Radio-Frequency Integrated Circuits

Harald Pretl
Johannes Kepler University
harald.pretl@jku.at

2025-10-13

Table of contents

1 Introduction 2
1.1 Wireless Transmission 2
1.2 Wireless Standards 6
2 Fundamentals 9
2.1 Channel Capacity 9
2.2 Linearity 10
2.2.1 Single-Tone Linearity 10
2.2.2 Multi-Tone Linearity 12
2.3 Noise 16
2.3.1 Types of Noise Generation 17
2.3.2 Noise in Impedance-Matched Systems 18
2.3.3 Noise Figure 19
2.3.4 Sensitivity 21
2.4 Modulation 22
2.5 Pulse Shaping and Spectral Efficiency 24
2.6 Orthogonal Frequency-Division Multiplexing (OFDM) 26
2.7 Multiple Access Techniques 28
3 Transceivers 29
3.1 Direct-Conversion Transceiver 29
3.2 Modulation and Demodulation 31
3.3 Filtering 32
3.4 Direct-Conversion Architecture 35
3.5 Duplexing 35
3.5.1 Frequency-Division Duplex (FDD) 35
3.5.2 Time-Division Duplex (TDD) 36
3.5.3 Comparison of FDD and TDD 37
3.6 Specialty Architectures 38
3.6.1 Super-Heterodyne Architecture 38
3.6.2 Low-IF Architecture 40
3.6.3 Super Simple Architecture 41
3.7 I/Q Imbalance 41
4 Low Noise Amplifiers 42
4.1 Resistively Matched Common-Source LNA 44
4.2 Common-Gate LNA 47
4.3 Inductively-Degenerated Common-Source LNA 48

4.4 Feedback LNA 50
5 Mixers 52
6 Oscillators 52
7 Phase-Locked Loops 52
8 Power Amplifiers 52
Bibliography 52

1 Introduction

This is the material for an introductory radio-frequency integrated circuits course. The contents are largely based on [1] and [2]; these two books are an excellent introduction into this topic and are highly recommended! For a general introduction into RF and microwave [3] is a great read!

It is assumed that readers are familiar with the contents of this Analog Circuit Design course.

> ! Important
>
> All course material (source code of this document, Jupyter notebooks for calculations, Xschem circuits, etc.) is made publicly available on GitHub (follow this link) and shared under the Apache-2.0 license.
>
> Please feel free to submit pull requests to fix typos or add content! If you want to discuss something that is not clear, please open an issue.
>
> The production of this document would be impossible without these (and many more) great open-source software products: `VS Code`, `Quarto`, `Pandoc`, `LaTeX`, `Typst`, `Jupyter Notebook`, `Python`, `Xschem`, `ngspice`, `CACE`, `pygmid`, `schemdraw`, `Numpy`, `Scipy`, `Matplotlib`, `Pandas`, `Git`, `Docker`, `Ubuntu`, `Linux`, ...

1.1 Wireless Transmission

In wireless transmission, we usually want to transmit data via a transmitter (TX) and a connected antenna to a receiver (RX) using an electromagnetic (EM) wave. This arrangement is shown in Figure 1.

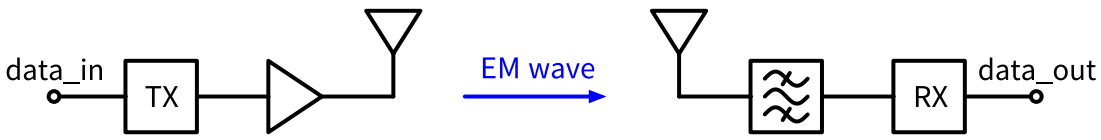

Figure 1: The block diagram of a simple wireless system.

Unfortunately, wireless transmission is hard. The wireless channel, i.e., the usage of electromagnetic waves to transmit information from a transmitter to a receiver, while tremendously useful, unfortunately has quite a few undesired features:

- The wireless channel is shared between all users.
- As a consequence, the available bandwidth is shared; this means that bandwidth is a scarce resource.

- The wireless channel has significant losses.
- The channel is time variant, as usually the transmitter and/or the receiver move, and/or the environment changes.

In order to estimate the power P_{R} of the wireless transmission at the receiver we can use Friis' transmission formula [3]:

$$P_{\mathrm{R}} = \frac{P_{\mathrm{T}}}{4\pi d^2} \cdot A_{\mathrm{R}} = P_{\mathrm{T}} \cdot \frac{A_{\mathrm{R}} \cdot A_{\mathrm{T}}}{d^2 \lambda^2} \tag{1}$$

Here, A_{R} (and A_{T}) is the effective area of the receive/transmit antenna, while d is the distance (line of sight) between the two antennas. The effective area of an antenna depends on the type and construction, but generally we can say that

$$A \propto \lambda^2$$

For an isotropic antenna (a theoretical construct where the radiation is equal in all directions) $A = \lambda^2/(4\pi)$, while for a $\lambda/2$-dipole $A = 0.13\lambda^2$. Of course, the speed of light c relates frequency f and wavelength λ of an electromagnetic wave by

$$c = \lambda f.$$

Generally speaking, the size of an electromagnetic antenna is proportional to the wavelength of the EM wave use for transmission. For man devices, we seek antennas on the order of a few centimeters, and this is why frequencies in the hundreds of MHz to GHz are so popular. Table 1 lists a few typical applications and their frequency and wavelength.

Table 1: Typical RF applications with their operating frequencies and corresponding wavelengths

Application	Frequency	Wavelength
FM Radio	88–108 MHz	2.8-3.4 m
WiFi (lowband)	2.4 GHz	12.5 cm
WiFi (highband)	5 GHz	6 cm
Bluetooth	2.4 GHz	12.5 cm
Cellular	0.6–5 GHz	6-50 cm
GNSS	1.575 GHz	19 cm

As you can see in Table 1 many of these antennas would not fit into the used device form factors, i.e., often we have to use electrically small antennas.

Note 1: Wavelength Calculation

Let's calculate the wavelength for a Bluetooth signal at 2.4 GHz. Given:

- Frequency $f = 2.4\ \text{GHz} = 2.4 \times 10^9\ \text{Hz}$
- Speed of light $c = 3 \times 10^8\ \text{m/s}$

Using the relationship $c = \lambda f$, we can solve for wavelength:

$$\lambda = \frac{c}{f} = \frac{3 \times 10^8\ \ \text{m/s}}{2.4 \times 10^9\ \ \text{Hz}} = 0.125\ \ \text{m} = 12.5\ \ \text{cm}$$

This means that a quarter-wavelength monopole antenna for 2.4 GHz Bluetooth would be approximately 3.1 cm long, which easily fits into most mobile devices.

In order to get a feeling for the attenuation experienced in wireless communication, we now calculate the following exemplary transmission. We will use the unit of dBm with is often used in RF design and is defined as

$$P\ |_{\text{dBm}} = 10 \cdot \log_{10}\left(\frac{P\ |_{\text{W}}}{1\,\text{mW}}\right) \tag{2}$$

Note 2: Wireless Transmission

We use the following parameters:

- Transmit power $P_{\text{T}} = 1\ \text{W}$
- Frequency $f = 2.4\ \text{GHz}$
- Communication distance $d = 10\ \text{km}$
- Using $\lambda/2$ dipoles on both ends

Using Equation 1 we calculate

$$P_{\text{R}} = P_{\text{T}} \cdot \frac{0.13\lambda^2 \cdot 0.13\lambda^2}{d^2\lambda^2} = P_{\text{T}} \cdot 0.13^2 \left(\frac{\lambda}{d}\right)^2 = 2.64\,\text{pW} = -85.8\,\text{dBm}$$

With the transmit power of 1 W = 30 dBm we have an attenuation of 116 dB! This is a very large number!

The loss we calculated in Note 2 is called the free-space path loss (FSPL). It is the minimum loss we can expect in a wireless communication system. In reality, the situation is often even worse. The free-space path loss FSPL (in dB) can be calculated as

$$\text{FSPL} = 20 \cdot \log_{10}(d/\text{m}) + 20 \cdot \log_{10}(f/\ \text{Hz}) + 20 \cdot \log_{10}\left(\frac{4\pi}{c}\ \text{m/s}\right). \tag{3}$$

Equation 3 can be readily derived from Equation 1 and Equation 2 assuming isotropic antennas at transmitter and receiver. Using Equation 3 we can easily calculate the FSPL for different distances and frequencies. The results are shown in Figure 2. It should be noted that the FSPL

increases by 20 dB per decade of distance and 20 dB per decade of frequency, making higher frequencies and longer distances very challenging.

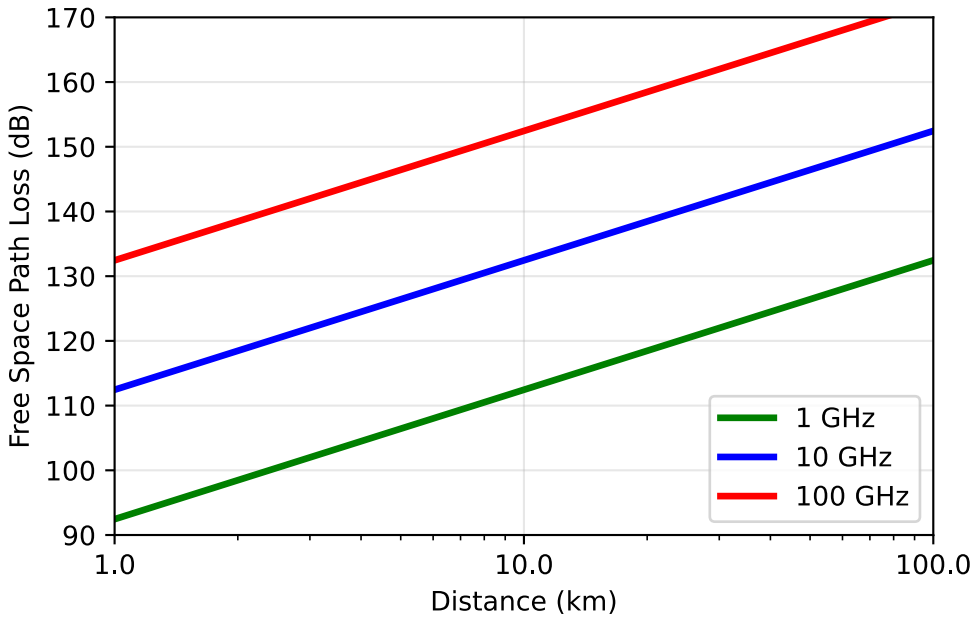

Figure 2: Free space path loss vs. distance for different frequencies (1 GHz, 10 GHz, and 100 GHz).

As dire as the situation of Figure 2 already looks, this is not even all factors considered:

- The given attenuation is for line-of-sight paths; often, the attenuation is significantly higher than this due to blockage by buildings, mountains, rain, or foliage.
- In lack of a direct line-of-sight path, the EM wave is redirected by reflections, causing additional attenuation, and the potential destructive interference by multi-path reception.

The consequences of this are (among others):

- The transmitter needs to generate enough **transmit power** to overcome the transmission loss; this has to be done often with high **efficiency**, as the transmit device is battery operated or limited by cooling.
- The receiver has to be able to process **weak signals**, i.e., the **noise** level of the signal processing has to be very low.
- Often, the receive signal is very weak, while there are strong signals at other frequencies (i.e., other wireless transmitters are located close to the receiver). This means the receiver has to be able to process a weak signal while simultaneously tolerate **large interfering signals** (called blockers).
- Since the frequency spectrum is shared among many users and wireless applications, the transmit information has to be packed efficiently into a **small bandwidth**.
- Very often, wireless devices are battery-operated. This means transmit and receive functions have to be implemented using **minimum power consumption**.

As stated in the beginning, designing wireless systems is hard.

1.2 Wireless Standards

In order to allow communication between different devices, different operators, and different manufacturers, wireless communication is standardized. There are many different standards, each with its own characteristics. Wireless standards define every aspect of the wireless communication, and be documents with hundreds or thousands of pages. Here, we mainly focus on the radio-frequency and analog aspects of wireless standards. Summarized in Table 2 are a few popular wireless standards with their main characteristics.

Table 2: Comparison of wireless communication standards

Standard	GSM (2G)	WCDMA (3G)	LTE (4G)**5G NR**	WiFi	Bluetooth	GNSS
Frequency range (MHz)	850, 900, 1800, 1900	850, 900, 1700, 1900, 2100	Multiple bands 450…7100 **(FR1)**, **24000… 48000 (FR2)**	2400, 5000, 6000	2400	1500, 1200
Modulation	GMSK, 8PSK (EDGE)	QPSK (DL), BPSK (UL), 16QAM (HSPA), 64QAM (HSPA)	QPSK, 16QAM, 64QAM (DL+UL) 256QAM (DL**+UL**)	BPSK, QPSK, 16QAM, 64QAM, 256QAM, 1024QAM, 4096QAM	GFSK (m=0.28…0.35), π/4-DQPSK, 8DPSK	BPSK, QPSK
Transmission/ multiple access	TDMA, FDMA	DS-CDMA	OFDMA (DL**+UL**), SC-FDMA/**DFT-s-OFDM (UL)**	OFDM, CSMA/CA	FHSS	CDMA
Duplex	FDD	FDD	FDD, TDD	TDD	TDD	n/a
Channel bandwidth	200 kHz	5 MHz	1.4, 3, 5, 10, 15, 20, …, **100 MHz (FR1), 400 MHz (FR2)**	10, 20, 40, 80, 160, 320 MHz	1 MHz	16…24 MHz
Symbol rate	270.833 ksym/s	3.84 Msym/s	15/**30/60** ksym/s	312.5 ksym/s	1 Msym/s	50 sym/s
Pulse shaping	Gaussian (BT=0.3)	Root Raised Cosine (α =0.22)	Rectangular	Rectangular	Gaussian (BT=0.5)	Rectangular
Transmit power	1…2 W	250 mW	200 mW (FDD), 400 mW (TDD)	100 mW	1…100 mW	n/a
PAR (UL)	0 dB (GMSK), 3 dB (8PSK)	3…8 dB	6…8 dB	Up to 12 dB	0 dB (GFSK), 3 dB (8DPSK)	n/a
MIMO	no	Not realized (DL 2x2)	DL 4x4 (up to 8x8)	2x2 (up to 8x8)	no	no
Channel bond	no	Up to 4x5 MHz	Up to 7x20/4x100 MHz	Up to 80+80+80+80 MHz	no	no

During this course, we will learn what these terms mean and how they impact the design of RF integrated circuits.

As you can see in Table 2, since LTE (4G) and 5G NR, many additional bands have been defined in the sub-6 GHz range (FR1) and also in the mm-wave range (FR2, 24.25 to 52.6 GHz). This means that modern wireless devices have to support many different frequency bands, which makes the design of RF frontends even more challenging. A good overview of the different frequency bands is given here for LTE and 5G NR.

RFIC design is a multidisciplinary field, requiring knowledge from various engineering domains, as shown in Figure 3. This makes RFIC design challenging, but also very interesting!

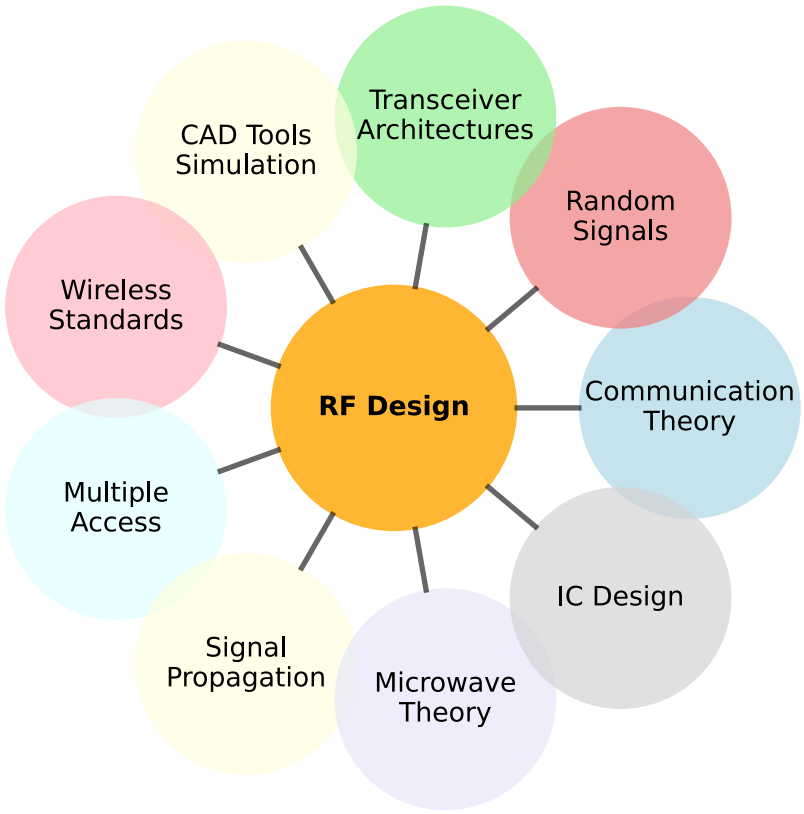

Figure 3: RF design as a multidisciplinary field requiring knowledge from various engineering domains (adapted from [1]).

Further, RFIC design requires careful consideration of many different aspects, as shown in Figure 4. Many parameters are often tightly coupled, requriring careful trade-offs during the design process.

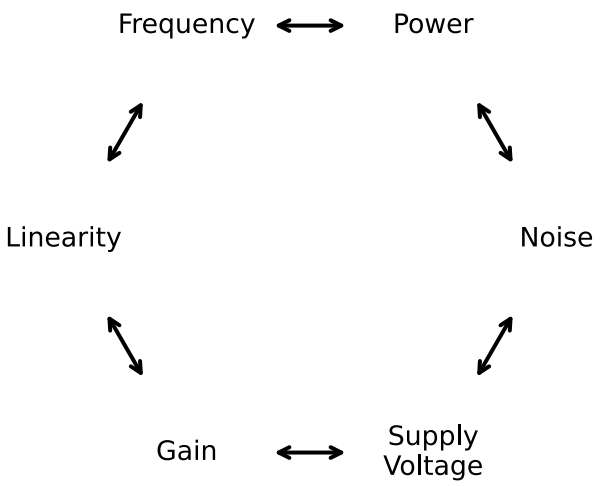

Figure 4: RFIC require careful design considerations and trade-offs (adapted from [1]).

2 Fundamentals

In this section, we will discuss a few important concepts which will be instrumental in the further study of RF circuits and systems. As signals in RF circuits and systems are often limited on the top end by linearity, and on the bottom end by noise, we will discuss these two topics in some detail.

