Chapter II.2

RF engineering

Christine Vollinger and Manfred Wendt
CERN, Geneva, Switzerland

Radio-frequency (RF) engineering is a subset of electrical engineering, here focused on the application of resonant accelerating structures (cavities), transmission-lines, and a variety of RF components, based on electromagnetic field principles, operating in the frequency range between a few MHz up to tens of GHz.

The JUAS *RF engineering* lecture, and this summary of the lecture, tries to close the gap between theoretical electromagnetism and radio-frequency (RF) theory on one side and practical RF engineering problems on the other. Among the broad range of RF engineering topics only a limited choice of items could be covered, which include some general aspects on RF engineering discussed in the introduction and the most essential topics related to radio-frequency technologies used in particle accelerators outlined in the following sections.

Among the various RF topics presented here, a good understanding of a single cylindrical, so-called "pill-box" cavity resonator is of particular importance, as it is the foundation for charged particle acceleration using radio-frequencies. This write-up of the RF engineering lectures is also intended to help the interested student to prepare for the RF exam with a summary of the most important equations and relationships.

II.2.1 Introduction

In this lecture, the basic concepts of RF engineering are discussed as they apply in particle accelerators of today. Since the field of RF is large, it will only be possible to scratch the surface, and we will give our own selection of topics. Also, we rely on the explanations of basic concepts of electromagnetism (see Chapter I.1) as well as the general introduction into RF (see Chapter II.1) which are presented in greater detail elsewhere in these proceedings, therefore reading and understanding of these two chapters are a prerequisite for this RF chapter. The aim of this chapter is to introduce basic RF concepts in the framework of the JUAS series to an audience of Master students and PhD students coming from different fields of studies and with different backgrounds. While we may repeat a few topics for better understanding, and while there also may be some overlap on topics presented in Chapter II.5 on superconducting RF cavities, we try to give a valuable, condensed introduction to RF engineering used in the field of particle accelerator technology. Evidently, this RF engineering section is not comprehensive and is not indented to replace any text books or special literature, instead, it focuses on the introduction of RF principles

This chapter should be cited as: RF engineering, C. Vollinger and M. Wendt, DOI: 10.23730/CYRSP-2024-003.765, in: Proceedings of the Joint Universities Accelerator School (JUAS): Courses and exercises, E. Métral (ed.),
CERN Yellow Reports: School Proceedings, CERN-2024-003, DOI: 10.23730/CYRSP-2024-003, p. 765.

relevant for the design of RF systems in particle accelerators. For more detailed reading, we will refer to our bibliography.

Before we dip into the radio frequency engineering details, it is worth to look into a table of the frequency ranges used in RF-engineering, see Table II.2.1. The onset of what is called **RF** starts around 3 kHz in the range of extremely low frequency – however, this is not what most people consider RF when thinking about particle accelerators. Here, the frequency range starts with what is commonly called "high frequency" at about 1 MHz.

Let's recall from very basics that *Hertz* (Hz) is the unit of frequency: $1\,\text{Hz} = 1\,\text{sec}^{-1}$. Equally, remember that 1 MHz = 10^6 Hz, and 1 GHz = 10^9 Hz, and that frequency f and wavelength λ are inversely proportional: **large frequency means small wavelength.**

In vacuum, f and λ are coupled by the speed of light, so $f \cdot \lambda = c_0$. Finally, some numbers you should know by heart: A signal with a frequency of 1 GHz has a wavelength of 30 cm, whereas a frequency of 3 MHz corresponds to about 100 m.

Table II.2.1: Example frequency ranges in RF engineering. Numbers taken from Wikipedia.

RF band name	**Abbreviation**	ITU band number	Frequency range
High frequency	HF	7	3–30 MHz
Very high frequency	VHF	8	30–300 MHz
Ultra high frequency	UHF	9	300–3000 MHz
Super high frequency	SHF	10	3–30 GHz
Extremely high frequency	EHF	11	30–300 GHz

The RF-system in an accelerator is one of the most central elements as it is the part which does the actual acceleration. Most pronouncedly visible is the cavity where the acceleration takes place, however, other beam operations like beam chopping, bunch compression etc. are equally carried out with the help of the electromagnetic fields of RF cavities. It will not be possible to explain an entire RF system in two lectures, therefore, we will explain the topic of RF transmission lines in section II.2.3, and concentrate in section II.2.4 on the simplest accelerating structure, the so-called pill-box cavity. It is important to un-

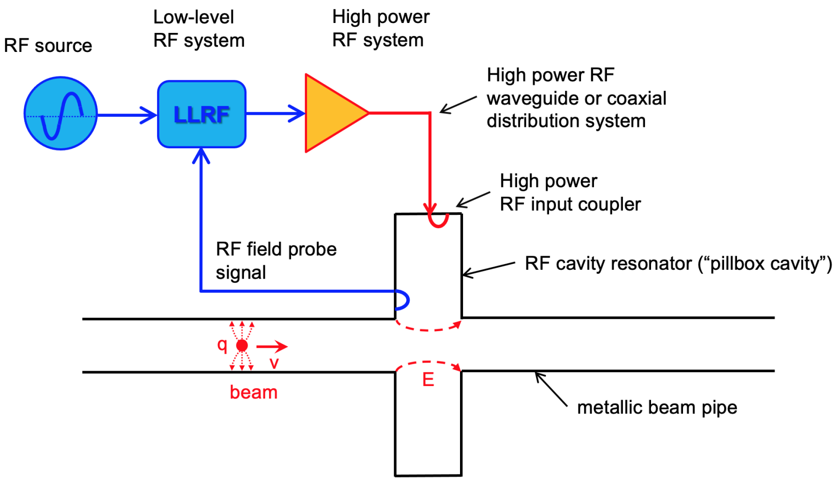


Fig. II.2.1: Schematic of accelerating single gap cavity with external components.

derstand that cavity design is an engineering branch in its own and a cavity that is used for acceleration is not a self-sufficient system. As can be seen in Fig. II.2.1, the resonant structure of a single gap cavity is hooked up to a high power RF system, consisting of a complex amplifier structure which is connected to the cavity via a feeder line. The feeder line is used for transmitting the amplifier power to the cavity. It typically consists of a waveguide or a coaxial line which terminates in a high power RF coupler to bring in the power that is needed to excite the cavity resonance. In addition, an RF-source and a low-level RF system are required for cavity operation, e.g. to provide the RF power with the correct frequency, amplitude and phase. Note that the cavity is operated under vacuum and the beam passage takes place by connecting the cavity body to the metallic vacuum beam pipe. The accelerating force is provided by the electric field inside the cavity.

II.2.2 RF Engineering Terms: deci-Bel

II.2.2.1 Decibel, dB, ("dee-bee"), or not to be . . .

Decibels are used to express large number ranges by using the base 10 logarithm of numbers. Hence, decibels are very handy to cover several orders of magnitude, for example power from mW to MW…

In this context, the *Bel* is a logarithmic unit expressing ratios between values, and particular popular is the tenth fraction, *deci-Bel*:

$$1\,\mathrm{dB} = \frac{1}{10}\mathrm{B} = 0.1\,\mathrm{B}\ \text{(Bel)}$$

While used in many engineering disciplines, the dB used in electrical and RF engineering usually expresses ratios between two electrical *power* values

$$P_{\mathrm{dB}} = 10\log_{10}\left(\frac{P_1}{P_2}\right)$$

Table II.2.2: Some important values of dB ratios.

dB ratio	P_1/P_2	V_1/V_2
$n\times$ dB	10^n	$10^{n/2}$
40 dB	10000	100
20 dB	100	10
10 dB	10	∼3.16
6 dB	∼4	∼2
3 dB	∼2	∼1.41
0 dB	**1**	**1**
−3 dB	∼0.5	∼0.71
−20 dB	0.01	0.1

or between two voltages or currents:

$$V_{\text{dB}} = 20\log_{10}\left(\frac{V_1}{V_2}\right) \qquad I_{\text{dB}} = 20\log_{10}\left(\frac{I_1}{I_2}\right)$$

with:

$$\frac{P_1}{P_2} = 10^{\left(\frac{P_{\text{dB}}}{10}\right)} \qquad \frac{V_1}{V_2} = 10^{\left(\frac{V_{\text{dB}}}{20}\right)} \qquad \frac{I_1}{I_2} = 10^{\left(\frac{I_{\text{dB}}}{20}\right)}$$

Please note, the 3 dB ratio (resulting from the half power definition) is common specification for the bandwidth!

II.2.2.2 "dB" is not "dBm"

dBm is defined as a logarithmic power unit, based on dB (deci-Bel) and a reference power of $P_{\text{ref}} = 1\,\text{mW}$

$$P_{\text{dBm}} = 10\log_{10}\left(\frac{P}{P_{\text{ref}}}\right), \qquad \text{with: } P = P_{\text{ref}}10^{\left(\frac{P_{\text{dBm}}}{10}\right)}$$

dBm may also be used as logarithmic voltage unit, e.g. for the popular $Z_0 = 50\,\Omega$ impedance we calculate $V_{\text{ref}} = \sqrt{Z_0 P_{\text{ref}}} = \sqrt{0.05}\,\text{V} \approxeq 0.2236\,\text{V}$ (RMS).

$$V_{\text{dBm}} = 20\log_{10}\left(\frac{V}{V_{\text{ref}}}\right), \qquad \text{with: } V = V_{\text{ref}}10^{\left(\frac{V_{\text{dBm}}}{20}\right)}$$

The use of dBm as logarithmic voltage (or current) unit **strictly requires** the waveform to be **sinusoidal!**

Table II.2.3: Some important values for dBm values

dBm	***P***	***V*** (RMS)
90 dBm	1 MW	7.07 kV
60 dBm	1 kW	223.6 V
30 dBm	1 W	7.07 V
20 dBm	100 mW	2.24 V
10 dBm	10 mW	707 mV
6 dBm	4 mW	446 mV
0 dBm	**1.0 mW**	**224 mV**
−20 dBm	10 µW	22.4 mV
−60 dBm	1 nW	224 µV
−120 dBm	1 fW	224 nV
−174 dBm	4×10^{-21} W	0.446 nV

Please note, $-174\,\mathrm{dBm}$ is the equivalent noise power in a bandwidth $BW = 1$ Hz at room temperature.

II.2.3 Transmission lines

So far, you have learned about Maxwell's equations which describe the macroscopic electromagnetic theory. The lectures given by Henke (see Chapter I.1) and Mostacci (see Chapter II.1), explain about the propagation of waves in unbounded media. The possibility of using guided wave propagation by means of transmission lines such as coaxial lines or waveguides opened a new era and the use of waveguides was an enormous milestone in RF-engineering [1].

Coaxial lines and hollow wave guides are fully described by electromagnetic field theory. This kind of electromagnetic analysis is heavy, and it can be observed that today, a considerable number of RF engineers design components without any direct use of the underlying electromagnetic analysis. Simulation software tools are used instead, even for advanced calculation and sophisticated measurement tools, like spectrum and network analysers, for the experimental testing of the obtained design.

In this section, general transmission line theory will be introduced and will be used to understand the use of waveguides and coaxial lines. Transmission line theory bridges the gap between classical electromagnetic field analysis (Maxwell's theory) and basic circuit theory of lumped elements. Circuit theory assumes that the dimensions of the elements (like resistors or simple cable lengths) are much smaller than the wavelength of the signal. This means voltage or current do not change along the lumped element dimension.

Transmission lines, however, cover a fraction of the wavelength which means that the associated voltages and currents vary in magnitude and phase *along* the element. Such elements can be expressed by distributed circuits to allow a wave description along geometrical dimensions of the transmission line. Consequently, transmission lines are modelled *piecewise by a lumped-element circuit*. Consider a transmission line of a certain length aligned in the z-direction as is shown in Fig. II.2.2. Voltage and current along the line depend on the coordinate z and their variation along the line depends on their wavelength. This variation increases with increasing frequency since high frequency signals have short wavelengths. The line is now split into infinitesimal increments of length dz, and each length dz is modeled by means of lumped elements. The entire line can then be seen as a series of lumped element circuits as is shown in Fig. II.2.3 (left).

The lumped elements of each line length dz are expressed per unit length as follows:

$$\text{series resistor } R' = \tfrac{R}{dz} \text{ in } [\Omega/\text{m}], \qquad \text{series inductor } L' = \tfrac{L}{dz} \text{ in } [\text{H/m}]$$

$$\text{parallel conductance } G' = \tfrac{G}{dz} \text{ in } [\text{S/m}], \qquad \text{parallel capacitance } C' = \tfrac{C}{dz} \text{ in } [\text{C/m}].$$

The goal is now to derive wave-describing equations from the equivalent circuit model with lumped elements which so far only describe a line of length dz.

By applying *Kirchhoff's laws* to derive the voltages and currents in our equivalent circuit, as is

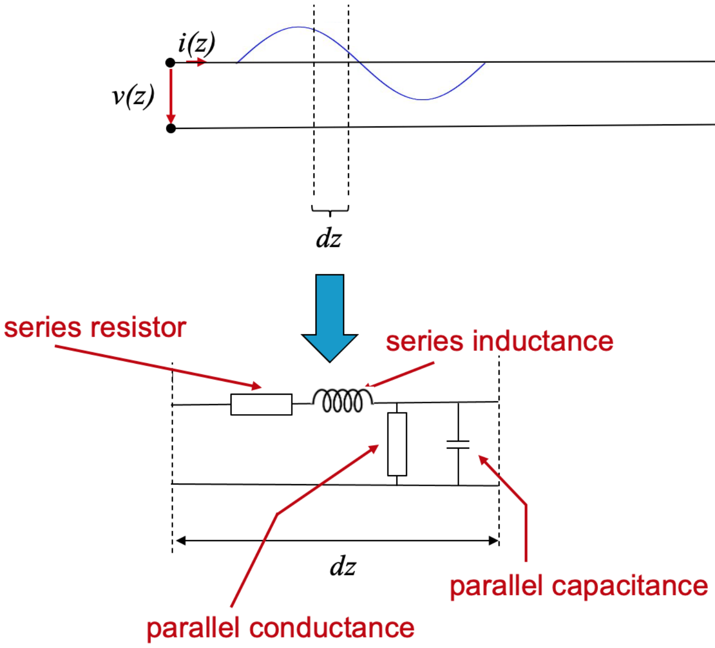


Fig. II.2.2: Transmission line split in infinitesimal small increments of dz, and equivalent circuit representation by lumped elements.

