequency circuit design. Effectively, in combination with the source resistance, Miller capacitance creates a frequency-dependent voltage divider, which, as we already know, behaves as a LP filter.

Example 7.5. Assume an ideal, single-stage, common-emitter amplifier (i.e., the input resistance $R_{in} = \infty$) with voltage gain of $A_v = -99$, as shown in Fig. 7.28 (left). The amplifier is driven by a voltage source whose output resistance is $R_S = 50\,\Omega$. In addition, there is a capacitor connected between the transistor's collector and base $C_{CB} = 1$ pF. Estimate the useful range of input signal frequencies.

Solution 7.5. The single-stage, common-emitter amplifier satisfies all three conditions required for the Miller effect. It is an inverting amplifier, it has gain larger than one, and it has a capacitive

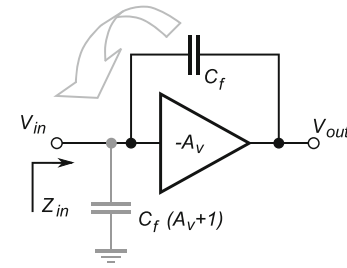


Fig. 7.27 A general inverting voltage amplifier with Miller capacitance

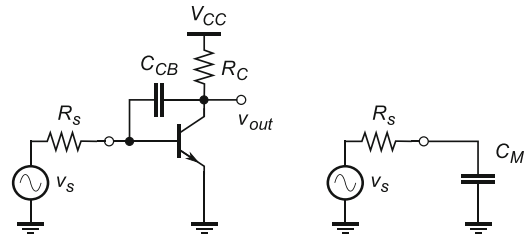


Fig. 7.28 A CE voltage amplifier with Miller capacitance caused by C_{CB} (*left*) and its equivalent LP filter network (*right*)

component that creates an AC path between the input and output terminals. Therefore, the equivalent schematic diagram of the signal source–amplifier network, Fig. 7.28 (right), is analyzed as a voltage divider that consists of the source resistance $R_S = 50\,\Omega$ and the Miller capacitance $C_M = C_{CB}(|A_V| + 1) = 100\,\text{pF}$.

The useful range of input signal frequencies, in this case, means simply the bandwidth of the (R_S,C_M) LP filter, which is defined at the $-6\,\text{dB}$ voltage point. We already know that 3 dB power point or, equivalently, the 6 dB voltage point refers to the frequency f_3 dB where the output voltage is exactly half the input voltage, i.e., the two resistances of the voltage divider are equal

$$R_S = Z_{C_M} \quad \therefore \quad R_S = \left|\frac{1}{j\omega C_M}\right| = \left|\frac{1}{j\,2\pi\, f_{3\,\text{dB}}} C_M\right|,$$

$$\therefore$$

$$f_{3\,\text{dB}} = \left|\frac{1}{j\,2\pi R_S C_M}\right| = \frac{1}{2\pi\, 50\,\Omega\, 100\,\text{pF}} = 31.831\,\text{MHz}, \tag{7.106}$$

which is a significant reduction in signal bandwidth, considering that we started from relatively small component values of $50\,\Omega$ and 1 pF, which by themselves would allow a bandwidth of 3.183 GHz.

Additional assumptions usually made are that the voltage gain A_v is not a function of the frequency, $A_v \neq f(\omega)$, and that collector–base capacitance C_{CB} is independent of collector–base voltage, i.e., $C_{CB} \neq f(V_{CB})$. Both assumptions simplify the analysis, otherwise we would have to use numerical solvers to reach any conclusion.

The fundamental reason for a CE amplifier's sensitivity to the Miller effect is that base–collector capacitance C_{CB} is real and unavoidable. The parasitic capacitance exists because of the reverse-biased, base–collector pn junction that behaves as a voltage-controlled capacitor. There are three main parasitic capacitances inside a BJT, Fig. 7.29 (left): collector–base C_{CB}, base–emitter C_{BE}, and collector–emitter C_{CE}. By inspection of the equivalent BJT model, Fig. 7.29 (right), for the case of a CE amplifier, we find that capacitance C_{BE} is connected across the input terminals while capacitance C_{CE} is across the output terminals. Both capacitances are safely grounded on one side, and therefore introduce only minor frequency limitation to the overall amplifier behaviour. However, the collector–base C_{CB} capacitance is floating and provides a feedback path to the signal, which gives rise to the Miller effect.

To complete the set of Miller effect requirements, the CE stage inherently inverts the signal and has large voltage gain. Because of that weakness of CE amplifiers, a CB amplifier is often used in RF designs in cases where its low input resistance is compatible with the previous stage. In the case of a CB amplifier, everything else being equal, there is no significant capacitance that connects the CB stage I/O terminals. In most textbooks, the collector–emitter capacitance C_{CE} is labelled as C_μ; it is made of two pn junction capacitances C_{CB} and C_{BE} in series and, hence, it is *very* small. At the same time, the C_{CB} capacitance is safely connected between the output node and the small signal ground.

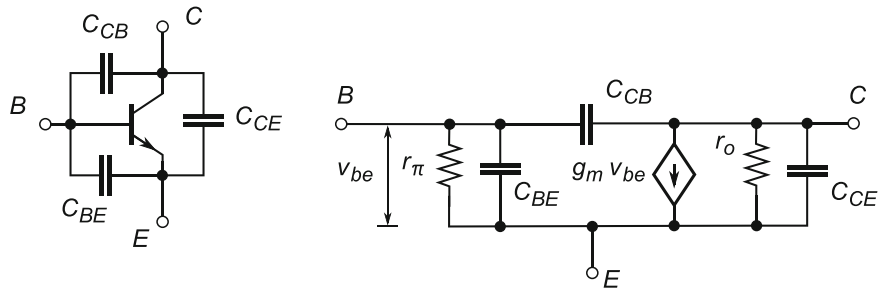


Fig. 7.29 BJT parasitic capacitances (*left*) and the equivalent small signal model (*right*)

In general, the Miller effect is not desirable in RF circuits. However, in Sect. 11.3.3, we find that even this apparent weakness of CE amplifiers has been exploited in a very interesting and important RF application. Although we focused on a capacitive feedback path, we note that any general impedance in the feedback path (in combination with the other two conditions) creates the Miller effect.

7.7 Tuned Amplifiers

In our treatment of "baseband amplifiers" we assumed low-frequency operation, which is to say that all single-tone signals from DC to "not so high" frequencies were treated as equally interesting, desirable, and possible to amplify. As a result, aside from frequency limitations caused by LP filtering effects at the input stage of CE amplifiers due to the Miller effect or by high-pass filtering effects on the AC coupling between the driver stage and the load, we merrily kept spending energy to amplify all possible tones that showed up at the amplifier's input terminal nodes.

As the reader is already guessing, this generous approach to signal amplification, aside from strictly technical difficulties, has at least two serious drawbacks if it is to be used for amplification of RF signals:

- The frequency of the content of the message, for example the human voice, is limited to the range 20 Hz to 20 kHz. That is, high-fidelity (HiFi) sound reproduction does not require any of the single tones outside of this frequency range and, in other words, it is a waste of energy to amplify them. The amplification energy must come from somewhere and, by doing so, we unnecessarily drain the amplifier's batteries. To make the things worse, we may need to provide an additional cooling mechanism to dissipate the excess heat generated by the components, not to mention the impact on the environment of disposing of the drained batteries.
- A less obvious, but equally important, drawback of wideband amplification is that all unwanted tones accepted into the amplifier contribute only to the increased noise level. After all, these tones are not needed and not desirable, hence they represent noise. The amplifier cannot possibly know what tones the user wants to hear, hence all tones are equally amplified. Because the overall noise energy is collected over a wider band than is necessary, it means the overall SNR is lowered.

As we will find out, there are other reasons why baseband amplifiers are not used in RF sections of radios. For the time being, the above two arguments should be enough to convince the reader that the overall result of using a baseband amplifier for radio signal amplification results in an expensive,

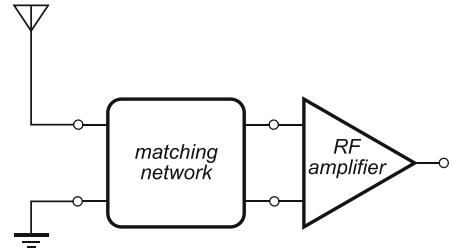


Fig. 7.30 The first three main stages of RF radio receiver

bulky, power-hungry, and less good quality RF amplifier. All that assumes we somehow managed to make the amplifier's bandwidth wide enough to start with.[3]

In Sect. 6.4, we learned that in order to efficiently transport EM energy collected by an antenna to the input terminals of an amplifier, we need to design a matching network. At that time, we quietly accepted that the maximum power transfer was possible, in theory, at only one frequency, and in practice over a very narrow range of frequencies determined by the Q factor of the matching network. In Sect. 5.9 we learned that any realistic RLC network behaves as a "bandpass filter".

The first three stages of a radio receiver, the antenna, the matching network, and the RF amplifier (see Fig. 7.30), are often referred to as the front-end of the RF radio receiver. It is now time to ask how exactly an RF amplifier is different from a baseband amplifier. Before answering this question, let us first state that:

- The frequency range of operation of an RF amplifier must be "aligned" (i.e., tuned) with the centre frequency of the matching network, which, in turn, is tuned with the antenna.
- The bandwidth of the RF amplifier should be similar to the bandwidth of the incoming message, approximately 20 kHz in the case of music. That is, the RF amplifier bandwidth should be not too wide, causing a decrease of SNR, or too narrow, introducing signal distortions (keep in mind Fourier).

Of the three single transistor amplifier types (i.e., CE, CB, and CC), the emitter follower is the only one that has voltage gain slightly less than unity, therefore we focus on the other two variants.

7.7.1 Single-Stage CE RF Amplifier

In principle, a single-stage CE amplifier is easily turned into a CE RF amplifier with two modifications: the collector's resistive load R_C is replaced by an $L_C C_C$ resonator; and an $L_B C_B$ resonator is connected between the input node and the ground, Fig. 7.31. Both resonators are tuned to the same resonant frequency ω_0. The input AC signal is then injected into the base through decoupling capacitor C_0, whose impedance is negligible at the resonator frequency ω_0, Fig. 7.31.

7.7.1.1 Intuitive View of CE RF Amplifier Operation

At the resonant frequency, $\omega_0 = 1/\sqrt{LC}$, both LC resonators are effectively equivalent to their respective dynamic resistances R_D. We already know that in the case of theoretically ideal LC

[3] Keep in mind that modern RF carrier frequencies are in the order of MHz or GHz. A wideband amplifier would need to be able to work from DC to the RF carrier frequency, which is not always possible to do with the current technology.

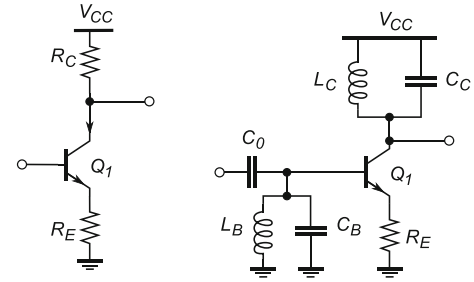


Fig. 7.31 A CE amplifier (*left*) and its equivalent RF CE amplifier (*right*)

components, a resonator's dynamic resistance R_D is infinite, while in the case of real LC components, the dynamic resistance R_D is calculated as $R_D = QZ_L(\omega_0)$.[4] Consequently, in the ideal case, the input side of the amplifier does not "feel" any additional resistive load, i.e., there is no current splitting at the base node and 100% of the AC signal current is injected into the base. The only ramification of adding an ideal L_BC_B resonator is that out of all possible frequencies only a single tone at ω_0 is able to pass through that L_BC_B "entrance door" and enter the transistor gate, all the other tones with frequencies $\omega \neq \omega_0$ are simply not aligned with the door and they "hit the wall", i.e., they are attenuated down to zero amplitude. In the case of real L_BC_B components, the entrance door is wider than a single frequency; hence not only ω_0 passes through but also the adjacent frequencies that are within the "width of the door". In technical terms, the input L_BC_B resonator works as a narrowband bandpass filter, whose centre frequency is ω_0 and bandwidth $BW = \Delta\omega$, where $\Delta\omega = \omega_0/Q$ (note that if $Q = \infty \Rightarrow BW = 0$). This is how the amount of noise entering the amplifier through the input terminal is controlled.

At the same time, at the output side of the CE RF amplifier, the collector is experiencing very large resistive load R_D, which translates into large voltage gain $A_v = g_m R_D$. The transistor's transconductance g_m is set by the biasing network (not shown). Therefore, in the ideal case ($R_D = \infty$), the amplifier would achieve an infinite voltage gain, i.e., it would be able to amplify even an infinitely small single-tone signal at exactly ω_0 frequency and would "ignore" all other tones. In the real case, the gain is limited by finite R_D within the finite bandwidth BW but it is, nevertheless, still very high.[5] Although the input side L_BC_B resonator blocks all unwanted frequencies from entering the amplifier, we already know that there is internally generated noise that also needs to be filtered out by the L_CC_C resonator. By means of these two LC resonators (effectively a double bandpass filter) along the signal path, the gain of the CE RF amplifier is optimized so that only the frequencies of interest within the bandwidth are amplified.

Although the intuitive picture of CE amplifier operation painted so far ignores quite a few fine details, it is definitely useful in terms of understanding the overall functionality of a theoretical CE RF amplifier. With its high input resistance and high output current gain, the CE amplifier is considered one of the most important structures for voltage signal amplification. Let us now find out about the limitations of a simple CE RF amplifier and what needs to be done in order to make it truly practical at frequencies of interest.

[4] We could have used Z_C instead. Remember that, at the resonance, $Z_L = Z_C$; hence, the dynamic resistance R_D is calculated as a product of the Q factor and either of the two impedances.

[5] That is why it is possible to see voltages across the LC resonator that are higher than the amplifier's power supply voltage level.

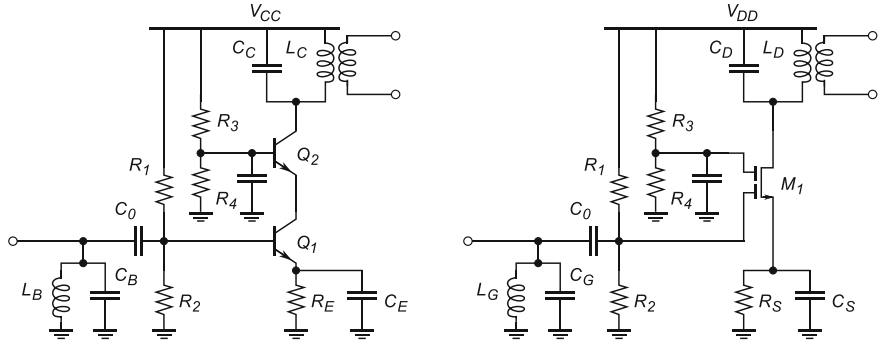


Fig. 7.32 A cascode BJT RF amplifier (*left*) and its equivalent dual-gate FET RF version (*right*)

7.7.1.2 Miller Effect

We already found out that the small bridging capacitance C_{CB} creates a feedback loop from the output terminals back to the input terminals. As it turned out, the Miller capacitance is perceived by the input terminals as approximately A_V times greater than the real collector–base C_{CB} capacitance, with the consequence of a LP filter effect on the input side and drastic reduction of the signal bandwidth.

It would appear from the analyses in Sect. 7.6 that a CE amplifier is hopelessly lost for all but low- to mid-frequency range RF applications. However, we can improve the frequency-dependent behaviour of a CE amplifier in RF applications by looking at the three conditions for the Miller effect one by one. Obviously, we cannot do anything about its inherent inverting signal nature and we do need to keep the high voltage gain. For all practical purposes, we cannot remove the Miller effect by modifying these two conditions. The only option left is to find out if we can do anything about the bridging I/O capacitance.

As a matter of fact, we learned at the end of Sect. 7.6 that not having the I/O bridging capacitance protects a CB amplifier configuration from the Miller effect. That gives us an idea of how to modify a simple CE stage and improve its bandwidth by turning it into a cascode amplifier, Sect. 7.3. An additional, and not so obvious, feature of a cascode amplifier architecture is that the insertion of a CB stage between the CE output node (the collector of Q_1) and the load resistor R_C effectively removes the capacitive connection between the input and the output nodes of the cascode amplifier, Fig. 7.32.

The collector–base capacitance C_{CB} of Q_1 connects the input terminal of the cascode amplifier with one of its internal nodes, while at the same time the cascode amplifier output terminal is taken from the CB amplifier output node (the collector of Q_2), which is safely disconnected (i.e., "buffered") from the input terminal. This property makes the cascode amplifier immune to the Miller effect and further increases the importance of cascode amplifier architecture.

In practice, a very common way to implement a cascode RF amplifier is by using a dual-gate MOSFET (often JFET) device, Fig. 7.32 (right). The two FET devices are manufactured on the same silicon substrate and packaged in the same package. This means that the manufactured dual device has exactly the same functionality as two cascode devices, with the advantage of reduced parasitic capacitances and greatly improved high-frequency (HF)performance compared to a configuration with two discrete devices.

7.7.1.3 CE RF Amplifier Stability

Careful observation of an RF CE amplifier's operation reveals a second problem. As we already know, with a resistive load, a CE is an inverting amplifier, i.e., the input and output signals are in "antiphase".

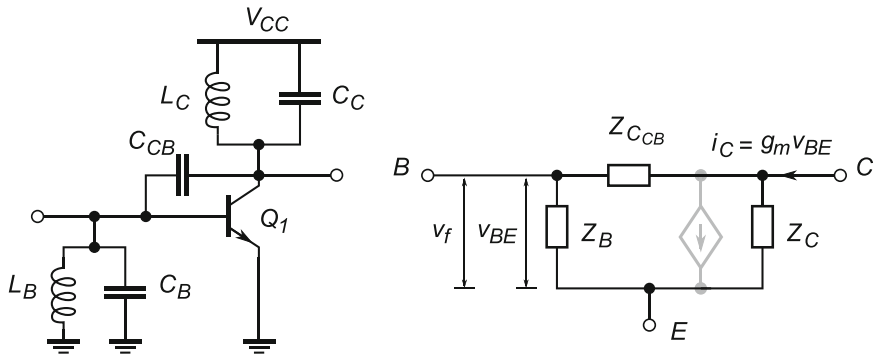


Fig. 7.33 A cascode CE RF amplifier with Miller impedance (*left*) and its equivalent model (*right*)

Equivalently, the input and output current are in phase, that is, the output current and the output voltage are in antiphase. At resonance, an LC resonator is the dynamic resistance R_D (Z_L and Z_C are same and have the opposite signs), hence from the amplifier's perspective it is the same as any other resistive load. However, the statement is valid only at one frequency, ω_0. At any other frequency than ω_0, the load becomes either capacitive or inductive. And that is where the trouble with using a CE amplifier for amplifying HF signals starts.

At the resonant frequency ω_0, the collector–base capacitor C_{CB} provides a feedback path for part of the collector current and adds $+90^\circ$ phase shift and, therefore, the base current and the fed-back current are in "quadrature".[6]

Below its resonant frequency ω_0, the LC tank behaves as an inductor (e.g. take a look at Fig. 5.14) and the collector voltage leads the collector current by $+90^\circ$. At the same time, the feedback capacitor C_{CB} provides additional 90° phase shifts, which makes the fed-back current in phase with the base current, which is effectively positive feedback that further increases the output current. Depending upon the transistor gain, within a few signal cycles the fed-back current becomes greater than the input current and the amplifier becomes unstable.

Above its resonant frequency ω_0, the LC tank behaves as a capacitor (again, take a look at Fig. 5.14) and the collector voltage lags the collector current by -90°. At the same time, the feedback capacitor C_{CB} provides additional 90° phase shifts, which makes the fed-back current in antiphase with the base current, which is effectively negative feedback that reduces the gain, however, the amplifier is stable.

In conclusion, if the resonator network introduces a 90° phase shift (in addition to the 90° phase shift created by the feedback capacitor) then the instability condition is created. Even the boundary condition, where the amplifier constantly switches between the stable and the unstable states, is itself an unstable condition. Keep in mind that, although we have used a capacitor as the feedback element, in general a feedback inductor also introduces a 90° phase shift and it may create instability conditions. Assuming that both the input and output LC resonators are centred at the same resonant frequency, we can conclude that the instability point for the amplifier is when each LC resonator by itself introduces a 45° phase shift, which leads to the negative feedback condition.

From the circuit designer's perspective, it is important to quantify under what conditions the instability of a CE tuned amplifier occurs. One way of looking at the CE RF amplifier circuit in Fig. 7.33 (left) is to draw its equivalent impedance network, Fig. 7.33 (right). Note that the impedances include all internal and external impedances associated with the BJT nodes and that the total collector current i_C is split between Z_C and $Z_{C_{CB}}$.

[6] "In quadrature" is a fancy way of saying that two variables are 90° apart in phase, i.e., orthogonal to each other.

The collector current i_C is set by v_{BE} voltage through the $g_m v_{BE}$ relation, and it flows through the parallel combination of impedances $Z_C||(Z_B + Z_{C_{CB}})$. Hence, the voltage v_{Z_C} across the collector resistance is

$$v_{Z_C} = i_C \ \left[Z_C||(Z_B + Z_{C_{CB}})\right] = g_m v_{BE} \ \frac{Z_C(Z_B + Z_{C_{CB}})}{Z_C + Z_B + Z_{C_{CB}}}, \tag{7.107}$$

We note that this is the same voltage that also appears across the input voltage divider $(Z_B + Z_{C_{CB}})$, hence the feedback voltage v_f generated across the input impedance Z_B (i.e., at the input node) due to the feedback current through $Z_{C_{CB}}$ path is calculated as

$$\begin{aligned} v_f &= \frac{v_{Z_C}}{Z_B + Z_{C_{CB}}} Z_B = g_m v_{BE} \ \frac{Z_C(Z_B + Z_{C_{CB}})}{Z_C + Z_B + Z_{C_{CB}}} \frac{Z_B}{Z_B + Z_{C_{CB}}} \\ &= g_m v_{BE} \ \frac{Z_C Z_B}{Z_C + Z_B + Z_{C_{CB}}} \approx g_m v_{BE} \ \frac{Z_C Z_B}{Z_{C_{CB}}}, \end{aligned} \tag{7.108}$$

where, the approximation $(Z_C + Z_B + Z_{C_{CB}}) \approx Z_{C_{CB}}$ is valid because capacitance C_{CB} is usually very small, which means that its impedance $Z_{C_{CB}} = 1/j\omega C_{CB}$ is very large relative to Z_C and Z_B. The amplifier is stable as long as the feedback voltage v_f is less than the base voltage v_{BE}, i.e.,

$$\begin{aligned} & v_f < v_{BE}\,; \qquad g_m v_{BE} \ \frac{Z_C Z_B}{Z_{C_{CB}}} < v_{BE}\,, \\ & \therefore \\ & g_m < \frac{Z_{C_{CB}}}{Z_C Z_B}. \end{aligned} \tag{7.109}$$

The specific impedances in (7.109) are determined as follows. Collector impedance Z_C is the collector resistance R_C in parallel with the difference between the inductor and capacitor reactance $X_C = |Z_{C_C} - Z_{L_C}|$. That is, the condition for 45° phase, dictates that $R_C = X_C$,[7] i.e.

$$Z_C = R_C||X_C = \frac{jX_C R_C}{R_C + jX_C} = \frac{jR_C R_C}{R_C + jR_C} = \frac{jR_C}{1 + j} \tag{7.110}$$

and similarly, for Z_B we write

$$Z_B = \frac{jR_B}{1 + j} \tag{7.111}$$

and for the collector–base capacitance impedance we have

$$Z_{C_{CB}} = \frac{1}{\omega_0 C_{CB}}. \tag{7.112}$$

[7]Imagine two vectors of equal length, R_C and X_C, that are 90° relative to each other. From the right-angle triangle rule, the total phase must be 45°.

After substituting these three impedances back into (7.109), we have

$$g_{\mathrm{m}} < \frac{(1+\mathrm{j})^2}{\omega_0 C_{\mathrm{CB}}\,\mathrm{j}R_{\mathrm{C}}\,\mathrm{j}R_B} = \left(\frac{1+\mathrm{j}}{\mathrm{j}}\right)^2 \frac{1}{\omega_0 C_{\mathrm{CB}} R_{\mathrm{C}} R_{\mathrm{B}}}$$

$$= \left(\frac{-\mathrm{j}(1+\mathrm{j})}{-\mathrm{j}\,\mathrm{j}}\right)^2 \frac{1}{\omega_0 C_{\mathrm{CB}} R_{\mathrm{C}} R_{\mathrm{B}}} = (1-\mathrm{j})^2 \frac{1}{\omega_0 C_{\mathrm{CB}} R_{\mathrm{C}} R_{\mathrm{B}}},$$

$$\therefore$$

$$g_{\mathrm{m}} < \frac{2}{\omega_0 C_{\mathrm{CB}} R_{\mathrm{C}} R_{\mathrm{B}}} \tag{7.113}$$

because[8] $|1-\mathrm{j}| = \sqrt{2}$ and $|\mathrm{j}| = 1$. Simplified (7.113) is the condition for stability of a double-tuned CE amplifier around its centre frequency ω_0. For example, for a transistor with $R_{\mathrm{C}} = 1\,\mathrm{M\Omega}$, $R_{\mathrm{B}} = 3\,\mathrm{k\Omega}$, $C_{\mathrm{CB}} = 1\,\mathrm{pF}$, at $f = 10\,\mathrm{MHz}$, application of (7.113) suggests that $g_{\mathrm{m}} < 10\,\mu\mathrm{S}$, which is really not useful amplification, which further demonstrates the stability issue of a simple CE amplifier.

7.7.1.4 Cascode RF and IF Amplifiers

From the previous discussion and aside from the reasons related to Miller effect, for practical purposes we conclude that using a cascode amplifier is also recommended from the stability perspective because the CE feedback path is broken and the cascode amplifier is inherently stable.

We conclude that a combination of the two single transistor amplifier stages, CE and CB, is inherently stable because the output of the CE stage drives the input of the CB stage. We already know that the input resistance of a CB stage R_{in} is very low, therefore the overall gain of CE stage $g_{\mathrm{m}}R_{\mathrm{in}}$ is very low, making the CE stage stable. At the same time, the CB stage is inherently stable because there is no feedback path. Hence, two stable cascaded amplifiers are unconditionally stable.

Schematic diagrams of two commonly used cascode RF amplifier structures are shown in Fig. 7.32. The advantage of using BJT devices is the higher g_{m} gains compared to MOSFET devices. On the other hand, MOSFET devices have very high input resistance, which makes the FET input stage almost an ideal load for a voltage source driver. To take advantage of both devices, in modern BiCMOS integrated technologies, a cascode amplifier is a combination of common-source (CS) and CB amplifiers.

7.7.1.5 Unilateralisation of CE Amplifier

For quite a while, using a cascode RF amplifier instead of a simple CE RF amplifier has been almost an automatic choice because of all the good qualities that we have learned about so far. However, not everything is lost for the CE amplifier. As we advance towards lower power consumption of wireless electronics, which is achieved mostly by lowering the power supply voltage, the main drawback of a cascode amplifier is becoming more visible. In order to keep both of the transistors in forward active mode, a higher power supply voltage is needed because at least two CE voltages v_{CE} must fit between the power rails.

Fortunately, since the times of tube amplifiers (which are still used in some very high-power RF amplifiers), at least two techniques have been known that help improve the stability of CE RF

[8] Again, use Pythagoras' theorem.

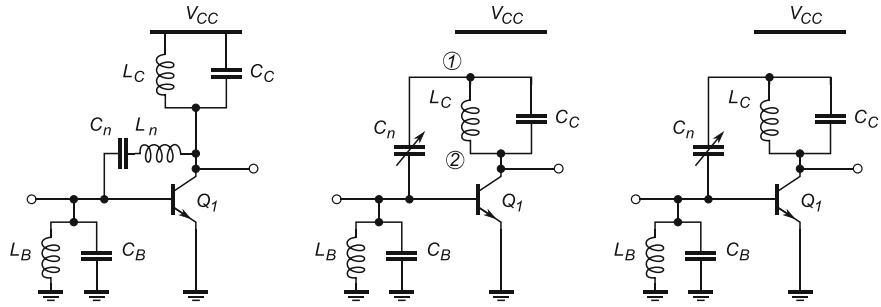


Fig. 7.34 Unilaterisation by resonating out the internal capacitance (*left*), by tunable capacitive feedback (*centre*), and by transformer tapping of the feedback signal (*right*)

amplifiers. In principle, the idea is very simple and general. Once we realized that main cause of CE amplifier instability is the fed-back signal with the right phase relative to the input signal phase and recognized that this fed-back signal is due to parasitic reactances inside the transistor that connect the output and input terminals of the amplifier, the solution to the problem came naturally. If another feedback path, external to the transistor, is created with a signal that is exactly the same as the parasitic fed-back signal, but with the opposite phase, then the sum of the parasitic feedback signal and the external feedback signal can be made zero. In other words, the parasitic feedback signal is "neutralized" at the input terminal node. Cancelation of the feedback signal makes the transistor a truly unidirectional device (i.e., no feedback path); the process is sometimes referred to as "unilaterisation".

With that idea in mind, the unilaterisation process becomes a matter of tapping the feedback signal at the output node and using the right components in the external feedback path to establish the right phase and amplitude of the cancelation signal. In the first variant of signal neutralization, where the idea of "resonating out" is used, in parallel to the internal parasitic C_{CB} , a serial inductor–capacitor L_nC_n path is added so that the overall reactance of all parasitic and external components is removed. In addition, the serial C_n capacitor removes the DC path from the output to the input terminal, Fig. 7.34 (left).

A second variant of the same idea, shown in Fig. 7.34 (centre), applies an external capacitive feedback C_n in which the feedback signal is tapped from node ① at the top of L_CC_C tank, where the phase is opposite compared to node ②. The overall effect is, again, that the parasitic feedback signal is neutralized.

As another variant, the feedback signal can be tapped using inductive coupling (i.e., a transformer), as shown in Fig. 7.34 (right).

It should be noted that the neutralization techniques based on discrete passive components in the feedback path are limited by the component's self-resonant frequencies. At some point, it becomes necessary to replace the discrete components with distributed components based on transmission lines. Indeed, modern HF transceivers are designed mostly using HF IC technology.

7.7.2 Single-Stage CB RF Amplifier

In cases when the source RF signal is in the form of current (i.e., a very high impedance source), it is beneficial to have an RF amplifier with low input resistance. We know that CB amplifiers satisfy the input impedance requirement (Fig. 7.35) because $R_{in} \approx {}^{1}/_{g_m}$ input resistance in the forward active mode. Both the input, L_EC_E and output L_CC_C resonators are tuned to the same frequency. However, in this configuration there is no feedback path from the output to the input node, hence the CB

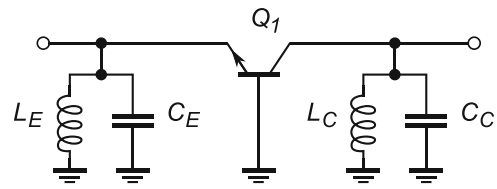


Fig. 7.35 A CB RF amplifier

amplifier is inherently stable. Use of matching transformers makes it possible to design optimal loading impedance.[9]

7.7.3 Insertion Loss

Careful analysis of a parallel LC loading tank interaction with the active amplifying device in a tuned RF amplifier uncovers another interesting and important phenomenon, which is actually the impedance-matching problem in disguise. At low frequencies, i.e., below resonance ω_0, impedance of the inductor Z_L is very low. At the same time, the voltage output is the collector current multiplied by the overall loading impedance, which is to say that there is *insertion loss* of the voltage signal compared with the signal level at resonance ω_0, where the loading resistance R_D is very high. At frequencies above resonance, impedance of the capacitor Z_C is low with the same voltage dividing effect on the output voltage level. Intuitively, we conclude that in the case of the ideal LC resonator, i.e., $R_D \to \infty$, there would be no insertion loss. However, resonating tanks are real and their dynamic resonances are finite and they appear in parallel with the collector resistance R_C (which is usually high).

Being effectively a bandpass filter, insertion loss (IL) is an important figure of merit of an LC resonator. In general, the overall RF amplifier output resistance Z_C consists of the parallel combination $R_C||R_D$,

$$Z_C = \frac{R_C R_D}{R_C + R_D} = R_C \frac{R_D}{R_C + R_D} = R_C \times IL, \tag{7.114}$$

where, insertion loss IL is defined by the resistive ratio $R_D/(R_C + R_D)$. It is common practice to express the insertion loss in units of dB, as

$$IL_{dB} = 20 \log \frac{R_D}{R_C + R_D} \quad \text{dB}, \tag{7.115}$$

where, the ideal case of $R_D \to 0$ leads to $IL \to 0$ dB; in any other case, the IL is a negative number of dBs, with the other extreme of $IL \to -\infty$ indicating no power transfer through the LC tank.

7.8 Summary

In this section, we reviewed the fundamental concepts of LF amplifiers and developed intuitive views of the internal amplifier operation. In our review, we concluded that the important parameters for any amplifier are its input and output resistance and its gain. We also realized that two basic electrical

[9]For matching transformers, see Sect. 4.1.7.2.

variables, voltage and current, determine the total of four possible amplifier transfer functions: voltage gain A_v, current gain A_i, voltage-to-current gain G_m, and current-to-voltage gain A_R.

As the first step in amplifier design, gain of the active devices (either BJT or FET) is set by their DC operating point, and the subsequent signal analysis is simplified by omitting details of the biasing circuit, i.e., it is simply assumed that the active devices somehow had their gain set. We reviewed basic circuit configurations for setting up stable DC operating points for the active devices. In the first approximation, a BJT device is seen as a current amplifier, where the current gain β serves as the multiplication factor in the relationship between the base and collector currents. After the collector current is passed through a resistive load R_C, which is effectively seen as a current-to-voltage amplifier, the combination of the two is seen as a G_m amplifier, i.e., input base current is amplified into voltage across the loading resistor.

Transformation of LF baseband amplifiers into RF amplifiers is done by adding bandpass filtering stages both at the input and output sides of the LF amplifier. Frequency analysis of RF amplifiers introduced concepts of Miller capacitance, amplifier stability, and insertion loss.

Problems

7.1. For a single NPN BJT, draw the schematic symbol and indicate potentials at the three terminals, i.e., the V_C, V_B, and V_E, and their relationship assuming the transistor is turned on, i.e., it is operating in the forward active region. Repeat the exercise using a PNP BJT.

7.2. Estimate the voltage gain A_v for the circuit in Fig. 7.36d if $R_C = 10\,\text{k}\Omega$ and $R_E = 100\,\Omega$. Express the result in dB.

7.3. The voltage gain of the circuit in Fig. 7.36d, for $R_C = 10\,\text{k}\Omega$, $R_E = 100\,\Omega$, $I_S = 100\,\text{fA}$, and $V_{BE} = 768.78\,\text{mV}$ at temperature $T = 25°\text{C}$, is recalculated at an operating frequency of $f = 10\,\text{MHz}$. In addition, in parallel with the emitter resistor R_E a capacitor C is connected. Estimate the new voltage gain A_v for: (a) $C = 1\,\mu\text{F}$ and (b) $C \to \infty$. How large is the gain difference for these two cases, in percentages? How large is the gain difference in comparison with the gain calculated in Problem 7.2? Can you draw any useful conclusions?

7.4. A signal generator is coupled with the CE amplifier in Fig. 7.36d through a serial capacitor $C = 1\,\mu\text{F}$. Estimate the range of frequencies where the CE amplifier should be used, if $R_C = 9.9\,\text{k}\Omega$, $R_E = 100\,\Omega$, $C_{CB} = (1/\pi)\text{pF}$, $R_1 = 2\,\text{k}\Omega$, $R_2 = 2\,\text{k}\Omega$.

7.5. Estimate the range of frequencies where the CE amplifier in Fig. 7.36c should be used, if the base side inductor $L = 2.533\,\text{pF}$, $R_C = 9.9\,\text{k}\Omega$, $R_E = 100\,\Omega$, $C_{CB} = 1\,\text{pF}$.

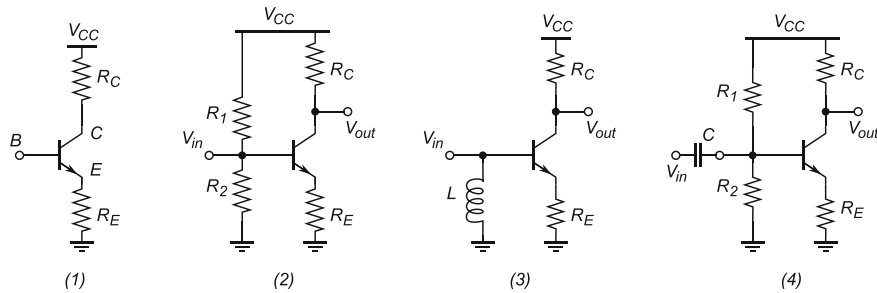


Fig. 7.36 Schematic networks for Problems 7.2, 7.3, 7.4, 7.5, 7.6, and 7.7

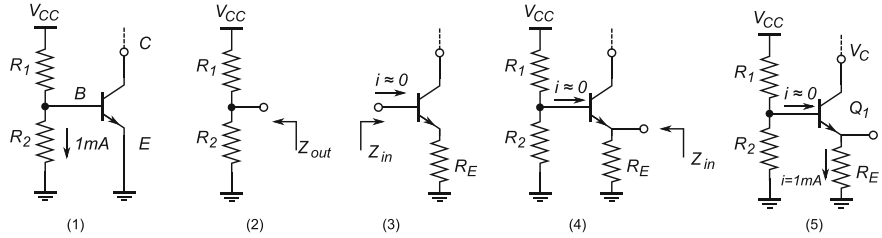


Fig. 7.37 Schematic networks for Problems 7.8, 7.9, and 7.10

7.6. For the CE amplifying circuit in Fig. 7.36a, estimate the Miller capacitance C_M if $R_C = 9.9\,k\Omega$, $R_E = 100\,\Omega$, $C_{CB} = 1\,pF$.

7.7. Estimate the input side bandwidth of the CE amplifier in Fig. 7.36b if $R_C = 9.9\,k\Omega$, $R_E = 100\,\Omega$, $C_{CB} = (1/\pi)pF$, $R_1 = 2\,k\Omega$, $R_2 = 2\,k\Omega$.

7.8. For a network shown by the schematic diagram in Fig. 7.37a:

(a) Assuming the base–emitter diode threshold voltage is $V_{th}(BE) = 0\,V$, i.e., an ideal BE diode, find value(s) of R_2 so that the transistor Q_1 is turned on. What potential V_C is required at the collector node C to maintain the saturation mode of operation?
(b) Assuming the base–emitter diode threshold voltage is $V_{th}(BE) = 1\,V$, i.e., a more realistic BE diode, find value(s) of R_2 so that the transistor Q_1 is turned on. What potential is required at the collector node V_C to maintain the saturation mode of operation?

7.9. Estimate impedances looking into the networks in Fig. 7.37b–d.

7.10. What is the required resistor ratio R_1/R_2 for the network in Fig. 7.37e, so that the transistor Q_1 is operating in saturation, if $V_{CC} = 10\,V$ and $R_E = 1k\Omega$?

(a) assuming the base–emitter diode threshold voltage is $V_{th}(BE) = 0\,V$, i.e., an ideal BE diode, find value(s) of R_2 so that the transistor Q_1 is turned on. What potential is required at the collector node V_C to maintain the saturation mode of operation?
(b) assuming the base–emitter diode threshold voltage is $V_{th}(BE) = 1\,V$, i.e., a realistic BE diode, find value(s) of R_2 so that the transistor Q_1 is turned on. What potential is required at the collector node V_C to maintain the saturation mode of operation?

7.11. In the amplifier in Fig. 7.23, resistors R_1 and R_2 make Q_1 a base voltage divider which should be set such that their current $I_{R_{1,2}} \approx 1/10$ of I_E (ignore the base current).

Data: Voltage gain $A = -8$, $V_{CC} = 9\,V$, $I_E \approx 2mA$, $\beta = 100$, $V_B = 1/3 V_{CC}$, $R_L = 2\,k\Omega$, $V_{BE} = 0.7\,V$, $R_{sig} = 10\,k\Omega$, and $C = \infty$. Estimate:

(a) The base voltage V_B
(b) R_1 and R_2
(c) Théveninresistance at the base node
(d) R_E
(e) I_C
(f) g_m
(g) r_e
(h) R_C

7.12. For the amplifier in Fig. 7.23 and the data below, estimate:

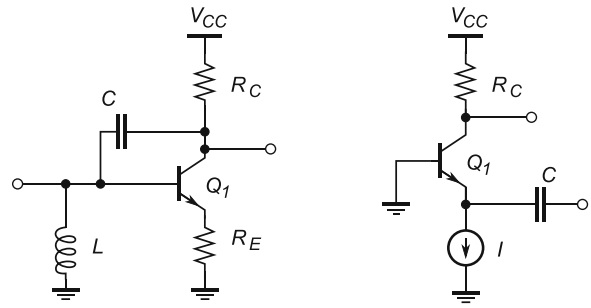


Fig. 7.38 Schematic for Problems 7.17 (*left*) and 7.13 (*right*)

(a) Collector resistance R_C
(b) Emitter resistance R_E
(c) Voltage at the base node V_B
(d) Resistance looking into the base R_{in}
(e) Gate-biasing resistors R_1 and R_2
(f) Small emitter resistor r_e
(g) Small resistor R_0 for given A_v
(h) Emitter capacitor C_E for given 3 dB point at the output
(i) Input capacitor C for given 3 dB point at the input
(j) For the component values found in parts (a) to (i), find the voltage gain A_v when $V_{out} = 2.5\,V$.

Data: $V_{CC} = 10\,V$, $V_{BE} = 0.6\,V$, $R_{th} = R_1||R_2 = 0.1R_{in}$, $V_T = 25\,mV$, $f_{3\,dB} = 20\,Hz$ at the output side, $f_{3\,dB} = 10\,Hz$ at the input side, $\beta = 99$, voltage gain $A_v = -100$ when $I_C = 1\,mA$, $V_E = 1\,V$, and $V_{out} = {}^1/_2 V_{CC}$.

7.13. The circuit shown in Fig. 7.38 (right) is a CB amplifier with $\beta = \infty$, $R_C = 7.5\,k\Omega$, $I = 0.5\,mA$, $C = \infty$, and $V_{CC} = 5\,V$. Estimate:

(a) DC voltage at the collector
(b) $g_m(Q_1)$
(c) AC voltage gain, $A = v_C/v_i$

7.14. The resistance seen by looking into a BJT emitter is $R_{out} = 100\,\Omega$. The resistance looking into the base is $R_{in} = 100\,k\Omega$. For $\beta = 99$, find the reflected resistance at the base node and R_E. (Note: ignore r_e.)

7.15. For a grounded emitter amplifier powered by $V_{CC} = 10\,V$ with collector resistor $R_C = 5.1\,k\Omega$, estimate the voltage gain for: (a) $V_{out} = 7.5\,V$, (b) $V_{out} = 5\,V$, (c) $V_{out} = 0.2\,V$.

7.16. If V_{BE} voltage of a BJT changes by 18 mV, what is the change of I_C, expressed in dB? What if V_{BE} changes by 60 mV? Note: Use $kT/q = 25\,mV$.

7.17. For the amplifier in Fig. 7.38 (left) with $C = 1\,pF$, the small signal voltage gain is $A = -99$. Estimate the value of the inductor L so that the input stage resonates at $f_0 = 15.915\,MHz$. Assume base current to be zero.

Chapter 8
Sinusoidal Oscillators

Abstract Communication transceivers require *oscillators* that generate pure electrical sinusoidal signals ("tones") for further use in modulators, mixers, and other circuits. Although oscillators may be designed to deliver other waveforms as well, e.g. square, triangle, and sawtooth waveforms, if intended for applications in wireless radio communications, the sinusoidal waveform is probably the most important one. A good sinusoidal oscillator is expected to deliver either a voltage or a current signal that is stable both in amplitude and frequency. Because a variety of oscillator structures are available that are suitable for generation of sinusoidal waveforms, circuit designers make the choice mostly based on their personal preference for one particular type of oscillator. In this chapter, we study several oscillator circuits, with emphasis on understanding the underlying principles, rather than very detailed analysis of any special oscillator type.

8.1 Criteria for Oscillations

Because the amplitude of the signal inside an oscillator circuit may (theoretically) increase infinitely, that is, the signal amplitude becomes large, we have to accept the conclusion that small signal circuit analysis is not an applicable method. Large signals imply *nonlinear circuits*, which means that we have to apply numerical methods in order to estimate the circuit's internal states. Consequently, oscillator design is as much an art as it is engineering. The good news, however, is that almost all oscillator circuits may be evaluated intuitively using a general block diagram (see Fig. 8.1), which portrays an oscillator as a closed loop system. In the forward signal path, there is an amplifier with gain A, which may be either non-inverting or inverting. The feedback path contains a passive network with gain of $\beta < 1$ that controls the overall phase shift around the loop.

In order to develop intuition about the loop operation, we start with the closed loop system that is shown as an *open loop*, Fig. 8.1, where the feedback signal v_f is disconnected from the signal v_i at the input terminal of the amplifier. By inspection, we write equations,

$$v_\mathrm{o} = A\, v_\mathrm{i}, \tag{8.1}$$

$$v_\mathrm{f} = \beta\, v_\mathrm{o} = \beta A\, v_\mathrm{i}, \tag{8.2}$$

which show that, on its journey around the loop, the signal v_i first becomes amplified by the amplifier and then attenuated by the feedback network. From the signal's perspective, the forward path with gain A and the feedback path with gain β are perceived as two gain stages in series. Hence, the total gain along the path is the product of two gains, i.e., $A \times \beta$. Simply put, in the case when the feedback

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8_8,

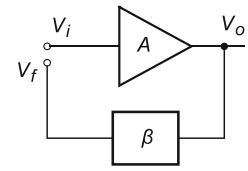


Fig. 8.1 Block diagram of a basic oscillator feedback loop

signal v_f is in phase (i.e., there is non-destructive addition) with the signal v_i and when its amplitude has increased, i.e., $v_\mathrm{f} > v_\mathrm{i}$, then the initial signal v_i is said to be amplified. It is easy to see that under the given conditions and after the loop is closed, the signal's amplitude keeps increasing indefinitely on each cycle around the loop. Therefore, it is logical to make the conclusion that, if the total gain around the loop is no less than one, the closed loop becomes *unstable*. That is, the amplifier gain A must be large enough to compensate for the signal loss in the passive feedback network.

Although control theory offers several commonly used methods for evaluating the stability of closed loop systems (for instance, the Bode plot, the Routh–Hurwitz stability criterion, root–locus analysis, and the Nyquist stability criterion are applicable to linear, time-invariant (LTI) systems and the Lyapunov stability criterion applies to nonlinear dynamic systems), none of these criteria is universal. In practice, we usually use more than one of the criteria to reach a conclusion about the system's stability. For the purposes of determining under what conditions a linear electronic circuit oscillates, we introduce the intuitive (and also non-perfect) "Barkhausen Stability Criterion", which states that, if a feedback circuit is to maintain oscillations, then

- The net gain around the feedback loop must be no less than one, i.e., $|A\beta| \geq 1$.
- The net phase shift around the loop must be a positive integer multiple of 2π radians, or $n \times 360^\circ$ (where n is an integer).

The Barkhausen Criterion is a necessary but not sufficient condition for oscillation. Both $A = A(\omega)$ and $\beta = \beta(\omega)$ are frequency dependent, therefore the conditions listed in the Barkhausen Criterion are satisfied at the same time only at a single frequency. There are, therefore, two necessary conditions for sustaining the loop oscillations: one related to the loop gain and one to the phase shift. In practical designs, of course, the initial loop gain must be greater than unity in order to increase the probability that the circuit can actually start oscillating, i.e., in a well-designed oscillator there should be no problem for the circuit in starting the oscillations on its own. In addition, it is necessary to build in some kind of mechanism to keep limiting the amplitude of the oscillation, so that the output signal does not become clipped or distorted.

From the block diagram of a general feedback amplifier, Fig. 8.2 (left), by inspection we write the loop transfer function as:

$$v_\mathrm{o} = A(v_\mathrm{in} + \beta v_\mathrm{o}),$$

$$\therefore$$

$$\frac{v_\mathrm{o}}{v_\mathrm{in}} = \frac{A}{1 - \beta A}. \tag{8.3}$$

Naturally, this loop transfer function reveals the instability condition (i.e., oscillation) if $\beta A = 1$ and thus gives the same result as the previous intuitive analysis. Hence, an oscillator may be thought of as a positive feedback amplifier that is intentionally made unstable. This means that any small voltage internally generated, for example, by thermal noise, rapidly builds up to a large amplitude voltage at the output node of the closed loop system.

Our simplified methodology for closed loop analysis is based on establishing the open loop parameters first. At the end of the active forward signal path, the amplifier perceives the passive

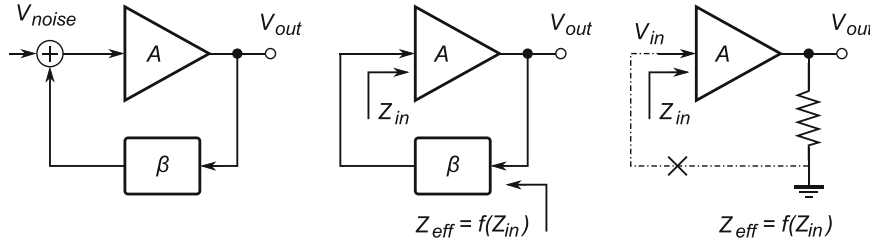


Fig. 8.2 A general closed loop showing injected noise (*left*); an oscillator closed loop (*centre*); and its equivalent open loop model (*right*)

feedback network as the load Z_{eff}, which is equivalent to the feedback network's input impedance, Fig. 8.2 (centre). This is a good time to note that a closed loop system is a one-port network, i.e., the only output terminal is at the oscillator circuit. We keep in mind that under the right conditions the internal thermal noise is sufficient to serve as the initial "input signal" until the loop starts to oscillate. To continue our simplified methodology, as we already know, in order to correctly calculate Z_{eff} impedance, we must take into account the amplifier's input impedance because it does affect the value of $Z_{\text{eff}} = f(Z_{\text{in}})$, Fig. 8.2 (centre). Using the open loop model enables us to estimate the loop gain by applying a signal that varies (in both amplitude and frequency) to the amplifier's input terminal and measuring the signal v_{out} at the output node, Fig. 8.2 (right).

8.2 Ring Oscillators

A simple example of a closed loop circuit that can generate a square pulse waveform is based on the signal propagation delay through a chain of inverters by using the principle of an inverting amplifier that is driving its own input terminal. If we observe the input terminal of the first inverter at an arbitrary point in time t_0 and if, for the sake of argument, we observe a positive pulse, we could "join" the pulse on its trip around the loop, Fig. 8.3 (top). After propagating through the first inverter the pulse becomes negative; after propagating through the second inverter the pulse is switched to positive again. It is straightforward to generalize and conclude that after every even inverter stage the pulse has as same polarity as the one at the first input, while after every odd inverter stage the pulse has the opposite polarity. Therefore, we conclude that for a chain of $(2n+1)$ inverters, a signal with opposite polarity takes $\Delta t = (2n+1)\,t_{\text{d}}$ seconds to travel around the loop and change the signal polarity at the input of the first inverter. The full period of a periodic signal is measured between two falling or two rising edges, hence a ring oscillator produces a square signal whose period is $T = 2\,(2n+1)\,t_{\text{d}}$ seconds, where t_{d} is the signal propagation time through each stage, Fig. 8.3 (bottom). We should note that the "output" terminal is chosen arbitrary: it could be taken from any point around the loop.

Example 8.1. An average propagation delay through a single inverter gate is estimated as $t_{\text{d}} = 0.998\,\text{ns}$. How many inverter gates are needed to design a ring oscillator working at $f = 1\,\text{MHz}$?

Solution 8.1. A 1 MHz signal is equivalent to the period of $T = 1\,\mu\text{s}$. Therefore, we calculate the number of required gates as

$$T = 2\,(2n+1)\,t_{\text{d}} \quad \therefore \quad n = \frac{1}{2}\left(\frac{T}{2t_{\text{d}}} - 1\right) = \frac{1}{2}\left(\frac{1\,\mu\text{s}}{2\times 0.998\,\text{ns}} - 1\right) = 250,$$

where, the "average" delay is determined by characterization of a large number of manufactured digital gates in a given process.

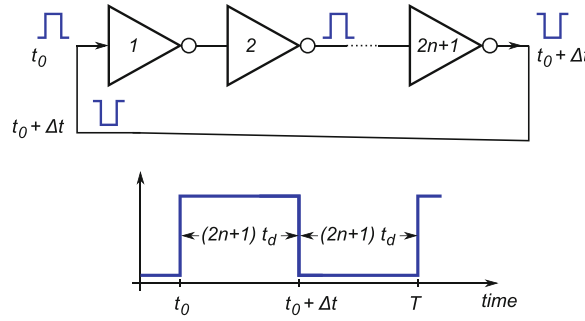


Fig. 8.3 Ring oscillator schematic diagram

Ring oscillators are often used in IC technology as sensors of process variations. The oscillator frequency depends upon the propagation delay of each inverting stage. Further, the stage propagation time depends upon the internal capacitances and resistances, which are very much process dependent. Therefore, by measuring the output frequency we are able to quantify the process variation.

8.3 Phase-Shift Oscillators

The oscillator architecture that probably best illustrates the use of the Barkhausen Criterion is known as a "phase-shift oscillator" (see Fig. 8.4). An inverting amplifier with gain $A = -a$ that is used in the forward signal path is assumed to have infinite input resistance, i.e., there is no current flow into its input terminal. The feedback network with gain β consists of a classical RC ladder network of at least three RC sections. Although, this feedback network arrangement is also occasionally found with R and C interchanged, the arrangement shown here is more common. To satisfy the Barkhausen Criterion, a feedback path phase shift of exactly 180° is required in order to align the feedback signal with its initial phase, because the inverted amplifier gain introduces a first signal inversion of 180°. Therefore, the frequency of oscillation is equal to the frequency at which the phase shift introduced by the RC network is exactly 180°.

Systematic analysis of the ladder network starts at output node ③ of the network, which can be treated as the output node of the passive feedback path, and progresses back to input terminal ⓪ of the feedback path. Hence, by inspection of network in Fig. 8.4, we write

$$i_3 = \frac{v_3}{R} \qquad v_2 = v_3 + \frac{1}{\mathrm{j}\omega C} i_3 = v_3 + \frac{v_3}{\mathrm{j}\omega RC}, \tag{8.4}$$

$$i_2 = \frac{v_2}{R} \qquad v_1 = v_2 + \frac{1}{\mathrm{j}\omega C}(i_2 + i_3) = v_3 + \frac{3v_3}{\mathrm{j}\omega RC} - \frac{v_3}{(\omega RC)^2}, \tag{8.5}$$

$$i_1 = \frac{v_1}{R} \qquad v_0 = v_1 + \frac{1}{\mathrm{j}\omega C}(i_1 + i_2 + i_3) = v_1 + \frac{(v_1 + v_2 + v_3)}{\mathrm{j}\omega RC}$$

$$= v_3 + \frac{6v_3}{\mathrm{j}\omega RC} - \frac{5v_3}{(\omega RC)^2} - \frac{v_3}{\mathrm{j}(\omega RC)^3}, \tag{8.6}$$

therefore,

$$v_0 = \left[v_3 - \frac{5v_3}{(\omega RC)^2}\right] + \mathrm{j}\left[\frac{v_3}{(\omega RC)^3} - \frac{6v_3}{\omega RC}\right] = \Re(v_0) + \mathrm{j}\Im(v_0). \tag{8.7}$$

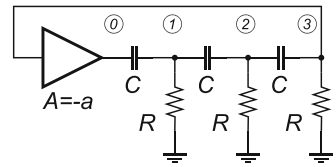


Fig. 8.4 Simplified schematic diagram of a phase-shift oscillator

The Barkhausen Criterion requires that the total phase shift around the loop should be exactly 2π, which means that the imaginary term $\Im(v_0)$ in (8.7) must equal zero, i.e.

$$\Im(v_0) = 0 \quad \therefore \quad \frac{v_3}{(\omega RC)^3} - \frac{6v_3}{\omega RC} = 0 \quad \therefore \quad \omega_0 = \frac{1}{\sqrt{6}RC}, \tag{8.8}$$

which defines the oscillation frequency. Substituting (8.8) into (8.7) gives:

$$v_0 = v_3 - \frac{5\,v_3}{(1/\sqrt{6}RC)^2\,(RC)^2} \quad \therefore \quad \frac{v_3}{v_0} = -\frac{1}{29} = \beta. \tag{8.9}$$

This is a very surprising result indeed: (8.9) states that the feedback path has a gain that is independent of the component values, i.e., $\beta = 1/29$. Following the Barkhausen Criterion $|\beta A| = 1$, we conclude that for this type of phase-shift oscillator, we must design the amplifier with inverting gain of at least $A = -29$. If the amplifier used inside the phase-shift oscillator has less than infinite input impedance, as would be the case for a real BJT amplifier, the derivations above would need to be modified. The modified equations are more difficult to solve and do not provide any further insight into this oscillator, hence they are omitted here.

Example 8.2. Estimate the minimum gain in dB of an inverting amplifier used in a phase-shift oscillator (Fig. 8.4).

Solution 8.2. To a first approximation, the oscillator loop gain must be at least one, hence the amplifier must compensate for the passive network attenuation of $\beta = 1/29 = -29.25\,\text{dB}$, by adding its own gain of $A = +29.25\,\text{dB}$.

Aside from being very good educational examples, phase-shift oscillators are used mostly at audio frequencies because the ladder network becomes impractical at higher radio frequencies.

8.4 RF Oscillators

Our introduction and treatment of RF oscillators follows a slightly different path from most textbooks. We first analyze four general RLC types of passive feedback network and then we use them inside some of the most common RF oscillator topologies. Although there is an infinite number of feedback network topologies that could be used in oscillators, if we impose the constraint that a minimal number of components is used, then only a small number of network topologies are suitable in the feedback path of an RF oscillator.

8.4.1 Tapped L, Centre-Grounded Feedback Network

Let us consider an RLC feedback network (see Fig. 8.5), making the following assumptions:

- The network operates near its resonant frequency .
- The Q factor is high, i.e., ten or higher.

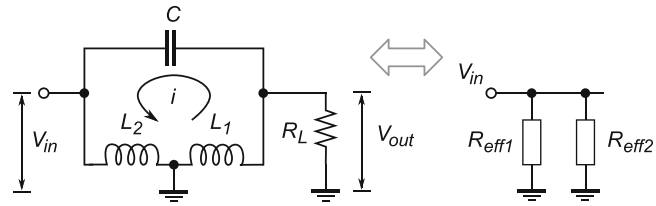


Fig. 8.5 Tapped L, centre-grounded network (*left*), and its equivalent representation for the energy dissipation calculations (*right*)

- Inductors L_1 and L_2 are not coupled.
- The network's Q factor is the effective Q of $L_1 + L_2$.

A further assumption is that the inductors have equal values of Q factor.

In Sect. 8.1, we discussed a general model of a feedback loop (Fig. 8.2) and concluded that, in order to characterize the feedback network, the following three parameters are required: the loop's resonant frequency ω_0; the passive feedback path voltage gain β; and the effective input resistance R_{eff} of the feedback network.

In order to evaluate the resonant frequency ω_0, we need to recognize that the resonating current i stays within the L_1, L_2, C loop (see Fig. 8.5). The inductance of two inductors in series equals the sum of the two inductances, therefore we write:

$$\omega_0^2 = \frac{1}{(L_1 + L_2)C}. \tag{8.10}$$

The fact that the LC loop is tapped at two points (at the small signal ground between the two inductors and at the top of the loading resistor) does not influence the value of the resonant frequency—it is set by the total LC in the loop, as we concluded in Sect. 5.2.

The voltage gain $\beta = v_{\text{out}}/v_{\text{in}}$ of the feedback network can be evaluated as follows. At the resonant frequency, assuming a high Q factor, most of the power just circulates around the loop between the inductors and the capacitor (due to low thermal losses). The circulation of power around the resonant circuit may be represented by the continuous current i shown in Fig. 8.5 (left).