2.1 Channel Capacity

In Section 1 we have already discussed the fact that we need to pack information into a minimum bandwidth, as the available spectrum is limited. To appreciate the limits of information transfer, we need to understand how much information can be transmitted over a given bandwidth. This limit is given by the *Shannon-Hartley theorem*, which gives the maximum data rate C (in bit/s) that can be transmitted over a communication channel with bandwidth B (in Hz) and signal-to-noise ratio SNR (with linear units):

$$C = B \cdot \log_2(1 + \mathrm{SNR}) \tag{4}$$

This formula gives us a theoretical upper limit on the data rate that can be achieved with a given bandwidth and SNR *under optimal conditions*. It is important to note that this limit is only achievable with ideal coding and modulation schemes, which are not practical in real-world systems. However, it provides a useful benchmark for evaluating the performance of communication systems.

i Note 3: Channel Capacity Example

Let us calculate the channel capacity for a system with a bandwidth of 2 MHz and an SNR of 7 dB (the BW and minimum SNR of Bluetooth LE for 1 Mbps). First, we need to convert the SNR from dB to linear units:

$$\text{SNR} = 10^{7/10} \approx 5$$

Now we can use Equation 4 to calculate the channel capacity:

$$C = B \cdot \log_2(1 + \text{SNR}) = 2\,\text{MHz} \cdot \log_2(1 + 5) = 2\,\text{MHz} \cdot 2.585 \approx 5.2\,\text{Mbps}$$

This sounds reasonable, as the user data rate for Bluetooth LE is 1 Mbps for the given SNR, which allows for quite some overhead for coding and protocols.

2.2 Linearity

As we have already seen in Section 1.1 the transmitter has to process large signals without distorting them, while the receiver has to process small signals in the presence of large signals. Both situations mean we need metrics and models to quantify and discuss linearity properties.

We are going to use a very simple, time-invariant model to study linearity, based on a Taylor polynomial.

! Linearity and Time Invariance in RF Systems

We use time invariance to simplify the mathematics. In practice, many circuits and systems will show time variant behavior which leads to quite a few very interesting and important phenomena! A time-invariant nonlinear system is also called a "memoryless" system, as the output at time t only depends on the input at time t.

In contrast, a system with memory (i.e., time-variant) will have an output at time t which depends on the input at time t and also on past inputs (e.g., at times $t - \Delta T$, $t - 2\Delta T$, etc.). Examples of systems with memory are filters, which have a frequency-dependent response, or power amplifiers with thermal memory effects.

We model a nonlinear circuit block with the following Taylor polynomial:

$$y(t) = \alpha_0 + \alpha_1 x(t) + \alpha_2 x(t)^2 + \alpha_3 x(t)^3 + \ldots \tag{5}$$

Usually, the blocks under study will have higher order nonlinear terms but we often stop at 3rd order to keep things simple. For practical work, higher order terms should be included if necessary.

Which $x(t)$ should we use to study wireless systems? Often, the bandwidth f_{BW} of a transmit signal is much smaller than the center frequency f_0, i.e., $f_{\text{BW}} \ll f_0$. In this case using a sinusoidal signal as a model is both simple to handle and approximately correct.

2.2.1 Single-Tone Linearity

We thus set (with A being the amplitude of the input signal and $\omega = 2\pi f$ the angular frequency)

$$x(t) = A\cos(\omega t)$$

and insert it into Equation 5. After some simple trigonometric manipulations we are at

$$\begin{aligned} y(t) = & \underbrace{\frac{1}{2}\alpha_2 A^2}_{\text{dc component}} + \underbrace{\left(\alpha_1 A + \frac{3}{4}\alpha_3 A^3\right)\cos(\omega t)}_{\text{fundamental}} \\ & + \underbrace{\frac{1}{2}\alpha_2 A^2 \cos(2\omega t)}_{\text{2nd harmonic}} + \underbrace{\frac{1}{4}\alpha_3 A^3 \cos(3\omega t)}_{\text{3rd harmonic}} \end{aligned} \tag{6}$$

Looking at Equation 6 we can make a few interesting observations:

- Even-order nonlinearity (α_2) creates low-frequency components; it effectively adds frequency components related to the envelope A. If A is a constant then this results in a dc term; if $A(t)$ is time variant it will created a squared version of it at low frequencies.
- The α_1 term is the gain of the circuit block.
- Odd-order nonlinearity (α_3) can impact the gain of the fundamental term passing through the block. Depending on the sign of α_3 this can lead to gain contraction or expansion.
- Even- and odd-order nonlinearities creates additional frequency components, so-called harmonics of the fundamental frequency. These harmonics are often unwanted, as they are far outside the wanted transmission frequency range, and need to be minimized, by either
  1. use a lowpass filter to filter these harmonics, or
  2. increase the linearity, i.e., make the α_2, α_3, etc., small enough.

The created harmonics are illustrated in Figure 5. Note that measuring harmonics to quantify the nonlinearity metrics like α_2 and α_3 is often not very accurate, as these harmonics are often filtered in bandwidth-limited systems.

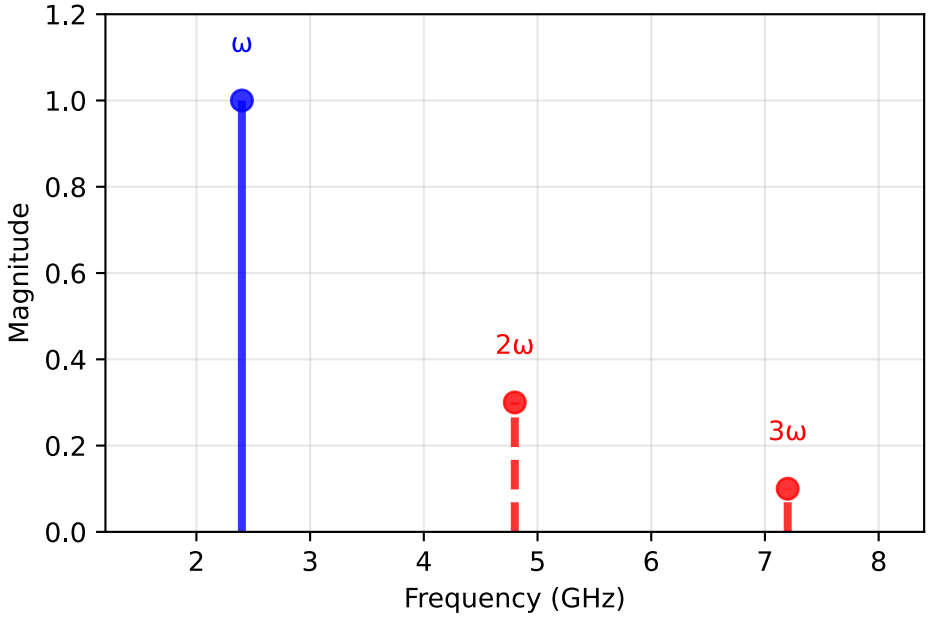

Figure 5: Single-tone test showing created harmonics at 2ω and 3ω.

How can we quantify the nonlinearity with a one-tone test? We can sweep the input signal $x(t)$ in amplitude, and observe the output $y(t)$. If the observed gain drops by 1 dB from the small-signal value we note the input power, and call this point the **1dB compression point** ($P_{1\text{dB}}$). We should always add whether this 1dB compression point is input- or output-referred to avoid ambiguity. The diagram in Figure 6 shows this test ($\alpha_1 = 100$, $\alpha_3 = -0.2$).

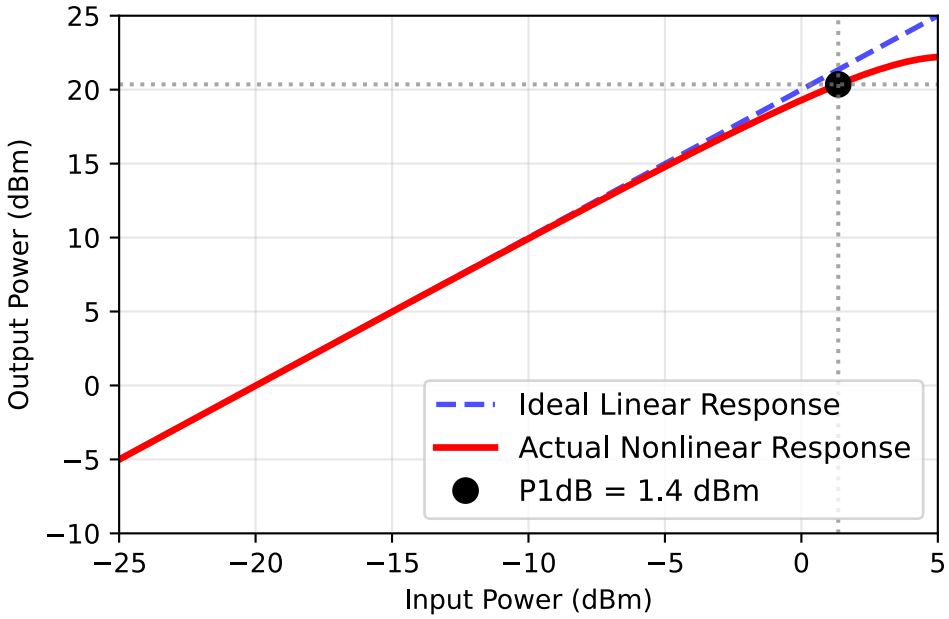

Figure 6: 1dB compression point test showing input vs output power relationship and the definition of P1dB.

! Compressive vs. Expansive Behavior

Note that for compressive behaviour, α_3 and α_1 have different signs, while for expansive behaviour, they have the same sign.

At some point, every circuit block will show compressive behavior, as the maximum signal amplitude will be limited by power supply voltages, device breakdown voltages, etc.

2.2.2 Multi-Tone Linearity

We now elevate our investigations and apply two sinusoids with different frequencies and different amplitudes and see which signals we get at the output of the nonlinear block. The two-tone test and resulting third-order intermodulation products (IM3) are illustrated in Figure 7.

$$x(t) = A_1 \cos(\omega_1 t) + A_2 \cos(\omega_2 t)$$

We apply the above stimulus to our nonlinear model described by Equation 5 and again, after some trigonometric manipulations, arrive at:

$$y(t) = y'(t) + y''(t) + y'''(t) \tag{7}$$

As many different frequency components are created by this simple two-tone test (and nonlinearity only up to 3rd order) we split the result into different equations and look at the result separately.

First, we start with the fundamental tones:

$$\begin{aligned} y'(t) = &\left(\underbrace{\alpha_1 A_1 + \frac{3}{4}\alpha_3 A_1^3}_{\text{compression/expansion}} + \underbrace{\frac{3}{2}\alpha_3 A_1 A_2^2}_{\text{cross-modulation/desens}} \right) \cos(\omega_1 t) \\ &+ \left(\underbrace{\alpha_1 A_2 + \frac{3}{4}\alpha_3 A_2^3}_{\text{compression/expansion}} + \underbrace{\frac{3}{2}\alpha_3 A_2 A_1^2}_{\text{cross-modulation/desens}} \right) \cos(\omega_2 t) \end{aligned} \tag{8}$$

As shown in Equation 8, interesting things happen:

- We (again) have the gain compression/expansion effect as already discussed in Section 2.2.1.
- In addition, we have **cross-modulation**, i.e., the envelope of one tone (e.g., $A_2(t)$ of the tone at ω_2) impacts the envelope of the other tone at ω_1. This can lead to unwanted signal distortation, even if there is a large frequency separation between ω_1 and ω_2!
- Further, since the sign of α_3 is usually opposite to α_1, this can also lead to **desensitization** ("desens"). If, for example, $A_2 \gg A_1$, then there would be no compression due to the tone ω_1 itself, however, the large tone at ω_2 will lead to gain compression of the tone at ω_1; this effect is called desense.

We now look at the next class of generated tones:

$$\begin{aligned} y''(t) = &\frac{1}{2}\alpha_2 A_1^2 + \frac{1}{2}\alpha_2 A_2^2 \\ &+\alpha_2 A_1 A_2 \cos[(\omega_1 - \omega_2)t] \\ &+\alpha_2 A_1 A_2 \cos[(\omega_1 + \omega_2)t] \end{aligned} \tag{9}$$

As we can see in Equation 9 new tones are created (besides the low frequency components we already know from the single-tone test) at the sum and difference of ω_1 and ω_2. These new frequency components are called "**intermodulation products of second order**" (IM2). These tones are created by the even-order nonlinearity (α_2). These IM2 products are far away from the wanted tones, so are often not very problematic in amplifiers (but there can be exceptions!). However, they can be very problematic in frequency conversion blocks like mixers. We will come back to this point when discussing zero-IF receivers.

We now investigate the next couple of tones:

$$
\begin{aligned}
y'''(t) = &\frac{3}{4}\alpha_3 A_1^2 A_2 \cos[(2\omega_1 + \omega_2)t] \\
&+\frac{3}{4}\alpha_3 A_1^2 A_2 \cos[(2\omega_1 - \omega_2)t] \\
&+\frac{3}{4}\alpha_3 A_1 A_2^2 \cos[(2\omega_2 + \omega_1)t] \\
&+\frac{3}{4}\alpha_3 A_1 A_2^2 \cos[(2\omega_2 - \omega_1)t]
\end{aligned} \tag{10}
$$

The tones shown in Equation 10 are called "i**ntermodulation products of third order**" (IM3), and are caused by the odd nonlinearities (like α_3). While the IM3 tones located at $2\omega_1 + \omega_2$ and $\omega_1 + 2\omega_2$ are similar to the sum IM2 tone and far away from ω_1 and ω_2, the other two tones are concerning.

Expressing $\Delta\omega = \omega_2 - \omega_1$ (and assuming $\omega_1 < \omega_2$), the building law of $2\omega_1 - \omega_2 = \omega_1 - \Delta\omega$ and $2\omega_2 - \omega_1 = \omega_2 + \Delta\omega$ results in new tones right besides ω_1 and ω_2, with a frequency separation only defined by $\Delta\omega$. This situation is illustrated in Figure 7.

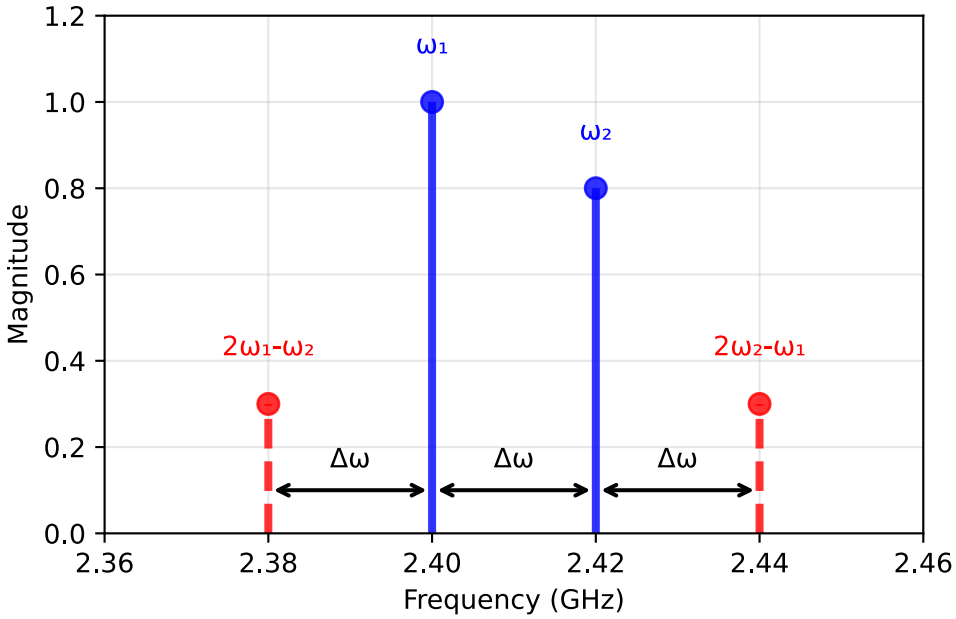

Figure 7: Two-tone test showing fundamental frequencies ω_1, ω_2 and third-order intermodulation products (IM3) at $2\omega_1$-ω_2 and $2\omega_2$-ω_1.

This close localization of the IM3 tones can also be utilized to characterize nonlinear performance. Using gain compression or harmonic generation (H3) it can be very difficult to extract nonlinearity of third order (α_3). However, using a two-tone test, the IM3 tones can be readily measured, even if the measured signal path shows a **bandpass characteristic**! As RF systems frequently employ bandpass filters to suppress out-of-band signals, this is a very important property of the two-tone test.

The resulting test is called a two-tone test yielding the third-order intercept point (IP3). This test is widely used in RF design to characterize the linearity of amplifiers, mixers, and complete

transceiver systems. The power relationship between fundamental tones and IM3 products as a function of input power is shown in Figure 8.

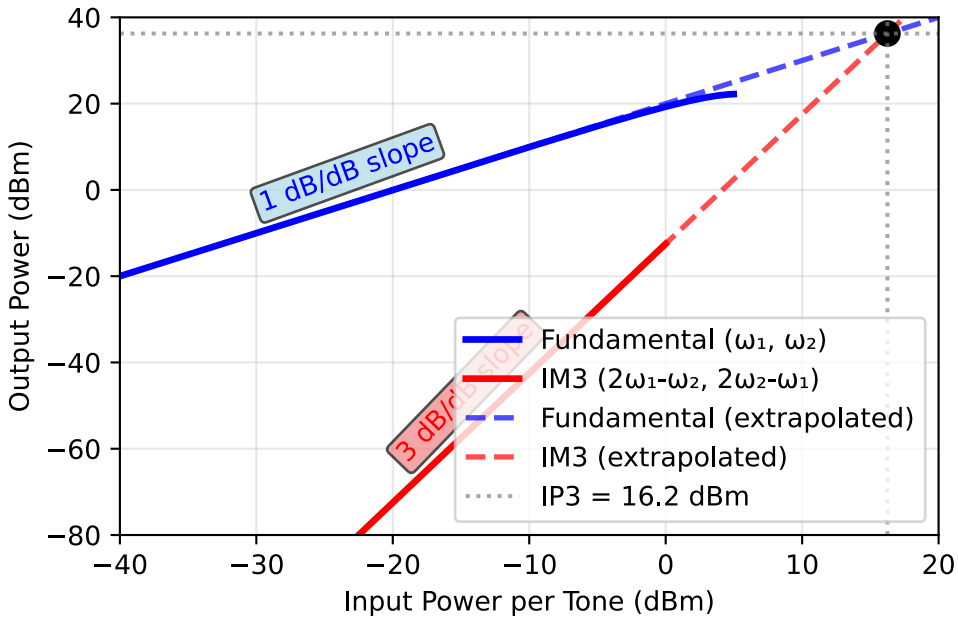

Figure 8: Two-tone IM3 test showing fundamental and IM3 product power vs. input power, with IP3 intercept point definition. Equal input power per tone is assumed.