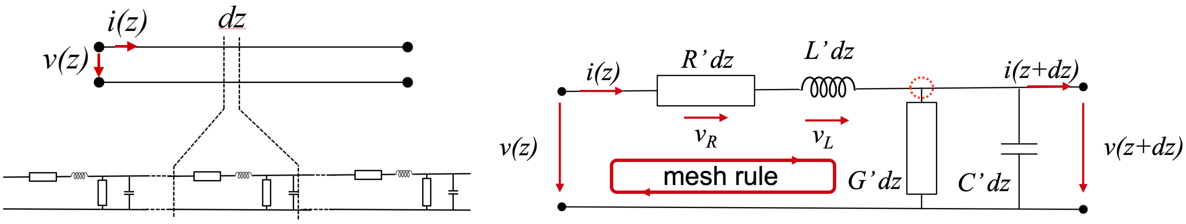


Fig. II.2.3: Transmission line increment with voltage and current distribution as equivalent circuit with lumped elements.

illustrated in Fig. II.2.3, we get from the *mesh rule*

$$
\begin{aligned}
v(z) &= v_{\mathrm{R}} + v_{\mathrm{L}} + v(z+dz) \\
v(z) &= R'\,dz\,i(z) + j\omega L'\,dz\,i(z) + v(z+dz)
\end{aligned}
$$

$$
\underbrace{v(z+dz) - v(z)}_{\frac{dv}{dz}\,dz} = -R'\,dz\,i(z) - j\omega L'\,dz\,i(z).
$$

Hence an equation for the voltage of our equivalent circuit

$$
\frac{dv}{dz} = -(R' + j\omega L')\,i(z) \quad . \tag{II.2.1}
$$

Equally, the *junction rule* gives an equation for the current within our equivalent circuit, as is illustrated in Fig. II.2.3 by evaluating the sum of the currents in the node (see red dashed circle)

$$\begin{aligned} i(z) &= i_{\mathrm{C}} + i_{\mathrm{G}} + i(z+dz) \\ i(z) &= j\omega C'\,dz\,v(z+dz) + G'\,dz\,v(z+dz) + i(z+dz) \quad , \end{aligned}$$

from which we get an equation for the current within the equivalent circuit

$$\underbrace{i(z+dz) - i(z)}_{\frac{di}{dz}\,dz} = -j\omega C' dz\, v(z+dz) - G' dz\, v(z+dz)$$

$$\frac{di}{dz} = -(G' + j\omega C')\,v(z) \quad . \tag{II.2.2}$$

The two differential equations (II.2.1) and (II.2.2) are the so-called *telegrapher equations*, expressed for harmonic time dependence here, i.e. the time-derivative is indicated as $j\omega$, as already explained in greater detail in Chapter I.1. By means of these two equations, it is possible to describe wave propagation for voltage v and current i on the transmission line by taking the 2nd derivative of the terms, and combining the two. Mathematically, one obtains the one-dimensional, scalar Helmholtz' equation for EM-fields:

$$\frac{d^2 v}{dz^2} = (R' + j\omega L')(G' + j\omega C')\,v(z) \tag{II.2.3}$$

$$\frac{d^2 i}{dz^2} = \underbrace{(R' + j\omega L')(G' + j\omega C')}_{\gamma^2}\,i(z) \tag{II.2.4}$$

It is now possible to define a propagation constant γ and a characteristic line impedance Z_c for the transmission line length dz by using the lumped elements of Eqs. (II.2.3) or (II.2.4):

$$\gamma = \sqrt{(R' + j\omega L')(G' + j\omega C')} \quad \text{and} \quad Z_c = \sqrt{\frac{R' + j\omega L'}{G' + j\omega C'}}.$$

With the complex propagation constant and the characteristic line impedance, we have described the general case for all two-conductor lines in TEM-mode. Note that if you compare transmission line theory with general EM-field theory, the complex propagation constant is mathematically derived as the wave number. Unfortunately, the notation is not unique, in particular the signs of γ and the wave number can differ. It is therefore important to check the definitions used in the respective literature. The propagation constant γ can be split into two parts, where the real part α is denoted *attenuation constant* and describes the attenuation of the line, whereas β denotes the *phase constant*

$$\gamma = \alpha + j\beta.$$

The notation used here leads to the following expressions for the attenuation and phase constants

$$\alpha = \sqrt{\frac{1}{2}\left[\sqrt{(R'^2 + \omega^2 L'^2)(G'^2 + \omega^2 C'^2)} - (\omega^2 L'C' - R'G')\right]}$$

$$\beta = \sqrt{\frac{1}{2}\left[\sqrt{(R'^2+\omega^2 L'^2)(G'^2+\omega^2 C'^2)}+(\omega^2 L'C'-R'G')\right]}$$

The definition of the general case for all two conductor lines allows one to consider three typical cases: the lossless line, the small loss case, and the high loss case. For the case of a lossless line, $\alpha = 0$ and the terms simplify greatly. The small loss case is the most common case and it is defined by the reactive elements dominating the transmission line description over the lossy elements, i.e. $\omega L' \gg R'$ and $\omega C' \gg G'$ - this feature allows the line to be used as an RF-element!

Finally, in the high-loss case, wave propagation on the line is suppressed. The table below summarises these cases (see [2], [3] for details and further reading).

Table II.2.4: Summary of different cases grouped by losses for a two-conductor line.

Lossless line	**Small loss case** (high frequency)	**High loss case** (low frequency)
$R' = 0 \quad G' = 0$ $\alpha = 0 \quad \beta = \omega\sqrt{L'C'}$	$\omega L' \gg R' \quad \omega C' \gg G'$ $\alpha \approx \frac{1}{2}\frac{R'}{\sqrt{L'/C'}} + \frac{G'\sqrt{L'/C'}}{2}$ $\beta = \omega\sqrt{L'C'}$	$\omega L' \ll R' \quad \omega C' \ll G'$ $\alpha \approx \sqrt{R'G'} \quad \beta \approx 0$
$Z_0 = \sqrt{\frac{L'}{C'}}$	$Z_0 \approx \sqrt{\frac{L'}{C'}}\left[1 + j\left(\frac{G'}{2\omega C'} - \frac{R'}{2\omega L'}\right)\right]$	

II.2.3.0.1 The coaxial line in TEM-mode

The coaxial line is largely used in laboratories, and as a two-conductor system the line carries TEM-signals. Most common are the SMA and the N-connector. SMA is the acronym for SubMiniature Version A connector and the N-connector is named after its inventor Paul Neill of Bell Labs/USA. Depending on the application, BNC connectors (Bayonet Neill-Concelman) are also used, which provide a quick connector/de-connector clamp, or SMB connectors (SubMiniature Version B) which are slightly smaller than SMA connectors. Also, other connector types exist to be used in different specific working fields. The choice of the cable and the connector depends on the frequency range of the work carried out.

Coaxial cables carry signals from DC frequencies onwards, however, they are limited towards higher frequencies when higher order (non-TEM) modes start propagating. While there are some applications in which coaxial lines are used for propagating higher order modes, this feature is commonly unwanted as it can lead to a mixture of modes and superposition of the related EM-fields (see, e.g. [3] for further reading). Figure II.2.4 shows a variety of coaxial lines and their names (left side), as well as different connectors, including transition pieces to connect different cable types, as well as elbows and coaxial T-junctions (right side).

The coaxial line consists of an inner conductor, a dielectric separator and an outer conductor that also has a shielding function. Figure II.2.5 (right) shows the illustration of coaxial cross-section, whereas the right side of Fig. II.2.5 shows a cut through a coaxial line with flexible inner and outer conductors.

Fig. II.2.4: Coaxial lines and their names (left side), different coaxial connectors, transition pieces and other coaxial parts (right side). Source: www.pinpoint.com.tw

The coaxial shielding is achieved since the EM-fields are carried in the area between the inner and the outer conductor such that radiation is fully suppressed. The inner conductor or the outer conductor (or both) can be made out of braided or stranded material, or it can be solid, and is usually made out of copper, and in some cases, a thin foil layer is added to the outer conductor before it is surrounded by a protective non-conducting jacket. Sandwiched between the two conductors is a dielectric that isolates the two conductors and holds the inner conductor in place. Typical dielectrics are polyethylene (PE) or polytetraflourethylene (PTFE). Also so-called air-lines exist, where the dielectric is omitted and the inner conductor is held in place by spacers that are inserted as supports. These air-lines without dielectric are used for example, when signal propagation speed is critical, i.e. to avoid signal delay. The best known application in every day life is the coaxial antenna cable for video, radio or TV distribution.

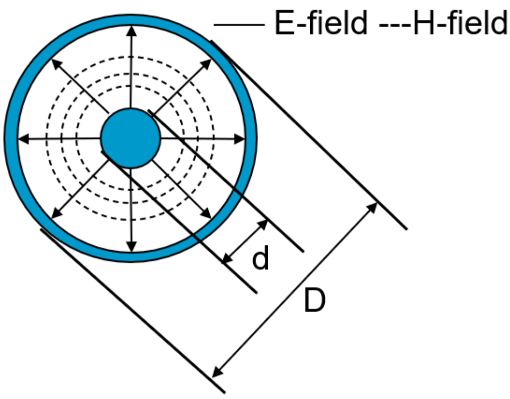


Fig. II.2.5: Illustration of the coaxial cross-section (left), and cut-through a coaxial cable with braided outer conductor and thin protection foil (right).

The characteristic impedance of a coaxial line is calculated from the ratio of inner to outer conductor

$$Z_0 = \sqrt{\frac{\mu}{\varepsilon}} \frac{\ln\left(\frac{D}{d}\right)}{2\pi}, \tag{II.2.5}$$

with d being the outer diameter of the inner conductor and D being the inner diameter of the outer conductor, and $\mu = \mu_0\mu_r$ being the permeability, and $\varepsilon = \varepsilon_0\varepsilon_r$ the permittivity of the dielectric. For low-loss dielectric without any magnetic properties ($\mu_\mathrm{r} = 1$), the equation simplifies to

$$Z_0 = \frac{60\ \Omega}{\sqrt{\varepsilon_\mathrm{r}}} \ln\left(\frac{D}{d}\right), \tag{II.2.6}$$

and for a so-called *airline*, i.e. a coaxial line without any dielectric, the characteristic impedance reads

$$Z_0 = \sqrt{\frac{\mu_0}{\varepsilon_0}} \frac{\ln\left(\frac{D}{d}\right)}{2\pi} = 120\pi\Omega. \tag{II.2.7}$$

In these cables, the inner conductor is supported by small centering rings, usually in star shape, of which the impact on the impedance and on other cable parameters can be neglected.

Also take note that the value $120\pi\Omega \approx 377\ \Omega$ is the so-called vacuum impedance and is often indicated as η_0 in literature. For measurements, cables with characteristic impedance of 50 Ω are used in most cases, to have the cable matching with the characteristic impedance of the instruments used. The ratio of outer-to-inner diameter is ~ 2.3, but also other characteristic impedances are used. Table Table II.2.5 gives some examples. However, the ratio between inner and outer diameters of the conductors

Table II.2.5: Characteristic impedance vs. conductor ratio for coaxial cable structures.

$Z_0/\sqrt{\varepsilon_r}$ [Ω]	ratio $\frac{D}{d}$
24.3	1.5
41.56	2.0
50	~2.3
75	~3.5

is not only important to obtain a certain Z_0, the optimum ratio D/d which is chosen also depends on the application, for example:

- optimum ratio to minimise losses (minimum of cable damping),
- equally for maximum voltage carriage of the cable to avoid RF cable break-down, and
- maximum power transition, that is e.g. used for connecting a cavity to the amplifier.

While all cables have losses, the amount of losses can limit the performance of the entire system (e.g. limit the power that can be transported to an accelerating cavity by the feeder-line), and these losses are also a function of cable length. For lower frequencies, typically below 3 GHz, losses in coaxial cables are usually small, and the main loss at low frequencies is due to the skin effect in the inner conductor. With increasing frequencies, dielectric losses of the insulating medium between inner and outer conductor become dominant. Figure Fig. II.2.6 illustrates a typical loss contribution of the different parts of a coaxial line, and their frequency dependence (see e.g. [4] for details).