By inspection, we write the network equations as:

$$v_{\text{in}} = i\,\mathrm{j}\omega L_2 \qquad v_{\text{out}} = -i\,\mathrm{j}\omega L_1 \quad \therefore \quad \beta = \frac{v_{\text{out}}}{v_{\text{in}}} = \frac{-i\,\mathrm{j}\omega L_1}{i\,\mathrm{j}\omega L_2} = -\frac{L_1}{L_2}, \tag{8.11}$$

that is, the voltage gain β of a tapped L, centre-grounded feedback network is set by the inductive voltage divider.

Calculation of the effective resistance R_{eff} is a bit more complicated, due to its dependence upon the amplifier's input resistance value, which is modelled as the loading resistor R_{L} in Fig. 8.5 (left). Keep in mind that the input node of a feedback network is the one which is connected to the output node of the amplifier, while the output node of the feedback network is loaded by the input impedance of the amplifier, Fig. 8.2 (left). As already found, at resonance, the effective resistance R_{eff} of an LC resonator is purely resistive. Moreover, the resonator is loaded by impedance R_{L} (i.e., input impedance of the amplifier). One way of calculating the effective input impedance R_{eff} of the feedback network, which is a function of the load impedance, is by power calculation.

The total RMS power that is being put into the network is,

$$P_{\text{rms}}(in) = \frac{v_{\text{in}}^2}{2\,R_{\text{eff}}}. \tag{8.12}$$

Mathematically, we can imagine that the total input power is split between two effective resistances as follows.

One part of the input power is delivered to the external load impedance R_L at the output. After substituting (8.11), we write

$$P_{ext} = \frac{v_{out}^2}{2\,R_L} = \frac{\left[v_{in}\left(-\frac{L_1}{L_2}\right)\right]^2}{2\,R_L} = \frac{v_{in}^2\left(\frac{L_1}{L_2}\right)^2}{2\,R_L} = \frac{v_{in}^2}{2\,R_L\left(\frac{L_2}{L_1}\right)^2} = \frac{v_{in}^2}{2\,R_{eff}},$$

$$\therefore$$

$$R_{eff\,1} = R_L\left(\frac{L_2}{L_1}\right)^2. \tag{8.13}$$

Due to the finite Q factor, some of the total input power is dissipated in the internal LC circuit. At resonance, the LC resonator is equivalent to its dynamic resistance $R_D = Q\omega(L_1 + L_2)$ that is effectively connected between the top and the bottom of the LC circuit, i.e., between the input and output nodes in Fig. 8.5. Hence, the power P_{int} dissipated in this resistor is

$$P_{int} = \frac{v_{R_D}^2}{2\,R_D} = \frac{(v_{in} - v_{out})^2}{2Q\omega(L_1 + L_2)} = \frac{\left[v_{in} + v_{in}\left(\frac{L_1}{L_2}\right)\right]^2}{2Q\omega(L_1 + L_2)}$$

$$= \frac{v_{in}^2(L_1 + L_2)}{2\,Q\omega L_2^2} = \frac{v_{in}^2}{2\,\frac{Q\omega L_2^2}{L_1 + L_2}},$$

$$\therefore$$

$$R_{eff\,2} = \frac{Q\omega L_2^2}{L_1 + L_2}. \tag{8.14}$$

Thus, from the input power perspective, because we referenced both powers P_{int} and P_{ext} relative to the input voltage v_{in}, it is being dissipated into two separate, parallel effective impedances, R_{eff1} and R_{eff2}. Because both of these power dissipations occur simultaneously, the input power must be the sum of the two effective powers:

$$P_{in} = P_{in1} + P_{in2} = \frac{v_{in}^2}{2R_{eff}} = \frac{v_{in}^2}{2(R_{eff\,1}||R_{eff\,2})} = \frac{v_{in}^2}{2\left(\frac{1}{R_{eff\,1}} + \frac{1}{R_{eff\,2}}\right)}. \tag{8.15}$$

Note that, from the power distribution perspective, the effective resistors are combined in parallel. Hence, the effective input impedance is estimated as

$$R_{eff} = R_{eff\,1}||R_{eff\,2} = R_L\left(\frac{L_2}{L_1}\right)^2 \,||\, \frac{Q\omega L_2^2}{L_1 + L_2}. \tag{8.16}$$

Equations (8.10), (8.11), and (8.16) define the three main parameters of a tapped L, centre-grounded feedback network. The oscillator design process now can be split into two parts: the amplifier design for the forward signal path and the passive RLC network design. In order to acquire a complete set of commonly used feedback networks, we need to define three additional network

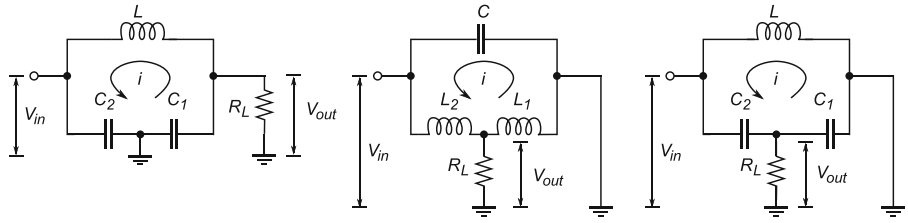


Fig. 8.6 RLC network configurations: tapped C, centre-grounded (*left*); tapped L, bottom-grounded (*middle*); and tapped C, bottom-grounded network (*right*)

configurations. The derivation process for the three main parameters of those network configurations is the same as the derivation process for the tapped L, centre-grounded feedback network. Therefore, the derivations of the formulas for ω_0, β, and R_{eff} for these RLC networks (see Fig. 8.6) are left as an exercise to the reader. Equations (8.17)–(8.25) complete the set of design parameters for passive RLC feedback networks that enable us to design commonly used RF oscillators.

8.4.2 Tapped C, Centre-Grounded Feedback Network

Figure 8.6 (left) shows this type of feedback network. The equations for its main parameters are:

$$\omega_0^2 = \frac{C_1 + C_2}{LC_1C_2}, \tag{8.17}$$

$$\beta = -\frac{C_2}{C_1}, \tag{8.18}$$

$$R_{\text{eff}} = R_{\text{L}} \left(\frac{C_1}{C_2}\right)^2 \,||\, Q\omega L \left(\frac{C_1}{C_1 + C_2}\right)^2. \tag{8.19}$$

8.4.3 Tapped L, Bottom-Grounded Feedback Network

Figure 8.6 (centre) shows this type of feedback network. The equations for its main parameters are:

$$\omega_0^2 = \frac{1}{C(L_1 + L_2)}, \tag{8.20}$$

$$\beta = \frac{L_1}{L_1 + L_2}, \tag{8.21}$$

$$R_{\text{eff}} = R_{\text{L}} \left(\frac{L_1 + L_2}{L_1}\right)^2 \,||\, Q\omega(L_1 + L_2). \tag{8.22}$$

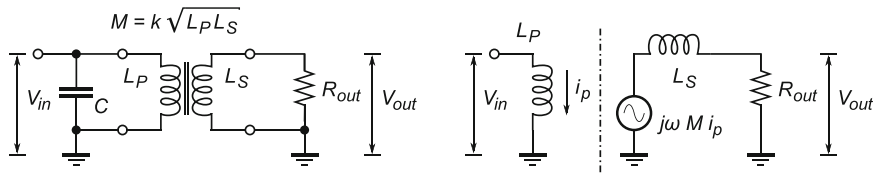


Fig. 8.7 A tuned primary transformer network (*left*) and its equivalent network (*right*)

8.4.4 Tapped C, Bottom-Grounded Feedback Network

Figure 8.6 (right) shows this type of feedback network. The equations for its main parameters are:

$$\omega_0^2 = \frac{C_1 + C_2}{LC_1C_2}, \tag{8.23}$$

$$\beta = \frac{C_2}{C_1 + C_2}, \tag{8.24}$$

$$R_{\text{eff}} = R_{\text{L}} \left(\frac{C_1 + C_2}{C_2} \right)^2 \,||\, Q\omega L. \tag{8.25}$$

8.4.5 Tuned Transformer

An additional type of feedback network that is a member of the same family of RLC networks is the "tuned transformer", which is also very often used in RF sinusoidal oscillators. A tuned transformer is a type of passive feedback network that uses the primary, the secondary, or both transformer coils in parallel with their respective capacitors to create LC resonator tanks. A tuned transformer is said to be either inverting or non-inverting depending upon the relative orientation of the primary and secondary coils. These properties indicate that a transformer is very versatile device that allows for almost arbitrary ratios of the primary inductance L_{P} to the secondary inductance L_{S}, with the additional level of freedom to introduce a phase shift of either 0° or 2π between the primary and secondary sides.

In our brief analysis of non-inverting, primary tuned transformers, Fig. 8.7 (left), we make the following assumptions:

- The coupling factor k $(0 \leq k \leq 1)$ between the primary and secondary is low, i.e., $k \ll 1$. Therefore, the loading effect of the primary on the secondary, and vice versa, may be ignored.
- The output load resistance is much greater than the secondary impedance, i.e., $R_{\text{out}} \gg \omega L_{\text{S}}$. If this condition is not true, an additional phase shift is induced and the frequency of oscillation is different from ω_0.

In the equivalent circuit diagram, Fig. 8.7 (right), coupling between the primary and the secondary is represented by the AC voltage source in the secondary branch. The dashed vertical line indicates that the secondary and the primary are separate circuits. The three parameters of tuned transformer feedback network are determined as follows.

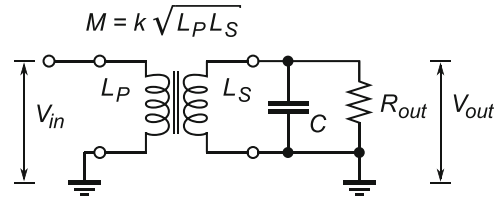


Fig. 8.8 Tuned secondary transformer network

In order to determine the voltage gain factor β, we start by writing an expression for current in the primary network as

$$i_{\mathrm{P}} = \frac{v_{\mathrm{in}}}{Z_{\mathrm{P}}} = \frac{v_{\mathrm{in}}}{\mathrm{j}\omega L_{\mathrm{P}}}, \tag{8.26}$$

which induces voltage in the secondary:

$$v_{\mathrm{ind}} = \pm \mathrm{j}\omega M\, i_{\mathrm{P}} = \pm \mathrm{j}\omega M \frac{v_{\mathrm{in}}}{\mathrm{j}\omega L_{\mathrm{P}}} = \pm \frac{M}{L_{\mathrm{P}}}\, v_{\mathrm{in}}, \tag{8.27}$$

where the $\pm$ sign indicates the phase difference between the primary and the secondary, which depends on the orientation of the transformer coils, and $M = k\sqrt{L_{\mathrm{P}} L_{\mathrm{S}}}$ is the mutual inductance. If the condition $R_{\mathrm{out}} \gg \mathrm{j}\omega L_{\mathrm{s}}$ is satisfied, then it follows that

$$v_{\mathrm{out}} = v_{\mathrm{ind}}, \tag{8.28}$$

which, after substituting (8.28) into (8.27) and rearranging, yields an expression for the voltage gain of the tuned amplifier as

$$\beta = \pm \frac{M}{L_{\mathrm{P}}}. \tag{8.29}$$

The effective resistance R_{eff} of this network is approximately just the impedance of the primary, because we assumed $k \ll 1$. Therefore, at resonance, we write

$$R_{\mathrm{eff}} \approx Q_{\mathrm{P}} \omega_0 L_{\mathrm{P}}, \tag{8.30}$$

where, Q_{P} is the Q factor of the primary transformer.

The resonant frequency ω_0, we simply write as

$$\omega_0^2 = \frac{1}{L_{\mathrm{P}} C}. \tag{8.31}$$

Let us take a brief look at the case of a tuned secondary transformer network (Fig. 8.8). The analysis is similar to the previous case and it leads to the following expressions for the three parameters of the network:

$$\omega_0^2 = \frac{1}{L_{\mathrm{S}} C}, \tag{8.32}$$

$$\beta = \pm \frac{\mathrm{j} M Q_{\mathrm{S\,eff}}}{L_{\mathrm{P}}}, \tag{8.33}$$

$$R_{\mathrm{eff}} = \mathrm{j}\omega L_{\mathrm{P}}. \tag{8.34}$$

The assumptions made in this section are often not used in textbooks, which are usually for a specific type of oscillator network.

8.5 Amplitude-Limiting Methods

Let us stop for a moment here and ask the following questions:

- If noise, with its infinite frequency spectrum, is responsible for starting the oscillation process, how is it that at the output terminal we see only a single tone?
- If the output signal is amplified with each pass through the loop, what keeps its amplitude stable and finite in real circuits?

To answer the first question, we recall that although the internal thermal noise is responsible for providing the initial stimulus to the input terminals of the forward path amplifier, the feedback RLC path is designed to be a very selective bandpass network (by means of the high Q factor). Hence, of all possible tones from the noise frequency spectrum only the one with frequency equal to ω_0 is actually amplified, while all other tones are suppressed. That frequency-selection behaviour of oscillator feedback is the key property of any oscillator circuit.

Now that we have determined intuitively how an oscillator locks on a single tone, the second question needs to be answered. We have already concluded that, in order to start oscillations, it is necessary to design the loop gain larger than unity. Consequently, immediately after powering up the oscillator, the output signal amplitude increases with each passing cycle. Eventually, if the gain stays constant, the output signal becomes a non-sinusoidal (i.e., square) waveform with amplitude that theoretically increases indefinitely. Therefore, in order to generate a non-distorted sinusoidal waveform, some form of amplitude-limiting mechanism is required that prevents the signal amplitude from becoming too large. A few amplitude-limiting schemes are introduced below.

8.5.1 Automatic Gain Control

This is one of the most elaborate and complicated methods for limiting the signal amplitude and it produces the best amplitude control overall. It consists of additional circuitry that measures the output signal amplitude and compares it to the desired amplitude. The error signal is fed back to the amplifier part of the oscillator causing it to either increase or decrease its gain. There are a number of ways to implement this scheme. Indeed, many commercial IC amplifiers provide automatic gain control (AGC) mechanism indicating that this method is very attractive for high-quality designs. We leave details of AGC circuits for a more advanced course.

8.5.2 Clamp Biasing

This method is particularly useful in FET oscillators where, by design, it is possible to limit, i.e., to "clamp", the positive voltage peaks at a level of one diode voltage drop above ground. Depending on the device and the actual implementation, the method may introduce a slight distortion of the waveform that needs to be "cleaned up" by a tuned circuit in order to achieve a good sine wave. The importance of this form of biasing is that it provides a form of AGC which is particularly useful in an oscillator.

8.5.3 Gain Reduction with Temperature-Dependent Resistors

This method is commonly used in audio oscillators and is a bit more complicated, however the output waveform is less distorted. It is, therefore, suitable for oscillators without a tuned circuit (e.g. phase-shift oscillators). It uses devices whose resistance is dependent on temperature, such as thermistors or small light bulbs. As these devices become heated by the increasing amplitude of the oscillations, they reduce the feedback signal. By careful design, schemes using these devices can produce amplitude-stable oscillators with clean sine waveforms.

8.5.4 Device Saturation with Tuned Output

The gain transfer characteristics of many active devices is such that, for a given biasing point, the gain is somewhat reduced as the signal amplitude becomes greater. By designing the feedback network such that the gain is adequate only for low amplitude signals, the device may have inherent properties that provide suitable amplitude limiting as the signal becomes larger. The most drastic type of distortion occurs when the output signal becomes much too large and the nonlinear characteristics of active devices lead into some degree of clipping or squaring of the waveform. However, the frequency spectrum of a square wave whose amplitude is $1V_{\text{pp}}$ contains the fundamental tone whose amplitude is $4/\pi V_{\text{pp}}$, i.e., slightly greater than the PP amplitude of the square wave. Provided there is a tuned, high-Q, resonant circuit after the node where the clipping occurred, a reasonably clean sine wave can be extracted from the square wave. Thus, in many applications the device itself contains a suitable built-in form of amplitude limiting.

8.6 Crystal-Controlled Oscillators

Piezoelectric crystalline materials, quartz being one of the best known, exhibit reciprocal properties relative to their mechanical and electrical behaviour. That is, if an electric potential is applied across a thin sheet of piezoelectric crystal, it physically bends. In return, if a piezoelectric crystal is physically deformed, then the internal electrical charges are separated and a voltage is produced across its plates. Consequently, if a sinusoidal electrical signal with frequency that is equal to the crystal's mechanical resonant frequency is applied across its plates, then a sheet of piezoelectric crystal exhibits both electrical and mechanical resonance. Moreover, the mechanical resonant frequency is very stable and can be controlled over several orders of magnitude by precisely cutting the quartz sheet into specific shapes and dimensions. Typical values of the fundamental tone resonant frequency are from low amounts of kHz to about 50 MHz. For higher frequencies, the physical dimensions of the crystal become too small and higher-order resonant tones are used instead. Crystal with a fundamental resonance at 30 MHz can also be used at 60 MHz, 90 MHz, 120 MHz, and sometimes even at 150 MHz.

The resonant frequency stability of ordinary crystals at room temperature is in the order of about one part in a million (1 ppm); an order of magnitude improvement in stability can be achieved if the crystal is mounted inside a temperature-controlled oven. By using special technologies, the upper achievable limit of frequency stability for crystals is about 0.1–1 ppb, i.e., less than one part in a billion. To put it in perspective, 0.1 ppb is equivalent to a ratio of 10^{10}, which is equivalent to approximately 1 s in 300 years. Not surprisingly, the main application of piezoelectric crystals in electronics is to serve as timing references for clock signals. Although the manufacturing process

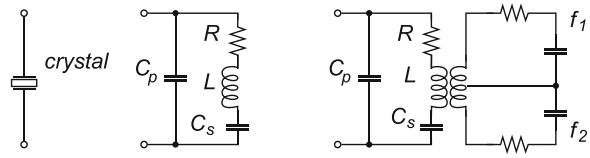


Fig. 8.9 The symbol for a quartz crystal (*left*); a basic electrical model (*centre*); and a model for dual-frequency overtone operation (*right*)

of crystals is not difficult by modern standards, in practice it is common that the crystals are manufactured to precisely match several "standard" reference frequencies that are commonly used in wired and wireless communications. Of course, modern circuits operate at frequencies far higher than the above-mentioned 150 MHz overtone. In addition, the crystal's resonant frequency is fixed by its physical dimensions (i.e., the resonant frequency is precise but is not tunable); by itself, crystal produces very tiny currents meaning that it always needs some active buffering circuit that improves its driving capability. Various frequencies are derived from the crystal reference frequency by means of a closed loop circuit known as a "phase-locked loop" (PLL). In Chap. 10, we study this very important circuit topology in more detail.

Accurate behavioural modelling of piezoelectric crystals requires that a set of differential equations describing both mechanical and electrical properties is solved, which is usually done numerically using modern multiphysics simulators. In our work, however, we are concerned only about the electrical properties of the crystals, which may be described in terms of a passive RLC electrical network (Fig. 8.9). A typical simple electrical model is based on a serial RLC branch in parallel with a small capacitor, Fig. 8.9 (centre); more complex models include the overtone modes of crystal operation. For instance, typical values used in this model for a 1.6 MHz crystal are as follows: $L = 250\,\text{mH}$, $r = 25\,\Omega$, therefore $Q = \omega L/r = 2{,}000$. Note that, for the frequency of operation, this is a very high value of the inductance, which is combined with a small internal resistance r to yield a high Q factor. However, it is the mechanical property of the crystal, rather than the electrical property, that gives this equivalent high inductance value. The remaining model parameters are $C = 0.04\,\text{pF}$ and $C_1 = 4\,\text{pF}$.

Because of the two parallel branches, there are two possible resonant frequencies: a series resonance that is determined only by the serial branch of the equivalent circuit in Fig. 8.9 and a parallel resonance determined by both the series branch and the parallel capacitance C_1 (Fig. 8.10). In order to demonstrate these two important resonant modes of a crystal, let us evaluate our example model. The series resonance is approximately

$$\omega_s = \frac{1}{\sqrt{LC}} = 1.000 \times 10^7 \ \text{rad/s} \quad \therefore \quad f_s = 1.592\,\text{MHz}. \tag{8.35}$$

In the parallel resonance mode, the capacitances are perceived as being in series around the resonant loop, which yields resonant frequency of

$$\omega_0 = \frac{1}{\sqrt{\dfrac{LC_1C}{(C_1+C)}}} = 1.005 \times 10^7 \ \text{rad/s} \quad \therefore \quad f_0 = 1.599\,\text{MHz}. \tag{8.36}$$

Note that these frequencies are very close to one another, about 0.5% apart in this example. This narrow separation is widened when the crystal is used in an LC resonant tank to increase its Q factor and that clearly determines whether the crystal is used in series or in parallel mode. Aside from this important property of dual resonant frequency, we also need to examine how the impedance of a piezoelectric element behaves at frequencies in proximity to these two resonant frequencies, Fig. 8.10 (right). At frequencies below the serial resonance ω_s the crystal reactance is dominated by

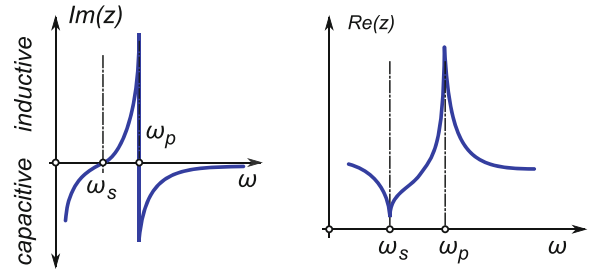


Fig. 8.10 Impedance of a crystal showing series and parallel resonances

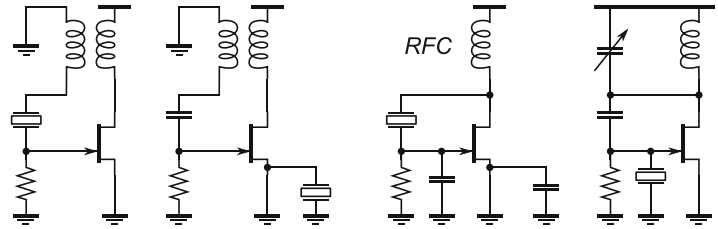


Fig. 8.11 Crystal-controlled oscillators: in series (i.e., low impedance) mode (*left*) and in parallel (i.e., high impedance) mode (*right*)

the large serial capacitance, i.e., it is negative. At the serial resonant frequency ω_s the reactance is zero (i.e., $Z_L = Z_C$) and the overall impedance is at its minimum close to zero, i.e., $Z = r$. Between the two resonant frequencies, the reactance is inductive and tends to infinity (in reality, a very high value), while the overall impedance follows the trend. At the parallel resonant frequency, ω_0 the overall impedance is at its maximum. Above the parallel resonant frequency, the reactance is again negative and the overall impedance decreases. It is very important to recognize that serial resonance is associated with minimum overall impedance and parallel resonance is associated with very high impedance; this determines how crystals are used in oscillator circuits.

Many different oscillator circuit arrangements use crystals. However, the general rule is that the low impedance mode (i.e., at series resonance) is used when the crystal is connected in series with the other elements (the two diagrams on the left of Fig. 8.11) and the high impedance mode is used in parallel with other circuit elements (the two diagrams on the right of Fig. 8.11). In other words, aside from controlling the oscillator's resonant frequency, insertion of a crystal into an oscillator should cause minimum interference with its internal voltages and currents.

8.7 Voltage-Controlled Oscillators

The ability to generate a single-tone, sinusoidal waveform with precisely controlled frequency is of vital importance for wireless communication systems and much engineering effort has gone into designing various forms of oscillator. However, communication systems require more than just one specific value of the frequency. For instance, every radio and TV receiver is capable of receiving signals from more than one transmitting station. As we already know, in order to select the desired station, we must tune the receiver to the particular frequency associated with the station. Another station uses another frequency. If we were only able to design an oscillator circuit capable of delivering a single frequency, we would either need to carry one radio receiver unit for each station to which we would like to listen or our receivers would be very bulky and complicated indeed. Obviously, that is not the case; we invented frequency-tunable oscillators whose output frequency depends on a control variable, either voltage or current, i.e., $\omega_0 = f(V_{ctrl}, I_{ctrl})$. Tunable oscillators are a key component of PLL circuits (Chap. 10). The resonant frequency of an LC resonator is determined by

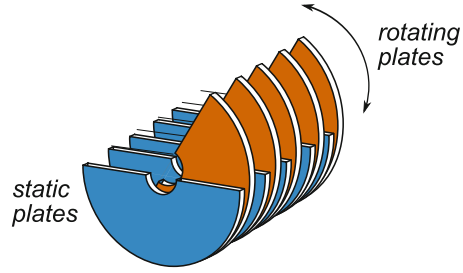


Fig. 8.12 Rotary air-gap variable capacitor used on radios for tuning the RF stage and the local oscillator. The overlapping capacitor area varies with the angle of rotation. Each pair of static and rotating plates makes one capacitor—there are five pairs in this image

component values of the inductor L and the capacitor C that are used in the resonator tank. Therefore, frequency tunability (or simply, "tunability") is achieved by varying the value of the inductance, the value of the capacitance or both. In principle, there are two possible ways of implementing a variable capacitor or inductor.

The "discrete" method simply means that a bank of serial or parallel components is connected and each component is independently switched in or out of the network. Obviously, this method is feasible only with a finite number of components and switches, hence it can deliver only a discrete set of component values. If a relatively fine change of the capacitive or inductive value is achieved with each switching step, then it is sometimes referred to as a "quasi-continuous" method.

A truly "continuous" method means that a component is capable of physically changing its value smoothly in response to the control variable. For instance, a rotary variable capacitor (Fig. 8.12) is created by mechanically controlling the overlapping area between two plates of a capacitor. According to (4.14) the capacitance of a plate capacitor is a linear function of the capacitor's surface area S, which means the overlapping area between the two plates.

Both methods of creating variable components (discrete and continuous) are used in practice. From a practical perspective, it is much easier to design and manufacture tunable capacitors than tunable inductors. Indeed, for over 100 years, the rotary variable capacitor was used almost exclusively for continuous tuning of LC resonators in commercial radio receivers. As you have already noticed, this kind of capacitor is very bulky and long ago it became the largest component by far inside a radio receiver. This implies that the mechanical rotary capacitor is not suitable for miniaturization and higher frequencies. This is further limited because tunability is achieved by manual control of a knob. Although miniature versions of fundamentally the same design, the trimmer capacitor, are still in use for semi-permanent tuning of radio receiver sections, modern high-frequency (HF) oscillators are based on a semiconductor device known as a "varicap diode" (*varactor*). At the same time, design of miniature variable inductors is still in the research domain, with some progress being made mostly due to advances in micro-electro-mechanical system (MEMS) technologies.

Capacitance C_D of a reverse biased p–n junction is a nonlinear function of the applied junction voltage V_D as

$$C_\mathrm{D} = \frac{C_0}{\left(1 - \frac{V_\mathrm{D}}{\phi}\right)^{\alpha}} \quad \therefore \quad C_\mathrm{D} \approx \frac{C_0}{\sqrt{1 + \frac{|V_\mathrm{D}|}{0.5}}}, \tag{8.37}$$

where C_0 is the diode capacitance at zero bias $V_\mathrm{D} = 0$, $\phi \approx 0.5$ is the contact potential of the p–n junction, and α is either ½ for idealized abrupt p–n junctions or ⅓ if the p–n junction is approximated by a linear function. As a first approximation, $\alpha = 1/2$ is assumed. Even though (8.37) is a nonlinear function $C_\mathrm{D} = f(V_\mathrm{D})$, it is important to note that the capacitance is electronically

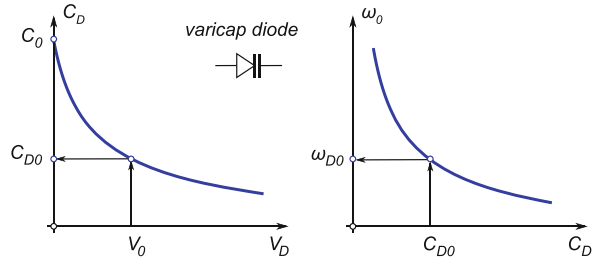


Fig. 8.13 Varicap diode functional dependencies of capacitance and its reverse biasing voltage V_D (*left*); resonant frequency dependence of an LC resonator and varicap capacitance (*right*)

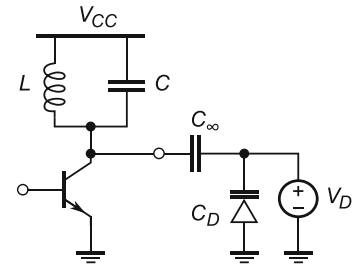


Fig. 8.14 LC resonator tuning by means of a varicap control voltage V_D

controlled by its biasing voltage, Fig. 8.13 (left), which is then used to control the resonance of an LC resonator, Fig. 8.13 (right). Hence, electronic control of LC tank resonant frequency $\omega_0 = f(V_D)$ is achieved. In addition, the p–n junction capacitance C_D is relatively small and is manufactured using IC technologies. Hence it is suitable for applications in HF integrated *voltage-controlled oscillators* (VCO).

A simplified schematic diagram (Fig. 8.14) of an electronically tunable LC resonator shows a varicap capacitor C_D and its DC biasing voltage V_D connected to an LC resonator tank through a large capacitor. The role of the C_∞ capacitor is to decouple the DC voltage V_D that keeps varicap C_D reverse biased without interfering with the biasing of the BJT current source. From an AC perspective, the varicap C_D is connected in parallel with the resonator capacitor C, hence the resonating frequency is calculated as

$$\omega_0 = \frac{1}{\sqrt{L(C + C_D(V_D))}} \qquad \therefore \qquad \omega_0 = f(V_D). \tag{8.38}$$

Now that we understand how a varicap diode is used in an LC resonator tank to control its resonant frequency, let us take a look at a simplified schematic diagram (Fig. 8.15) of a VCO circuit that incorporates, for instance, a tapped C bottom-grounded passive feedback network (see Fig. 8.6 (right)) with a CB amplifier (see Fig. 7.6). In order to implement electronic control of its resonant frequency, a varicap diode is added[1] in the LC resonating loop, which is perceived by the loop as being in series with the C_1 and C_2 resonator capacitors. The CB amplifier is biased through an RFC inductor that provides DC connection and, at the same time, AC decoupling from the power supply line. In textbooks, an oscillator configuration that uses a tapped C, bottom-grounded feedback network is usually referred to as a "Clapp oscillator".

In the first approximation, i.e., ignoring parasitic elements, the resonant frequency of the oscillating current is approximately set by the passive feedback network only. Capacitors C_1 and C_2 are in series, hence their equivalent capacitance C_S is

[1] For simplicity, varicap DC biasing is not shown.

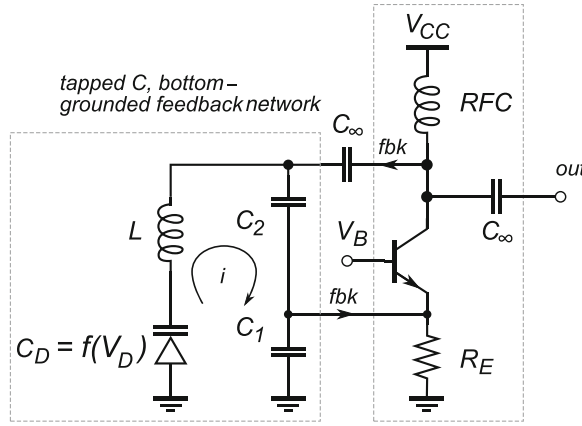


Fig. 8.15 A VCO that uses a CB amplifier in the forward path and a tapped C, bottom-grounded network in the feedback path

$$C_S = \frac{C_1 C_2}{C_1 + C_2}. \tag{8.39}$$

It, in turn, is in series with the varicap, hence the total equivalent capacitance C in the LC loop is written as

$$C = \frac{C_D C_S}{C_D + C_S} \quad \therefore \quad \omega_0 = \frac{1}{\sqrt{LC}}. \tag{8.40}$$

For all practical purposes, we can look at a VCO as a voltage-to-frequency converter. That being the case, we need to find out how sensitive the output frequency is, relative to the change of varicap biasing voltage V_D. Indeed, this sensitivity is one of the most important parameters of a VCO; it is known as the "frequency deviation constant". Mathematically speaking, the frequency deviation constant is determined by derivative of the frequency with respect to the varicap biasing voltage, i.e.

$$k = \frac{\mathrm{d}\omega}{\mathrm{d}V_D} = \frac{\mathrm{d}\omega}{\mathrm{d}C_D}\frac{\mathrm{d}C_D}{\mathrm{d}V_D}, \tag{8.41}$$

where, the two derivative terms are derived separately. After substituting (8.37), the second derivative term is[2]

$$\begin{aligned} \frac{\mathrm{d}C_D}{\mathrm{d}V_D} &= \frac{\mathrm{d}}{\mathrm{d}V_D}\left(\frac{C_0}{\sqrt{1+2V_D}}\right) \\ &= -\frac{C_0}{(1+2V_D)\sqrt{1+2V_D}} = -\frac{C_D}{1+2V_D} \quad \therefore \quad = -\frac{C_{D0}}{1+2V_0} \end{aligned} \tag{8.42}$$

after varicap diode capacitance C_D is used at a specific biasing voltage V_0. The first derivative term in (8.41), after substituting (8.40) and rearranging the terms and setting a specific biasing voltage V_0, is written as

[2]We keep in mind that the absolute value of biasing voltage is $|V_D| = V_D$.

$$\frac{\mathrm{d}\omega}{\mathrm{d}C_\mathrm{D}} = -\frac{\omega_0}{2C_{\mathrm{D0}}\left(1+\frac{C_{\mathrm{D0}}}{C_\mathrm{S}}\right)}. \tag{8.43}$$

It is handy to express (8.43) in terms of the ratio n of varicap capacitance relative to the series capacitance C_S, i.e., as $C_{\mathrm{D0}} = n\,C_\mathrm{S}$ and rewrite it as

$$\frac{\mathrm{d}\omega}{\mathrm{d}C_\mathrm{D}} = -\frac{\omega_0}{2\,(1+n)\,C_{\mathrm{D0}}} \tag{8.44}$$

then, after substituting (8.42) and (8.44) back into (8.41), we write

$$\begin{aligned} k &= \frac{\mathrm{d}\omega}{\mathrm{d}C_\mathrm{D}}\,\frac{\mathrm{d}C_\mathrm{D}}{\mathrm{d}V_\mathrm{D}} = \left[-\frac{C_{\mathrm{D0}}}{1+2V_0}\right]\left[-\frac{\omega_0}{2\,(1+n)\,C_{\mathrm{D0}}}\right] \\ &= -\frac{\omega_0}{2\,(1+n)\,(1+2V_0)}. \end{aligned} \tag{8.45}$$

The last expression is very useful for estimating the voltage-to-frequency conversion factor for a Clapp oscillator.

Example 8.3. For a given LC Clapp oscillator whose resonant frequency is set to $\omega_0 = 2\pi\,10\,\mathrm{MHz}$ by varicap diode voltage of $V_0 = 6\,\mathrm{V}$. The ratio of varicap capacitance to the serial capacitance in the resonator tank is $n = 0.1$. Estimate the frequency deviation constant of this oscillator.

Solution 8.3. A straight implementation of (8.45) results in

$$k = -\frac{\omega_0}{2\,(1+n)\,(1+2V_0)} = -\frac{10\,\mathrm{MHz}}{2\,(1+0.1)\,(1+2\times 6\,\mathrm{V})} = 349.650\frac{\mathrm{kHz}}{\mathrm{V}}.$$

We conclude that, for each volt of change in varicap bias, the oscillator resonant frequency moves about 350 kHz, which means that, for the given power supply of 6 V, we can count on no more than approximately 2 MHz or so of total frequency change, i.e., the expected frequency range approximately 9–11 MHz for this particular VCO design.

8.8 Time and Amplitude Jitter

A realistic periodic waveform produced by an oscillator suffers from a short-term frequency fluctuation that is referred to as "phase noise". At the same time, amplitude variations of the waveform are always present to a certain extent. For instance, an oscillator's sinusoidal output with phase variations $\theta(t)$ and amplitude variations $A(t)$ may be expressed as

$$v_\mathrm{s}(t) = V_\mathrm{s}[1+A(t)]\,\sin[\omega_c t + \theta(t)], \tag{8.46}$$

where V_s is the average peak voltage of the output signal. The phase and amplitude variations may be random or discrete or both. Individual spectral components at the oscillator output are referred to as "spurious responses"; noise, in this context, refers to the random variations of both frequency and phase. The engineering term for these random variations is "jitter". A very useful and practical method for estimating the amplitude and time jitter of a periodic signal with period T is to create an "eye plot" (Fig. 8.16). Instead of plotting a waveform from time zero to the last data point, the full data

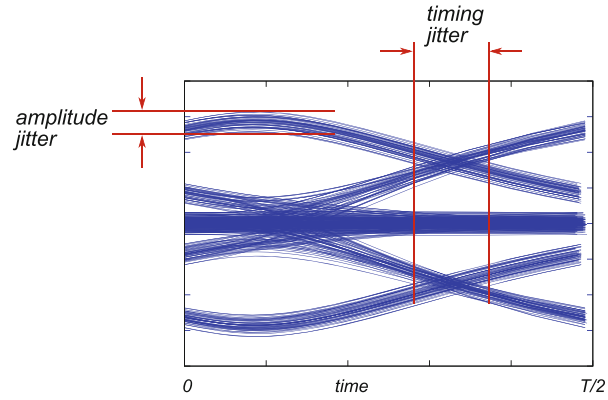


Fig. 8.16 Eye diagram of a long waveform showing time and amplitude jitter. The periodic waveform is split in sections each half a period long, i.e., $T/2$, and overlapped

vector is split into sections so that each section contains a set of data only half a period long. Then, all sections are overlapped (similar to a deck of playing cards). The newly created plot looks similar to an open eye if the amount of jitter is not too excessive. Amplitude jitter becomes clearly visible and easily measurable on the vertical axis, while timing jitter is easily measured on the horizontal axis around the cross-over point between rising and falling edges of the waveform (see Fig. 8.16). Commonly, timing jitter t_{jitter} is expressed relative to the waveform period T, e.g. $t_{\text{jitter}} = T/8$. Almost all modern oscilloscopes have a built-in eye diagram function, which makes it extremely easy for the user to create the plot in real time.

Detailed statistical analysis of phase noise is the subject of more advanced courses in communication theory, hence it is omitted in this text.

8.9 Summary

Basic techniques for the analysis of general oscillator circuits are presented in this chapter in which we have learned about closed loop feedback networks that are used to generate sinusoidal waveforms. Circuit conditions that lead into steady oscillations are specified in terms of the loop gain and the phase shift. Because oscillators are, in general, large-signal systems, small-signal techniques are used only to estimate the initial conditions that are required to establish steady oscillations. Nonlinear numerical analysis techniques are required for detailed circuit design. Because four types of passive RLC network are used in most typical oscillators, we accepted the open loop design methodology where the forward active path, consisting of an amplifier, is designed to compensate for the gain of the passive RLC feedback network. At the same time, the feedback network is responsible for setting up the correct phase shift around the loop. VCOs were introduced as very important for practical radio communication systems.

Problems

8.1. Derive an expression for loop gain for the general case of a loop consisting of a forward path amplifier with gain A and a feedback circuit path with gain.

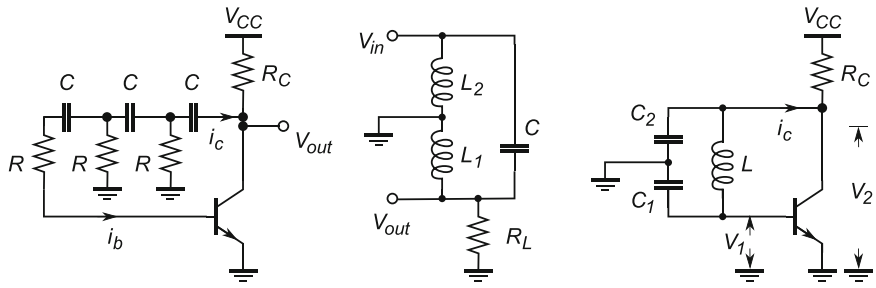


Fig. 8.17 Simplified schematic of a phase oscillator for Problems 8.2, 8.3, and 8.7

8.2. One of several phase oscillators, Fig. 8.17 (left), is based on a CE amplifier and three RC stages in the feedback loop. Derive an expression for the minimal transistor gain factor $\beta_{\min}$ (not to be confused with the feedback loop parameter) and resonant frequency ω, under the following assumptions:

- The transistor's output resistance r_o is infinite.
- All capacitors have the same value.
- All resistors have the same value and the transistor's base resistance r_b is absorbed into the left-most resistor R.
- Details of the biasing network are not shown.
- All elements are ideal (i.e., ignore the small emitter resistance r_e and base collector capacitance C_{BC}).

Then, calculate values for the resistors R and capacitors C, if $R_C = 10\,\text{k}\Omega$.

8.3. Estimate the resonant frequency ω_0 of an oscillator whose feedback network is shown in Fig. 8.17 (centre), if $L_1 = 0.5\,\mu\text{H}$, $L_2 = 1.5\,\mu\text{H}$, and $C = 126.65\,\text{pF}$.

8.4. Using the same data as in Problem 8.3, estimate the feedback network's gain factor β.

8.5. Using the same data as in Problem 8.3, estimate the effective resistance R_{eff} that this feedback network presents to the output of the oscillator's amplifier whose input impedance is $R_{\text{in}} = 10\,\text{k}\Omega$. The Q factor of the effective inductor L_{eff} is $Q = 50$.

8.6. Repeat Problems 8.3 to 8.5 for the other three types of feedback network shown in the textbook.

8.7. For the circuit shown Fig. 8.17 (right), derive expressions for: (a) the resonant frequency ω and (b) g_m of a BJT. Use these two formulas to calculate the resonant frequency and g_m using the following data: $R_C = 10\,\text{k}\Omega$, BJT output resistance $r_c = 10\,\text{k}\Omega$, $L = 2\,\mu\text{H}$, $C_1 = C_2 = 253.30\,\text{pF}$, and $Q_L \to \infty$. Details of the biasing network are omitted, for simplicity.

Assuming finite Q_L, derive new equations for the resonant frequency ω and g_m. Using, for example, $Q_L = 50$, recalculate these two values and compare with the ideal case solutions.

8.8. For the Clapp oscillator shown in Fig. 8.15, calculate the oscillating frequency at: (a) zero bias of the varicap diode and (b) $V_D = -7\,\text{V}$.

Data: $L = 100\,\mu\text{H}$, $C_1 = C_2 = 300\,\text{pF}$, and $C_0 = 20\,\text{pF}$.

Chapter 9
Frequency Shifting

Abstract In this chapter, we focus on the mathematical operation of "frequency shifting" that is fundamental to wireless communication systems. Frequency shifting (or "translation") is complementary to the frequency tuning mechanism used in VCOs. However, as will be shown, it is a much broader concept with a much wider range of applications. As it turns out, mathematical multiplication of two sinusoidal waveforms with given frequencies results in waveforms that contain both higher and lower frequencies. This phenomenon is known as "frequency shifting", where the term "up-conversion" refers to the process of shifting a lower frequency tone to the upper frequency range (used in RF transmitters), while "down-conversion" refers to the frequency shifting from higher to lower frequency ranges (used in RF receivers). Hence, in a complete wireless communication system, the information-carrying signal is shifted in both directions.

9.1 Signal-Mixing Mechanism

An electronic circuit that can multiply two AC signals is called a *mixer*. A mixer in RF systems always refers to a circuit with a *nonlinear* component that, for two input single-tone signals ω_1 and ω_2, produces single-tone output signals that are the sum (i.e., $\omega_1 + \omega_2$) and the difference (i.e., $|\omega_1 - \omega_2|$)[1] of the input frequencies. Note that, in audio systems, operators refer to "mixing" two sound tracks, which is not mixing in the RF sense: the sum and difference of the input frequencies are not generated and no nonlinear component is involved in the circuit. Rather, it is a *linear addition* of two signals so that the two sound tracks are heard simultaneously. Symbolic representation of these two operations (see Fig. 9.1) illustrates the difference between linear addition and the mixing of two AC signals.

Because ideal LTI systems cannot possibly produce output signals with spectral components not present at the input, mixers must be either nonlinear or time-varying elements in order to provide the frequency translation. Historically, many devices (e.g. electrolytic cells, magnetic ribbons, brain tissue, and rusty scissors, in addition to more traditional devices such as vacuum tubes and transistors) have been used as the nonlinear elements, demonstrating that virtually any nonlinear device can be used as a mixer. Of course, some nonlinearities work better than others, so we focus only on practical RF mixer types.

[1] The absolute value is equivalent to a geometrical distance on the horizontal axis, i.e., $|\omega_1 - \omega_2| = |\omega_2 - \omega_1|$.

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8_9,

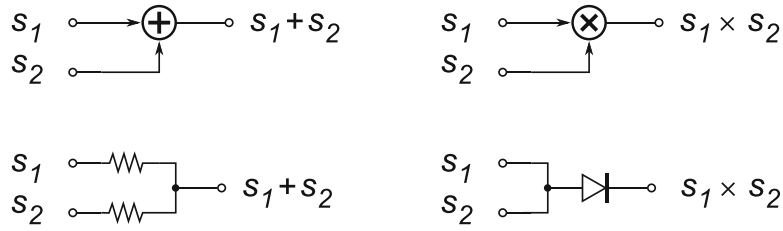


Fig. 9.1 Summing (*left*) and mixing (*right*) functions

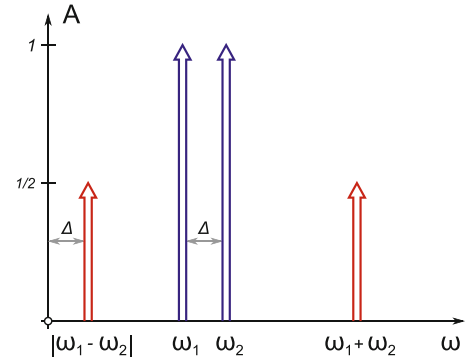


Fig. 9.2 Multiplication of ω_1 and ω_2 tones with amplitudes of 1 results in two new tones $\omega_1 + \omega_2$ and $|\omega_1 - \omega_2|$ with amplitudes of ½

At the core of all modern mixers is the multiplication of two sinusoidal signals in the time domain. The fundamental usefulness of the multiplication may be understood from the basic trigonometric identities[2]

$$\sin(\omega_1 t) \times \sin(\omega_2 t) = \frac{1}{2}\left[\cos(|\omega_1 - \omega_2|t) - \cos(\omega_1 + \omega_2)t\right], \tag{9.1}$$

$$\cos(\omega_1 t) \times \cos(\omega_2 t) = \frac{1}{2}\left[\cos(|\omega_1 - \omega_2|t) + \cos(\omega_1 + \omega_2)t\right], \tag{9.2}$$

which shows that the multiplication of two sinusoidal functions[3] results in two new sinusoids (Fig. 9.2). Note that the linear sum of the new sinusoidal functions on the right of (9.1) and (9.2) does not further affect arguments of the sinusoidal functions: one argument is the sum and the other argument is the difference of the original arguments, and those are the only two tones in the output spectrum.

It should be easy to see that, if the arguments define time-domain waveforms $(\omega_1 t)$ and $(\omega_2 t)$, a low-frequency signal multiplied by a high-frequency signal results in one signal higher than the higher signal, i.e., $(\omega_1 + \omega_2)$, and one signal that is lower than the lower signal, i.e., $|\omega_1 - \omega_2|$. Because a time-domain signal shape does not reveal information about its frequency content, it is much more practical to observe signals in the frequency domain (Fig. 9.2). The signal multiplication results (9.1) and (9.2) are a bit unfortunate because the goal was to shift one tone with the help of another and have one and only one tone as the result. Instead the multiplication operation delivers two tones—one up-converted and one down-converted. Hence, if this approach is to be used, one of the two new tones must be removed by additional processing.

[2] In strict mathematical syntax: $\sin x \cdot \sin y = 1/2\ [\cos(|x-y|) - \cos(x+y)]$ and $\cos x \cdot \cos y = 1/2\ [\cos(|x-y|) + \cos(x+y)]$.

[3] Keep in mind that sin and cos functions have the same shape, they are just phase-shifted versions of each other, i.e., they have different starting points.

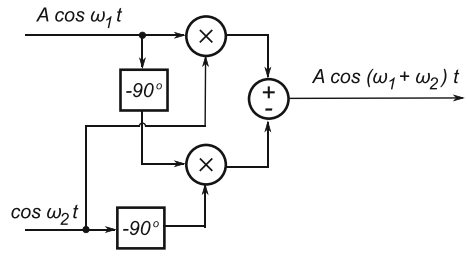


Fig. 9.3 Ideal frequency shifting circuit based on a literal implementation of (9.3)

An ideal theoretical model that delivers only one shifted tone is derived, for instance, by subtracting (9.1) from (9.2), which after substituting the arguments for frequency results and assuming that the product of the two amplitudes is A, we write as

$$A\cos[(\omega_1+\omega_2)t] = A[\cos(\omega_1 t)\cos(\omega_2 t) - \sin(\omega_1 t)\sin(\omega_2 t)], \tag{9.3}$$

which could be directly synthesized by a circuit whose block diagram is shown in Fig. 9.3. However, a practical implementation of circuit based on the block diagram in Fig. 9.3 is not trivial. It would be relatively straightforward to build the adder and multiplier blocks as IC devices. However, the wideband 90° phase shift circuit is where the problem arises. It is relatively easy to design a narrowband 90° phase shift at a given fixed frequency, but over a wide range of frequencies there is no convenient way of producing an exact phase shift over the whole range. For this reason, the scheme in Fig. 9.3, which in theory works over any range of frequencies, is seldom attempted. Most frequency shifting is done using a single multiplier, as already indicated, in less than perfect relationship between (9.1) and (9.2).

The unwanted component, which could be either $(\omega_1+\omega_2)$ or $(\omega_1-\omega_2)$ is removed by filtering. Most practical frequency translation circuits combine the processes of multiplying and filtering in order to achieve frequency translation. Hence, this methodology performs a quasi-multiplication because, besides the wanted product, one or more other frequency components are generated and must be removed by some form of filtering. In the following sections we take a look at some of the most commonly used mixer circuits.

9.2 Diode Mixers

A diode mixer is a very simple circuit (Fig. 9.4) that is useful up to very high frequencies. Because it works at almost any frequency, it is commonly used in measuring equipment which is expected to work over a range of frequencies. Two voltage single-tone signals

$$v_1 = V_1\cos(\omega_1 t), \tag{9.4}$$

$$v_2 = V_2\cos(\omega_2 t) \tag{9.5}$$

are first added and then passed through an ideal diode whose voltage–current function is given as

$$i_D = I_S\left\{\exp\left(\frac{v_D}{V_t}\right) - 1\right\}, \tag{9.6}$$

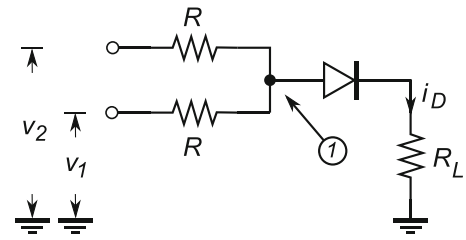


Fig. 9.4 Simplified schematic diagram of a diode mixer

which is the nonlinear element that is required for frequency shifting. In the following analysis, for the simplicity, we assume a small diode current and ignore the voltage drop across the loading resistor R_L. That is, the diode voltage V_D is approximately equal to the voltage at node ①, i.e., $V_D \approx V(1)$.

Two equal resistors R serve as a linear voltage adder. Because of their voltage-dividing property, the voltage at node ① is half the sum of the two inputs, i.e.,

$$v_D = v_1 = \frac{1}{2}(v_1 + v_2) = \frac{1}{2}[V_1 \cdot \cos(\omega_1 t) + V_2 \cdot \cos(\omega_2 t)]. \tag{9.7}$$

Following the signal path after node ①, the diode voltage v_D is converted into current i_D. The diode voltage is assumed to be small (implying that $v_D < V_t$ so that the higher-order terms in (9.9) are approximately zero), which allows the exponential term in (9.6) to be expanded into the Taylor series around the diode's biasing point, where series expansion for an exponential function is well known as

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots, \tag{9.8}$$

hence, after substitution of $x = v_D/V_t$ into (9.8) and application of the exponential term in (9.6), we write

$$i_D = I_S \left\{ \left[1 + \frac{v_D}{V_t} + \frac{1}{2}\left(\frac{v_D}{V_t}\right)^2 + \frac{1}{6}\left(\frac{v_D}{V_t}\right)^3 + \frac{1}{24}\left(\frac{v_D}{V_t}\right)^4 + \cdots \right] - 1 \right\}. \tag{9.9}$$

We now examine each of the terms on the right of (9.9) separately and find out about the signal's total spectrum (note that 1 and −1 cancel). Obviously, the exact series includes an infinite number of terms. In the first approximation, because the assumption is that the signal is small, the third and higher-orders terms can be ignored (they are smaller and smaller numbers divided by larger and larger numbers). After substituting (9.7) into (9.9), we focus on the first two terms:

- The linear term:

$$\frac{v_D}{V_t} = \frac{1}{2V_t}[V_1 \cdot \cos(\omega_1 t) + V_2 \cdot \cos(\omega_2 t)] = f(\omega_1, \omega_2). \tag{9.10}$$

  We conclude that the linear term of the series expansion has a frequency spectrum that is equal to the original spectrum of the signal v_D, i.e., ω_1 and ω_2. We already had that spectrum, hence this term is not much use.
- The square term:

$$\frac{1}{2}\left(\frac{v_D}{V_t}\right)^2 = \frac{1}{2V_t^2}\left\{\frac{1}{2}[V_1 \cdot \cos(\omega_1 t) + V_2 \cdot \cos(\omega_2 t)]\right\}^2$$

$$= \frac{1}{8V_t^2}\left[V_1^2\cos^2(\omega_1 t) + 2V_1V_2\cos(\omega_1 t)\cos(\omega_2 t) + V_2^2\cos^2(\omega_2 t)\right]$$

$$= \frac{1}{8V_t^2}\left[V_1^2\,\frac{1}{2}(1+\cos(2\omega_1 t)) + V_1V_2(\cos(|\omega_1-\omega_2|t) + \cos((\omega_1+\omega_2)t)\right.$$

$$\left. + V_2^2\,\frac{1}{2}(1+\cos(2\omega_2 t))\right], \tag{9.11}$$

which states that the output frequency spectrum due to the second (nonlinear) term contains

$$\frac{1}{2}\left(\frac{v_D}{V_t}\right)^2 = f\left[(\omega_1-\omega_2), 2\omega_1, 2\omega_2, (\omega_1+\omega_2)\right]. \tag{9.12}$$

In other words, aside from the desired (ω_1–ω_2) and (ω_1+ω_2) terms, there are additional tones ($2\omega_1$ and $2\omega_2$) present that are not part of the ideal multiplication operation.

Therefore, a diode is the simplest active device that serves as a mixer and works over a wide range of frequencies. In addition, a more specific small-signal condition should be stated as $V_1V_2 < V_t^2$. The additional, unwanted tones are usually filtered out with an LC resonator.

In conclusion, using a diode as a nonlinear element for the purpose of multiplying two single-tone signals does produce the desired theoretical tones the $(\omega_1 - \omega_2)$ and $(\omega_1 + \omega_2)$. However, it also produces tones that are not part of the ideal solution (i.e., ω_1, ω_2, $2\omega_1$, $2\omega_2$, ...). In addition, if the higher-order harmonics in (9.9) are not neglected, many more tones are observed in the output frequency spectrum. Therefore, for good performance this mixer is restricted to quite low input signal levels. Because all tones that are not needed must be filtered out afterwards, a diode is a simple but very inefficient multiplying element.

9.3 Transistor Mixers

Active mixers are based on nonlinear exponential functions of BJTs and metal-oxide semiconductor field-effect transistors (MOSFETs). If two voltage signals

$$v_1 = V_1\cos(\omega_1 t), \tag{9.13}$$

$$v_2 = V_2\cos(\omega_2 t) \tag{9.14}$$

are added and then applied to the gate of an ideal BJT Q_1 (Fig. 9.5 (left)) or v_1 is applied to the gate of Q_2 and v_2 is applied to the emitter node by means of a 1:1 ratio transformer (Fig. 9.5 (right)), then the two signals are mixed. Assuming ideal transistors with a current gain of β, the two variants of BJT mixer are very similar, therefore the following two equations are written by inspection:

$$V_{BE}(Q_1) = \frac{1}{2}(v_1 + v_2), \tag{9.15}$$

$$V_{BE}(Q_2) = v_1 - v_2. \tag{9.16}$$

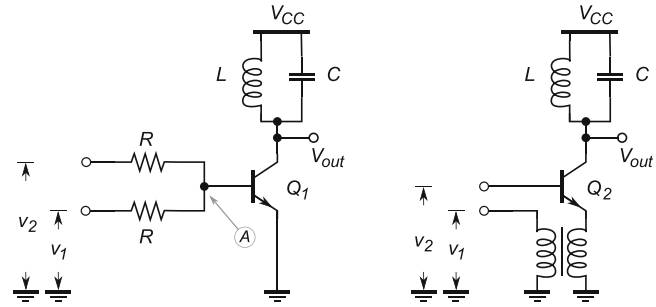


Fig. 9.5 Simplified schematic of two versions of BJT mixers

The relationship between the collector current I_C of a BJT versus the base-emitter voltage V_{BE} is the same as for a forward-biased diode,

$$i_C = I_S \left\{ \exp\left(\frac{v_{BE}}{V_t}\right) - 1 \right\}, \tag{9.17}$$

which is to say that the expression for the square term of interest in the output frequency spectrum of the circuit in Fig. 9.5 (left) is similar to the one for a diode, with the addition of the β factor:

$$I_{Cs} = \beta I_S \frac{V_1 V_2}{8 V_t^2} \left[\cos(|\omega_1 - \omega_2|t) + \cos((\omega_1 + \omega_2)t)\right].$$

The corresponding expression for the circuit in Fig. 9.5 (right) is only slightly different. It is important to note that, because of the β factor, a BJT mixer has much better efficiency than a simple diode mixer and it is possible even to have a "conversion gain". That means that it is possible for the output tone (usually the low-frequency tone, $|\omega_1 - \omega_2|$) to have more power than the input signal. On the other hand, a BJT mixer has the same limitation as the diode in terms of the input signal amplitude relative to the V_t voltage. Both circuits in Fig. 9.5 use an LC resonator in the collector branch that is tuned to either of the two tones of interest, i.e., either to $|\omega_1 - \omega_2|$ or to $\omega_1 + \omega_2$, and they filter out all unwanted tones in the frequency spectrum.

9.4 JFET Mixers

In RF amplifiers, it is common practice to replace BJTs with JFETs. JFET gate current is much less than the base current and has higher transconductance than a MOSFET transistor. Therefore it is often used in the front end of low-noise, high-input-impedance RF amplifiers.

Two input voltage signals

$$v_1 = V_1 \cos(\omega_1 t), \tag{9.18}$$

$$v_2 = V_2 \cos(\omega_2 t) \tag{9.19}$$

are applied to JFET transistors J_1 and J_2 (used instead of Q_1 and Q_2 in the topology of in Fig. 9.5). We use a procedure similar to that in the previous sections; the main difference is that the current–voltage characteristics between the drain current I_D and the gate-source voltage v_{GS} of a JFET obey the following relationship

$$I_D = I_{DSS} \left(1 - \frac{v_{GS}}{V_p}\right)^2, \tag{9.20}$$

where I_{DSS} is the JFET saturation drain current, V_{GS} is the gate-source voltage, and V_p is the pinch-off voltage. In the JFET case, there is no exponential term, which makes the derivation a bit simpler. Therefore, a straightforward expansion of (9.20) leads to

$$I_D = I_{DSS}\left[1 - 2\frac{v_{GS}}{V_p} + \frac{v_{GS}^2}{V_p^2}\right]. \tag{9.21}$$

By focusing only on the nonlinear terms in (9.21), the square term is

$$\begin{aligned} I_D &\sim -I_{DSS}\frac{1}{4}\frac{[V_1 \cdot \cos(\omega_1 t) + V_2 \cdot \cos(\omega_2 t)]^2}{V_p^2} \\ &\sim -I_{DSS}\frac{V_1 V_2}{2V_p^2}\left[\cos(|\omega_1 - \omega_2|t) + \cos((\omega_1 + \omega_2)t)\right], \end{aligned} \tag{9.22}$$

where (9.22) focuses only at the cos product term from the previous step. It is interesting to note that, because there was only a second-order term in (9.21) and no higher-order terms, there was no need to apply power series expansion as in the cases of the diode and BJT. That is, there is no strict limitation to the amplitudes of V_1 and V_2, as long as the JFET is not cut off or becomes forward biased. Again, similar to the BJT circuits from Sect. 9.3, the LC resonator simultaneously filters out all harmonics except the desired one. JFETs are commonly used in RF mixer applications because of their tolerance for high signal levels and good conversion efficiency.