Note that, as shown in Figure 8, the IM3 products rise with a slope of 3 dB/dB, i.e., if the input power is increased by 1 dB, the IM3 products increase by 3 dB. The fundamental tones rise with a slope of 1 dB/dB (as long as we are in the linear region). The IP3 point is defined as the intersection of the **extrapolated** linear lines of fundamental and IM3 products. As both lines have different slopes, this intersection point is usually far outside the actual operating range of the circuit block under test!

When calculating the IIP3 (input-referred IP3) we can use the following formula, assuming equal input power per tone. It is important to always check the slope of the IM3 products to ensure that we are indeed in the third-order region! If the input power per tone is P_{in} (in dBm) and the input-referred power of one IM3 tone is P_{IM3} (in dBm), then the input-referred IP3 is given by

$$\text{IIP3} = P_{\text{in}} + \frac{P_{\text{in}} - P_{\text{IM3}}}{2} \tag{11}$$

Further, for midly nonlinear systems (i.e., α_3 is dominating), the IIP3 can be approximated from the 1dB compression point as

$$\text{IIP3}\,|_{\text{dBm}} \approx P_{\text{1dB}}|_{\text{dBm}} + 9.6\,\text{dB}\ . \tag{12}$$

If we have two blocks which are cascaded, and we know the gain and IIP3 of both blocks, we can calculate the overall IIP3 of the cascade with the following approximation. An exact calculation is very involved, as the nonlinearities of the first block (and the resulting tones) will be processed by the second block, creating even more tones; this process escalates very quickly. However, for practical purposes, the following approximation is often sufficient:

$$\frac{1}{\mathrm{IIP3}_{\mathrm{total}}} \approx \frac{1}{\mathrm{IIP3}_1} + \frac{G_1}{\mathrm{IIP3}_2} + \frac{G_1 G_2}{\mathrm{IIP3}_3} \tag{13}$$

Here G_1 is the linear gain of the first block, and $\mathrm{IIP3}_1$, $\mathrm{IIP3}_2$ are the input-referred IP3 of the first and second block, respectively. Note that all powers have to be in linear units (i.e., Watts) when using Equation 13. An even more simplified version of Equation 13 can be used with all quantities given in dBm and dB, respectively:

$$\mathrm{IIP3}_{\mathrm{total}} \approx \min\{\mathrm{IIP3}_1, \mathrm{IIP3}_2 - G_1, \mathrm{IIP3}_3 - G_1 - G_2\} \tag{14}$$

A typical RF system cascade with multiple blocks and their individual IIP3 contributions is shown in Figure 9.

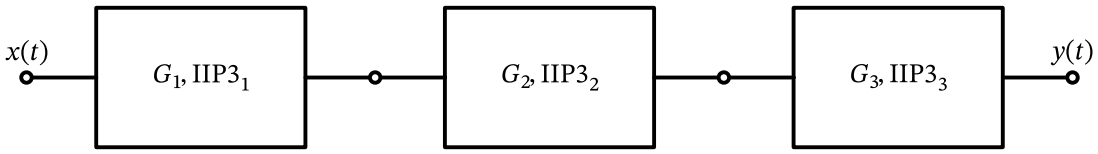

Figure 9: Block cascade for IIP3 calculation showing multiple stages with gains and individual IIP3 values.

i Note 4: Simple IIP3 Cascade Calculation

Let's calculate the overall IIP3 of two cascaded blocks. The first block is a low-noise amplifier with an IIP3 of −10 dBm and a gain of 20 dB. The second block is a mixer that has a gain of 10 dB and an IIP3 of 5 dBm. What is the overall IIP3?

Using Equation 14 we can quickly estimate:

$$\mathrm{IIP3}_{\mathrm{total}} \approx \min\{-10\,\mathrm{dBm}, 5\,\mathrm{dBm} - 20\,\mathrm{dB} = -15\,\mathrm{dBm}\} = -15\,\mathrm{dBm}$$

We see that the overall IIP3 is limited by the linearity of the second block, as the first block amplifies all signals (including blockers) by 20 dB before they reach the second block.

2.3 Noise

Just as nonlinearity is a limiting factor for large signals, noise is the limiting factor for small signals. Noise is present in all electronic circuits and systems, and it is impossible to avoid it. However, we can try to minimize its impact on the system performance.

Noise is usually characterized by its power spectral density (PSD) in units of Watts per Hertz (W/Hz). For example, thermal noise at room temperature has a PSD of approximately $kT = 4 \times 10^{-21}$ W/Hz, or −174 dBm/Hz (with the Boltzmann constant $k = 1.38 \times 10^{-23}$ J/K). This means that if we have a bandwidth of 1 MHz, the total thermal noise power would be:

$$P_{\mathrm{thermal}} = \mathrm{PSD} \cdot B = -174\,\mathrm{dBm/Hz} + 10\log_{10}\left(\frac{1\,\mathrm{MHz}}{1\,\mathrm{Hz}}\right) = -114\,\mathrm{dBm}$$

The PSD of noise can be flat vs. frequency (which is called "white noise"), or can decrease with frequency (e.g., "flicker noise" or "1/f noise"). Further, noise can be generated by resistors (thermal noise), semiconductors (shot noise, generation-recombination noise), etc. A detailed discussion of noise sources can be found in [4] or [5].

2.3.1 Types of Noise Generation

Resistors generate thermal noise, which is white noise with a PSD of $4kTR$ (in V^2/Hz) when looking at the voltage across the resistor, or $4kT/R$ (in A^2/Hz) when looking at the current through the resistor. This noise is generated by the random thermal motion of charge carriers in the resistor.

! Thermal Noise

Note that the simple approximation given above is only valid for reasonably high frequencies and typical temperatures, and is known as the Rayleigh-Jeans approximation of Planck's blackbody radiation accounting for quantum effects and is given by [3]

$$\text{PSD} = \frac{4Rhf}{e^{hf/kT} - 1}$$

where h is the Planck constant ($h = 6.626 \times 10^{-34}$ Js) and f is the frequency. The Rayleigh-Jeans approximation is valid for $f \ll kT/h$, which is approximately 6 THz at room temperature (290 K).

We can integrate the above PSD over the full frequency range and show the rms noise voltage of a resistor R is bounded to

$$\overline{v_\text{n}^2} = \int_0^\infty \frac{4Rhf}{e^{hf/kT} - 1} df = \frac{2(\pi kT)^2}{3h} \cdot R$$

which equates to approximately 13 mVrms noise voltage for a 1 kΩ resistor at room temperature (which is impossible to measure in practice, as there will be some form of bandwidth limitation in any real measurement setup).

MOSFETs generate several types of noise, the most important ones being the thermal noise of the channel and flicker noise.

The thermal noise of the channel can be modeled as a current noise source between drain and source with a PSD of $\overline{I_\text{n}^2} = 4kT\gamma g_{d0}$ (in A^2/Hz), where γ is a process-dependent parameter (usually between 2/3 and 2). The parameter g_{d0} is the small-signal output conductance of the MOSFET in triode, i.e., $g_{d0} = g_\text{ds}$, or equal to $g_{d0} = g_\text{m}$ when in saturation.

In saturation, it is often useful to express the thermal noise as a voltage noise source at the gate with a PSD of $\overline{V_\text{n}^2} = 4kT\gamma/g_\text{m}$ (in V^2/Hz). We can see that we can lower this noise of the MOSFET by increasing the transconductance g_m, which can be achieved by increasing the bias current.

In addition, at high frequencies, the MOSFET also has induced gate-current noise, which is correlated with the channel thermal noise. A detailed discussion of this noise source can be found in [5].

Flicker noise is usually modeled as a voltage noise source at the gate with a PSD of $K_f/(C'_\text{ox}WLf)$ (in V^2/Hz), where K_f is a process-dependent parameter, C'_ox is the oxide capacitance per unit area, L and W are the length and width of the MOSFET, and f is the frequency. Note that we can lower the flicker noise by increasing the area of the MOSFET

(WL), however, this increases the parasitic capacitances associated with the MOSFET, and this often prohibitive for RF operation!

In **bipolar junction transistors (BJTs)**, the most important noise source is the shot noise due to the diffusion current in the base-emitter junction. Its PSD can be modeled as a current noise source between collector and emitter with a PSD of $2qI_\mathrm{C}$ (in A^2/Hz), where q is the elementary charge ($q = 1.6 \times 10^{-19}$ C) and I_C is the dc collector current.

> ! Equivalence of Shot and Thermal Noise
>
> Note that it has been shown in [6] that thermal noise and shot noise are actually equivalent, as both are generated by the random, thermally agitated motion of charge carriers!

Ideal **capacitors** and **inductors** do not generate noise, however, real capacitors and inductors have parasitic resistances which generate thermal noise.

In RF systems additional noise sources can be present. One noteworthy example is the **cosmic microwave background** radiation, which can be modeled as a noise temperature of approximately 3 K. While this is negligible compared to thermal noise at room temperature (approximately 290 K), it can be significant in very low-noise systems, such as radio telescopes pointing to the sky. An other important noise source in RF systems is the **atmospheric noise**, which is generated by natural phenomena like lightning or in the ionosphere.

> i A Note on Circuit Noise Calculations
>
> When doing circuit noise calculations, it is instructive to keep the following points in mind:
>
> - For circuit calculations involving noise sources it is convenient to replace the power spectral density by equivalent sinusoidal generators in small bandwidths.
> - The noise power spectral density in a small bandwidth Δf is given by $\overline{V_\mathrm{n}^2} = \overline{v_\mathrm{n}^2}/\Delta f$ and $\overline{I_\mathrm{n}^2} = \overline{i_\mathrm{n}^2}/\Delta f$.
> - The quantities $\overline{V_\mathrm{n}^2}$ and $\overline{I_\mathrm{n}^2}$ can be considered the mean-square value of sinusoidal generators. Using these values, network noise calculations reduce to familiar sinusoidal circuit-analysis calculations using V_n and I_n.
> - Multiple *independent* noise sources can be calculated individually at the output, and the total noise in bandwidth Δf is calculated as a mean-square value by adding the individual mean-square contributions from each sinusoid.

2.3.2 Noise in Impedance-Matched Systems

We now want to calculate the maximum noise power that can be extracted from a noisy source. We assume the following situation as shown in Figure 10. Note that the voltage source $\overline{V_\mathrm{n,s}^2}$ models the thermal noise of the source resistor R_s resulting in a Thevenin equivalent circuit.

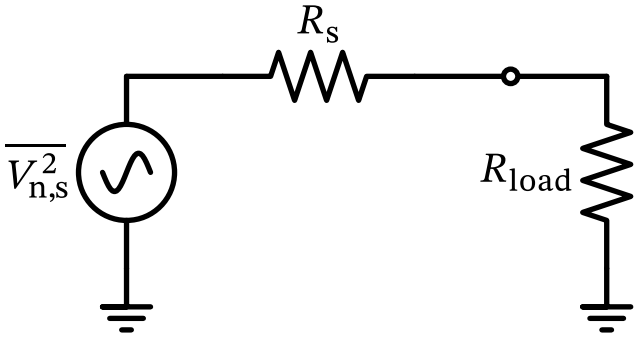

Figure 10: A noise-matched system with source and load impedances.

We know that the noise of the source resistor is given by $\overline{V_\mathrm{n,s}^2} = 4kTR_\mathrm{s}$. We assume the load resistor R_load as noiseless and matched to the source resistor, i.e., $R_\mathrm{load} = R_\mathrm{s}$ for maximum power transfer. The noise power spectral density delivered to the load resistor is then given by

$$P_\mathrm{n,load} = \frac{\overline{V_\mathrm{n,load}^2}}{R_\mathrm{load}} = \frac{\overline{V_\mathrm{n,d}^2}}{4R_\mathrm{s}} = kT \tag{15}$$

The calculation of Equation 15 confirms the initial statement that the maximum noise power spectral density that can be extracted from a noisy source is kT (in W/Hz). This result is independent of the actual value of the source resistance R_s.

We can further generalize the thermal noise of any impedance as

$$\overline{V_\mathrm{n}^2} = 4kT\Re\{Z\} \tag{16}$$

as for example in the complex impedance Z_ant of an antenna. Since an antenna is a reciprocal device, if we measure its radiation impedance Z_rad (for example with a vector network analyzer), we can calculate its thermal noise with Equation 16 to $\overline{V_\mathrm{n}^2} = 4kT\Re\{Z_\mathrm{rad}\}$.

2.3.3 Noise Figure

In RF systems, we often want to quantify the noise performance of a circuit block or a complete system. The most widely used metric is the **noise factor (F)**, which is defined as the ratio of the signal-to-noise ratio (SNR) at the input to the SNR at the output of a circuit block or system. If we express the noise factor in dB, we call it the **noise figure (NF)** [3]. The noise factor is given by

$$F = \frac{\mathrm{SNR}_\mathrm{in}}{\mathrm{SNR}_\mathrm{out}} = \frac{\left(P_\mathrm{s}/P_\mathrm{n}\right)_\mathrm{in}}{\left(P_\mathrm{s}/P_\mathrm{n}\right)_\mathrm{out}} \tag{17}$$

where P_s is the signal power and P_n is the noise power. The noise factor is always larger than or equal to 1 (or 0 dB), as no circuit can improve the SNR!

! SNR Improvement

Note that the SNR can be improved by filtering, as filtering reduces the noise power. If the noise bandwidth is larger than the signal bandwidth, then the SNR can be improved without affecting the signal. However, this is not considered in the noise factor, as the noise factor assumes that both signal and noise pass through the same bandwidth.

Let us look at a simple model of a noise circuit block as shown in Figure 11. The input signal S_{in} is accompanied by noise N_{in}. By definition it is assumed that the input noise power results from a matched resistor at $T_0 = 290\,\text{K}$, so that $N_{\text{in}} = kT_0$. The circuit block has a power gain G and adds its own noise N_{dut} to the output signal. For simplicity, we assume that the input and output of the circuit block are impedance matched to avoid reflections.

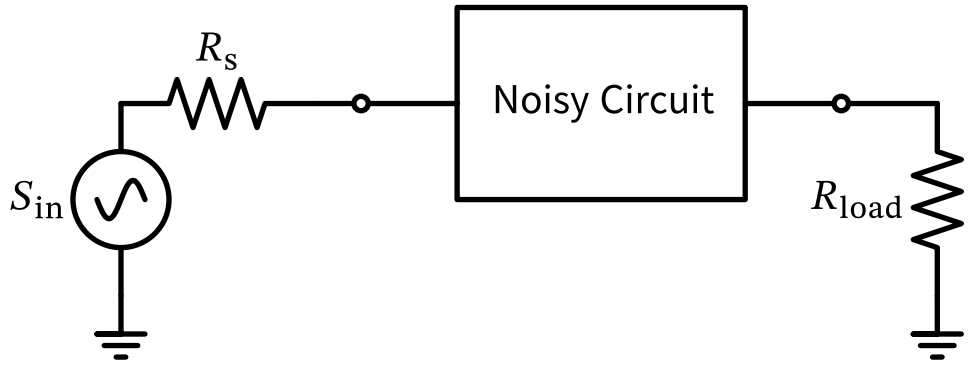

Figure 11: A noise-matched system with source and load impedances and a noisy circuit block.

The output signal and noise powers are then given by

$$S_{\text{out}} = GS_{\text{in}}$$

$$N_{\text{out}} = GN_{\text{in}} + N_{\text{dut}}$$

The resulting noise factor can then be calculated as

$$F = \frac{S_{\text{in}}/N_{\text{in}}}{S_{\text{out}}/N_{\text{out}}} = \frac{1}{G}\frac{GN_{\text{in}} + N_{\text{dut}}}{N_{\text{in}}} = 1 + \frac{N_{\text{dut}}}{GN_{\text{in}}},$$

in other words, the noise factor is 1 plus the ratio of the noise added by the device under test (DUT) to the amplified input noise.

Note that a noiseless block ($N_{\text{dut}} = 0$) has a noise factor of $F = 1$. A passive block with loss factor L (and impedance matched at input and output) has a noise factor of $F = L$ (in linear units), as it attenuates the signal and $N_{\text{out}} = N_{\text{in}} = kT$ if everything is in thermal equilibrium.

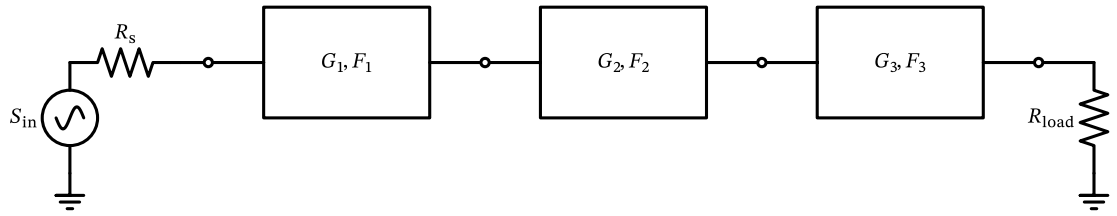

Figure 12: Block cascade for noise factor calculation showing multiple stages with gains and individual noise factors.

If we have a cascade of multiple blocks, as shown in Figure 12, we can calculate the overall noise factor with the **Friis formula** [3]

$$F_{\text{total}} = 1 + (F_1 - 1) + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} \tag{18}$$

where F_i and G_i are the noise factor and power gain of the i-th block, respectively. Note that all gains have to be in linear units (not dB) when using Equation 18. We can interpret Equation 18 as follows:

- The overall noise factor F_{total} is always larger than or equal to the noise factor of the first block (F_1).
- The noise factor of the first block is the most important one, as the noise factors of the following blocks are reduced by the gain of all preceding blocks. This is especially important in RF receivers, where the first block is usually a low-noise amplifier (LNA) with a very low noise figure (e.g., 1 dB or less) and a high gain (e.g., 10 dB or more). This ensures that the noise of the following blocks is negligible.
- The noise factor of the last block is reduced by the gain of all preceding blocks, so it is usually not very important.

Here we also see a trade-off between noise and linearity, as shown by Equation 13 and Equation 18. For low noise, we should try to maximize G_1, however, this will affect linearity (IIP3) in a negative way. As in many other situation in RF design, we have to find a good compromise between conflicting requirements.