Losses are not the only concern in coaxial cable optimisation. A power line, used to connect the RF-amplifier to the accelerating cavity, for example, needs to be optimized for efficient power transport and to avoid voltage breakdown, whereas for a measurement cable, phase-stability is very often most important. The attenuation constant of a coaxial line can be split into two terms, the first term is indicating resistive losses, α_R and the second term shows dielectric losses α_D. The total attenuation is then

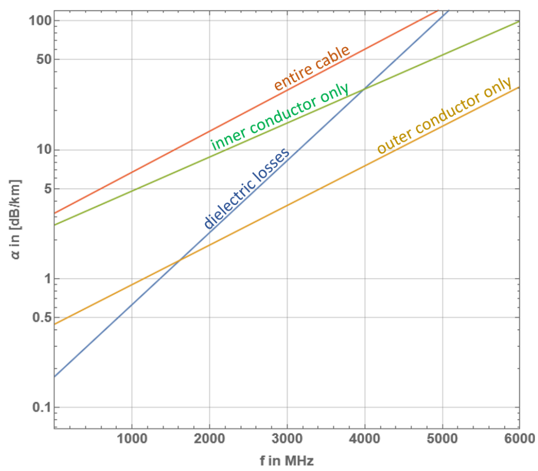


Fig. II.2.6: Illustration of loss contribution of the different parts for a coaxial cable, and frequency dependence.

calculated from

$$\alpha = \underbrace{\frac{\sqrt{\varepsilon_r}}{Z_0 \ln\left(\frac{D}{d}\right)} \left(\frac{1}{d} + \frac{1}{D}\right) \sqrt{\frac{\mu\,\omega}{2\,\sigma}}}_{\alpha_{\mathrm{R}}} + \underbrace{\pi f \frac{\sqrt{\varepsilon_r}}{c_0} \tan\delta_\varepsilon}_{\alpha_{\mathrm{D}}} \tag{II.2.8}$$

Note that the expression $\sqrt{\frac{\mu\,\omega}{2\,\sigma}}$ denotes the frequency-dependent surface resistance R_s, and Z_0 is the vacuum impedance, often also denoted as $\eta_0 = 120\pi\Omega$.

Coaxial line with minimum damping - Example

1. How to get the *optimal ratio D/d* of outer conductor diameter to inner conduction diameter for a coaxial cable to **obtain minimum loss?**

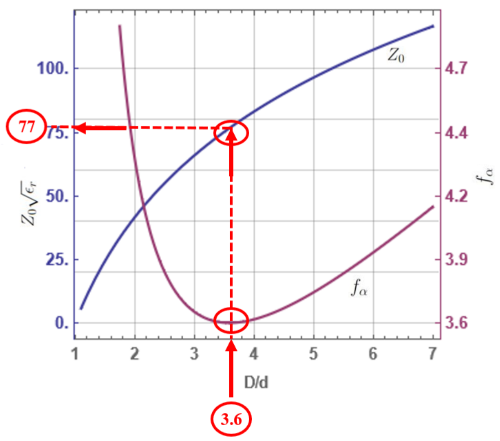


Fig. II.2.7: Characteristic impedance Z_0 and form factor of the loss calculation vs. ratio of diameters outer to inner conductor D/d for a coaxial line.

Since the term of the dielectric losses is independent of the ratio of the two diameters, we will carry out the following steps to find a ratio D/d, such that the first term reaches a minimum:

(a) Re-write the resistive (ohmic) losses to single out the ratio D/d reads:

$$\alpha_{\mathrm{R}} = \frac{\sqrt{\varepsilon_r}\, R_s}{Z_0 \ln\left(\frac{D}{d}\right)} \left(\frac{1}{d} + \frac{1}{D}\right) = \frac{\sqrt{\varepsilon_r}\, R_s}{Z_0 \ln\left(\frac{D}{d}\right)} \frac{1}{D} \left(\frac{D}{d} + 1\right)$$

We note that the resulting term not only depends on the ratio, but also on $1/D$. This is an indicator for the obvious result that damping of the line is reduced, if the line has a very large outer diameter.

(b) Group all terms that depend on the ratio D/d in a form factor which we denote f_α

$$f_\alpha = \frac{1 + \frac{D}{d}}{\ln \frac{D}{d}} = \alpha_{\mathrm{R}} \frac{Z_0 D}{R_s \sqrt{\varepsilon_r}}. \tag{II.2.9}$$

This form factor is of the type $\frac{x+1}{\ln(x)}$ and its dependence on the ratio D/d is illustrated in Fig. II.2.7. The plot also shows the characteristic impedance for a coaxial line without dielectric between inner and outer conductor (Z_0).

(c) From the plot, we can see that f_α reaches a minimum for the ratio $D/d = 3.6$ which corresponds to a characteristic impedance of 77 Ω.

(d) If we select PTFE for dielectric which has a permittivity of ε_{r} = 2.1, we obtain for the characteristic impedance of the line ≈ 53 Ω.

Note that the value obtained is valid only if the inner and the outer conductor of the line are of the same material. In the case that different materials are used, the calculation can follow the same principle, but will become slightly more complex. It is obvious that optimisations for coaxial lines can be carried out for other applications as well, e.g. to maximise the power transport capability or for minimum sparking (see e.g. [5] for details).

II.2.3.0.2 Planar Transmission Lines

Beyond coaxial lines and hollow wave guides, the group of planar transmission lines comprises all types of wave guides where at least one conductor has a flat (planar) shape. This conductor is usually supported by or embedded in dielectric material, denoted a substrate. The idea of using this type of geometry for signal transmission is from the 1950s [6], however, only once the technology was developed to produce planar transmission lines for microwave integrated circuits (MICs), planar lines got tremendously popular. Nowadays, this is the standard in circuit design, making use of the fact that circuit elements at higher frequencies can be built in planar transmission line techniques by means of thickfilm or thinfilm technologies or the more modern photolithographic processes. Three main versions of the planar transmission line exist, these are the stripline, the microstripline, and the slotline, but many other variations of the three main versions exist, depending on the application [7] As a caveat of this technology, it should be mentioned that open planar transmission lines start to radiate as soon as the transit time of the transported RF signal comes close to the range of the time period of the RF signal.

A particular type of planar transmission lines is the so-called *stripline* which consists of a flat conductor, in the simplest case sandwiched symmetrically between two large ground plates, see fig:stripline. A dielectric substrate between the two ground plates supports the inner conductor. A stripline can be easily produced by etching a conductor strip on a substrate and then covering the set with a second substrate layer. Afterwards, the second ground plate is added in a metallisation process. The fundamental mode of propagation for the stripline is transverse electromagnetic (TEM) mode, provided that the conductors can be considered perfectly conducting, the embedding substrate is homogeneous and lossless, the line is not exposed to external fields, and the dimensions are chosen such that the ground plates are large compared to the flat conductor. Further, the distance between the two conducting ground plates and the conductor width needs to be kept below $\lambda/2$ to avoid that higher order modes are building up in the structure. Figure II.2.8 illustrates a cross-section of a stripline with the TEM-field distribution. To calculate the EM-fields for TEM-mode operation, electrostatic field expressions are sufficient, and often conformal mapping is used to solve for the field distributions. As soon as the structure becomes slightly asymmetric, the exact EM-calculation is mathematically getting heavy, however, closed-form expressions can be found in good handbooks (see, e.g. [3,7,8]).

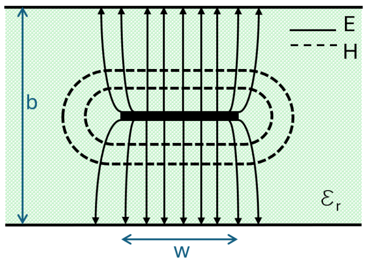


Fig. II.2.8: Symmetric stripline on substrate with electromagnetic field configuration.

Similar to the coaxial line, described in Sec. II.2.3.0.1, the loss-less stripline in TEM-mode is characterised by the three parameters v_p, Z_0, and β. Phase velocity v_p, and phase constant β can be either calculated from the material parameters, or from the equivalent circuit of the line:

$$\begin{aligned} v_p &= \frac{1}{\sqrt{\mu_0 \varepsilon_0 \varepsilon_r}} = \frac{c_0}{\sqrt{\varepsilon_r}} = \frac{1}{\sqrt{L'C'}}, \\ \beta &= \frac{\omega}{v_p} = \omega\sqrt{\mu_0 \varepsilon_0 \varepsilon_r}, \end{aligned} \tag{II.2.10}$$

where L' and C' denote inductance and capacitance of the lumped elements in the equivalent circuit expression of a line of length dz, and have the units [H/m], and [F/m], respectively (see Sec. II.2.3 for details).

II.2.3.0.3 Waveguides in TE- and TM-modes

As was already mentioned, we generally distinguish between two types of wave propagation: So-called free wave propagation that is carried out by waves in unbounded media (see Chapters I.1 and II.1), and so-called guided waves, which means that the wave propagation is confined by material boundaries, typically metallic walls or dielectric materials. Recall that in the case of TEM wave propagation, the magnetic and the electric field components only exist in the plane transverse to the propagation direction and that the resulting wave patterns depend on the number of conductors that are used.

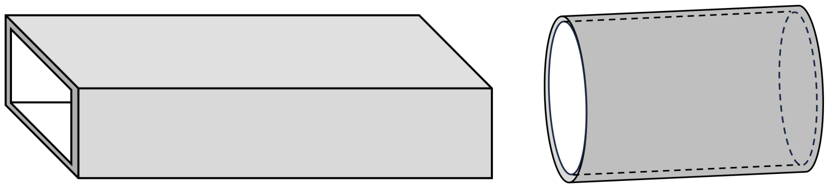


Fig. II.2.9: Illustration of waveguides with rectangular (left) and circular cross-section (right).

Contrary to coaxial lines and striplines, waveguides are a single conductor system, i.e. propagation of TEM-waves is prohibited by design. Geometrically, waveguides are just hollow metallic tubes with uniform cross-sections of which different cross-sectional shapes exist. Figure II.2.9 shows the illustration of waveguides with a rectangular and a circular cross-section. A major advantage for wave propagation in waveguides and coaxial lines is that both build an enclosed system without any radiation losses.

Propagating waves inside a waveguide are following the waveguide shape, even if it is bend. Figure II.2.10 shows a picture of a rectangular waveguide with a connecting flange and a bending shape, as well as a waveguide termination with a coaxial N-connector attached that works as a transmission from waveguide to coaxial line. Inside the waveguide, the propagating waves need to fulfill the boundary conditions on the waveguide walls. From the electromagnetic field description, we know that on *perfectly conducting surfaces*, the tangential electric field has to vanish to fulfill the boundary condition. We speak of an *electric wall*, where $E_{\text{tang}} = 0$. This boundary condition for the electric field can be fulfilled with the use of trigonometric functions, where the roots of the sinusoidal pattern matches the position of the conducting walls. The derivation of the EM-fields can be found in Chapter II.1. Except for the case of loss calculation, the well conducting walls of a waveguide can be approximated as a perfect conductor.

Fig. II.2.10: Picture of a rectangular waveguide (left) and a waveguide termination with a coaxial N-connector attached (images: www.pasternak.com).

Wave propagation inside a waveguide is building up in discrete frequency patterns, so-called *waveguide modes*, and depending on their field distribution, it is distinguished between

– TE-modes (transverse electric), where the **electric field** component in the waveguide's cross-section is building up only in the plane transverse to the propagation direction, and no electric field component in propagation direction exists, and

– TM-modes (transverse magnetic), where the **magnetic field** component in the waveguide's cross-section is building up only in the plane transverse to the propagation direction, and no magnetic field component in propagation direction exists.

Each mode has a so-called *cut-off frequency*, a lower frequency limit below which wave propagation in the waveguide is not possible. The cut-off frequency is depending on the waveguide cross-sectional dimensions, only, and the mode with the lowest cut-off frequency is called the *dominant mode*. Note that for higher frequencies, also so-called *hybrid waves* will build up their field patterns in the waveguide. Hybrid mode propagation can be considered as a combination of the resulting field patterns of TE- and TM-modes, and will be omitted here.

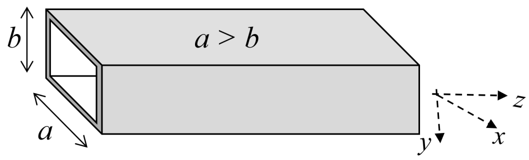


Fig. II.2.11: Cross-sectional dimensions of rectangular waveguide with wave propagation in z-direction.

The cut-off frequency of **rectangular waveguides** can be easily calculated from the waveguide cross-sectional dimensions shown in Fig. II.2.11. It is identical for TE- and TM-mode propagation

$$f_{c,mn} = \frac{1}{2\sqrt{\mu\varepsilon}}\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2}.$$

Note that if $n = m = 0$, all field components become zero, as there is no TE_{00} mode. The dominant mode is TE_{10}-mode as it has the lowest cut-off frequency which simplifes to

$$f_{c,10} = \frac{1}{2a\sqrt{\mu\varepsilon}}.$$

Figure II.2.12 shows the electric field of the TE_{10}-mode in the cross-section, as obtained from EM-simulation codes. We can clearly see the field maximum building up in the center of the cross-section, while vanishing on the conducting waveguide walls. Equally interesting are the EM-field patterns as illustrated in RF-engineering books (see pictures in the lower part of Fig. II.2.13). Here, the electric field is shown by continuous lines, and the increased field strength is indicated by using a denser pattern of arrows in the center, and no field lines next to the conducting walls. The corresponding magnetic field is shown in dashed lines, and the views 1 and 2 are the EM-field patterns in the respective longitudinal cuts (pictures courtesy: Pozar [3]). Figure II.2.13 shows also the next higher modes field patterns [3].

Equally to the rectangular waveguide, the propagation patterns of the EM-fields build up in discrete modes, and the cut-off frequency of a **circular waveguide** (see Fig. II.2.14) depends only on the inner radius R of the tube.