9.5 Dual-Gate MOSFET Mixers

We have already learned (Sect. 7.7.1.4 and Fig. 7.32) that a cascode amplifier configuration is very useful in RF applications because of its high output impedance and its resilience to the Miller effect. Putting the two transistors into a single package and creating a dual-gate device was a natural development. In this section, we learn about the application of dual-gate transistors to the design of RF mixers. The additional important property of a dual-gate transistor is that the two devices are almost perfectly "matched"—they are "twins" in respect of their electrical properties. Consequently, two independent input signals applied to the two gates control the drain current at the same time and equally well.

Two input voltage signals

$$v_1 = V_{DC1} + V_1 \sin(\omega_1 t), \tag{9.23}$$

$$v_2 = V_{DC2} + V_2 \sin(\omega_2 t) \tag{9.24}$$

are applied to a dual-gate FET transistor that is used as a mixer (Fig. 9.6). In a standard cascode configuration, transistor M_1 is set as the CS amplifier for the v_1 signal, while M_2 serves as the CG current buffer. Therefore, assuming $v_2 = \text{const} = V_{DC2}$, we write equations for the M_1 transistor in saturation (ignoring its nonlinear effects) as

$$I_D = k(V_{GS} - V_{th})^2 = k(v_1 - V_{th})^2,$$

$$\therefore$$

$$g_m{}' \equiv \frac{dI_D}{dV_{GS}} = 2k(v_1 - V_{th}) = 2k[V_1 \sin(\omega_1 t) - V_{th}], \tag{9.25}$$

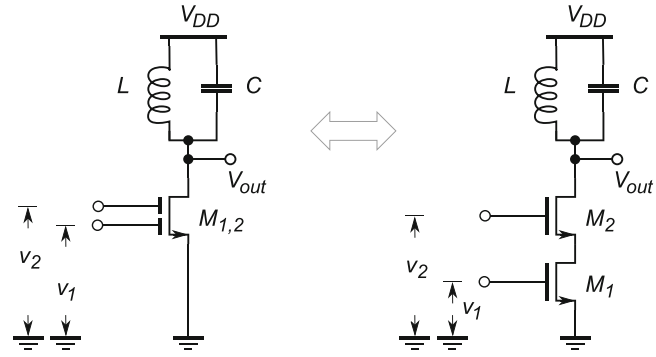


Fig. 9.6 Simplified schematic of a dual-gate FET mixer and its equivalent circuit diagram, where M_1 and M_2 are assumed to be identical

where, $V_{DC1,2}$ are constant biasing DC voltages, $k = (\mu_n C_{ox} W)/(2L)$, V_{th} is the MOS threshold voltage, and $g_m{}'$ is the circuit's overall g_m under the condition that the gate of M_2 is at its small signal ground. Drain current I_D passes through the current buffer M_2 with no loss (i.e., the two transistors have the same drain current), which is followed by LC load. Therefore, the voltage across the LC load at resonance is approximately bounded by its dynamic resistance R_D multiplied by the drain current, which is the same as the output voltage V_{out} relative to V_{DD}. In other words,

$$V_{out} = I_D R_D = g_m{}' v_1 R_D = g_m{}' R_D V_1 \sin(\omega_1 t). \tag{9.26}$$

That is to say, when the gate of M_2 is at the small signal ground, in respect of signal v_1 the circuit works as a CS cascoded amplifier. However, when the gate of M_2 is used as the input terminal for the second signal, for example when the dual-gate MOS mixer signal v_2 comes from a local oscillator (LO), the common drain current is additionally controlled by v_2. Variation of the drain current because of variation of the v_2 voltage is manifested as a change in the circuit's overall g_m, as

$$\begin{aligned} I_D &= k[(V_{DC2} + V_{GS2}) - V_{th}]^2, \\ &\therefore \\ g_m &\equiv \frac{dI_D}{d(V_{DC2} + V_{GS2})} = 2k[(V_{DC2} + V_{GS2}) - V_{th}] \\ &= 2kV_{DC2} + 2k(V_2 \sin(\omega_2 t) - V_{th}) \\ &\sim g_m{}' + g_{m\Delta} \sin(\omega_2 t), \end{aligned} \tag{9.27}$$

where, $g_m{}'$ is part of the circuit's g_m due to v_1 (9.25), while $g_{m\Delta}$ is a variation of the circuit's g_m due to v_2, whose common mode is at (i.e., it is centred around) V_{DC2}. It is important to note that V_{DC2} is not constant any more and that this arrangement works because the two transistors are identical with the same drain current.

After replacing $g_m{}'$ in (9.26) with g_m from (9.27), it follows that

$$\begin{aligned} V_{out} &= [g_m{}' + g_{m\Delta} \sin(\omega_2 t)] R_D V_1 \sin(\omega_1 t) \\ &= g_m{}' R_D V_1 \sin(\omega_1 t) + g_{m\Delta} R_D V_1 \sin(\omega_2 t) \sin(\omega_1 t) \\ &\sim g_{m\Delta} R_D V_1 \left[\cos(|\omega_1 - \omega_2| t) + \cos((\omega_1 + \omega_2) t)\right], \end{aligned} \tag{9.28}$$

where (9.28) focuses only on the cos product term and the LC resonator is tuned to either of the two desired tones and filters out all the other harmonics.

In conclusion, a dual-gate FET mixer is commonly used in the design of an RF mixer to multiply the incoming RF signal with the LO. Setting the appropriate LO frequency, the RF signal is then precisely shifted in the frequency domain, i.e., either "down-converted" or "up-converted". In addition, from (9.28), it becomes obvious that the v_2 signal amplitude should be as large as possible so that the $g_{m\Delta}$ term is maximized, which is one of the advantages of this circuit.

9.6 Image Frequency

A less obvious, but very important, consequence of signal multiplication (9.1) and (9.2) is that for any given frequency ω_1 there are two separate single tones ω_2 and ω_3, that produce exactly the same $|\omega_1 - \omega_2| = |\omega_1 - \omega_3|$ tones (Fig. 9.7).[4] At the same time, the higher frequency tones $(\omega_1 + \omega_2)$ and $(\omega_1 + \omega_3)$ are easily distinguished. This phenomenon of dual frequencies entering the mixer and producing the same output tone is so important in wireless communication systems that it is commonly referred to as "image frequency", or a "ghost image". For instance, if the original intent was to multiply frequencies ω_1 and ω_2, then signal ω_3 would be declared a ghost image. Similarly, ω_2 would be declared a ghost image if the original intent was to do frequency shifting of the ω_1 and ω_3 tones.

9.6.1 Image Rejection

The problem of ghost images is very real and is dealt with by using the following two methods. First, the transmitting frequencies are licensed and assigned at a national level—some frequencies are forbidden for communications because they would represent ghost images to their dual frequencies, which are already in use. Second, radio receiver front-end electronics are required to be able to suppress image frequencies relative to the desired tones by a specified amount.

The front end of a receiver consists of one or more parallel tuned resonant circuits that act as a bandpass filter centred around the resonant frequency. To really appreciate the need for high Q resonators, let us try to find out what happens when the incoming signal frequency is not exactly the same as the LC tank resonant frequency. In other words, at what distance $\Delta\omega$ is the image frequency found from the resonant frequency suppressed by the Q factor of the front-end LC tank?

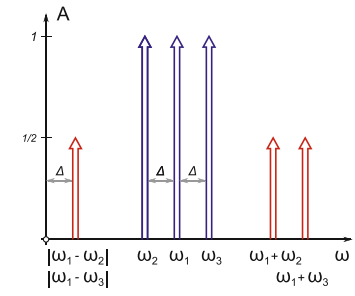


Fig. 9.7 Frequency domain diagram of the relative positions of the main and image frequencies

[4] Although all three signals ω_1, ω_2, and ω_3 are shown as having the same amplitude, in general it does not have to be the case.

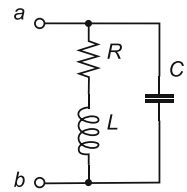


Fig. 9.8 Realistic LC tank model

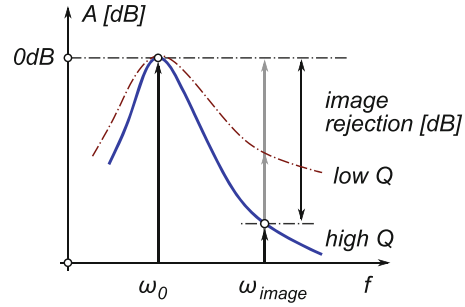


Fig. 9.9 Graphical representation of an image-rejection measure

9.6.2 LC Tank Admittance

In order to estimate the amount of signal suppression for a tone that is not centred at the resonant LC frequency we need to evaluate the frequency dependence of a realistic LC tank model (Fig. 9.8). A realistic inductor is modelled as a serial combination of an ideal inductor L and an ideal resistor R that embodies the total wire resistance in the resonant loop (including the inductor's DC resistance), while the capacitor is still assumed to be ideal. We already derived an expression, (5.61), that is repeated here for convenience

$$|Y| = Y_0\sqrt{1+(\delta Q)^2}, \tag{9.29}$$

where ω is close to ω_0 (i.e., it is less than one decade away), so that $\omega/\omega_0 \approx 1$ and

$$\delta = \frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}. \tag{9.30}$$

The graph in Fig. 9.9 demonstrates the relationship between the resonant tone ω_0 and the nearby tone ω_{image} for cases of high and low Q resonators. Lower Q means a wider bandwidth, which means that signal (ω_{image}) is more attenuated than the resonant tone normalized to 0 dB level. Higher Q provides a narrower bandwidth, which means that the same image signal is even more attenuated, hence, it is more suppressed relative to the desired tone at ω_0.

A straightforward implementation of (9.29) to calculate the output voltage at the image frequency, assuming the signal current I_S, yields

$$|V_0| = \frac{|I_S|}{|Y|} = \frac{I_S}{Y_0\sqrt{1+(\delta Q)^2}}, \tag{9.31}$$

which, at resonance, reduces to $V_0(\omega_0) = I_S/Y_0$. Hence, the relative voltage amplitude between the non-resonant and resonant voltages is given by

$$A_r \triangleq \frac{|V_0|}{V_0(\omega_0)} = \frac{1}{\sqrt{1+(\delta Q)^2}} \tag{9.32}$$

for a single tuned circuit. If several tuned circuits are included and isolated by amplifiers, then the overall response is given by the product $A_r(tot) = A_{r1}A_{r2}\ldots A_{rn}$. Further improvement of the image rejection is usually achieved by using a double-conversion radio receiver architecture.

Example 9.1. An AM broadcast receiver is tuned to 500 kHz with an LC resonator whose $Q = 50$. Calculate the signal rejection in dB of unwanted signal being transmitted at 1,430 kHz.

Solution 9.1. Because the front-end LC tank is tuned at $f_0 = 500$ kHz, a radio transmitter emitting at that frequency is the desired signal. It is then straightforward to implement (9.29) as follows

$$\delta Q = \left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right) Q = \left(\frac{1430}{500} - \frac{500}{1430}\right) 50 = 126,$$

which, after substituting in (9.32) and approximating $\sqrt{1+126^2} \approx 126$, yields

$$A_r = 20 \log \frac{1}{126} = -42\,\text{dB}.$$

Therefore, if a second radio station is transmitting at 1,430 kHz, its signal is received as 126 times weaker than the signal from the desired radio station. Using two tuned amplifiers would double the selectivity and further suppress the image signal down to −84 dB.

9.7 Summary

In this section, we have learned about the frequency-shifting mechanism that is fundamental to radio communication systems. The underlying mathematics is based on multiplying two sinusoidal forms, while the practical realization is based on passing the two single tones through a nonlinear element. Because of imperfect multiplication in realistic systems based on diodes, BJT or FET devices and additional filtering is required to remove unwanted tones. As a side product of the multiplication operation, we learned about the existence of the ghost image and its influence on the wanted signals. Image suppression is an important requirement for the front end of radio receivers, hence we worked out a formula for estimating the image (or any other side signal, for that matter) suppression relative to the desired tone that is aligned with the LC resonant frequency.

Problems

9.1. For this problem, use these four single-tone signals:
$S_1 = V_1 \sin(\omega_1 t)$, $S_2 = V_2 \sin(\omega_2 t)$, $S_3 = \cos(|\omega_1 - \omega_2|t)$, and $S_4 = \cos((\omega_1 + \omega_2)t)$.
Assuming $f_1 = 1$ MHz, $f_2 = 20$ MHz, $V_1 = 2$ V, and $V_2 = 3$ V, do the following:

(a) Find an expression for $S = S_1 S_2$. Using graphing software of your choice, plot S, $(V_1V_2)S_1$, and $-(V_1V_2)S_1$ in the same window. Observe the relative relationships between these signals.
(b) Plot $S_o = {}^1/_2 \cdot (V_1V_2) \cdot (S_3 - S_4)$. What can you conclude?

9.2. Starting from $S_1 = \sin(2\pi \times 10\,\text{MHz} \times t)$, find two other single tones that could be used to generate a single tone at $f = 1$ kHz. Explain the process and the result.

9.3. A large number of radio stations transmit their programs at various carrier frequencies. A radio receiver is tuned to receive an AM wave transmitted at a carrier frequency of $f_{RF} = 980\,\text{kHz}$. The LO inside the receiver is set at $f_{LO} = 1{,}435\,\text{kHz}$. Find:

(a) The frequencies coming out of the receiver's mixer.
(b) Which frequency is IF.
(c) The frequency of a radio station which would represent an image frequency to the radio station.
(d) The frequency graph of the frequencies involved.

9.4. A tuned RF amplifier has an LC tank with $Q = 20$ and it is tuned at RF frequency f_0. Estimate the attenuation of the image signal, if the image frequency is 10% higher than the RF signal.

Chapter 10
Phase-Locked Loops

Abstract Arguably the most important circuit in modern electronics is a PLL. A PLL is embedded into virtually every piece of modern electronic equipment: computers, wireless communication systems, and test equipment, to name a few. What is more, each of the subsystem blocks inside these systems may contain their own PLL sub-circuit manufactured in an IC technology. With the advent of IC technologies, embedding PLL circuits into higher-level systems became a routine matter, where the embedding is done "on-chip", i.e on the same silicon die as the rest of the analog, digital, or mixed-signal ICs. Modern RF transceivers that include PLL circuits are realized as a single IC, implying that cost, size and power consumption are very important design parameters that contributed to the widespread use of PLLs. In this chapter, we introduce only the terminology, the basic principles of operation, and the applications of PLL; detailed analysis and study of PLL are far beyond the scope of this book.

10.1 PLL Operational Principles

A PLL is a closed loop feedback circuit that employs an external stable reference signal to control both frequency and phase of its internal VCO circuit. In other words, PLL tries to exactly follow in time the leading reference signal. In general, two signals may be either completely unrelated to each other, for instance the frequencies of a fly's flapping wings and the phases of the moon, or they may be in some fixed relationship between the two frequencies and phases. It is important to realize that two related periodic signals may be synchronized in phase, in frequency, or in both phase and frequency (Fig. 10.1). A PLL has a goal to synchronize both phase and frequency of its VCO wave with the reference wave.

In a typical application, an external stable crystal oscillator is used as a reference that provides an input signal to the PLL circuit, which is then accurately replicated in the loop, both in terms of its frequency and phase. The PLL continuously keeps self-adjusting its internal VCO wave to match the reference signal over long periods of time, i.e., it stays "locked" to the reference. However, the role of the PLL does not stop there, a PLL is capable of providing multiple copies of the same reference signal to various circuits in the system, while at the same time it provides a driving capability that is far beyond what the crystal itself can provide. In fact, a crystal is barely capable of serving as a signal source to a very light load that is physically located close by, i.e., it cannot provide high current drive. What is more, a PLL can create not only multiple copies of the same frequency reference, but also signal copies that are at any fractional multiple of the reference frequency. Considering that stable crystal references are manufactured in a limited range of frequencies, that particular capability of a PLL by itself would be sufficient to justify the use of PLL circuits. In its basic form, when used

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8_10,

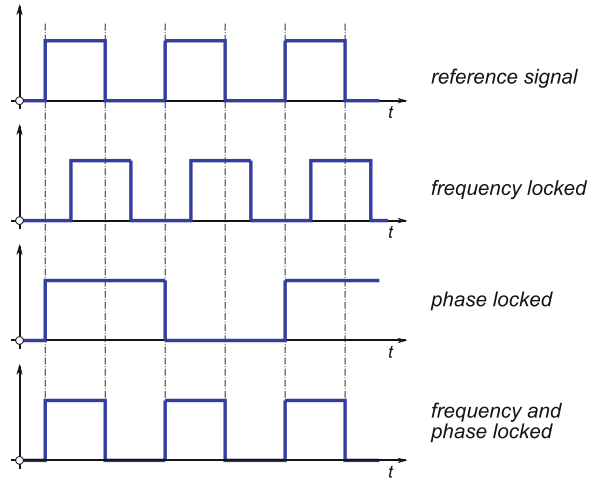


Fig. 10.1 Time-domain examples of signals "locked" (i.e. synchronized) in frequency, phase, and frequency and phase to the reference signal

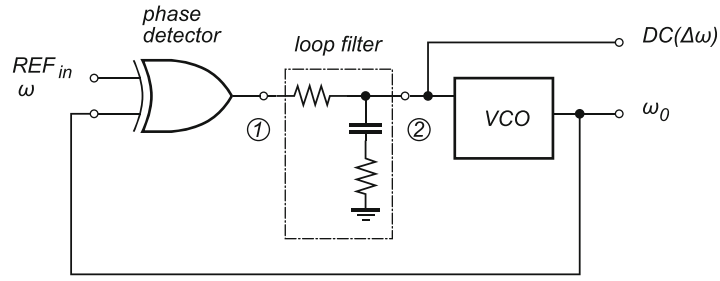


Fig. 10.2 Basic PLL circuit block diagram

to generate on-chip frequency references, the PLL is said to work as a *clock synthesis unit* (CSU). However, as we will se in the following sections, a PLL is an extremely versatile circuit that is used for number of different applications.

In its basic form, a PLL consists of three main blocks (Fig. 10.2): a phase detector (PD), a loop filter, and a VCO. The three blocks are arranged into a feedback loop that operates as follows. The PD compares the frequency and phase of the local VCO's wave with the frequency and phase of the incoming reference signal. Technically, it should be called a "phase comparator" because it measures the phase difference θ_d and frequency difference ω_d between the two waves. The output waveform of the PD at node ① takes the shape of a pulse-width modulated (PWM) stream whose average value is a function of the phase and frequency differences. This averaged, i.e. DC, PWM stream voltage at node ② is applied as a control voltage to the VCO, as we discussed in Sect. 8.7.

10.2 Linear Model of PLL

In order to make PLL analysis manageable within the scope of this book, we adopt linear models of its basic blocks, even though the linearities are valid only for a limited range. In addition, even though the input v_{in} and output v_{out} signals of a PLL are not always pure sinusoidal waveforms (in fact, more often they are pulse streams), for the purposes of our analysis we assume that they are sinusoidal waves, i.e.

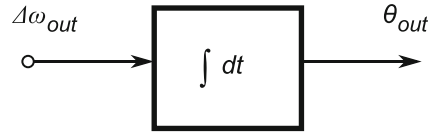


Fig. 10.3 Phase detector block diagram of (10.5)

$$v_{\text{in}} = \sin(\omega_{\text{in}} t + \theta_{\text{in}}), \tag{10.1}$$

$$v_{\text{out}} = \sin(\omega_{\text{in}} t + \theta_{\text{out}}), \tag{10.2}$$

where, ω_{in} is the constant input reference frequency, ω_{out} is the locked output frequency, and θ_{in} and θ_{out} are the reference and output phases. By definition, frequency is a derivative of phase, hence the output frequency is also written as

$$\omega_{\text{out}} \equiv \frac{\text{d}}{\text{d}t}(\omega_{\text{in}} t + \theta_{\text{out}}) = \omega_{\text{in}} + \frac{\text{d}}{\text{d}t}\theta_{\text{out}}, \tag{10.3}$$

$$\therefore$$

$$\Delta\omega_{\text{out}} = \omega_{\text{out}} - \omega_{\text{in}} = \frac{\text{d}}{\text{d}t}\theta_{\text{out}}, \tag{10.4}$$

$$\therefore$$

$$\theta_{\text{out}} = \int \Delta\omega_{\text{out}}\,\text{d}t. \tag{10.5}$$

Therefore, (10.5) implies that PLL requires a basic integrator block that converts the change of the VCO output frequency $\Delta\omega_{\text{out}}$ into the output phase θ_{out} (see Fig. 10.3).

These are the basic equations of the PLL linear model. We are now going to take a closer look at its basic blocks.

10.2.1 Phase Detector Model

The main purpose of a PD is to measure the difference between the input phase θ_{in} and the VCO phase θ_{out} (i.e., the PLL output phase) and to convert it into a proportional DC voltage level. Therefore, we define the instantaneous *phase error* θ_{err} as

$$\theta_{\text{err}} = \theta_{\text{in}} - \theta_{\text{out}}. \tag{10.6}$$

The measured phase error information is used to generate a proportional DC error voltage v_{err} as

$$v_{\text{err}} = K_{\text{PD}}\,\theta_{\text{err}}, \tag{10.7}$$

where K_{PD} is the proportionality constant. The DC error voltage is to be converted into frequency by the VCO circuit. However, VCOs are designed to oscillate at some nominal, i.e free-running, frequency ω_0 that is usually set in the middle of its frequency tuning range, i.e.

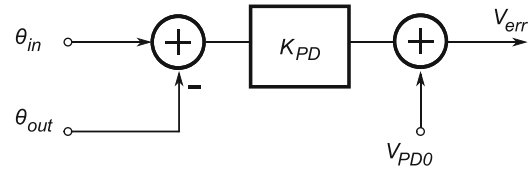


Fig. 10.4 Phase detector block diagram of (10.9)

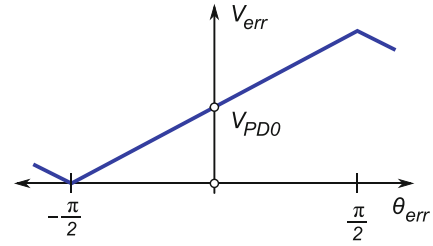


Fig. 10.5 Phase detector characteristic

$$\omega_0 = \frac{\omega_{\max} + \omega_{\min}}{2}. \tag{10.8}$$

For practical reasons, the free-running frequency ω_0 is associated with a positive DC control voltage $v_{c0} \neq 0$. The choice of the free-running control voltage simplifies the biasing setup of the VCO and enables its frequency to shift both above and below the free-running frequency, while the control voltage always stays positive. Therefore, if the phase error equals zero then the DC error voltage equals $\theta_{\text{err}} = 0 \Rightarrow v_{\text{err}} = v_{\text{PD0}}$, which is reflected by a shift of function (10.7) up by $v_{\text{PD0}} \neq 0$, i.e.

$$v_{\text{err}} = K_{\text{PD}}\,\theta_{\text{err}} + v_{\text{PD0}} = V_{\text{c}}, \tag{10.9}$$

which is synthesized by the signal flow shown in Fig. 10.4. By definition, the PD characteristic, i.e., v_{err} against θ_{err}, is periodic with the basic range between $-\pi/2 \leq \theta_{\text{err}} \leq \pi/2$ (see Fig. 10.5).

10.2.2 VCO Model

The main purpose of a VCO is to convert a DC control voltage level V_{c} into proportional frequency ω_{out} of the output wave. It is desirable to have a linear relationship between the DC control level and the output frequency, although it is not a mandatory requirement (Fig. 10.6). Every VCO is designed for a certain range of control voltage $\omega_{\max}$ to $\omega_{\min}$ that corresponds to the range of the control voltage V_{cmax} to V_{cmin}. Outside that region, behaviour of the VCO is not considered valid.

We define output frequency deviation $\Delta\omega_{\text{o}}$ as the difference between the output frequency ω_{out} and the input reference frequency ω_{in}, i.e.,

$$\Delta\omega_{\text{o}} \equiv \omega_{\text{out}} - \omega_{\text{in}}. \tag{10.10}$$

The slope of the VCO characteristic around the lock frequency is called the "VCO gain" K_{VCO}, therefore it is defined as

$$K_{\text{VCO}}\frac{\omega_{\text{out}}}{V_{\text{c}}}, \tag{10.11}$$

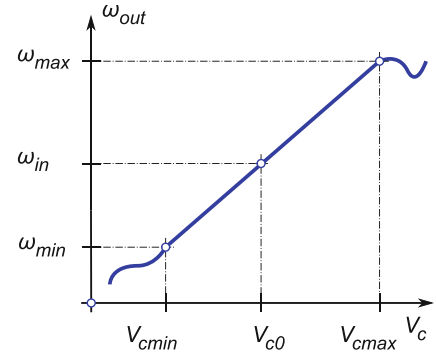


Fig. 10.6 VCO characteristic

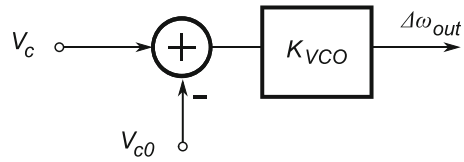


Fig. 10.7 VCO block diagram of (10.12)

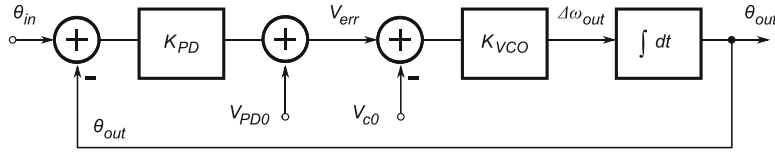


Fig. 10.8 Block diagram of the basic PLL linear model

that is, the frequency deviation (10.10) is approximated as

$$\Delta\omega_{\mathrm{o}} = K_{\mathrm{VCO}}\,(V_{\mathrm{c}} - V_{\mathrm{c0}}), \tag{10.12}$$

which is directly synthesized by the block diagram in Fig. 10.7.

When the output frequency equals the input frequency, the output frequency deviation equals zero and the PLL is said to be "in lock". In other words, the output frequency deviation is a measure of how far the output frequency is from the input reference frequency. Therefore, by definition, when $\Delta\omega_{\mathrm{o}} = 0$, the control voltage equals V_{c0}, Fig. 10.6.

10.2.3 PLL Bandwidth

We are now ready to put these three linear models (the PD, the VCO, and the integrator) together in a simple PLL loop (Fig. 10.8). In its simplest form, the PLL loop starts at the input of PD where the two waves are compared in terms of their respective phases. The PD compares the instantaneous phase θ_{out} of the VCO wave with the reference phase θ_{in} of the input wave and generates the error signal in accordance with (10.9). If, for instance, the reference wave phase is ahead of the VCO phase, a positive error signal V_{err} is generated. The error signal is added to the voltage control signal V_{c0} which forces the VCO to increase its current frequency by $\Delta\omega_{\mathrm{out}}$ relative to the free-running frequency, i.e., the VCO is forced to "speed up" and try to "catch" the reference signal. The frequency increase is converted into the new phase by the integrator block, which is again evaluated at the PD input terminals. After the first cycle is finished, the two input waves are closer to each other; the loop

forces the VCO to keep increasing its frequency until the two phases are matched and the PD detector generates $V_{\text{err}} = 0$. That is interpreted by the loop as the "stop" command, i.e., the VCO is expected to hold its frequency unchanged. If the reference phase was behind ("lagging") the VCO phase, then the sign of V_{err} is negative and the VCO is forced to "slow down" until the two phases match.

Once the PD generates the zero error signal, this whole process is continuously repeated; the PLL stays locked to the reference signal and only wiggles the minimal amount around the value of the input reference phase.

For the sake of argument, let us now allow the input reference signal to change its frequency, without being concerned about why and how it happened. We should be able to understand by now that the PLL loop always tries to follow the input reference signal; it does not know that it is running after a "moving target". The decision on whether to "speed up" or "slow down" is made at the end of each cycle.

Which brings us to the main point of this discussion: the change of the reference frequency is an AC signal by itself. That is, the PLL loop must have sufficient loop bandwidth to enable the PLL to accurately follow the input frequency variations and to keep the lock. For instance, if the integrator block takes too much time to process the current error information about the $\Delta\omega_{\text{out}}$, then the PLL would never be able to accurately follow the input frequency changes. Indeed, the PLL would only sluggishly lock to some average value of the AC signal because that is what an integrator does. From control theory, we already know that making the integrator much faster than the input AC signal only makes the loop unstable, i.e., the "speed up" and "slow down" variations amount to self oscillations.

The careful reader should have recognized by now that the feedback loop in Fig. 10.8 is equivalent to a general feedback loop with the single forward gain block $G(s)$ defined as

$$G(s) = \frac{K_{\text{PD}} K_{\text{VCO}}}{s} = \frac{K_0}{s}, \tag{10.13}$$

where $K_0 = K_{\text{PD}} K_{\text{VCO}}$, after the integrator block is replaced with $1/s = 1/\text{j}\omega$ function. Then, by inspection, we write the phase transfer function as

$$\frac{\theta_{\text{out}}}{\theta_{\text{in}}} = \frac{G(s)}{1 + G(s)} = \frac{K}{K + s}. \tag{10.14}$$

Therefore, for small variations, the loop bandwidth is determined by the frequency point when the absolute value of the loop gain $|G(\text{j}\omega)| = 1$, i.e., at the 3 dB point, hence

$$1 = \frac{K_{\text{PD}} K_{\text{VCO}}}{\omega_{3\,\text{dB}}},$$

$$\therefore$$

$$\omega_{3\,\text{dB}} = K_{\text{PD}} K_{\text{VCO}}, \tag{10.15}$$

which give us a hint of how to control the loop bandwidth. We have to alter the loop gain without changing the parameters of the PD and VCO blocks: their gains are the result of circuit topologies used in the design. The obvious solution is to add a simple voltage divider in the signal path and proportionally reduce the loop gain, which reduces the bandwidth as well. As designers, we have full freedom to determine the voltage divider resistive ratio, i.e., it means that (10.13) takes shape

$$G(s) = \frac{K_{\text{PD}} K_{\text{VCO}} K_{\text{R}}}{s}, \tag{10.16}$$

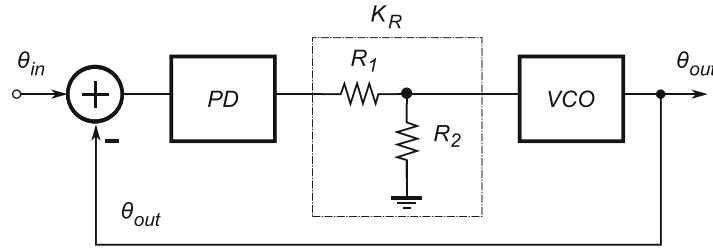


Fig. 10.9 Block diagram of PLL with added R_1, R_2 voltage divider for the purpose of reducing the loop bandwidth

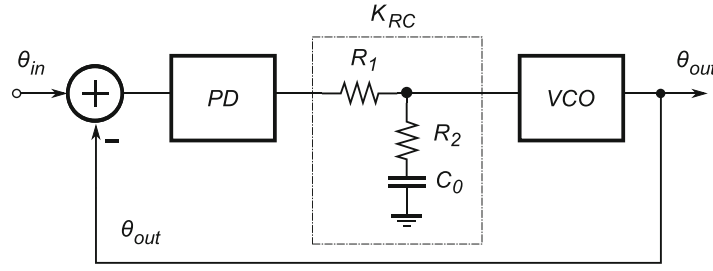


Fig. 10.10 Block diagram of PLL with added R_1, R_2, C_0 lead–lag loop filter for the purpose of controlling the loop bandwidth

which describes the topology in Fig. 10.9. A PLL with a simple voltage divider attenuator that is used to control the loop bandwidth is usually referred to as a "first-order PLL" because its transfer function (10.14) has only a first-order polynomial of s in the denominator. With the added resistive divider, the loop bandwidth becomes,

$$1 = \frac{K_{\mathrm{PD}} K_{\mathrm{VCO}} K_{\mathrm{R}}}{\omega_{3\,\mathrm{dB}}},$$

$$\therefore$$

$$\omega_{3\,\mathrm{dB}} = K_{\mathrm{PD}} K_{\mathrm{VCO}} K_{\mathrm{R}}, \tag{10.17}$$

where K_{R} is the resistive divider gain. Therefore, by careful design of the divider, we keep control of the PLL bandwidth.

10.2.4 The Loop Filter Model

The first-order PLL represents the simplest circuit topology that gives control of the loop bandwidth to the designer. Unfortunately, its main drawback is that it inherently reduces the DC gain and, therefore, reduces the VCO's control voltage. Consequently, in accordance with (10.12), the frequency tuning range is also proportionally reduced. And that is a problem.

Instead of using the simple voltage divider that indiscriminately reduces the loop DC gain, the most commonly used solution is to add a large capacitor and simply brake the newly introduced DC path through the R_2 resistor (see Fig. 10.10), which is commonly known as a "lead–lag" filter. A straightforward analysis of the lead–lag network yields its transfer function $F(s)$ as

$$F(s) = \frac{R_2}{R_1 + R_2} \frac{s + \frac{1}{R_1 + R_2}}{s + \frac{1}{R_2 C_0}}, \tag{10.18}$$

where, after substituting (10.18) in (10.16), we write the new expression for the forward path gain as

$$G(s) = \frac{K_{\text{PD}} K_{\text{VCO}} F(s)}{s}. \tag{10.19}$$

It is straightforward, although a bit involved, to write an expression for the loop gain transfer function as

$$\frac{\theta_{\text{out}}}{\theta_{\text{in}}} = \frac{Ks + \frac{1}{R_1+R_2}}{s^2 + \left(K + \frac{1}{R_2 C_0}\right) s + \frac{K}{R_1+R_2}}, \tag{10.20}$$

where $K = K_{\text{PD}} K_{\text{VCO}} F(s)$. Because the denominator of the transfer function consists of a second-order polynomial, this version of the PLL transfer function is commonly referred to as a "second-order PLL".

10.3 PLL Applications

There are several typical applications of PLL circuits in modern communication systems and new ones are still being invented. Most often, a PLL is used as a frequency synthesizer, a clock recovery unit, or a tracking filter. Its uses as a phase and frequency modulator/demodulator are mentioned in Sects. 11.4 and 12.3.3.

10.3.1 Frequency Synthesizers

Probably most obvious application of a PLL is in the design of a multiple-frequency synthesizer. The main motivation for this application comes from, for instance, the FM radio broadcasting system. In most countries, the frequency spectrum of 87.5 MHz to 108.0 MHz is used for this purpose and each radio station is required to fit into one of 101 channels, numbered from 200 (87.9 MHz) to 300 (107.9 MHz) in increments of 200 kHz. For successful signal reception, the synthesized frequency needs to be accurate to within 10 parts per million (ppm), which is to say that the accuracy of a crystal oscillator is required. Considering that the crystal itself is one of the bulkiest parts that is used in modern radio electronics, having 101 crystals in parallel would be a very impractical radio indeed.

Instead, a single crystal reference is used to provide a stable input reference to PLL that includes a "programmable divider" in its feedback path (see Fig. 10.11). Operation of the frequency synthesizer is based on a single 10 MHz crystal reference whose frequency is divided by 50, hence the 200 kHz reference is perceived by the PD. A tunable VCO is designed to generate a wave in the 87.9 MHz to 107.9 MHz frequency range as its output, which is then taken back through the feedback path into the programmable divider whose division ratio is $N = 439.5$ to 539.5. Consequently, any of the 101 discrete frequencies can be divided down to a 200 kHz wave, which is then compared with the 200 kHz reference wave. We note that the trick in this circuit is in the fractional programmable divider that reduces any of the 101 different frequency values to the one value, which enables the use of the single crystal reference.

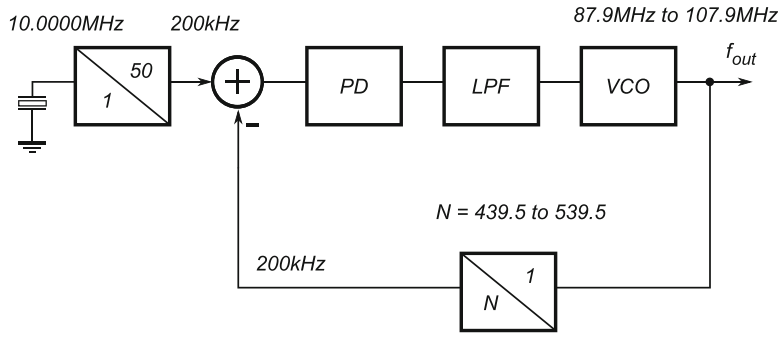


Fig. 10.11 Block diagram of a PLL frequency synthesizer with added programmable divider in the feedback path

10.3.2 Clock and Data Recovery Units (CRU)

Without going into design details, we note that in many communication systems there is a need to recover both the original clock frequency and the data by looking only into the received data stream. This process is called "clock and data recovery" and it is routinely done in disk players, floppy disk readers, and telephone and satellite data links. The basic PLL circuit is usually used for this purpose, with special attention paid to the design of the PD and the loop filter. Because the clock recovery process is data dependent (which for all practical purposes is a random process), the recovered clock suffers from increased timing jitter. Hence, the design of clock and data recovery units (CRU) presents a significant challenge to circuit designers.

10.3.3 Tracking Filters

Narrowband filters are usually designed for a given fixed centre frequency by using fixed value components. However, there are applications where the carrier frequency "drifts" in time, for instance due to the Doppler effect. If a fixed narrowband filter were used to track signals suffering from the Doppler effect, it would very quickly lose the signal because the carrier frequency would drift out of the filter's bandwidth. Instead, a PLL behaves as a very narrowband filter whose centre frequency moves with the carrier frequency of the input wave.

10.4 Summary

Although the linear model PLL is valid only over a very limited range and the treatment presented in this chapter is incomplete, it is intuitively simple and sufficient as a first introduction to this fascinating circuit. The basic terminology, functionality, and applications presented should be sufficient for the reader to develop an initial understanding of how and why PLL circuits fit into the topic of wireless communication systems. Because the detailed study of PLL circuits is far beyond the scope of this book, the reader is encouraged to consult the references and continue exploring the details of PLL topologies.

Problems

10.1. For the PLL model in Fig. 10.9, given that the gain of the PD is $K_{\mathrm{PD}} = 1.27\,\mathrm{V/rad}$ and the VCO gain is $P_{\mathrm{VCO}} = 2\,\mathrm{Mrad/s/V}$, the goal is to design a resistive divider to satisfy the bandwidth requirement of $\omega_{3\,\mathrm{dB}} = 0.73\,\mathrm{Mrad/s}$. Determine R_1 and R_2 values by using your engineering judgment.

10.2. The centre frequency of a PLL is set to 10 MHz. After power up, the minimum input signal frequency that causes the PLL to start following it is 9.9 MHz; after that, the PLL follows the input signal as long as it is within $10\,\mathrm{MHz} \pm 500\,\mathrm{kHz}$. Outside that range, the PLL loses the lock again. Estimate the lock range and the capture range of this PLL.

Chapter 11
Modulation

Abstract In the broad sense, the term "modulation" implies a change in time of a certain parameter. For instance, while listening to a steady single-tone signal with constant amplitude and frequency coming out of a speaker, we merely receive the simplest message that conveys information only about the existence of the signal source and nothing else. If the source is turned off, then we cannot even say if there is a signal source out there or not. For the purpose of transmitting a more sophisticated message, the communication system must use at least the simplest modulation scheme, based on time divisions, i.e., turning on and off the signal source. By listening to short and long beeps, we can decode complicated messages letter by letter. When you think about it, smoke signals are based on the same principle. As slow and inefficient as it is, Morse code does work and is used even today in special situations, for example in a very low SNR environment.

In this chapter, we study the main modulation techniques for wireless communications, which are based on the time variation of periodic electrical signals.

11.1 The Need for Modulation

The main purpose of a communication system is to transmit a message from one point in space to another. It would be very inefficient, to say the least, if we had a single communication system that is capable of transmitting only one message at a time. Just imagine if the whole world's phone system used a single metallic wire and users had to line up at the two ends for a chance to communicate with the other side. An evolutionary improvement is the development of a network with multiple communication points. Indeed, the phone system is based on the existence of a temporary physical wire connection between the sending and the receiving points, which, of course, necessitates the existence of switching circuits. Today, direct transmission of relatively LF signals, such as our voice, is routinely done through the phone network.

However, it is easy to recognize that the wire-based communication network is very expensive to build and maintain. That is mostly because the wire itself needs a supporting medium, in this case the Earth's surface, which presented huge technical challenges, for instance, when the intercontinental cables were laid at the bottom of the Atlantic and Pacific oceans.

A wireless transmission system does not have this issue, however its own problem immediately becomes visible in the case of direct transmission of audio signals. One commonly used antenna is

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8_11,

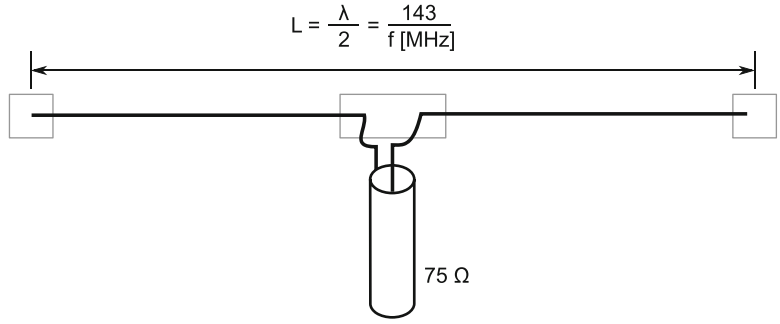


Fig. 11.1 Dipole antenna

known as a "half-wave dipole antenna",[1] a name derived from the requirement that the antenna wire length L is approximately equal to half the wavelength λ (Fig. 11.1), which is calculated as

$$L = \frac{1}{2}\lambda = \frac{1}{2}cT = \frac{1}{2}\frac{c}{f} \approx \frac{300 \times 10^{6\,\mathrm{m/s}}}{2}\frac{1}{f[1/\mathrm{s}]} = \frac{150}{f\,[\mathrm{MHz}]}\,[\mathrm{m}], \tag{11.1}$$

where λ is the wavelength of the incoming EM wave, T is its period, and c is the speed of light. Antenna designers usually fudge (11.1) by approximately 5%, so that the commonly cited rule of thumb for the length L of the wire intended to receive a signal at frequency f is written as

$$L = \frac{143}{f[\mathrm{MHz}]}\,[\mathrm{m}], \tag{11.2}$$

which, for a simple example of audio frequency $f = 1\,\mathrm{kHz} = 0.001\,\mathrm{MHz}$ leads to a required antenna $L = 143\,\mathrm{km}$ long. For all practical purposes, we already have that antenna in the form of the phone cables laid in trenches all around the world. Which is to say that direct radio transmission of audio signals is not practical. A straightforward solution to this problem is to apply the frequency shifting principle and move the audio signals higher in the frequency domain. For instance, if the signal frequency is $f = 10\,\mathrm{MHz}$ then we calculate, from (11.2), that the required antenna must be approximately $L = 143/10 = 14.3\,\mathrm{m}$, while for a signal $f = 1\,\mathrm{GHz}$, the required dipole antenna is $L = 0.143\,\mathrm{m}$ long. Obviously, the use of higher frequency signals results in more practical antenna sizes. A further study of EM wave propagation properties showed that transmission losses through various materials are dependent on the frequency. Hence, for the given transmission medium (air in this case), not all wavelengths travel the same distance for the same initial signal power, which means that the choice of operating frequency is very important in regard to how much energy is used for the transmission.

An efficient communication system needs to be capable of transmitting multiple transmissions at the same time. Considering that the audio bandwidth requires approximately 20 kHz, if the RF equipment is capable of working from, say, 1 GHz to 2 GHz, then the $(2-1)\,\mathrm{GHz} = 1\,\mathrm{GHz}$ frequency bandwidth can be viewed as a wide cable that consists of $1\,\mathrm{GHz}/20\,\mathrm{kHz} = 50\mathrm{E}3$ parallel "wires", i.e., 50,000 separate "channels", where each channel can carry one full audio signal. If each of the 50,000 audio sources is precisely frequency shifted and aligned next to each other within the 1 GHz bandwidth, then by means of frequency shifting and filtering, the wireless system is enabled to transmit multiple signals at the same time. In practical terms, a wirelessly transmitted signal consists

[1]For a detailed theory of quarter-wave and dipole antennas see, for example, *Antennas and Propagation for Wireless Communication Systems* by S. Saunders and A. Aragón-Zavala.

of two signals: a high-frequency signal that serves as the carrier and a low-frequency information signal that is somehow embedded into the carrier by the transmitter circuitry and de-embedded by the receiver.

We can now summarize the reasons why modulation is needed:

- To enable practical wireless transmission of audio signals.
- To enable power-efficient transmission that depends on the carrier frequency.
- To serve as the mechanism of embedding low-frequency information into the high-frequency carrier.

For a given periodic signal, a natural question is what exactly we can modulate. A general, time-domain, periodic signal $c(t)$ is described as

$$c(t) = C \sin(\omega t + \phi), \tag{11.3}$$

where C is its maximum amplitude, ω is its radial frequency, and ϕ is its initial phase. By inspection of (11.3), we conclude that there are three possible ways to embed information into the carrier:

- Vary the amplitude C in time so that $C(t)$ becomes equal to the time variation of the information signal (this method is called "amplitude modulation").
- Vary the frequency ω in time so that $\omega(t)$ becomes equal to the time variation of the information signal (this method is called "frequency modulation").
- Vary the phase ϕ in time so that $\phi(t)$ becomes equal to the time variation of the information signal (this method is called "phase modulation").

Although, theoretically any combination of the three parameters C, ω, and ϕ could be used to modulate the carrier $c(t)$, including modulating all three at the same time, in practice a communication system is designed to work with only one type of modulation. Hence, we only talk about either amplitude, frequency, or phase modulation (PM) systems. In the following sections, we take a closer look at each of these three main modulation methods.

11.2 Amplitude Modulation

Conceptually, amplitude modulation (AM) is the simplest form of carrier modulation. In this methodology, the amplitude of the carrier signal is made to replicate the shape of the baseband modulating signal, i.e., the information-carrying signal (for instance, the voice signal). While keeping in mind that a complicated signal consists of multiple single tones, we derive an analytical expression for the AM signal as described in this section.

The AM technique is based on the existence of two time-varying signals (assuming zero initial phase)

$$b(t) = B \sin \omega_{\mathrm{b}} t, \tag{11.4}$$

$$c(t) = C \sin \omega_{\mathrm{c}} t, \tag{11.5}$$

where $b(t)$ is the LF information (i.e., the modulating signal), B is its maximal amplitude, ω_{b} is its angular frequency, and $c(t)$ is the high-frequency carrier, C is its maximal amplitude, ω_{c} is its angular frequency.

The sum of the modulating signal $b(t)$ and the carrier's maximum amplitude C is called the "envelope wave" $e(t)$, which is described as

$$e(t) = C + b(t). \tag{11.6}$$

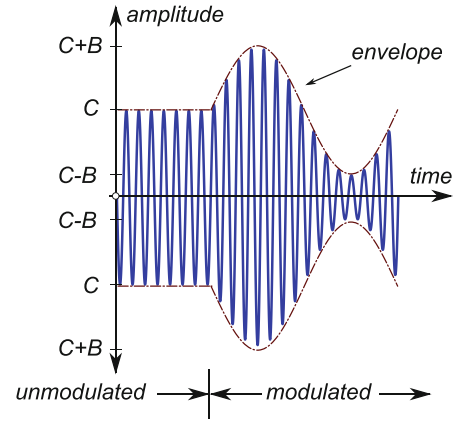


Fig. 11.2 Amplitude modulated signal, the unmodulated carrier, and the information signal in the form of an envelope

Note that for the unmodulated AM signal, i.e., $b(t) = 0$, the envelope $e(t)$ equals the carrier's maximum amplitude C and it is, therefore, constant (see Fig. 11.2).

An analytical expression for the modulated carrier amplitude $c_{\mathrm{AM}}(t)$ is derived by replacing the carrier's amplitude C in (11.5) the with expression for the envelope (11.6), which is the equivalent of saying that the carrier amplitude is modulated by the baseband signal, i.e.

$$\begin{aligned} c_{\mathrm{AM}}(t) &= e(t)\sin\omega_{\mathrm{c}}t \\ &= (C + B\sin\omega_{\mathrm{b}}t)\sin\omega_{\mathrm{c}}t \\ &= C\left(1 + \frac{B}{C}\sin\omega_{\mathrm{b}}t\right)\sin\omega_{\mathrm{c}}t \\ &= C(1 + m\sin\omega_{\mathrm{b}}t)\sin\omega_{\mathrm{c}}t \qquad (11.7) \\ &= \sin\omega_{\mathrm{c}}t + m\sin\omega_{\mathrm{b}}t\sin\omega_{\mathrm{c}}t \\ &= \sin\omega_{\mathrm{c}}t + \frac{m}{2}\left[\cos|\omega_{\mathrm{c}} - \omega_{\mathrm{b}}|\,t - \cos(\omega_{\mathrm{c}} + \omega_{\mathrm{b}})\,t\right], \qquad (11.8) \end{aligned}$$

where the modulation index is defined as $m = B/C$. Without losing in generality, after setting $C = 1$ in (11.7) everything afterwards is normalized to the carrier maximal amplitude.

The AM index m is an important communication parameter that shows the ratio of the baseband and the carrier maximum amplitudes (see Fig. 11.3). In the interests of efficient power transfer and high SNR, it is desirable to have the amplitude of the modulating signal as high as possible relative to the carrier's amplitude. If the carrier's maximal amplitude is greater than the modulating signal's amplitude, i.e., $m < 1$, then the embedded envelope is a faithful representation of the information (in this case, a clean sinusoidal shape). If the carrier's maximal amplitude is equal to the modulating signal's amplitude, i.e., $m = 1$, the embedded envelope is still a faithful copy of the information. However, in the case of modulation index $m > 1$ the envelope is not a faithful copy of the information. Keep in mind that the amplitude-modulated signal has two symmetrical envelopes, one positive and one negative, that carry the same information. As long as the two envelopes are kept separate and do not overlap (i.e., the positive envelope stays positive and the negative envelope stays negative), it is possible to recover the information from either of the two envelopes, Fig. 11.3 (left and centre). However, once the two envelopes overlap, Fig. 11.3 (right), the information is distorted because sections of the positive envelope cross over and become part of the negative envelope, and vice versa, which causes signal clipping. This is referred to as "over-modulation": neither the positive nor the

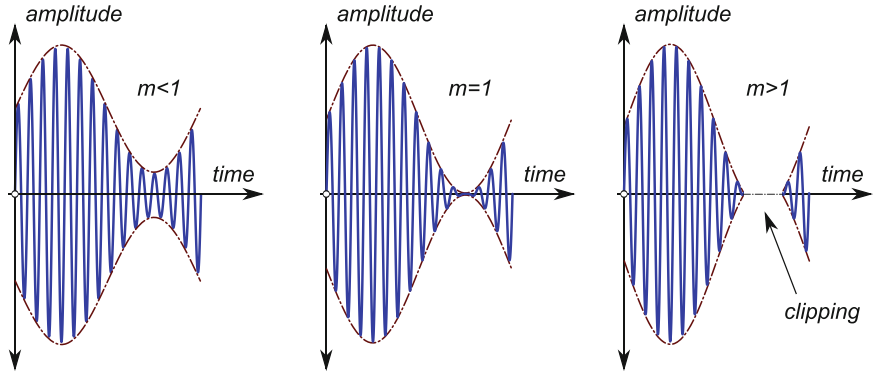


Fig. 11.3 Time domain plot of sinusoidal amplitude modulation (11.7) for three values of the AM modulating index. For $m \leq 1$, the modulating signal is correctly embedded in the envelope, however, for $m > 1$ signal clipping occurs and the sinusoidal envelope shape is distorted at the zero amplitude level

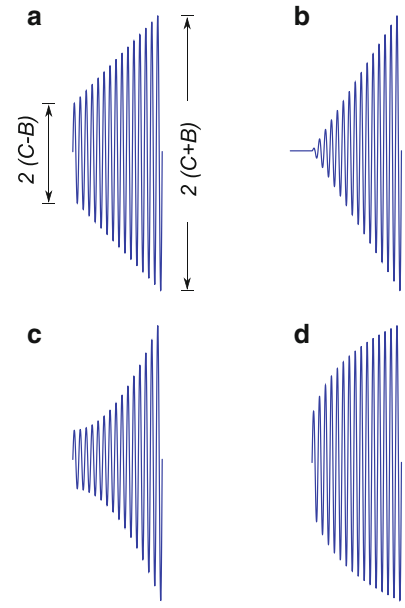


Fig. 11.4 Trapezoidal patterns of an AM signal: (**a**) for $m < 1$; (**b**) for $m > 1$, with the clipping section easily visible as the straight line tail; (**c**) for $m < 1$ and weak RF driver, i.e., the carrier signal is too strong; and (**d**) for $m < 1$ and nonlinear modulator, with visible nonlinear gain for high amplitudes

negative envelope looks like the original information (it looks like a clipped sinusoidal). Note that in case of over-modulation, (11.7) is not valid within the clipping region. In practical systems, the modulation index is held close to one for most of the time.

11.2.1 Trapezoidal Patterns and the Modulation Index

Except for the trivial case of a single-tone modulating signal (Fig. 11.3), observing amplitude-modulated signals in the time domain using an oscilloscope is cumbersome because it is difficult to synchronize the time sweep. Instead, non-periodic signals, for instance voice, are observed using the "trapezoidal method", which is usually used to plot Lissajous curves. In this method, the AM signal $c_{\mathrm{AM}}(t)$ is fed into channel A and the modulating signal $b(t)$ is fed into channel B of the oscilloscope.

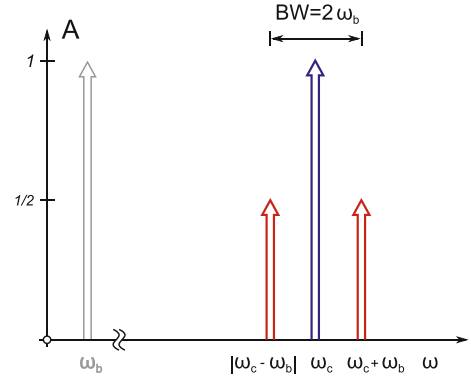


Fig. 11.5 Frequency spectrum of an AM signal containing a harmonic at the carrier frequency ω_c in addition to the two side harmonics $(\omega_c + \omega_b)$ and $|\omega_c - \omega_b|$

By setting the plotting mode so that channel A is on the vertical axis and channel B is on the horizontal axis, the amplitude-modulated signal plots trapezoidal patterns similar to the ones in Fig. 11.4.

It is straightforward to expand the definition of the AM index as

$$m = \frac{B}{C} = \frac{(C+B)-(C-B)}{(C+B)+(C-B)} = \frac{2(C+B)-2(C-B)}{2(C+B)+2(C-B)}, \tag{11.9}$$

which is easily correlated with the geometrical sizes of the plots in Fig. 11.4. Hence, by measuring lengths of the long and short trapezoidal sides directly on the oscilloscope and applying (11.9), we calculate the modulation index.

11.2.2 Frequency Spectrum of Amplitude-Modulated Signal

By inspection of (11.7) and (11.8), we realize that amplitude modulation is equivalent to the multiplication operation discussed in Sect. 9.1. Therefore, the frequency content of the amplitude-modulated signal contains the two side tones, the upper-side $(\omega_c + \omega_b)$ and the lower-side $|\omega_c - \omega_b|$ harmonics. In addition, the AM signal also contains harmonic ω_c at the carrier frequency. It is important to notice that the amplitude of the side tones is multiplied by $m/2$, which (in the best case of $m = 1$) means that amplitude of the side tones is half of the carrier amplitude tone (see Fig. 11.5). We also observe that, for a baseband signal whose highest harmonic is ω_b, the amplitude-modulated signal occupies the bandwidth $BW = 2\,\omega_b$ that is centred around the carrier frequency.

11.2.3 Average Power

It is important to quantify the amount of energy contained in each of the three harmonics of an AM signal (11.8) as

$$\begin{aligned} c_{AM}(t) &= \sin\omega_c t + \frac{m}{2}\cos\omega_L t - \frac{m}{2}\cos\omega_U t \\ &= c_C + c_L - c_U, \end{aligned} \tag{11.10}$$

where c_C is the instantaneous carrier voltage, c_L is the instantaneous voltage of the lower-side harmonic $|\omega_c - \omega_b|$, and c_U is the instantaneous voltage of the upper-side harmonic $(\omega_c + \omega_b)$. Hence, the instantaneous power of the AM wave across a resistor R is

$$p_{\mathrm{AM}} = \frac{c_{\mathrm{AM}}^2}{R}$$
$$= \frac{(c_{\mathrm{C}} + c_{\mathrm{L}} - c_{\mathrm{U}})^2}{R} = \frac{c_{\mathrm{C}}^2}{R} + \frac{c_{\mathrm{L}}^2}{R} + \frac{c_{\mathrm{U}}^2}{R} + \frac{2}{R}(c_{\mathrm{C}}c_{\mathrm{L}} - c_{\mathrm{L}}c_{\mathrm{U}} - c_{\mathrm{U}}c_{\mathrm{C}}), \tag{11.11}$$

where, the three squared terms denote the instantaneous power of each of the wave components: the carrier, the lower-side harmonic, and the upper-side harmonic.

Let us first evaluate the cross-product term $(c_{\mathrm{C}}c_{\mathrm{L}} - c_{\mathrm{L}}c_{\mathrm{U}} - c_{\mathrm{U}}c_{\mathrm{C}})$. As we discussed in Sect. 2.6.1, the product of two sinusoidal terms is another sinusoidal term (it is irrelevant for our discussion that it is frequency shifted). Furthermore, the average value of a sinusoidal waveform is zero, therefore all three cross products have zero average values and do not contribute to the total average power calculations.

With reference to (11.5) and (11.10), for each of the three squared terms in (11.11), starting with the carrier voltage, we write expressions for their average power as

$$P_{\mathrm{Cavg}} = \frac{c_{\mathrm{Crms}}^2}{R} = \frac{\left(\frac{C}{\sqrt{2}}\right)^2}{R} = \frac{C^2}{2R}, \tag{11.12}$$

$$P_{\mathrm{Lavg}} = \frac{c_{\mathrm{Lrms}}^2}{R} = \frac{\left(\frac{mC/2}{\sqrt{2}}\right)^2}{R} = \frac{m^2}{4}\frac{C^2}{2R} = \frac{m^2}{4}P_{\mathrm{Cavg}}, \tag{11.13}$$

$$P_{\mathrm{Uavg}} = \frac{c_{\mathrm{Urms}}^2}{R} = \frac{\left(\frac{mC/2}{\sqrt{2}}\right)^2}{R} = \frac{m^2}{4}\frac{C^2}{2R} = \frac{m^2}{4}P_{\mathrm{Cavg}} = P_{\mathrm{Lavg}}. \tag{11.14}$$

Hence, the total average power P_{T} of an AM waveform is therefore

$$P_{\mathrm{Tavg}} = P_{\mathrm{Cavg}} + \frac{m^2}{4}P_{\mathrm{Cavg}} + \frac{m^2}{4}P_{\mathrm{Cavg}} = P_{\mathrm{Cavg}}\left(1 + \frac{m^2}{2}\right). \tag{11.15}$$

In order to simplify the syntax, we keep in mind that (11.15) refers to the average power and we simply write

$$P_{\mathrm{T}} = P_{\mathrm{C}}\left(1 + \frac{m^2}{2}\right). \tag{11.16}$$

Again, the value of the AM factor m is important for the overall power transfer efficiency, with (11.16) showing that the total average power required for $m = 1$ is $P_{\mathrm{T}} = 1.5P_{\mathrm{C}}$, while each of the sidebands transfers only $1/4 P_{\mathrm{C}}$. We conclude that even for 100% AM scheme, i.e., $m = 1$, only $1/6$ of the total power is present in each of the sidebands (each contains its own copy of the useful information), while $2/3$ of the total power is in the carrier (which contains no information whatsoever).