2.3.4 Sensitivity

In RF receivers, we often want to know the minimum input signal power that can be detected with a certain SNR. This minimum input signal power is called the **sensitivity** of the receiver. The sensitivity can be calculated as

$$P_{\text{in,min}} = P_{\text{n}} \cdot \text{SNR}_{\text{min}} \cdot F \tag{19}$$

where P_{n} is the noise power at the input, SNR_{min} is the minimum detectable SNR, and F is the noise factor of the receiver. The input noise power can be calculated as

$$P_{\text{n}} = kTB$$

where k is the Boltzmann constant, T is the temperature in Kelvin, and B is the bandwidth of the receiver. Expressing Equation 19 in dBm we get the following formula:

$$P_{\text{in,min}}|_{\text{dBm}} = -174\,\text{dBm/Hz} + \text{NF} + 10\log_{10}(B/\ \text{Hz}) + \text{SNR}_{\text{min}}|_{\text{dB}} \tag{20}$$

where −174 dBm/Hz is the thermal noise PSD at room temperature (290 K). We can see that the sensitivity improves with lower noise figure, smaller bandwidth, and lower minimum detectable SNR.

Note 5: Sensitivity Calculation for WiFi

Let's calculate the sensitivity of a WiFi receiver operating at 5 GHz with a bandwidth of $B = 80\,\text{MHz}$, a noise figure of $NF = 7\,\text{dB}$, and a minimum detectable SNR of 25 dB. This high SNR means that a high-order modulation scheme (like 64-QAM) is used for high data rates.

Using Equation 20 we get:

$$P_{\text{in,min}} = -174\,\text{dBm/Hz} + 7\,\text{dB} + 10\log_{10}(80 \times 10^6) + 25\,\text{dB} \approx -63\,\text{dBm}$$

This means that the minimum input signal power that can be detected by the WiFi receiver is approximately −63 dBm.

2.4 Modulation

In order to transmit information via an EM wave, we need to modulate the EM wave with the information signal. Looking at a simple sinusoidal carrier wave

$$s(t) = A\cos(\omega_0 t + \varphi)$$

we see that we can change one or more of the following parameters to encode information:

- Amplitude $A(t)$ (amplitude modulation, AM; the digital form is called amplitude-shift keying, **ASK**)
- Frequency $\omega_0(t)$ (frequency modulation, FM; the digital form is called frequency-shift keying, **FSK**)
- Phase $\varphi(t)$ (phase modulation, PM; the digital form is called phase-shift keying, **PSK**)
- Amplitude $A(t)$ and phase $\varphi(t)$ (quadrature amplitude modulation, **QAM**)

The modulation formats FM and PM have the advantage that the carrier amplitude is constant, which makes them more robust against nonlinear distortion.

QAM is widely used in modern communication systems, as it allows to transmit more bits per symbol by combining amplitude and phase modulation. The form with 4 different symbols is called QPSK. Higher-order modulation like 16-QAM, for example, uses 16 different symbols, which can encode 4 bits per symbol (as $2^4 = 16$). Even higher-order QAM formats like 64-QAM (6 bits per symbol), 256-QAM (8 bits per symbol), 1024-QAM (10 bits per symbol), or 4096-QAM (12 bits per symbol) are also used in modern systems like WiFi or LTE.

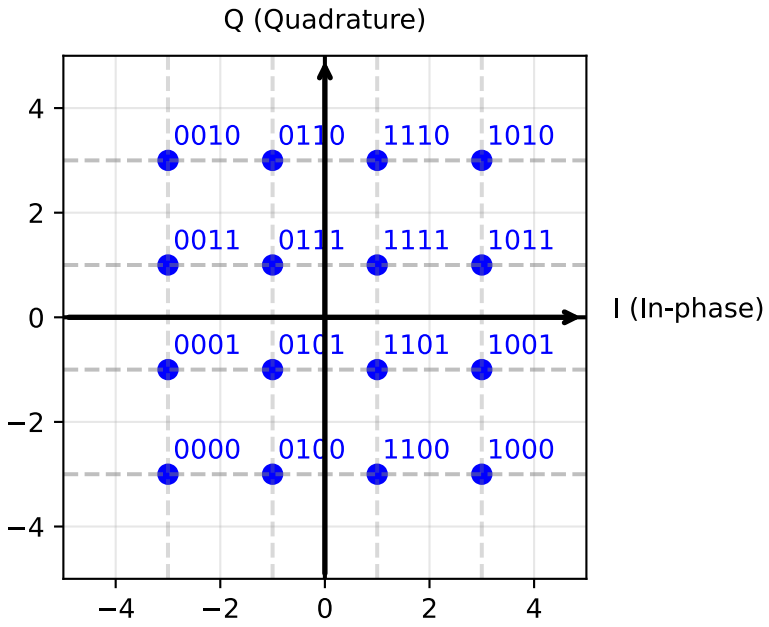

Figure 13: 16-QAM constellation diagram with Gray code labeling of constellation points.

Shown in Figure 13 is the "constellation diagram" of a 16-QAM modulation format. The constellation points are arranged in a square grid, with each point representing a unique combination of amplitude and phase. The distance between the constellation points determines the robustness against noise and interference; larger distances result in better performance, but also require more bandwidth. The mapping of bits to constellation points is called "bit mapping" or "symbol mapping". The example in Figure 13 uses a Gray code mapping, which minimizes the number of bit errors in case of a symbol error.

The constellation diagram can be imagined as a complex plane, where the x-axis represents the in-phase component (I) and the y-axis represents the quadrature component (Q) of the modulated signal. During transmission of a specific symbol, the RF carrier is modulated to the corresponding amplitude and phase, resulting in a specific point in the constellation diagram.

The table below shows the SNR requirements for different modulation formats to achieve a bit error rate (BER) of 10^{-5} in an additive white Gaussian noise (AWGN) channel. As we can see, higher-order modulation formats require higher SNR to achieve the same BER.

Table 3: SNR requirements for different modulation schemes to achieve BER = 10^{-5} in AWGN channel

Modulation	Bits/Symbol	Required SNR (dB)
BPSK	1	9.6
QPSK	2	12.6
16-QAM	4	18.2
64-QAM	6	24.4
256-QAM	8	30.6
1024-QAM	10	36.9
4096-QAM	12	43.2

2.5 Pulse Shaping and Spectral Efficiency

When we modulate symbols onto a carrier, we usually do not transmit the symbols as pure sinusoids, but rather as pulses with a certain shape. The pulse shape determines the bandwidth of the transmitted signal and its spectral efficiency. A common pulse shape is the rectangular pulse, which has a sinc-shaped spectrum. However, the sinc function $\sin(\pi x)/\pi x$ has side lobes that extend to infinity, which can cause interference with adjacent channels.

For reference, the spectrum of a random binary sequence with equal probability of 0s and 1s, using rectangular pulses with a duration of T_b is given by ($S(f)$ is the two-sided power spectral density):

$$S(f) = \frac{T_\mathrm{b}}{4}\mathrm{sinc}^2(fT_\mathrm{b}) + \frac{1}{4}\delta(f) = \frac{T_\mathrm{b}}{4}\left(\frac{\sin(\pi f T_\mathrm{b})}{\pi f T_\mathrm{b}}\right)^2 + \frac{1}{4}\delta(f)$$

To avoid this, we can use pulse shapes that have better spectral properties, such as the raised cosine pulse or the root-raised cosine pulse. The raised-cosine pulse has a roll-off factor α that determines the excess bandwidth beyond the Nyquist bandwidth. The **root-raised cosine (RRC)** pulse is used in practical systems, as it can be implemented with a matched filter at the receiver.

The raised-cosine pulse $p(t)$ (with a spectrum shaped like a raised cosine) is given by:

$$p(t) = \frac{\sin(\pi t/T_\mathrm{b})}{\pi t/T_\mathrm{b}} \cdot \frac{\cos(\alpha\pi t/T_\mathrm{b})}{1-(2\alpha t/T_\mathrm{b})^2}$$

Setting $\alpha = 0$ results in a sinc pulse in the time domain (with a perfect bandwidth containment in the frequency domain), while $\alpha = 1$ results in a pulse with double the Nyquist bandwidth. The pulse shape for $\alpha = 0$ and $\alpha = 0.22$ (used in 3G) is shown in Figure 14.

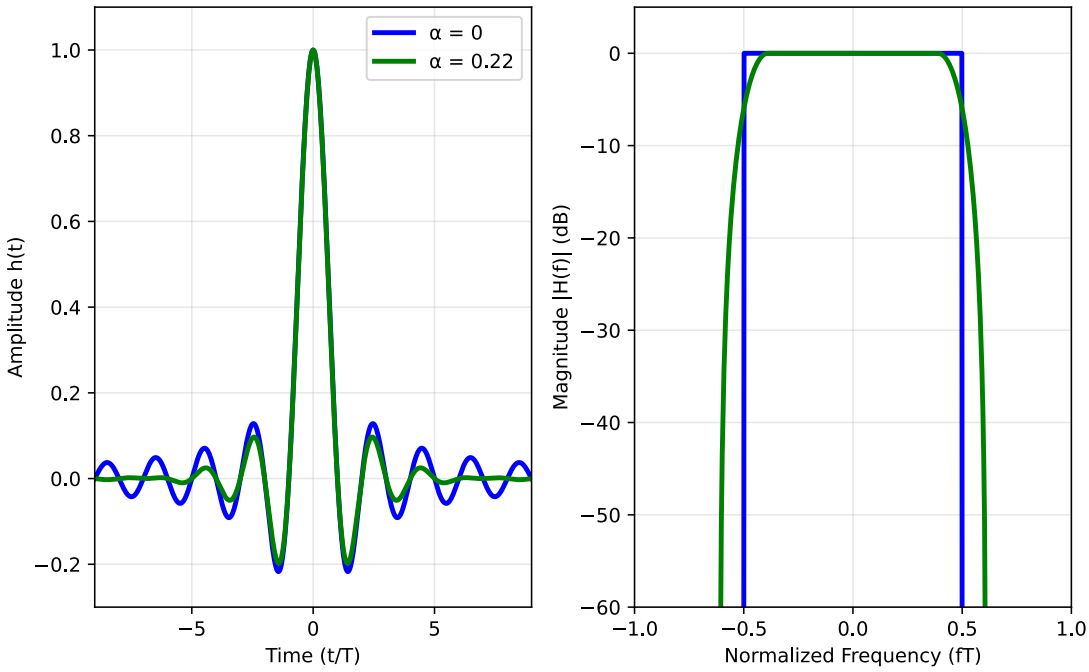

Figure 14: Raised cosine pulse shaping in time and frequency domain for different roll-off factors α.

Another often-used pulse shape is the Gaussian pulse, which is used in Gaussian minimum-shift keying (GMSK, used in 2G) modulation, or in Gaussian frequency-shift keying (GFSK, used in Bluetooth). The Gaussian pulse has a smooth shape and a narrow spectrum. The Gaussian pulse is given by:

$$p(t) = \frac{\sqrt{\pi}}{\alpha} e^{-(\pi t/\alpha)^2} \quad \text{with} \quad \alpha = \frac{\sqrt{\ln 2}}{\sqrt{2}} \cdot \frac{T_\mathrm{b}}{BT_\mathrm{b}}$$

where BT_b controls the width of the pulse. The spectrum of the Gaussian pulse is also Gaussian-shaped, which helps to minimize inter-symbol interference (ISI).

The Gaussian pulse for $BT = 0.5$ as used in Bluetooth is shown in Figure 15.

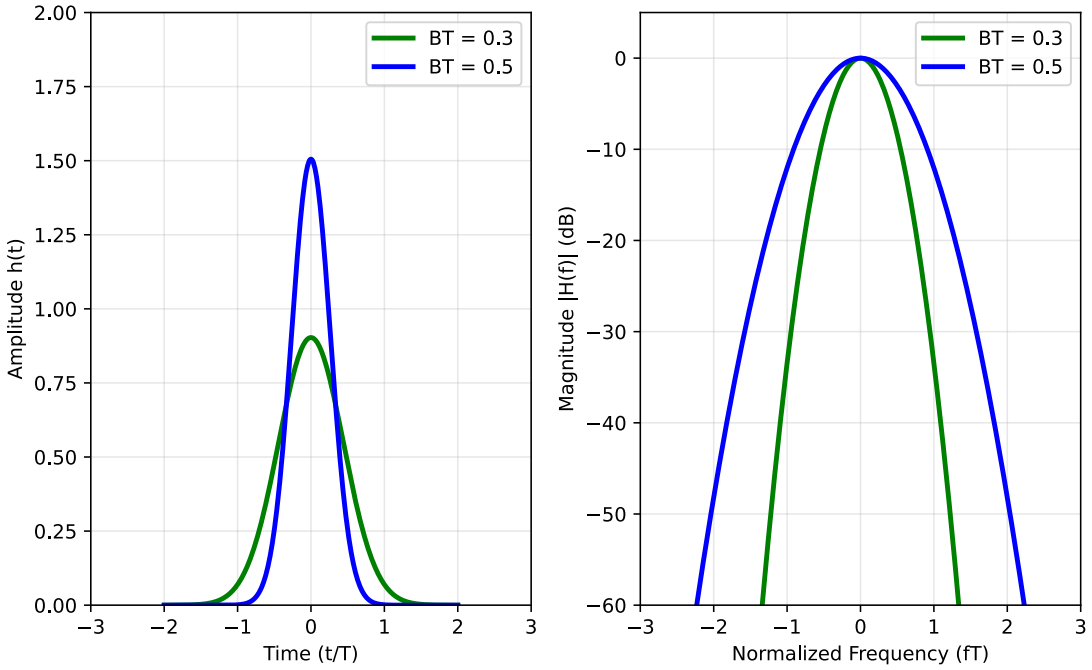

Figure 15: Gaussian pulse shaping in time and frequency domain for different bandwidth-time products BT.

For both the raised-cosine and Gaussian pulse, the trade-off between time- and frequency-domain containment is clearly visible. This is also captured in "**Küpfmüller's uncertainty principle**", which states that the product of the time duration and the bandwidth of a pulse is lower-bounded by a constant. In other words, if we want to have a pulse that is very short in time, it will have a wide bandwidth, and vice versa.

2.6 Orthogonal Frequency-Division Multiplexing (OFDM)

As we have seen in the previous section, if we make the symbol rate high, we need to use pulses with a wide bandwidth. The problem with a wide bandwidth in wireless communication is **multi-path propagation**, which causes frequency-selective fading. This means that some frequencies are attenuated more than others, which can cause errors in the received signal. Equalizing such a frequency-selective channel can be very complex, especially if the channel changes rapidly (as in mobile communication). We now face a dilemma: How can we achieve high data rates (which require high symbol rates and thus wide bandwidth) while avoiding frequency-selective fading? The key idea, implemented in **OFDM**, is to split the wideband channel into multiple narrowband sub-channels (subcarriers), each with a low symbol rate. This way, each subcarrier experiences flat fading, which is much easier to equalize.

The key question is now how to implement this idea efficiently, as we now have to apply modulation to hundreds or thousands of individual subcarriers. The solution is to use the **inverse fast Fourier transform (IFFT)** at the transmitter to generate the time-domain OFDM signal from the frequency-domain symbols, and the **fast Fourier transform (FFT)** at the receiver to recover the frequency-domain symbols from the time-domain OFDM signal. This is illustrated in Figure 16.

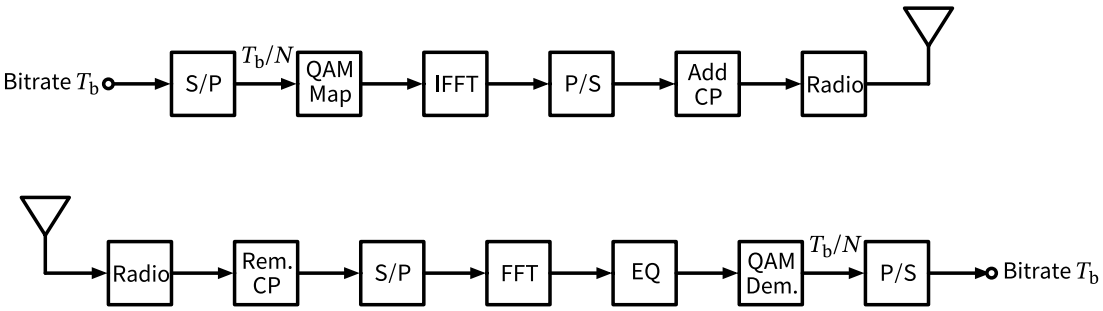

Figure 16: OFDM transmission system block diagram showing transmitter and receiver processing chains.

The OFDM transmitter takes a block of N symbols (e.g., 64-QAM symbols) and maps them onto N subcarriers, thereby reducing the symbol rate for each subcarrier to T_b/N. The IFFT then generates the time-domain OFDM signal, which is transmitted over the wireless channel. Before transmission the **cyclic prefix (CP)** is added to each OFDM symbol.

At the receiver, first the CP is removed, and then the FFT recovers the frequency-domain symbols, which can then be equalized (fairly simply by multiplying each subcarrier with a complex factor to correct amplitude and phase) and demodulated.

A key property of OFDM is that the subcarriers are **orthogonal** to each other, which means that they do not interfere with each other, even if they overlap in frequency. This is achieved by choosing the subcarrier spacing Δf such that it is equal to the symbol rate $1/T_\mathrm{b}$, i.e., $\Delta f = 1/T_\mathrm{b}$. This way, the integral of the product of two different subcarriers over one symbol period is zero, which means that they are orthogonal.

To further improve the robustness against multi-path propagation, a CP is added to each OFDM symbol. The CP is a copy of the last part of the OFDM symbol, which is added to the beginning of the symbol. This way, if there are delayed copies of the OFDM symbol due to multi-path propagation, they will still fall within the CP and will not cause inter-symbol interference (ISI). The length of the CP should be longer than the maximum delay spread of the channel.

Note 6: OFDM in LTE

In LTE OFDM is used for the downlink (base station to user equipment) with the following parameters:

- Subcarrier spacing: 15 kHz
- CP length: 5.2 µs (normal), 4.7 µs (extended)
- Number of subcarriers: 1200 (for 20 MHz bandwidth)
- Modulation: QPSK, 16-QAM, 64-QAM, 256-QAM

From the subcarrier spacing we can calculate the symbol duration as $T_\mathrm{b} = 1/\Delta f = 1/15\,\mathrm{kHz} \approx 66.7\,\mu\mathrm{s}$.

We can calculate the raw bitrate for a 20 MHz LTE channel as

$$\mathrm{Bitrate} = N_\mathrm{sc} \cdot N_\mathrm{sym} \cdot \frac{1}{T_\mathrm{b} + T_\mathrm{CP}} = 1200 \cdot 8 \cdot \frac{1}{66.7\,\mu\mathrm{s} + 5.2\,\mu\mathrm{s}} \approx 133\,\mathrm{Mbps}$$

Without the overhead for control channels and error correction coding a user datarate of approximately 100 Mbps can be achieved in a 20 MHz LTE channel.