Due to the cylindric geometry, trigonometic functions cannot be used any more, instead we need *Bessel functions* of 1st type to fulfill the boundary conditions on the cylindric wall. In the case of a perfectly conducting tube with radius R with no resistive attenuation, we obtain different propagation

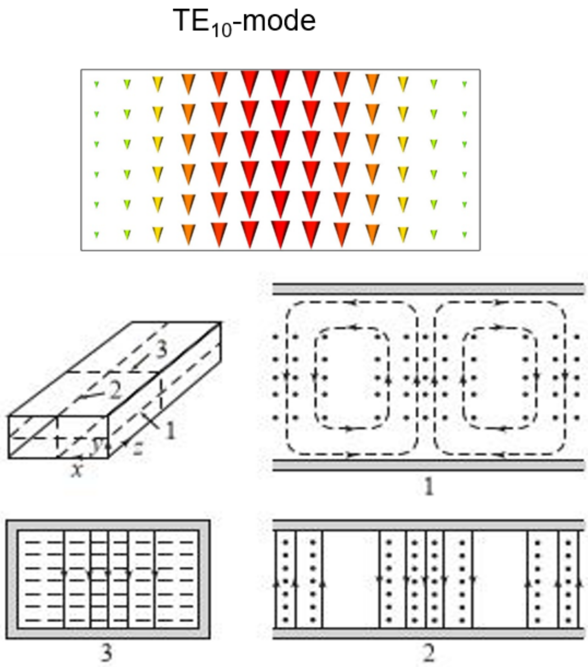


Fig. II.2.12: Field pattern of the TE_{10}-mode as obtained from EM-simulations together with the field patterns as illustrated in RF-engineering books (simulation: E. Jensen, picture courtesy: Pozar [3]).

constants for the TE-mode and the TM-mode as follows:

$$\beta_{nm} = \sqrt{\omega^2 \varepsilon \mu - \left(\frac{p'_{nm} \text{ or } p_{nm}}{R}\right)^2},$$

where p_{nm} denote the roots of the Bessel function for a TM-mode, i. e. $J_n(p_{nm}) = 0$, and p'_{nm} denote the roots of the derivative of the Bessel function for the TE-mode, i. e. $J'_n(p'_{nm})$. This leads to the cut-off frequencies for the different modes (see Fig. II.2.15)

$$f_{c,nm} = \frac{1}{2\pi\sqrt{\mu\varepsilon}} \left(\frac{p'_{nm} \text{ or } p_{nm}}{R}\right).$$

Figure II.2.15 shows the Bessel functions of first type J_0, J_1, J_2, as well as the derivative of the Bessel function of first type J'_0 together with their 1st and 2nd roots. The notation for the roots of the Bessel function is as follows:

$$\begin{aligned} p_{01} &:= 1^{\text{st}} \text{ root of Bessel function of first type, } J_0 \\ p_{02} &:= 2^{\text{nd}} \text{ root of Bessel function of first type, } J_0 \\ p'_{01} &:= 1^{\text{st}} \text{ root of derivative of Bessel function of first type, } J'_0, \end{aligned}$$

and the values of the roots of the Bessel function, as well as the derivative of the Bessel function can be found in Table II.2.6 below.

Let's see an example: calculate the cut-off frequencies for a circular waveguide with inner radius $R = 60$ mm. Our waveguide is filled with vacuum, hence we take the vacuum material parameters

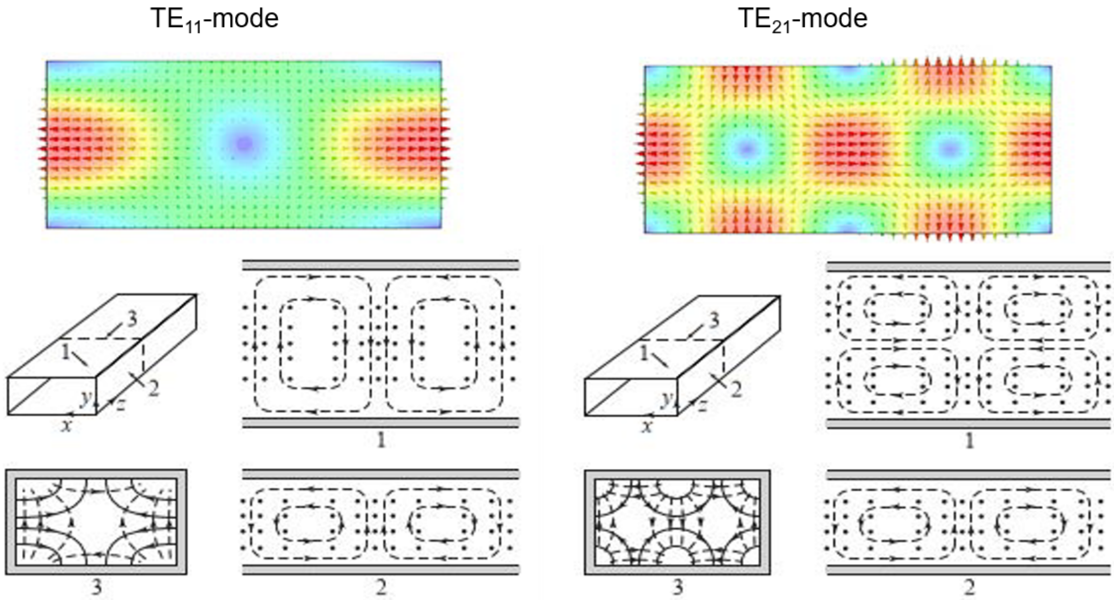


Fig. II.2.13: Field pattern of the TE_{11}-, and the TE_{21}-mode as obtained from EM-simulations together with the field patterns as illustrated in RF-engineering books (simulation: E. Jensen, picture courtesy: Pozar [3]).

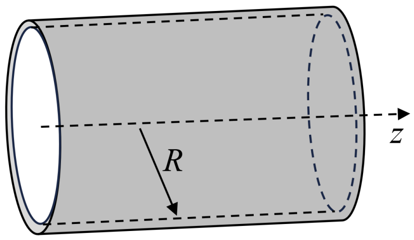


Fig. II.2.14: Waveguide with circular cross-section of inner radius R, and wave propagation in z-direction.

Table II.2.6: Values of roots of the Bessel functions for TE- and TM-modes of a circular waveguide.

	Root values p'_{nm} for TE-modes		Root values p_{nm} for TM-modes	
n	p'_{n1}	p'_{n2}	p_{n1}	p_{n2}
0	3.832	7.016	2.405	5.520
1	1.841	5.331	3.832	7.016

$\mu_0 = 4\pi \cdot 10^{-7}$ Vs/Am, and $\varepsilon_0 = 8.854 \cdot 10^{-12}$ As/Vm, and from Table II.2.6, we insert the correct roots of the Bessel functions, to obtain:

$$f_c(\mathrm{TE}_{11}) = \frac{1}{2\pi\sqrt{\mu_0\varepsilon_0}} \frac{p'_{11}}{R} = \frac{2.9 \cdot 10^8}{2\pi} \frac{1.841}{0.06} = 1.416 \text{ GHz},$$

$$f_c(\mathrm{TM}_{01}) = \frac{1}{2\pi\sqrt{\mu_0\varepsilon_0}} \frac{p_{01}}{R} = \frac{2.9 \cdot 10^8}{2\pi} \frac{2.405}{0.06} = 1.850 \text{ GHz},$$

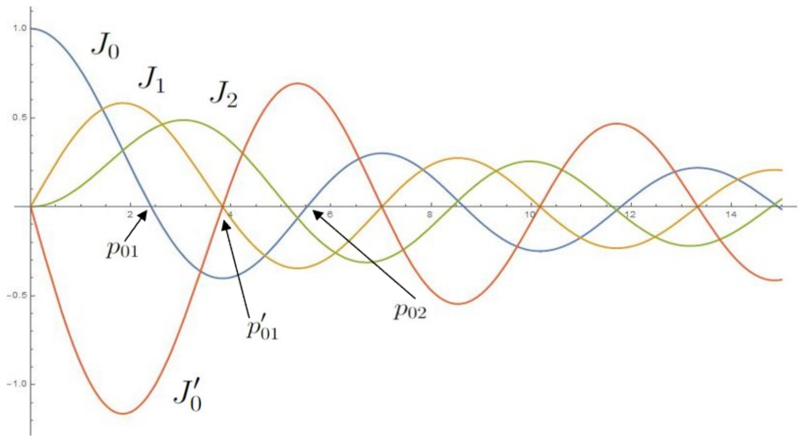
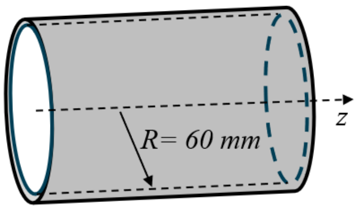


Fig. II.2.15: Bessel function of first type, J_0 and derivative of Bessel function of first type, J'_0. Root values of the Bessel functions are indicated.



$$f_{c,nm} = \frac{1}{2\pi\sqrt{\mu\epsilon}}\left(\frac{\overset{\text{TE-mode}}{p'_{nm}} \text{ or } \overset{\text{TM-mode}}{p_{nm}}}{R}\right)$$

Fig. II.2.16: Calculation of cut-off frequencies for TE- and TM-modes in a circular waveguide for a given inner radius.

as the frequencies where these modes start propagating. Figure II.2.17 shows the electric field of the TE_{11}-mode, which is the fundamental mode in a circular waveguide. Due to the circular symmetry, this mode comes with two possible orientations of the electrical field, we speak of *mode polarisations*. The magnetic field of the mode with the next higher frequency is shown in Fig. II.2.18.

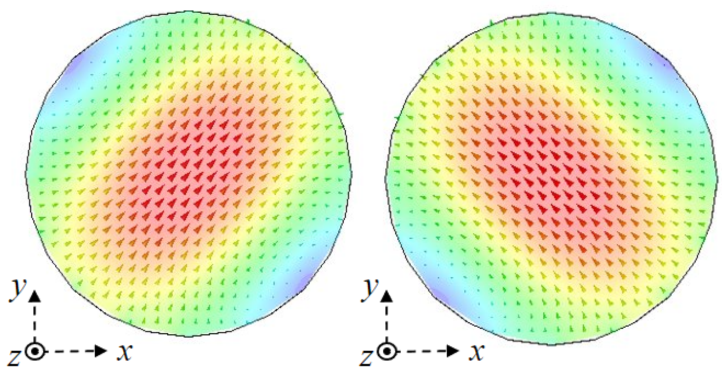


Fig. II.2.17: Simulated electric field pattern of TE_{11}-mode in the cross-section of a round waveguide with two polarisations.

Note that occasionally, more than one mode is propagating in a waveguide at the same time. Such a waveguide is called *overmoded*. This case is typically unwanted during waveguide operations, as it can

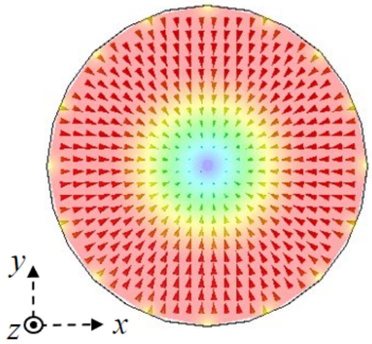


Fig. II.2.18: Simulated magnetic field pattern of TM_{01}-mode in the cross-section of a round waveguide.

lead to an undesired propagation pattern, and the transported energy is split between the two modes.

II.2.4 The standing wave “pill-box” resonator

Microwave resonators can be built from waveguides by closing the two open ends. We call such an RF-resonator a *cavity*. The cavity interacts directly with the beam, providing acceleration or energy increase to the beam, it is also used for all kind of RF manipulations within an accelerator. A cavity stores electric and magnetic energy inside the hollow body, and the frequency of its electromagnetic field resonance depends on the cavity dimensions. In the same way as for the EM-field distributions that we know from the rectangular and the circular waveguide, the electromagnetic field has to fulfil the boundary conditions on the resonator walls, and also in this case, the field builds up in so-called *resonant modes*.

Cavities which are used for the acceleration of particles in our accelerators are mostly of a cylindrical and flat shape. This is why we call them pill-box cavities. Generally, resonators are classified by their quality factor Q (often simply called the Q-value) which can be used as a measure of “how well the cavity is resonating”.

The Q-value is defined as the ratio between the energy stored in the resonator vs. the energy dissipated in the resonator’s wall:

$$Q_0 = 2\pi f_0 \frac{\text{energy stored in the resonator}}{\text{energy dissipated in the resonator}}.$$

We distinguish between Q_0, denoting the so-called *unloaded* Q-value that describes the pure resonator, independent of the external world, and Q-values that consider the cavity being connected to its surroundings, for example by a feeder-line, the amplifier, vacuum beam pipes, or even the impact of the beam passing through the cavity. The Q-value of a cavity reduces, if loss mechanisms are introduced, including losses that are simply due to the power that is dissipated in the metallic wall of a cavity. Equally, all the abovementioned connections to the cavity resonating body leads to a reduction of the unloaded Q_0, and we speak of *cavity loading* when the cavity is introduced to external, not resonator-intrinsic loss mechanisms. In the case that the cavity is filled with a dielectric, these losses will have to be considered as well, as can be done in this case via the dielectric’s material parameters (not considered here).

As an example, we can compare measurements taken on the PS 80 MHz pill-box cavity (shown in Fig. II.2.19) which was tested for Q-value deterioration by connecting a lossy tuner geometry that shifted the cavity out of its fundamental mode. In this case, the tuner is the external system (with its external Q_{ext}), and the connection of the cavity to the tuner is a cavity loading mechanism.

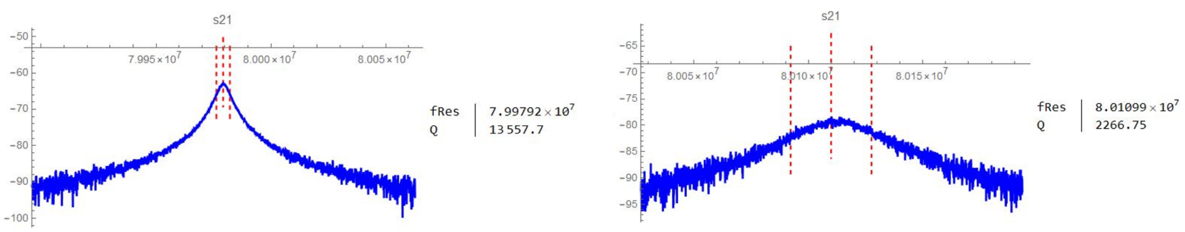


Fig. II.2.19: Measurement of unloaded Q_0 and loaded Q_{L} on 80 MHz cavity after introducing an external loading mechanism.