Although the above analysis focused on a single-tone signal, we keep in mind that a non-sinusoidal modulating signal consists of a number of sine waves, not necessarily harmonically related. The overall average power is then the sum of the individual single-tone average powers:

$$P_{\mathrm{T}} = P_{\mathrm{C}}\left(1 + \frac{m_1^2}{2} + \frac{m_2^2}{2} + \cdots\right), \tag{11.17}$$

where m_i $(i = 1, 2, \ldots, n)$ is the modulation index of tone i. A detailed analysis of random signals similar to speech involves statistical mathematical models and is the subject of advanced courses in signal processing, hence it is omitted in this book.

11.2.4 Double-Sideband and Single-Sideband Modulation

The amplitude modulation scheme described so far is the most straightforward form of signal modulation. Its full name is "double-sideband–full carrier" (DSB–FC) modulation. In summary, it has the advantage of being very simple to modulate and demodulate, but has the following disadvantages:

- It can be over-modulated, which causes signal distortion (generation of new tones that can end up being shifted outside the assigned bandwidth).
- It is inefficient in the use of power (most of the transmitted power is in the carrier, which contains no information).
- Its required bandwidth is twice the modulation signal's bandwidth, that is, it does not use the frequency bandwidth efficiently, which is important because the overall available bandwidth is limited and directly controls the maximum number of the users.

The power inefficiency and bandwidth requirement of the DSB–FC modulating scheme present serious concerns for modern battery-powered RF equipment. By inspection of (11.8) and (11.16), we conclude that major power savings could be achieved by removing the carrier tone from the AM signal frequency spectrum before it reaches the transmitting antenna. After some further analysis, we also conclude that the two side tones are entirely down to the product of the carrier and the modulating signal, while the carrier term is caused only by the presence of the DC offset in the modulating signal. Therefore, if the modulating signal is somehow balanced so that the DC term cancels, the output spectrum would contain only the two sideband terms with no carrier term. Such a modulation scheme is known as "double-sideband–suppressed carrier" (DSB–SC) modulation.

Indeed, the whole class of symmetrical modulating circuits, commonly known as "balanced modulators", was developed to perform carrier tone removal. Before we proceed into specific circuit examples of AM circuits, let us summarize the possibilities from the conceptual point of view. In general, transmission of a single tone is rarely used, instead complicated signals such as speech or music are most often transmitted. From the conceptual perspective, our main concern is to determine the frequency bandwidth that is required by the signal. For instance, an audio signal occupies an approximately 20 kHz-wide bandwidth, which means that we must allocate a 20 kHz frequency bandwidth for its transmission. Hence, we need to use the terms "upper sideband" (USB) and "lower sideband" (LSB) to indicate that the transmitted signal consists of more than a single tone, where the sideband is bounded by the frequencies of its lowest and highest tones.

Once the carrier tone is removed by the use of balanced modulators, the modulated signal still consists of both upper and lower sidebands and occupies double the required bandwidth. This issue is resolved by using bandpass filters in the last stages of the AM transmitter, which removes either the lower or the higher sideband. This AM scheme is known as "single-sideband–suppressed carrier" (SSB–SC) modulation using either the USB or the LSB frequency range. The frequency spectrums of these four major AM schemes is summarized in Fig. 11.6.

11.2.4.1 Bandpass Filters for SSB Modulation

Although functionally very simple, the design of bandpass filters suitable for SSB suppression is limited by the available technology. The main problem becomes more obvious after we take a look at the commonly cited formula for the required Q factor of the bandpass filter, which is expressed in terms of the amount of required suppression A_{dB} as

$$Q = \frac{f_{\mathrm{C}}}{\Delta f} \frac{\sqrt{10^{\left(\frac{A_{\mathrm{dB}}}{20}\right)}}}{4}, \tag{11.18}$$

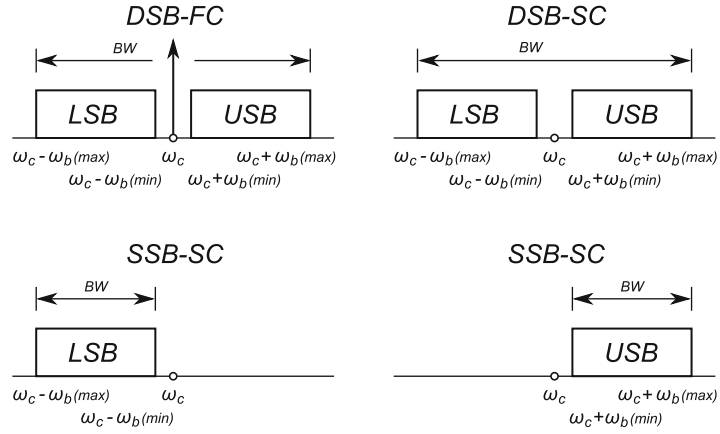


Fig. 11.6 Amplitude modulated signal spectrum: (**a**) double-sideband–full carrier (DSB–FC); (**b**) double-sideband–suppressed carrier (DSB–SC); (**c**) single-sideband–suppressed carrier (SSB–SC) using lower sideband (LSB); (**d**) single-sideband–suppressed carrier (SSB–SC) using upper sideband (USB)

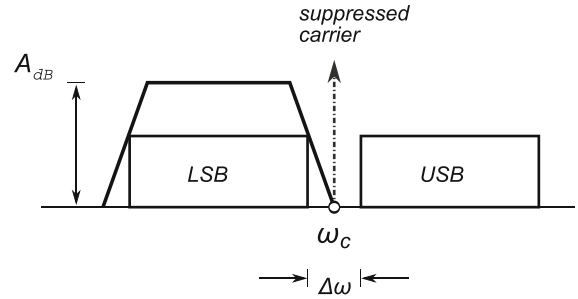


Fig. 11.7 Bandpass filter definition for LSB suppression, (11.18), of a DSB–SC signal

where f_C is the carrier frequency, Δf is the separation between the USB and the LSB, and A_{dB} is the required attenuation expressed in units of dB (see Fig. 11.7). A practical implementation of SSB suppression bandpass filters is based on the following options:

- Surface acoustic wave (SAW) filters are hybrid filters based on electromechanical signal conversion using piezoelectric materials. In the AM frequency range, this kind of filter may be able to achieve a Q factor as high as 35,000 (the literature data is not always conclusive).
- Crystal filters are another form of quartz crystal that can achieve a Q factor in the order of 20,000 (again, we take this number as indicative rather than definitive).
- Mechanical filters are based on the mechanical resonance of various metallic materials. We assume that their Q factor is of the order of 10,000, however some nickel–iron alloys apparently achieve a Q factor as high as 25,000.
- Ceramic filters are made of ceramic alloys and may be able to achieve Q factors in the order of 2,000.
- RLC filters are based on discrete components. With careful design, they may provide Q factors in the order of 500.

The Q factor numbers here are only for illustrative purposes. Keep in mind that the Q factor is defined for a specific centre frequency and bandwidth, which means that it may not always be possible to compare fairly the various types of SSB bandpass filter. Nevertheless, for the sake of the exercise, we assume that all these filters are comparable and on an equal footing.

Example 11.1. In a typical AM radio system, the signal bandwidth is $\Delta f = \pm 100\,\text{Hz}$. Estimate the type of SSB filter that is needed to suppress the LSB by $A_{dB} = 80\,\text{dB}$, if the centre frequency is: (a) $f_C = 100\,\text{kHz}$ and (b) $f_C = 1\,\text{MHz}$.

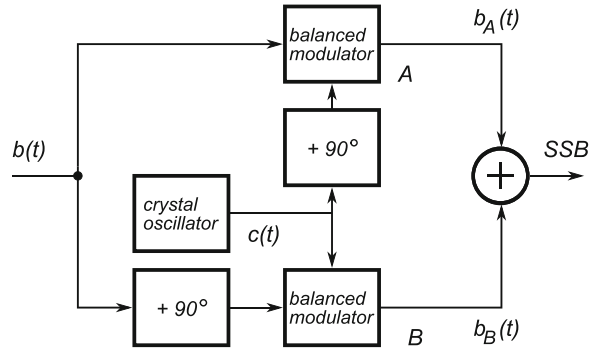


Fig. 11.8 Block diagram of a *phase shifter* generating an SSB AM signal

Solution 11.1. By direct implementation of (11.18), we write

(a)

$$Q = \frac{f_C}{\Delta f}\,\frac{\sqrt{10^{\left(\frac{A_{dB}}{20}\right)}}}{4} = \frac{100\,\text{kHz}}{200\,\text{Hz}}\,\frac{\sqrt{10^{\left(\frac{80}{20}\right)}}}{4}$$
$$= 12{,}500 \quad \text{i.e., we need a crystal filter or better.}$$

(b)

$$Q = \frac{1\,\text{MHz}}{200\,\text{Hz}}\,\frac{\sqrt{10^{\left(\frac{80}{20}\right)}}}{4}$$
$$= 125{,}000 \quad \text{i.e., we need several SAW filters in cascade.}$$

Aside from the "brute force approach" of using a high Q bandpass SSB suppression filter to directly remove one of the sidebands, there are number of more sophisticated techniques in use. To illustrate the possibilities, let us take a look at a typical representative method known as the "phase shift method".

In the phase shift method, two identical balanced modulators are used in parallel (Fig. 11.8). The message signal $b(t)$ is fed directly into modulator A and with a phase shift of 90° into modulator B. The carrier frequency is provided by the crystal oscillator and it is fed into modulator A with the 90° phase shift. The two output waves $b_A(t)$ and $b_B(t)$ are first added in the summing block and the output wave is an SSB with suppressed lower sideband. We keep in mind that balanced modulators cancel the carrier frequency. In order to see how the LSB cancellation happens, we need to take a look at the mathematics of this system.

Assuming that the message signal is $b(t) = \sin(\omega_b t)$ and the carrier signal is $c(t) = \sin(\omega_c t)$, then the two balanced modulator output signals are

$$b_A(t) = \cos[(\omega_c t + 90^\circ) - \omega_b t] - \cos[(\omega_c t + 90^\circ) + \omega_b t]$$
$$= \cos[\omega_c t - \omega_b t + 90^\circ] - \cos[\omega_c t + \omega_b t + 90^\circ], \tag{11.19}$$

$$b_B(t) = \cos[\omega_c t - (\omega_b t + 90^\circ)] - \cos[\omega_c t + (\omega_b t + 90^\circ)]$$
$$= \cos[\omega_c t - \omega_b t - 90^\circ] - \cos[\omega_c t + \omega_b t + 90^\circ]. \tag{11.20}$$

The first of the two cosine terms in (11.19) and (11.20) are LSB terms that are opposite in phase, the one in (11.19) is leading by 90° while the one in (11.20) is lagging by 90°. Therefore, they cancel when added in the summing block[2] and the output of the summing block is

$$SSB = b_{\mathrm{A}}(t) + b_{\mathrm{B}}(t) = -2\cos[\omega_{\mathrm{c}}t + \omega_{\mathrm{b}}t + 90^{\circ}] = 2\sin[\omega_{\mathrm{c}}t + \omega_{\mathrm{b}}t], \tag{11.21}$$

which is the USB signal at $\omega_{\mathrm{c}}t + \omega_{\mathrm{b}}t$ frequency. We note that this circuit produces true SSB AM output spectrum, that by controlling the phase of the LO we can easily choose to cancel either USB or LSB, and that there is no more need to have the SSB filter. On the negative side, the need for a wide band phase shifting circuit limits the practical range of this scheme. In addition, the demodulator on the receiving side needs to be synchronized with the incoming SSB wave; any mismatch in the transmitter and the receiver local waveforms will cause unwanted tones. Considering the number of SSB schemes that have been developed over time, we limit our discussion to the two basic techniques presented in this section.

11.2.5 The Need for Frequency and Phase Synchronization

In a transmitter's AM circuit, the information-carrying signal $b(t)$ is up-converted using the local carrier tone $c(t)$ that is generated by the transmitter's VCO. Once the amplitude-modulated signal departs from the antenna into space, the receiving circuit must tune in the appropriate frequency band and down-convert the incoming RF signal to the baseband. This frequency shifting is done by the receiver's mixer, which uses the receiver's VCO as the source of the high-frequency reference for its multiplication operation. Aside from technical details of the tuning process itself, there is another important issue that at first is not obvious.

A common assumption is that the receiver's local VCO generates exactly the same frequency ω_{c} as the one generated by the transmitter's VCO (which is located faraway from the receiver). Considering the vast distances between the transmitter and the receiver, it is natural to ask how these two frequencies are synchronized and if there is any consequence if they are not equal. To answer these questions, let us take a look what happens when the receiver's local VCO generates a tone that is only slightly off both in frequency and in phase relative to the tone generated by the transmitter's VCO, i.e., instead of the correct carrier wave $c(t) = f(t)\cos\omega_{\mathrm{c}}t$ generated by the transmitter, the receiver's VCO generates a slightly incorrect $c'(t) = \cos[(\omega_{\mathrm{c}} + \Delta\omega_{\mathrm{c}})t + \theta]$, which has an error both in frequency $\Delta\omega_{\mathrm{c}}$ and in phase $\theta \neq 0$. For simplicity, the modulating signal is $f(t)$, so that the multiplication operation is performed by the receiver as[3]

$$\begin{aligned}
&\{f(t)\cos\omega_{\mathrm{c}}t\} \times \{\cos[(\omega_{\mathrm{c}} + \Delta\omega_{\mathrm{c}})t + \theta]\} \\
&\quad = f(t)\,\{\cos\omega_{\mathrm{c}}t\cos\omega_{\mathrm{c}}t\cos(\Delta\omega_{\mathrm{c}}t + \theta) - \cos\omega_{\mathrm{c}}t\sin\omega_{\mathrm{c}}t\sin(\Delta\omega_{\mathrm{c}}t + \theta)\} \\
&\quad \therefore \\
&\quad = \frac{1}{2}f(t)\cos(\Delta\omega_{\mathrm{c}}t + \theta) + \frac{1}{2}f(t)\cos 2\omega_{\mathrm{c}}t\cos(\Delta\omega_{\mathrm{c}}t + \theta) \\
&\qquad - \frac{1}{2}f(t)\sin 2\omega_{\mathrm{c}}t\sin(\Delta\omega_{\mathrm{c}}t + \theta),
\end{aligned} \tag{11.22}$$

[2] See Sect. 2.6.4.

[3] Trigonometric identity: $\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta$.

where the last three terms denote the frequency spectrum due to the frequency and phase mismatch. The last two terms are at a high frequency, close to $2\omega_c t$ because $2\omega_c \gg \Delta\omega_c$, i.e., $2\omega_c \pm \Delta\omega_c \approx 2\omega_c$, and are easily removed by a bandpass filter centred at $\Delta\omega_c$. The first of the terms, however, shows that instead of correctly recovering the information signal $f(t)$, the resulting waveform is a function of $\cos(\Delta\omega_c t + \theta)$. This is a rather serious issue since each time the cosine argument equals an odd number multiple of $\pi/2$, i.e., $(\Delta\omega_c t + \theta) = (2n+1)\pi/2$, the entire $f(t)$ signal disappears. The user perceives this as a strong audio effect known as "beating" of the output signal.

Therefore, in order to correctly demodulate the incoming DSB signal, it is necessary to multiply it by using the local tone with exactly correct phase and frequency. This means that a DSB radio system requires quite a complicated receiver. However, on the positive side, if there is no frequency and phase mismatch, it is possible to simultaneously transmit at the same frequency on a second DSB channel but with carrier phase $\theta = 90^\circ$ with respect to the first channel. These two signals can be received independently of one another. This transmitting scheme is known as quadrature multiplexing.

In Chap. 10, we introduced phase-locked-loop (PLL) circuit topology. PLL circuits are fundamental for wireless communication because they are capable of synthesizing periodic waveforms (either sinusoidal or square) that are both phase- and frequency-locked to the waveform of the RF carrier. Hence, they can (ideally) eliminate problems related to the phase and frequency offsets between the local VCO and the RF carrier waveforms.

In summary, DSB and SSB transmitting schemes have the following main advantages:

- They are efficient in respect to the signal power, there is no waste due to the carrier tone.
- DSB can transmit two channels simultaneously by using the quadrature multiplexing methodology.
- SSB modulation is efficient in terms of its frequency bandwidth requirements.

The main disadvantages of these AM schemes are that a much more complicated transceiver is required and that SSB is not suitable for pulse (digital) communication or music. Today, a number of various schemes are used in either DSB or SSB modulation receivers.

11.2.6 Amplitude Modulator Circuits

From the mathematical point of view, amplitude modulation operation is equivalent to the frequency-shifting operation described in Chap. 9. Therefore, the mixer circuits described in Sects. 9.3–9.5 are perfectly valid as amplitude modulator circuits. The diode mixer introduced in Sect. 9.2 does not have gain; hence, it is used only at very high frequencies where the active devices also lose their gain.

Modulation is done inside the transmitter circuit and a wide range of circuits are used for AM. A detailed study of transmitter circuits is left for another occasion; in this book, we focus only on the general principles and a few typical modulation circuits.

Amplitude modulation may be done either at a low level in the transmitter, i.e., where the signal power is still relatively low, for example at the base of an input BJT or at a high level in the transmitter hierarchy, i.e., both the carrier and the modulating signals are combined close to the antenna.

There are at least three possible low-level schemes used to generate AM waves, as shown by the block diagrams in Fig. 11.9:

1. A literal implementation of the AM waveform model (11.7).
2. A nonlinear device that produces an AM waveform, which is mathematically approximated by a polynomial as

$$v_o = a_0 + a_1 v_i + a_2 v_i^2 + \cdots . \tag{11.23}$$

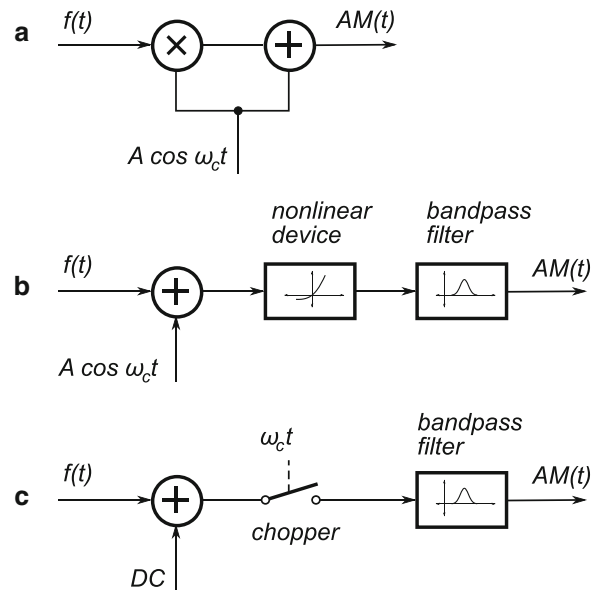


Fig. 11.9 The principles of low-level AM modulators

After substituting $v_i = f(t) + A\cos\omega_c t$ it follows that

$$v_o = a_0 + a_1 f(t) + a_1 \cos\omega_c t + a_2 f^2(t) + 2a_2 A f(t)\cos\omega_c t + a_2 A^2 \cos^2\omega_c t + \cdots, \tag{11.24}$$

which, after expanding the squared cosine terms,[4] shows that a nonlinear device generates spectrum components that do not exist in the original signal. Hence, a bandpass filter is needed to remove the frequency components that are not close to the carrier frequency, so that

$$v(t) \approx a_1 \cos\omega_c t + 2a_2 A f(t)\cos\omega_c t, \tag{11.25}$$

which is the desired AM waveform.

3. Nonlinearity does not have to be provided by an active component. Inherently, any switching function is also nonlinear, in fact it is very nonlinear. A switching device that is controlled by a periodic signal ω_c is usually referred to as a "chopper" and may be used to generate an AM waveform. Effectively, the chopper works as a multiplier for its input signal $b(t)$ and the switching square wave (i.e., the switch is a binary device) at ω_c frequency. Again, the chopper is followed by a bandpass filter that is needed to filter out harmonics from the pulse spectrum.

One of main disadvantages of low-level AM modulators is that the modulation must be followed by a linear amplifier. Linear amplifiers are relatively inefficient for power transfer applications, hence these modulating schemes are not used for high-power RF transmitters in commercial broadcasting radio stations or for modern battery-powered wireless devices. Instead, some topology based on a high-level scheme that employs a class C amplifier is used more often. In the following pages, we briefly introduce several commonly used circuits for amplitude modulation.

[4] Use trigonometric identities for $\cos^2\theta = \dfrac{1+\cos 2\theta}{2}$ and $\sin^2\theta = \dfrac{1-\cos 2\theta}{2}$.

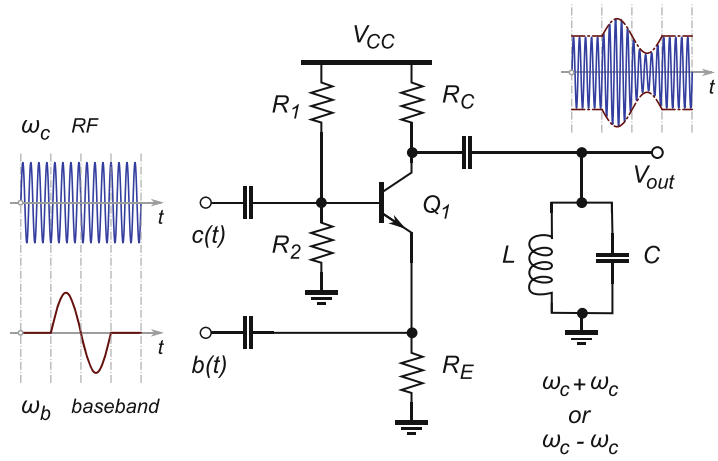


Fig. 11.10 A typical BJT amplitude modulator circuit

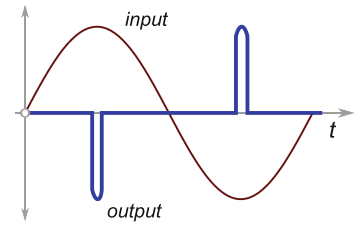


Fig. 11.11 Time domain plot of input and output signal shapes for a class C amplifier that illustrates the nonlinear mode of operation

11.2.6.1 BJT AM Circuit

At least in principle, one of the simplest ways to do amplitude modulation is to feed the carrier tone and the modulation signal to a single active device as shown in Fig. 11.10. The base serves as the input terminal for the RF signal $c(t)$, while the emitter serves as the input terminal to the modulation signal $b(t)$. The nonlinear characteristics of the active device provides frequency shifting, while the LC resonant circuit at the output node is tuned to either of the two sidebands. As we already know, the resonator tank presents high impedance at the chosen sideband frequency, while effectively shorting to the ground all other tones in the signal spectrum.

11.2.6.2 Class C AM Circuit

Most modern, highly efficient, low-power, battery-powered wireless devices employ a class C amplifier (or some other switching amplifier configuration) in their transmitting stages. One of the main advantages of using a class C amplifier in a broadcast AM transmitter is that only the final stage needs to be modulated; all the preceding stages can be driven at a constant signal level. However, in order to obtain a better quality AM waveform sometimes the last two stages of the transmitter are modulated. The power efficiency usually associated with switching amplifiers comes from their nonlinear mode of operation, which is in contrast to linear amplifiers. Instead of faithfully reproducing the input signal throughout the whole period, switching amplifiers are turned one only for a short period of time, Fig. 11.11.

The main disadvantage of the class C amplifier approach is that a large audio amplifier is also needed for the modulation stage, with power consumption at least equal to the transmitter output itself. Traditionally, the modulation is applied using an audio transformer and this device is also bulky.

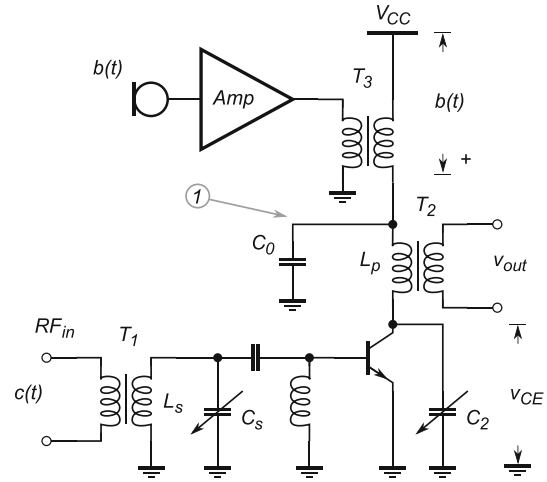


Fig. 11.12 A BJT collector modulator

Direct coupling from the audio amplifier is also possible (known as a cascode arrangement), though this usually requires quite a high DC supply voltage (around 30 V or even more), which is not suitable for mobile units.

A simplified schematic diagram of a typical class C modulator circuit is shown in Fig. 11.12. The input RF carrier signal $c(t)$ enters through transformer T_1 whose secondary inductance L_S together with capacitor C_S creates an LC resonating bandpass filter. The BJT is biased in class C operation, which means that the transistor conducts only for a very short period of time close to the peak amplitudes of the carrier signal, Fig. 11.11. These short pulses cause the collector current which, in turn, drives the output resonator circuit which consists of a primary inductor L_p of the output transformer T_2 and capacitor C_2. Decoupling capacitor C_0 forces node ① to become a small signal ground.

The modulating signal $b(t)$ is added to the supply voltage V_{CC} through the secondary of the transformer T_3. For sinusoidal modulation, the collector supply voltage is

$$\begin{aligned} v_{CE} &= V_{CC} b(t) = V_{CC} + B \sin \omega_b t \\ &= V_{CC} (1 + m \sin \omega_b t), \end{aligned} \tag{11.26}$$

where $m = B/V_{CC}$ is the AM index. When $m = 1$ the peak modulation signal voltage must equal the supply voltage, i.e., $B = V_{CC}$. Consequently, when there is no modulation, $b(t) = 0$, the collector voltage would be V_{CC}, while with the maximum RF input the voltage swing is $2V_{CC}$, which maintains the collector voltage average of V_{CC}. Therefore, the positive peak of RF swing can reach $4V_{CC}$. This is an important design parameter that requires that the BJT must have high breakdown voltage.

11.2.6.3 Balanced AM Circuits

The main property of a balanced modulator is that it outputs the product of the two input signals $b(t)$ and $c(t)$ while suppressing one or both of them. Based on whether only one or both of the input tones is removed from the output spectrum, the balanced modulator is said to be either single- or double-balanced. That is, in the ideal case of a double-balanced modulator, the output spectrum contains $\omega_c \pm \omega_b$ tones but neither ω_c nor ω_b themselves. If signal-suppressing input is used by the carrier $c(t)$ then a balanced modulator produces the DSB–SC spectrum. A large number of balanced modulator designs are in use; we review the operation of three typical circuits: diode ring modulators, balanced FET modulators, and IC balanced modulators.

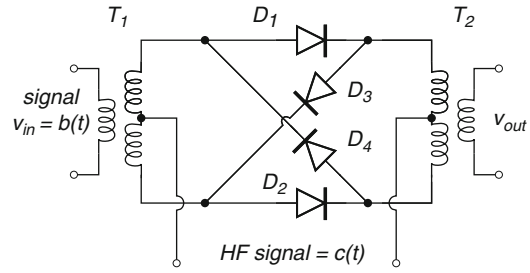


Fig. 11.13 Double-balanced diode ring modulator

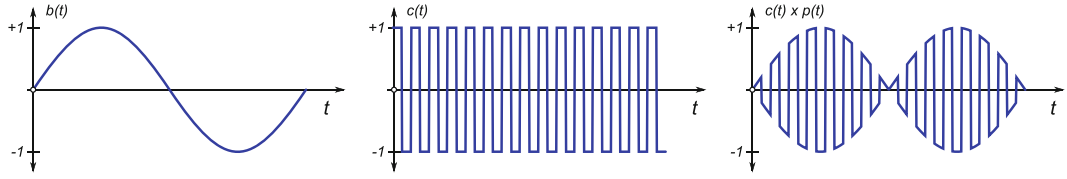


Fig. 11.14 Multiplication of signal $b(t)$ with ± 1 pulse stream $c(t)$ results in an amplitude-modulated carrier signal $c_{\mathrm{AM}}(t) = b(t) \times c(t)$

11.2.6.4 Double-Balanced Diode Ring Modulator

A simple circuit commonly used for low-frequency applications in telephone networks is based on four diodes and two transformers (see Fig. 11.13). The bulkiness of the two transformers is probably the main reason this modulator is not used in more applications. The amplitude-modulated output produced by the diode ring modulator is "double-balanced with suppressed carrier". Diode pairs D_1–D_2 and D_3–D_4 are alternatively switched on and off by the HF carrier $c(t)$, where the carrier signal could be either a sinusoidal or a square wave at frequency ω_{c} whose amplitude is much larger than that of the modulation signal $b(t)$. Therefore, the output signal $c_{\mathrm{AM}}(t)$ is produced by multiplying the modulation signal $b(t)$ with ± 1 pulse stream, Fig. 11.14.

In the ideal circuit, all four diodes are perfectly matched and the two transformers are perfectly symmetrical, hence, when HF signal is zero, $c(t) = 0$, then the output signal is also zero, $c_{\mathrm{AM}}(t) = 0$. By inspection, we follow the HF current entering the centre tap of transformer T_1, which then splits and passes in parallel through diodes D_1 and D_2, only to converge again at the centre tap of T_2 and return to the HF source. The two HF currents that enter the primary side of transformer T_2 in opposite directions induce voltages of equal magnitude and opposite polarity in the T_2 secondary, which therefore cancel each other and produce zero voltage output.

The square wave switching function $c(t)$ can be written using its Fourier transformation with amplitude $A = \pi/2$ as

$$c(t) = \sin \omega_{\mathrm{c}} t + \frac{1}{3} \sin 3\,\omega_{\mathrm{c}} t + \cdots + \frac{1}{n} \sin n\,\omega_{\mathrm{c}} t + \cdots, \tag{11.27}$$

where $n = 2k + 1$ is an odd number. Further mathematical analysis shows that multiplication of the time-varying signal $c(t)$ and the sinusoidal modulation $b(t) = \sin \omega_{\mathrm{b}} t$, after substituting (11.27), results in

$$\begin{aligned} c_{\mathrm{AM}} &= b(t) \times c(t) \\ &= \sin \omega_{\mathrm{b}} t \, \sin \omega_{\mathrm{c}} t + \sin \omega_{\mathrm{b}} t \, \frac{1}{3} \sin 3\,\omega_{\mathrm{c}} t + \cdots + \sin \omega_{\mathrm{b}} t \, \frac{1}{n} \sin n\,\omega_{\mathrm{c}} t + \cdots, \end{aligned} \tag{11.28}$$

which, after expanding all products, shows that the output spectrum contains only $(n\omega_{\mathrm{c}} \pm \omega_{\mathrm{b}})$ terms with neither ω_{c} nor ω_{b} terms by themselves, hence this is a double-balanced modulator.

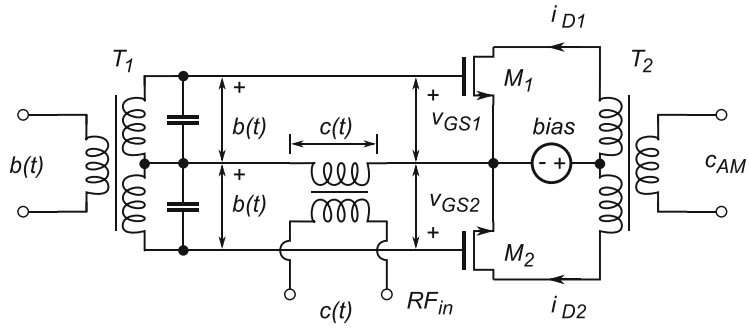


Fig. 11.15 Single-balan -ced FET modulator

11.2.6.5 Single-Balanced FET Modulator

Instead of using diodes, two FET transistors are connected as in Fig. 11.15 to produce a DSB–SC amplitude-modulated waveform. The modulation signal $b(t)$ is split into two copies at the symmetrical secondary of T_1, while the carrier signal $c(t)$ is injected into the circuit through its own 1:1 transformer. The two FET transistors are perfectly matched, which is usually achieved by using IC components.

By inspection, we write

$$v_{\mathrm{GS1}} = c(t) + b(t) = C\cos\omega_{\mathrm{c}}t + B\cos\omega_{\mathrm{b}}t, \tag{11.29}$$

$$v_{\mathrm{GS2}} = c(t) - b(t) = C\cos\omega_{\mathrm{c}}t - B\cos\omega_{\mathrm{b}}t, \tag{11.30}$$

hence, the drain currents of the two FET transistors are approximated by the second-order polynomials as

$$i_{\mathrm{D1}} \approx I_0 + a_1 v_{\mathrm{GS1}} + a_2 v_{\mathrm{GS1}}^2, \tag{11.31}$$

$$i_{\mathrm{D2}} \approx I_0 + a_1 v_{\mathrm{GS2}} + a_2 v_{\mathrm{GS2}}^2, \tag{11.32}$$

hence, after substituting (11.29) and (11.30), we write

$$i_{\mathrm{D1}} = I_0 + a_1(C\cos\omega_{\mathrm{c}}t + B\cos\omega_{\mathrm{b}}t) + a_2(C\cos\omega_{\mathrm{c}}t + B\cos\omega_{\mathrm{b}}t)^2, \tag{11.33}$$

$$i_{\mathrm{D2}} = I_0 + a_1(C\cos\omega_{\mathrm{c}}t - B\cos\omega_{\mathrm{b}}t) + a_2(C\cos\omega_{\mathrm{c}}t - B\cos\omega_{\mathrm{b}}t)^2. \tag{11.34}$$

The total current in the primary of transformer T_2 is the current difference $(i_{\mathrm{D1}} - i_{\mathrm{D2}})$, which subsequently induces the output voltage AM waveform as

$$c_{\mathrm{AM}} = M\frac{\mathrm{d}(i_{\mathrm{D1}} - i_{\mathrm{D2}})}{\mathrm{d}t}, \tag{11.35}$$

where M is the mutual inductance between the primary and secondary inductances of transformer T_2 (in an ideal case $M = 1$). Therefore,

$$\begin{aligned}(i_{\mathrm{D1}} - i_{\mathrm{D2}}) &= 2a_1 B\cos(\omega_{\mathrm{b}}t) + 4a_2 BC\cos(\omega_{\mathrm{b}}t)\cos(\omega_{\mathrm{c}}t)\\ &= 2a_1 B\cos(\omega_{\mathrm{b}}t) + 2a_2 BC\cos(\omega_{\mathrm{c}} - \omega_{\mathrm{b}})t + 2a_2 BC\cos(\omega_{\mathrm{c}} + \omega_{\mathrm{b}})t.\end{aligned} \tag{11.36}$$

By inspection of (11.36) we conclude that its first term contains frequency component ω_{b} of the modulating signal $b(t)$, the second term is at frequency $(\omega_{\mathrm{c}} - \omega_{\mathrm{b}})$, and the third term is at $(\omega_{\mathrm{c}} + \omega_{\mathrm{b}})$.

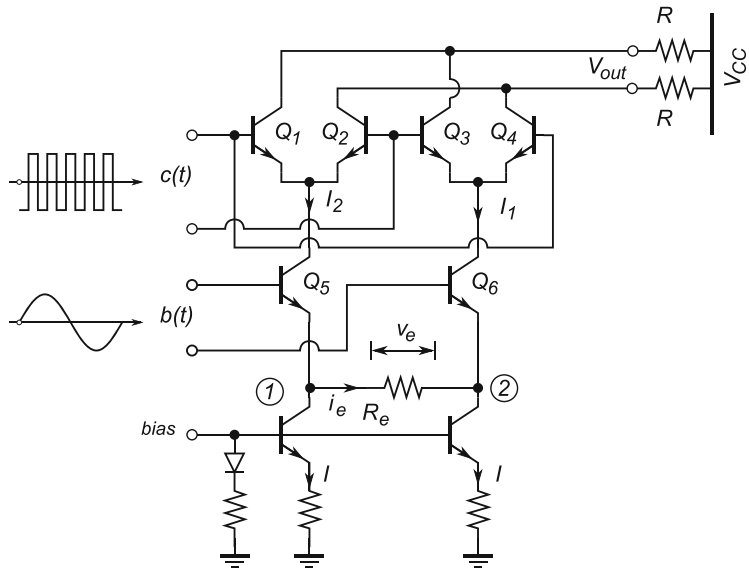


Fig. 11.16 Double-balanced IC modulator based on a Gilbert cell

The linear addition of these three terms and derivation (11.35) do not change the frequency spectrum, hence the output voltage c_{AM} waveform contains only the two sidetones $(\omega_{\mathrm{c}} \pm \omega_{\mathrm{b}})$ and the modulation signal $b(t)$. It should be noted that if waveforms $b(t)$ and $c(t)$ exchanged input terminals, then the carrier tone would have survived and $b(t)$ would have been suppressed from the output frequency spectrum.

11.2.6.6 Double-Balanced IC Modulator

In order to reduce the size of the electronic equipment, the use of transformers in modern mobile devices is avoided if possible. Ideally, the goal is to design all critical functions that are needed for radio equipment communication using IC technology. It is no surprise that balanced modulators are available in IC form (Fig. 11.16). The AM modulator circuit is a Gilbert-cell-based differential multiplier. A Gilbert cell is a very versatile circuit that has many applications. If used in switching mode, it works as a balanced AM multiplier circuit.

The circuit works as a multiplier of the two input signals. When it is used as a balanced modulator, an elementary analysis of the circuit is as follows. First, we find the output signal when the carrier signal is absent. Second, we add the carrier signal as the product. The carrier is considered to be a high-level switching voltage that alternately switches transistors pairs Q_1–Q_4 and Q_2–Q_3 on and off.

With no carrier signal applied, and assuming that the base currents are negligible, after summing the currents at junctions ① and ②, we write

$$\begin{aligned} I_2 &= I + i_{\mathrm{e}}, \\ I_1 &= I - i_{\mathrm{e}}. \end{aligned} \tag{11.37}$$

Hence, the output voltage v_{o} is

$$v'_{\mathrm{o}} = v_2 - v_1 = R(I_2 - I_1) = R(2i_{\mathrm{e}}). \tag{11.38}$$

Application of KVL to the loop that contains the modulating voltage signal $b(t)$ and resistance R_e yields

$$b(t) = V_{be5} + v_e - V_{be6}, \tag{11.39}$$

$$\therefore$$

$$b(t) \approx v_e \tag{11.40}$$

because the circuit operates with small signal current $I \gg i_e$ and keeps $V_{be5} \approx V_{be6}$. Therefore

$$i_e = \frac{v_e}{R_e} = \frac{b(t)}{R_e}, \tag{11.41}$$

$$\therefore$$

$$v'_{out} = \frac{2R}{R_e} b(t) = \frac{2R}{R_e} \sin \omega_b t \tag{11.42}$$

after substituting (11.41) back into (11.38). When the carrier signal $c(t)$ is added then the output is the product of the two. In the case of a square carrier signal, which contains an infinite number of odd harmonics, we approximate its function as

$$c(t) = \sin(\omega_c t) + \frac{1}{3}\sin(3\omega_c t) + \frac{1}{5}\sin(5\omega_c t) + \cdots \approx \sin(\omega_c t), \tag{11.43}$$

where all higher harmonics (at $3\omega_c$, $5\omega_c$, ...) are easily filtered out, which leads to an expression for the output voltage v_o when both the carrier $c(t)$ and the modulation $b(t)$ signals are present,

$$\begin{aligned} v_{out} &\approx v'_{out} \times c(t) = \frac{2R}{R_e} \sin(\omega_b t) \times \sin(\omega_c t) \\ &= \frac{R}{R_e} \left[\cos(\omega_c - \omega_b)t - \cos(\omega_c + \omega_b)t\right], \end{aligned} \tag{11.44}$$

that is, the output contains only the upper and lower side tones, while the carrier itself does not appear in the output. This modulator circuit is a typical example of how, by taking advantage of IC technology where the components are manufactured as perfect copies of each other, almost perfectly balanced voltages and currents are possible without the external bulky components.

11.3 Angle Modulation

Following the discussion in Sect. 11.1 in regard to (11.3), we now proceed to find out how the two angular parameters of the carrier signal $c(t)$, ω and phase ϕ can be modulated by the modulating signal $b(t)$. Although, frequency and PM are similar and are often studied together under the more inclusive name "angle modulation", the two are different in very important details. Frequency modulation (FM) is commonly used for HiFi broadcasting of music and speech, because of its lower sensitivity to noise, while PM requires a more complicated receiver and is used in some wireless LAN standards, and military and space applications.

11.3.1 Frequency Modulation

The modulating signal $b(t)$ in (11.4) is used to vary the frequency ω_c of the carrier waveform $c(t)$ in the time domain. Let the change in carrier frequency be

$$\Delta\omega_c = k\,b(t), \tag{11.45}$$

where k is a constant known as the "frequency deviation constant"; then the instantaneous carrier frequency is

$$\omega(t) = \omega_c + \Delta\omega_c = \omega_c + k\,b(t), \tag{11.46}$$

where ω_c is the unmodulated carrier frequency. After substituting $b(t) = B\cos\omega_b t$ in (11.46), the instantaneous frequency $f(t)$ of the FM waveform becomes

$$\omega(t) = \omega_c + kB\cos\omega_b t, \tag{11.47}$$

$$\therefore$$

$$f(t) = \frac{\omega(t)}{2\pi} = f_c + \frac{kB}{2\pi}\cos\omega_b t, \tag{11.48}$$

that is, the maximum and minimum values of the instantaneous frequency are

$$f_{max} = f_c + \frac{kB}{2\pi}, \tag{11.49}$$

$$f_{min} = f_c - \frac{kB}{2\pi}, \tag{11.50}$$

where the maximum swing of the instantaneous frequency from the unmodulated carrier frequency f_c is called "peak frequency deviation" Δf and is defined as

$$\Delta f \equiv f_{max} - f_c = \frac{kB}{2\pi}, \tag{11.51}$$

which enables us to define the FM index m_f and the deviation ratio δ as

$$m_f \equiv \frac{\Delta f}{f_m} = \frac{kB}{\omega_b}, \tag{11.52}$$

$$\delta \equiv \frac{\Delta f}{f_c} = \frac{kB}{\omega_c}. \tag{11.53}$$

An example sketch diagram of the instantaneous frequency $\omega_c(t)$ variation over time is shown in Fig. 11.17 (top). It is important to understand that this graph illustrates the frequency–time curve and not the amplitude–time curve, which is shown in Fig. 11.17 (bottom).

An analytical expression for the FM waveform may be derived as follows. The unmodulated carrier is a sine wave, therefore

$$c(t) = C\sin(\omega_c t + \phi). \tag{11.54}$$

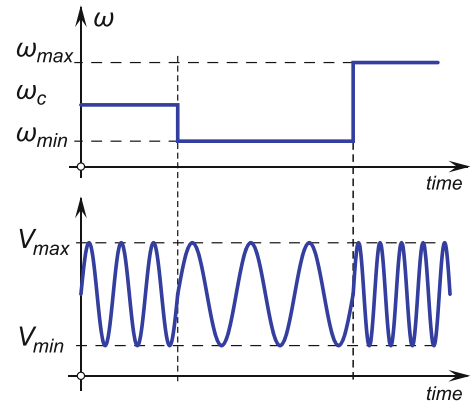


Fig. 11.17 FM waveform: Frequency over time (*top*) and amplitude over time (*bottom*). The modulating signal information is clearly embedded by the control of the carrier's frequency

Equation (11.54) is only a special case of the more general case

$$c(t) = C\sin[\theta(t)], \tag{11.55}$$

where $\theta(t)$ is an arbitrary time-dependent function. By definition, the angular frequency $\omega_c(t)$ is the rate of change in time of $\theta(t)$. Only when the frequency is constant is the particular form of (11.54) valid. When the frequency is time-dependent, as in FM, an instantaneous angular frequency may be defined as

$$\omega_c(t) = 2\pi f_c(t) = \frac{d\theta(t)}{dt}, \tag{11.56}$$

$$\therefore$$

$$\theta(t) = \int \omega_c(t)\,dt. \tag{11.57}$$

The instantaneous frequency $\omega_c(t)$ is related to the modulated frequency through relation (11.48). For instance, in the case of constant (unmodulated) angular frequency ω_c, we write

$$\theta(t) = \int \omega_c\,dt = \omega_c t + \phi, \tag{11.58}$$

where ϕ is the integration constant. In the case of sinusoidal modulation, after substituting (11.54) into (11.57) we have

$$\theta(t) = \int 2\pi\,(f_c + \Delta f\cos\omega_b t)\,dt = \omega_c t + \frac{\Delta f}{f_m}\sin\omega_b t + \phi. \tag{11.59}$$

The integration constant ϕ may be made equal to zero by an appropriate choice of time reference axis, while the expression for the sinusoidally modulated FM wave is obtained by substituting (11.59) into (11.55) as[5]

$$\begin{aligned} c_{FM} &= C\sin\left(\omega_c t + \frac{\Delta f}{f_m}\sin\omega_b t\right) \\ &= C\sin\left(\omega_c t + m_f\sin\omega_b t\right) \\ &= C\left[\sin(\omega_c t)\cos(m_f\sin\omega_b t) + \cos(\omega_c t)\sin(m_f\sin\omega_b t)\right]. \end{aligned} \tag{11.60}$$

[5]Use the trigonometric identity: $\sin(\alpha\pm\beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta$.

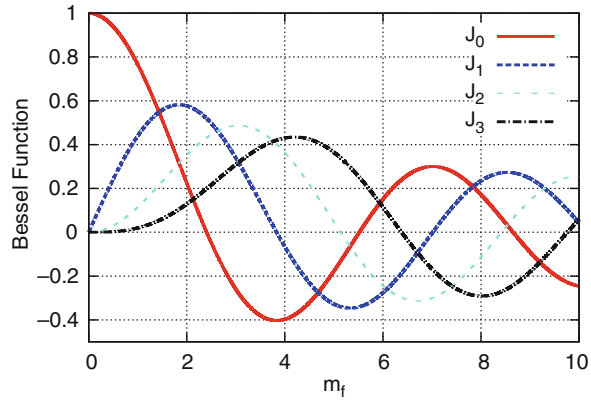


Fig. 11.18 Bessel functions $J_n(m_\mathrm{f})$ for $n = 0, 1, 2, 3$

We note that, unlike the AM index m, the FM index m_f can be greater than unity.

It turns out that mathematicians have already found suitable expansions for functions of type $\cos(x \sin y)$ using Bessel functions, i.e.,

$$\cos(m_\mathrm{f} \sin \omega_\mathrm{b} t) = J_0(m_\mathrm{f}) + 2 \sum_{n=1}^{\infty} J_{2n}(m_\mathrm{f}) \cos(2n\omega_\mathrm{b} t), \tag{11.61}$$

$$\sin(m_\mathrm{f} \sin \omega_\mathrm{b} t) = 2 \sum_{n=0}^{\infty} J_{2n+1}(m_\mathrm{f}) \sin[(2n+1)\omega_\mathrm{b} t], \tag{11.62}$$

where $J_n(m_\mathrm{f})$ is the Bessel function of the first kind and of n-th order. After substituting (11.61) and (11.62) into (11.60), and after expanding the sinusoidal products, the analytical expression of the FM waveform for the case of sinusoidal modulation (11.60) is written as

$$\begin{aligned} c_\mathrm{FM} = {} & J_0(m_\mathrm{f}) C \sin \omega_\mathrm{c} t \\ & + J_1(m_\mathrm{f}) C [\sin(\omega_\mathrm{c} + \omega_\mathrm{b}) t - \sin(\omega_\mathrm{c} - \omega_\mathrm{b}) t] \\ & + J_2(m_\mathrm{f}) C [\sin(\omega_\mathrm{c} + 2\omega_\mathrm{b}) t + \sin(\omega_\mathrm{c} - 2\omega_\mathrm{b}) t] \\ & + J_3(m_\mathrm{f}) C [\sin(\omega_\mathrm{c} + 3\omega_\mathrm{b}) t - \sin(\omega_\mathrm{c} - 3\omega_\mathrm{b}) t] \\ & + \cdots, \end{aligned} \tag{11.63}$$

where, Bessel function $J_n(m_\mathrm{f})$ is defined by the series

$$\begin{aligned} J_n(m_\mathrm{f}) = \frac{m_\mathrm{f}^n}{2^n n!} \Bigg[1 - \frac{m_\mathrm{f}^2}{2(2n+2)} + \frac{m_\mathrm{f}^4}{2(4)(2n+2)(2n+4)} \\ - \frac{m_\mathrm{f}^6}{2(4)(6)(2n+2)(2n+4)(2n+6)} + \cdots \Bigg]. \end{aligned} \tag{11.64}$$

It is handy to have Bessel functions (11.64) both in graphical form (Fig. 11.18) and in tabular form (Table 11.1). For instance, by reading values from the table for FM index $m_\mathrm{f} = 1.0$, we find that first five significant spectral component amplitudes are:

Table 11.1 Bessel functions of order 1–10, for modulation factors 0–10

m_f	J_0	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9	J_{10}
0.0	1.000	–	–	–	–	–	–	–	–	–	–
0.2	0.990	0.099	0.005	–	–	–	–	–	–	–	–
0.4	0.960	0.196	0.020	0.001	–	–	–	–	–	–	–
0.5	0.938	0.242	0.030	0.002	–	–	–	–	–	–	–
0.6	0.912	0.287	0.044	0.004	–	–	–	–	–	–	–
0.8	0.846	0.369	0.076	0.010	0.001	–	–	–	–	–	–
1.0	0.765	0.440	0.115	0.020	0.002	–	–	–	–	–	–
1.2	0.671	0.498	0.159	0.033	0.005	0.001	–	–	–	–	–
1.4	0.567	0.542	0.207	0.050	0.009	0.001	–	–	–	–	–
1.5	0.512	0.558	0.232	0.061	0.012	0.002	–	–	–	–	–
1.6	0.455	0.570	0.257	0.072	0.015	0.002	–	–	–	–	–
1.8	0.340	0.582	0.306	0.099	0.023	0.004	0.001	–	–	–	–
2.0	0.224	0.577	0.353	0.129	0.034	0.007	0.001	–	–	–	–
2.5	−0.048	0.497	0.446	0.217	0.074	0.019	0.004	0.001	–	–	–
3.0	−0.260	0.339	0.486	0.309	0.132	0.043	0.011	0.002	–	–	–
3.5	−0.380	0.137	0.459	0.387	0.204	0.080	0.025	0.008	0.001	–	–
4.0	−0.397	−0.066	0.364	0.430	0.281	0.132	0.049	0.015	0.004	0.001	–
4.5	−0.321	−0.231	0.218	0.425	0.348	0.195	0.084	0.030	0.009	0.002	0.001
5.0	−0.178	−0.328	0.467	0.365	0.391	0.261	0.131	0.053	0.018	0.005	0.001
6.0	0.151	−0.277	−0.243	0.115	0.358	0.362	0.246	0.130	0.056	0.021	0.007
7.0	0.300	−0.005	−0.301	−0.168	0.158	0.348	0.339	0.234	0.128	0.059	0.023
8.0	0.172	0.235	−0.113	−0.291	−0.105	0.186	0.338	0.321	0.223	0.126	0.061
9.0	−0.090	0.245	0.145	−0.181	−0.265	−0.055	0.204	0.327	0.305	0.215	0.125
10.0	−0.246	0.043	0.255	0.058	−0.220	−0.234	−0.014	0.217	0.318	0.292	0.207

carrier frequency	(f_c)	$J_0(1.0) = 0.765$
first-order side frequencies	$(f_c \pm f_b)$	$J_1(1.0) = 0.440$
second-order side frequencies	$(f_c \pm 2f_b)$	$J_2(1.0) = 0.115$
third-order side frequencies	$(f_c \pm 3f_b)$	$J_3(1.0) = 0.020$
fourth-order side frequencies	$(f_c \pm 4f_b)$	$J_3(1.0) = 0.002$

The fact that the spectrum components around the carrier frequency decrease in amplitude does not mean that the carrier wave is amplitude modulated. The carrier wave is the sum of all harmonics in its spectrum, and the harmonics add up to produce the constant amplitude FM waveform, Fig. 11.17 (bottom). The main distinction to note is that the FM modulated carrier is not a sine wave, whereas each of the spectrum components around the carrier frequency is. In addition, negative amplitudes in Table 11.1 only indicate the phase inversion.

Existence of the negative harmonic amplitudes implies that there are values of the FM index m_f for which the corresponding harmonic amplitude is zero, for instance if $m_f = 2.4, 5.5, 8.65, \ldots$ amplitude of the tone at the carrier frequency becomes zero. It is important to distinguish this case from its matching AM-balanced case of suppressed carrier. For the FM waveform, if the tone at the carrier frequency is suppressed it only means that its energy is redistributed to the side tones, while the FM waveform amplitude is always constant. This statement emphasizes the point that it is only the sinusoidal component in the FM spectrum that is at the carrier frequency, not the FM carrier itself, whose amplitude may become zero and varies from positive to negative peak values as the modulation index changes.

The ideal FM waveform frequency spectrum consists of the infinite number of harmonic tones uniformly spaced by the modulation frequency f_b (see Fig. 11.19). Therefore, the procedure for establishing the required FM waveform bandwidth involves approximate methods. One commonly

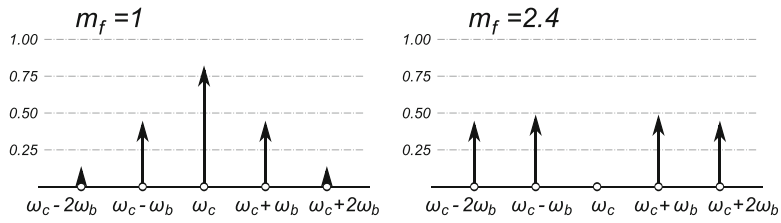


Fig. 11.19 Frequency spectrums of an FM waveform showing only the first two relevant side tones: $m_\mathrm{f} = 1$ (*left*) and $m_\mathrm{f} = 2.4$ (*right*)

used method for estimating FM waveform bandwidth is based on approximation that sets the bandwidth limits by inclusion of the highest relevant harmonics on both sides as

$$B_\mathrm{FM} = 2n f_\mathrm{b}, \tag{11.65}$$

where n is the highest order of the side frequency harmonic tone whose amplitude is significant (i.e., not negligible). By careful observation, using values from Table 11.1, it was found that if the order of the side frequency is greater than $m_\mathrm{f} + 1$, the FM waveform amplitude is within 5% of the unmodulated carrier amplitude. Using this approximation as the guide for estimating the bandwidth requirement, (11.65) is rewritten as

$$B_\mathrm{FM} = 2(m_\mathrm{f} + 1) f_\mathrm{b} = 2(\Delta f + f_\mathrm{b}). \tag{11.66}$$

Relation (11.66) is known as "Carson's rule". In order to illustrate the application of this rule, let us take a look at the following numerical examples:

- If $\Delta f = 75\,\mathrm{kHz}$ and $f_\mathrm{b} = 0.1\,\mathrm{Hz}$ then $B_\mathrm{FM} = 150\,\mathrm{kHz}$
- If $\Delta f = 75\,\mathrm{kHz}$ and $f_\mathrm{b} = 1.0\,\mathrm{kHz}$ then $B_\mathrm{FM} = 152\,\mathrm{kHz}$
- If $\Delta f = 75\,\mathrm{kHz}$ and $f_\mathrm{b} = 10\,\mathrm{kHz}$ then $B_\mathrm{FM} = 170\,\mathrm{kHz}$

Thus, although the modulation frequency changes by a factor of 100, the bandwidth occupied by the spectrum is almost constant.

Bessel functions relate the voltage amplitude of each of the sinusoidal side frequency components to the unmodulated carrier amplitude. That is

$$E_n = J_n E_\mathrm{c}, \tag{11.67}$$

where E_n is the amplitude of the n-th harmonic, J_n is Bessel's function of n-th order, and E_c is the amplitude of the carrier tone. Assuming that the amplitudes E_n and E_c are their RMS values, the power contained in the n-th sinusoidal component is given as

$$P_n = \frac{E_n^2}{R}. \tag{11.68}$$

After noticing that there is only one component at the carrier frequency and pairs of components for each frequency n, the total power in the FM waveform is simply the sum of all harmonics, i.e.

$$\begin{aligned} P_\mathrm{T} &= P_0 + 2P_1 + 2P_2 + \ldots \\ &= \frac{E_0^2}{R} + \frac{2E_1^2}{R} + \frac{2E_2^2}{R} + \ldots \\ &= \frac{J_0 E_\mathrm{c}^2}{R} + \frac{2J_1 E_\mathrm{c}^2}{R} + \frac{2J_2 E_\mathrm{c}^2}{R} + \ldots \end{aligned}$$

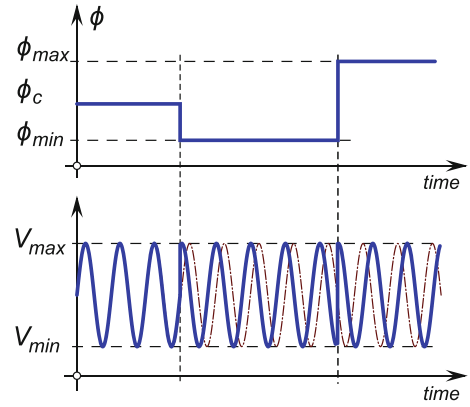


Fig. 11.20 PM waveform: Phase over time (*top*) and amplitude over time (*bottom*). The modulating signal information is clearly embedded by the control of the carrier's phase, which is shown by a *dashed line* for reference

$$= P_c \left(J_0^2 + 2\left(J_1^2 + J_2^2 + \ldots\right)\right), \tag{11.69}$$

where P_c is the power of the unmodulated carrier and J_n are Bessel's functions for the given value of modulation index m_f. Again, the total power in the modulated waveform remains constant for all values of the modulation index. This is illustrated by the fact that the sum of the squares of the Bessel function coefficients in (11.69) for a given value of m_f is always unity. For instance, if $m_f = 1.5$, the total power P_T relative to the unmodulated carrier power P_c is found using values from Table 11.1 as

$$\frac{P_T}{P_c} = 0.512^2 + 2(0.558^2 + 0.232^2 + 0.061^2 + 0.012^2 + 0.002^2) = 1.000258.$$

That is, if only the first five side harmonics are used then the rounding error is 0.026%.

11.3.2 Phase Modulation

The third method of RF carrier modulation is PM which is somewhat similar to the FM technique. In today's communication systems, it is often used for satellite and deep-space missions because, like FM, its noise properties are superior to AM but, unlike FM, it can be produced in a simple circuit driven from a frequency stable, crystal-controlled carrier oscillator. A VCO is intentionally made very variable with respect to frequency to produce high deviations and a high modulation index.

The derivation process of an analytical expression for the PM waveform (Fig. 11.20) is similar to the one used for deriving the FM waveform expression. Start with the unmodulated carrier that is given by

$$c(t) = \sin(\omega_c + \phi_c). \tag{11.70}$$

When phase-modulated, the carrier phase ϕ_c is replaced with the instantaneous phase $\phi(t)$, where

$$\phi(t) = \phi_c + K b(t), \tag{11.71}$$

where K is the phase deviation constant (analogous to k for FM) and $b(t)$ is the modulating signal. Because ϕ_c is constant, we can set its value to zero by choosing the appropriate reference point. After substituting $b(t) = Bm(t)$, where $m(t)$ is a general time-dependent function, (11.71) becomes

$$\phi(t) = \Delta\phi\, m(t), \tag{11.72}$$

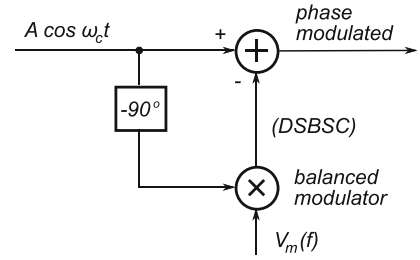


Fig. 11.21 Phase modulator block diagram

where $\Delta\phi = KB$ is the maximum phase deviation. After substituting (11.72) back into (11.70), the expression for the phase-modulated waveform is written as

$$c_{\mathrm{PM}}(t) = \sin[\omega_{\mathrm{c}} t + \Delta\phi\, m(t)] = \sin[\omega_{\mathrm{c}} t + m_{\mathrm{p}} \sin \omega_m t] \tag{11.73}$$

for sinusoidal modulation. After renaming the phase deviation $\Delta\phi$ to PM index m_{p}, a comparison of (11.73) and (11.60) shows a similarity between FM and PM schemes.