2.7 Multiple Access Techniques

In wireless communication systems, multiple users need to share the same frequency spectrum. This is achieved by using **multiple access techniques**, which allow multiple users to transmit and receive data simultaneously without interfering with each other. The most common multiple access techniques are:

1. **Time division multiple access (TDMA)**: Users are assigned specific time slots for transmission, allowing multiple users to share the same frequency channel by dividing the time into slots.
2. **Frequency division multiple access (FDMA)**: Users are assigned specific frequency bands within the overall frequency spectrum, allowing multiple users to transmit simultaneously on different frequencies.
3. **Code division multiple access (CDMA)**: Users are assigned unique spreading codes, allowing them to transmit simultaneously over the same frequency band. The receiver uses the code to extract the desired signal. A variant of CDMA is frequency-hopping spread spectrum (FHSS), where the carrier frequency is changed rapidly according to a pseudo-random sequence known to both the transmitter and receiver. This is used in Bluetooth.
4. **Orthogonal frequency division multiple access (OFDMA)**: A variant of OFDM, where multiple users are assigned different subcarriers for transmission, allowing for efficient use of the frequency spectrum. This is used in 4G LTE and 5G NR.
5. **Spatial division multiple access (SDMA)**: Uses multiple antennas to create spatially separated channels, allowing multiple users to transmit simultaneously in the same frequency band.

In addition, all of these techniques can be combined to create more efficient and flexible communication systems. For example, OFDMA can be used in conjunction with SDMA to

allow multiple users to share the same frequency resources while also taking advantage of spatial diversity. Also TDMA can be combined with FDMA to create a hybrid multiple access scheme (which has been used in 2G GSM).

3 Transceivers

Nowadays, the various small-signal RF functions for receive and transmit are integrated into so-called transceivers (TRX). A TRX is a device that can both transmit and receive signals, and is usually called an "RFIC". While high monolithic integration is certainly the norm for radio-frequency devices intended for standards like Bluetooth, WiFi, cellular, etc., it is increasingly used also for mm-wave frequencies for applications like automotive radar and 5G cellular.

Typically TRX include components like amplifiers, mixers, filters, oscillators, and phase-locked loops. When digital interfaces are used for the baseband data transport also functions like analog-to-digital conversion (ADC) and digital-to-analog conversion (DAC) are integrated together with digital signal processing (DSP) blocks and potentially high-speed interfaces.

In this lecture we will focus on the RF part of a TRX, which is responsible for the upconversion of baseband or intermediate frequency (IF) signals to the desired transmit frequency during transmission, and the downconversion of received signals from the carrier frequency to baseband or IF during reception. For filters, low-frequency amplifiers, ADCs, DACs, and DSP blocks we refer to related courses and literature, for example our analog circuit design course.

3.1 Direct-Conversion Transceiver

The following typical functions have to be performed by a TRX:

- Pulse-shaping filtering of the baseband signal (can be implemented analog or in most cases digital).
- Modulating the baseband signal onto a carrier frequency (upconversion) in the TX or downconversion in the RX.
- Contain the RF signal in a small bandwidth (TX), or single out the wanted signal in the RX.
- Adapt gain (and linearity) to the signal strength in the RX, and to the output power in the TX.
- Generate the carrier frequency (local oscillator, LO) with low phase noise.

The dominant architecture for the TRX is the so-called direct-conversion (or Zero-IF) architecture, where the upconversion and downconversion is performed in a single step. This is in contrast to superheterodyne architectures, where the signal is first converted to an intermediate frequency (IF) before being converted to baseband. The direct-conversion architecture has the advantage of reduced complexity and cost, as it requires fewer components and less filtering. However, it also has some disadvantages, such as increased susceptibility to DC offsets and I/Q imbalance. A typical TRX block diagram is shown in Figure Figure 17.

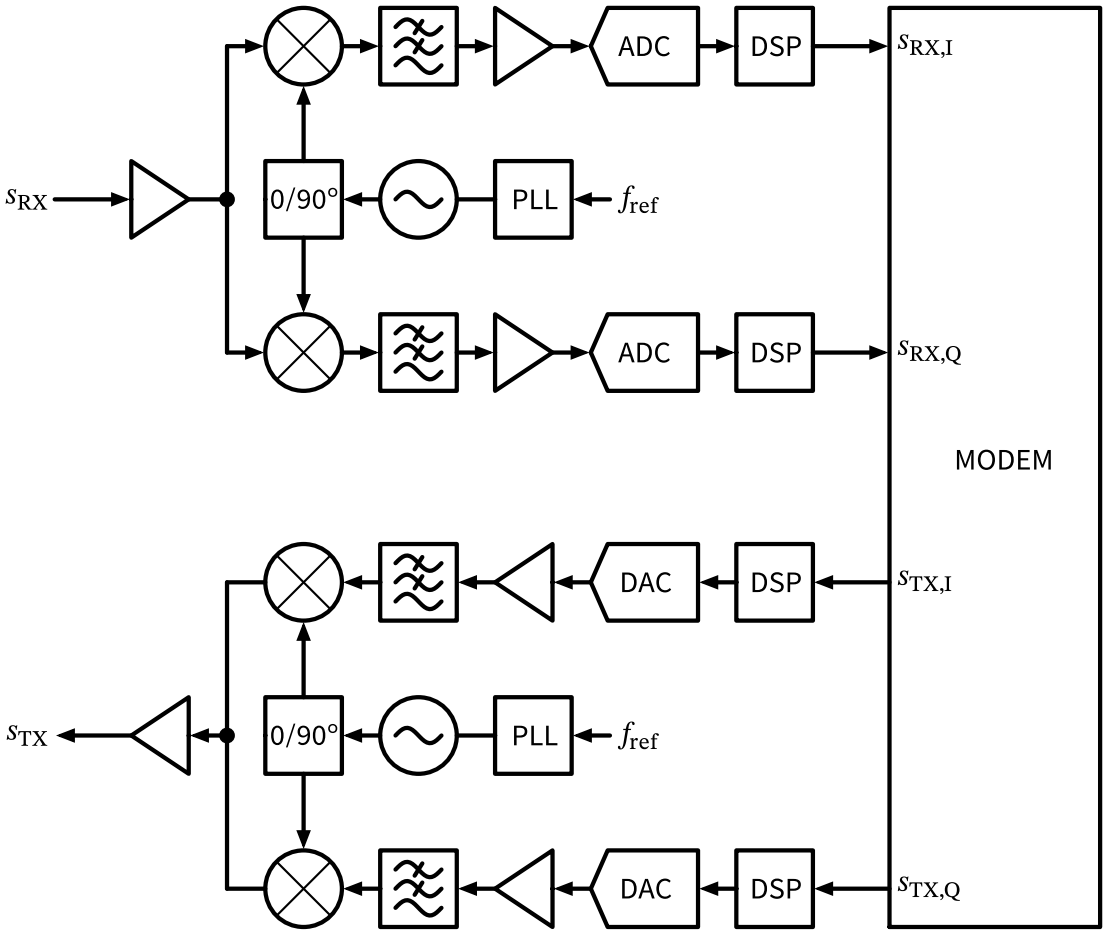

Figure 17: Block diagram of a typical transceiver (TRX) showing the main functional blocks of RX and TX. The modem provides the digital baseband processing and interfaces to the rest of the system.

As can be seen in Figure 17, this generic example can be adapted in various ways. Generally, the amplifier gains are adjustable to adapt to different signal levels. If various channel bandwidths are to be supported, the corner frequencies of the low-pass filters (LPF) can be adjusted, as well as (optionally) the sampling rate of the ADCs and DACs. The local oscillator (LO) frequency is generated by a phase-locked loop (PLL) synthesizer, which can be tuned to the desired carrier frequency. In case of frequency-division duplex (FDD) operation, two PLLs are used to generate the TX and RX LO frequencies, which are separated by the duplex distance. In time-division duplex (TDD) operation, a single PLL is sufficient, supplying the LO signal to both RX and TX.

The modem that is shown in Figure 17 is responsible for the digital baseband processing, including functions like channel coding/decoding, modulation/demodulation, equalization, and error correction. The modem is usually implemented as a digital System-on-Chip (SoC) consisting of (multiple) CPUS, DSPs, and fixed-function blocks for time-critical processing. For an in-depth discussion we recommend [7] or [8].

3.2 Modulation and Demodulation

Modulation is the process of varying a carrier signal at frequency f_c in order to transmit information. The complex baseband signal (after converting the real-valued digital s_I and s_Q signals to analog and pulse-shaping filtering) is represented as

$$s_\mathrm{BB}(t) = s_\mathrm{I}(t) + j s_\mathrm{Q}(t).$$

We want to shift this signal to the carrier frequency f_c, which can be done by multiplying with a complex exponential:

$$s_\mathrm{RF,complex}(t) = s_\mathrm{BB}(t) \cdot e^{j\omega_\mathrm{c} t} = \left[s_\mathrm{I}(t) + j s_\mathrm{Q}(t)\right] \cdot \left[\cos(\omega_\mathrm{c} t) + j \sin(\omega_\mathrm{c} t)\right].$$

The real-valued RF signal is obtained by taking the real part of this expression:

$$s_\mathrm{RF}(t) = \Re\left\{s_\mathrm{RF,complex}(t)\right\} = s_\mathrm{I}(t)\cos(\omega_\mathrm{c} t) - s_\mathrm{Q}(t)\sin(\omega_\mathrm{c} t). \tag{21}$$

The process formulated in Equation 21 is done in the TX, as shown in Figure 18.

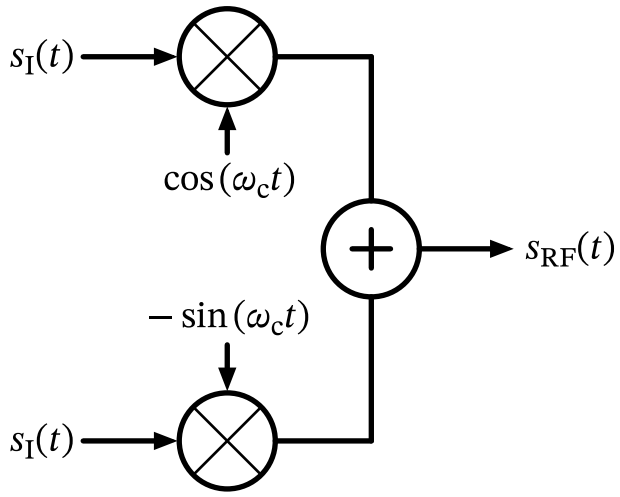

Figure 18: TX modulator.

The RF signal generation according to Equation 21 is called quadrature modulation. This is the modulation used most often in modern communication systems, as it allows to transmit two independent signals (I and Q) in the same bandwidth. The I and Q signals are also called quadrature components, as they are 90° out of phase with each other.

Alternatively, a modulation called polar modulation can be used, where the amplitude and phase of the carrier are varied according to the baseband signal. This is done by converting the I and Q signals to polar coordinates

$$s_\mathrm{RF}(t) = \Re\left\{A(t) \cdot e^{j\varphi(t)} \cdot e^{j\omega_\mathrm{c} t}\right\}$$

with

$$A(t) = \sqrt{s_\mathrm{I}^2(t) + s_\mathrm{Q}^2(t)}, \quad \varphi(t) = \tan^{-1}\left(\frac{s_\mathrm{I}(t)}{s_\mathrm{Q}(t)}\right).$$

As the mathematical operations required for the cartesian to polar transformation are quite nonlinear, the $A(t)$ and $\phi(t)$ signals are wideband. Some wireless standards allow efficient use of polar modulation, for example Bluetooth, where basically all TX are realized as polar modulators.

In the RX, the received RF signal is downconverted to baseband by a similar process, as shown in Figure 19.

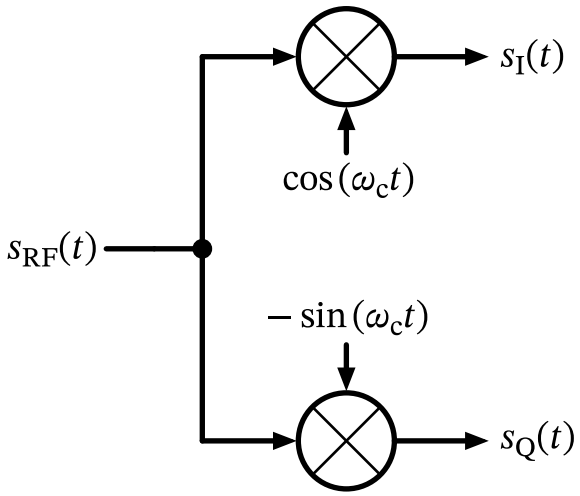

Figure 19: RX demodulator.

For demodulation we have to shift the RF signal down to baseband, which mathematically is done by multiplying with the complex conjugate of the carrier:

$$s_{\mathrm{BB,complex}}(t) = s_{\mathrm{RF}}(t) \cdot e^{-j\omega_{\mathrm{c}}t} = s_{\mathrm{RF}}(t) \cdot [\cos(\omega_{\mathrm{c}}t) - j\sin(\omega_{\mathrm{c}}t)] \tag{22}$$

3.3 Filtering

Filtering is an essential function in both TX and RX. In the TX, filtering is used to limit the bandwidth of the transmitted signal to the allocated channel bandwidth, and to suppress out-of-band emissions. In the RX, filtering is used to select the wanted signal from a crowded spectrum, and to suppress unwanted signals (blockers) that can cause interference or desensitization of the RX. A typical example of filtering in the RX is shown in Figure 20, where a bandpass filter is used to attenuate strong unwanted blockers while only slightly attenuating the wanted signal.

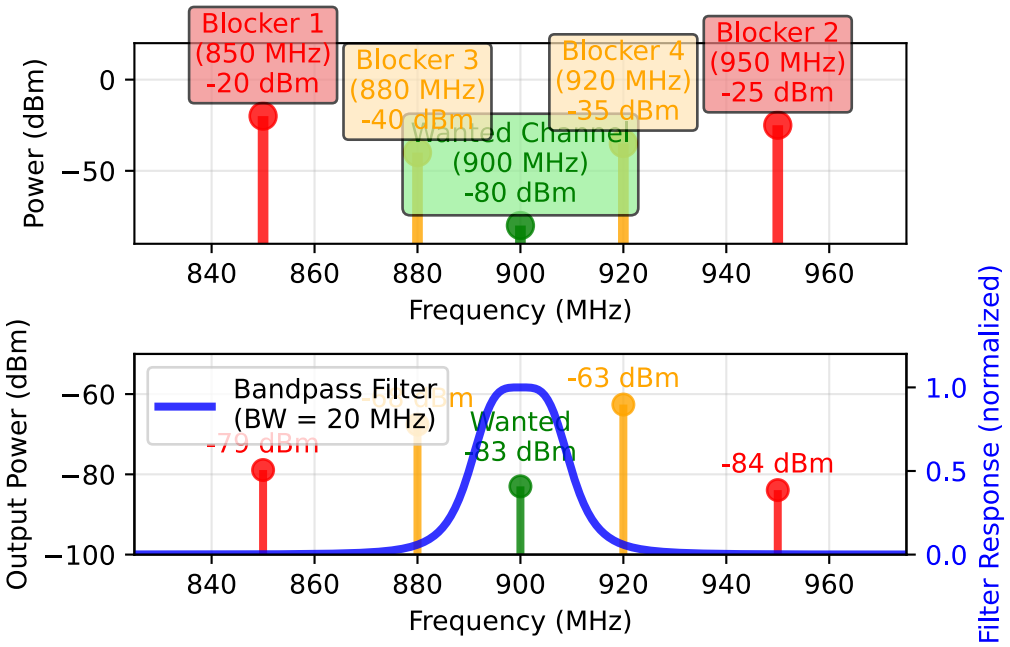

Figure 20: Filtering of wanted channel amid strong unwanted blockers. Exemplary shown in an RX scenario around 900 MHz. The strong blockers (top figure) are attenuated by an RF bandpass filter (bottom figure) with a bandwidth of 20 MHz, achieving more than 40 dB rejection of the blockers while only slightly attenuating the wanted signal.

In any filter there exists a fundamental trade-off between selectivity (steepness of the filter skirts), bandwidth, and insertion loss. A very selective filter with steep skirts and large BW will have a high insertion loss. Conversely, a filter with low insertion loss will have a gentle roll-off and may not sufficiently suppress unwanted signals. A useful metric to quantify the performance of a filter is the quality factor Q, defined as

$$Q = \frac{f_c}{\Delta f}$$

where f_c is the center frequency and Δf is the −3 dB bandwidth of the filter. A higher Q indicates a more selective filter.

The achievable Q depends on the filter technology used. For example, on-chip LC filters can achieve Q values of around 10-20, while off-chip SAW or BAW/FBAR filters can achieve Q values of several hundreds, and a crystal filter can achieve Q values of several thousands. The choice of filter technology depends on the application requirements, such as frequency range, bandwidth, insertion loss, and cost. Generally speaking, the required filtering to single out the wanted signal in the RX spectrum and decrease the power of strong blockers to a tolerable level is one of the most critical design choices, and is usually distributed at different locations in the RX chain:

- RF filters (between antenna and LNA) provide a first level of filtering, and are usually implemented as off-chip SAW or BAW/FBAR filters. They provide high Q and good selectivity, but have a fixed center frequency and bandwidth. They are used to pass the wanted band of interest, and to attenuate strong out-of-band blockers.

- IF filters (in case of a super-heterodyne receiver) provide additional filtering, and can be implemented as on-chip LC filters or off-chip SAW/BAW filters. They provide moderate Q and selectivity, and can be tuned to some extent.
- BB filters (after downconversion) provide the final level of filtering before entering the ADCs, and are usually implemented as on-chip active RC filters. They provide channel selection, and can be easily adjusted to different bandwidths.
- Digital filters (in the DSP block) provide the final level of filtering and signal processing, and can be implemented as FIR or IIR filters. They provide high flexibility and can be easily adapted to different standards and requirements. Digital filters show now variations, so they can be designed to be very selective.

It is important to note (because this dictates a lot of choices in RF design) that high-Q filters are usually fixed-frequency and fixed-bandwidth. Only baseband and digital filters can be easily adjusted to different bandwidths!

! Filter Technologies

Baseband filters (analog) are usually implemented as active RC filters on-chip. They are very flexible and can have adjustable bandwidth by either changing R and/or C. For medium frequencies $g_\mathrm{m} - C$ filters can be used, which are also tunable by changing the transconductance g_m and/or C. For even higher bandwidths, on-chip LC filters can be used, which have a limited Q of around 10-20.