Figure II.2.19 shows the measurement of unloaded $Q_0 \approx 13557$ taken in transmission at resonance, at its fundamental frequency $f_{\mathrm{res}} \approx 80$ MHz (left) and the loaded Q_{L} taken on the same cavity with the tuner connected as load mechanism. The connection of a load mechanism is leading to a small shift in resonance frequency to $f_{\mathrm{res}} \approx 79.98$ MHz, and at the same time causing a Q-reduction to Q_{L} = 2267. Measurement was taken with a Vector Netwerk Analyser (VNA) as S-parameter measurement in transmission, and Q-values were determined via the *3dB-method*.

II.2.4.0.1 Transit time effect

At this stage, it is important to introduce the so-called *transit time effect*, and the *transit time factor* T. Since the cavity is providing a time-varying electric field over gap of the type $V_0 \cos(\omega_0 t)$, the energy that can be given to the particle by this harmonically oscillating field always will be less than what the particle would gain if it passes through a DC field of voltage V_0, due to the limited particle velocity v. Hence, the transit time factor takes into account the time which is needed by the particle to pass through the cavity (and the fact that the cavity will not be always on maximum voltage during this passage). Figure II.2.20 illustrates this concept.

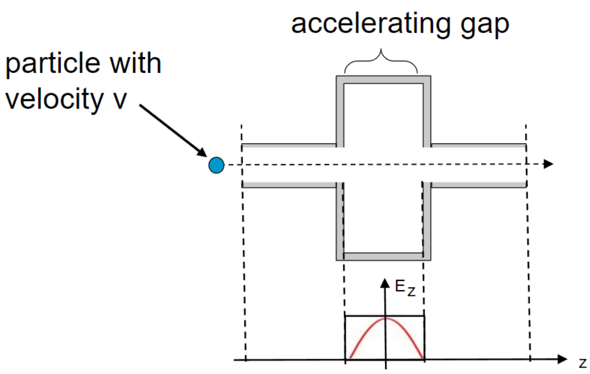


Fig. II.2.20: Concept of accelerating gap within a cavity of harmonically changing electric field compared to a DC electric field of voltage V_0.

This way, the transit time factor T can be considered as a parameter that describes the reduced energy gain for the particle passing the accelerating gap. T is defined as follows

$$T = \frac{\text{energy gained in time} - \text{varying RF}-\text{field}}{\text{energy gained in a DC} - \text{field of voltage } V_0},$$

and can be written as:

$$T = \frac{\int_{-L/2}^{L/2} E(0,z)\cos(2\pi z/\beta\lambda)dz}{\int_{-L/2}^{L/2} E(0,z)}.$$

Recall that the cosine argument can be derived from

$$\omega t = \omega\frac{z}{v} = 2\pi f\frac{z}{\beta c_0} = \frac{2\pi z}{\beta\lambda}.$$

From the explanations given above, it can be seen that to achieve the maximum energy gain from the voltage in the accelerating gap, we would like a transit time factor $T = 1$. From the formulae, however, this would mean that the gap length $g = 0$ inside the cavity. Although this is not possible, a cavity design request can be formulated from this, i. e. for a single gap cavity, keep **the gap as small as possible**. However, other considerations like the increasing risk of electric breakdown that scales inversely to the gap size also impacts on the optimum gap geometry. Note that for the consideration of the transit time factor, it is assumed that the particle does not change velocity along the gap length.

II.2.4.0.2 Accelerator efficiency figures-of-merit

As a means to compare different cavity designs, several figures-of-merit are commonly used to characterize accelerating cavities which will be introduced here:

– (unloaded) Q_0-value which was just mentioned above, is a measure of the quality of the resonator's resonance. It is dimensionless and calculated from

$$Q_0 = \frac{\omega_0 U}{P},$$

where ω_0 denotes the angular resonance frequency, and the term $\omega_0 U$ is the energy *stored* in the resonator, whereas P is the energy *dissipated* in the resonator.

– Shunt impedance r_s, given in [MΩ]. r_s is a measure of the effectiveness of the cavity to produce an axial voltage V_0, hence a typical design goal for a single-gap (pill-box) cavity is to read a high shunt impedance. The shunt impedance is calculated from

$$r_\mathrm{s} = \frac{V_0^2}{P}.$$

– Effective shunt impedance $r_{s,\mathrm{eff}}$, given in [MΩ/m]. This parameter is a measure of effectiveness of the cavity to deliver energy to a particle, normalised per unit power loss. The effective shunt impedance is calculated from, making use of the *transit time factor T*

$$r_{s,\mathrm{eff}} = \frac{(V_0\, T)^2}{P} = r_\mathrm{s} T^2.$$

– R-over-Q (also often called R-upon-Q), given in [Ω], is a measure of cavity acceleration efficiency at a given frequency. It is calculated from

$$r_\mathrm{s}/Q = \frac{(V_0\, T)^2}{\omega_0 U}.$$

II.2.4.0.3 Rectangular cavity resonators

Rectangular cavity resonators have less applications in accelerators, however, since the resonance pattern of rectangular cavity resonators can be easily derived from the already known field pattern of the TE_{10}-mode of the rectangular waveguide, we will start with this shape.

Figure II.2.21 shows a rectangular cavity resonator with a quadratic ground plate. The arrows illustrate the electric field pattern of the lowest mode, TE_{101} which can be derived from the field pattern of the TE_{10}-mode of the rectangular waveguide. TE_{101} is the fundamental mode of this type of resonator, i. e. the mode with the lowest resonant frequency which is also the *dominant mode*. From the boundary

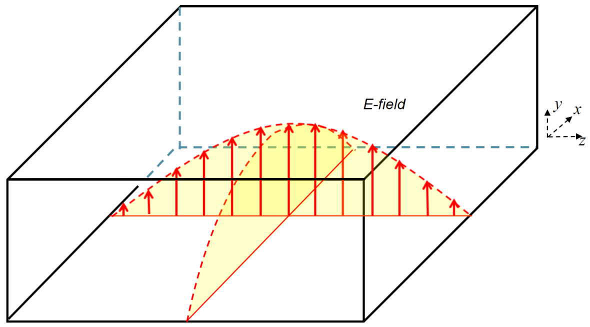


Fig. II.2.21: Rectangular cavity resonator with quadratic bottom plate. Electric field patterns are illustrated.

condition on the well-conducting walls in the x/y-planes and the y/z-planes, the electric field needs to be zero, consequently, multiples of the electric field maxima are building up inside the resonator. The mode indices can be obtained by counting the *maxima* of the electric field along one axis (we see either a TE_{xyz} or a TM_{xyz}- mode). From the side lengths of the resonator, the resonant frequencies for the TE_{mnl}-modes and the TM_{mnl}-modes can be easily calculated from

$$f_{mnl} = \frac{1}{2\pi\sqrt{\mu\varepsilon}}\sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 + \left(\frac{l\pi}{c}\right)^2},$$

and the resonant wavelength for the fundamental mode results to

$$\lambda_0 = \frac{2}{\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2 + \left(\frac{l}{c}\right)^2}}.$$

Note that the equation for the resonant frequency is given for the general case, including the possibility to fill the resonator with a dielectric by making use of the filling material parameters μ and ε. The use of a dielectric will lead to a frequency increase. The Q-value for the TE_{101}-mode is given by

$$Q_{\mathrm{TE101}} = \frac{\lambda_0}{\delta}\frac{b}{2}\frac{(a^2+b^2)^{3/2}}{c^3(a+2b)+a^3(c+2b)}.$$

In all cases, the parameters a, b, c are indicating the resonator's dimensions in x, y and z-directions. Recall that δ denotes the skin depth by which the EM-fields penetrate the resonator wall, defined as $\delta = \sqrt{\frac{2}{\omega\sigma\mu}}$ where σ is the material's conductivity and μ the permeability.

In the case of a quadratic ground plate, when $a = c$, these formulae for the TE_{101}-mode simplify to

$$\lambda_0 = \sqrt{2}a,$$

and

$$Q_0 = \frac{1}{\delta}\frac{ab}{a+2b}.$$

II.2.4.0.4 Cylindric cavity resonators

Compared to rectangular cavity resonators, cylindrically shaped "pill-box" cavity resonators are much more common in accelerator design. Indeed, the so-called "pill-box" cavity is the best visible and best recognised element in an accelerator chain. As it is the **one element** that transforms the actual string of magnets into an accelerator, it is on top of the RF system food chain. However, as was already mentioned earlier, the cavity requires power to be obtained from the high power RF system via a high power RF coupler to excite the cavity resonance, as well as an RF-source and a low-level RF system for cavity operation (also see Fig. II.2.1).

Figure II.2.22 shows the 80 MHz single gap cavity of the CERN Proton Synchrotron (PS) installed in the PS accelerator (left), and seen from the installation side with its RF amplifier connected to the cavity (right). The pill-box type shape of the cavity is clearly visible in both pictures.

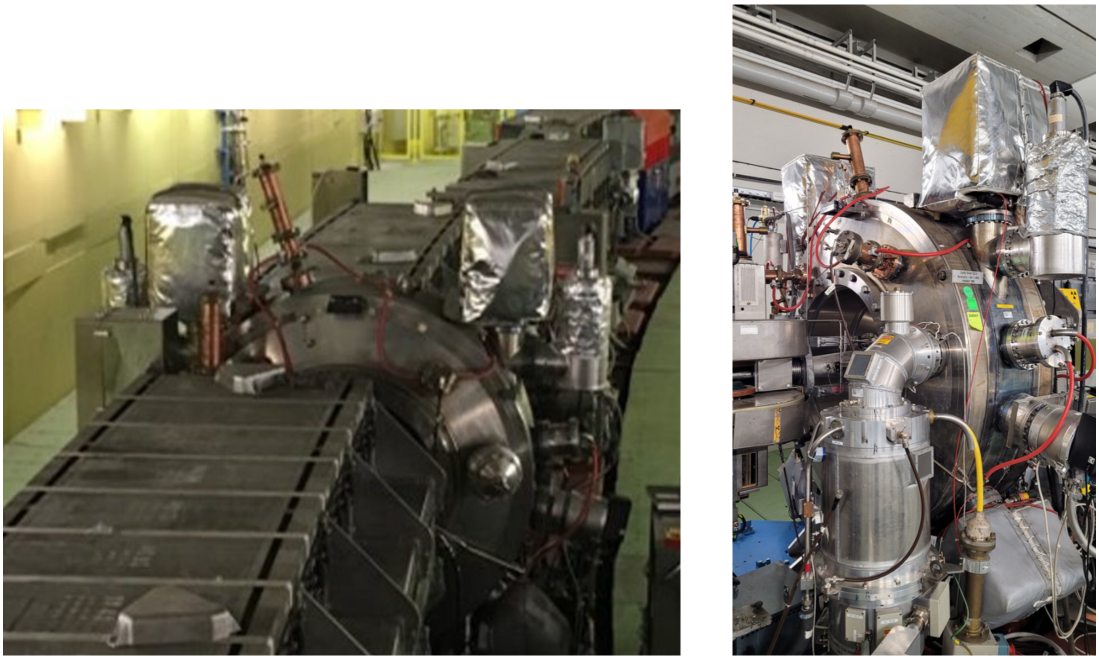


Fig. II.2.22: 80 MHz cavity of pill-box type installed in the CERN PS accelerator, sandwiched between normal conducting dipole magnets (left), and same cavity in side view with its RF amplifier and vacuum pumps connected (right).

Same as for the rectangular shaped cavity, cylindric cavity resonators can be derived from the circular (round) waveguide by closing the waveguide openings. Recall that the TE_{11}-mode is the dominant mode for a circular waveguide - by shortening both waveguide ends, we obtain a cylindrical cavity for which the TE_{111}-mode is the **dominant TE-mode**, whereas the TM_{010}-mode is the **dominant TM-mode**. Our requirement for acceleration of particles is a strong electric field in direction of the particle's trajectory, hence the TM_{010}-mode with a transversal magnetic field, and a longitudinal electric field is to be used.

Figure II.2.23 shows an illustration of the electric and the magnetic field distribution in a pill-box cavity resonator. The red arrows indicate the electric field strength described by the Bessel function J_0 which has its maximum value along the z-axis, whereas the magnetic field which has its maximum close

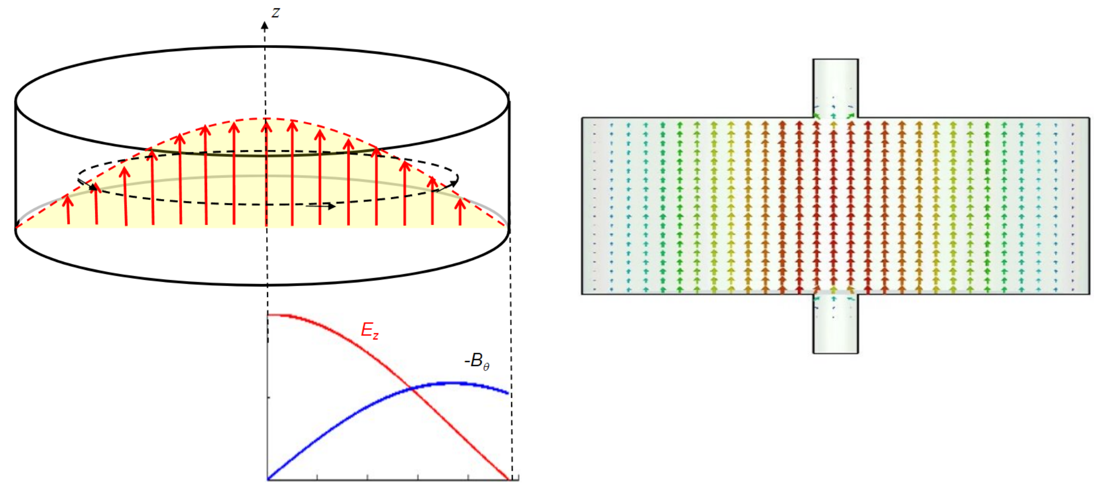


Fig. II.2.23: Left: illustration of electric and magnetic field distributions in a cylindrical cavity of "pill-box" type. Right: simulated electric field in the cross-section of a "pill-box" cavity with vacuum beam pipes attached.

to the outer boundary is indicated by the black dashed line. Note the negative sign of the magnetic field resulting from the derivative of the Bessel function J_0', having opposite sign on either side of the axis (see Bessel functions in Fig.II.2.15 for details). For a pill-box cavity with radius a and height h, the resonant frequencies for the TE- and TM-modes can be calculated from:

$$f_{\mathrm{TE},mnl} = \frac{c_0}{2\pi}\sqrt{\left(\frac{p'_{nm}}{a}\right)^2 + \left(\frac{l\pi}{h}\right)^2}$$

$$f_{\mathrm{TM},mnl} = \frac{c_0}{2\pi}\sqrt{\left(\frac{p_{nm}}{a}\right)^2 + \left(\frac{l\pi}{h}\right)^2}.$$

Note that the TM_{010}-mode is independent of the cavity height h, and as long as the cavity remains in its pill-box shape, i. e. the cavity remains in a maximum diameter-to-height-ratio of ≈ 1, most of the formulae used for the calculation of cavity parameters can be simplified from the general case. The validity of this criteria can be cross-checked by using a so-called *mode chart* for cylindrical resonators. Mode charts are used in RF-practice as a fast means to identify which mode(s) are currently building-up in a cavity. Figure II.2.24 shows a mode chart for a general cylindrical cavity indicating the resonantly excited modes as a function of cavity dimensions, taken from [3].