A simplified block diagram of a PM waveform generator is derived after expanding (11.73) and using the narrowband approximation $\Delta\phi < 0.25$, i.e., $\cos\Delta\phi \approx 1$ and $\sin\Delta\phi \approx \Delta\phi$. After applying these approximations, the following result is obtained (remember that cosine and sinusoidal functions are the same, except for a 90° phase difference):

$$v(t) = A\cos\omega_{\mathrm{c}} t - A\,(m_{\mathrm{p}}\cos\omega_{\mathrm{b}})\,\sin\omega_{\mathrm{c}} t\,. \tag{11.74}$$

Equation (11.74) is presented in a graphical form in Fig. 11.21, which suggests one possible block-level diagram of a simple PM transmitter implementation scheme.

11.3.3 Angle Modulator Circuits

Frequency modulation is done by means of a VCO circuit whose varicap biasing control terminal serves as the input terminal for the modulation voltage $V_{\mathrm{D}} = b(t)$. Therefore, any circuit described in Sect. 8.7 is perfectly suited for the role of an angle modulator circuit. In this section, we study two additional principles that are used in VCO circuits for modulating the frequency of the output waveform: the reactance modulator and the varicap-based phase modulator.

11.3.3.1 Reactance Modulator

The main disadvantage of varicap-based FM circuits is their narrow tuning range, which is due to the diode's very narrow small-signal operational zone. Instead, at frequencies that are not ultra high, wider tunable impedance variation can be achieved with a circuit known as "reactance modulator". Its operation is based on the intentional increase and exploitation of Miller's effect by a circuit that effectively behaves as voltage-controlled capacitive impedance (Fig. 11.22) between the output node and the ground.

Use of a FET device indicates an assumption that the gate current $i_{\mathrm{g}} = 0$, which is the first approximation used in the analysis of a reactance modulator circuit. In addition, the use of an RFC, i.e., "RF choke", inductor enables DC biasing of the FET while blocking AC currents. The equivalent schematic circuit diagram for such an arrangement is shown in Fig. 11.22 (right).

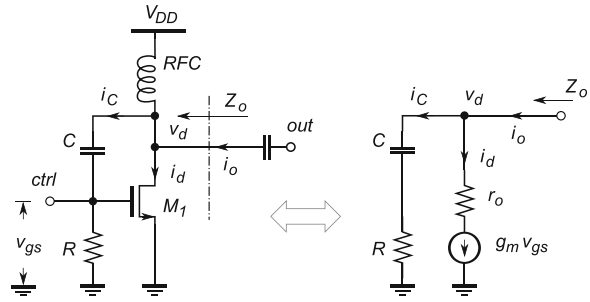


Fig. 11.22 A reactance modulator circuit and its equivalent small signal network

Effective output impedance Z_o, as seen by looking into the output node, is found by definition. That is, by the ratio of the output voltage v_o to the output current i_o, as

$$i_C = \frac{v_o}{R + Z_C} = \frac{v_o}{R - j\frac{1}{\omega C}}, \tag{11.75}$$

$$\therefore$$

$$v_{gs} = R i_C = \frac{R v_o}{R - j\frac{1}{\omega C}}, \tag{11.76}$$

which leads to an expression for M_1 drain current i_d as

$$i_d = g_m v_{gs} = g_m \frac{R v_o}{R - j\frac{1}{\omega C}}. \tag{11.77}$$

It is time to introduce the following assumptions: current i_C through the capacitor C branch needs to be much smaller than the M_1 drain current i_d, i.e $i_d \gg i_C$ or $i_d + i_C \approx i_d$; and the capacitor C impedance X_C is much greater than resistance R, that is $R - X_C \approx -X_C$.

Application of these approximation enables us to write an expression for the output current i_o, as

$$i_o = i_C + i_d \approx i_d = g_m \frac{R v_o}{R - j\frac{1}{\omega C}} \approx g_m \frac{R v_o}{-j\frac{1}{\omega C}} = \frac{v_o}{-j\frac{1}{\omega g_m RC}},$$

$$\therefore$$

$$Z_o \equiv \frac{v_o}{i_o} = -j\frac{1}{\omega (g_m RC)} = -j\frac{1}{\omega C_{RM}}, \tag{11.78}$$

where $C_{RM} = (g_m \omega RC)$ depicts the effective capacitance as seen at the output node of the reactance modulator. Through (11.78), we have shown that disposition of the output impedance Z_o of the reactance modulator is very well approximated by a tunable capacitance $C_{RM} = f(g_m)$, which is controlled by the M_1 gate-source voltage, i.e., $C_{RM} = f(v_{gs})$. Hence, the reactance modulator behaves as a voltage-controlled capacitor, which can be connected to an LC resonator tank for purposes of controlling its resonant frequency and, therefore, enabling FM. As a closing note, if the positions of the resistor R and capacitor C are swapped inside the network, the output impedance would effectively become inductive. As an exercise, the reader is encouraged to derive that expression.

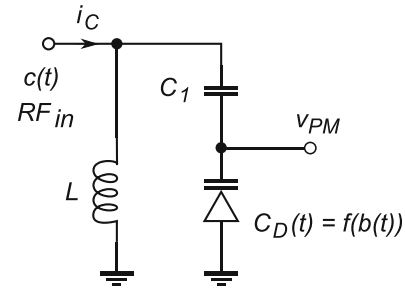


Fig. 11.23 Phase modulator circuit

11.3.3.2 Varicap Diode-Based Phase Modulator

A steady RF waveform $c(t)$ generated by a crystal oscillator with constant amplitude and phase is injected into the PM circuit whose straightforward implementation is shown in Fig. 11.23. For simplicity, the varicap biasing control voltage $b(t)$ is not shown. It is implicitly assumed that the varicap capacitance is a function of the modulation voltage, i.e., $C_\mathrm{D}(t) = f(b(t))$. The variation of diode capacitance C_D alters the phase angle of the phase modulator's tuned-circuit admittance and then the phase angle of its RF voltage.

Time-dependence of the RF waveform phase is implemented by adding voltage-controlled phase variation on the phase of the constant RF waveform, as already seen in (11.71), which is repeated here for convenience

$$\phi(t) = \phi_\mathrm{c} + Kb(t) = K\,b(t), \tag{11.79}$$

where K is the phase deviation constant, $b(t)$ is the modulating signal, and ϕ_c is the phase of the RF waveform $c(t)$. By setting $\phi_\mathrm{c} = 0$ in (11.79), the phase variation $\phi(t) = K\,b(t)$ is expressed relative to ϕ_c.

Using a procedure similar to that used in Sect. 8.7, the phase deviation constant K of the simple phase modulator (Fig. 11.23), after the derivative term is expanded into three terms, is derived as follows.

$$K = \frac{\mathrm{d}\phi}{\mathrm{d}V_\mathrm{D}} = \frac{\mathrm{d}\phi}{\mathrm{d}C}\,\frac{\mathrm{d}C}{\mathrm{d}C_\mathrm{D}}\,\frac{\mathrm{d}C_\mathrm{D}}{\mathrm{d}V_\mathrm{D}}, \tag{11.80}$$

where the total tuning capacitance C consists of the varicap capacitance C_D and capacitance C_1 in serial connection, i.e.

$$C = \frac{C_\mathrm{D}\,C_1}{C_\mathrm{D} + C_1},$$

$$\therefore$$

$$\frac{\mathrm{d}C}{\mathrm{d}C_\mathrm{D}} = \left(\frac{C_1}{C_1 + C_{\mathrm{D}0}}\right)^2 = \frac{1}{(1+n)^2}, \tag{11.81}$$

where n is the ratio of varicap capacitance $C_{\mathrm{D}0}$ to fixed capacitance C_1 at varicap biasing voltage V_0.

As we already know, admittance Y of a tuned LC circuit with dynamic resistance R_D and phase[6] is

[6] Use the Pythagorean theorem on complex numbers.

$$Y = \frac{1}{R_D} + j\left(\omega_0 C - \frac{1}{\omega_0 L}\right), \tag{11.82}$$

$$\therefore$$

$$\tan\phi = \left[R_D\left(\omega_0 C - \frac{1}{\omega_0 L}\right)\right] \approx \phi, \tag{11.83}$$

where, for small angles we applied the approximation $\tan\phi \approx \phi$, hence

$$\frac{d\phi}{dC} = \omega_0 R_D = \frac{Q}{C} = \frac{Q(C_1 + C_D)}{C_1 C_D} = \frac{Q(1+n)}{C_{D0}} \tag{11.84}$$

after substituting $R_D = Q/\omega_0 C$ into (11.84) and after applying biasing voltage V_0. We already found the sensitivity of varicap capacitance versus diode voltage, hence, after substituting (8.42), (11.84) and (11.81) into (11.80), we write

$$\begin{aligned} K &= \left[\frac{Q(1+n)}{C_{D0}}\right]\left[\frac{1}{n+1}\right]^2\left[-\frac{C_{D0}}{1+2V_0}\right] \\ &= -\frac{Q}{(1+n)(1+2V_0)}. \end{aligned} \tag{11.85}$$

Example 11.2. Estimate the PM deviation constant K for the phase modulator in Fig. 11.23 for the following data: $V_0 = 15$ V, $C = 10C_{D0}$, and $Q = 70$.

Solution 11.2. A straightforward implementation of (11.85) yields

$$K = -\frac{Q}{(1+n)(1+2V_0)} = -\frac{70}{(1+10)(1+2\times 15V)} = -0.2^{\text{rad}}/\text{v}, \tag{11.86}$$

which means that, if the biasing voltage across the varicap diode changes by 1 V, the phase of the output waveform changes by 11.46°.

11.4 PLL Modulator

By careful inspection of the PLL circuit in Fig. 11.24, we note that by adding a modulation signal $b(t)$ to the original VCO control signal v_D, we effectively push the VCO away from the reference point. If the loop bandwidth is wide enough, the loop quickly responds and generates the cancelling signal because the two phases at the input terminals of PD are forced by the loop to be equal. Hence, the total control signal is then $v_c = v_D + b(t)$ and must be constant because the input reference phase θ_{in} is constant. Therefore, the input phase $-\theta_{in}$ is proportional to v_D and θ_{out} is proportional to the modulation signal $b(t)$.

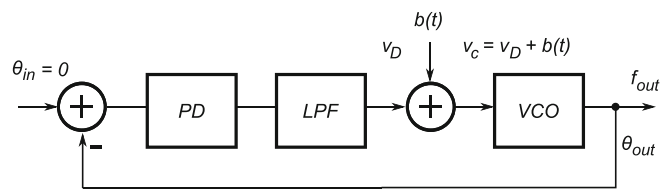


Fig. 11.24 A PLL phase modulator circuit

11.5 Summary

The three main modulation schemes, AM, FM, and PM, are briefly introduced in this chapter. Modulation is a process by which the useful baseband signal $b(t)$ (i.e., the information) is embedded into the HF single-tone signal $c(t)$ by altering its amplitude, frequency, or phase and is frequency-shifted upward and centred around the HF carrier frequency ω_0. Therefore, very loosely, the mixer circuits discussed in Chap. 9 may be considered as a special case of modulators because mixers are analyzed only from the perspective of the frequency-shifting operation (i.e., signal multiplication). Frequency upshifting is primarily done in the transmitting part of the wireless radio system. Because they are similar, frequency and PM schemes are often studied together under the common designation of "angle modulation techniques". Although modern wireless systems employ mostly FM and PM schemes, because of their lower sensitivity to amplitude noise, AM is the simplest and oldest of the three schemes and is still widely used in long-distance Earth-bound broadcasting systems.

Problems

11.1. An audio signal $A_a = 15\sin(2\pi 1{,}500t)$ modulates a carrier $A_c = 60\sin(2\pi 100000t)$.

1. Sketch the audio signal and the carrier waveform.
2. Construct the modulated wave.
3. Determine the modulation factor and percentage modulation.
4. What are the frequencies of the audio signal and the carrier?
5. What frequencies would show up in a spectrum analysis of the modulated wave?

11.2. How many AM broadcasting stations can be accommodated in a 100 kHz bandwidth if the highest frequency modulating the carriers is 5 kHz?

11.3. A large number of radio stations transmit their programs at various carrier frequencies. A radio receiver is tuned to receive amplitude-modulated waves transmitted at a carrier frequency of $f_{RF} = 980$ kHz. The LO inside the receiver is set at $f_{LO} = 1{,}435$ kHz. Estimate,

1. The frequencies coming out of the receiver's mixer.
2. Which frequency is IF.
3. The frequency of a radio station which would represent the image frequency.
4. The frequency graph of the frequencies involved.

11.4. A tuned RF amplifier has an LC tank with $Q = 100$ tuned at RF frequency f_0. Estimate the attenuation of the image signal, if the image frequency is 5% lower than the RF signal.

11.5. For a phase-modulation circuit that consists of a middle-tapped inductor in parallel with a serial combination of capacitor C and varicap diode C_{d0}, find the phase deviation constant K. Data: $Q = 70$, $C = 10C_{d0}$, $V_0 = 15$a,V.

11.6. For a FM index of $m_f = 1.5$ and modulation signal $f_b = 10$ kHz, find:

1. The estimated required bandwidth B_{FM} (using Carson's rule).
2. The ratio of the total power P_T to the power in the FM unmodulated waveform.
3. Which harmonic has the highest amplitude.

11.7. Determine the power content of each of the sidebands and of the carrier of an AM signal that has modulation of 85% and contains 1,200 W of total power.

11.8. Using the plots in Fig. 11.3, sketch the corresponding trapezoidal forms.

11.9. An AM signal whose carrier waveform is modulated 70% contains 1,500 W at the carrier frequency. Determine the power content of the upper and lower sidebands for this percentage modulation. Calculate the power at the carrier and the power of each of the sidebands when the percentage modulation drops to 50%.

11.10. An AM standard broadcast receiver is to be designed with an IF (IF) of 455 kHz.

1. Calculate the required frequency that the local oscillator f_{LO} should be at when the receiver is tuned to $f_c = 540$ kHz, if the LO tracks above the frequency of the received signal.
2. Repeat (a) if the LO tracks below the frequency of the received signal.

11.11. A $f_c = 107.6$ MHz carrier is frequency modulated by a $f_m = 7$ kHz sine wave. The resultant FM signal has frequency deviation of $\Delta f = 50$ kHz.

1. Find the carrier swing of the FM signal.
2. Determine the highest and the lowest frequencies attained by the modulated signal.
3. What is the modulation index of the FM wave?

11.12. An FM transmitter has total power of $P_T = 100$ W and modulation index of $m_f = 2.0$.

1. Find the power levels contained in all frequency components.
2. Estimate the bandwidth requirement if the modulation signal is $f_m = 1.0$ kHz.

11.13. For the circuit in Fig. 11.22, find the value of capacitor C. Data: $f_{out} = 3.5$ MHz, $C_T = 83.4$ nF, $L_T = 20$ nH, $R = 100\,\Omega$, $g_m(M_1) = 10$ mS.

Chapter 12
AM and FM Signal Demodulation

Abstract When a modulated signal arrives at the receiving antenna, the embedded information must somehow be extracted by the receiver and separated from the HF carrier signal. This information recovery process is known as "demodulation" or "detection". It is based on an underlying mechanism similar to the one used in mixers, where a nonlinear element is used to multiply two waves and accomplish the frequency shifting. However, the demodulation process is centred around the carrier frequency ω_0 and the signal spectrum is shifted downward to the baseband and returned to its original position in the frequency domain. Both modulation and demodulation involve a frequency-shifting process; both processes shift the frequency spectrum by a distance ω_0 on the frequency axis; and both processes require a nonlinear circuit to accomplish the task. Although very similar, the two processes are different in very subtle but important details. In the modulating process the carrier wave is generated by the LO circuit, and then combined with the baseband signal inside the mixer. In the demodulating process, however, the carrier signal is already contained in the incoming modulated signal and it can be recovered at the receiving point.

12.1 AM Demodulation Principles

In order to introduce the AM demodulation process analytically, let us consider a simple square law device, with one input and one output terminal, whose voltage–current characteristic is

$$i(t) = a_2\, c_{\mathrm{AM}}^2(t), \tag{12.1}$$

where a_2 is a constant and $c_{\mathrm{AM}}(t)$ is an amplitude-modulated (AM) wave of the following form

$$c_{\mathrm{AM}}(t) = C\,(1 + m\cos\omega_{\mathrm{b}} t)\cos\omega_{\mathrm{c}} t, \tag{12.2}$$

where $b(t)$ is the information baseband signal and m is the amplitude modulation index that is presented at its input terminal. Then, the output signal contains the following terms

$$\begin{aligned} i(t) &= a_2 C^2\,[1 + m\cos\omega_{\mathrm{b}} t]^2\,\cos^2\omega_{\mathrm{c}} t \\ &= \frac{a_2 C^2}{2}\,[1 + m\cos\omega_{\mathrm{b}} t]^2\,[1 + \cos 2\omega_{\mathrm{c}} t] \end{aligned}$$

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8_12,

$$= a_2 C^2 \left[\frac{1}{2} + m\cos\omega_b t + \frac{m^2}{4} + \frac{m^2}{4}\cos 2\omega_b t \right.$$
$$+ \frac{1}{2}\cos 2\omega_c t + \frac{m}{2}\cos(2\omega_c + \omega_b)t + \frac{m}{2}\cos(2\omega_c - \omega_b)t$$
$$\left. + \frac{m^2}{2}\cos 2\omega_c t + \frac{m^2}{8}\cos(2\omega_c + 2\omega_b)t + \frac{m^2}{8}\cos(2\omega_c - 2\omega_b)t \right]. \quad (12.3)$$

That is, the output spectrum contains tones at ω_b, $2\omega_b$, $2\omega_c$, $(2\omega_c + \omega_b)$, $(2\omega_c - \omega_b)$, $2(\omega_c + \omega_b)$, and $2(\omega_c + \omega_b)$, with the carrier frequency ω_c being absent. We keep in mind that there is a wide separation between the baseband frequency ω_b and the HF carrier ω_c (i.e., $\omega_c \gg \omega_b$), let alone the separation between ω_b and $n\omega_c$, or between ω_b and any other tone $(n\omega_c \pm \omega_b)$ for that matter. The point is that, even with a relatively simple LP filter, we are able to suppress all higher-frequency tones and approximate the output current signal $i(t)$ with

$$i(t) \approx a_2 m C^2 \left[\cos\omega_b t + \frac{m}{4}\cos 2\omega_b t \right], \quad (12.4)$$

which consists only of the desired information signal ω_b and its second harmonic $2\omega_b$, with DC terms (i.e., $1/2$, $m/2$, etc.) removed. It is now matter of designing an LP filter with steep frequency transfer curve so that the attenuation of the second harmonic is "good enough" relative to the first harmonic.

12.2 Diode AM Envelope Detector

There is a debate in the literature about the distinction between the terms "detector" and "demodulator" in respect to whether the diode AM envelope detector is a real demodulator or not. The argument is mostly semantic, with claims that a "true" demodulator must involve two input signals, the local carrier signal and the AM signal, not just the AM signal. Having acknowledged the argument, we proceed into analysis of the simplest possible AM envelope detector (also known as the "peak detector") for extracting information from the envelope of the AM signal. It also happens to be one of the most versatile little circuits in electronics and is used in a wide range of applications.

The diode peak detector circuit (see Fig. 12.1) has a built-in timing constant $\tau = RC$ that is fundamental to its operation. The diode serves as an ideal switch that controls the flow of the AM signal. On the positive swing of the sinusoidal input voltage V_{AM}, the diode is forward biased and the capacitor voltage V_C follows as $V_C = V_{AM}$ because the AM signal source is directly connected to the capacitor. When the input AM voltage reaches its maximum value V_m, the diode becomes reverse biased and it turns off. That is, the capacitor voltage is at $V_C = V_m$ and the capacitor is disconnected from the AC source. Hence, it starts to discharge exponentially through resistor R with timing constant $\tau = RC$. The discharging process lasts as long as $V_C > V_{AM}$, i.e., until the next upswing of the input AM voltage when the diode turns on again and the cycle repeats (see Fig. 12.2).

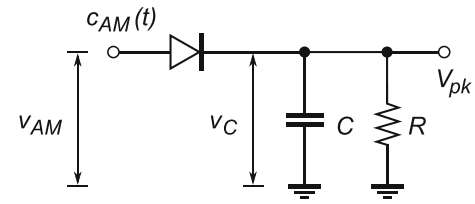


Fig. 12.1 Diode AM envelope detector

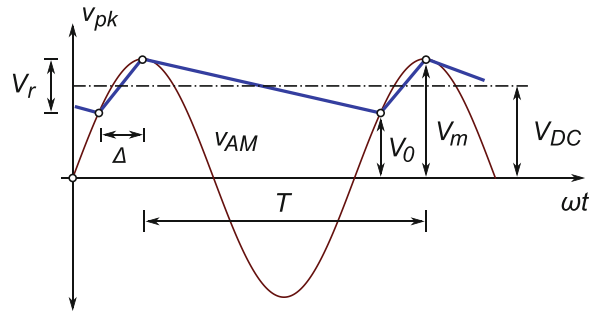


Fig. 12.2 The piece-wise approximate shape of the envelope wave as decoded by the diode AM envelope detector; the value of voltage drop V_r is exaggerated relative to the maximum amplitude V_m; in reality $V_m \gg V_r$ and $\Delta \to 0$

An exact analysis of the recovered signal wave $c_{pk}(t)$ is a bit more involved, both numerically and in using calculus, and is covered in the literature. For the purposes of our analysis, we use the approximate engineering approach, which yields reasonably accurate results.

12.2.1 Ripple Factor

With reference to Fig. 12.2, it is assumed that the output voltage V_{pk} is approximated with a linear function within the time window that is labelled as Δ in each cycle. In reality, during that time period $V_{pk} = V_{AM}$. Indeed, the approximation is valid because the maximum amplitude V_m of the two signals (V_{AM} and V_{pk}) is $V_m \gg V_r$, where V_r is the amount of voltage by which the capacitor is discharged before the diode turns again on the subsequent upswing of the input wave. Consequently, the value of the time window $\Delta \to 0$ is very small, which also means that period T of the sawtooth V_{pk} function is approximately equal to the period of the sinusoidal V_{AM} function. In addition, it is assumed that the timing constant $\tau = RC \gg T$, hence the exponential capacitor discharge[1] is approximated with the linear function within one period T time window. For the sake of clarity, in Fig. 12.2 the value of amplitude V_r is exaggerated relative to the amplitude of V_m.

With these approximations in mind, we write an expression for the average value V_{DC} of the extracted sawtooth voltage as

$$V_{DC} = V_m - \frac{V_r}{2} = I_{DC} R, \tag{12.5}$$

where I_{DC} is the average discharge capacitor current. The value of I_{DC} is approximated as follows. Starting with a fully charged capacitor C whose voltage is $V_C = V_m$, the diode is turned off and the capacitor discharge current I_{DC} is controlled by resistor R; it is assumed constant because of the $\tau \gg T$ approximation. The value of the constant discharging current is easily calculated as the initial current at the beginning of the discharging cycle when the capacitor voltage is $V_C = V_m$, hence

$$I_{DC} = \frac{V_m}{R}. \tag{12.6}$$

The RMS amplitude value of the sawtooth voltage wave is

$$V_{rms} = \frac{V_r}{2\sqrt{3}}. \tag{12.7}$$

[1] See Sect. 4.1.5.3.

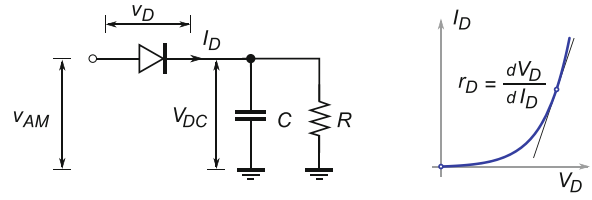


Fig. 12.3 Diode resistance r_D and variables used in an approximate analysis of diode detector efficiency

After one full time period T, the capacitor voltage has dropped by V_r (Fig. 12.2), which is controlled by the time constant $\tau = RC$ as

$$V_r = V_m - V_0 = V_m - V_m e^{-\frac{T}{RC}} = V_m \left[1 - e^{-\frac{T}{RC}}\right]$$

$$\therefore$$

$$\approx V_m \left[1 - \left(1 - \frac{T}{RC}\right)\right] = \frac{V_m}{R}\frac{T}{C} = I_{DC}\frac{T}{C} = \frac{I_{DC}}{fC}, \tag{12.8}$$

where $T = 1/f$ is the period of the carrier wave and the exponential function was approximated by (9.8), using only the linear terms. After substituting (12.8) into (12.7), we write

$$V_{rrms} = \frac{I_{DC}}{fC2\sqrt{3}}. \tag{12.9}$$

The ripple factor r_F of the extracted V_{pk} signal is defined as

$$r_F \equiv \frac{V_{rrms}}{V_{DC}} = \frac{\frac{I_{DC}}{fC2\sqrt{3}}}{I_{DC}R} = \frac{1}{fRC2\sqrt{3}}. \tag{12.10}$$

Naturally, the ripple factor reduces if:

- The frequency of the input signal is reduced—in the limiting case for DC input (of course), there is no ripple.
- A larger capacitor, which stores more charges and increases the timing constant τ, is used—in the limiting case of $C \to \infty$ the capacitor voltage never changes.
- A larger resistor is used—in the limiting case $R \to \infty$ no current flows, therefore $\tau \to \infty$ and the voltage across the capacitor never changes.

The ripple factor increases at higher frequencies or if smaller RC components are used. In most cases, (12.10) is adequately accurate, especially in the case of small ripple values.

12.2.2 Detection Efficiency

Internal diode resistance r_D causes, effectively, a voltage divider formed by the diode resistance and the load resistor R. Hence, the amplitude of the incoming AM wave v_{AM} is proportionately reduced, which is quantified by the "detection efficiency factor". A reasonably accurate analytical expression for detection accuracy may be derived after making the following assumptions. First, let us approximate the current–voltage characteristic of the diode with a linear function, Fig. 12.3 (right),

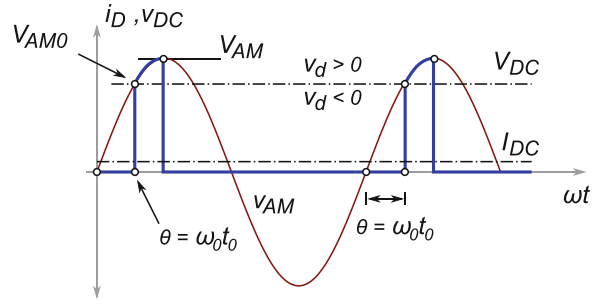


Fig. 12.4 Definitions used in the approximate analysis of diode current i_D

which is reasonable around the diode's biasing point. Second, let us assume that the voltage V_{DC} across the RC load is constant over the AM wave period, i.e., ripple factor $r_F = 0$ (see Fig. 12.3 (left) and Fig. 12.4). Third, the diode current i_D is assumed to be

$$i_D = \begin{cases} \dfrac{v_D}{r_D}, & \text{for } v_D \geq 0 \\ 0, & \text{for } v_D < 0 \end{cases}, \tag{12.11}$$

where the last two assumptions are valid for $\tau \gg RC$. At the same time, the value of DC through the diode is also calculated from the resistor load side as

$$I_D = \frac{V_{DC}}{R}. \tag{12.12}$$

The AM wave v_{AM} that enters the diode AM detector is described as

$$v_{AM} = C(1 + m\cos\omega_b t)\cos\omega_c t = V_{AM}\cos\omega_c t, \tag{12.13}$$

where $V_{AM} = C(1 + m\cos\omega_b t) = f(C, m, \omega_b, t)$ is the time-varying amplitude of the carrier wave $\cos\omega_c t$, which is introduced simply for convenience of writing the following equations, m is the amplitude modulation index, and $\cos\omega_b t$ is the modulation baseband wave.

With these assumptions, in reference to Fig. 12.4, we write an expression for voltage across the diode while in conducting mode, i.e., $v_D > 0$, as

$$v_D = v_{AM} - V_{DC} = V_{AM}\cos\omega_c t - V_{DC}, \tag{12.14}$$

$\therefore$

$$i_D = \begin{cases} \dfrac{V_{AM}}{r_D}\cos\omega_c t - \dfrac{V_{DC}}{r_D}, & \text{for } v_D \geq 0 \\ 0, & \text{for } v_D < 0 \end{cases}, \tag{12.15}$$

where r_D is the diode resistance for the given biasing point. The point in the AM signal cycle $\theta = \omega_c t_0$ corresponds to the crossover point when $v_{AM} > V_{DC}$ and the diode turns on, therefore the diode current $i_D > 0$ becomes larger than zero. That particular amplitude value when $v_{AM}(\theta) = V_{DC} = V_{AM0}$ is important for our analysis (Fig. 12.4) and we note that the following relation holds

$$V_{AM}\cos\omega_c t_0 = V_{AM}\cos\theta = V_{DC}. \tag{12.16}$$

By inspection of Fig. 12.4, we note that frequency spectrum of the instantaneous diode current i_D must contain an infinite number of harmonics because of its sharp switching characteristic, however

we will focus only on its DC and the first harmonic term at frequency ω_c. Therefore, we find the average value of the diode current I_{DC} by integrating i_D over one period, i.e., by definition

$$I_D = \frac{1}{2\pi}\int_0^{2\pi} i_D \, d\omega_c t,$$

$$\therefore$$

$$I_D = \frac{1}{\pi}\int_0^{\theta} i_D \, d\theta \tag{12.17}$$

because θ changes only within the $0, \pi$ window. Therefore, we continue the integration as

$$\begin{aligned} I_D &= \frac{1}{\pi r_D}\int_0^{\theta} (V_{AM}\cos\theta - V_{DC})\, d\theta \\ &= \frac{1}{\pi r_D}(V_{AM}\sin\theta - V_{DC}\theta), \end{aligned} \tag{12.18}$$

which, after substituting (12.16), becomes

$$I_D = \frac{1}{\pi r_D} V_{AM}(\sin\theta - \theta\cos\theta). \tag{12.19}$$

From (12.12), (12.16), and (12.19), we write

$$I_D = \frac{V_{DC}}{R} \tag{12.20}$$

$$= \frac{V_{AM}\cos\theta}{R} \tag{12.21}$$

$$= \frac{V_{AM}}{\pi r_D}(\sin\theta - \theta\cos\theta), \tag{12.22}$$

$$\therefore$$

$$\frac{r_D}{R} = \frac{1}{\pi}\frac{\sin\theta - \theta\cos\theta}{\cos\theta} \tag{12.23}$$

$$= \frac{1}{\pi}(\tan\theta - \theta), \tag{12.24}$$

which gives the ratio of the diode resistance and the load resistor (r_D/R) as a function of θ but not the two resistances by themselves.

Now we have all elements needed to define the detection efficiency η of the diode detector as the ratio of the average value of the load voltage V_{DC} relative to the peak AM input wave V_{AM} by using (12.16) as

$$\eta = \frac{V_{DC}}{V_{AM}} = \cos\theta. \tag{12.25}$$

From (12.20) and (12.22), we also write

$$\eta = \frac{R}{\pi r_D}(\sin\theta - \theta\cos\theta), \tag{12.26}$$

which provides the detection efficiency η as a function of $(R/r_D, \theta)$. It is important to note that it does not depend on V_{AM} by itself, which implies that the detection efficiency is also not a function of the

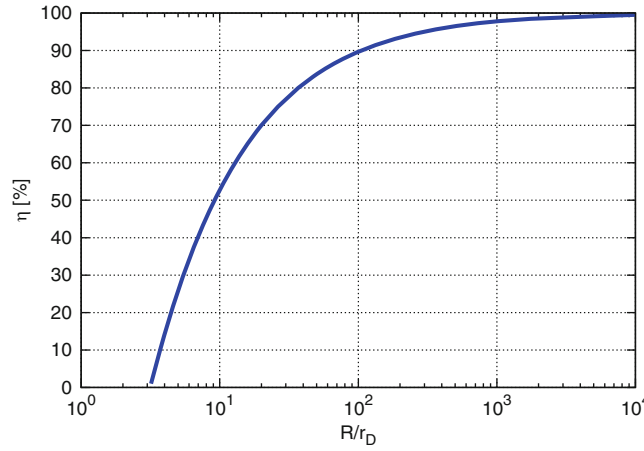


Fig. 12.5 Detection efficiency η against the R/r_{D} plot of (12.25) and (12.26)

amplitude modulation index m. These two equations, (12.25) and (12.26), provide the designer with a tool to determine the required type of diode (i.e., its resistance) and loading resistor for the desired detection efficiency. It is not easy to write explicit analytical expressions for $\eta = f(R/r_{\mathrm{D}})$; instead we use the two equations to produce a graphical relationship of $\eta = f(R/r_{\mathrm{D}})$ (see Fig. 12.5).

Example 12.1. For a given diode, whose resistance is $r_{\mathrm{D}} = 100\,\Omega$, determine the value of the loading resistor R if the desired detection efficiency for the diode AM detector is $\eta = 80\%$.

Solution 12.1. From (12.25), we find

$$\eta = \cos\theta \qquad \therefore \qquad \theta = \arccos(0.8) = 0.6435 \tag{12.27}$$

then we write

$$\frac{R}{r_{\mathrm{D}}} = \frac{\pi\,\eta}{(\sin\theta - \theta\cos\theta)} = 29.5, \tag{12.28}$$

which, for the given $r_{\mathrm{D}} = 100\,\Omega$, yields $R = 29.5 \times 100\,\Omega = 2.95\,k\Omega$.

12.2.3 Input Resistance

Similarly to any other electronic circuit, it is important to find an expression for the effective input resistance R_{eff} of the diode AM detector for given resistance R. Due to the nonlinear nature of the circuit, the most often used analytical method is based on analysis of power absorbed by the detector. In this case, we make an approximation by taking into account only the fundamental harmonic of the diode current i_{D}, whose maximum value is found by direct implementation of the Fourier coefficient as

$$\begin{aligned} I_{\mathrm{D\,max}} &= \frac{1}{\pi}\int_0^{2\pi} i_{\mathrm{D}}\cos\omega_{\mathrm{c}}t\;\mathrm{d}(\omega_{\mathrm{c}}t) \\ &= \frac{2}{\pi}\int_0^{\theta}\frac{1}{r_{\mathrm{D}}}\left(V_{\mathrm{AM}}\cos\alpha - V_{\mathrm{DC}}\right)\cos\alpha\;\mathrm{d}\alpha \end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{\pi r_{\mathrm{D}}} \int_0^{\theta} V_{\mathrm{AM}} \cos^2 \alpha \mathrm{d}\alpha - \int_0^{\theta} V_{\mathrm{DC}} \cos\alpha \, \mathrm{d}\alpha \\
&= \frac{2}{\pi r_{\mathrm{D}}} \left(\frac{1}{4} \sin 2\alpha + \frac{\alpha}{2} \right) \Bigg|_0^{\theta} - V_{\mathrm{DC}} \, \sin\alpha |_0^{\theta} \\
&= \frac{2V_{\mathrm{AM}}}{\pi r_{\mathrm{D}}} \left(\theta + \frac{1}{2} \sin 2\theta - \sin\theta \cos\theta \right) \\
&= \frac{V_{\mathrm{AM}}}{\pi r_{\mathrm{D}}} (\theta - \sin\theta \cos\theta) . \qquad (12.29)
\end{aligned}
$$

The power P dissipated in the diode and the resistor is, by definition

$$
P = \frac{1}{T} \int_0^T v_{\mathrm{AM}} \, i_{\mathrm{D}} \, \mathrm{d}t = \frac{V_{\mathrm{AM}} I_{\mathrm{D}\max}}{2} = \frac{V_{\mathrm{AM}}^2}{2\pi r_{\mathrm{D}}} (\theta - \sin\theta \cos\theta),
$$

$$
\therefore
$$

$$
\frac{R_{\mathrm{eff}}}{r_{\mathrm{D}}} \equiv \frac{V_{\mathrm{AM}}^2}{2P} = \frac{\pi}{\theta - \sin\theta \cos\theta}, \qquad (12.30)
$$

which only yields the ratio of the input resistance R_{eff} and the diode resistance r_{D} as a function of θ. In order to find out how resistor R influences the input resistance, we substitute (12.24) into (12.30) and write

$$
\frac{R_{\mathrm{eff}}}{R} = \frac{R_{\mathrm{eff}}}{r_{\mathrm{D}}} \frac{r_{\mathrm{D}}}{R} = \frac{\tan\theta - \theta}{\theta - \sin\theta \cos\theta}, \qquad (12.31)
$$

which only gives the ratio of the input effective resistance R_{eff} and resistor R as a function of θ. In the ideal case, detection efficiency is high, i.e., $\eta \to 1$, which implies very low $\theta \to 0$. We expand the sinusoidal terms into their respective power series

$$
\sin\theta = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots \approx \theta - \frac{\theta^3}{6},
$$

$$
\cos\theta = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n} = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \cdots \approx 1 - \frac{\theta^2}{2}
$$

and we take only the first two terms of the series. After substituting (12.25) into (12.31), we derive

$$
\begin{aligned}
\frac{R_{\mathrm{eff}}}{R} &= \frac{1}{\cos\theta} \frac{\sin\theta - \theta\cos\theta}{\theta - \sin\theta\cos\theta} \\
&= \frac{1}{\eta} \left[\frac{\left(\theta - \frac{\theta^3}{6}\right) - \theta\left(1 - \frac{\theta^2}{2}\right)}{\theta - \left(\theta - \frac{\theta^3}{6}\right)\left(1 - \frac{\theta^2}{2}\right)} \right] \\
&= \frac{1}{\eta} \left[\frac{\frac{\theta^3}{3}}{\frac{2\theta^3}{3} + \frac{\theta^5}{2}} \right] \approx \frac{1}{\eta} \left[\frac{\frac{\theta^3}{3}}{\frac{2\theta^3}{3}} \right] \\
&= \frac{1}{2\eta} \approx \frac{1}{2} \qquad (12.32)
\end{aligned}
$$

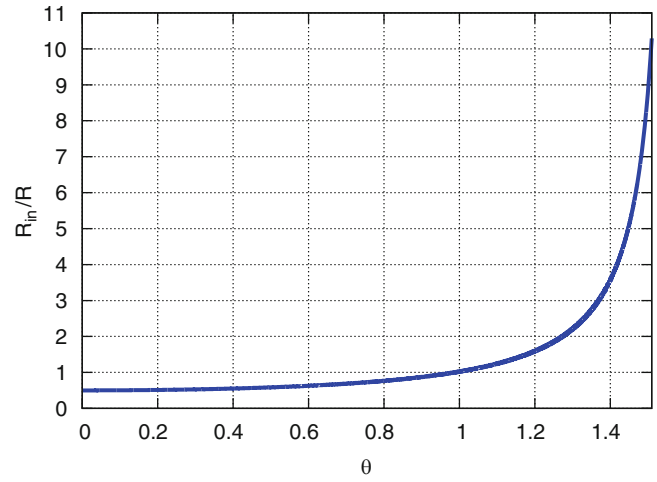


Fig. 12.6 Ratio of effective input resistance to resistor R_{eff}/R—plot of (12.31)

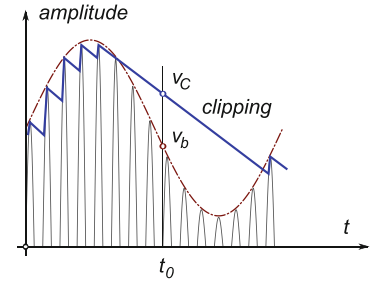


Fig. 12.7 AM wave: the wrong timing constant in the diode detector causes clipping

for the case of high detection efficiency $\eta \to 1$. Therefore, we approximate the effective input resistance as $R_{\text{eff}} \approx {}^1/_{2R}$. The detailed functional relationship of $R_{\text{eff}}/R = f(\theta)$, as derived in (12.31), is shown in Fig. 12.6.

12.2.4 Distortion Factor

So far we have used linearized approximations to estimate the parameters of a diode AM amplitude detector, which worked quite well and produced reasonably accurate expressions. One of the approximations that we made was linearization of the diode $I_{\text{D}}V_{\text{D}}$ characteristic, which is one source of distortion in the output wave due to the nonlinearity of the exponential function. For large decoding efficiency circuits and strong modulation signals, this distortion source contributes a few percentage points; for weak modulation signals, however, this source of distortion may be as high as about 25%.

The second, and less obvious, source of distortion is caused by the capacitor discharge current being constant and set by the timing constant $\tau = RC$. The problem is that the choice of timing constant is always a compromise between opposing requirements. In an AM wave, peaks and troughs of the envelope signal may come almost randomly, hence it is realistic that "clipping" may occur (see Fig. 12.7) when the timing constant is too long and the slope of the discharging current is not steep enough to follow the envelope downslope accurately. Consequently, the recovered waveform does not accurately follow the embedded AM envelope and the recovered signal is distorted.

In order to reduce the ripple factor, the timing constant needs to be long relative to the period T of the carrier signal. However, if it is made too long, clipping occurs. We need to estimate the maximum allowable value of the timing constant in order to prevent clipping and yet to make the response of the diode detector fast enough to follow the slope of the envelope signal, where the clipping factor is determined by making this compromise.

The most critical condition for the peak detector occurs when the modulation frequency ω_b is highest. The envelope wave equation is given by

$$b(t) = C_0\,(1 + m\cos\omega_\mathrm{b} t), \tag{12.33}$$

where C_0 is the maximum amplitude of the carrier signal and m is the amplitude modulation index. At any moment in time $t = t_0$, the value and slope of the modulation envelope of the modulation signal are

$$b(t_0) = C_0\,(1 + m\cos\omega_\mathrm{b} t), \tag{12.34}$$

$$\left(\frac{\mathrm{d}b(t)}{\mathrm{d}t}\right)\bigg|_{t_0} = -\omega_\mathrm{b} m C_0 \sin\omega_\mathrm{b} t. \tag{12.35}$$

By setting the potential across the capacitor equal to the modulation voltage at $t = t_0$, we write,

$$v_\mathrm{C} = C_0\,(1 + m\cos\omega_\mathrm{b} t). \tag{12.36}$$

After considering $t > t_0$, the capacitor signal decays at the following rate

$$v_\mathrm{C} = V_{\mathrm{C}0}\,\mathrm{e}^{-\dfrac{t - t_0}{RC}}, \tag{12.37}$$

$$\therefore$$

$$\left(\frac{\mathrm{d}v_\mathrm{C}}{\mathrm{d}t}\right)\bigg|_{t_0} = -\frac{1}{RC}\,v_\mathrm{C} = -\frac{C_0}{RC}\,(1 + m\cos\omega_\mathrm{b} t). \tag{12.38}$$

In order to avoid diagonal clipping, the capacitor voltage v_C must be equal to or less than the envelope voltage v_b for time $t > t_0$ and the slope must be equal to or less than the envelope slope at $t = t_0$ (which is clearly not the case in Fig. 12.7). These conditions are written as

$$-\frac{C_0}{RC}\,(1 + m\cos\omega_\mathrm{b} t) \le -\omega_\mathrm{b} m C_0 \sin\omega_\mathrm{b} t, \tag{12.39}$$

$$\therefore$$

$$\frac{1}{RC} \ge \omega_\mathrm{b}\,\frac{m\sin\omega_\mathrm{b} t}{1 + m\cos\omega_\mathrm{b} t}. \tag{12.40}$$

The fastest RC constant is at the point when

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\frac{m\sin\omega_\mathrm{b} t}{1 + m\cos\omega_\mathrm{b} t} = 0,$$

$$\therefore$$

$$m\,\omega_\mathrm{b}\,\frac{[\cos\omega_\mathrm{b} t\,(1 + m\cos\omega_\mathrm{b} t) + m\sin^2\omega_\mathrm{b} t]}{(1 + m\cos\omega_\mathrm{b} t)^2} = 0, \tag{12.41}$$

which is equivalent to the following conditions:

$$\cos\omega_b t(1+m\cos\omega_b t)+m\sin^2\omega_b t=0, \tag{12.42}$$

$$\cos\omega_b t+m(\cos^2\omega_b t+\sin^2\omega_b t)=0 \quad \therefore \quad \cos\omega_b t=-m. \tag{12.43}$$

After substituting $\cos\omega_b t=-m$ into (12.42), we give the second solution as

$$-m(1-m^2)+m\sin^2\omega_b t=0 \quad \therefore \quad \sin\omega_b t=\sqrt{1-m^2}. \tag{12.44}$$

Values of the two sinusoidal functions, (12.43) and (12.44), at this particular instance in time are substituted back into (12.40), which yields the boundary condition for the timing constant RC where the capacitor voltage has the greatest difficulty in following the modulation signal as

$$\frac{1}{RC}\geq\omega_b\frac{m\sqrt{1-m^2}}{1-m^2},$$

$$\therefore$$

$$\frac{1}{RC}\geq\omega_b\frac{m}{\sqrt{1-m^2}}, \tag{12.45}$$

which is the commonly cited condition that needs to be satisfied if the output voltage v_C is to follow the AM waveform envelope even under the worst conditions. The formula is very approximate in the sense that, for instance, it implies that for the maximum modulation index $m=1$, the RC time constant would have to be zero, which further implies that the output waveform is equal to the input AM carrier waveform, i.e., no envelope detection is possible. A more conservative condition, which was found experimentally, modifies (12.45) to

$$\frac{1}{RC}\geq m\omega_b, \tag{12.46}$$

which gives a guide to the designer in how to select the passive component values for the design of a diode AM envelope decoder.

12.3 FM Wave Demodulation

The recovery process for information embedded into an FM wave carrier is based on a two-step procedure where the frequency variation of the carrier is first converted into an amplitude variation, which is then converted back into the baseband modulation signal by conventional AM demodulators.

In principle, an FM demodulation system includes a chain of processing sub-blocks (see Fig. 12.8): a frequency-to-amplitude converter, an AM envelope detector, and an LP filter. The transfer function of a frequency-to-amplitude converter is

$$H(j\omega)=\frac{V_{AM}(j\omega)}{V_{FM}(j\omega)}, \tag{12.47}$$

$$\therefore$$

$$v_{AM}(t)=\frac{dv_{FM}}{dt}, \tag{12.48}$$

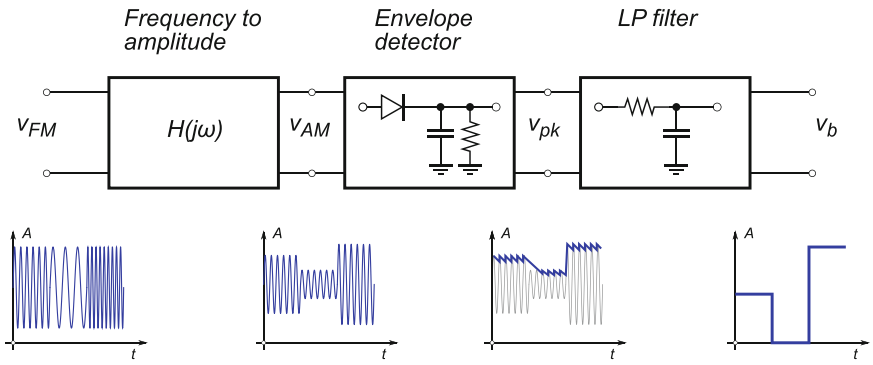


Fig. 12.8 FM wave demodulation chain

which serves as a time-domain differentiator of the FM wave. For a frequency-modulated signal at carrier frequency ω_0, we write

$$v_{\mathrm{FM}}(t) = A\cos(\omega_0 t + \theta(t)), \tag{12.49}$$

where A is the FM wave's fixed amplitude and $\theta(t)$ is the time-varying phase angle. Therefore, using (12.48), the output of the frequency-to-amplitude converter is

$$v_{\mathrm{AM}}(t) = -A\left[\omega_0 + \frac{\mathrm{d}\theta}{\mathrm{d}t}\right]\sin(\omega_0 t + \theta(t)), \tag{12.50}$$

whose amplitude portion is first approximately detected by the envelope detector as

$$v_{\mathrm{pk}}(t) = A\left[\omega_0 + \frac{\mathrm{d}\theta}{\mathrm{d}t}\right], \tag{12.51}$$

where, the first term $\omega_0 A$ is a DC component and is to be removed by the LP filter. The second term contains the embedded information signal $b(t)$ through (11.59) that may be written as

$$\theta(t) = m_{\mathrm{f}}\int b(t)\,\mathrm{d}t, \tag{12.52}$$

where m_{f} is the FM index. Therefore, output of the envelope detector (12.51) contains the information $b(t)$ that is subsequently "cleaned up" and fully recovered by the LP filter.

There are three main types of FM demodulator circuit that are reviewed in the following sections:

- Slope detectors and FM discriminators
- Quadrature detectors
- PLL demodulators

They are used to implement the general system shown in Fig. 12.8 and described by (12.47)–(12.52).

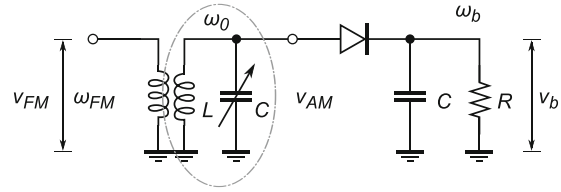


Fig. 12.9 A slope detector circuit using a simple LC resonator (*circled*) and an AM slope detector

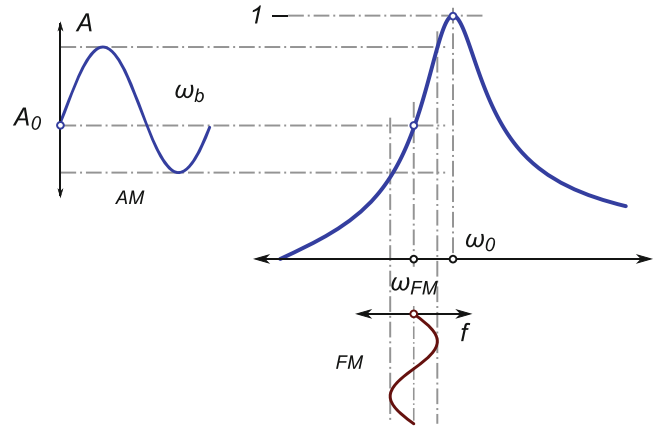


Fig. 12.10 Slope detection using a simple LC resonator

12.3.1 Slope Detectors and FM Discriminators

Although slope detectors are not much used any more, their simplicity and obvious operation allows us to easily understand the basic principle of frequency-to-amplitude conversion, therefore they serve the educational purpose well. At its core, a slope detector employs a simple LC resonator tank and a diode AM detector in series (see Fig. 12.9). Although it is a very simple network, the exact analysis of a slope detector circuit is very complicated because the input signal v_{FM} is frequency modulated and, therefore, a simple steady-state analysis does not apply; instead, transient analysis is required.

Nevertheless, we illustrate its operation graphically in Fig. 12.10. The resonant LC tank is tuned at ω_0 frequency and the carrier frequency of the incoming FM wave v_{FM} is ω_{FM} not equal to ω_0, i.e., $\omega_{\text{FM}} \neq \omega_0$. Because of that arrangement, the non-modulated tone of the incoming FM wave falls on the slope of the LC resonator's frequency characteristic, Fig. 12.10. As we discussed in Sect. 9.6.2, the vertical axis of the LC resonator frequency characteristic shows the amplitude of the incoming tone relative to the amplitude of the resonant tone at ω_0 that is normalized to one. Thus, the amplitude of the incoming tone at ω_{FM} is reduced to A_0. As the frequency of the incoming FM wave changes to $\omega_{\text{FM}} \pm \Delta\omega$, the amplitude of the recovered FM wave also changes in accordance with the slope of the LC tank characteristic. Once the FM wave passes through the resonator tank it becomes amplitude modulated in accordance with its embedded modulation signal $b(t)$, which is to say that the conventional diode AM detector is now able to extract the modulation signal $b(t)$ as usual.

We note that either side of the frequency characteristic may be used for frequency-to-amplitude conversion. On the right side of the characteristic, the increase in frequency corresponds to a decrease in amplitude; on both sides of the characteristic, the tuning range is very narrow. In addition, we also note that the recovered signal is distorted by the nonlinear characteristic. For example, the amplitude of the recovered sinusoid is not symmetrical around the A_0 point—the positive side is slightly larger (Fig. 12.10).

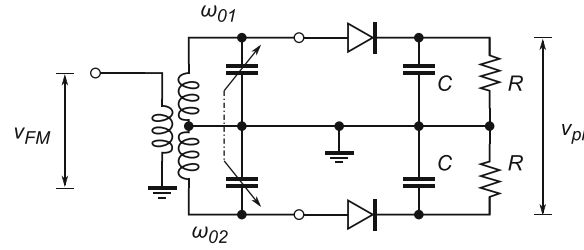


Fig. 12.11 A dual slope detector using symmetrical LC resonators and AM slope detectors

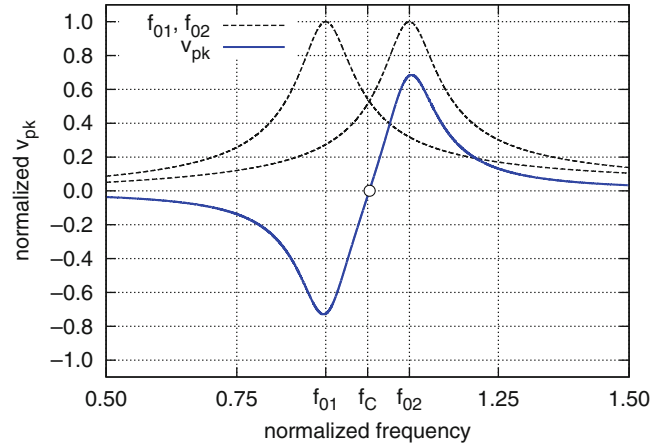


Fig. 12.12 Dual slope FM detection using two slightly offset LC resonators

12.3.1.1 Dual Slope Detector

A simple evolutionary step for improving the slope detector's performance is to design a "dual slope detector" by creating a symmetrical circuit that literally contains two mirrored versions of the slope detector (see Fig. 12.11). The resonant frequencies of the two resonators are tuned to two separate frequencies that are slightly off on each side relative to the ω_{FM} value. Due to the symmetrical topology of the circuit, i.e., signals flowing through the two sides of the circuit are opposite in phase, the newly created frequency characteristic has a wider linear region centred around the FM carrier frequency ω_{FM} (Fig. 12.12). The main strength of the dual slope FM decoder, namely the linearity that is created by the two offset resonators, is also the source of its main weakness. The circuit depends on three key frequencies (the FM carrier frequency and the two side frequencies), which means that it enables the reception of radio signals at each of these three frequencies instead of only one. In addition, it requires two variable capacitors, which further increases its complexity.

12.3.1.2 Foster–Seeley Dual Slope Detector

A modified version of a dual slope FM detector, known as "Foster–Seeley" (see Fig. 12.13), includes a shunting capacitor C_0 between the primary L_1 and the centre tap of the secondary inductance L_2 of the input transformer, a shunting capacitor C_2 across the secondary inductance L_2, and an RF choke *RFC*. It would be possible to use a resistor instead of *RFC*, however, that would reduce detection efficiency of the peak detectors (the *RFC* blocks the RF signals and provides a DC path). The input transformer is dual side tuned, i.e., both the L_1C_1 and L_2C_2 resonators are tuned to the non-modulated carrier frequency ω_0 of the incoming FM wave.

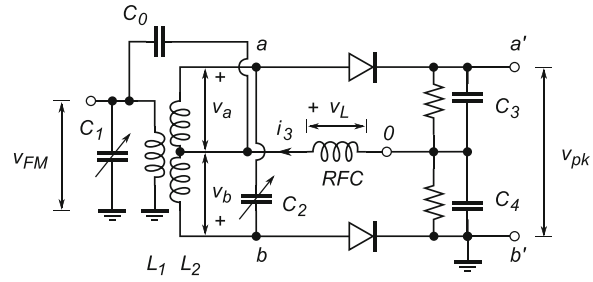


Fig. 12.13 A Foster–Seeley discriminator

A limited analysis of a Foster–Seeley decoder is possible and is educational. We start by assuming that reactances X_{C0} and X_{C4} are very small, while reactance $X_L = X_{RFC}$ are very high. With these two assumptions, we approximate the voltage across *RFC* by following the (C_0, *RFC*, C_4) path, as

$$\mathbf{v}_L = \mathrm{j}X_L\,\mathbf{i}_3 = \mathrm{j}X_L\,\frac{\mathbf{v}_{FM}}{-\mathrm{j}X_{C_0} + \mathrm{j}X_L - \mathrm{j}X_{C_4}} \approx \mathrm{j}X_L\,\frac{\mathbf{v}_{FM}}{+\mathrm{j}X_L} = \mathbf{v}_{FM}, \tag{12.53}$$

where bold letters indicate vector variables. By assuming a small mutual inductance between primary and secondary transformer inductances, we can approximate the resonant current $\mathbf{i}_1$ inside the L_1C_1 resonant tank as

$$\mathbf{i}_1 \approx \frac{\mathbf{v}_{FM}}{r_1 + \mathrm{j}X_{L1}} \approx \frac{\mathbf{v}_{FM}}{\mathrm{j}X_{L1}} \tag{12.54}$$

for a high Q primary inductor, where r_1 is the internal resistance of the L_1 inductor. This current induces EM potential $\mathbf{v}_2$ inside the secondary coil. After substituting (12.54), we write

$$\mathbf{v}_2 = \pm\mathrm{j}\omega M\,\mathbf{i}_1 = \pm\frac{M}{L_1}\,\mathbf{v}_{FM}. \tag{12.55}$$

The secondary coil is loaded by capacitor C_2, hence, after ignoring the loading effect of the diodes (and keeping the low mutual coupling approximation, i.e., the term $\omega^2M^2/r_p \approx 0$), we write an expression for the secondary resonant current $\mathbf{i}_{C2} = \mathbf{i}_{L2} = \mathbf{i}_2$ as

$$\mathbf{i}_2 \approx \frac{\mathbf{v}_2}{r_2 + \mathrm{j}X_{L2} - \mathrm{j}X_{C2}} = \pm\frac{\mathbf{v}_{FM}}{r_2 + \mathrm{j}(X_{L2} - X_{C2})}\,\frac{M}{L_1} \tag{12.56}$$

after substituting (12.55), where r_2 is the internal resistance of the L_2 inductor.

The same secondary current causes voltages $\mathbf{v}_{a0}$ and $\mathbf{v}_{b0}$ across their respective half L_2 inductance, after arbitrarily picking the positive sign of the current, as

$$\mathbf{v}_{a0} = \mathbf{i}_2\,\frac{\mathrm{j}X_{L2}}{2} = \frac{\mathbf{v}_{FM}}{r_2 + \mathrm{j}(X_{L2} - X_{C2})}\,\frac{\mathrm{j}X_{L2}M}{2L_1}, \tag{12.57}$$

$$\mathbf{v}_{b0} = -\mathbf{i}_2\,\frac{\mathrm{j}X_{L2}}{2} = \frac{\mathbf{v}_{FM}}{r_2 + \mathrm{j}(X_{L2} - X_{C2})}\,\frac{\mathrm{j}X_{L2}M}{2L_1}, \tag{12.58}$$

$$\therefore$$

$$\mathbf{v}_{ab} = \mathbf{i}_2\,X_{L2} = \frac{\mathbf{v}_{FM}}{r_2 + \mathrm{j}(X_{L2} - X_{C2})}\,\frac{\mathrm{j}X_{L2}M}{L_1}, \tag{12.59}$$

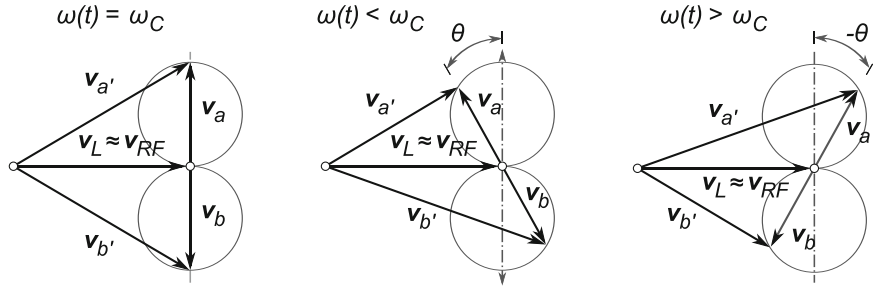


Fig. 12.14 Vector diagram for Foster–Seeley discriminator internal voltages

which means that $\mathbf{v}_{a0}$ and $\mathbf{v}_{b0}$ have 180° phase difference and $|\mathbf{v}_{ab}|$ has the same shape as in Fig. 12.12. We note that the DC potential at the output terminals $V_{a'}$ is proportional to the peak voltage V_a, while potential $V_{b'}$ is proportional to V_b, relative to the ground, therefore the total output DC voltage is

$$V_{a'b'} = V_{a'0} + V_{0b'} = V_{a'0} - V_{b'0}. \tag{12.60}$$

Now, we consider the following three cases:

- *A non-modulated FM wave, i.e., the instantaneous frequency $\omega(t)$ is equal to the FM carrier ω_0 frequency*: The secondary circuit is at resonance and the two reactances are equal $(X_{L2} = X_{C2})$, therefore, from (12.57) and (12.58), it follows that $|\mathbf{v}_{a0}| = |\mathbf{v}_{b0}|$, which after taking into account (12.60) leads to the conclusion that the output DC voltage is $V_{a'b'} = 0$.