Baseband filters (digital) are implemented as FIR or IIR filters in the DSP block. They are very flexible and can be easily adapted to different standards and requirements. Digital filters show now variations, so they can be designed to be very selective.

Surface acoustic wave (SAW) and **bulk acoustic wave (BAW/FBAR)** filters are off-chip components that can achieve high Q values of several hundreds. They have a fixed center frequency and bandwidth. Usually 1-2 such filters are required per supported band of interest.

Crystal filters can achieve very high Q values of several thousands, but are usually bulky and expensive.

LC filters can be either implemented off-chip (using discrete components) or on-chip. Off-chip LC filters can achieve higher Q values than on-chip LC filters, but are usually larger and more expensive. On-chip LC filters are limited in Q (around 10-20), but are very compact and can be integrated into the RFIC. Off-chip LC filters can achieve Q values of around 50-100, depending on the frequency and component quality.

Ceramic filters are another off-chip filter technology that can achieve moderate to high Q values (up to several hundreds). They are usually smaller and less expensive than SAW or BAW/FBAR filters, but also lower performance.

Waveguide filters are used at very high frequencies (above 10 GHz) and can achieve very high Q values (up to several thousands). They are usually bulky and expensive, and are not commonly used in mobile applications, but rather in fixed installations like base stations or satellite communication.

Fundamentally, the choice of filter technology is a trade-off between performance, size, cost, and flexibility. In most cases, a combination of different filter technologies is used to achieve the desired performance.

3.4 Direct-Conversion Architecture

The transceiver architecture shown in Figure 17 is called direct-conversion or zero-IF architecture, as the downconversion in the RX and upconversion in the TX is done in a single step. This architecture several advantages:

- Per RX and TX a single LO is required (which can even be shared between RX and TX in TDD operation).
- There are a minimum number of RF blocks, which good for cost and power consumption.
- This architecture is very flexible and can be easily adapted to different standards and requirements, and shows generally very good performance if the disadvantages can be overcome by good design.
- This architecture allows a high integration level, as basically all blocks can be implemented on-chip.
- Direct conversion is the de-facto standard architecture for cellular, WiFi, Bluetooth (with the exception of the TX), and GNSS.

However, the direct-conversion architecture also has some disadvantages:

- LO-RF coupling can cause self-mixing and desensitization of the RX, as well as LO leakage in the TX. This is an issue because the LO frequency is the same as the RF frequency.
- Even-order distortion products (especially IIP2) cause sensitivity degradation due to strong amplitude-modulated blockers.
- LO pulling can occur in the TX (again, LO and RF are at the same frequency).
- IQ errors (gain and phase mismatch) of the I and Q paths can cause constellation distortion leading to increased error vector magnitude (EVM).
- DC offsets can occur due to self-mixing of LO leakage and even-order distortion products.
- Flicker noise (1/f noise) upconversion can cause increased phase noise close to the carrier, as well as increased RX noise figure.

Nowadays there exist good design techniques to mitigate these disadvantages. However, in some cases (for example very high linearity requirements, or very high frequencies) other architectures like low-IF or super-heterodyne may be preferred.

3.5 Duplexing

In the block diagram of Figure 17, we have not yet considered how to share the antenna between RX and TX. Essentially, there are two main methods to achieve this: **frequency-division duplex (FDD)** and **time-division duplex (TDD)**.

3.5.1 Frequency-Division Duplex (FDD)

In FDD, the RX and TX operate at different frequencies, separated by a duplex distance. This allows simultaneous transmission and reception, which is beneficial for applications like voice communication where low latency is required. However, FDD requires two separate frequency bands, which can be a limitation in terms of spectrum availability. Additionally, FDD requires two PLLs to generate the RX and TX LO frequencies, which increases complexity and power consumption.

The RF RX and TX paths are connected to the antenna via a duplexer, which is a three-port device that allows signals to pass between the antenna and the RX or TX path, while isolating the RX and TX paths from each other. A typical FDD TRX block diagram is shown in Figure 21.

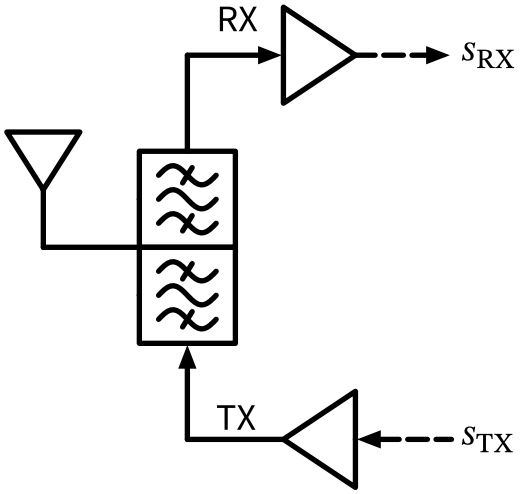

Figure 21: Block diagram of an FDD RF front-end.

Advantages of FDD:

- RX and TX can operate simultaneously, which is beneficial for low-latency applications.
- There is no need for fast switching between RX and TX, which simplifies the design.
- Relaxed synchronization requirements between RX and TX and different users.

Disadvantages of FDD:

- Duplexers are costly components, which significant insertion loss depending on filtering requirements.
- Requires two separate frequency bands, which can be a limitation in terms of spectrum availability, and MIMO channel estimation.
- The large TX causes severe desensitization of the RX, which requires high linearity and good filtering (50 dB to 60 dB).

3.5.2 Time-Division Duplex (TDD)

In TDD, the RX and TX share the same frequency band but operate at different times. This allows for more efficient use of the available spectrum, as the same frequency can be used for both transmission and reception. TDD is particularly well-suited for applications with asymmetric traffic patterns, where the data rate in one direction is significantly higher than in the other. However, TDD requires precise timing control to avoid interference between RX and TX periods, which can increase complexity. In TDD, a single PLL can be used to generate the LO frequency for both RX and TX, which reduces complexity and power consumption. The RF RX and TX paths are connected to the antenna via a switch, which alternates between connecting the antenna to the RX path and the TX path. A typical TDD TRX block diagram is shown in Figure 22.

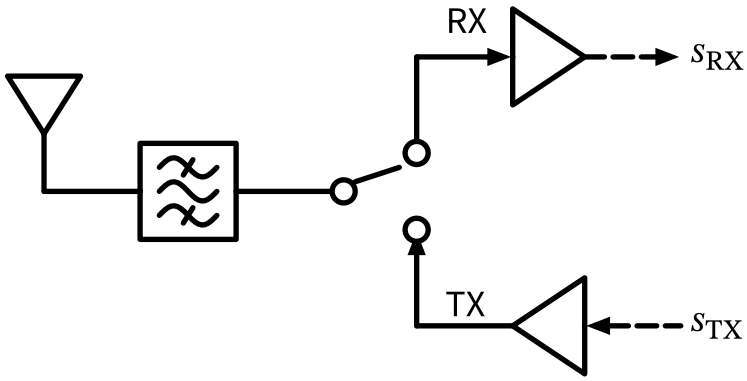

Figure 22: Block diagram of a TDD RF front-end.

Advantages of TDD:

- More efficient use of the available spectrum, as the same frequency can be used for both RX and TX.
- A single PLL can be used for both RX and TX, which reduces complexity and power consumption.
- No duplexer is required (just a single band filter), which reduces cost and insertion loss.
- No RX blocking by own TX, which relaxes linearity and filtering requirements.
- Easier to implement MIMO, as all antennas can operate in the same frequency band.

Disadvantages of TDD:

- RX and TX cannot operate simultaneously, which can be a limitation for low-latency applications.
- Requires precise timing control to avoid interference between RX and TX periods, which can increase complexity.
- Synchronization between RX and TX and different users is required, which can be challenging in some scenarios.

3.5.3 Comparison of FDD and TDD

Below is a summary of important wireless standards and their duplexing method as shown in Table 4:

Table 4: Comparison of duplexing methods used by major wireless standards

Wireless Standard	Duplexing Method	Comments
GSM (2G)	FDD & TDMA	TX and RX operate at different frequencies (FDD) and different times (TDMA)
UMTS (3G)	FDD	Traditional cellular standard using paired spectrum
LTE (4G)	FDD/TDD	FDD is used mostly <2.7 GHz, TDD is used >2.3 GHz
5G NR	FDD/TDD	FDD is used mostly <2.7 GHz, TDD is used >2.3 GHz
WiFi (802.11)	TDD	Unlicensed spectrum operation
Bluetooth	TDD	Short-range personal area network
Zigbee	TDD	Low-power IoT applications

As you can see in Table 4, there is a tendency to use FDD for lower frequencies and large communication distances, while TDD is preferred for higher frequencies and smaller distances.

3.6 Specialty Architectures

In some cases, other architectures may be preferred over the direct-conversion architecture. During the evolution of wireless communication, many different architectures have been proposed and used. However, only a few of them are still relevant today. Some examples are shown next.

3.6.1 Super-Heterodyne Architecture

The super-heterodyne architecture is a widely used approach in radio. It works by mixing the incoming/outgoing RF signal with an LO to produce an intermediate frequency (IF) signal. This IF signal is then amplified and processed, allowing for better selectivity and sensitivity compared to direct-conversion architectures. Super-heterodyne receivers/transmitters are known for their excellent performance in terms of image rejection and dynamic range, making them suitable for a variety of applications, including traditional analog TV and radio broadcasting. As simplified block diagram of a super-heterodyne transceiver is shown in Figure 23.

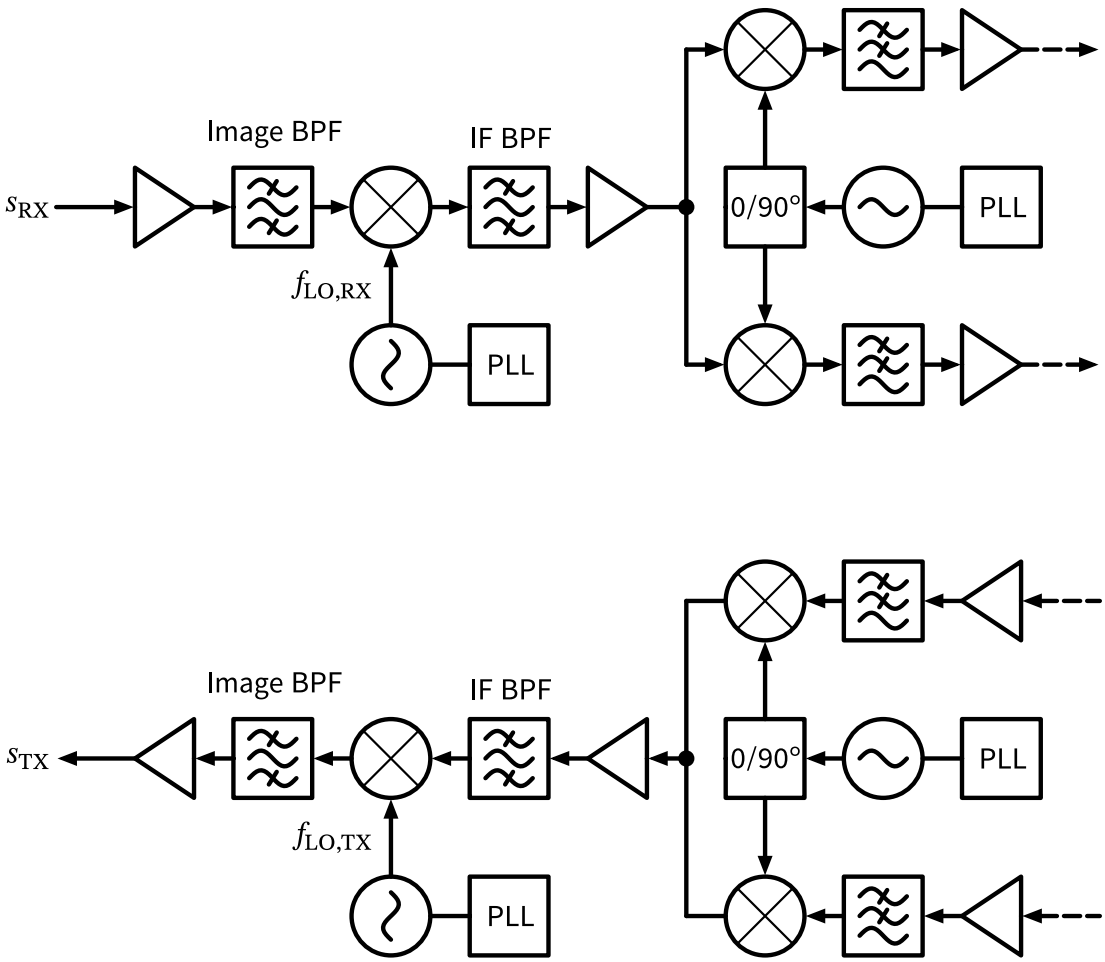

Figure 23: Block diagram of a super-heterodyne transceiver (TRX) showing the main functional blocks of RX and TX.

When you compare Figure 17 with Figure 23, you can immediately appreciate the increased complexity of the super-heterodyne architecture. It requires two PLLs to generate the RX and TX LO frequencies, as well as additional mixers and filters for the IF stage. This increases cost, power consumption, and size. However, the super-heterodyne architecture can provide better performance in terms of selectivity and sensitivity, especially in challenging RF environments with strong blockers, as it allows filtering at RF, IF, and baseband frequencies.

One important aspect of super-heterodyne receivers is the choice of the intermediate frequency (IF). The IF should be high enough to allow for effective filtering and **image rejection**, but low enough to avoid excessive complexity and power consumption. Common IF frequencies range from a few MHz to several hundred MHz, depending on the application and frequency band.

An important issue in super-heterodyne receivers is the **image frequency**. The image frequency is a spurious frequency that can interfere with the desired signal, and is located at $f_{\text{image}} = f_{\text{RF}} \pm 2f_{\text{IF}}$ (the signs depends on the choice of high-side or low-side mixing). To suppress the image frequency, an image-reject filter is either placed before (RX) or after (TX) the mixer. The design of this filter is critical, as it must provide sufficient attenuation of the image frequency while maintaining low insertion loss for the desired signal.

An alternative to image filtering is the use of active image rejection techniques, such as the **Hartley** or **Weaver** architectures. These techniques use additional mixers and phase shifters to cancel out the image frequency, allowing for improved performance without the need for a dedicated image-reject filter.

3.6.2 Low-IF Architecture

To avoid some of the issues of direct-conversion architectures (like dc offsets and flicker noise), a low-IF architecture can be used. In a low-IF architecture, the RX and TX signals are mixed to a low intermediate frequency (typically a few MHz to tens of MHz) instead of directly to baseband. This allows for easier filtering of DC offsets and flicker noise, while still maintaining the benefits of a single LO and reduced complexity compared to super-heterodyne architectures. A low-IF architecture is shown in Figure 24.

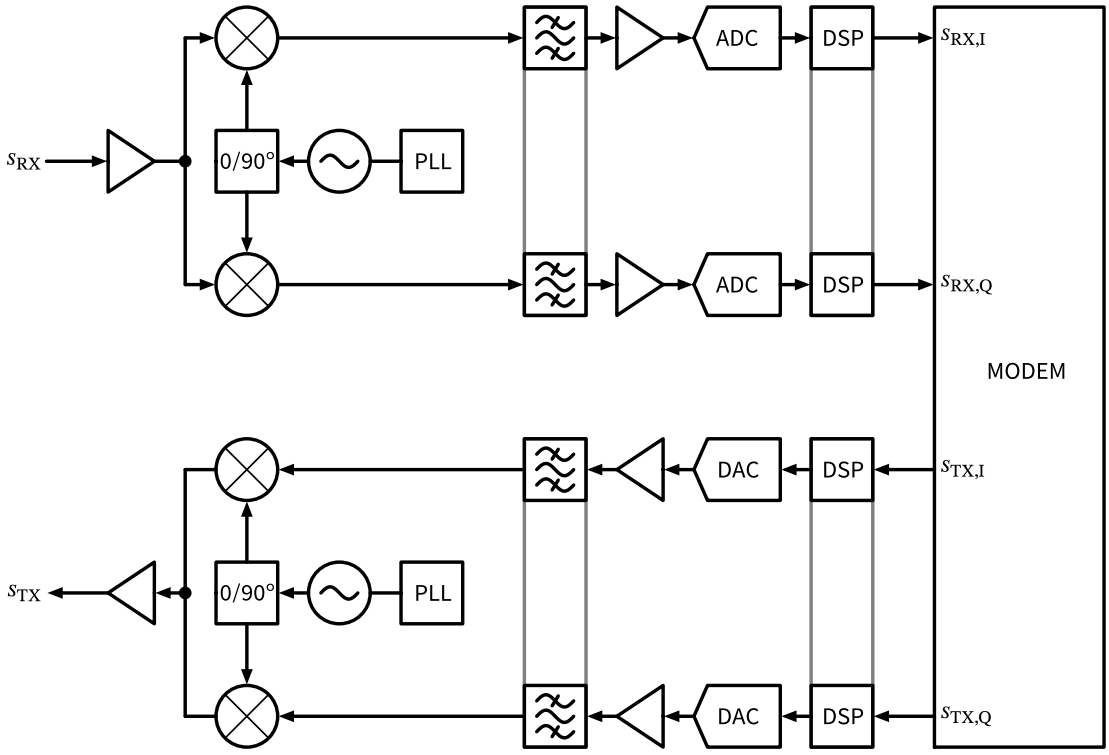

Figure 24: Block diagram of a low-IF transceiver (TRX) showing the main functional blocks of RX and TX. Note the usage of complex analog and digital baseband filters. Otherwise, the structure is similar to a zero-IF TRX as shown in Figure 17.

The low-IF architecture is the defacto standard for Bluetooth receivers. Its advantage compared to direct-conversion vanishes for larger channel bandwidths, this is why it is not used for cellular or WiFi (GSM receivers might be an exception).

One noteworthy disadvantage of low-IF architectures is the required 2xBW compared to direct-conversion. This might cause increased power consumption in the analog baseband filters and ADCs/DACs. Additionally, the low-IF architecture still requires careful design to mitigate issues like IQ imbalance and LO leakage, although these issues are generally less severe than in direct-conversion architectures.

3.6.3 Super Simple Architecture

For some applications with very low cost and low performance requirements, a super simple architecture can be used (think garage door opener). In this architecture, the RX and TX paths are stripped down to the bare minimum. A super simple receiver just uses a bandpass filter and an envelope detector, while a super simple transmitter uses an oscillator and power amplifier. These simplified architectures are shown in Figure 25.