The same mode chart is also used in section II.2.5, where the numerical calculation of a cavity is carried out, and reads as follows:

– moving along the ordinate axis, where the ratio of the cavity diameter to cavity height is indicated $(2R/h)^2$, the first mode showing up is TE_{111} for all values roughly below 1;

– the TE-modes have no electric field along the particle's trajectory, and hence are not used for acceleration;

– First TM-mode is TM_{010} showing up for ratios $(2R/h)^2 > 1$. The larger the values for this ratio

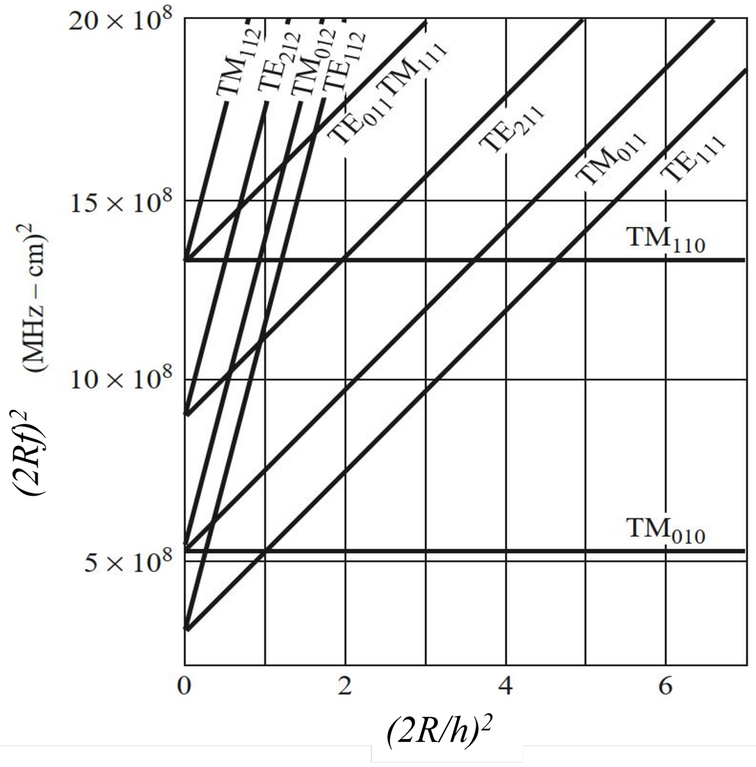


Fig. II.2.24: Resonant mode chart for a general cylindrical cavity showing the excited modes as function of cavity dimensions, R is the cavity radius, and h is the cylinder height (picture courtesy Pozar [3]).

become, the "flatter" our cavity will be, and the more the modes TE_{111} and TM_{010} are separated in frequency. A sufficient frequency separation of modes is usually desired as it reduces the risk that two different modes show up simultaneously in the cavity.

The calculation of a single gap cavity which fulfils the geometric criteria of pill-box cavity when we have no dependence of the resonance frequency on the cavity height h, allows the use of simpler formulae to calculate the accelerator figures-of-merit. For a TM_{010}-cavity with radius a, the wavelength can be calculated from the simple expression

$$0.383\,\lambda_{\mathrm{TM},010} = a,$$

and the calculation of unloaded Q_0 equally simplifies to the expression which can often be found in handbooks of accelerator physics

$$Q_0 = \frac{0.383\,\lambda_{\mathrm{TM},010}}{\delta}\left[1+\frac{a}{h}\right]^{-1} = \frac{a}{\delta}\left[1+\frac{a}{h}\right]^{-1}.$$

Recall that δ denotes the skin depth by which the EM-fields penetrate the resonator wall, defined as

$$\delta = \sqrt{\frac{2}{\omega_0 \sigma \mu_0}}, \tag{II.2.11}$$

where ω_0 is the resonance frequency, σ is the material's conductivity, and μ_0 the vacuum permeability.

The accelerator figure-of-merit R/Q for a pill-box cavity resonator is calculated from

$$R/Q = \frac{4\eta_0}{\pi\, p_{01}^3\, J_1^2(p_{01})} \frac{\sin^2\left(\frac{p_{01}}{2}\frac{h}{a}\right)}{h/a},$$

where η_0 denotes the vacuum impedance which was already introduced in Sec. II.2.3.0.1, and $p_{01} = 2.405$ and $J_1(p_{01}) = 0.51911$ are to be found and calculated with the help of Table II.2.6, showing the roots of the Bessel functions for a circular waveguide. Inserting all numbers, we obtain for R/Q the simple expression

$$R/Q = 128\, \frac{\sin^2\left(\frac{p_{01}}{2}\frac{h}{a}\right)}{h/a}, \tag{II.2.12}$$

which for small numbers of the sinusoidal function can be even further simplified to

$$R/Q \approx 185\, h/a.$$

Depending on the cavity type, the values for the accelerator efficiency figure-of-merit can vary considerably. For orientation, some typical numbers are given in Table II.2.7 below:

Cavity type	R/Q-value	Q_0	r_s
Ferrite loaded cavity (low frequency, rapid cycling)	4 kΩ	50	200 kΩ
Room temperature copper cavity (type 1 with nose cone)	192 Ω	30×10^3	5.75 MΩ
Superconducting cavity (type 2 with large iris)	50 Ω	1×10^{10}	500 GΩ

Table II.2.7: Typical values of accelerator efficiency figures-of-merit (following Caspers et al., JUAS 2021).

Figures II.2.25 and II.2.26 show some pill-box cavities (single gap) which where used in different accelerators. From these pictures, it can be clearly seen that the resonance frequency of the fundamental mode scales directly with the size of the cavity, and this can lead to very large diameters, taking precious space in the accelerator and making the cavities inconveniently large to handling.

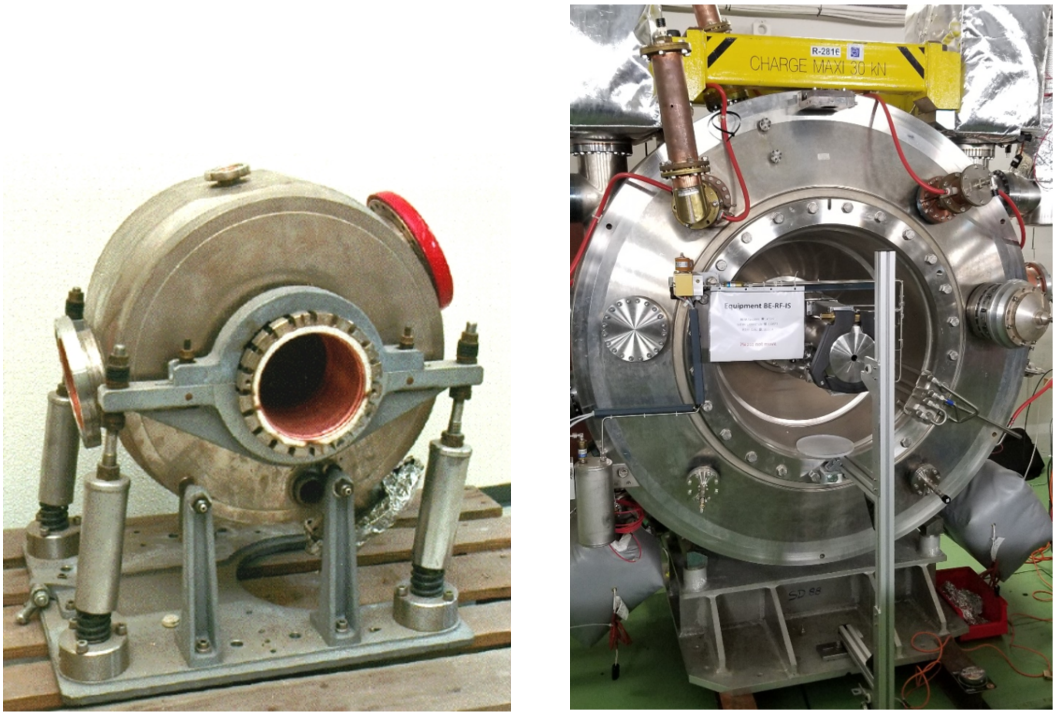


Fig. II.2.25: Large pill-box cavities. Left: cavity from DORIS Storage ring (1970, picture courtesy: H. Damerau), right: PS 80 MHz cavity (currently installed).

Fig. II.2.26: Large pill-box cavities. Example of PS 19 MHz cavity (1966, picture courtesy: E. Jensen).

It should be mentioned that in practice, a *pure* pill-box cavity is unfortunately not very efficient for particle acceleration. However, a simple shape modification can be done by using so-called *nose cones* within the resonating body. A nose cone is a protrusion on the cavity wall that leads to a concentration of the electrical field in the accelerating gap.

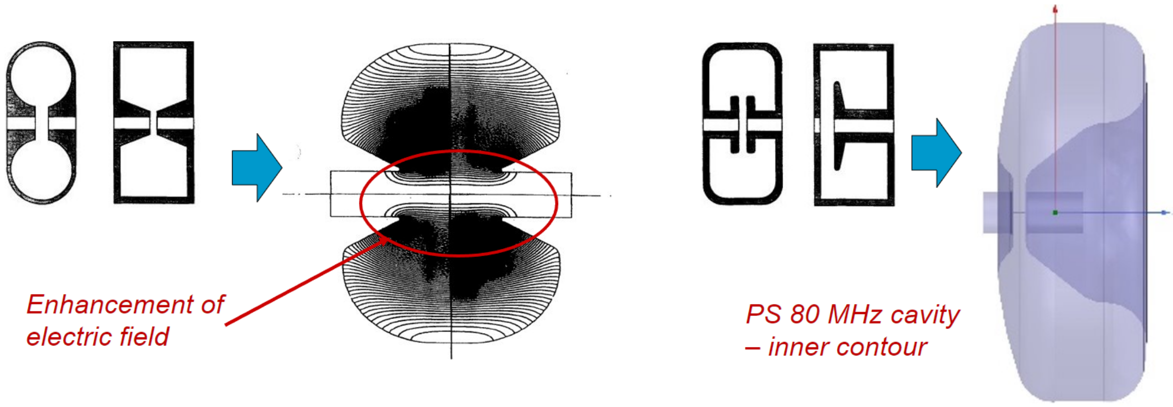


Fig. II.2.27: Concept of nose cones to be used on a pill-box cavity (left) and implementation with EM-simulation code (right). The goal is to enhance the electric field in the accelerating gap. Illustrations taken from Puglisi, RF-CAS 92, CERN.

II.2.5 Numerical analysis

This document aims to cover the topics discussed and exercises performed in the RF tutorial concerning electromagnetic simulations. The simulations covered involve the analysis of Eigenmodes in a resonant cylindrical cavity and are performed using the free student version of the software ANSYS Electronics Desktop (AEDT) 2022 [9]. The latest available student version is easily accessible and it can be downloaded from the ANSYS website: `https://www.ansys.com/academic/students/ansys-electronics-desktop-student`. This tutorial is based on the write-ups performed for the CERN Accelerator School (CAS) for the Advanced Accelerator Physics and Radio Frequency training [10, 11].

II.2.5.1 Ansys HFSS

Ansys HFSS is a 3D electromagnetic (EM) simulation software for designing and simulating high-frequency electronic products such as antennas, RF or microwave components. It uses the Finite Element Method (FEM) and it subdivides the structure in tetrahedral elements where Maxwell's equations are solved.

Settings

1. Open ANSYS Electronics Desktop.
2. Select the solver **HFSS** from the *Desktop* toolbar on top. A new project and the new HFSS Design will be accessible from the Project Manager window on the left.