  It is instructive to present the internal voltages on a vector diagram. At resonance, both the primary and the secondary resonators have real impedances and there is, therefore, 90° difference between voltages $\mathbf{v}_{RF}$ and $\mathbf{v}_{a0}$. At the same time, voltage $\mathbf{v}_L$ across RFC is in phase with $\mathbf{v}_{RF}$ (12.53), as shown in Fig. 12.14 (left).
- *A modulated FM wave for which the instantaneous frequency $\omega(t)$ is lower than the FM carrier ω_0 frequency*: The secondary circuit is at resonance and reactance $(X_{L2} < X_{C2})$, therefore, from (12.57) and (12.58) it follows that $|\mathbf{v}_{a0}| < |\mathbf{v}_{b0}|$ (Fig. 12.14 (centre)).
- *A modulated FM wave for which the instantaneous frequency $\omega(t)$ is higher than the FM carrier ω_0 frequency*: The secondary circuit is at resonance and reactance $(X_{L2} > X_{C2})$, therefore, from (12.57) and (12.58) it follows that $|\mathbf{v}_{a0}| > |\mathbf{v}_{b0}|$ (Fig. 12.14 (right)).

In order to find phase angle θ between vectors $\mathbf{v}_{FM}$ and $\mathbf{v}_{ab}$, i.e., between the input FM voltage and the inducted voltage across L_2, we rearrange and approximate (12.59) assuming high Q value, as

$$\begin{aligned}
\frac{\mathbf{v}_{ab}}{\mathbf{v}_{FM}} &= \frac{1}{r_2 + \mathrm{j}(X_{L2} - X_{C2})}\frac{M}{L_1} = \frac{1}{1 + \mathrm{j}\left[\frac{1}{r_2}(X_{L2} - X_{C2})\right]}\frac{\mathrm{j}X_{L2}M}{r_2 L_1} \\
&= \frac{1}{1 + \mathrm{j}\left[\frac{1}{r_2}(X_{L2} - X_{C2})\right]}\frac{\mathrm{j}X_{L2}M}{r_2 L_1} = \frac{1}{1 + \mathrm{j}\frac{1}{r_2}\left(\omega L_2 - \frac{1}{\omega C_2}\right)}\frac{\mathrm{j}X_{L2}M}{r_2 L_1} \\
&= \frac{1}{1 + \mathrm{j}\frac{\omega L_2}{r_2}\left(1 - \frac{1}{\omega^2 L_2 C_2}\right)}\frac{\mathrm{j}X_{L2}M}{r_2 L_1} = \frac{1}{1 + \mathrm{j}\frac{\omega L_2}{r_2}\left(1 - \frac{\omega_0^2}{\omega^2}\right)}\frac{\mathrm{j}X_{L2}M}{r_2 L_1}
\end{aligned}$$

$$= \frac{1}{1+\mathrm{j}\dfrac{\omega_0 L_2}{r_2}\left(\dfrac{\omega}{\omega_0}-\dfrac{\omega_0}{\omega}\right)}\frac{\mathrm{j}X_{\mathrm{L}2}M}{r_2 L_1} = \frac{1}{1+\mathrm{j}\dfrac{\omega_0 L_2}{r_2}\delta}\frac{\mathrm{j}X_{\mathrm{L}2}M}{r_2 L_1}$$
$$= \frac{1}{1+\mathrm{j}Q\delta}\frac{\mathrm{j}X_{\mathrm{L}2}M}{r_2 L_1} = \frac{1}{1+\mathrm{j}Q\delta}\frac{\mathrm{j}\omega L_2 M}{r_2 L_1}$$
$$= \frac{1}{1+\mathrm{j}Q\delta}\frac{\mathrm{j}QM}{L_1} = \frac{1}{1+\mathrm{j}Q\delta}\mathrm{j}Qk\sqrt{\frac{L_2}{L_1}}, \tag{12.61}$$

where Q is the Q factor of the secondary coil, $M = k\sqrt{L_1 L_2}$, and δ is sometimes referred to as the "detuning factor".

At resonance when $\omega = \omega_0$, then $\delta = 0$, leading to

$$\frac{\mathbf{v}_{\mathrm{ab}}}{\mathbf{v}_{\mathrm{FM}}} = \frac{\mathrm{j}\omega L_2 M}{r_2 L_1} = \mathrm{j}Qk\sqrt{\frac{L_2}{L_1}}, \tag{12.62}$$

that is, at resonance when the primary and the secondary resonators are left only with their real impedances, there is a 90° phase difference between the $\mathbf{v}_{\mathrm{ab}}$ and $\mathbf{v}_{\mathrm{FM}}$ vectors due to the C_0 capacitor. For any other case, when the instantaneous frequency is off from the resonant frequency, there will be a positive or negative phase angle θ added to the 90° average phase angle (see Fig. 12.14). Phase shift θ is small and is caused by the first term of (12.61), which is approximated[2] as

$$\theta = \arg\frac{1}{1+\mathrm{j}Q\delta} \approx \arg(1-\mathrm{j}Q\delta) = -\arctan(Q\delta) \approx -\arctan\frac{2Q\Delta\omega}{\omega_0} \tag{12.63}$$

after detuning factor δ was approximated by using substitution $\omega = \omega_0 + \Delta\omega$, as

$$\delta = \frac{\omega_0+\Delta\omega}{\omega_0} - \frac{\omega_0}{\omega_0+\Delta\omega} = \frac{2\Delta\omega}{\omega_0+\Delta\omega} \approx \frac{2\Delta\omega}{\omega_0}. \tag{12.64}$$

The last parameter that we need to define for the discriminator is its sensitivity factor k_{d}. In other words, we are interested in finding out how much change in the output DC voltage V_{ab} is generated for a unit change in frequency, i.e.

$$k_{\mathrm{d}} \equiv \frac{\mathrm{d}V_{\mathrm{ab}}}{\mathrm{d}f} \quad \left[\frac{\mathrm{V}}{\mathrm{Hz}}\right] \tag{12.65}$$

and we determine Foster–Seeley parameters using both analytical results in this section and the vector diagram.

Therefore, the overall function of the Foster–Seeley discriminator is to produce a changing DC voltage at the output terminals whose amplitude is proportional to the amplitude of the FM embedded signal. Very often in the literature, we find another version of the Foster–Seeley discriminator called a "ratio detector". With a minor tweak, the ratio detector achieves better AM rejection than the discriminator with about 6 dB (theoretically) lower sensitivity. Subsequently, a number of modified versions of the ratio detector have been designed and used.

[2] For $b \ll 1$, it follows that $b^2 \approx 0$, hence, $\dfrac{1}{1+\mathrm{j}b} = \dfrac{1}{1+\mathrm{j}b}\dfrac{1-\mathrm{j}b}{1-\mathrm{j}b} = \dfrac{1-\mathrm{j}b}{1+b^2} \approx 1-\mathrm{j}b$.

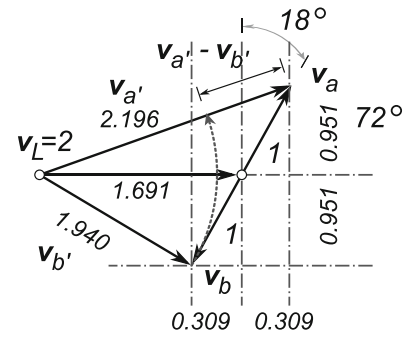


Fig. 12.15 Solution vector diagram for Example 12.2 (not to scale)

Example 12.2. One of most common IF in FM receivers is $f_0 = 10.7\,\text{MHz}$ while the maximum allowed frequency deviation from the carrier frequency is $\Delta f = 75\,\text{kHz}_{\text{pk}}$, i.e., $\Delta f = 150\,\text{kHz}_{\text{pp}}$. The internal components of a Foster–Seeley discriminator are scaled so that

$$K = \frac{QM}{2L_1} = 0.5$$

and it has $Q = 23.259$. The output voltage is measured as $V_{\text{a}} = 1V_{\texttt{rms}}$. Determine the peak output voltage V_{ab} and discriminator sensitivity.

Solution 12.2. Phase shift θ is calculated from (12.64) as

$$\theta = -\arctan\frac{2Q\Delta\omega}{\omega_0} = -\arctan\frac{2\times 23.259\times 75\,\text{kHz}_{\text{pk}}}{10.7\,\text{MHz}} = -18^\circ. \tag{12.66}$$

Knowing θ, we construct a vector diagram that contains a right triangle with 90°, 18°, and 72°, as shown in Fig. 12.15. From the $K = 0.5$ data point and (12.61), we conclude that $|\mathbf{v}_{\text{L}}| = |\mathbf{v}_{\text{a}}/K = 2$ (keep in mind that $\mathbf{v}_{\text{L}} = \mathbf{v}_{\text{RF}}$). With that information, it is straightforward to apply Pythagoras' theorem on right triangles to Fig. 12.15 and conclude that $V_{\text{ab}} = V_{\text{a}'} - V_{\text{b}'} = 2.196V_{\texttt{rms}} = \sqrt{2}\times 2.196V_{\texttt{rms}} = 3.106\,\text{VDC}$.

The calculated voltage $V_{\text{ab}} = 3.106\,\text{V DC}$ is generated by the circuit due to the maximum frequency deviation therefore $k_{\text{d}} = 3.106\,\text{VDC}/75\,\text{kHz} \approx 41.5^{\mu\text{V}}/_{\text{Hz}}$.

12.3.2 Quadrature Detector

A quadrature detector is based on a similar principle to a slope detector. However, instead of converting the incoming FM wave into an AM wave, a quadrature detector first converts an FM wave into a PM wave. This conversion is done by means of a phase-shifting network whose phase versus frequency characteristic is linear, hence variation of the carrier frequency $\Delta\omega_{\text{c}}$ creates a frequency-dependent phase shift. In principle, the function of a quadrature decoder is relatively simple. As its name implies, it is based on two signals whose phase difference is 90° (hence, they are "in quadrature"), where the incoming FM wave is split and transmitted through two separate paths. The first path is a simple short connection, while the second path leads through a phase-shifting network (see Fig. 12.16), which adds first a fixed 90° phase shift due to capacitor C_0 and then an additional

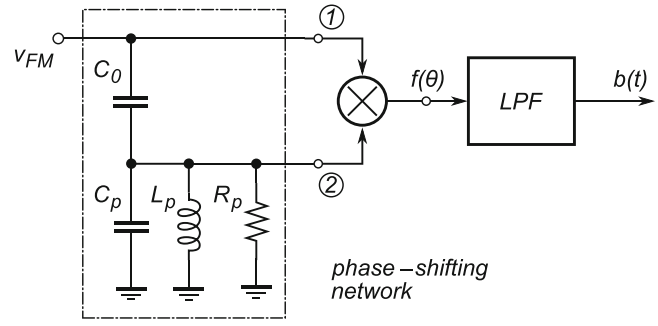


Fig. 12.16 A quadrature decoder circuit

$\theta = k_\omega \Delta\omega_0$ shift due to the $R_pL_pC_p$ resonator network. Consequently, the original FM wave arrives at node ① with its original $\theta = 0$ phase, while its copy arrives at node ② with a new phase of $\theta = \pi/2 + k_\omega \Delta\omega_0$, where k_ω is the proportionality constant.

In the following simplified analysis, we extract the term causing the frequency-dependent phase shift and show that it is indeed linear for small frequency variations. Along the way, we are not concerned about exact expressions for the amplitude of a sinusoidal wave, because it is always set by passive component values of an RLC resonant circuit and it may contain a large number of polynomial terms if calculated exactly. Instead, we focus only on the phase-altering terms. In addition, we assume high Q values, which simplifies the serial to parallel transformation of RLC resonating networks. We note that the resonant frequency of the phase-shifting network is

$$\omega_0 = \frac{1}{\sqrt{(C_0 + C_p)L_p}} = \omega_c \tag{12.67}$$

because, looking into node ②, the two capacitors appear in parallel (C_0 is connected to the AC ground through the FM signal source).

To show how the linear $\theta = f(\Delta\omega)$ phase shift characteristic is implemented, let us take a look at a simplified schematic diagram of a quadrature decoder and its phase-shifting network (Fig. 12.16). First, we note that reactance X_{C0} of capacitor C_0 and impedance Z_{RLC} of parallel $R_pL_pC_p$ resonator are in series, and they effectively create a voltage divider at node ②, hence we write

$$v_2 = \frac{Z_{RLC}}{X_{C0} + Z_{RLC}} v_1 = A_0 v_1, \tag{12.68}$$

where, term A_0 is determined by the exact set of values (C_0, R_p, L_p, C_p) of what is just another RLC resonating circuit. Exact derivation of A_0 involves a set of serial–parallel transformations and contains a number of polynomial terms. However, we already showed details of a similar derivation in (12.61), where we concluded that the transfer function A_0 of any resonant RLC network may be simplified into the following general form

$$\begin{aligned} A_0 &= \frac{jK}{1 + j\alpha} = \frac{jK}{\sqrt{1+\alpha^2}}; \quad \angle(\arctan\alpha) \\ &= \frac{K}{\sqrt{1+\alpha^2}}; \quad \angle\left(\frac{\pi}{2} + \arctan\alpha\right), \end{aligned} \tag{12.69}$$

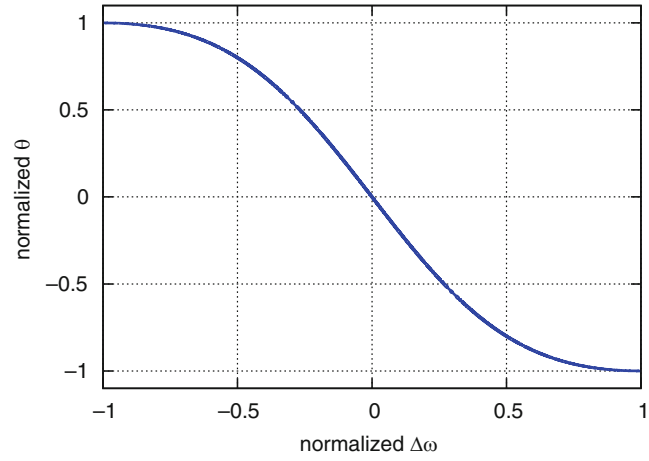


Fig. 12.17 Quadrature decoder phase characteristic showing linear conversion from frequency variations $\Delta\omega$ of an FM wave to phase $\theta(\omega)$

where $K = f(Q)$ is a real constant that controls the amplitude. However, $\alpha = f(\omega, \omega_0, Q)$ is a function of the passive component values[3] and the instantaneous frequency ω. The general form (12.69) is very handy because we can determine its amplitude and phase[4] simply by inspection. After substituting (12.69) into (12.68), we write

$$\begin{aligned} v_2 &= \frac{K}{\sqrt{1+\alpha^2}}\, v_1; \quad \angle\left(\frac{\pi}{2} + \arctan\alpha\right) \\ &= \frac{K}{\sqrt{1+\alpha^2}}\, V_1 \cos\omega_c t; \quad \angle\left(\frac{\pi}{2} + \arctan\alpha\right) \\ &= \frac{K_1}{\sqrt{1+\alpha^2}} \cos\left(\omega_c t + \frac{\pi}{2} + \arctan\alpha\right), \end{aligned} \tag{12.70}$$

where $K_1 = KV_1$ is the new amplitude proportionality constant. The time domain term affects the phase $\theta(t)$ and we need to find its average value relative to the sinusoidal variation. Therefore, we evaluate the time-dependent term by integrating it over carrier time $T/2$, as

$$\begin{aligned} I &= \frac{\omega}{\pi} \int_0^{\frac{\pi}{\omega}} \cos\left(\omega_c t + \frac{\pi}{2} + \arctan\alpha\right) \mathrm{d}t \\ &= -\frac{2}{\pi}\, \frac{\alpha}{\sqrt{1+\alpha^2}}, \\ &\therefore \\ phase(v_2) &\propto -\frac{2}{\pi}\, \frac{K_1}{\sqrt{1+\alpha^2}}\, \frac{\alpha}{\sqrt{1+\alpha^2}} = -K_2\, \frac{\alpha}{1+\alpha^2}, \end{aligned} \tag{12.71}$$

where $K_2 = (K_1 \times 2/\pi)$ is the new amplitude proportionality constant. A plot of (12.71) shows the linear phase dependence against the frequency variations (see Fig. 12.17).

[3] Keep in mind that $\omega_0 = f(RLC)$ and $Q = f(\omega_0 L, R)$.

[4] Keep in mind that $\phi = \arctan \Im/\Re$, which reduces to $\phi = \arctan \Im$ when $\Re = 1$.

Now that we have found out how the $\Delta\omega$ to phase conversion is implemented, we need to find out how the embedded information signal $b(t)$ is extracted. That extraction is done by the multiplier circuit in the frequency domain after signals v_1 and v_2 have reached its input terminals. Hence, output of the multiplier circuit, after substituting $k_\omega \Delta\omega = \arctan\alpha$ and assuming an ideal multiplier, is

$$\begin{aligned} f(\theta) = v_1 \times v_2 &= K_0 \left[\cos\omega_c t \times \cos\left(\omega_c t + \frac{\pi}{2} + k_\omega \Delta\omega\right)\right] \\ &= -K_0 \left[\cos\omega_c t \times \sin\left(\omega_c t + k_\omega \Delta\omega\right)\right] \\ &= -\frac{K_0}{2}\left[\sin\left(2\omega_c t + k_\omega \Delta\omega\right) - \sin\left(k_\omega \Delta\omega\right)\right] \\ &\approx \frac{K_0}{2}\sin\left(k_\omega \Delta\omega\right), \end{aligned} \tag{12.72}$$

where the approximation was introduced after the signal passed through the LP filter and the high-frequency tone close to $2\omega_c$ was removed from the signal spectrum. As the last step of signal recovery, we note that

$$b(t) \propto \sin\left(k_\omega \Delta\omega\right) \approx k_\omega \Delta\omega \tag{12.73}$$

for small variations of the sine argument. Therefore, for small frequency shifts, a quadrature decoder has a reasonably linear characteristic. We note that implementation of the multiplier and the LP filter is very important for operation of an analog quadrature decoder. However, if v_1 wave is digital, i.e., a square pulse stream, then a simple digital multiplier (in form of an AND gate) is employed.

12.3.3 PLL Demodulator

By careful inspection of the PLL circuit (Fig. 10.2), we note that if, instead of looking into the VCO's output node that generates either a sinusoidal or a square wave at frequency ω_0, we probe the VCO's input node ② where the (quasi) DC voltage level is generated, then without any additional circuitry we have realized a phase or a frequency demodulator. We keep in mind that the voltage at node ② is directly proportional to the change of frequency $\Delta\omega$ of the wave entering the input of the PD. If the input wave is an FM wave, then the VCO control voltage accurately tracks the FM, in other words the envelope of the modulation signal $b(t)$ that is embedded into the FM wave.

There are two slightly different cases of PLLs for FM demodulation. In the first case, when the loop bandwidth is wide enough to match the bandwidth of the modulation signal, then the PLL works as a frequency demodulator. On the other hand, if the loop bandwidth is very narrow, then the PLL is locked to the unmodulated carrier signal ω_0 so that the reference phase is averaged, that is, the phase detector holds almost constant phase that serves as a reference for comparison with the VCO phase.

12.4 Summary

Basic techniques for recovery of the received information $b(t)$ are based on a very simple diode-rectifying circuit that is a fundamental component of both AM and FM demodulators. Accurate reproduction of the envelope wave is obviously important and the amount of imperfection of the recovered information signal $b(t)$ is referred to as "distortion". When the recovered signal carries

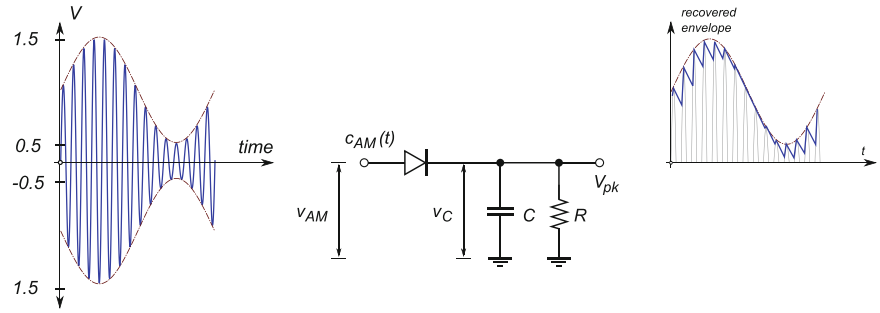


Fig. 12.18 Simplified schematic diagram for Problem 12.1

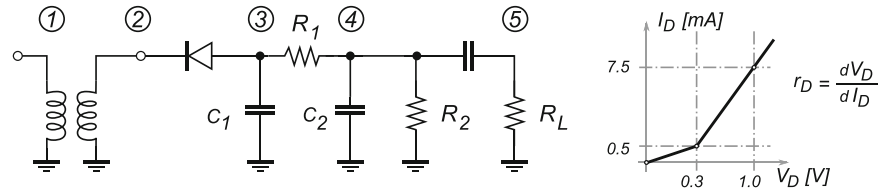


Fig. 12.19 Simplified schematic diagram and voltage current characteristics for Problem 12.2

audio information, the distortion from a low-quality demodulator is perceived by our hearing system as "bad sound". Similarly, if the recovered signal carries digital information, the distortion may cause "bit errors" if the binary signal levels shift too far from their acceptable levels, as defined by digital noise margins. Modern, more sophisticated, integrated versions of radio transceivers heavily employ PLL circuits both for modulation and demodulation of digital waves.

Problems

12.1. Using the simplified schematic in Fig. 12.18 and if $R = 2\,\text{k}\Omega$, estimate:

(a) Detector input impedance
(b) Total power delivered to the detector
(c) $v_0(max)$, $v_0(min)$, and $V_0(DC)$
(d) Average output current $I_0(DC)$
(e) An appropriate capacitor value C to prevent diagonal clipping distortion for maximal modulation frequency $f_m(max) = 5$ kHz and maximal modulation index $m_a = 0.9$

12.2. Assume that the AM diode detector in Fig. 12.19 (left) is receiving a 665 kHz IF carrier modulated with a 5 kHZ tone as the input signal V_{in}. Component values are: $C_1 = 220\,\text{pF}$, $C_2 = 22\,\text{pF}$, $R_1 = 470\,\Omega$, $R_2 = 4.7\,\text{k}\Omega$, $R_L = 50\,\text{k}\Omega$. The characteristics of diode I_D against V_D are shown in Fig. 12.19 (right).

(a) Sketch qualitatively the detector output tones along an ω axis showing relative amplitudes of the tones.
(b) Sketch the AM waveform shape at nodes ① to ⑤.
(c) Sketch the equivalent circuit at 5 kHz. Calculate the amplitude ratio of the input signal and the signal at node ③.
(d) Sketch equivalent circuit at 665 kHz. Calculate the amplitude ratio of the input signal and the signal at node ③. Comment on the result relative to the result in part (c).

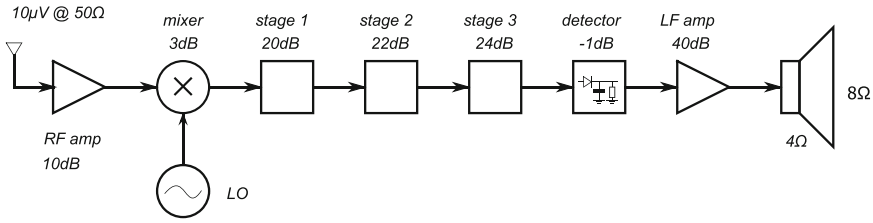


Fig. 12.20 Block diagram for Problem 12.3

12.3. The voltage signal received by a 50Ω antenna has an amplitude of 10μV (Fig. 12.20). Gain contributions are noted next to each block of the system. Estimate:

(a) The input signal power in W and dBm units.
(b) The power delivered to the speaker in dBm and W.

12.4. A modulation wave signal has a symmetrical triangular shape with zero DC component and an amplitude of $V_b = 2V_{pp}$ while the carrier wave has amplitude of $V_C = 2V_p$. Calculate the modulation index and find the ratio of the side lengths in the corresponding trapezoidal pattern.

12.5. For an unmodulated signal, the AM current in the antenna is $I_0 = 1$ A, while the sinusoidal modulation wave causes the antenna current to be $I_m = 1.1$ A. Calculate the modulation index.

Chapter 13
RF Receivers

Abstract In the general sense, a radio receiver is an electronic system that is expected to detect the existence of a single, very specific EM wave in the overcrowded air space, separate it from the rest of the frequency spectrum, and extract a message. Hence, the literal implementation of the receiver function, which is known as a TRF receiver, consists only of a receiving antenna, an RF amplifier, and an audio amplifier. In addition, advanced radio receiver versions include one or more mixers and VCO blocks, which are meant to perform either a single-step frequency down-conversion (also known as a "heterodyne receiver") or multiple step frequency down-conversions (also known as a "super-heterodyne receiver") in order to shift the HF wave down to the baseband.

In this chapter, we study basic radio receiver topologies, the nonlinear effects caused by less than ideal electronic circuitry used to implement the receiver, and receiver specification parameters.

13.1 Basic Radio Receiver Topologies

In its simplest form, a radio receiver is just an LC resonator with an envelope detector. The simplest possible implementation is known as "crystal radio" (Fig. 13.1) and it consists of the antenna (a long wire), an inductor with several taps (i.e., a quasi-tunable inductor), a diode, and high-impedance headphones. The resonance is achieved by the antenna–inductor connection.

To understand how this works, we keep in mind that, for instance, the commercial AM radio band is in the 530–1,710 kHz range, that is, the associated wavelengths are from 566 to 174 m, or equivalently 141 to 44 m quarter wavelength. Using an antenna of quarter of the wavelength ($\lambda/4$) is common practice, which means that even for the upper AM band we would need a wire at least 44 m long. Usually, we settle for a wire about 20 m long (we have to be *very careful* with the trees and houses in the neighbourhood), which means that at these frequencies the antenna is mostly capacitive. Indeed, a 20 m long antenna behaves like a 250–300 pF capacitor. Knowing the resonance equation, it is straightforward to calculated the required inductive size. The envelope detector is built of a diode and high-impedance headphones that serve as the resistive load in combination with the antenna and parasitic capacitances. Because of the high impedances within the circuit, i.e., small currents, the amount of energy collected in the antenna is sufficient to generate an audio signal in the headphones. Hence, there is no need for an external power supply, which was the reason why this kind of radio receiver was used very much by soldiers during World War I (when the radio was named a "foxhole radio").

The most direct and oldest implementation of a commercial radio receiver was based on the TRF topology (see Fig. 13.2). Although a TRF receiver may contain more than one RF tuned amplifier, each RF amplifier must be directly tuned to its carrier frequency ω_c, with subsequent stages tuned

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8_13,

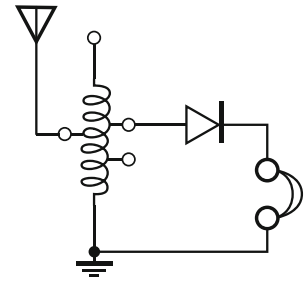


Fig. 13.1 A "crystal" radio receiver topology

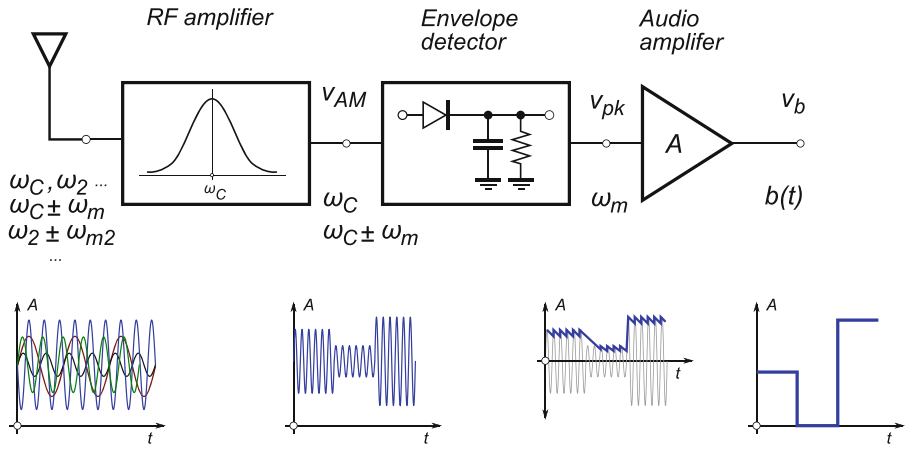


Fig. 13.2 TRF receiver topology

appropriately. This makes it a very impractical system to work with, especially with a large number of radio stations. Nevertheless, this topology was the main technology until it was replaced by the heterodyne receiver.

Tuning the resonator stage in a TRF receiver to the carrier implies that the envelope detector must decode the message at the HF carrier frequency, which means that the carrier frequency must be relatively low. In addition, a relatively wide bandwidth (i.e., low Q factor) of a single front-end LC resonator allows a number of tones to pass through and enter the envelope detector, which directly affects the overall SNR of the receiver as well as its selectivity. To make things worse, the RF amplifier gain is a function of the signal frequency, hence different carrier frequencies were received with different gains. At times when only a handful of radio stations were broadcasting, it was relatively easy to separate their carrier frequencies so that they did not interfere with each other, even with the low Q resonator tanks being used. As the radio broadcasting industry grew, the air space became more crowded and the only way to increase Q, and therefore selectivity, was to add a cascade of LC tanks (see Sect. 5.10). However, this was at a cost of increased complexity and increased effort to keep all the resonators properly tuned.

The solution was to introduce, first, the heterodyne receiver topology (with one mixer/VCO stage), as shown in Fig. 13.3, and then the super-heterodyne receiver topology (with two or more mixer/VCO stages, also known as "double conversion"). Similarly to a TRF receiver, heterodyne receivers first tune to the HF carrier frequency. However, after the RF amplifier stage separates the desired carrier frequency from the crowded frequency spectrum, the carrier-centred signal is down-shifted in frequency to some *IF* that is fixed for the given receiver. This makes it much easier to design the downstream stages: they always work at the same frequency regardless to what carrier frequency

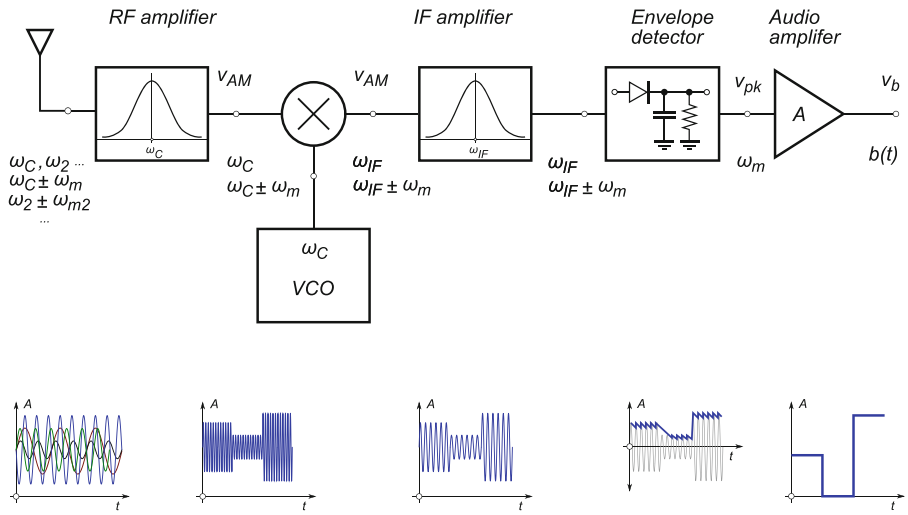


Fig. 13.3 Heterodyne receiver topology. The super-heterodyne topology contains two or more cascaded mixer/VCO stages that perform frequency down-shifting in more than one step

the RF stage is tuned to. The amount of shifting is determined by the current frequency of the local VCO, which is tuned in tandem with the RF stage.

At the system level, a radio receiver is analyzed and characterized using common metrics, so that we can compare the performance of various designs. Some of the most common parameters that are compared are:

- *Selectivity*: the minimum separation between the desired carrier frequency and its first neighbouring frequency, under the condition that the receiver can safely receive the intended signal.
- *Sensitivity*: the minimum amplitude of the incoming RF signal that the receiver can decode, under the condition of the required SNR.
- *Dynamic range*: the amplitude ratio of the strongest and weakest signals that the receiver can decode.

We establish detailed metrics for each of these parameters in the following sections, however, in the meantime, we need to familiarize ourselves with terminology and several key consequences of the fact that RF circuits are *nonlinear systems*.

13.2 Nonlinear Effects

Understanding the characterization of general systems is very important for understanding the behaviour of radio systems. Let us review basic terminology from systems theory. We loosely define a *linear system* as one in which the output signal consists of the sum of proportionally scaled input signals. In mathematical terminology, it satisfies the superposition law, i.e.,

$$F[a_1 x_1(t) + a_2 x_2(t)] = a_1 F(x_1(t)) + a_2 F(x_2(t)), \tag{13.1}$$

where a_1 and a_2 are constants independent of time. If a system does not satisfy the superposition law, then it is *nonlinear*.

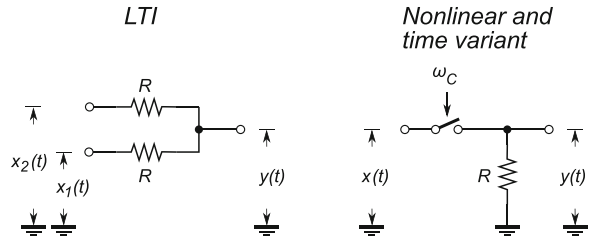


Fig. 13.4 An LTI system (*left*) and a nonlinear and time-variant system (*right*)

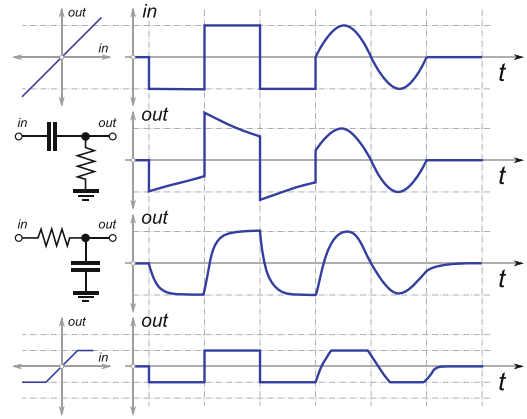


Fig. 13.5 Distorted waveforms caused by nonlinear transfer functions

A system is time invariant if a time shift in the input results in the same time shift at the output, i.e., in mathematical terminology, if $x(t) \rightarrow (t)$, then $x(t-\tau) \rightarrow (t-\tau)$ for all values of τ. Systems that are both linear and time invariant are known as "LTI systems".

We define a memoryless system as one whose output does not depend on the past values of its input. For instance, a memoryless linear system obeys the relation $y(t) = a\,x(t)$, where a is a constant. If $a(t)$ is a function of time, then relation $y(t) = a(t)x(t)$ describes a memoryless time-variant system. We can define a memoryless nonlinear system by using the general polynomial relation

$$y(t) = a_0 + a_1 x(t) + a_2 x^2(t) + a_3 x^3(t) + \cdots, \tag{13.2}$$

where a_i is constant in time (otherwise we are defining a time-variant memoryless nonlinear system). Clearly, if in practice all terms in (13.2) disappear (or are negligibly small) except the first two, then the linear approximation of $y(t)$ would be valid. Figure 13.4 shows two networks, one that is LTI and one that is both nonlinear and time variant. Note that, in Fig. 13.4 (right), the switch itself is a nonlinear element: because of the dependence of the output variable $y(t)$ on the switching frequency ω_c, time invariance is broken. General radio systems are analyzed using (13.2) where a_i is a constant because they are approximated as memoryless time-invariant nonlinear systems.

When the output amplitude is not a linear function of the input amplitude, we describe it as "amplitude distortion". In general, distortion is any difference between the original and output forms of the signal (see Fig. 13.5). The following effects caused by nonlinearity of the transfer characteristic are most commonly studied: harmonic distortion, gain compression, intermodulation, and desensitization.

13.2.1 Harmonic Distortion

After injecting a single-tone signal $x(t) = B\cos\omega t$ into a nonlinear system whose transfer function is described as (13.2) with the DC term removed (i.e., $a_0 = 0$), the output shows at the output node as

$$\begin{aligned} y(t) &= a_1 B\cos\omega t + a_2 B^2\cos^2\omega t + a_3 B^3\cos^3\omega t + \cdots \qquad (13.3)\\ &= a_1 B\cos\omega t + \frac{a_2 B^2}{2}(1+\cos 2\omega t) + \frac{a_3 B^3}{4}(3\cos\omega t + \cos 3\omega t) + \cdots \\ &= \frac{a_2 B^2}{2} + \left(a_1 B + \frac{3a_3 B^3}{4}\right)\cos\omega t + \frac{a_2 B^2}{2}\cos 2\omega t + \frac{a_3 B^3}{4}\cos 3\omega t + \cdots \\ &= b_0 + b_1\cos\omega t + b_2\cos 2\omega t + b_3\cos 3\omega t + \cdots, \qquad (13.4) \end{aligned}$$

where b_0 is the output signal's DC term. In addition, we make the following observations: the input signal spectrum $x(\omega)$ contains only one tone ω, while the output signal spectrum $y(\omega)$ contains higher-order harmonics 2ω, 3ω, etc. that did not exist in the input spectrum. They have been created by the nonlinear transfer function. Even-order harmonics (i.e., terms with $2\omega t$, $4\omega t$, ...) are associated with the even-order constants $a_i, i = 2k$, therefore these terms disappear if the system transfer function has odd symmetry, for instance, transfer functions of differential circuits. For large amplitude, $B \gg 1$ the n-th harmonics is approximately proportional to B^n. These are very important observations that give clues about the frequency spectrum of a nonlinear system.

One of the commonly used quantitative measures of nonlinearity is "total harmonic distortion" (THD). The individual percentage distortions are calculated as

$$D_2 = \frac{b_2}{b_1} \quad D_3 = \frac{b_3}{b_1} \quad D_4 = \frac{b_4}{b_1} \quad \cdots \qquad (13.5)$$

relative to the first harmonic coefficient. Then, by definition, we calculate THD for voltage or current as

$$THD = \sqrt{D_2^2 + D_3^2 + D_4^2 \cdots}. \qquad (13.6)$$

Example 13.1. A cosine current was measured at the output of a non-inverting amplifier. The three experimentally determined pairs of the input voltage V_{in} and the matching output current I_{out} are: $V_{\max} \Rightarrow I_{\max} = 1\,\text{mA}$, $V_{\text{b}} \Rightarrow I_{\text{b}} = 0.01\,\text{mA}$, $V_{\min} \Rightarrow I_{\max} = -0.95\,\text{mA}$ where V_{b} is the biasing voltage at the midpoint between the maximum and minimum input voltage amplitudes. Based on the available data, estimate the THD of the system.

Solution 13.1. The collected experimental data correspond to the cosine wave input voltage function (the non-inverting amplifier), therefore we know the associated ωt angles.[1] After substitution back into (13.4), this results in:

$$\begin{aligned} &V_{\text{in}} = V_{\max} \quad \therefore \quad \omega t = 0 \quad \therefore \quad I_{\max} = b_0 + b_1 + b_2,\\ &V_{\text{in}} = V_{\text{b}} \quad \therefore \quad \omega t = \frac{\pi}{2} \quad \therefore \quad I_{\text{b}} = b_0 - b_2,\\ &V_{\text{in}} = V_{\min} \quad \therefore \quad \omega t = \pi \quad \therefore \quad I_{\max} = b_0 - b_1 + b_2, \end{aligned}$$

[1] Simply a plot cosine function and find arguments for its maximum, minimum, middle, and $\pm 1/2$ amplitude points.

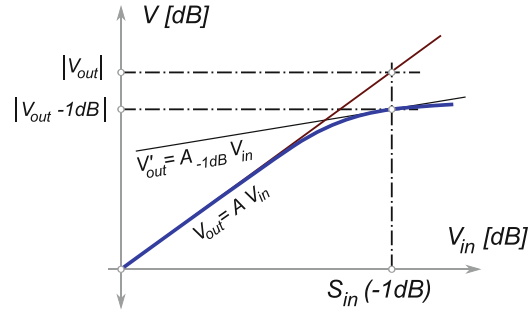


Fig. 13.6 Output signal level against input signal level and the 1 dB compression point

which is solved as

$$b_0 = \frac{I_b}{2} + \frac{I_{\max}}{4} + \frac{I_{\min}}{4} = 17.5\,\mu\text{A},$$
$$b_1 = \frac{I_{\max}}{2} - \frac{I_{\min}}{2} = 975\,\mu\text{A},$$
$$b_2 = -\frac{I_b}{2} + \frac{I_{\max}}{4} + \frac{I_{\min}}{4} = 7.5\,\mu\text{A}.$$

By definition, we write

$$D_2 = \frac{b_2}{b_1} \times 100\% = 0.77,$$
$$THD = \sqrt{D_2^2 + D_3^2 + D_4^2 \cdots} = \sqrt{D_2^2} = D_2 = 0.77\%$$

because with only three measurements we can solve for up to the second order term in (13.4). If more detailed measurement was done, for instance with five measured points that would add amplitudes at $V_{\text{in}}(\pm 1/2)$, then we would have $\omega t = \pi/3$ and $\omega t = 2\pi/3$ corresponding angles as well, which would enable us to calculate b_0, b_1, b_2, b_3, and b_4 constants.

13.2.1.1 Gain Compression

A common property of most amplifier circuits is that as the input signal power level increases, at first the output signal level increases proportionally. That is, for low-power signals, the output–input relationship is linear $P_{\text{out}} = A\,P_{\text{in}}$, where A is the gain that is calculated as the $A = dP_{\text{out}}/dP_{\text{in}}$ derivative. However, eventually, the output signal level is limited by the circuit's power supply level or the reduced biasing current of its active devices. In other words, the small signal linearity relationship does not hold for large input signal levels.

We define the *1 dB compression point* as the input signal power level $S_{\text{in}}(-1\,\text{dB})$ which corresponds to the gain $A_{(-1\,\text{dB})}$ for which the output signal level is 1 dB lower relative to the linear model (see Fig. 13.6, noting that the plot is in log–log scale).

The 1 dB compression point is determined both analytically and experimentally. Let us take a nonlinear system described by (13.4) and try to find the 1 dB compression point. The first term in (13.4) is the DC term, hence its derivative is zero and it is not part of the gain equation. The second term describes the output signal of the input $x(t) = A\cos\omega t$, hence we write the equations for the

linear gain function (i.e., if all nonlinear terms in (13.3) are ignored) and the nonlinear gain function as

$$
\begin{aligned}
|V_{\text{out}}| &\approx a_1 B \cos \omega t, \\
|V'_{\text{out}}| &\approx \left(a_1 B + \frac{3 a_3 B^3}{4} \right) \cos \omega t, \\
&\therefore \\
\left| \frac{V'_{\text{out}}}{V_{\text{out}}} \right| &= \left(1 + \frac{3 a_3 B^2}{4 a_1} \right),
\end{aligned}
$$

where the ratio $V'_{\text{out}}/V_{\text{out}}$ is the apparent gain between the linear and the nonlinear functions. Clearly, if $a_3/a_1 < 0$ and $\left| \frac{3 a_3 B^2}{4 a_1} \right| < 1$ then there is compression in the gain. After conversion into the dB scale,[2] we write an expression for the apparent gain as

$$
\begin{aligned}
20 \log V'_{\text{out}} - 20 \log V_{\text{out}} &= 20 \log \left(1 + \frac{3 a_3 B^2}{4 a_1} \right), \\
-1\,\text{dB} &= 20 \log \left(1 + \frac{3 a_3 B^2}{4 a_1} \right), \\
10^{-1/20} - 1 &= \frac{3 a_3 B^2}{4 a_1}, \\
&\therefore \\
B(-1\,\text{dB}) &= \sqrt{0.145 \left| \frac{a_1}{a_3} \right|},
\end{aligned}
\tag{13.7}
$$

where the input signal level $S_{\text{in}}(-1\,\text{dB})$ was introduced in dB, therefore

$$
S_{\text{in}}(-1\,\text{dB}) = 20 \log \left[B(-1\,\text{dB}) \right] \ \text{dB}. \tag{13.8}
$$

Interestingly enough, (13.7) shows that the 1 dB compression point of the first harmonic is, through a_3, intimately connected to the third harmonic of the input signal. We formalize this connection in the following sections.

13.2.2 Inter-Modulation

As opposed to harmonic distortion, which is caused by self-mixing of one input signal and where the higher-order harmonics in (13.4) are relatively easy to suppress by LP filtering, *intermodulation* involves two input tones with close frequencies ω_{a} and ω_{b}. Consequently, in case of any nonlinearity the output spectrum must contains various harmonics of the fundamental tones, however, it also contains tones that are not harmonics of the input frequencies.

[2] Keep in mind that $\log a/b = \log a - \log b$.

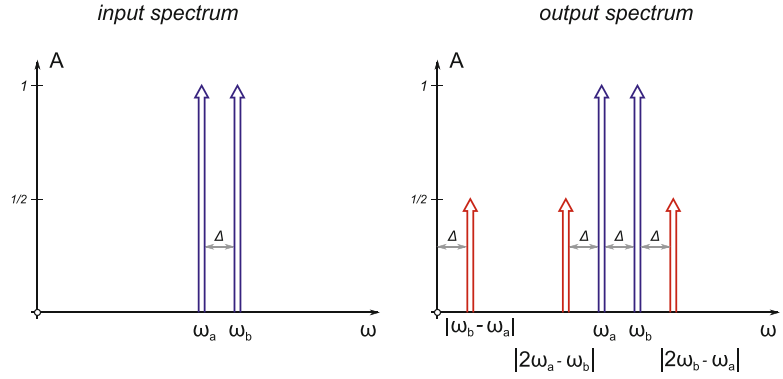


Fig. 13.7 Part of the intermodulation frequency spectrum showing the third-order terms $2\omega_a \pm \omega_b$ close to the fundamental tones

Let us assume the input signal is the sum of $x(t) = B_1 \cos \omega_a t + B_2 \cos \omega_b t$; then (13.3) becomes

$$\begin{aligned} y(t) = {} & a_1 \left(B_1 \cos \omega_a t + B_2 \cos \omega_b t\right) \\ & + a_2 \left(B_1 \cos \omega_a t + B_2 \cos \omega_b t\right)^2 \\ & + a_3 \left(B_1 \cos \omega_a t + B_2 \cos \omega_b t\right)^3 + \cdots , \end{aligned} \tag{13.9}$$

which, after expanding and collecting the frequency terms, yields the following terms

$$\begin{aligned} y(t) = {} & \frac{a_2(B_1^2 + B_2^2)}{2} && \textit{(DC term)} \\ & + \left(a_1 B_1 + \frac{3}{4} a_3 B_1^3 + \frac{3}{2} a_3 B_1 B_2^2\right) \cos \omega_a t && \textit{(fundamental terms)} \\ & + \left(a_1 B_2 + \frac{3}{4} a_3 B_2^3 + \frac{3}{2} a_3 B_2 B_1^2\right) \cos \omega_b t \\ & + \frac{a_2}{2} \left(B_1^2 \cos 2\omega_a t + B_2^2 \cos 2\omega_b t\right) && \textit{(second-order terms)} \\ & + a_2 B_1 B_2 \left[\cos(\omega_a + \omega_b)t + \cos|\omega_a - \omega_b|t\right] \\ & + \frac{a_3}{4} \left(B_1^3 \cos 3\omega_a t + B_2^3 \cos 3\omega_b t\right) && \textit{(third-order terms)} \\ & + \frac{3a_3}{4} \left\{B_1^2 B_2 \left[\cos(2\omega_a + \omega_b)t + \cos(2\omega_a - \omega_b)t\right]\right. \\ & \qquad \left. + B_1 B_2^2 \left[\cos(2\omega_b + \omega_a)t + \cos(2\omega_b - \omega_a)t\right]\right\} , \end{aligned} \tag{13.10}$$

which shows that the output spectrum contains the two fundamental tones, ω_a, ω_b, the second-order terms, $2\omega_a$, $2\omega_b$, $|\omega_a \pm \omega_b|$, and the third-order terms $3\omega_a$, $3\omega_b$, $2\omega_a \pm \omega_b$, $2\omega_b \pm \omega_a$. It is of particular interest that we found third-order tones such as $2\omega_a \pm \omega_b$ that are not harmonics of the fundamental tones. The problem is that if the two input tones are close to each other, i.e., $\omega_a \approx \omega_b$, then $2\omega_a - \omega_b \approx 2\omega_a - \omega_a \approx \omega_a$! That is, there are cases in the frequency spectrum when the third-order terms are too close to the fundamental tones, Fig. 13.7 and cannot easily be filtered out.

Frequency spectrum analysis (13.10) comes in handy for the "two-tone test" that uses two slightly different tones with the same small amplitude $B_1 = B_2 = B$, which means that the higher harmonics are negligible and (13.10) simplifies to

$$
\begin{aligned}
y(t) = a_2 B + B\left(a_1 + \frac{9}{4}a_3 B^2\right)\cos\omega_a t + B\left(a_1 + \frac{9}{4}a_3 B^2\right)\cos\omega_b t \\
+ \frac{B^2 a_2}{2}\left(cos2\omega_a t + \cos 2\omega_b t\right) + a_2 B^2\left[\cos(\omega_a + \omega_b)t + \cos|\omega_a - \omega_b|t\right] \\
+ \frac{B^3 a_3}{4}\left(\cos 3\omega_a t + \cos 3\omega_b t\right) \\
+ \frac{3B^3 a_3}{4}\left\{\left[\cos(2\omega_a + \omega_b)t + \cos(2\omega_a - \omega_b)t\right]\right. \\
\left. + \left[\cos(2\omega_b + \omega_a)t + \cos(2\omega_b - \omega_a)t\right]\right\}. \qquad (13.11)
\end{aligned}
$$

With an assumption of a small amplitude B, i.e., $B^2 \to 0$, the amplitudes of the fundamental terms are approximated as

$$\left(a_1 + \frac{9}{4}a_3 B^2\right) \approx a_1. \qquad (13.12)$$

We are especially interested in the power of tones at $(2\omega_b \pm \omega_a)$ relative to the power of the fundamental tones. In a similar fashion to the derivation of the 1 dB compression point, let us take a look at the input signal level that causes the power of the third-order term to be equal to the power of the fundamental, i.e.

$$
\begin{aligned}
a_1 B &= \frac{3B^3 a_3}{4}, \\
&\therefore \\
B(IIP3) &= \sqrt{\frac{4}{3}\left|\frac{a_1}{a_3}\right|}, \qquad (13.13)
\end{aligned}
$$

where the amplitude of the fundamental was approximated as (13.12) and $B(IIP3)$ refers to the input signal level known as the *third-order intercept point* (IIP3). By comparing (13.13) with (13.7), it is straightforward to write

$$B(-1\,\text{dB}) = \sqrt{\frac{4}{3}\left|\frac{a_1}{a_3}\right| 0.11} = IIP3 - 9.6\,\text{dB}. \qquad (13.14)$$

We note that IIP3 gives nonlinearity because of the third-order terms and that the initial assumption was that the two input tones had small amplitude. That is, expression (13.13) is not valid for strong signals. Because of that, IIP3 is the theoretical point that is extrapolated from the linear portions of the gain plot, shown in Fig. 13.8. It is interesting to note that the slope of the third-order term is three times the slope of the fundamental. This observation leads to a graphical solution for the third-order IIP3 from the experimental data (Fig. 13.9). The input power of the fundamental tones is measured and compared with the power of the third-order term on a spectrum analyzer, where the power difference ΔP is in dB, Fig. 13.9 (left), which is translated into the I/O power plot, Fig. 13.9 (right). Using similar triangles, we conclude that the IIP3 point must be at

$$IIP3 = P_{\text{in}} + \frac{\Delta P}{2} \quad \text{dB}, \qquad (13.15)$$

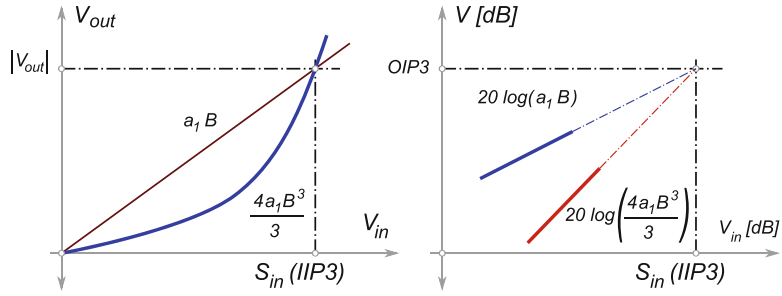


Fig. 13.8 Third-order intercept point extrapolation

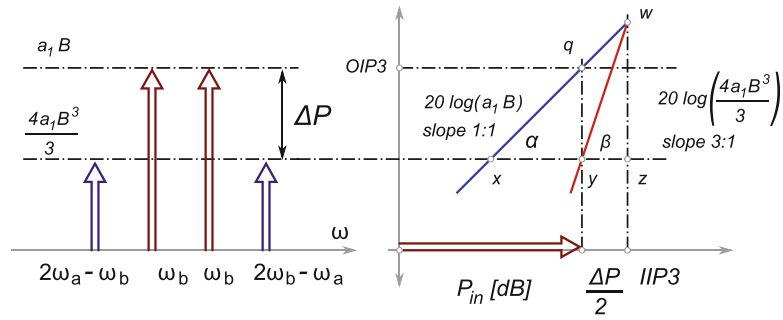


Fig. 13.9 Graphical solution for third-order intercept point

which is a practical way of estimating the IIP3 by measurement. In this analysis, we have ignored any effects of the second-order terms. They have less influence in narrowband systems relative to the third-order terms, however in the case of low IF or direct conversion systems, the second-order terms come very close to the baseband signal. If not taken care of, they may even "overwrite" the desired tone. More detailed study of intermodulation terms is beyond the scope of this book.

13.2.3 Cross-Modulation

There are two important cases of cross-modulation that we need to become familiar with. In the first scenario, two signals arrive at the antenna, one much stronger than the other. The problem is that the desired signal is the "weak" one. As an illustration, imagine using a cell phone in a crowded bus with another cell phone user very close by. The signal leaving the neighbouring cell phone is very strong, but unfortunately it is not for you. The one that you are trying to hear is already at the end of its journey and is very weak, barely dumping its leftover energy into the antenna. Unfortunately for you, the other user is doing the same and your signal may be "blocked" or "jammed".

Let us take a closer look at the case from the mathematical perspective. The incoming signal

$$x(t) = B_1 \cos \omega_a t + B_2 \cos \omega_b t; \qquad B_2 \gg B_1 \tag{13.16}$$

is processed by a nonlinear circuit whose gain equation is given by (13.2), which, after substitution of (13.16), becomes

$$\begin{aligned} y(t) &\approx \left(a_1 B_1 + \frac{3}{4} a_3 B_1^3 + \frac{3}{2} a_3 B_1 B_2^2\right) \cos \omega_a t + \cdots, \\ &(B_2 \gg B_1) \\ &\approx \left(1 + \frac{3}{2} \frac{a_3}{a_1} B_2^2\right) a_1 B_1 \cos \omega_a t + \cdots, \end{aligned} \tag{13.17}$$

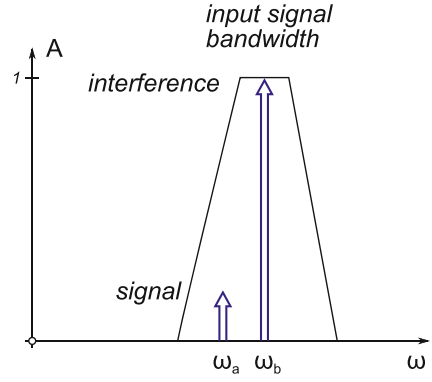


Fig. 13.10 Strong interference and weak signal in the same band

where, we focus only on the first fundamental term of the desired tone at ω_a. Most circuits are compressive, therefore $a_3/a_1 < 0$, which leads to the conclusion that, under the right circumstances and large amplitude B_2 of the blocking signal, the amplitude of the desired signal ω_a may be reduced to zero, i.e.

$$0 = \left(1 - \frac{3}{2}\left|\frac{a_3}{a_1}\right| B_2^2\right) \quad \therefore \quad B_2 = \sqrt{\frac{2}{3}\left|\frac{a_1}{a_3}\right|}. \tag{13.18}$$

Modern RF equipment is expected to correctly decode the desired signal in presence of an interfering signal that may be 60–70 dB stronger.

In the second scenario (see Fig. 13.10), the receiving antenna is exposed to two signals, the desired one at frequency ω_a and a strong AM signal, i.e.

$$x(t) = B_1 \cos \omega_a t + B_2(1 + m \cos \omega_b t) \cos \omega_c t. \tag{13.19}$$

Using the same approach again, we focus only on the main harmonic of the desired signal, i.e.

$$\begin{aligned} y(t) &\approx \left[a_1 B_1 + \frac{3}{2} a_3 B_1 B_2^2 \left(1 + \frac{m^2}{2} + \frac{m^2}{2} \cos 2\omega_b t + 2m \cos \omega_b t\right)\right] \cos \omega_a t + \cdots \\ &= f(\omega_b, 2\omega_b) \cos \omega_a t \end{aligned} \tag{13.20}$$

in other words, the receiving signal is modulated by the AM signal, which is superimposed on the original message. Depending on the exact circumstances, the desired signal may be completely blocked by the strong AM signal.

13.2.4 Image Frequency

The main limitation of a TRF receiver, its limited selectivity over a wide range of receiving frequencies, was a strong motivation for development of heterodyne receiver topology. Even though it is much more complicated than the simple TRF receiver structure, advances in IC technology enable very sophisticated heterodyne and super-heterodyne receivers to be manufactured as a sub-circuit of even more complex communication integrated systems. Indeed, it is a standard expectation for modern equipment to have one or more integrated RF transceivers included for a fractional increase in the overall cost.