Figure 25: Block diagram of a super simple TX and RX.

Despite the simple architecture, digital amplitude-shift-keying (ASK) or on-off-keying (OOK) can be used. If the receiver is able to discriminate between frequencies (e.g., by using two RF filters with an envelope detector each), also frequency-shift-keying (FSK) can be used.

3.7 I/Q Imbalance

In direct-conversion and low-IF architectures, the I and Q paths are used to process the in-phase and quadrature components of the signal. Ideally, these paths should have identical gain and a 90° phase difference. However, in practice, there are always some mismatches between the I and Q paths, leading to **I/Q imbalance**. This imbalance can cause constellation distortion, leading to increased error vector magnitude (EVM) and degraded system performance.

I/Q imbalance can be characterized by two parameters: gain mismatch (ΔG) and phase mismatch ($\Delta\varphi$). Gain mismatch refers to the difference in gain between the I and Q paths, while phase mismatch refers to the deviation from the ideal 90° phase difference. The impact of I/Q imbalance on system performance depends on the modulation scheme used, with higher-order modulations being more sensitive to these impairments.

There are two ways to quantify I/Q imbalance:

- **Image rejection ratio (IRR)**: The IRR is a measure of how well the receiver can reject the image frequency caused by I/Q imbalance. It is defined as the ratio of the power of the desired signal to the power of the image (unwanted) signal, typically expressed in dB. A higher IRR indicates better performance, with values above 30 dB to 40 dB generally considered acceptable for most applications.
- **Error vector magnitude (EVM)**: The EVM is a measure of the difference between the ideal transmitted signal and the received signal, expressed as a percentage of the signal's magnitude. It quantifies the overall distortion in the received signal, including the effects of I/Q imbalance. Lower EVM values indicate better performance, with typical requirements ranging from 1% to 10% depending on the modulation scheme and application.

The EVM (in rms) is defined as

$$\mathrm{EVM} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} \left|s_{\mathrm{ideal}}(i) - s_{\mathrm{meas}}(i)\right|^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} \left|s_{\mathrm{ideal}}(i)\right|^2}} \tag{23}$$

where $s_{\text{ideal}}(i)$ is the ideal transmitted symbol, $s_{\text{meas}}(i)$ is the measured received symbol, and N is the number of symbols. EVM is expressed either in percent or in dB using

$$\text{EVM}\,|_{\text{dB}} = 20 \cdot \log_{10}(\text{EVM}).$$

In order to make the I/Q mismatch sufficiently small, among the possible techniques are:

- Careful layout and matching of the components in the I and Q paths to minimize gain and phase mismatches. This usually involves good layout techniques. Further, the LO I/Q generation should be done with high accuracy.
- Calibration techniques can be used to measure and compensate for I/Q imbalance. This can be done either in the analog domain (e.g., using variable gain amplifiers and phase shifters) or in the digital domain (e.g., using digital signal processing algorithms). Digital compensation is usually preferred, as it is more flexible and can adapt to changing conditions. A CORDIC can be readily used for this purpose.

4 Low Noise Amplifiers

As shown in Section 2.3.4, the sensitivity of a receiver is determined (besides the channel bandwidth) mainly by the noise figure of the receiver. The noise figure is in turn determined by the noise figure of the first active component in the receive chain, which is usually a low noise amplifier (LNA), as exemplified in Equation 18. Hence, as shown in Figure 21 and Figure 22, low noise amplifiers (LNAs) are the first active building block in a receiver after the antenna and some initial RF filtering.

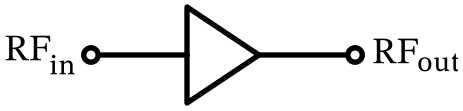

Figure 26: Block diagram of an LNA. Typically, the LNA input is impedance matched to 50 Ω, while the output is often not matched if the LNA is kept on chip. Often, the LNA gain is adjustable to allow for gain control in the receiver depending on the signal conditions. The LNA also might have a low-power bypass mode to reduce the power consumption of the LNA for sufficiently strong signals.

The LNA as a building block is shown in Figure 26. The main purpose of the LNA is to amplify the received signal with as little additional noise as possible. The LNA is usually designed for a specific frequency band, e.g., the 2.4 GHz ISM band or the 5 GHz WLAN band, and is typically designed for a specific impedance, e.g., 50 Ω, which is the standard impedance for RF systems; impedance matching is usually required at the input; the output impedance matching is only required if the output of the LNA goes off chip—if it is kept on chip impedance matching is often not required. The LNA is also designed to be sufficiently linear, i.e., to not introduce significant distortion to the amplified signal, however, compared to the noise requirements, linearity is often less critical.

Why is impedance matching at the input of the LNA so important?

- The antenna is usually designed for a specific impedance, e.g., 50 Ω, and if the LNA input is not matched to this impedance, a significant portion of the received signal will be reflected back to the antenna, resulting in a loss of signal power.

- Filters at the input of the LNA need to be terminated with the correct impedance to achieve the desired filter characteristics.
- Any transmission line in front of the LNA needs to be matched to avoid reflections and standing waves.

To quantify the "quality" of an impedance match, the reflection coefficient Γ is often used, which is defined as [3]:

$$\Gamma = \frac{Z_{\text{in}} - Z_0}{Z_{\text{in}} + Z_0}$$

where Z_0 is the characteristic impedance of the system (usually 50 Ω) and Z_{in} is the input impedance of the LNA. The reflection coefficient Γ is a complex number with a magnitude between 0 and 1, where 0 indicates a perfect match and 1 indicates a complete mismatch. This reflection coefficient can be represented on a Smith chart shown in Figure 27.

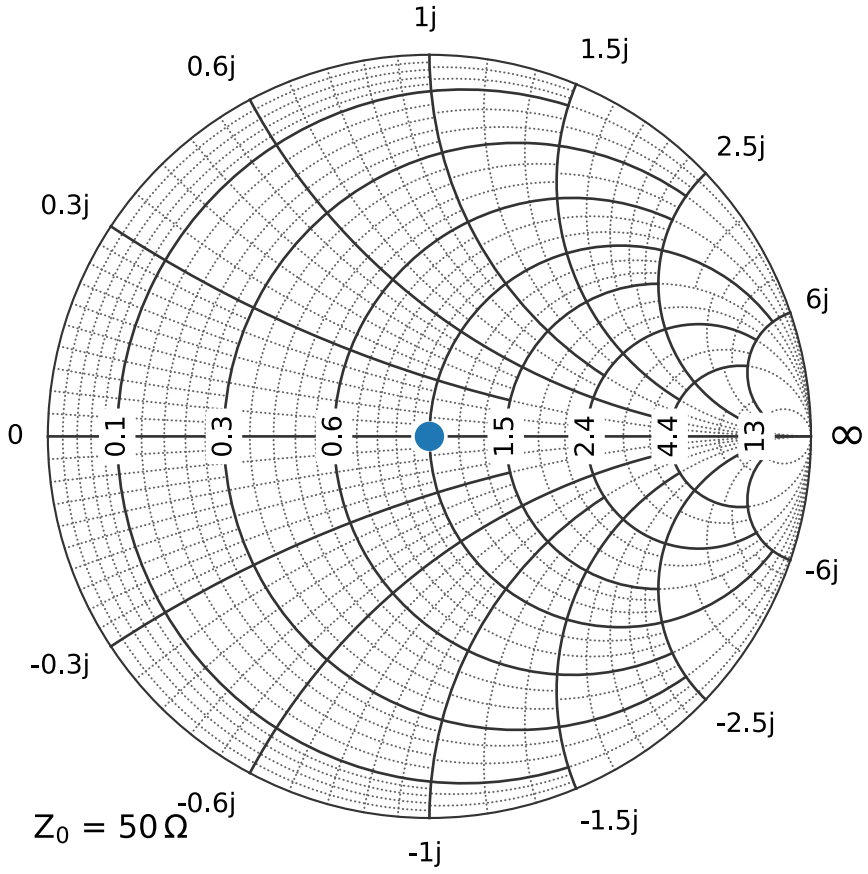

Figure 27: Smith chart showing constant resistance and reactance circles for impedance matching in RF circuits.

However, in practice, the more commonly used metric for impedance matching is the return loss (RL), which is defined as [3]:

$$\text{RL} = -20\log_{10}|\Gamma| = 20\log_{10}\left|\frac{Z_{\text{in}} + Z_0}{Z_{\text{in}} - Z_0}\right|.$$

A higher return loss indicates a better impedance match. A return loss of 10 dB indicates that 10% of the signal power is reflected back, while a return loss of 20 dB indicates that only 1% of the signal power is reflected back. In practice, a return loss of at least 10 dB is desired, with higher (positive!) values being better [9].

4.1 Resistively Matched Common-Source LNA

The key question is now how to design an LNA with low noise figure and an input impedance matched to 50 Ω? In order to appreciate this design challenge, we will first try a naive approach, using a common-source amplifier with resistive termination, as shown in Figure 28.

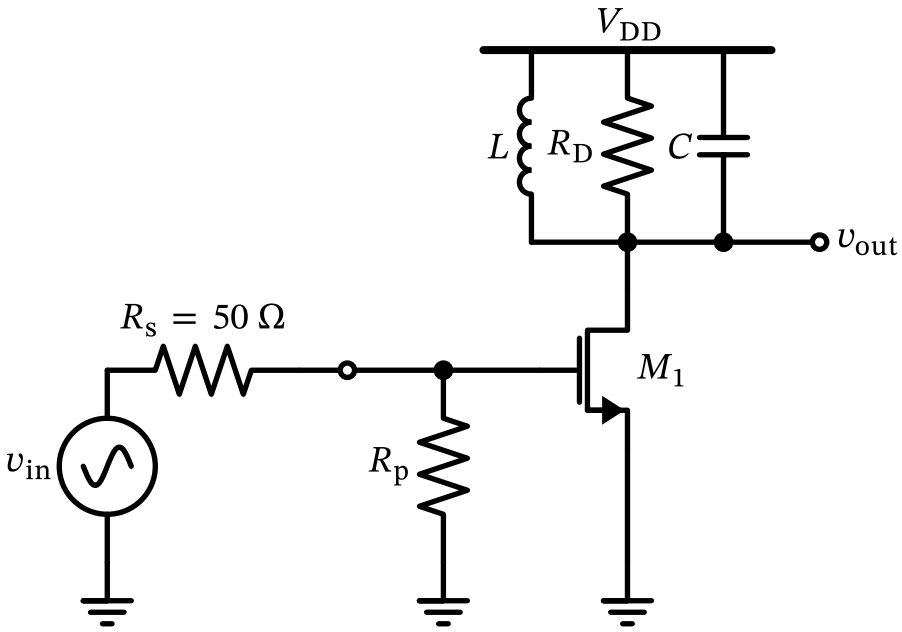

Figure 28: A simple LNA with resistive input matching and a tank circuit as a load (biasing details are omitted). The LNA is driven by a 50 Ω source.

If we assume the gate capacitance of M_1 negligible, we can achieve good input impedance matching by choosing $R_s = R_p = 50\,\Omega$. The voltage gain of this simple common-source LNA is given by $A_v = -g_m R_D$, neglecting capacitances and g_{ds} of M_1 (we assume that the load tank is tuned to the desired frequency with $\omega_0 = 1/\sqrt{LC}$).

How can we calculate the noise figure of this simple LNA? We formulate

$$F = \frac{\text{total noise at output}}{\text{noise at output due to source only}}. \tag{24}$$

We derive a small-signal equivalent circuit of Figure 28, which is shown in Figure 29, to calculate the total noise at the output of the LNA.

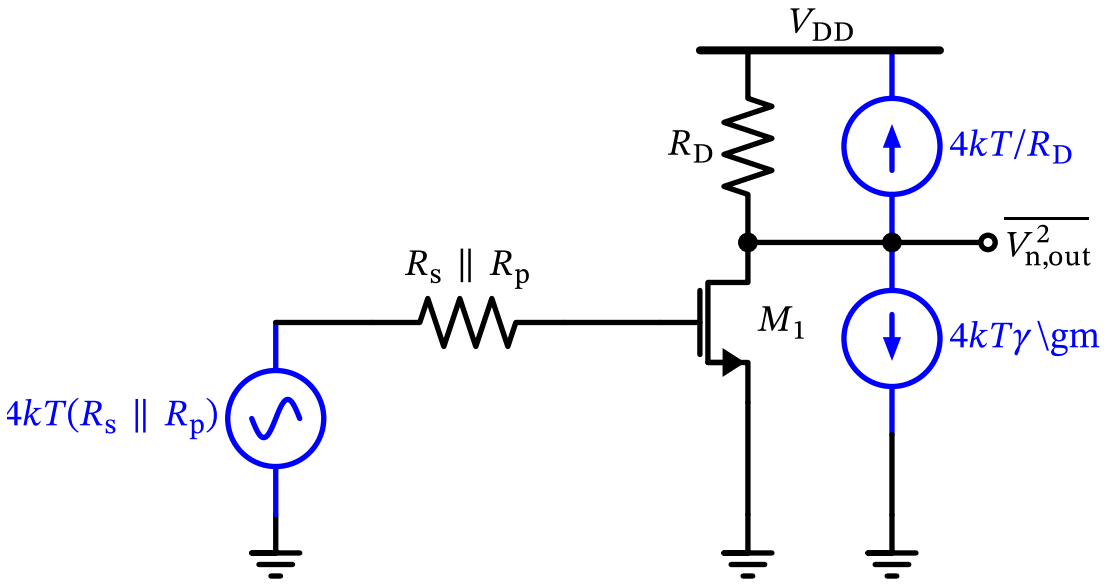

Figure 29: Equivalent circuit of resistively matched common-source LNA.

Withe the help of Figure 29, we can calculate the total output noise power spectral density as

$$\overline{V_{\mathrm{n,out,1}}^2} = A_\mathrm{v}^2 \cdot 4kT\left(R_\mathrm{s} \parallel R_\mathrm{p}\right) = \left(g_\mathrm{m} R_\mathrm{D}\right)^2 \cdot 4kT\left(R_\mathrm{s} \parallel R_\mathrm{p}\right)$$

and

$$\overline{I_{\mathrm{n,out}}^2} = 4kT\gamma g_\mathrm{m} + \frac{4kT}{R_\mathrm{D}}$$

$$\overline{V_{\mathrm{n,out,2}}^2} = R_\mathrm{D}^2 \cdot \overline{I_{\mathrm{n,out}}^2} = 4kT\gamma g_\mathrm{m} R_\mathrm{D}^2 + 4kTR_\mathrm{D}$$

so that in total

$$\overline{V_{\mathrm{n,out}}^2} = \overline{V_{\mathrm{n,out,1}}^2} + \overline{V_{\mathrm{n,out,2}}^2} = 4kT\left[\left(g_\mathrm{m} R_\mathrm{D}\right)^2\left(R_\mathrm{s} \parallel R_\mathrm{p}\right) + \gamma g_\mathrm{m} R_\mathrm{D}^2 + R_\mathrm{D}\right]. \tag{25}$$

We now need to find the output noise coming from the source only. For this we can use the equivalent circuit in Figure 30, to formulate the output noise due to the source only.

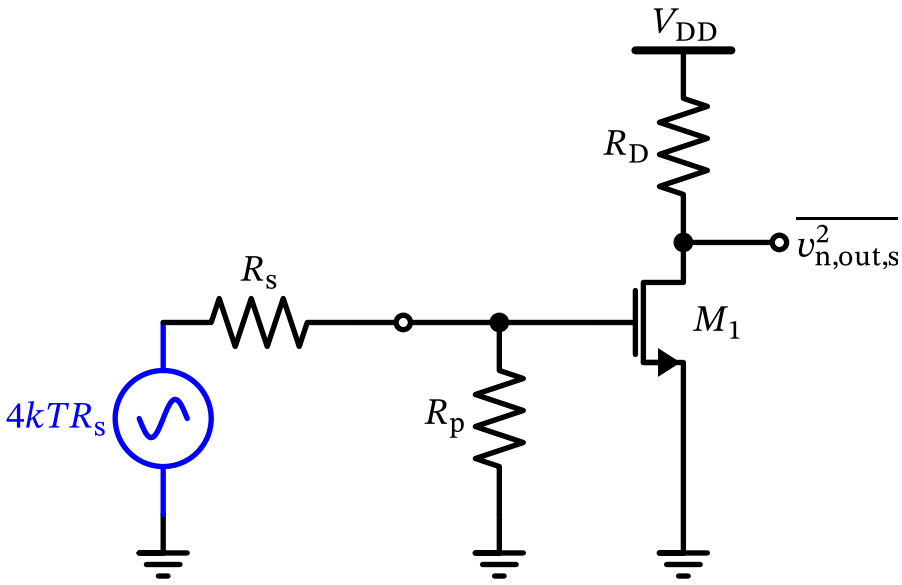

Figure 30: Eqivalent circuit to calculate the output noise from the input.

We find that

$$\overline{V_{n,out,s}^2} = A_v^2 \cdot 4kTR_s \cdot \left(\frac{R_p}{R_s + R_p} \right)^2. \tag{26}$$

Finally, we can use Equation 25 and Equation 26 with Equation 24 to calculate the noise figure of the simple resistively matched common-source LNA as

$$F = \frac{\overline{V_{n,out}^2}}{\overline{V_{n,out,s}^2}} = 1 + \frac{R_s}{R_p} + \frac{\gamma R_s}{g_m (R_s \parallel R_p)^2} + \frac{R_s}{g_m^2 (R_s \parallel R_p)^2 R_D}. \tag{27}$$

i Common-source LNA with Resistive Matching

As an exercise to calculate circuits with noise, re-confirm and derive yourself the result of Equation 27.

How can we interpret Equation 27? We see that we can minimize the noise factor by making g_m large. Then we have a noise factor of

$$F = 1 + \frac{R_s}{R_p} = 2$$

so we see that we are limited to a minimum noise figure of 3 dB, even if we spend the bias current to make g_m very large. We can go below 3 dB noise figure only if we choose $R_p > R_s$, however, this means that the input is no longer matched to 50 Ω, which is usually not acceptable. Hence, this simple resistively matched common-source LNA is not a good choice for a low-noise amplifier, with one exception: For very wideband amplifiers, where a NF of larger then 3 dB is acceptable, this configuration might be a good choice.

We see that we are stuck at high noise figures if we realize the real part of the input impedance with a resistor. This leaves us with the question on how to realize a real part of the input impedance then? We will answer this question in the next section.