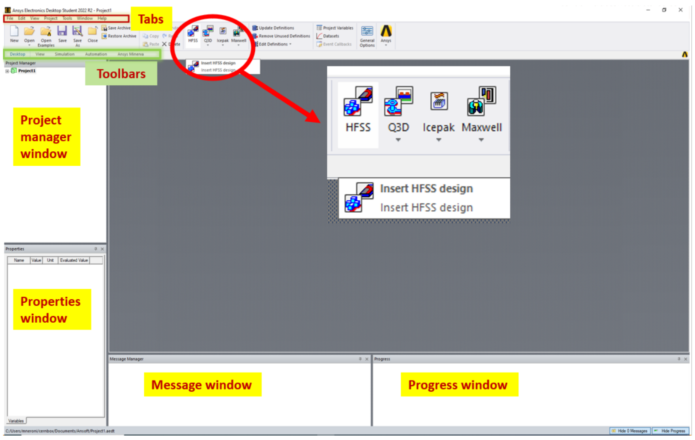


Fig. II.2.28: AEDT Graphical User Interface - Creation of HFSS design

3. After the creation of the HFSS design, the *Eigenmode* solver must be selected from the "HFSS" Tab (top menu) $\rightarrow$ Solution types $\rightarrow$ *Eigenmode* 1.
4. The project units can be chosen from "Modeler" $\rightarrow$ **Units** $\rightarrow$ *cm* 2.
5. Finally, the project can be saved with the name '*Cavity-JUAS23*'. A HFSS project will have a file extension *.aedt 3.

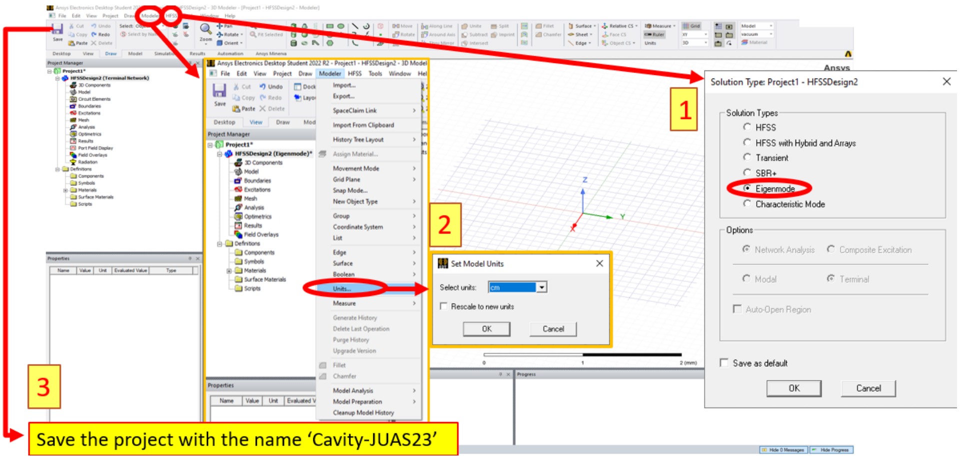


Fig. II.2.29: Solver definition, project units and saving of HFSS project.

II.2.5.2 Modelling

The Eigenmodes analysis is carried out on a cylindrical cavity with given dimensions which will be defined as parameters in the HFSS project and which are indicated in Fig. II.2.30.

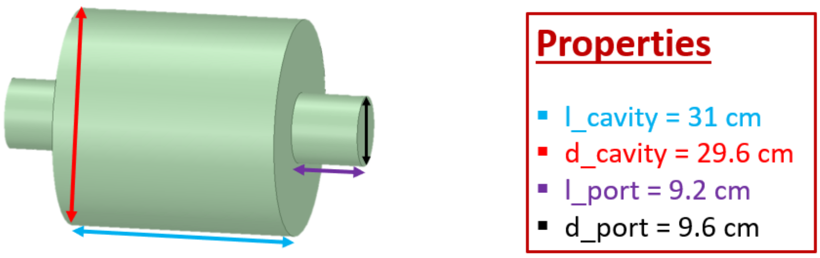


Fig. II.2.30: Model geometry and cavity dimensions.

The geometry can be built using three cylinders, one representing the cavity and two small ones for the input and output ports, by following the steps below.

Model geometry

1. Select the cylinder shape from the *Draw* toolbar (top panel).
2. Bring the cursor at the origin of axis and click to fix the **center**.
3. Drag and click to fix the cylinder **radius**.
4. Drag and click to fix the cylinder **length**. Note that the dimensions can be defined in a following step (Fig.II.2.31).

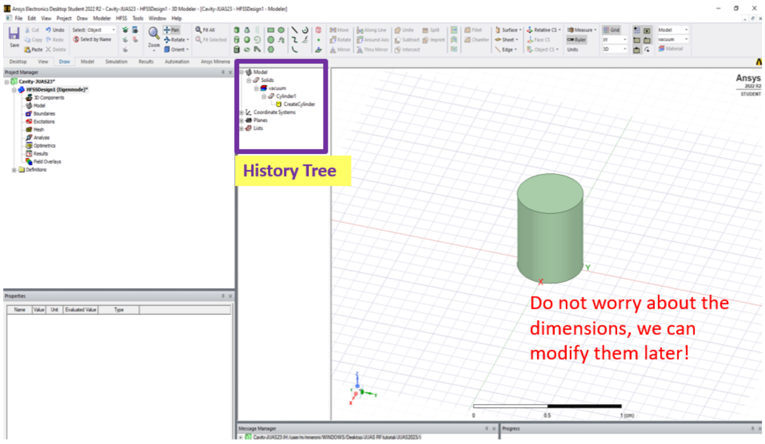


Fig. II.2.31: Modelling of the cylindrical cavity.

5. From the History Tree window, under Model → Solids → vacuum → *Cylinder1*, double-click on *CreateCylinder* - **1** - Fig. II.2.32.

6. Type the coordinates of the center position, the symmetry axis, radius and length (height) of the cylindrical resonator, by using the variables as defined in Fig. II.2.30 - **2**.

7. When typing a variable, a new window will pop up and you will be able to select the variable type, unit and value - **3**.

8. The new variables will appear under properties when the correspondent HFSSDesign is selected in the project manager window - **4**.

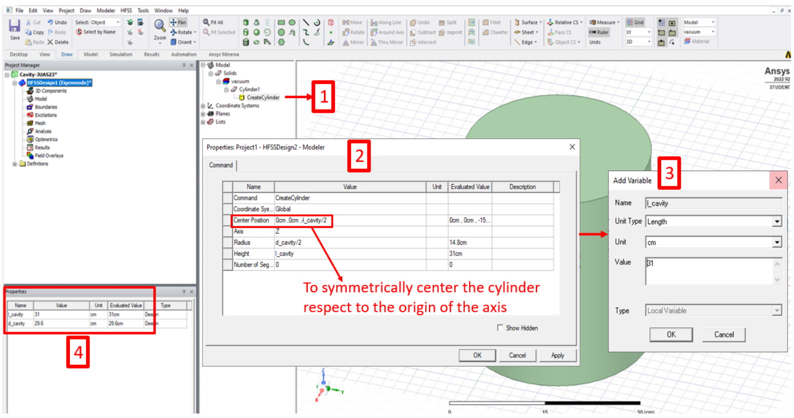


Fig. II.2.32: Defining cavity dimensions.

9. The input and output beam-ports can be built following the same procedure. The output port center coordinates will be: (0, 0, $l_{\text{cavity}}/2$).

10. The input port can be created mirroring the output port with respect to the XY plane. Select the *Cylinder2* from the History Tree. From the *Draw* toolbar, select **Thru Mirror**.

11. One click in the center of the axis (0,0,0).

12. Hold 'Z' on the keyboard and drag the pointer until you don't reach the point with coordinate $z = -l_{\text{cavity}}/2$. When the cavity input surface, and therefore the desired coordinate, is reached, a big black dot will appear, like in Fig. II.2.33.

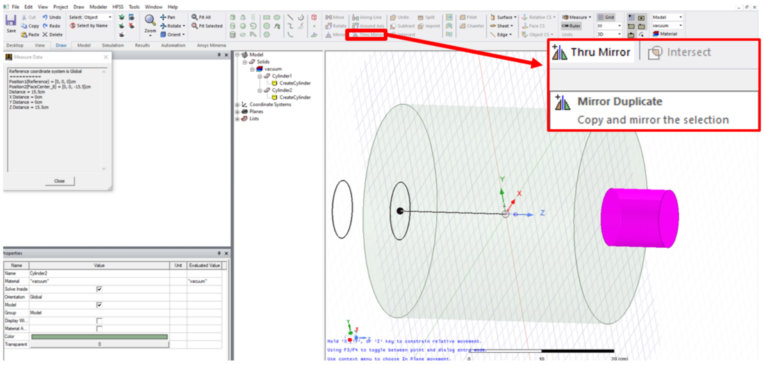


Fig. II.2.33: Mirroring output port.

13. After creating the three cylinders, it is convenient to unite them as one solid. For doing this, select *Cylinder1*, *Cylinder2* and *Cylinder2_1* from the History Tree window and click on **Unite** from the *Draw* toolbar.

14. In order to have a better visualization of the field plot later, it is useful to increase the transparency of the solid. Select the *Cylinder* in the History Tree and modify the **transparency** value to 0.8 from the Properties window (bottom left) - Fig. II.2.34.

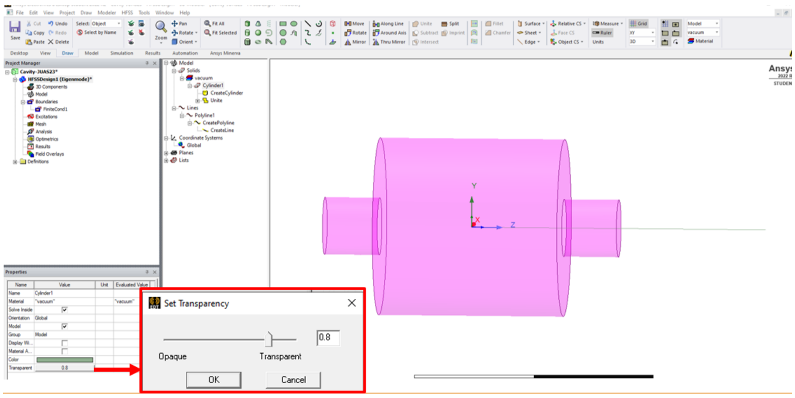


Fig. II.2.34: Increasing the transparency of the created geometry.

II.2.5.3 Eigenmode simulation

In this section, we will show how to set-up an Eigenmode simulation in HFSS. The cylindrical cavity will be simulated considering the case first without losses and then including the finite conductivity of the wall material. The result of the eigenmode simulation gives all eigen-frequencies of the resonant modes, allowed by the geometry and the boundaries. The mode type (TE, TM or TEM) can be identified by looking at the field map for each frequency of resonance. Finally, by adding the losses, given by the finite conductivity of the cavity wall, two figures of merit can be computed: the Q factor and the R/Q.

Simulation set-up

1. From the Project manager window, under the HFSS Design, right click on **Analysis** → **Add solution setup**. A pop-up window will appear (Fig. II.2.35).
2. Under the menu *General*, give the name 'Eigenmode' to the simulation - 1.
3. Set the starting frequency to 1 MHz - 2.
4. We want to see the first three eigenmodes- 3.
5. HFSS does an adaptive mesh refinement, automatically refining the mesh until convergence criterion is satisfied. Maximum number of passes and maximum delta frequency per pass (maximum percentage of error - accuracy indicator) can be defined to guide the analysis. We choose Maximum Number of Passes = 6, Maximum Delta Frequency Per Pass = 10% - 4.

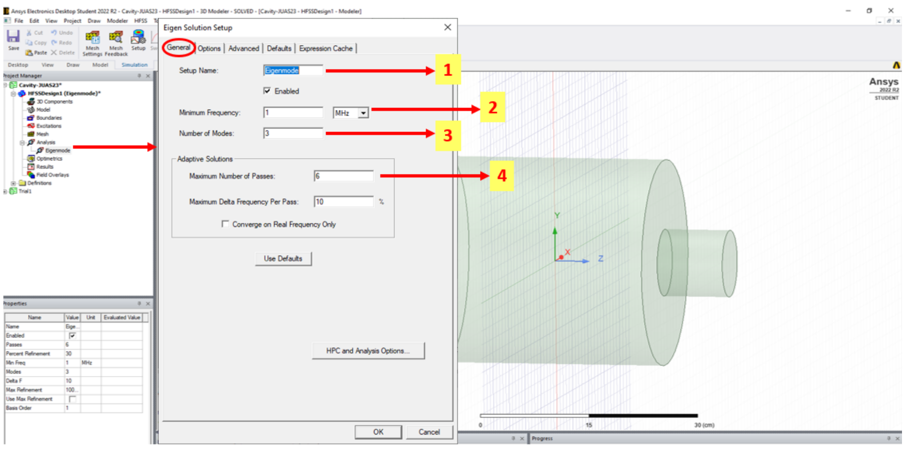


Fig. II.2.35: Setting up Eigenmode simulation.

6. Other settings for the adaptive mesh analysis can be set-up under the menu *Options* (Fig. II.2.36). We select 2 Minimum Converged Passes, the analysis will stop when at least consecutive 2 iterations will stay under the Delta Frequency limit - 5.
7. Select **Second order** under Order of Basis Functions, which is preferable when dealing with curved structures - 6.

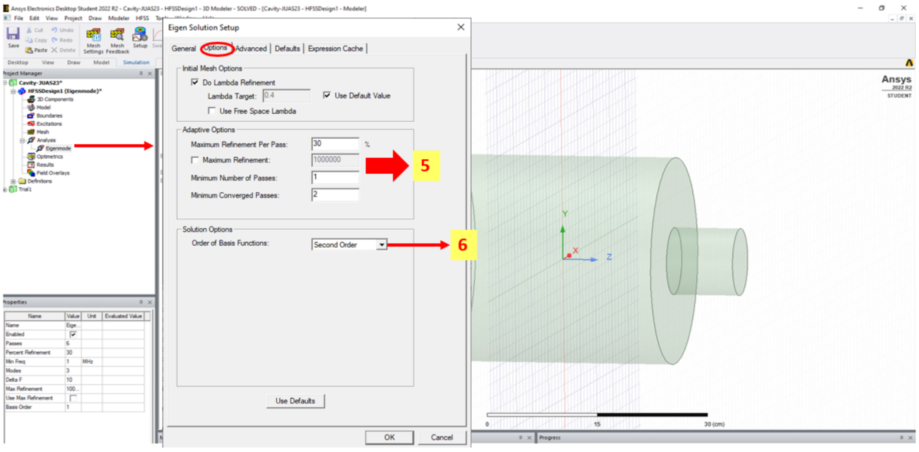


Fig. II.2.36: Refining the adaptive mesh analysis.