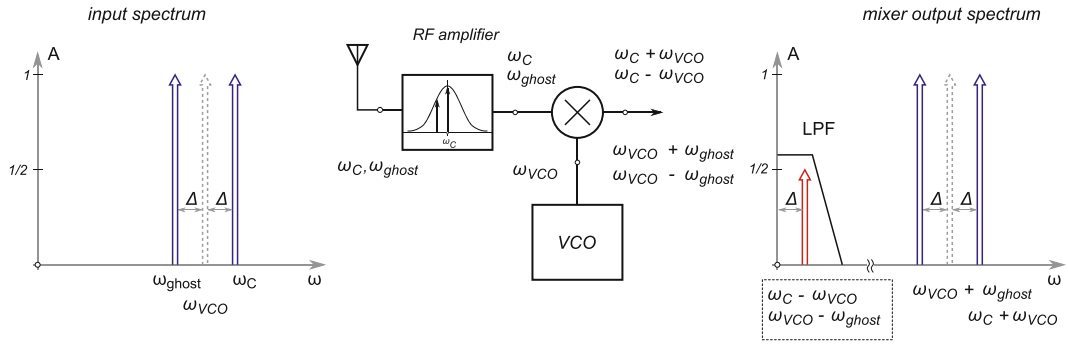


Fig. 13.11 The frequency domain relation among the carrier frequency ω_c, the image frequency ω_{ghost}, the LO frequency ω_{VCO}, and the sum and difference tones generated by mixer. This illustration assumes that $\omega_{VCO} < \omega_c$. If $\omega_{VCO} > \omega_c$, then the roles of the carrier and the image frequency are swapped

However, the solution to the selectivity problem, which was enabled by the addition of a VCO–mixer combination, comes with its own issue, known as the "image frequency", which is sometimes referred to as a "ghost frequency". This inherent issue comes from the fact that a mixer generates two tones, $\omega_a \pm \omega_b$, at its output terminal (see Fig. 13.11). In order to see how the ghost frequency issue arises, let us take a look at the following scenario. Let us say that an audio signal with $f_m = 1\,\text{kHz}$ is embedded into a carrier signal, $f_C = 10\,\text{MHz}$. At the receiving side, the LO is tuned to $f_{VCO} = 9.999\,\text{MHz}$. Routinely, we state that the frequency spectrum at output of the ideal mixer must contain

$$\begin{aligned} f_1 &= f_C + f_{VCO} = 10\,\text{MHz} + 9.999\,\text{MHz} = 19.999\,\text{MHz}, \\ f_2 &= f_C - f_{VCO} = 10\,\text{MHz} - 9.999\,\text{MHz} = 1\,\text{kHz}, \end{aligned} \tag{13.21}$$

where $f_2 = 1\,\text{kHz}$ is the desired signal, and $f_1 = 19.999\,\text{MHz}$ is the high-frequency tone that is easily removed by an LP filter. However, a more careful analysis reveals that, in the case of another signal arriving at the receiving antenna, we may have the following scenario. Let us take a look at the frequency that is located, in this case, at two times the modulation frequency f_m below the carrier frequency, i.e.

$$f_{ghost} = f_C - 2f_m = 10\,\text{MHz} + 2 \times 1\,\text{kHz} = 9.998\,\text{MHz}, \tag{13.22}$$

which is close enough to the carrier frequency and, therefore, passes through the RF amplifier's resonator and enters the mixer. Consequently, output of the mixer must contain the following tones

$$\begin{aligned} f_3 &= f_{VCO} + f_{ghost} = 9.999\,\text{MHz} + 9.998\,\text{MHz} = 19.997\,\text{MHz}, \\ f_4 &= f_{VCO} - f_{ghost} = 9.999\,\text{MHz} - 9.998\,\text{MHz} = 1\,\text{kHz}. \end{aligned} \tag{13.23}$$

To our surprise, we find out that we have received not the desired message but another message carried by another carrier at the image frequency. Indeed, the second message was generated by a second (real) transmitter working at $f_{ghost} = 9.998\,\text{MHz}$ frequency, and it is irreversibly mixed with the desired message.

The issue of image frequency must be dealt with before the first mixer stage. The following methods are most often used to deal with it:

- Increasing the Q factor of the input front-end resonator and further rejecting the image (see Sect. 9.6.1).
- Keeping a minimum distance between any two neighbouring radio-transmitting frequencies.
- Declaring "forbidden" frequencies within the frequency spectrum.
- Introducing super-heterodyne receiver topology with a second VCO–mixer pair that further separates the troubling tones from the desired one.

In reality, the radio system design process involves a number of specifications and standards that provide guidelines and working boundaries to the designer.

Example 13.2. For a standard AM receiver that is tuned to a carrier signal of $f_C = 620\,\text{kHz}$ and uses IF frequency of $f_{IF} = 455\,\text{kHz}$, determine the image frequency f_{image} if the receiver is designed to have $f_{VCO} > f_C$.

Solution 13.2. With reference to Fig. 13.11, we write an expression for the frequency of the LO f_{VCO} as the difference between

$$f_{IF} = f_{VCO} - f_C \quad \therefore \quad f_{VCO} = f_{IF} + f_C = 1{,}075\,\text{kHz},$$
$$\therefore$$
$$f_{IF} = f_{image} - f_{VCO} \quad \therefore \quad f_{image} = f_{IF} + f_{VCO} = 1{,}530\,\text{kHz}. \tag{13.24}$$

13.3 Radio Receiver Specifications

System-level radio designers aim to improve the selectivity of the systems by designing architectures that are better equipped to deal with the intermodulation and image frequency issues. It is common for modern radio receiver architectures that are implemented using IC technologies to be able to select a signal from a wide range of carrier frequencies that span over several "standard" frequency bands. For example, the latest cell phones are capable of covering up to three GSM frequency bands, such as the 2,100–1,900–850 MHz combination. The rule is that each user must conform to its assigned channel boundaries, i.e., just as it is not desirable to have signal cross-talk within a multi-wire bundle, it is not desirable to have "spilling over" of frequency spectrum among wireless channels.

13.3.1 Dynamic Range

The term *dynamic range* refers to the ratio of the largest and smallest values that the system is capable of processing. For instance, if the lowest signal amplitude that an amplifier can detect and amplify is 1 mV and the largest amplitude is 1 V, then its dynamic range is 1:1,000. It is common practice in technical and science literature to describe 1 V relative to 1,000 V as a 60 dB dynamic range; that is, dynamic range is a *dimensionless* number.

State-of-the-art electronic equipment often exhibits a dynamic range of more than 100 dB. In order to put a perspective on these numbers, 100 dB is a ratio of 100,000:1 (the equivalent of 1 mV relative to 100 V). That is equivalent to a ratio of, for instance, the height of the CN Tower in Toronto relative to an ant.

13.3.1.1 Noise Floor

The upper limit of the dynamic range is set by circuit nonlinearities. The most commonly used metric for quantifying the dynamic range of a circuit is by specifying its 1 dB compression point or, equivalently, its IIP3. Therefore, control of the upper signal limit is, to a large extent, under the control of the designer. For instance, a straightforward way of increasing the upper signal limit is to design the circuit to operate with increased power supply voltage level.

Determining the minimum signal level that can be detected against the background noise starts by establishing the total amount of noise in the system. We introduced thermal noise in (3.2), which is repeated here

$$P_{\mathrm{n}} = kT\Delta f \quad \mathrm{W}, \tag{13.25}$$

which can be normalized for $\Delta f = 1$ Hz, as

$$P_{\mathrm{n}} = kT \quad [\mathrm{W/Hz}]. \tag{13.26}$$

Unless specifically stated, we assume "room temperature" for the environment, i.e., $T = 290$ K, and write

$$P_{\mathrm{n}} = kT = 1.38 \times 10^{-23}\,\mathrm{J/K} \times 290\,\mathrm{K} = -174\,\mathrm{dBm}. \tag{13.27}$$

This number is commonly used to set the "noise floor" at room temperature. Reducing the environment temperature, of course, reduces the noise floor. That approach is used in high-end receivers for radio astronomy where the incoming signal is very low. Indeed, the approximate power of the radio signal that was transmitted by the Galileo space probe and arrived at Earth was in the order of 10×10^{-21} W or -170 dBm and requires a 70-meter-long DSN antenna. However, the cooling system for temperatures close to 0 K is not suitable for general use. Circuit designers reduce the system noise by controlling the frequency bandwidth Δf.

13.3.1.2 Sensitivity

Defining the sensitivity of a receiver requires that we put together all the knowledge that we have collected in this book and apply the following reasoning. The receiver input signal is referenced relative to the noise floor. Depending upon the circuit bandwidth, there is additional $10\log\Delta f$ noise added into the system. Narrowband systems are the obvious conclusion, however, this opportunity for noise reduction can be exploited only so much. Therefore, for any bandwidth above 1 Hz (13.25) is extended as

$$P_{\mathrm{n}} = -174\,\mathrm{dBm} + 10\log\Delta f \quad \mathrm{dBm}. \tag{13.28}$$

Progressing through the receiver circuit, the internally generated noise is quantified by the noise figure NF, which needs to be added into the noise budget, hence

$$P_{\mathrm{n}} = -174\,\mathrm{dBm} + 10\log\Delta f + NF \quad \mathrm{dBm}, \tag{13.29}$$

which sets the "real" noise floor for the receiver. In order to be useful, the receiver must be able to process signals above the real noise floor; in other words, it has to be designed for a certain desired signal-to-noise ratio, SNR_{desired}.

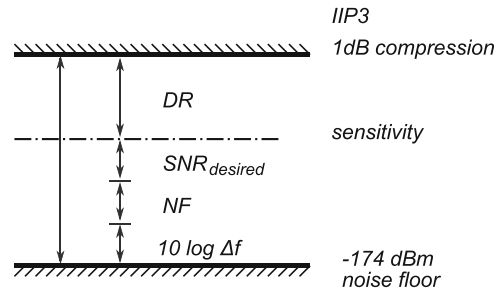


Fig. 13.12 Elements of dynamic range at room temperature

We now define the receiver sensitivity (S) as the signal level

$$S_{\mathrm{n}} = P_{\mathrm{n}} + SNR_{\mathrm{desired}} \quad \mathrm{dBm}, \tag{13.30}$$

where P_{n} represents the level, for the given bandwidth Δf, at which the signal power is equal to the noise power. That is, the level is equivalent to the case when the SNR of the receiver is 0 dB (see Fig. 13.12).

With this discussion in mind, we define the ideal dynamic range as the difference between the 1 dB compression point and the receiver's sensitivity, i.e.

$$DR = 1\,\mathrm{dB}_{\mathrm{point}} - S_{\mathrm{n}} \quad \mathrm{dBm}, \tag{13.31}$$

which is a somewhat optimistic result. In practice, it is often adjusted by about 30% down to $2/3\,DR$. Clearly, it is a goal to design a receiver with as wide a dynamic range as possible. The current state of the art is about 100 dB.

Example 13.3. Determine the sensitivity of a receiver at room temperature whose $NF = 5\,\mathrm{dB}$, $BW = 1\,\mathrm{MHz}$, and desired $SNR = 10\,\mathrm{dB}$.

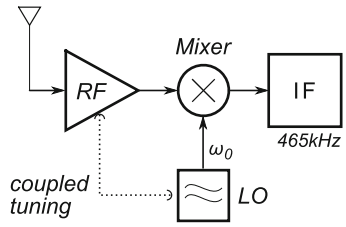
Solution 13.3. A straightforward implementation of (13.29) yields

$$S = -174\,\mathrm{dBm} + 10\log 1\,\mathrm{MHz} + 5\,\mathrm{dB} + 10\,\mathrm{dB} = -99\,\mathrm{dBm},$$

which is a relatively typical number for state-of-the-art receivers.

13.4 Summary

Figures of merit serve the purpose of comparing various design solutions and looking for ways to improve them. Radio receivers deal with very low signal powers; a cell phone, for instance, receives signals as low as $-110\,\mathrm{dBm}$. Thermal noise presents the lower power limit under which the desired signal becomes irretrievably drowned in the background noise. On the upper side limit, nonlinear effects in the receiver circuit and signal distortion become determining factors for establishing the dynamic range.



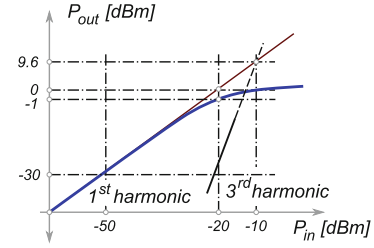


Fig. 13.13 AM receiver block diagram for Problem 13.2 (*left*) and diagram for Problem 13.4 (*right*)

Problems

13.1. An AM receiver is designed to receive RF signals in the 500–1,600 kHz frequency range with the required bandwidth of $BW = 10\,\text{kHz}$ at $f_0 = 1{,}050\,\text{kHz}$. The RF amplifier uses inductor $L = 1\,\mu\text{H}$.

1. Calculate the bandwidth at $f = 1{,}600\,\text{kHz}$ and capacitance C.
2. Calculate the bandwidth at $f = 500\,\text{kHz}$ and capacitance C.
3. Comment on the results.

13.2. An AM receiver is designed to receive RF signals in the 500–1,600 kHz frequency range. All incoming RF signals are shifted to IF $IF = 465\,\text{kHz}$. AM receiver tuning is commonly done by a knob that simultaneously tunes resonating capacitors in the RF and LO oscillator sections. For the receiver architecture in Fig. 13.13 (matching network not shown),

1. Calculate the tuning ratio $C_{\text{RF}}(max)/C_{\text{RF}}(min)$ of the resonator capacitor in the RF amplifier.
2. Calculate the tuning ratio $C_{\text{LO}}(max)/C_{\text{LO}}(min)$ of the resonator capacitor in the local oscillator LO.
3. Recommend the resonating frequency for the local oscillator.

13.3. The LO oscillator frequency is 11 MHz and the RF signal frequency is 10 MHz. What is the image frequency?

13.4. The I/O power characteristic of an amplifier is given in Fig. 13.13 (right). Estimate the gain, the 1 dB compression point, and the IIP3.

13.5. A receiver whose IF frequency is 455 kHz is tuned to a 950 kHz signal. Find all the interference signals including their second harmonics. Is any of them within the range $950\,\text{kHz} \pm 200\,\text{kHz}$? If yes, what Q factor of the front-end LC resonator is needed in order to suppress the interference signal to -80 dB below the desired tone?

13.6. A receiver operates in the 3–30 MHz range while using 10.7 MHz IF frequency. Estimate the range of oscillator frequencies and the range of image frequencies. Can you suggest filters to be used with this receiver?

13.7. A double-conversion receiver architecture is based on two IF frequencies, $IF_1 = 10.7\,\text{MHz}$ and $IF_2 = 455\,\text{kHz}$. If the receiver is tuned to a 20 MHz signal, find the frequencies of the LOs and the image frequencies.

Appendix A
Physical Constants and Engineering Prefixes

Table A.1 Basic physical constants

Physical constant	Symbol	Value
Speed of light in vacuum	c	$299\,792\,458\,\mathrm{m/s}$
Magnetic constant (vacuum permeability)	μ_0	$4\pi \times 10^{-7}\,\mathrm{N/A^2}$
Electric constant (vacuum permittivity)	$\varepsilon_0 = 1/(\mu_0 c^2)$	$8.854\,187\,817 \times 10^{-12}\,\mathrm{F/m}$
Characteristic impedance of vacuum	$Z_0 = \mu_0 c$	$376.730\,313\,461\,\Omega$
Coulomb's constant	$k_e = 1/4\pi\varepsilon_0$	$8.987\,551\,787 \times 10^{9}\,\mathrm{Nm^2/C^2}$
Elementary charge	e	$1.602\,176\,565 \times 10^{-19}\,\mathrm{C}$
Boltzmann constant	k	$1.380\,648\,8 \times 10^{-23}\,\mathrm{J/K}$

Table A.2 Basic engineering prefix system

Tera	Giga	Mega	Kilo	Hecto	Deca	Deci	Centi	Milli	Micro	Nano	Pico	Femto	Atto
T	G	M	k	h	da	d	c	m	μ	n	p	f	a
10^{12}	10^{9}	10^{6}	10^{3}	10^{2}	10^{1}	10^{-1}	10^{-2}	10^{-3}	10^{-6}	10^{-9}	10^{-12}	10^{-15}	10^{-18}

Table A.3 SI system of fundamental units

Name	Symbol	Quantity	Symbol
Meter	m	Length	l
Kilogram	kg	Mass	m
Second	s	Time	t
Ampere	A	Electric current	I
Kelvin	K	Thermodynamic temperature (-273.16°C)	T
Candela	cd	Luminous intensity	Iv
Mole	mol	Amount of substance	n

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8,

Appendix B
Maxwell's Equations

The complete set of Maxwell's equations is listed here for reference.

1. Gauss's law for electric fields:

$$\oint_S \mathbf{D} \cdot \mathrm{d}\mathbf{s} = q_{\text{free, enc}} \qquad \text{integral form,} \tag{B.1}$$

$$\nabla \cdot \mathbf{D} = \rho_{\text{free}} \qquad \text{differential form.} \tag{B.2}$$

2. Gauss's law for magnetic fields:

$$\oint_S \mathbf{B} \cdot \mathrm{d}\mathbf{s} = 0 \qquad \text{integral form,} \tag{B.3}$$

$$\nabla \cdot \mathbf{B} = 0 \qquad \text{differential form.} \tag{B.4}$$

3. Faraday's law:

$$\oint_L \mathbf{E} \cdot \mathrm{d}\mathbf{l} = -\frac{\mathrm{d}}{\mathrm{d}t} \int_S \mathbf{B} \cdot \mathrm{d}\mathbf{s} \qquad \text{integral form,} \tag{B.5}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \text{differential form.} \tag{B.6}$$

4. Ampere–Maxwell law:

$$\oint_L \mathbf{H} \cdot \mathrm{d}\mathbf{l} = I_{\text{free, enc}} + \frac{\mathrm{d}}{\mathrm{d}t} \int_S \mathbf{D} \cdot \mathrm{d}\mathbf{s} \qquad \text{integral form,} \tag{B.7}$$

$$\nabla \times \mathbf{H} = \mathbf{J}_{\text{free}} + \frac{\partial \mathbf{D}}{\partial t} \qquad \text{differential form.} \tag{B.8}$$

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8,

Appendix C
Second-Order Differential Equation

The three basic elements have voltages at their respective terminals as:

$$v_{\mathrm{R}} = iR \quad v_{\mathrm{L}} = L\frac{\mathrm{d}i}{\mathrm{d}t} \quad v_{\mathrm{C}} = \frac{q}{C}. \tag{C.1}$$

If they are put together in a series circuit that includes a voltage source $v(t)$, after applying KVL, the circuit equation is

$$v(t) = v_{\mathrm{L}} + v_{\mathrm{R}} + v_{\mathrm{C}},$$

$$\therefore$$

$$v(t) = L\frac{\mathrm{d}i}{\mathrm{d}t} + iR + \frac{q}{C}. \tag{C.2}$$

However, we know that a current is a derivative of charge in respect to time, hence we have the second-order differential equation

$$v(t) = L\frac{\mathrm{d}^2q}{\mathrm{d}t^2} + R\frac{\mathrm{d}q}{\mathrm{d}t} + \frac{1}{C}q,$$

$$\therefore$$

$$v(t) = \frac{\mathrm{d}^2q}{\mathrm{d}t^2} + \frac{R}{L}\frac{\mathrm{d}q}{\mathrm{d}t} + \frac{1}{LC}q. \tag{C.3}$$

This is solved, starting with its auxiliary quadratic equation

$$0 = x^2 + \frac{R}{L}x + \frac{1}{LC} \tag{C.4}$$

and its general solution with complex roots is

$$r_{1,2} = \frac{1}{2}\left(-\frac{R}{L} \pm \sqrt{\left(\frac{R}{L}\right)^2 - \frac{4}{LC}}\right). \tag{C.5}$$

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8,

Appendix D
Complex Numbers

A complex number is a neat way of presenting a point in (mathematical) *space* with two coordinates or, equivalently, it is a neat way to write two equations in the form of one. A general complex number is $Z = a + \mathrm{j}b$, where a and b are real numbers referred to as real and imaginary parts, i.e. $\Re(Z) = a$, and $\Im(Z) = b$. Here is a reminder of the basic operations with complex numbers. Keep in mind that $\mathrm{j}^2 = -1$.

$$(a+\mathrm{j}b)+(c+\mathrm{j}d) = (a+c)+\mathrm{j}(b+d), \tag{D.1}$$

$$(a+\mathrm{j}b)-(c+\mathrm{j}d) = (a-c)+\mathrm{j}(b-d), \tag{D.2}$$

$$(a+\mathrm{j}b)\,(c+\mathrm{j}d) = (ac-bd)+\mathrm{j}(bc+ad), \tag{D.3}$$

$$\frac{(a+\mathrm{j}b)}{(c+\mathrm{j}d)} = \frac{(a+\mathrm{j}b)}{(c+\mathrm{j}d)}\,\frac{(c-\mathrm{j}d)}{(c-\mathrm{j}d)} = \frac{ac+bd}{c^2+d^2}+\mathrm{j}\,\frac{bc-ad}{c^2+d^2}, \tag{D.4}$$

$$(a+\mathrm{j}b)^* = (a-\mathrm{j}b), \tag{D.5}$$

$$|(a+\mathrm{j}b)| = \sqrt{(a+\mathrm{j}b)(a-\mathrm{j}b)} = \sqrt{(a^2+b^2)}. \tag{D.6}$$

It is much easier to visualize complex numbers and operations if we use vectors and the trigonometry of a right triangle, i.e. Pythagoras' theorem. The imaginary part always takes its value from the y axis and the real part is always on the x axis (see Fig. D.1).

Therefore, an alternative view of complex numbers is based on geometry, i.e.

$$(a+\mathrm{j}b) \equiv (|Z|, \theta), \tag{D.7}$$

where, of course, the absolute value of Z is the length of the hypotenuse and the real and imaginary parts are the two legs of the right-angled triangle, i.e.

$$|Z| = \sqrt{ZZ^*} = \sqrt{(a^2+b^2)}, \qquad heta = \arctan\left(\frac{b}{a}\right), \tag{D.8}$$

where θ is the phase angle. After using Euler's formula, this becomes

$$\mathrm{e}^{\mathrm{j}x} \equiv \cos x + \mathrm{j}\cos x \tag{D.9}$$

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8,

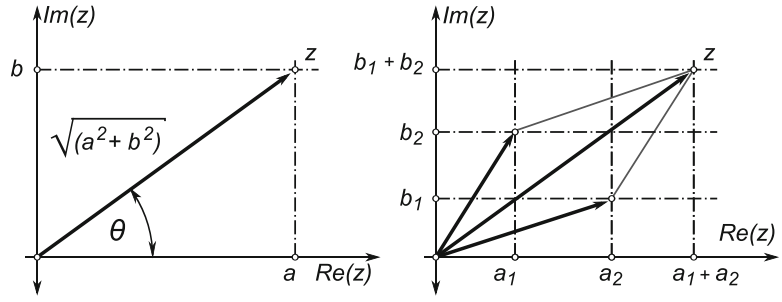


Fig. D.1 Complex numbers in $[\Re(Z), \Im(Z)]$ space, their equivalence to Pythagoras' theorem and vector arithmetic

and enables us to write a really compact form of complex numbers

$$Z = a + \mathrm{j}b = |Z|\,\mathrm{e}^{\mathrm{j}\theta}, \tag{D.10}$$

which leads into another simple way of doing complex arithmetic, by using the absolute values and the arguments in combination with the algebraic rules of exponential numbers, for example

$$\left(A\,\mathrm{e}^{\mathrm{j}\theta_A}\right)\,\left(B\mathrm{e}^{\mathrm{j}\theta_B}\right) = AB\,\mathrm{e}^{\mathrm{j}(\theta_A+\theta_B)} \tag{D.11}$$

and we have the final link,

$$A\,\mathrm{e}^{\mathrm{j}\theta} \equiv A\,(\cos\theta + \mathrm{j}\sin\theta), \tag{D.12}$$

where

$$\Re\left(A\,\mathrm{e}^{\mathrm{j}\theta}\right) = A\cos\theta \qquad \Im\left(A\,\mathrm{e}^{\mathrm{j}\theta}\right) = A\sin\theta. \tag{D.13}$$

Appendix E
Basic Trigonometric Identities

$$\sin(\alpha + \pi/2) = +\cos\alpha \tag{E.1}$$

$$\cos(\alpha + \pi/2) = -\sin\alpha \tag{E.2}$$

$$\sin(\alpha + \pi) = -\sin\alpha \tag{E.3}$$

$$\cos(\alpha + \pi) = -\cos\alpha \tag{E.4}$$

$$\sin(\alpha \pm \beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta \tag{E.5}$$

$$\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta \tag{E.6}$$

$$\sin^2\alpha = 1/2\,(1 - \cos 2\alpha) \tag{E.7}$$

$$\cos^2\alpha = 1/2\,(1 + \cos 2\alpha) \tag{E.8}$$

$$\sin^3\alpha = 1/4\,(3\sin\alpha - \sin 3\alpha) \tag{E.9}$$

$$\cos^3\alpha = 1/4\,(3\cos\alpha + \cos 3\alpha) \tag{E.10}$$

$$\sin^2\alpha\cos^2\alpha = 1/8\,(1 - \cos 4\alpha) \tag{E.11}$$

$$\sin^3\alpha\cos^3\alpha = 1/32\,(3\sin 2\alpha - \sin 6\alpha) \tag{E.12}$$

$$\cos\alpha\cos\beta = 1/2\,(\cos(\alpha - \beta) + \cos(\alpha + \beta)) \tag{E.13}$$

$$\sin\alpha\sin\beta = 1/2\,(\cos(\alpha - \beta) - \cos(\alpha + \beta)) \tag{E.14}$$

$$\sin\alpha\cos\beta = 1/2\,(\sin(\alpha + \beta) + \sin(\alpha - \beta)) \tag{E.15}$$

$$\cos\alpha\sin\beta = 1/2\,(\sin(\alpha + \beta) - \sin(\alpha - \beta)) \tag{E.16}$$

$$\sin\alpha \pm \sin\beta = 2\sin\left(\frac{\alpha \pm \beta}{2}\right)\cos\left(\frac{\alpha \mp \beta}{2}\right) \tag{E.17}$$

$$\cos\alpha + \cos\beta = 2\cos\left(\frac{\alpha + \beta}{2}\right)\cos\left(\frac{\alpha - \beta}{2}\right) \tag{E.18}$$

$$\cos\alpha - \cos\beta = -2\sin\left(\frac{\alpha + \beta}{2}\right)\sin\left(\frac{\alpha - \beta}{2}\right) \tag{E.19}$$

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8,

Appendix F
Useful Algebraic Equations

1. *Binomial formula*

$$(x \pm y)^2 = x^2 \pm 2xy + y^2, \tag{F.1}$$

$$(x \pm y)^3 = x^3 \pm 3x^2y + 3xy^2 \pm y^3, \tag{F.2}$$

$$(x \pm y)^4 = x^4 \pm 4x^3y + 6x^2y^2 \pm 4xy^3 + y^4, \tag{F.3}$$

$$(x \pm y)^n = x^n + nx^{n-1} + \frac{n(n-1)}{2!}x^{n-2}y^2 + \frac{n(n-1)(n-2)}{3!}x^{n-3}y^3 \cdots + y^n, \tag{F.4}$$

where, $n! = 1 \cdot 2 \cdot 3 \cdots n$ and $0! \equiv 1$.

2. *Special cases*

$$x^2 - y^2 = (x-y)(x+y), \tag{F.5}$$

$$x^3 - y^3 = (x-y)(x^2 + xy + y^2), \tag{F.6}$$

$$x^3 + y^3 = (x+y)(x^2 - xy + y^2), \tag{F.7}$$

$$x^4 - y^4 = (x^2 - y^2)(x^2 + y^2) = (x-y)(x+y)(x^2 + y^2). \tag{F.8}$$

3. *Useful Taylor series*

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots, \tag{F.9}$$

$$\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}x^{2n+1} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots \quad \text{for all } x, \tag{F.10}$$

$$\cos x = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}x^{2n} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots \quad \text{for all } x, \tag{F.11}$$

$$\tan x = \sum_{n=1}^{\infty} \frac{B_{2n}(-4)^n(1-4^n)}{(2n)!}x^{2n-1} = x + \frac{x^3}{3} + \frac{2x^5}{15} + \cdots \quad \text{for } |x| < \frac{\pi}{2}. \tag{F.12}$$

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8,

Appendix G
Bessel Polynomials

1. *Bessel differential equation*

$$x^2\frac{\mathrm{d}^2y}{\mathrm{d}x^2}+x\frac{\mathrm{d}y}{\mathrm{d}x}+(x^2-\alpha^2)y=0. \tag{G.1}$$

2. *Relation with trigonometric functions*

$$\cos(x\sin\alpha)=J_0(x)+2\left[J_2(x)\cos 2\alpha+J_4(x)\cos 4\alpha+\cdots\right], \tag{G.2}$$

$$\sin(x\sin\alpha)=2\left[J_1(x)\sin\alpha+J_3(x)\sin 3\alpha+J_5(x)\sin 5\alpha+\cdots\right], \tag{G.3}$$

$$\begin{aligned}\cos(x\cos\alpha)=&J_0(x)-2\left[J_2(x)\cos 2\alpha-J_4(x)\cos 4\alpha\right.\\&\left.+J_6(x)\cos 6-J_8(x)\cos 8\alpha\cdots\right],\end{aligned} \tag{G.4}$$

$$\sin(x\cos\alpha)=2\left[J_1(x)\cos\alpha-J_3(x)\sin 3\alpha+J_5(x)\sin 5\alpha+\cdots\right]. \tag{G.5}$$

3. *Bessel series*

$$J_0(x)=1-\frac{x^2}{2^2}+\frac{x^4}{2^2\cdot 4^2}-\frac{x^6}{2^2\cdot 4^2\cdot 6^2}+\cdots, \tag{G.6}$$

$$J_1(x)=\frac{x}{2}\left(1-\frac{x^2}{2^2\cdot 2}+\frac{x^4}{2\cdot 2^4\cdot 2\cdot 3}+\cdots\right), \tag{G.7}$$

$$\begin{aligned}J_n(x)=\frac{x^n}{2^n n!}\left(1-\frac{x^2}{2^2\cdot(n+1)}+\frac{x^4}{2\cdot 2^4\cdot(n+1)\cdot(n+2)}\right.\\\left.+\frac{(-1)^p x^{2p}}{p!\,2^{2p}(n+1)(n+2)\cdots(n+p)}+\cdots\right).\end{aligned} \tag{G.8}$$

4. *Bessel approximations*

For very large x, the Bessel function reduces to

$$J_n(x)=\sqrt{\frac{2}{\pi x}}\cos\left(x-\frac{n\pi}{2}-\frac{\pi}{4}\right). \tag{G.9}$$

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8,

Bibliography

[Amo90] S. W. Amos. *Principles of Transistor Circuits*. Number 0–408–04851–4. Butterworths, 1990.

[BG03a] Les Besser and Rowan Gilmore. *Practical RF Circuit Design for Modern Wireless Systems I*. Number 1-58053-521-6. Artech House, 2003.

[BG03b] Les Besser and Rowan Gilmore. *Practical RF Circuit Design for Modern Wireless Systems II*. Number 1-58053-522-4. Artech House, 2003.

[BMV05] J. S. Beasley, G. M. Miller, and J. K. Vasek. *Modern Electronic Communication* . Number 0–13–113037–4. Pearson, Prentice Hall, 2005.

[Bro90] James J. Brophy. *Basic Electronics for Scientist*. Number 0–07–008147–6. McGraw–Hill Inc., 1990.

[Bub84] P. Bubb. *Understanding Radio Waves*. Number 0–7188–2581–0. Lutterworth Press, 1984.

[CC03a] David Comer and Donald Comer. *Advanced Electronic Circuit Design*. Number 0–471–22828–1. John Wiley & Sons Inc., 2003.

[CC03b] David Comer and Donald Comer. *Fundamentals of Electronic Circuit Design*. Number 0–471–41016–0. John Wiley & Sons Inc., 2003.

[CL62] D. R. Corson and P. Lorrain. *Introduction to Electromagnetic Fields and Waves*. Number 62–14193. Freeman Co., 1962.

[DA01] W. A. Davis and K. K. Agarwal. *Radio Frequency Circuit Design*. Number 0–471–35052–4. Wiley Interscience, 2001.

[DA07] W. A. Davis and K. K. Agarwal. *Analysis of Bipolar and CMOS Amplifiers*. Number 1–4200–4644–6. CRC Press, 2007.

[Ell66] R. S. Elliott. *Electromagnetics*. Number 66–14804. McGraw Hill, 1966.

[Fle08] D. Fleisch. *A Student's Guide to Maxwell's Equations*. Number 978–0–521–87761–9. Cambridge University Press, 2008.

[FLS05] R.P. Feynman, R.B. Leighton, and M. Sands. *The Feynman Lectures on Physics*. Number 0–8053–9047–2. Pearson Addison Wesley, 2005.

[GM93] Paul G. Gray and Robert G. Meyer. *Analysis and Design of Analog Integrated Circtuits*. Number 0–471–57495–3. John Wiley & Sons Inc., 1993.

[Gol48] S. Goldman. *Frequency Analysis, Modulation and Noise*. Number TK6553.G58 1948. McGraw-Hill, 1948.

[Gre04] B. Green. *The Fabric of Cosmos*. Number 0–375–72720–5. Vintage Books, 2004.

[Gri84] J. Gribbin. *In Search of Schrödinger's Cat, Quantum Physics and Reality*. Number 0–553–34253–3. Bantam Books, 1984.

[HH89a] T. C. Hayes and P. Horowitz. *Student Manual for The Art of Electronics*. Number 0–521–37709–9. Cambridge University Press, 1989.

[HH89b] P. Horowitz and W. Hill. *The Art of Electronics*. Number 0–521–37095–7. Cambridge University Press, 1989.

[Hur10] P. G. Huray. *Maxwell's Equations*. Number 978–0–470–54276–7. Wiley, 2010.

[II99] U. S. Inan and A. S. Inan. *Electromagnetic Waves*. Number 0–201–36179–5. Prentice Hall, 1999.

[JK93] W. H. Hayt Jr. and J. E. Kemmerly. *Engineering Circuit Analysis*. Number 0–07–027410–X. McGraw Hill, 1993.

[JN71] R. H. Good Jr. and T. H. Nelson. *Classical Theory of Electric and Magnetic Fields*. Number 78–137–628. Academic Press, 1971.

[Jr.89] W.H Hayt Jr. *Engineering Electromagnetics*. Number 0-07-024706-1. McGraw Hill, 1989.

[KB80] H. L. Krauss and C. W. Bostian. *Solid State Radio Engineering*. Number 0–471–03018–X. Wiley, 1980.

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8,

[Kin09] G. C. King. *Vibrations and Waves*. Number 978–0–470–01189–8. Wiley, 2009.
[Kon75] J. A. Kong. *Theory of Electromagnetic Waves*. Number 0–471–50190–5. Wiley, 1975.
[LB00] R. Ludwig and P. Bretchko. *RF Circuit Design, Theory and Applications*. Number 0–13–095323–7. Prentice Hall, 2000.
[Lee05] Thomas H. Lee. *The Design of CMOS Radio–Frequency Integrated Circtuis*. Number 0–521–63922–0. Cambridge University Press, 2005.
[Lov66] W.F. Lovering. *Radio Communication*. Number TK6550.L546 1966. Longmans, 1966.
[PP99] Z. Popovic and D. Popovic. *Electromagnetic Waves*. Number 0–201–36179–5. Prentice Hall, 1999.
[Pur85] E. M. Purcell. *Electricity and Magnetism*. Number 0–07–004908–4. McGraw Hill, 1985.
[Rad01] M.M. Radmanesh. *Radio Frequency and Microwave Electronics*. Number 0–13–027958–7. Prentice Hall, 2001.
[Raz98] Behzad Razavi. *RF Microelectronics*. Number 0–13–887571–5. Prentice Hall, 1998.
[RC84] Dennis Roddy and John Coolen. *Electronic Communications*. Number 0-8359-1598-0. Reston Publishing Company, 1984.
[RR67] J. H. Reyner and P. J. Reyner. *Radio Communication*. Sir Isaac Pitman & Son Ltd, 1967.
[Rut99] David B. Rutledge. *The Electronics of Radio*. Number 0-521-64136-5. Cambridge University Press, 1999.
[SB00] Ben Streetman and Sanjay Banerjee. *Solid State Electronic Devices*. Number 0-13-025538-6. Prentice Hall, 2000.
[Sch92] R. J. Schoenbeck. *Electronic Communications Modulation and Transmission*. Number 0–675–21311–8. Prentice Hill, 1992.
[Scr84] M.G. Scroggie. *Foundations of Wireless and Electronics*. Number 0–408–01202–1. Newnes Technical Books, 10 edition, 1984.
[See56] S. Seely. *Radio Electronics*. Number 55–5696. McGraw Hill, 1956.
[Sim87] Robert E. Simpson. *Introductory Electronics for Scientist and Engineers*. Number 0–205–08377–3. Allyn and Bacon Inc., 1987.
[Sze81] S. M. Sze. *Physics of Semiconductor Devices*. Number 0–471–05661–8. John Wiley and Sons, 1981.
[Ter03] David Terrell. *Electronics for Computer Technology*. Number 0–7668–3872–2. Thompson Delmar Learning, 2003.
[Tho06] M. T. Thompson. *Intuitive Analog Circuit Design*. Number 0–7506–7786–4. Newnes, 2006.
[Wik10a] Wikipedia.org. Electromagnetic wave equation.
URL: http://en.wikipedia.org/wiki/electromagnetic_wave_equation, September 2010.
[Wik10b] Wikipedia.org. Waves, wavelength.
URL: http://en.wikipedia.org/wiki/wave, July 2010.
[Wol91] D. H. Wolaver. *Phase–Locked Loop Circuit Design*. Number 0–13–662743–9. Prentice Hall, 1991.
[You04] Paul H. Young. *Electronic Communication Techniques*. Number 0–13–048285–4. Pearson, Prentice Hill, 2004.

Glossary

This glossary of technical terms is provided for reference only. The reader is advised to further study the terms in appropriate books, for example, a technical dictionary.

1 dB gain compression point The point at which the power gain at the output of a nonlinear device or circuit is reduced by 1 dB relative to its small signal linear model predicted value.

Absolute zero The theoretical temperature at which entropy would reach its minimum value. By international agreement, absolute zero is defined as 0 K on the Kelvin scale and as $-273.15°$C on the Celsius scale.

Active device An electronic component that has signal gain larger than one, for example a transistor. Compare to *passive device*.

Active mode A condition for a BJT in which the emitter-base junction is forward biased, while the collector-base junction is reverse biased.

Admittance The measure of how easily AC current flows in a circuit (in Siemens [S]). The reciprocal of *impedance*.

Ampere (A) The unit of electric current defined as the flow of one coulomb of charge per second.

Ampère's Law A current flowing into a wire generates a magnetic flux that encircles the wire following the "right hand rule' (the right thumb points in the direction of the current flow and the curled fingers show the direction of the magnetic field). Study *Maxwell's equations* for more details.

Amplifier A linear device that implements the mathematical equation $y = Ax$, where y is the amplified output signal, A is the gain coefficient, and x is the input signal.

Analogue The general class of devices and circuits meant to process a continuous signal. Compare with *digital* and sampled signals.

Attenuation Gain lower than one.

Attenuator A device that reduces gain without introducing phase or frequency distortion.

Automatic gain control A closed-loop feedback system designed to hold the overall gain as constant as possible.

Average power The power averaged over one time period.

Bandwidth The difference between upper and lower frequencies at which the amplitude response is 3 dB below the maximum. It is equivalent to half-power bandwidth.

Base The region of a BJT between the emitter and the collector.

Bel (B) A dimensionless unit used to express the ratio of two powers. A more practical unit is the *dB*.

Beta β The current gain of a BJT. It is the ratio of the change in collector current to the change in base current, $\beta = dI_C/dI_B$.

Bias A steady current or voltage used to set the operating conditions of a device.

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8,

Breakdown voltage The voltage at which the reverse current of a reverse-biased p–n junction suddenly rises. If the current is not limited, the device is destroyed.

Capacitance The ratio of the electric charge and voltage between two conductors.

Capacitor A device made of two conductors separated by an insulating material for the purpose of storing an electric charge, i.e., energy.

Celsius (°C) A unit increment of temperature unit defined as 1/100 between the freezing point (0°C) and boiling point (100°C) of water. Compare with *Kelvin* and Fahrenheit.

Characteristic curve A family of I–V plots shown for several parameter values.

Characteristic line impedance The entry point impedance of an infinitely long transmission line.

Charge A basic property of elementary particles of matter (electrons, protons, etc.) responsible for creating a force field.

Circuit The interconnection of devices, both passive and active, for the purpose of synthesizing a mathematical function.

Common base A single BJT amplifier configuration in which the base potential is fixed, the emitter serves as the input and the collector as the output terminal. Also known as a "current buffer". Equivalent to a "common-gate" configuration for MOS amplifiers.

Common collector A single BJT amplifier configuration in which the collector potential is fixed, the base serves as the input and the emitter as the output terminal. Also known as a "voltage buffer" or voltage follower. Equivalent to a "common-drain" configuration for MOS amplifiers.

Common emitter A single BJT amplifier configuration in which the emitter potential is fixed, while the base serves as the input and the collector as the output terminal. Also known as the "g_m stage". Equivalent to a "common-source" configuration for MOS amplifiers.

Common mode The average value of a sinusoidal waveform.

Conductivity The ability of a matter to conduct electricity.

Conductor A material that easily conducts electricity.

Coulomb (C) The unit of electric charge defined as the charge transported through a unity area in one second by an electric current of one ampere. An electron has a charge of $1.602 \times 10^{-19} C$.

Coulomb's Law A definition of the force between two electric charges in space.

Current A transfer of electrical charge through a unity size area per unit of time.

Current gain The ratio of current at the output terminals to the current at the input terminals of a device or circuit.

Current source A device capable of providing constant current value regardless of the voltage at its terminals.

DC See *Direct current*.

DC analysis A mathematical procedure to calculate the stable operating point.

DC biasing The process of setting the stable operating point of a device.

DC load line A straight line across a family of I–V curves that shows movement of the operating point as the output voltage changes for a given load.

Decibel (dB) A dimensionless unit used to express the ratio of two powers. A decibel is ten times smaller than a *bel* (B).

Device A single discrete device, for instance a resistor, a transistor, or a capacitor.

Dielectric A material that is not good in conducting electricity, i.e. the opposite of a conductor. Characterized by the dielectric constant.

Differential amplifier An amplifier that operates on differential signals.

Differential signal A difference between two sinusoidal signals of the same frequency, same amplitude, same common mode, and with phase difference of 180°.

Digital The general class of devices and circuits meant to process a sampled signal. Compare with *analogue* and continuous signals.

Diode A nonlinear, two-terminal device that obeys the exponential transfer function. Used as unidirectional switch.

Direct current (DC) Current that flows in one direction only.

Discrete device An individual electrical component that exhibits behaviour associated with a resistor, a transistor, a capacitor, an inductor, etc. Compare with distributed components.

Dynamic range The difference between the maximum acceptable signal level and the minimum acceptable signal level.

Electric field A field generated by an electric charge, detected by the existence of the electric force within a space surrounding the charge.

Electrical noise Any unwanted electrical signal.

Electromagnetic (EM) wave A phenomenon exhibited by a flow of electromagnetic energy through space. In the special case of a *standing wave*, this definition may need more explanation.

Electron A fundamental particle that carries negative charge.

Electronics The branch of science and technology which makes use of the controlled motion of electrons through different media and a vacuum.

Electrostatics The branch of science that deals with the phenomena arising from stationary or slow-moving electric charges.

Emitter A region of a BJT from which charges are injected into the base. One of the three terminal points of a BJT device.

Energy A concept that can be loosely defined as the ability of a body to perform work.

Equivalent circuit A simplified version of a circuit that performs the same function as the original.

Equivalent noise temperature The absolute temperature at which a perfect resistor would generate the same noise as its equivalent real component at room temperature.

Fall time The time during which a pulse decreases from 90% to 10% of its maximum value (sometimes defined between the 80% and 20% points).

Farad (F) The unit of capacitance of a capacitor. One farad is very large; the capacitance of the Earth's ionosphere with respect to the ground is around 50 mF.

Faraday cage An enclosure that blocks out external static electric fields.

Faraday's Law The law of electromagnetic induction. See also *Faraday cage*.

Feedback The process of coupling output and input terminals through an external path. Negative feedback increases the stability of an amplifier at the cost of reduced gain, positive feedback boosts gain and is needed to create oscillating circuits.

Field A concept that describes a flow of energy through space.

Field-effect transistor (FET) A transistor controlled by two perpendicular electric fields used to change the resistivity of the semiconductor material underneath the gate terminal and force current between the source and drain terminals.

Flicker noise A random noise in semiconductors whose power spectral density is, to the first approximation, inverse to frequency ($1/f$ noise).

Frequency The number of complete cycles per second.

Frequency response A curve showing the gain and phase change of a device as a function of frequency.

Gain The ratio of signal values measured at output and input terminals.

Gauss's Law A law relating the distribution of electric charge to the resulting electric field.

Ground An arbitrary potential reference point that all other potentials in a circuit are compared against. The difference between the ground potential and the node potential is expressed as voltage at that node. The ground node may or may not have the lowest potential in the circuit.

Henry (H) The unit of measurement for self and mutual inductance.

Hertz (Hz) The unit of measurement for frequency, equal to one cycle per second.

Impedance Resistance of a two-terminal device at any frequency.

Inductance A property whereby a change in the electrical current through a circuit induces an electromotive force (EMF) that opposes the change in current.

Inductor A passive electrical component that can store energy in a magnetic field created by an electric current passing through it.

Input Current, voltage, power, or another driving force applied to a circuit or device.

Insertion loss The attenuation resulting from inserting a circuit between source and load.

Insulator A material with very low conductivity.

Intermediate frequency (IF) A frequency to which a carrier frequency is shifted as an intermediate step in transmission or reception.

Intermodulation products Additional harmonics created by a nonlinear device processing two or more single-tone signals.

Junction A joining of two semiconductor materials.

Junction capacitance Capacitance associated with a p–n junction region.

Kelvin (K) The unit increment of temperature on the absolute temperature scale.

Kirchhoff's current law (KCL) The law of conservation of charge: at any instant, the total current entering any point in a network is equal to the total current leaving the same point.

Kirchhoff's voltage law (KVL) The law of conservation of energy given or taken by a potential field (not including energy taken by dissipation): at any instant, the algebraic sum of all electromotive forces and potential differences around a closed loop is zero.

Large signal A signal with an amplitude large enough to move the operating point of a device far from its original biasing point. Hence, a nonlinear model of the device must be used.

Large-signal analysis A method used to describe the behaviour of devices stimulated by large signals. It describes nonlinear devices in terms of the underlying nonlinear equations.

Law of conservation of energy The fundamental law of nature. It states that energy can neither be created nor destroyed, it can only be transformed from one state to another.

Linear network A network in which the parameters of resistance, inductance, and capacitance are constant with respect to current or voltage, and in which the voltage or current of sources is independent of or directly proportional to other voltages and currents, or their derivatives, in the network.

Load A device that absorbs energy and converts it into another form.

Local oscillator (LO) An oscillator used to generate a single-tone signal that is needed for upconversion and downconversion operations.

Lossless A theoretical device that does not dissipate energy.

Low noise amplifier (LNA) An electronic amplifier used to amplify very weak signals captured by an antenna.

Lumped element A self-contained and localized element that offers one particular property, for example, resistance over a range of frequencies.

Magnetic field A field generated by magnetic energy, detected by the existence of a magnetic force within the space surrounding a magnet.

Matching A concept of connecting two networks to enable maximum energy transfer between them.

Matching circuit A passive circuit designed to interface two networks to enable maximum energy transfer between the two networks.

Maxwell's equations A set of four partial differential equations that relate electric and magnetic fields to their sources, charge density and current density. These equations can be combined to show that light is an electromagnetic wave. Individually, the equations are known as Gauss's law, Gauss's law for magnetism, Faraday's law of induction, and Ampère's law with Maxwell's correction. These four equations and the Lorentz force law make up the complete set of laws of classical electromagnetism.

Metal-oxide semiconductor field-effect transistor (MOSFET) Originally, a sandwich of aluminum–silicone dioxide–silicon was used to manufacture FET transistors. Although metal is no longer used to create gates for FE transistors, the name has stuck.

Microwaves Waves in the frequency range of 1–300 GHz, i.e. with a wavelength of 300–1 mm.

Mixer A nonlinear, three-port device used for frequency-shifting operations.
Negative resistance The resistance of a device or circuit where an increase in the current entering a port results in a decrease in voltage across the same port.
Noise Any unwanted signal that interferes with a wanted signal.
Noise figure (NF) A measure of degradation of the signal-to-noise ratio (SNR), caused by components in a radio frequency (RF) signal chain.
Nonlinear circuit A system that does not satisfy the superposition principle or whose output is not directly proportional to its input.
Norton's Theorem Any collection of voltage sources, current sources, and resistors with two terminals is electrically equivalent to an ideal current source in parallel with a single resistor. This is the twin of *Thévenin's theorem*.
NPN transistor A transistor with a p-type base and n-type collector and emitter.
Octave The interval between any two frequencies having a ratio of 2:1.
Ohm (Ω) Unit of resistance, as defined by Ohm's law.
Ohm's Law The change of current through a conductor between two points is directly proportional to the change of voltage across the two points and inversely proportional to the resistance between them.
One-dB gain compression point See *1 dB gain compression point*.
Open-loop gain The ratio of the output signal and the input signals of an amplifier with no feedback path present.
Oscillator An electronic device that generates a single tone (or some other regular shape) signal at predetermined frequency.
Output Current, voltage, power, or a driving force delivered at the output terminals.
Passive device A component that does not have a gain larger than one. Compare to *active device*.
Phase The angular property of a wave.
Phase shifter A two-port network that provides a controllable phase shift of the RF signals.
Phasor A mathematical representation of a sine wave by a rotating vector.
Power The rate at which work is performed.
Quality factor (Q factor) A dimensionless parameter that characterizes a resonator's bandwidth relative to its centre frequency.
Radio frequency (RF) Any frequency at which coherent electromagnetic radiation of energy is possible.
Reactance The opposition of a circuit element to a change of current, caused by the build-up of electric or magnetic fields in the element.
Reactive element An inductor and capacitor.
Reflected waves The waves reflected from a discontinuity in the medium in which they are travelling.
Resistance A measure of an object's opposition to the passage of a steady electric current.
Resistor A lumped element designed to have a certain resistance.
Resonant frequency The frequency at which a given system or circuit responds with maximum amplitude when driven by an external single tone.
Root mean square (RMS) The square root of the arithmetic mean (average) of the squares of the original values.
Saturation A condition in which an increase of the input signal to a circuit does not produce an expected change at the output.
Self-resonant frequency The frequency at which all real devices or circuits start to oscillate due to the internal parasitic inductances and capacitances.
Signal An electrical quantity containing information that is carried by a voltage or current.
Single-ended circuit A circuit operating on single-ended (as opposed to differential) signals.

Skin effect The tendency of an alternating current (AC) to distribute itself within a conductor so that the current density near the surface of the conductor is greater than at its core. That is, the electric current tends to flow at the "skin" of the conductor, at an average depth called the "skin depth".

Small signal A low-amplitude signal that occupies a very narrow region that is centred at the biasing point. Hence, the linear model always applies.

Small-signal amplifier An amplifier that operates only in the linear region.

Space The boundless, three-dimensional extent in which objects and events occur and have relative position and direction.

Stability The ability of a circuit to stay away from the self-resonating frequency.

Standing wave A wave that remains in a constant position. It can arise in a stationary medium as a result of interference between two waves travelling in opposite directions. For waves of equal amplitude travelling in opposite directions, there is no net propagation of energy on average.

Standing wave ratio (SWR) The ratio of the maximum to the minimum value of current or voltage in a standing wave.

Thévenin's theorem Any combination of voltage sources, current sources, and resistors with two terminals is electrically equivalent to a single voltage source and a single series resistor. This is the twin of *Norton's theorem*.

Third-order intercept point (IP3) A measure of weakly nonlinear systems and devices, for example, receivers, linear amplifiers, and mixers.

Time A concept used to order a sequence of events.

Transmission line Any system of conductors capable of efficiently conducting electromagnetic energy.

Tuned circuit A circuit consisting of inductance and capacitance that can be adjusted for resonance at a desired frequency.

Tuning The process of adjusting the resonant frequency of a *tuned circuit*.

Varactor A two-terminal p–n junction used as a voltage-controlled capacitor.

Volt (V) A unit of measurement for potential difference.

Voltage-controlled oscillator (VCO) An oscillator whose output frequency is controlled by a voltage.

Voltage divider A simple linear circuit that produces an output voltage that is a fraction of its input voltage.

Voltage follower amplifier An amplifier that provides electrical impedance transformation from one circuit to another. Also known as a "voltage buffer amplifier".

Voltage source A device capable of providing a constant voltage value regardless of the current at its terminals.

Wave A disturbance that progresses from one point in space to another.

Wavefront A surface having constant phase.

Wavelength A distance in space between two consecutive points having the same phase.

Wave propagation The journey of a wave through space.

White noise A random signal that consists of all possible frequencies from zero to infinity.

Work The advancement in space of a point under application of a force.

Solutions

Solutions to Selected Problems in Chapter 1

1.1 The relative permeability μ_r and permittivity ε_r of free space are equal to unity, while free space permeability μ_0 and permittivity ε_0 are measured directly, resulting in the intrinsic impedance as

$$Z_0 = \sqrt{\frac{\mu}{\varepsilon}} = \sqrt{\frac{\mu_r \mu_0}{\varepsilon_r \varepsilon_0}} = \sqrt{\frac{\mu_0}{\varepsilon_0}} = \sqrt{\frac{4\pi \times 10^{-7}}{8.85 \times 10^{-12}}} \approx 377\Omega. \tag{1}$$

The phase velocity is calculated as

$$v_p = \frac{1}{\sqrt{\mu \varepsilon}} = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} = 2.998 \times 10^8 \text{m/s} \approx 3 \times 10^8 \text{m/s}, \tag{2}$$

i.e., the phase velocity of an EM wave v_p in free space is equal to the speed of light c.
The wavelength is calculated from (1.9) as

$$\lambda = \frac{2\pi}{\beta} = \frac{2\pi v_p}{\omega} = \frac{v_p}{f} = \begin{cases} 30\,\text{m} & \text{for} \quad f_1 = 10\,\text{MHz} \\ 3\,\text{m} & \text{for} \quad f_2 = 100\,\text{MHz} \\ 3\,\text{mm} & \text{for} \quad f_3 = 10\,\text{GHz} \end{cases}. \tag{3}$$

Discussion: This example shows how the wavelength becomes comparable to the sizes of the discrete components (let alone the size of the PCB) as the signal frequency goes up. Note that for typical RF signal frequencies used in this book (e.g. $f = 10\,\text{MHz}$), the size of the PCB (which is of the order of several centimetres per side) is much less than the 30 m calculated wavelength of a 10 MHz signal. Therefore, the use of approximate Maxwell's equations is justified.

On the other hand, for signals whose frequency is of the order of multiples of GHz, even IC designs that typically occupy only a few millimeters per side must account for phase differences along the signal path.

1.2 After applying Kirchhoff's voltage law (KVL) around a circuit loop that includes $V(z)$ and $V(z+\Delta z)$ combining the resistive and inductive impedances, and applying the limit $\Delta z \to 0$,

$$V(z) = (R + j\omega L) I(z) \Delta z + V(z + \Delta z),$$

$$\therefore$$

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8,

$$-\frac{\mathrm{d}V(z)}{\mathrm{d}z} = \lim_{\Delta z \to 0}\left(-\frac{V(z+\Delta z)-V(z)}{\Delta z}\right),$$

$$\therefore$$

$$-\frac{\mathrm{d}V(z)}{\mathrm{d}z} = (R+\mathrm{j}\omega L)\,I(z), \tag{4}$$

similarly, applying Kirchhoff's current law (KCL) to node a

$$I(z+\Delta z) = I(z) - V(z+\Delta z)(G+\mathrm{j}\omega C)\,\Delta z,$$

$$\therefore$$

$$\frac{\mathrm{d}I(z)}{\mathrm{d}z} = \lim_{\Delta z \to 0}\left(\frac{I(z+\Delta z)-I(z)}{\Delta z}\right),$$

$$\therefore$$

$$\frac{\mathrm{d}I(z)}{\mathrm{d}z} = -(G+\mathrm{j}\omega C)\,V(z). \tag{5}$$

Equations (4) and (5) are coupled first-order differential equations. A solution to this system of equations is found by decoupling the two equations, which is accomplished by differentiating both sides, first of (4) and then of (5). By doing this, explicit solutions for $V(z)$ and $I(z)$ are found.

Therefore, starting with spatial differentiation of (4) and substituting (5),

$$-\frac{\mathrm{d}^2V(z)}{\mathrm{d}z} = (R+\mathrm{j}\omega L)\frac{\mathrm{d}I(z)}{\mathrm{d}z},$$

$$\frac{\mathrm{d}I(z)}{\mathrm{d}z} = -(G+\mathrm{j}\omega C)\,V(z),$$

$$\therefore$$

$$\frac{\mathrm{d}^2V(z)}{\mathrm{d}z} = (R+\mathrm{j}\omega L)(G+\mathrm{j}\omega C)\,V(z),$$

$$\frac{\mathrm{d}^2V(z)}{\mathrm{d}z} - (R+\mathrm{j}\omega L)(G+\mathrm{j}\omega C)\,V(z) = 0,$$

$$\frac{\mathrm{d}^2V(z)}{\mathrm{d}z} - k^2\,V(z) = 0, \tag{6}$$

where a complex propagation constant k is defined as

$$k = \Re(k) + j\Im(k) = \sqrt{(R+\mathrm{j}\omega L)(G+\mathrm{j}\omega C)} \tag{7}$$

and is a function of the transmission line geometry (keep in mind that R, L, C, and G are all distributed parameters calculated separately for the specific shape of the conductor).

Repeating the same procedure, starting with the spatial differential of (5), a similar solution to (6) is found for the current spatial dependence, i.e.

$$\frac{\mathrm{d}^2I(z)}{\mathrm{d}z} - k^2\,I(z) = 0. \tag{8}$$

Equations (6) and (8) show explicitly the voltage or current spatial dependence along the transmission line. Solutions to these two decoupled equations are well known to be of the following form (assuming, of course, that the transmission line is aligned with the z axis):

$$V(z) = V^{+}\mathrm{e}^{-kz} + V^{-}\mathrm{e}^{+kz}, \tag{9}$$

$$I(z) = I^{+}\mathrm{e}^{-kz} + I^{-}\mathrm{e}^{+kz}. \tag{10}$$

By convention, each of these equations is interpreted as a combination of two waveforms, one propagating in the positive z direction and the other in the negative z direction.

The two equations (9) and (10) are correlated because they describe the same waveform, which means that they are connected through the transmission line impedance. For example, substituting (9) back into (4) results in an explicit relationship between the voltage $V(z)$ and current $I(z)$:

$$\begin{aligned} -\frac{\mathrm{d}V(z)}{\mathrm{d}z} &= (R + \mathrm{j}\omega L)\, I(z), \\ &\therefore \\ I(z) &= -\frac{1}{(R + \mathrm{j}\omega L)} \frac{\mathrm{d}V(z)}{\mathrm{d}z} \\ &= -\frac{1}{(R + \mathrm{j}\omega L)} \frac{\mathrm{d}(V^{+}\mathrm{e}^{-kz} + V^{-}\mathrm{e}^{+kz})}{\mathrm{d}z} \\ &= \frac{k}{(R + \mathrm{j}\omega L)} (V^{+}\mathrm{e}^{-kz} - V^{-}\mathrm{e}^{+kz}), \end{aligned} \tag{11}$$

which is to say that the expression connecting current $I(z)$ and voltage $V(z)$ must be impedance. Because it is an important parameter of a transmission line, it is named *characteristic line impedance* Z_0. After substituting (7)

$$Z_0 = \frac{(R + \mathrm{j}\omega L)}{k} = \sqrt{\frac{(R + \mathrm{j}\omega L)}{(G + \mathrm{j}\omega C)}}. \tag{12}$$

In the ideal lossless case, i.e., there is no thermal dissipation $R = G = 0$, this degenerates into

$$Z_0 = \sqrt{\frac{L}{C}}. \tag{13}$$

Discussion: Characteristic impedance of a lossless transmission line (13) is not a function of frequency. That fact should be contrasted with the general definition (12), which is a complex quantity and takes into account the always-present thermal losses (but not inter-component interference). However, characteristic impedance is a very strong function of the line geometry (through the distributed values L and C) and must be calculated for each type of transmission line, for example for two-wire line, coaxial line, parallel-plate line, etc.