4.2 Common-Gate LNA

We remember from our analog circuit design lecture that the common-gate configuration has an input impedance of $1/g_\mathrm{m}$, neglecting parasitic capacitances. Hence, if we choose $g_\mathrm{m} = 1/50\,\Omega = 20\,\mathrm{mS}$, we can achieve input matching to 50 Ω without using a resistor at the input. This is the key idea of the common-gate LNA, which is shown in Figure 31.

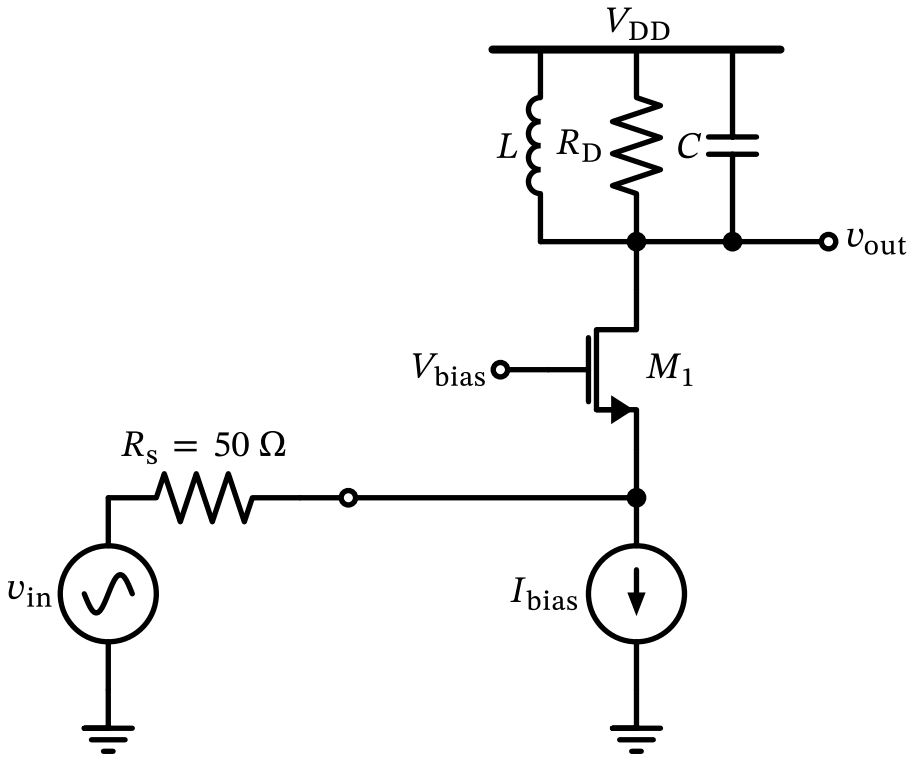

Figure 31: Circuit diagram of a common-gate LNA.

By inspecting Figure 31 and the practice from Section 4.1, we can directly write down the output noise voltage as (with $1/g_\mathrm{m} = R_\mathrm{s}$)

$$\overline{V_\mathrm{n,out}^2} = kT\left[(g_\mathrm{m}R_\mathrm{D})^2 R_\mathrm{s} + \gamma g_\mathrm{m} R_\mathrm{D}^2 + 4R_\mathrm{D}\right] = kT\left(\frac{R_\mathrm{D}^2}{R_\mathrm{s}} + \gamma\frac{R_\mathrm{D}^2}{R_\mathrm{s}} + 4R_\mathrm{D}\right). \tag{28}$$

The output noise due to the source only is given by

$$\overline{V_\mathrm{n,out,s}^2} = \frac{R_\mathrm{D}^2}{R_\mathrm{s}}. \tag{29}$$

Finally, we can use Equation 28 and Equation 29 with Equation 24 to calculate the noise figure of the common-gate LNA as

$$F = 1 + \gamma + \frac{4R_\mathrm{s}}{R_\mathrm{D}} \overset{R_\mathrm{D} \gg R_\mathrm{s}}{\rightarrow} F = 1 + \gamma. \tag{30}$$

With a classical long-channel $\gamma = 2/3$, we can achieve a minimum noise figure of 2.2 dB, which is already better than the resistively matched common-source LNA. However, with modern short-channel devices, γ is often larger than 1, so that the minimum noise figure of the common-gate LNA is often larger than 3 dB [1].

4.3 Inductively-Degenerated Common-Source LNA

As we have seen in Section 4.2, using circuit techniques can realize a real part of an input impedance without the associated thermal noise of a resistor. We now try something different, in the hope that it will result in an even lower noise figure. We construct an LNA based on a common-source MOSFET amplifier, but we add an impedance Z_{deg} into the source line. This arrangement is shown in Figure 32.

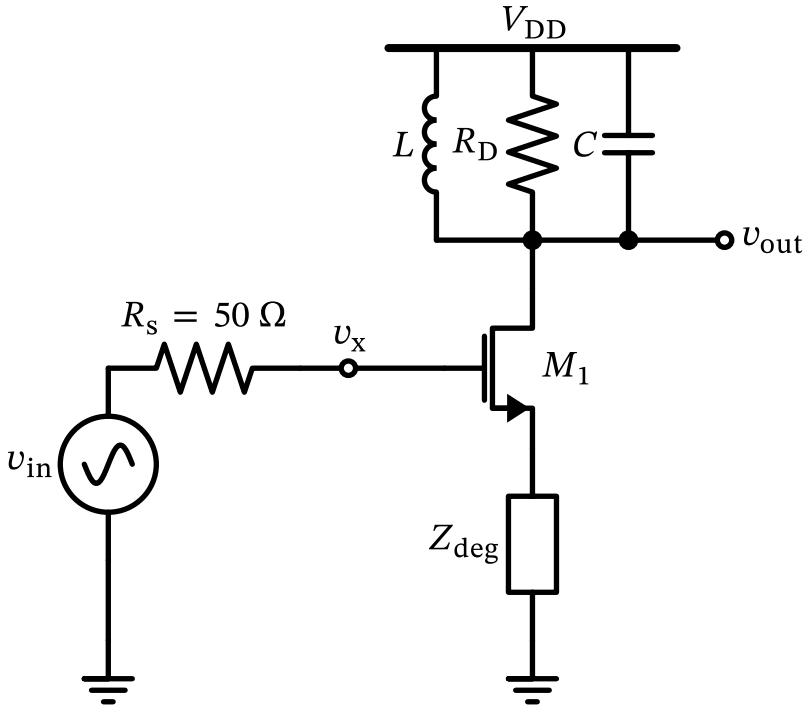

Figure 32: A common-source MOSFET stage with degeneration impedance.

We now extract the small-signal equivalent circuit of Figure 32, which is shown in Figure 33, to calculate the input impedance.

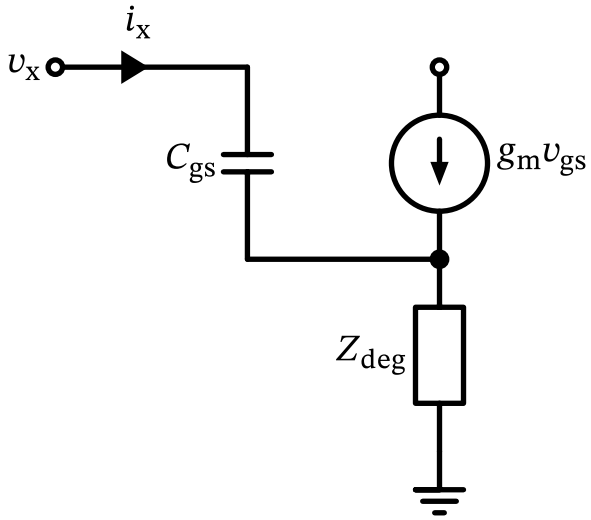

Figure 33: Equivalent small-signal circuit of the input stage around M_1.

We find that

$$v_x = v_{gs} + Z_{deg}(i_x + g_m v_{gs}), \quad v_{gs} = \frac{i_x}{sC_{gs}}$$

so that we can write the input impedance as

$$Z_{in} = \frac{v_x}{i_x} = \frac{1}{sC_{gs}} + Z_{deg} + \frac{g_m Z_{deg}}{sC_{gs}}. \tag{31}$$

The final term in Equation 31 is the interesting one: By choosing Z_{deg} to be inductive (which we can do by either use an on-chip or off-chip inductor), we can realize a real part of the input impedance. If we choose $Z_{deg} = sL$, we find that

$$Z_{in} = \frac{1}{sC_{gs}} + sL + \frac{g_m L}{C_{gs}}.$$

By proper choice of L and C_{gs}, we can achieve input matching to 50 Ω at the desired frequency ω_0. We find that the real part of the input impedance is given by

$$\Re\{Z_{in}\} = \frac{g_m L}{C_{gs}}$$

Without proof (refer to [1] or [2] for a derivation) we find for the noise factor of this input stage (with some simplification) as

$$F = 1 + \frac{\gamma R_s \omega_0^2 C_{gs}^2}{g_m}. \tag{32}$$

Finally we have an LNA input stage configuration which allows us to achieve a noise figure below 3 dB, even with $\gamma > 1$, by proper choice of g_m. Making g_m large (by spending more bias current) results in (first order) arbitrarily low noise figure. The inductively-degenerated common-source LNA is a widely used LNA input stage configuration in modern RFICs. A somewhat detailed schematic is shown in Figure 34.

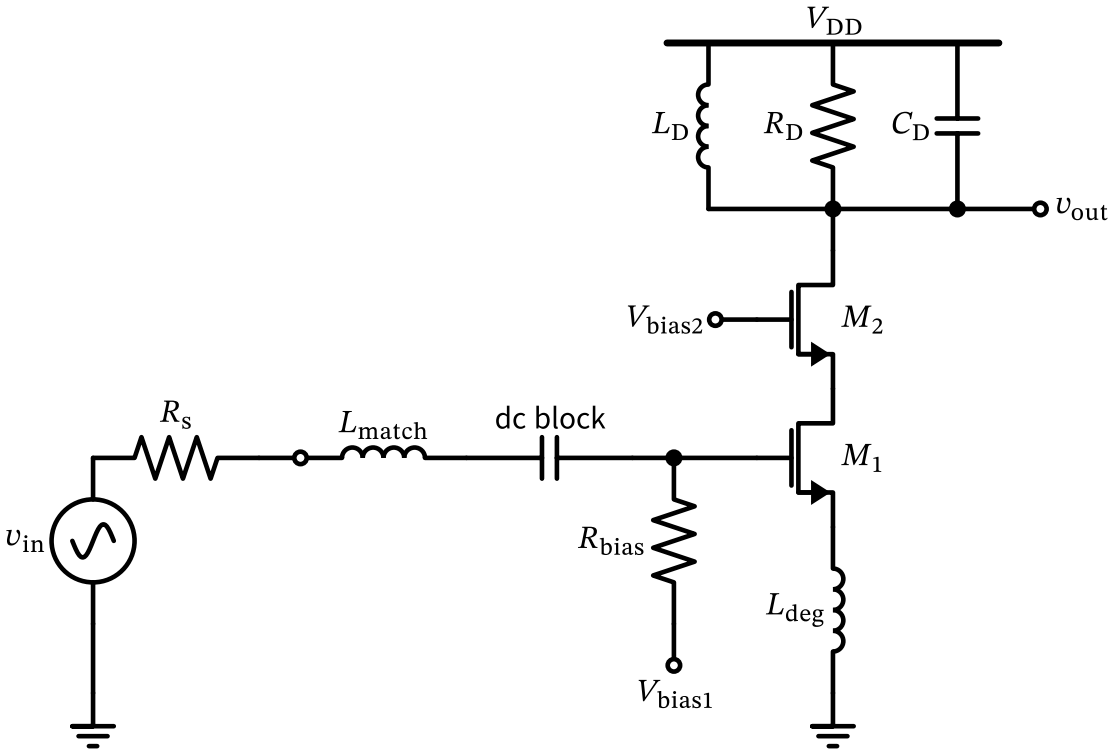

Figure 34: An (almost complete) common-source MOSFET stage with degeneration impedance and cascode.

The inductor L_{match} is used to match the input impedance to 50 Ω at the desired frequency, L_{deg} is used to realize the real part of the input impedance, R_{bias} is used to set the bias current of M_1, M_2 is a cascode transistor which increases the output impedance and thus the gain of the stage (plus it improves the reverse isolation), and R_{D}, L_{D}, and C_{D} form a load tank which provides high gain at the desired frequency. A dc block is used at the input so that the bias point of M_1 is not corrupted by the input signal source. The bias voltage V_{bias2} sets the operating point of the cascode transistor M_2.

What is missing in Figure 34 is any form of frequency tuning of the load to the frequency of interest, and the support of different gain modes. Apart from these details this LNA circuit is a good starting point for a practical LNA design.

4.4 Feedback LNA

One drawback of the inductively-degenerated common-source LNA is the usage of at least one inductor. If the inductor is placed on chip it has a comparatively large size, and if it is implemented in the package (via a bondwiew) or on the PCB it add to the bill-of-materials (BOM) cost.

If the CMOS technology is sufficiently fast, then a shunt feedback LNA, as shown in Figure 35, might be a good choice.

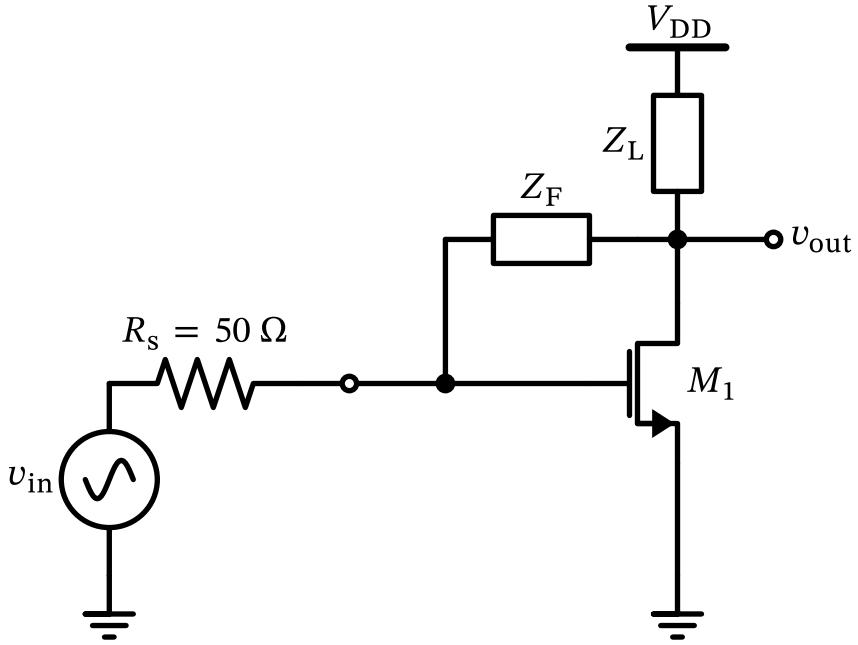

Figure 35: A shunt-feedback LNA.

Without proof, the input impedance of the shunt feedback LNA is given by

$$Z_{in} = \frac{Z_F + Z_L}{1 + g_m Z_L}. \tag{33}$$

The noise factor of the shunt feedback LNA is given by

$$F = 1 + \left| \frac{Z_F + R_s}{g_m Z_F + 1} \right|^2 \cdot \frac{\gamma g_m + \Re\{Y_L\}}{\Re\{Z_{in}\}} \tag{34}$$

As you can see from Equation 34, by making g_m large (by spending more bias current) the noise figure can be made arbitrarily small! Depending on the choice of Z_F and Z_L, the input impedance of this LNA can be changed in interesting ways.

By setting $Z_L \to \infty$ (e.g., by biasing with a current source and high-impedance loading), we find that

$$Z_{in} = \frac{1}{g_m}$$

which is independent of Z_F and is a well-known result for a common-source stage. The disadavntage of this configuration is the noise factor, which (given that Z_F is sufficiently large) trends to $F = 1 + \gamma$, which is the same as for the common-gate LNA.

A bit more interesting is the case when $g_m Z_L = A_0$ and $Z_L \gg Z_F$, which results in

$$Z_{in} = \frac{Z_L}{1 + A_0}$$

which is the well-known result that the input impedance of an amplifier with feedback is reduced by the factor $1 + A_0$, where A_0 is the open-loop gain of the amplifier. The noise factor can be made small by making g_m large, as we have already noted above.

A very interesting case can be achieved by choosing $Z_\mathrm{L} = 1/sC_\mathrm{L}$ and $Z_\mathrm{F} = 1/sC_\mathrm{F}$, which results in

$$Y_\mathrm{in} = \frac{1}{Z_\mathrm{in}} = \frac{g_\mathrm{m} C_\mathrm{F}}{C_\mathrm{L} + C_\mathrm{F}} + s\frac{C_\mathrm{L} C_\mathrm{F}}{C_\mathrm{L} + C_\mathrm{F}} \tag{35}$$

Looking at Equation 35, we see that the input admittance has a **real part**! By proper choice of components, we can achieve an input impedance matched to 50 Ω at the desired frequency. The noise factor can again be made small by making g_m large.

There is also an option, again by proper choice of Z_F and Z_L, to achieve an inductive input impedance component, which can be used to resonate out the input capacitance of the LNA, similar to the inductively-degenerated common-source LNA. However, in contrast to the inductively-degenerated common-source LNA, no inductor is required in this case. This configuration is called a reactance-cancelling LNA.

5 Mixers

6 Oscillators

7 Phase-Locked Loops

8 Power Amplifiers

Bibliography

[1] B. Razavi, *RF Microelectronics*, 2nd edition. Pearson, 2011.

[2] H. Darabi, *Radio Frequency Integrated Circuits and Systems*, 2nd edition. Cambridge University Press, 2020.

[3] D. M. Pozar, *Microwave Engineering*. Wiley, 2011.

[4] P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, Fifth. Wiley, 2009.

[5] B. Razavi, *Design of Analog CMOS Integrated Circuits*. McGraw-Hill, 2017.

[6] R. Sarpeshkar, T. Delbruck, and C. Mead, "White noise in MOS transistors and resistors," *IEEE Circuits and Devices Magazine*, vol. 9, no. 6, pp. 23–29, 1993, doi: 10.1109/101.261888.

[7] B. Sklar and F. J. Harris, *Digital Communications: Fundamentals and Applications*, 3rd edition. Pearson, 2020.

[8] A. F. Molisch, *Wireless Communications: From Fundamentals to Beyond 5G*, 3rd edition. Wiley-IEEE Press, 2022.

[9] T. S. Bird, “Definition and Misuse of Return Loss [Report of the Transactions Editor-in-Chief],” *IEEE Antennas and Propagation Magazine*, vol. 51, no. 2, pp. 166–167, 2009, doi: 10.1109/map.2009.5162049.