1.3 A straightforward implementation of Ampère's law, (B.7) and (B.8), is repeated here for convenience,

$$\oint_L \mathbf{H} \cdot \mathrm{d}\mathbf{l} = I_{\text{free, enc}} + \frac{\mathrm{d}}{\mathrm{d}t} \int_S \mathbf{D} \cdot \mathrm{d}\mathbf{s} \qquad \text{integral form,} \tag{14}$$

$$\nabla \times \mathbf{H} = \mathbf{J}_{\text{free}} + \frac{\partial \mathbf{D}}{\partial t} \qquad \text{differential form.} \tag{15}$$

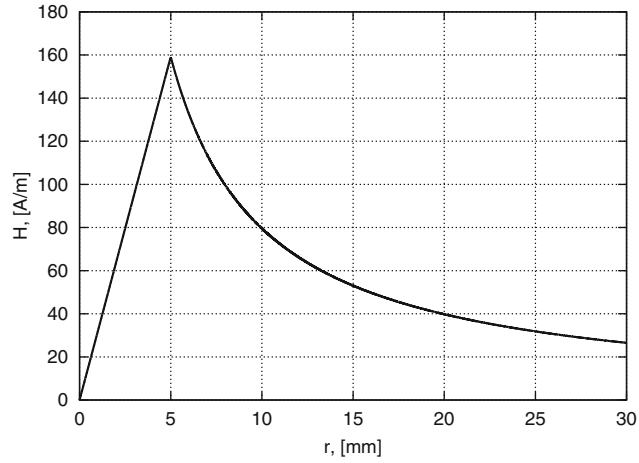


Fig. 1 The magnetic field distribution inside and outside an infinitely long wire of radius $a = 5\,\text{mm}$ carrying a current of 5 A

For a constant current $I(t) = const$ and all the charges contributing to the current flow, the current density $\mathbf{J}$ is uniform through any cross-section area of radius r inside the conductor, up to the radius a at the conductor's surface. Hence, the portion of the total current I_r flowing through any area with radius r ($0 \leq r \leq a$) inside the conductor is determined by the ratio between the full cross-section area πa^2 and the inside area πr^2, i.e., Ampère's law can be written as

$$H \cdot 2\pi r = I \frac{\pi r^2}{\pi a^2} \quad \therefore \quad H = \frac{Ir}{2\pi a^2}, \tag{16}$$

where $0 \leq r \leq a$. Outside the conductor, the current density is equal to zero, which simplifies (14) or (15) so that the magnetic field H outside the conductor is calculated as

$$H \cdot 2\pi r = I \quad \therefore \quad H = \frac{I}{2\pi r}, \tag{17}$$

where $r \geq a$. The total magnetic field outside and inside the infinitely long conductive wire is

$$H(r) = \begin{cases} \frac{Ir}{2\pi a^2} = 31.810 \times 10^3 r \ \text{A/m} & r \leq 5\,\text{mm} \\ \frac{I}{2\pi r} = \frac{0.798}{r} \ \text{A/m} & r \geq 5\,\text{mm} \end{cases}. \tag{18}$$

The graph of this radial magnetic field distribution is plotted in Fig. 1.

Discussion: It is important to notice that the magnetic field increases linearly inside the conductor because more current contributes to the magnetic field. Outside the wire, the magnetic field strength is inversely proportional to the distance because the whole current has been accounted for and there are no more contributors to the field. This problem is a typical application of Maxwell's equations without any approximations.

1.4 The voltage induced in the loop is equal to the line integral of the electric field $\mathbf{E}$ along the loop. Employing Faraday's law, (B.5) and (B.6), repeated here for convenience

$$\oint_L \mathbf{E} \cdot \mathrm{d}\mathbf{l} = -\frac{\mathrm{d}}{\mathrm{d}t} \int_S \mathbf{B} \cdot \mathrm{d}\mathbf{s} \qquad \text{integral form,} \tag{19}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \text{differential form} \tag{20}$$

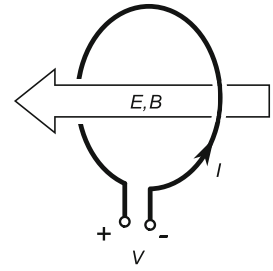


Fig. 2 The time rate of change of the magnetic flux density induces a voltage

results in

$$\begin{aligned} V &= -\oint_L \mathbf{E}\cdot \mathrm{d}\mathbf{l} = \frac{\mathrm{d}}{\mathrm{d}t}\int_S \mathbf{B}\cdot \mathrm{d}\mathbf{s} = \frac{\mathrm{d}}{\mathrm{d}t}\int_S \mu_0\mathbf{H}\cdot \mathrm{d}\mathbf{s} \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\int_S \mu_0 H_0 \cos(\omega t)\mathbf{n}\cdot \mathrm{d}\mathbf{s} \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\mu_0 H_0 \cos(\omega t)\,\pi a^2\mathbf{n} \\ &= -\pi a^2 \omega \mu_0 H_0 \sin(\omega t) \\ &= -0.31\sin(6.28\times 10^8 t)\ \ \mathrm{V}, \end{aligned}$$

where $\mathbf{B} = \mu_0\mathbf{H}$ is the magnetic flux density and $\mathbf{n}$ is the unity vector in the same direction as magnetic field $\mathbf{H}$ vector.

Discussion: This is a typical example of Maxwell's equation in a form of Faraday's law (also known as the "transformer law"), where the time-varying field induces a voltage response in a conductive loop (see Fig. 2).

Solutions to Selected Problems in Chapter 2

2.5 The instantaneous power exists during the times when the pulse amplitude is not zero. Because the pulse amplitude is constant, we write

$$p(t) = \frac{v^2(t)}{R} = \frac{(2V)^2}{100} = 40\,\mathrm{mW}.$$

Hence, the total energy is

$$W = Pt = 40\,\mathrm{mW}\times 1\,\mathrm{ms} = 40\,\mu\mathrm{J}.$$

2.6 4.4A; 6.6A; 17.5C; 3.5A.

2.7 −56 W; 16 W; −60 W; 160 W; −60 W.

2.8 3.94 kΩ; 1.890 W; −30.0 mA; 46.9 μS.

2.9 v_{R2}= 32 V; v_x=6 V.

2.10 −960 W; 1920 W; −1920 W; 960 W.

2.11 V_{Th}=8 V; R_{Th} = 10 kΩ.

2.12 $a; a/\sqrt{2}; a/\sqrt{3}$.

2.13 208.333 W; 6.455 A.

2.14 26 W; 2 W.

Solutions to Selected Problems in Chapter 3

3.1

(a) $$S_n = 1.38 \times 10^{-23} \times 300 = 4.14 \times 10^{-21}\,\text{W/Hz}.$$

(b) $$P_n = 4.14 \times 10^{-21} \times 10^6 = 4.14 \times 10^{-15}\,\text{W}.$$

(c) $$P_s = \frac{1 \times 10^{-6}/2}{50} = 5 \times 10^{-15}\,\text{W}.$$

(d) $$SNR = \frac{P_s}{P_n} = \frac{5 \times 10^{-15}}{4.14 \times 10^{-15}} = 0.82\,\text{dB}.$$

3.3 (a) For the two resistors separately, from (3.3) it follows that

$$E_n^2(R_1) = 4 \times 20\,\text{k}\Omega \times 1.38 \times 10^{-23} \times 290\,\text{K} \times 100\,\text{kHz} = 32 \times 10^{-12}\,\text{V}^2,$$
$$E_n^2(R_2) = 4 \times 50\,\text{k}\Omega \times 1.38 \times 10^{-23} \times 290\,\text{K} \times 100\,\text{kHz} = 80 \times 10^{-12}\,\text{V}^2,$$
$$\therefore$$
$$E_n(R_1) = 5.658\,\mu\text{V},$$
$$E_n(R_2) = 8.946\,\mu\text{V}.$$

(b) Serial resistance is $R_s = 70\,\text{k}\Omega \quad \therefore \quad E_n(R_s) = 10.59\,\mu\text{V}.$
(c) Parallel resistance is $R_p = 14.286\,\text{k}\Omega \quad \therefore \quad E_n(R_s) = 4.78\,\mu\text{V}.$

3.4 From (5.82), the dynamic resistance of the LC resonator at resonance is calculated as

$$R_D = \frac{Q}{\omega_0 C} = \frac{30}{2\pi \cdot 120\,\text{MHz} \cdot 25\,\text{pF}} = 1.59\,\text{k}\Omega$$

then from (3.12)

$$V_n^2 = 4Q^2 R_L k T \Delta f = 4 R_D k T \Delta f = 0.254 \times 10^{-12}\,\text{V}^2,$$
$$\therefore$$
$$V_n = 0.50\,\mu\text{V}.$$

3.5 Application of *Thévenin's theorem* on the E_s, R_s, and R_i network results in the following:

$$R_t = \frac{R_s R_i}{R_s + R_i} = 46.15\,\Omega,$$

$$V_t = V_s \frac{R_i}{R_s + R_i} = 0.923\,\mu V.$$

The equivalent noise voltage at the amplifier input is calculated for the serial combination of $R_t + R_n = 446.15\,\Omega$, which after applying (3.3), results in $V_n = 0.267\mu V$.

3.9 Using (3.39) we find

$$I_n = \sqrt{2q_e I_{DC}\Delta f} = \sqrt{2 \times 1.602 \times 10^{-19} C \times 1\,mA \times 1 \times 10^6 Hz} = 17.90\,nA.$$

Voltage across the p–n junction and dynamic diode resistance r_D are

$$V_T = \frac{kT}{q} = 25.843\,mV,$$

$$\therefore$$

$$r_D = \frac{V_T}{I_{DC}} = 25.843\,\Omega.$$

Therefore, the noise current I_n through the diode resistance r_D generates noise voltage $V_n = I_n r_D = 17.90\,nA \times 25.843\,\Omega = 462.564\,nV$.

3.10 The shot noise current goes through the source resistance, i.e.

$$I_n = \sqrt{2q_e I_{DC}\Delta f} = \sqrt{2 \times 1.602 \times 10^{-19} C\,5\,\mu A\,10 \times 10^6 Hz} = 4\,nA,$$

$$\therefore$$

$$V_{ns}(R_S) = I_n \times R_S = 4\,nA \times 150\,\Omega = 600\,nV,$$

while the noise across the amplifier resistance R_{in} is generated as

$$V_n(R_{in}) = \sqrt{4R_{in}kT\Delta f} = \sqrt{4 \times 300 \times 1.38 \times 10^{-23} \times 300 \times 10 \times 10^6} = 7.048\,\mu V$$

and the thermal noise from the source is

$$V_{nt}(R_S) = \sqrt{4R_S kT\Delta f} = \sqrt{4 \times 150 \times 1.38 \times 10^{-23} \times 300 \times 10 \times 10^6} = 4.984\,\mu V.$$

Therefore, the total noise at the input of the amplifier is

$$V_n = \sqrt{V_{ns}^2(R_S) + V_{nt}^2(R_S) + V_n^2(R_S)} = 8.653\,\mu V,$$

so that the $SNR(in)$ is calculated by definition as

$$SNR(in) = 20\log\frac{V_S}{V_n} = 20\log\frac{10 \times 10^{-6}}{8.653 \times 10^{-6}} = 1.257\,dB.$$

3.11 First convert the dB values into the numbers, i.e., $12\,dB = 15.85$ and $50\,dB = 10^5$. Hence, $T_{rec} = (15.85 - 1) \times 300 = 4455\,K$ and $T_{sys} = 90\,K + \frac{4455\,K}{10^5} \approx 90\,K$.

Solutions to Selected Problems in Chapter 5

5.7 First, let us determine the self-resonant frequency f_{L0} of this coil as

$$f_{L0} = \frac{1}{2\pi\sqrt{1\,\mu\text{H} \times 5\,\text{pF}}} = 71.2\,\text{MHz}$$

and the Q factor is

$$Q = \frac{2\pi\, 25\,\text{MHz}\, 1\,\mu\text{H}}{5\,\Omega} = 31.4.$$

Hence, using (5.64), we write

$$L_{\text{eff}} = \frac{L}{1 - \left(\frac{f}{f_{L0}}\right)^2} = \frac{1\,\mu\text{H}}{1 - \left(\frac{25\,\text{MHz}}{71.2\,\text{MHz}}\right)^2} = 1.14\,\mu\text{H}$$

and using (5.68), we write

$$Q_{\text{eff}} = Q\left[1 - \left(\frac{f}{f_{L0}}\right)^2\right] = 31.4\left[1 - \left(\frac{25\,\text{MHz}}{71.2\,\text{MHz}}\right)^2\right] = 27.5.$$

5.8 At $f = 10\,\text{kHz}$, we have,

$$\begin{aligned} X_L &= 2\pi\, 10\,\text{kHz}\, 3\,\text{mH} = 188.5\,\Omega, \\ X_C &= \frac{1}{2\pi\, 10\,\text{kHz}\, 100\,\text{nF}} = 159.2\,\Omega, \\ Z &= \sqrt{R^2 + (X_L - X_C)^2} = \sqrt{30^2 + (188.5 - 159.2)^2}\,\Omega = 41.9\,\Omega. \end{aligned}$$

We note that the serial RLC circuit looks more inductive at 10 kHz.

At $f = 5\,\text{kHz}$, we have,

$$\begin{aligned} X_L &= 2\pi\, 5\,\text{kHz}\, 3\,\text{mH} = 94.2\,\Omega, \\ X_C &= \frac{1}{2\pi\, 5\,\text{kHz}\, 100\,\text{nF}} = 318.3\,\Omega, \\ Z &= \sqrt{R^2 + (X_L - X_C)^2} = \sqrt{30^2 + (94.2 - 318.3)^2}\,\Omega = 226.1\,\Omega. \end{aligned}$$

We note that the serial RLC circuit looks more capacitive at 5 kHz.

5.11 The two networks must have the same Q factor, which is found by definition as

$$Q_S = \frac{X_S}{R_S} = \frac{1}{2\pi C_S R_S} = \frac{1}{2\pi\, 7.95\,\text{pF}\, 10\,\Omega} = 2. \tag{21}$$

Using (5.77) and (5.78), we write

$$R_p = R_s(1 + Q^2) = 10\,\Omega(1 + 2^2) = 50\,\Omega,$$

$$X_p = X_s\left(1+\frac{1}{Q^2}\right) = \frac{1}{2\pi\,1\,\text{GHz}\,7.95\,\text{pF}}\left(1+\frac{1}{2^2}\right) = 25.024\,\Omega \quad \therefore \quad C_P = 6.36\,\text{pF}.$$

Solutions to Selected Problems in Chapter 7

7.11

(a) V_B: $V_{CC} = 9V \quad \Rightarrow \quad V_B = \frac{1}{3}V_{CC} \quad \therefore \quad V_B = 3\,\text{V}.$

(b) R_1 *and* R_2*:*

$$(R_1+R_2) = \frac{V_{CC}}{I_{R_1R_2}} = \frac{9\,\text{V}}{\frac{1}{10}2\,\text{mA}} = 45\,\text{k}\Omega, \tag{22}$$

$$\frac{R_2}{R_1+R_2} = \frac{1}{3} \quad \text{(because } V_{R_2} = 3\,\text{V and } V_{R_1+R_2} = 9\,\text{V)} \tag{23}$$

$$\therefore$$

$$R_1 = 30\,\text{k}\Omega, \quad R_2 = 15\,\text{k}\Omega.$$

(c) Thévenin resistance: $R_{th} = R_1||R_2||R_{sig} = 5\,\text{k}\Omega$.

(d) R_E: Let us compare calculated values for R_E when we reflect R_{th} back to the emitter side and when we ignore the reflected resistance.

Including the reflected base resistance, resistance at the emitter side becomes:

$$R_E' = R_E + \frac{R_{th}}{\beta+1} = R_E + \frac{5\,\text{k}\Omega}{101} \approx R_E + 50\,\Omega. \tag{24}$$

In any case, we calculate

$$I_E = \frac{V_E}{R_E} = \frac{V_B - V_{BE}}{R_E} = \frac{2.3\,\text{V}}{R_E} \approx 2\,\text{mA}. \tag{25}$$

From the above equation, $R_E = 1.15\,\text{k}\Omega$ when ignoring the reflected base resistance and $R_E = 1.10\,\text{k}\Omega$ when we include it. Use your engineering judgement ...

The closest 10% standard value is $R_E = 1\,\text{k}\Omega$.

(e) I_C: For $\beta = 100$, it follows that $\alpha = \beta/(\beta+1) = 0.99$ and $I_C = \alpha I_E$=1.98 mA. Again, use your engineering judgement to decide if you would use $I_C \approx I_E$ instead.

(f) g_m:

$$g_m = \frac{I_C}{V_T} \approx \frac{2\,\text{mA}}{25\,\text{mV}} = 80\frac{\text{mA}}{\text{V}} = 80\,\text{mS}. \tag{26}$$

(g) r_e:

$$r_e = \frac{1}{g_m} = \frac{V_T}{I_C} \approx \frac{1}{80\,\text{mS}} = 12.5\,\Omega. \tag{27}$$

(h) R_C: Because of the capacitance in parallel with R_E, the voltage gain is set as

$$A_v = -\frac{R_C}{r_e} \quad \Rightarrow \quad R_C = -(-8)80\,\Omega = 640\,\Omega. \tag{28}$$

What is R_C if you use $R_E||r_e$ instead? Again, use your engineering judgement ....

7.12 *DC setup*:

(a) When output voltage equals 1/2 of the power supply voltage, we have:

$$V_{out} = \frac{1}{2}V_{CC} = 5\,V,$$

$$\therefore$$

$$R_C = \frac{V_{CC} - V_{out}}{I_C} = 5\,k\Omega.$$

(b) To set V_E=1 V we need:

$$I_E \approx I_C \quad \Rightarrow \quad R_E = \frac{V_E}{I_E} = 1\,k\Omega.$$

(c) After setting V_E, it follows that

$$V_B = V_E + V_{BE} = 1.6\,V.$$

(d) Resistance R_{in} looking into the base is

$$R_{in} \approx (\beta + 1)\,R_E = 100\,k\Omega.$$

(e) Voltage across the R_2 resistor is $V_{R_2} = V_B$, while voltage across the R_1 resistor is $V_{R_1} = V_{CC} - V_B$, or by writing full equations we have:

$$V_{R_2} = I_{R_2}R_2 = \frac{V_{CC}}{R_1 + R_2}R_2 \quad \Rightarrow \quad \frac{V_B}{V_{CC} - V_B} = \frac{R_2}{R_1} = \frac{1}{5.25}.$$

Acceptable solutions for setting the bias resistors and still meeting the requirement for V_B are achieved by any combination of R_1 and R_2 resistors with a ratio of 5.25. For example, one possible way is to say that $R_{th} \leq R_1||R_2 = 0.1R_{in}$, which leads to

$$R_{th} \leq R_1||R_2 = 0.1R_{in} \leq 10\,k\Omega.$$

One good choice is $R_2 = 10\,k\Omega$, which leads to $R_1 = 5.25R_2 = 52.5\,k\Omega$.

(f) The small emitter resistor is, by definition

$$r_e = \frac{V_T}{I_C} = 25\,\Omega.$$

AC setup:

(g) For the required A, we have

$$A = \frac{R_C}{R_E||(r_e + R_0)} \quad \Rightarrow \quad R_0 = 27.63\,\Omega. \tag{29}$$

(h) Output 3 dB point is set by C_E and $(R_0 + r_e)$, so

$$C_E = \frac{1}{2\pi f_{3\,dB}\,(R_0 + r_e)} \quad \Rightarrow \quad C_E = 151.2\,\mu F.$$

(i) At the input side, the high-pass filter is set by C and $R_{th}||(\beta+1)(r_e + R_0)$. Note that for small signal analysis, the relevant resistance is different from DC, i.e., the relevant resistance is

$$R_{eq} = R_{th}||(\beta+1)(r_e + R_0) = R_1||R_2||(\beta+1)(r_e + R_0) = 3.236\,k\Omega,$$

which leads to

$$C = \frac{1}{2\pi f_{3\,dB} R_{eq}} \quad \Rightarrow \quad C = 4.919\,\mu F.$$

(j) When output voltage drops to V_{out}=2.5 V the collector resistor current changes as

$$I_C = \frac{V_{R_C}}{R_C} = \frac{V_{CC} - V_{out}}{R_C} = 1.5\,mA.$$

Consequently, for the new collector current we have

$$r_e = \frac{V_T}{I_C} = 16.67\,\Omega,$$

which means that, using gain equation (29), for the new value of r_e, we have $A = -117.9$ V/V = 41.43 dB.

7.13 The assumption that $\beta = \infty$ means that $I_B = 0$ and $I_C = I_E$, so we write:

(a) Collector voltage is V_{R_C} below V_{CC}, i.e.,

$$I_C = I = 0.5\,mA, \tag{30}$$

$$V_C = V_{CC} - R_C I_C = 5\,V - 7.5\,k\Omega 0.5\,mA = 1.25\,V. \tag{31}$$

(b) By definition,

$$g_m = \frac{I_C}{V_T} = \frac{0.5\,mA}{25\,mV} = 20\,mS. \tag{32}$$

(c) We need to recognize that $V_{BE} = -v_i$. (The base is at ground, input is at the emitter, output is at the collector—this is a CB configuration.)

$$v_C = R_C\,i_C = -R_C\,g_m V_{BE} = R_C\,g_m\,v_i, \tag{33}$$

$$\therefore$$

$$A = \frac{v_C}{v_i} = g_m R_C = 20\frac{\text{mA}}{V}\,7.5\,\text{k}\Omega = 150\frac{\text{V}}{\text{V}} = 43.52\,\text{dB}. \tag{34}$$

7.14 From the emitter side, we see $100\,\Omega = R_b/(\beta+1)$.
From the base side, we see $100\,\text{k}\Omega = (\beta+1)\,R_e$.
Then, $R_b = 10\,\text{k}\Omega$ and $R_e = 1\,\text{k}\Omega$.

7.17 The resonant LC frequency is set as

$$f_0 = \frac{1}{2\pi\sqrt{LC_M}}. \tag{35}$$

In this case, capacitor C is reflected back to the input stage as Miller capacitance, which is calculated as

$$C_M = (A+1)\,C = 100\,\text{pF}.$$

Since there are no other capacitances, from (35) we get: $L = 1\,\mu\text{H}$.

Solutions to Selected Problems in Chapter 8

8.1 Following the signal through the loop, first along the forward path through the amplifier and then back through the feedback network, the loop equations are written by inspection as

$$\begin{aligned} V_{out} &= A\,(V_{in} + V_{fb}),\\ V_{fb} &= \beta\,V_{out},\\ &\therefore\\ V_{out} &= A\,(V_{in} + \beta\,V_{out}),\\ &\therefore\\ \frac{V_{out}}{V_{in}} &= \frac{A}{1-A\beta}, \end{aligned} \tag{36}$$

which implies that, for $A\beta = 1$, the loop gain becomes infinite, i.e., the loop is unstable and starts to oscillate.

Note: for the case of an oscillator, the role of the input voltage V_{in} is taken by the internal noise, which is sufficient to start the oscillations.

8.2 A common-emitter amplifier is inverting, therefore, by inspection, the total loading resistance R_o and output voltage v_{out} are

$$\begin{aligned} R_o &= R_C\\ v_{out} &= -R_o\,i_c = -R_o\,\beta\,i_b. \end{aligned}$$

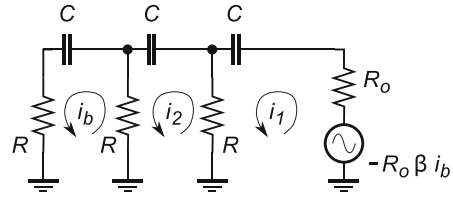


Fig. 3 Simplified schematic of a phase oscillator for Solution 8.2

From the perspective of the feedback network, the CE amplifier behaves as a voltage source with R_o source resistance, while the feedback loop is maintained through the branch with current i_b. Therefore, the circuit equations are set in accordance with the equivalent circuit in Fig. 3. The voltage loop equations are:

$$-R_\mathrm{o}\beta\, i_\mathrm{b} = \left(R_\mathrm{o} + R - \frac{\mathrm{j}}{\omega C}\right) i_1 - R\, i_2, \tag{37}$$

$$0 = -R\, i_1 + \left(R + R - \frac{\mathrm{j}}{\omega C}\right) i_2 - R\, i_\mathrm{b}, \tag{38}$$

$$0 = -R\, i_2 + \left(R + R - \frac{\mathrm{j}}{\omega C}\right) i_\mathrm{b}. \tag{39}$$

The system (37) to (39) can be solved in number of ways. One possible approach is to introduce substitution $Z = R - \mathrm{j}/\omega C$ and eliminate the three currents to arrive at

$$(R_\mathrm{o} + Z)(Z^2 + 2RZ) - R(R^2 + RZ - \beta R R_\mathrm{o}) = 0, \tag{40}$$

which, after a bit of straightforward algebra, results in the polynomial

$$Z^3 + (2R + R_\mathrm{o})Z^2 + (2RR_\mathrm{o} - R^2)Z + \beta R^2 R_\mathrm{o} - R^3 = 0. \tag{41}$$

A complex number equals zero if both its real and complex parts are zero, which is to say that the real and complex parts of (41) are

$$\Re: \quad \left[R^3 - \frac{3R}{(\omega C)^2}\right] + (2R + R_\mathrm{o})\left(R^2 - \frac{1}{(\omega C)^2}\right) + (2RR_\mathrm{o} - R^2)R + \beta R^2 R_\mathrm{o} - R^3 = 0, \tag{42}$$

$$\Im: \quad \mathrm{j}\left\{-\frac{3R^2}{\omega C} + \frac{1}{(\omega C)^3} - (2R + R_\mathrm{o})\frac{2R}{\omega C} - (2RR_\mathrm{o} - R^2)\frac{1}{\omega C}\right\} = 0. \tag{43}$$

Solving the imaginary part (43), after removing the substitution, delivers the formula for the resonant frequency as

$$\omega = \frac{1}{RC\sqrt{4\frac{R_\mathrm{o}}{R} + 6}}. \tag{44}$$

At the same time, the real part (42) is solved for β as

$$\beta = 23 + 4\frac{R_\mathrm{o}}{R} + 29\frac{R}{R_\mathrm{o}}. \tag{45}$$

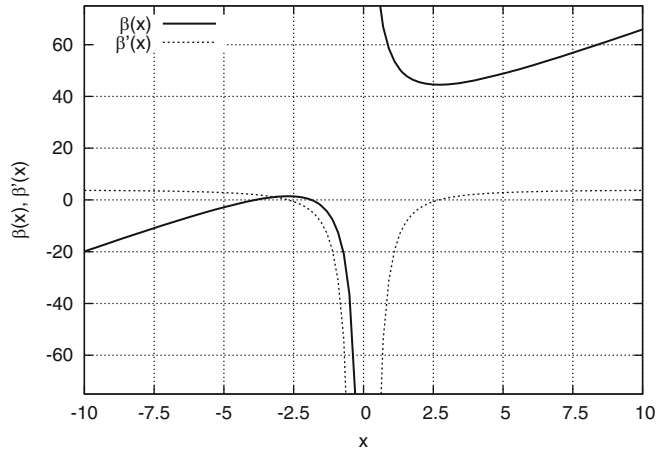


Fig. 4 Plot of $\beta(x)$ and $\beta'(x)$ functions for Solution 8.2

Equation (45) solves for β as a function of the resistor ratio $x = R_o/R$ (see Fig. 4). The minimum of this function is found easily by setting its first derivative to zero as

$$\beta(x) = \frac{29}{x} + 4x + 23,$$

$$\therefore$$

$$\beta'(x) = -\frac{29}{x^2} + 4,$$

$$\therefore$$

$$x \approx \pm 2.6926.$$

There are two possible solutions for x. In this case, the positive value is taken to calculate the resistors, hence $R_o/R \approx 2.6926$; after being substituted in (45), that produces

$$\beta_{\min} \approx 44.5. \tag{46}$$

It should be noted that the β value does not depend upon specific values of R_o and R, only on their ratio.

From (46) and $R_o = R_C = 10\,\text{k}\Omega$, it follows that $R = R_o/2.6926 = 3.714\,\text{k}\Omega$. At $f = 10\,\text{MHz}$ from (44), it follows that $C \approx 1\,\text{pF}$.

8.3 The circulating resonant current i_c in a tapped L, centre-grounded network perceives the L_1, L_2, and C components in series, therefore $L_{\text{eff}} = L_1 + L_2 = 2\,\mu\text{H}$. The resonant frequency is calculated as

$$f_0 = \frac{1}{2\pi\sqrt{L_{\text{eff}}C}} = 10\,\text{MHz}.$$

8.4 By inspection, the input voltage is distributed across the L_2 inductor, while the output voltage is distributed across the L_1 inductor. Since the same resonating current i_c is circulating through both components, it is straightforward to write

$$v_{\text{in}} = i_c\,\text{j}\omega L_2,$$

$$v_{\text{out}} = -i_c\,\text{j}\omega L_1,$$

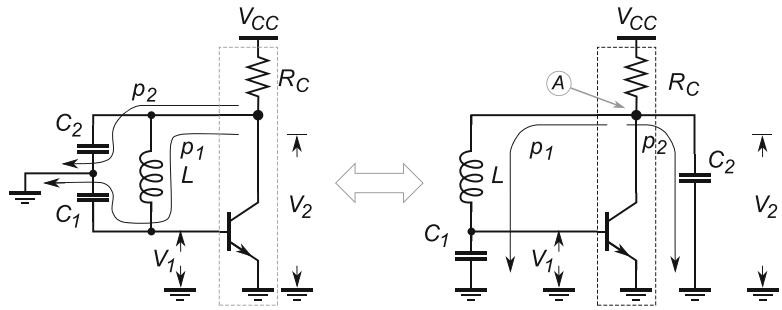


Fig. 5 Simplified schematic transformation of LC oscillator for Solution 8.7

$$\therefore \quad \beta = \frac{v_{\text{out}}}{v_{\text{in}}} = -\frac{i_c \, \mathrm{j}\omega L_1}{i_c \, \mathrm{j}\omega L_2} = -\frac{L1}{L2} = -\frac{0.5\,\mu\text{H}}{1.5\,\mu\text{H}} = -0.333. \tag{47}$$

8.5 Using the formula provided in the textbook and knowing that the R_L resistor of the feedback network (Fig. 8.5) is in fact the input resistance of the amplifier, a straightforward implementation of the formula yields

$$\begin{aligned} R_{\text{eff}} &= R_L \left(\frac{L_2}{L_1}\right)^2 || \frac{Q\omega_0 L_2^2}{L_1 + L_2} \\ &= 10\,\text{k}\Omega \left(\frac{1.5\,\mu\text{H}}{0.5\,\mu\text{H}}\right)^2 || \frac{50 \times 2\pi \times 10\,\text{MHz} \times (1.5\,\mu\text{H})^2}{2\,\mu\text{H}} \\ &= 90\,\text{k}\Omega \,||\, 3.534\,\text{k}\Omega = 3.4\,\text{k}\Omega. \end{aligned} \tag{48}$$

8.7 Although, by now we already know how to estimate the resonant frequency of this circuit, let us take the opportunity to develop a possible methodology to solve this kind of circuit in a more general way.

First, let us rearrange the circuit network so that it becomes more obvious how the network equations are going to be written. Figure 5 (left) shows two paths, p_1 and p_2, from the collector node through the feedback network to ground. Following the components on each of the two paths, it is straightforward to redraw the equivalent circuit diagram to look like Fig. 5 (right).

It then becomes easy to generalize components in each branch of the circuit and to introduce the equivalent subnetwork that represents the amplifier itself (grey box that contains BJT and R_C), whose function is to be a voltage-controlled current source, i.e., collector current $i_C = f(V_1)$.

By inspection, the equivalent resistance R_o at the collector node Ⓐ is

$$R_o = R_C || r_C || R_D, \tag{49}$$

where R_C is the given real resistor (looking up into the R_C from node Ⓐ), r_C is the output resistance of BJT at its biasing point (looking down into the collector), and R_D (looking left into the LC resonator) is the dynamic resistance at resonant frequency. From the signal perspective, the three resistances are in parallel, i.e., connected between node Ⓐ and signal ground. In addition, the condition $Q \to \infty$ implies that $R_D = f(Q^2) \to \infty$, i.e., it has no influence on R_o and can be ignored in (49).

The BJT amplifier (the grey box in Fig. 5) is then replaced with its equivalent current source whose current is $g_m v_1$ and output resistance is R_o, Fig. 6 (left). Note that the feedback loop is maintained through the controlling voltage v_1.

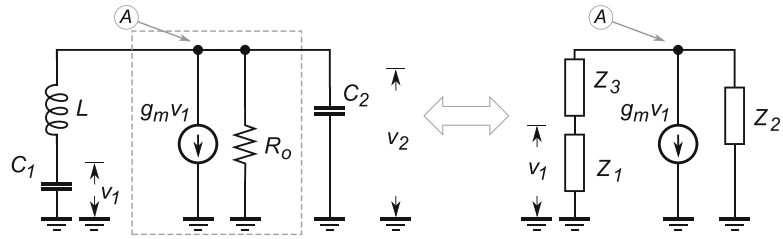


Fig. 6 Simplified schematic transformation of LC oscillator for Solution 8.7

Finally, in order to simplify the incoming analytical expressions, the last step in this network transformation is to substitute the real RLC components with general Z_1, Z_2, and Z_3 impedances, Fig. 6 (right), where

$$Z_1 = -\mathrm{j}\frac{1}{\omega C_1},$$
$$Z_2 = R_\mathrm{o} \,||\, \left(-\mathrm{j}\frac{1}{\omega C_2}\right),$$
$$Z_3 = \mathrm{j}\omega L.$$

With this last substitution and transformation, it becomes straightforward to write the KCL current equation at node Ⓐ , after recognizing that the same current flows through Z_1 and Z_3, as

$$-g_\mathrm{m} v_1 = \frac{v_2}{Z_2} + \frac{v_2 - v_1}{Z_3}, \tag{50}$$
$$\frac{v_1}{Z_1} = \frac{v_2}{Z_1 + Z_3}, \tag{51}$$

which leads to

$$-g_\mathrm{m} v_1 = \frac{v_2}{Z_2} + \frac{v_2}{Z_3} - \frac{v_1}{Z_3}, \tag{52}$$
$$v_1 = \frac{Z_1}{Z_1 + Z_3} v_2. \tag{53}$$

By substituting (53) into (52), it follows that

$$-g_\mathrm{m}\frac{Z_1}{Z_1 + Z_3} v_2 = v_2\left(\frac{1}{Z_2} + \frac{1}{Z_3}\right) - \frac{Z_1}{Z_1 + Z_3} v_2,$$
$$\therefore$$
$$-g_\mathrm{m} Z_1 = \frac{Z_1 + Z_3}{Z_2} + \frac{Z_1 + Z_3}{Z_3} - \frac{Z_1}{Z_3}\frac{Z_1 + Z_3}{Z_1 + Z_3},$$
$$-g_\mathrm{m} Z_1 = \frac{Z_1}{Z_2} + \frac{Z_3}{Z_2} + \frac{Z_1}{Z_3} + 1 - \frac{Z_1}{Z_3},$$
$$-g_\mathrm{m} Z_1 Z_2 = Z_1 + Z_2 + Z_3. \tag{54}$$

The resulting (54) is general, in the sense that the three impedances Z_1, Z_2, and Z_3 may be any combination of RLC components, not necessarily the ones we started with in this problem. Indeed, CE amplifiers that use the other feedback networks studied in this course could be solved by applying the same methodology as in this example.

In this particular example, it is easier to switch to *admittances* for the reactive components (of course, the final result must be the same). Following up (54), we write

$$-g_\mathrm{m} = \frac{Z_1+Z_2}{Z_1Z_2} + \frac{Z_3}{Z_1Z_2} = \frac{1}{Z_1} + \frac{1}{Z_2} + \frac{1}{Z_1}\frac{1}{Z_2}Z_3 = Y_1 + Y_2 + Y_1Y_2Z_3,$$

$$\therefore$$

$$-g_\mathrm{m} = \mathrm{j}\omega C_1 + \left(\frac{1}{R_\mathrm{o}} + \mathrm{j}\omega C_2\right) + (\mathrm{j}\omega C_1)\left(\frac{1}{R_\mathrm{o}} + \mathrm{j}\omega C_2\right)\mathrm{j}\omega L$$

$$= \left\{\frac{1}{R_\mathrm{o}} - \frac{\omega^2 LC_1}{R_\mathrm{o}}\right\} + \mathrm{j}\omega\left\{(C_1+C_2) - \omega^2 LC_1C_2\right\}. \tag{55}$$

A condition of resonance is that the imaginary part of (55) equals zero, which directly leads to the expression for the resonant frequency as,

$$C_1 + C_2 = \omega^2 LC_1C_2,$$

$$\therefore$$

$$\omega_0^2 = \frac{1}{L\frac{C_1C_2}{C_1+C_2}} = \frac{1}{LC_\mathrm{s}}, \tag{56}$$

where C_s is the equivalent series capacitance of C_1 and C_2. Expression (56) is what we have already seen by inspection of the resonant loop in Fig. 8.6; it is merely reconfirmed by this derivation. For the given data, from (56) it follows that $f_0 = 10\,\mathrm{MHz}$.

Under the condition of oscillation, (55) is left only with its real part, i.e., after substituting ω_0 from (56), we write

$$-g_\mathrm{m} = \frac{1}{R_\mathrm{o}} - \frac{\omega_0^2 LC_1}{R_\mathrm{o}},$$

$$\therefore$$

$$g_\mathrm{m} = \frac{1}{R_\mathrm{o}}\frac{C_1}{C_2}. \tag{57}$$

Therefore, from (57) for the given data, $R_\mathrm{o} = R_\mathrm{C}||r_\mathrm{c} = 5\,\mathrm{k\Omega}$, hence $g_\mathrm{m} = 1/5\,\mathrm{k\Omega} \times 1 = 200\,\mu\mathrm{S}$.

The case of finite $Q = 50$ is worked out by using $Z_3 = r + \mathrm{j}\omega L$ all along, where $r = \omega L/Q = 2.513\,\Omega$. Since the expressions become a bit more complicated, it may be beneficial to use analytical software tools, for example, MAPLE.

It is also recommended that the same problem is solved for other types of LC feedback network using the same methodology.

8.8 The capacitance of a biased varicap diode is calculated as

$$C_\mathrm{D} = \frac{C_0}{\sqrt{1+\frac{|V_\mathrm{D}|}{0.5}}} = \frac{20\,\mathrm{pF}}{\sqrt{1+\frac{|-0.7|}{0.5}}} = 5.16\,\mathrm{pF}.$$

(a) The three capacitances, C_1, C_2, and C_D are perceived by the resonating loop as being in series, hence the total loop capacitance C at zero bias is,

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_0} = 0.0567^{1}/_{\text{pF}} \quad \therefore \quad C = 17.65\,\text{pF},$$

which means that the zero biasing frequency is

$$f_0 = \frac{1}{2\pi\sqrt{17.65\,\text{pF} \times 100\,\mu\text{H}}} \approx 3.789\,\text{MHz}.$$

(b) At $V_D = -7\text{V}$, the total loop capacitance is

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_D} \quad \therefore \quad C = 4.998\,\text{pF},$$

which means that the zero biasing frequency is

$$f_0 = \frac{1}{2\pi\sqrt{4.998\,\text{pF} \times 100\,\mu\text{H}}} \approx 7.126\,\text{MHz}.$$

Solutions to Selected Problems in Chapter 9

9.1 Once the product of the two signals is overlapped with the required single-tone signals, note that the low-frequency tone was embedded into the high-frequency tone as its *envelope*. This property of multiplied signals is fundamental for wireless communications.

9.2 The trigonometric identity $\sin x \times \sin y = {}^{1}/_{2}[\cos(|x-y|) - \cos(x+y)]$ shows that the frequency spectrum of two multiplied tones contains another two tones (and not the original ones): a low-frequency tone with frequency $f_{LF} = |f_1 - f_2|$ and a high-frequency tone $f_{HF} = f_1 + f_2$.

In order to find the frequency of the unknown signals that, after multiplication with a 10 MHz tone produces a 1 kHz signal, we need to look at two possible differences, i.e., 10.001 MHz − 10.000 MHz = 1 kHz and |9.999 MHz − 10.000 MHz| = 1 kHz. Hence, multiplication of a 10 MHz tone with these two tones, i.e., 9.999 MHz and 10.001 MHz, results in two overlapping 1 kHz LF tones in the output frequency spectrum. (Note: the HF tones in the output spectrum are not identical: one is at 20.001 MHz and the other is at 19.999 MHz—their difference is double the LF tone.)

9.3

(a) The received signal and LO frequencies are mixed at the receiver's mixer, therefore the output is the sum and the difference of the two:

$$\text{sum: } 1435\,\text{kHz} + 980\,\text{kHz} = 2415\,\text{kHz},$$

$$\text{difference: } 1435\,\text{kHz} - 980\,\text{kHz} = 455\,\text{kHz}.$$

(b) The receiver is supposed to down-convert the incoming waveform, so IF is $f_{IF} = 455\,\text{kHz}$.

(c) Working backwards from the mixer output, it is straightforward to find the frequencies with the same f_{IF} for the given local oscillator f_{LO}:

$$\text{sum: } 1435\,\text{kHz} + 455\,\text{kHz} = 1890\,\text{kHz},$$

$$\text{difference: } 1435\,\text{kHz} - 455\,\text{kHz} = 980\,\text{kHz}.$$

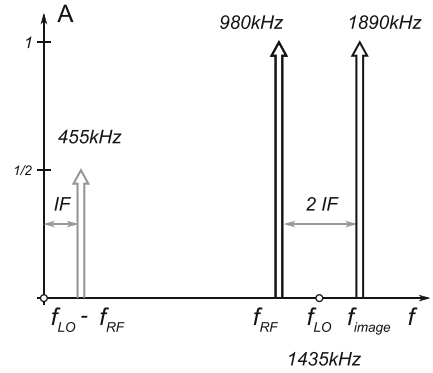


Fig. 7 Graph for Solution 9.3

In other words, a station operating at 1890 kHz will result in the same IF as the wanted station operating at 980 kHz. It would not be possible to separate the two, which means that the other frequency is the image of the wanted frequency.

A practical solution to this problem is to simultaneously tune the tuned RF amplifier at the receiver's input to the same frequency as the wanted frequency.

(d) A graphical representation for the case in which $f_{\mathrm{LO}} > f_{\mathrm{RF}}$ is shown in Fig. 7.

9.4 The formula for relative amplitudes A_{r} of LC tank frequencies is:

$$A_{\mathrm{r}} = \frac{1}{\sqrt{1 + Q^2 \left(\frac{f}{f_0} - \frac{f_0}{f}\right)^2}}.$$

Therefore, for $Q = 20$, $f/f_0 = 1.1/1 = 1.1$ and $f_0/f = 1/1.1 = 0.9090\ldots$. We find that $A_{\mathrm{r}} = 0.253 = -11.93\,\mathrm{dB}$. For comparison, if $Q = 200$, then $A_{\mathrm{r}} = -31.64\,\mathrm{dB}$.

Solutions to Selected Problems in Chapter 10

10.1 From (10.17), we have

$$K_{\mathrm{R}} = \frac{\omega_{3\,\mathrm{dB}}}{K_{\mathrm{PD}}\, K_{\mathrm{VCO}}} = \frac{0.73\,\mathrm{Mrad/s}}{1.27\,\mathrm{V/rad} \times 2\,\mathrm{Mrad/s/V}} = 0.287.$$

The voltage divider obviously has gain

$$K_{\mathrm{R}} = \frac{1}{1 + \frac{R_1}{R_2}} \quad \therefore \quad \frac{R_1}{R_2} = 2.5,$$

which has two unknowns. We pick arbitrary values, for example, if $R_2 = 10\,\mathrm{k\Omega}$, then it follows that $R_1 = 25\,\mathrm{k\Omega}$. We just note that there are additional constraints needed to reach a unique solution.

10.2 The first capture happened at 100 kHz from the centre frequency. Hence, we conclude that the capture range is $10\,\mathrm{MHz} \pm 100\,\mathrm{kHz}$. Similarly, once in lock, the PLL stays locked as long as the input frequency is within $10\,\mathrm{MHz} \pm 500\,\mathrm{kHz}$, which is to say that its lock range is 1 MHz around the centre frequency.

Solutions to Selected Problems in Chapter 11

11.1

(a) Use any plotting tool, if needed.
(b) The AM signal is

$$S = [60 + 15\cos(2\pi 1500t)]\cos(2\pi 100000t).$$

The modulated signal has maximum amplitude of $60 + 15 = 75$ and minimum amplitude of $60 - 15 = 45$.
(c) The modulation factor m (also known as percentage modulation or modulation index) is the ratio of the maximum frequency of the modulated signal to the amplitude of the carrier. Here,

$$m = \frac{15}{60} = 0.25 \qquad \text{or } 25\%.$$

(d) The carrier is at $f_c = 100\,\text{kHz}$ and the signal is at $f_s = 1.5\,\text{kHz}$.
(e) Three frequencies would show at the output in a spectrum analysis: the carrier $f_c = 100\,\text{kHz}$, the lower-sideband frequency $f_c - f_s = (100 - 1.5)\,\text{kHz} = 98.51,\text{kHz}$ and the upper-sideband frequency $f_c + f_s = (100 + 1.5)\,\text{kHz} = 101.5\,\text{kHz}$.

11.2

The AM signal occupies two times the signal frequency (the distance between the upper-sideband and lower-sideband frequencies), i.e., 10 kHz. Therefore only 10 stations can fit into a 100 kHz space.

11.3

(a) The received signal and LO frequencies are mixed at the receiver's mixer, therefore the output is the sum and the difference of the two:

$$\text{sum: } 1435\,\text{kHz} + 980\,\text{kHz} = 2415\,\text{kHz},$$

$$\text{difference: } 1435\,\text{kHz} - 980\,\text{kHz} = 455\,\text{kHz}.$$

(b) The receiver is supposed to down-convert the incoming waveform, so IF is $f_{IF} = 455\,\text{kHz}$.
(c) Working backwards from the mixer output, it is straightforward to find the frequencies with the same f_{IF} for the given local oscillator f_{LO}:

$$\text{sum: } 1435\,\text{kHz} + 455\,\text{kHz} = 1890\,\text{kHz},$$

$$\text{difference: } 1435\,\text{kHz} - 455\,\text{kHz} = 980\,\text{kHz}.$$

In other words, a station operating at 1890 kHz will result in the same IF as the wanted station operating at 980 kHz. It would not be possible to separate the two, which means that the other frequency is the image of the wanted frequency.

A practical solution to this problem is to simultaneously tune the tuned RF amplifier at the receiver's input to the same frequency as the wanted frequency.
(d) A graphical representation for the case in which $f_{LO} > f_{RF}$ is shown in Fig. 7.

11.5 For the phase modulator described here, the phase deviation constant is

$$K = -\frac{Q}{(1 + 2V_0)(1 + n)},$$

where $C = nC_{d0}$ and C_{d0} is the varicap diode capacitance for the biasing voltage V_0. A straightforward calculation gives $K = -0.2\,\text{rad/V}$.

11.6

(a) Using Carson's rule, we write:

$$B_{\mathrm{FM}} = 2(m_{\mathrm{f}}+1)f_{\mathrm{b}} = 2(1.5+1)\,10\,\mathrm{kHz} = 50\,\mathrm{kHz}.$$

(b) Using Bessel's function (see Table 11.1), we write:

$$\begin{aligned}\frac{P_{\mathrm{T}}}{P_{\mathrm{c}}} &= J_0^2 + 2\left(J_1^2 + J_2^2 + J_3^2 + J_4^2 + J_5^2\right)\\ &= 0.512^2 + 2\left(0.558^2 + 0.232^2 + 0.061^2 + 0.012^2 + 0.002^2\right)\\ &= 1.000258. \end{aligned} \tag{58}$$

In other words, the total power is constant, just redistributed between the carrier and sidebands.

(c) According to Table 11.1, for $m_{\mathrm{f}} = 1.5$, the first sideband frequency signal has the highest amplitude, $J_1 = 0.558$ relative to the amplitude of the unmodulated signal.

11.7 Using the equation that relates total power to carrier power, it is straightforward to write:

$$\begin{aligned} P_{\mathrm{T}} &= P_{\mathrm{c}}\left(1+\frac{m^2}{2}\right),\\ 1200\,\mathrm{W} &= P_{\mathrm{c}}\left(1+\frac{0.85^2}{2}\right),\\ &\therefore\\ P_{\mathrm{c}} &= 881.5\,\mathrm{W}. \end{aligned}$$

The sum of the carrier power and the power in the two sidebands P_{SB} equals the total power, therefore,

$$P_{\mathrm{SB}} = P_{\mathrm{T}} - P_{\mathrm{c}} = 318.5\,\mathrm{W}.$$

One half of the total sideband power P_{SB} is in the upper sideband (USB) and one half in the lower sideband (LSB), i.e.

$$P_{\mathrm{USB}} = P_{\mathrm{LSB}} = \frac{P_{\mathrm{SB}}}{2} = 159.25\,\mathrm{W}.$$

11.9 Using the equation that relates total power and the carrier power, the expression for power in one sideband is

(a) $m = 0.7$:

$$P_{\mathrm{USB}} = P_{\mathrm{LSB}} = \frac{m^2 P_{\mathrm{c}}}{4} = \frac{0.7^2\,1500\,\mathrm{W}}{4} = 183.75\,\mathrm{W}.$$

(b) $m = 0.5$: The Carrier power is same for all modulation indexes, however the sideband powers are:

$$P_{\mathrm{USB}} = P_{\mathrm{LSB}} = \frac{m^2 P_{\mathrm{c}}}{4} = \frac{0.5^2\,1500\,\mathrm{W}}{4} = 93.75\,\mathrm{W}.$$

11.10 The IF is the difference between the carrier and the local oscillator frequencies, i.e.,

(a) for $f_{LO} > f_c$

$$f_{IF} = f_{LO} - f_c,$$
$$\therefore$$
$$f_{LO} = 995\,\text{kHz}.$$

(b) for $f_c > f_{LO}$

$$f_{IF} = f_c - f_{LO},$$
$$\therefore$$
$$f_{LO} = 85\,\text{kHz}.$$

11.11

(a) The given frequency deviation is Δf =50 kHz, so the carrier swing is 100 kHz, (i.e., "deviating" on both sides of the carrier frequency).
(b) The highest frequency is one deviation above the carrier, i.e., 107.65 MHz, and the lowest frequency is one deviation below the carrier, i.e., 107.66 MHz.
(c) By definition,

$$m_f = \frac{\Delta f}{f_m} = \frac{50\,\text{kHz}}{7\,\text{kHz}} = 7.143.$$

11.12

For an unmodulated FM signal, the total power is equal to the carrier power, $P_T = P_c$, (i.e., for $m = 0$). Also, the total power does not change for various modulation indexes—it is only redistributed. The carrier power (which starts at 100 W) is reduced by the appropriate J_0 coefficient and the rest of the power is "assigned" to the sideband signals.

(a) Using the equation for FM power and Table 11.1 (for $m_f = 2.0$), it follows that:

$$P_T = P_c\left(J_0^2 + 2\left(J_1^2 + J_2^2 + J_3^2 + J_4^2 + J_5^2 + J_6^2\right)\right),$$

and

$$J_0 = 0.224, J_1 = 0.577, J_2 = 0.353,$$
$$J_3 = 0.129, J_4 = 0.034, J_5 = 0.007, J_6 = 0.001.$$

In other words,

$$P_0 = 100\,\text{W} \times 0.224^2 = 5.0176\,\text{W},$$
$$P_1 = 100\,\text{W} \times 2 \times 0.577^2 = 66.5858\,\text{W},$$
$$P_2 = 100\,\text{W} \times 2 \times 0.353^2 = 24.9218\,\text{W},$$
$$P_3 = 100\,\text{W} \times 2 \times 0.129^2 = 3.3282\,\text{W},$$
$$P_4 = 100\,\text{W} \times 2 \times 0.034^2 = 0.2312\,\text{W},$$

$$P_5 = 100\,\text{W} \times 2 \times 0.007^2 = 0.0098\,\text{W},$$

$$P_6 = 100\,\text{W} \times 2 \times 0.001^2 = 0.0002\,\text{W},$$

which gives, again, a total power of 100 W.

(b) Using Carson's rule, the estimated bandwidth (for $m_\text{f} = 2$) is

$$B_\text{FM} = 2(m_\text{f} + 1)\, f_\text{m} = 6 \times 1.0\,\text{kHz} = 6\,\text{kHz}.$$

11.13 The circuit attached to the $L_\text{T}C_\text{T}$ resonator is known as a reactance modulator. For a MOS transistor implementation, its equivalent capacitance at node ① is given as $C_\text{eq} = g_\text{m}RC$.

Therefore, for the given data, the resonant frequency is set by:

$$f_\text{out} = \frac{1}{2\pi\sqrt{L_\text{T}(C_\text{T} + C_\text{eq})}},$$

$$\therefore$$

$$C_\text{T} + C_\text{eq} = 103.4\,\text{nF}, \tag{59}$$

$$\therefore$$

$$C_\text{eq} = 20\,\text{nF}. \tag{60}$$

Solutions to Selected Problems in Chapter 12

12.1

(a) For the envelope detector, $Z_\text{in} = R/2$, therefore $Z_\text{in} = 1\,\text{k}\Omega$.

(b) The amplitude of the unmodulated input signal (i.e., carrier amplitude $v_\text{c}(pk)$) is the same as the average of the AM waveform envelope. From the data, we write $v_\text{i}(avg) = (1.5\,\text{V} + 0.5\,\text{V})/2 = 1.0\,\text{V} = v_\text{c}(pk)$. By definition, the RMS power of the carrier is

$$P_\text{c} = \frac{v_\text{c}^2(pk)}{2Z_\text{in}} = 0.5\,\text{mW}.$$

From the same data, we can find the modulation index as

$$m = \frac{1.5 - 0.5}{1.5 + 0.5} = 0.5.$$

Therefore,

$$P_t = [1 + (0.5)^2/2](0.5\,\text{mW}) = 562.5\,\mu\text{W}.$$

(c) Output of the envelope detector should be the same as the envelope of the input waveform, except for the diode voltage drop. From the data (Fig. 12.18 (right)) we see that the envelope is shifted by 0.2 V from the diode voltage drop.

Therefore, the maximum value of the envelope is $v_0(max) = 1.3\,\text{V}$, the minimum value of the envelope is $v_0(min) = 0.3\,\text{V}$, so the average (DC) output is $v_0(DC) = (1.3\,\text{V} + 0.3\,\text{V})/2 = 0.8\,\text{V}$.

(d) Knowing the average (DC) current and output resistance (R), the output current must be $I_0(DC) = 0.8\,\text{V}/2\,\text{k}\Omega = 400\,\mu\text{A}$.

(e) For the capacitor, we write

$$C = \frac{\sqrt{(1/m_\text{a})^2 - 1}}{2\pi R f_\text{m}(max)} = 7.7\,\text{nF}.$$

12.2

(a) The output signal spectrum, at R_L, is due to the nonlinear characteristics of the diode. Both the carrier and the signal are being processed by the diode and, therefore, producing side tones.

With the stated assumption of the problem, the incoming IF signal has a carrier at 665 kHz which was modulated by a 5 kHz signal. As a result, the incoming IF signal contains 665 kHz, as well as the sidetones 660 kHz and 670 kHz, note the relative amplitudes. These three tones are processed, i.e., multiplied again, by the diode nonlinear characteristics producing the following tones:

Sum frequencies:
$$660 + 665 = 1325\,\text{kHz},$$
$$660 + 670 = 1330\,\text{kHz},$$
$$665 + 670 = 1335\,\text{kHz}.$$

Difference frequencies:
$$670 - 665 = 5\,\text{kHz},$$
$$665 - 660 = 5\,\text{kHz},$$
$$670 - 660 = 10\,\text{kHz}.$$

Note that the 10 kHz tone can be easily filtered out.

(b) The fast-changing signal is the carrier; the slow-changing signal is the signal envelope. Note that, because of the diode orientation, the negative amplitude signal envelope is recovered. The last stage serves the purpose of removing the DC offset in the signal envelope by using the blocking capacitor C.

(c) The capacitors have resistors with the following values: $Z_{C1} = 1/(2\pi * 5\,\text{kHz} * 220\,\text{pF}) = 144.68\,\text{k}\Omega \approx 145\,\text{k}\Omega$, $Z_{C2} = 1/(2\pi * 5\,\text{kHz} * 22\,\text{pF}) = 1.4468\,\text{M}\Omega \approx 1.45\,\text{M}\Omega$, and diode resistance is $R_\text{D} = \Delta V/\Delta I = 0.7\,\text{V}/7\,\text{mA} = 100\,\Omega$ (from the diode transfer characteristics graph). The transformer is just an ideal voltage element. Hence, the equivalent voltage divider consists of $R_\text{D} = 100\,\Omega$ and $R = Z_{C1}||(R_1 + (Z_{C2}||R_2||R_\text{L})) = 4.6\,\text{k}\Omega$. That means that the voltage amplitude at node ③, relative to the voltage amplitude of the input signal, is the same ratio as $A = V(3)/V_\text{in} = R/(R + R_\text{D}) = 0.978$, i.e., the 5 kHz signal is almost not attenuated at all.

(d) The capacitors have resistors with the following values: $Z_{C1} = 1/(2\pi * 665\,\text{kHz} * 220\,\text{pF}) = 1087.86\,\Omega \approx 1.1\,\text{k}\Omega$, $Z_{C2} = 1/(2\pi * 665\,\text{kHz} * 22\,\text{pF}) = 10.878\,\text{k}\Omega \approx 11\,\text{k}\Omega$, and diode resistance is again $R_\text{D} = \Delta V/\Delta I = 0.7\,\text{V}/7\,\text{mA} = 100\,\Omega$ (from the diode transfer characteristics graph). The transformer is an ideal voltage element. The equivalent circuit is the same as the one in part (c), however this time $R = 840\,\Omega$ and $A = 0.894$, i.e., the 665 kHz carrier is attenuated a bit more than the 5 kHz signal.

By choosing appropriate component sizes, a designer has control over how much the carrier tone is attenuated relative to the envelope signal.

Note: Try repeating the exercise using the same components, except that C_2 =22 nF.

12.3

(a) Input power is calculated by definition as

$$P_\text{in} = \frac{V^2}{R} = \frac{(8\,\mu\text{V})^2}{50\,\Omega} = 1.28\,\text{pW},$$

$$\therefore$$

$$P_{\text{in}} \equiv 10\log\frac{1.28\,\text{pW}}{1\,\text{mW}} = -88.9\,\text{dBm} = -118.9\,\text{dBW}.$$

(b) Simple addition in dB along the system chain gives (note that we can add dB and dBm because of the definition of the units):

$$\begin{aligned} P_{\text{out}} &= -88.9\,\text{dBm} + 8\,\text{dB} + 3\,\text{dB} + 24\,\text{dB} + 26\,\text{dB} \\ &\quad + 26\,\text{dB} - 2\,\text{dB} + 34\,\text{dB} = 30.1\,\text{dBm}, \\ &\therefore \\ P_{\text{out}} &\equiv 1\,\text{W}. \end{aligned}$$

12.4

$$V_{\max} = V_{\text{C}} + \frac{V_{\text{b}}}{2} = 3\,\text{V}; \qquad V_{\min} = V_{\text{C}} - \frac{V_{\text{b}}}{2} = 1\,\text{V};$$
$$m = 0.5.$$

We keep in mind that the longer side b of the trapezoidal pattern is proportional to $V_{\max}$, while the shorter side a is proportional to $V_{\min}$, hence $b/a = 3/1 = 3$.

12.5

$$m = \sqrt{2\left[\left(\frac{1.1A}{1A}\right)^2 - 1\right]} = 0.648.$$

Solutions to Selected Problems in Chapter 13

13.4 By inspection of the graph, we conclude that in the linear part of the transfer characteristic for an input of -50 dBm, the output power is -30 dBm, hence the gain is 20 dB. The linear part of the characteristics extends to approximately -20 dBm of the input power, when the output power becomes -1 dBm instead of the expected 0 dBm. Therefore the 1 dB compression point is at -20 dBm of the input power. The-third order harmonics power is extrapolated until it intersects with the extrapolated linear part of the characteristics, and the crossing point is found at the output power of approximately $+9.6$ dBm, which is only the extrapolated point, not the real measurement point. Keep in mind that the amplifier output never reaches that level of output power, it has already saturated close to the 1 dB compression point level.

Index

R. Sobot, *Wireless Communication Electronics: Introduction to RF Circuits and Design Techniques*, DOI 10.1007/978-1-4614-1117-8,