8. Other mesh settings can be found under the *Simulation* toolbar (panel on top) → **Mesh settings**. Select 'Apply curvilinear meshing to all curved surfaces'. Higher order (curvilinear) elements will be used to represent the geometry and will give a better resolution than with rectilinear mesh elements - 7.

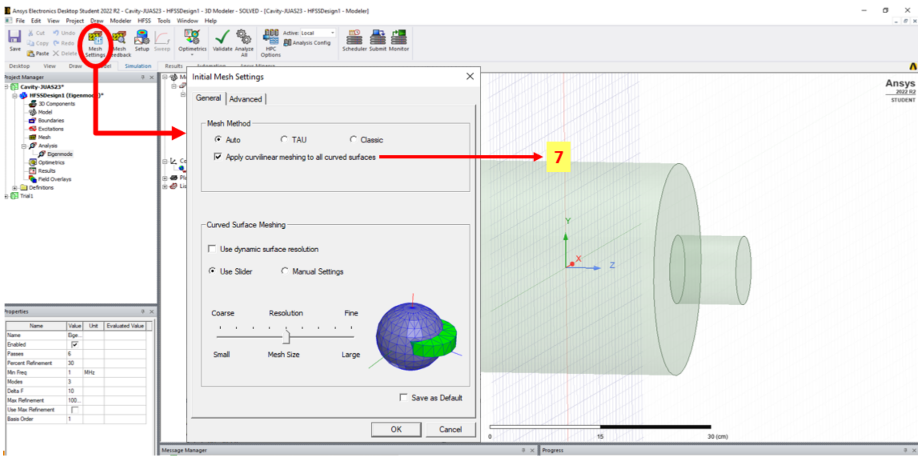


Fig. II.2.37: Using curvilinear meash elements for curved surfaces.

9. Before running the simulation, in HFSS it is possible to check if everything is correct, by clicking on the **Validation** button in the *Simulation* toolbar. Any error would be indicated by a red cross - 8.

10. Finally, if everything is green in the validation process, you can click on **Analyze** to start the simulation - 9.

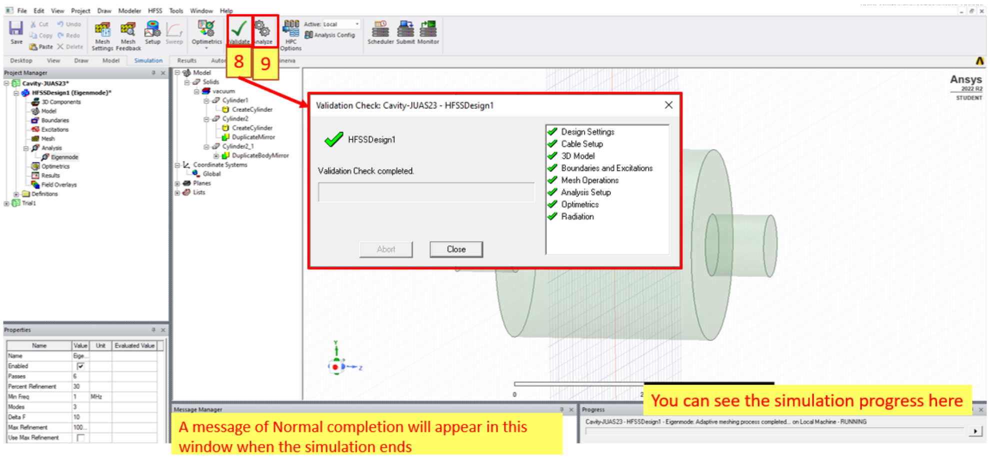


Fig. II.2.38: Validation and run of eigenmode analysis.

II.2.5.3.1 Eigenmode results

After the simulation is finished, the eigen frequencies of the modes can be inspected. Each mode can be recognised, distinguishing between TE, TM and TEM mode, by looking at the field plot.

Eigenmodes results: eigen frequencies and field plots

1. From the *Project Manager* window, under the HFSS Design → **Analysis**, right click on 'Eigenmode' and then select **Convergence**. The frequencies of the first three resonant modes can be read in the Eigenmode Data menu, as in Fig. II.2.39.

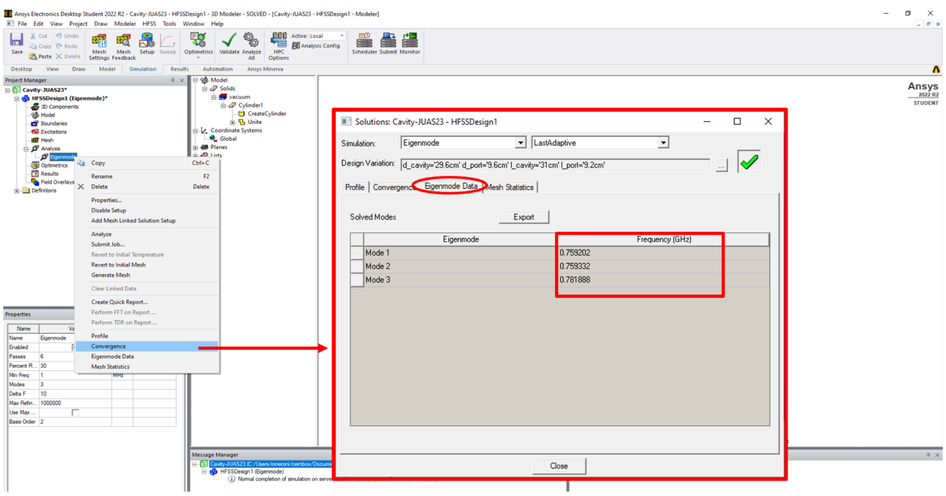


Fig. II.2.39: Eigen frequencies results.

To identify transverse electric (TE) and transverse magnetic (TM) modes, the electric field on the longitudinal plane can be plotted. The TM mode would have a longitudinal (on the z-axis) electric field component, while a TE mode would show a transverse electric field.

2. In order to plot the E-field, the source, and so the correspondent mode we want to see,

must be selected. This can be done in the Project Manager window, right click on Field Overlays → **Edit Sources**. Just one mode per time can be excited. Choose 'Stored Energy' as Excitation Type with magnitude 1 Joule for Mode 1 (Fig. II.2.40).

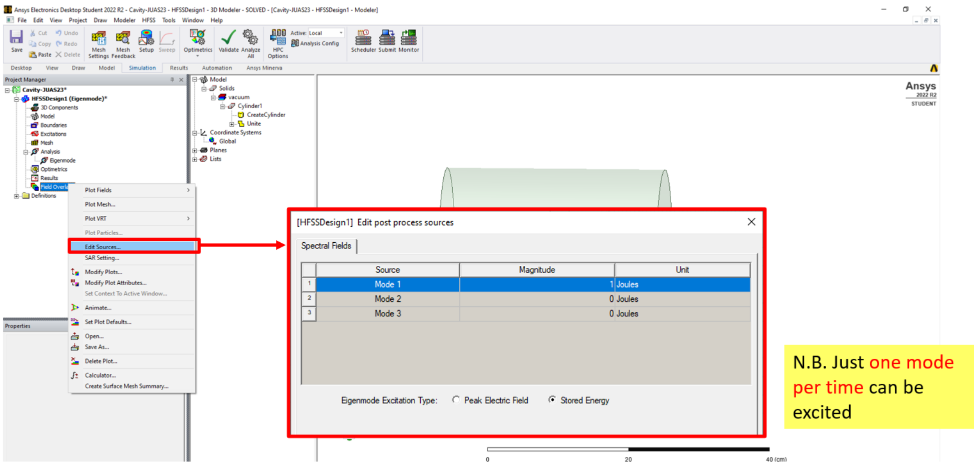


Fig. II.2.40: Selecting Eigenmode excitation source.

3. In the History Tree window, under Planes, select **YZ plane** (Fig. II.2.41) - 1.
4. In the Project Manager window, right click on Field Overlays → **Plot Fields** → **E** → **Vector_E** - 2.

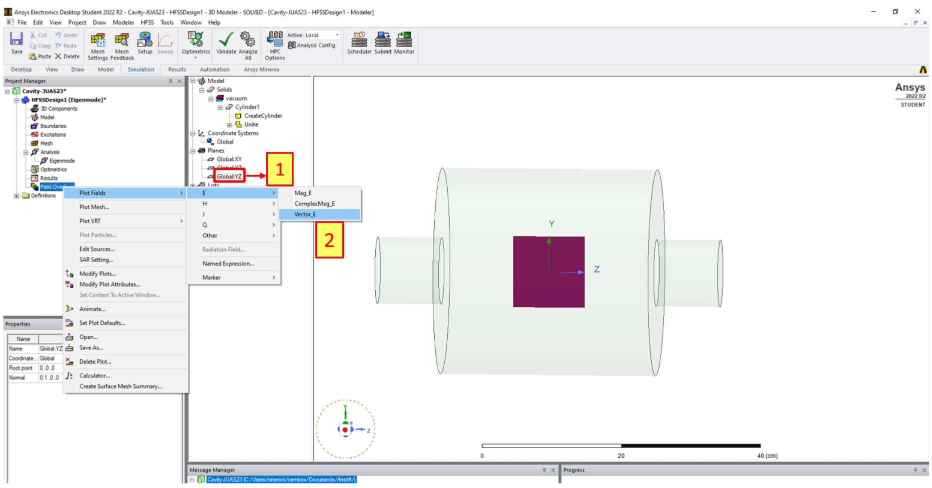


Fig. II.2.41: Setting up the plot of electric field on the longitudinal plane.

5. The instructions above will give, as a result, the plot of the vector electric field for the mode 1 (first eigen frequency) as in Fig. II.2.42. The procedure can be repeated by exciting a different mode (editing the source as in Fig. II.2.40) and looking at the field pattern.

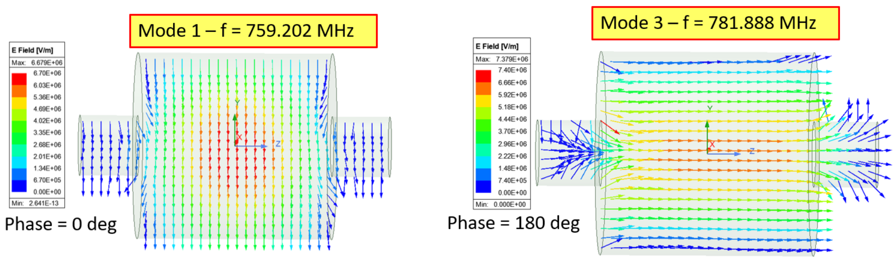


Fig. II.2.42: Plot of electric field on the longitudinal plane (YZ) for mode 1 and 3.

From Fig. II.2.42 the Mode 1 gives a transverse electric mode (TE) because the electric field has no longitudinal component. The same behaviour is given by Mode 2, which has not been plotted. It gives the same TE mode as mode 1, but with a different polarisation. Of much interest is the third eigenmode where the electric field is longitudinal and a TM mode can be recognised. The TM_{010} mode is the mode used for particle acceleration, and it is the fundamental mode of a pillbox cavity. Two important conclusions can be drawn by this analysis:

- The cavity fundamental mode (lowest cut-off frequency) is a TE mode.
- The simulated cylindrical resonator is not a pillbox cavity. If we look at the Mode Chart in Fig. II.2.43, we can see that the TM_{010} mode is the fundamental one only if the cavity is flat enough: ratio $\dfrac{2a}{h} > 1$ (diameter-over-length greater than 1). The mode chart shows what modes can be excited at a given frequency for a given cavity size.

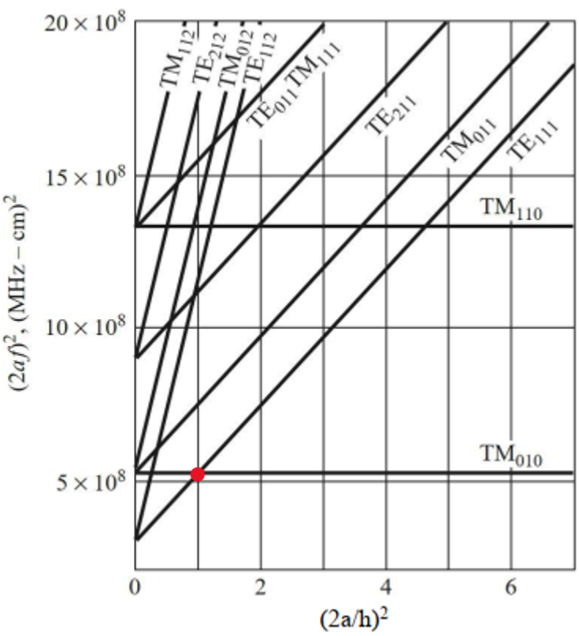


Fig. II.2.43: Mode chart for a cylindrical cavity [3]. On the y-axis is a quantity proportional to the resonant frequency of the cavity, and on the x-axis the cavity diameter/length ratio. The red dot is the working point obtained by using the cavity dimensions defined in section II.2.5.2.

II.2.5.3.2 Inclusion of material finite conductivity: Q factor and 'R-over-Q'

The cylindrical resonator considered so far was ideal, made of perfect conducting wall material. A more realistic case can be reproduced by adding a finite conductivity to the cavity walls. We consider stainless-steel walls in this study case. By including the electrical properties of the material, two figures of merit for the cavity can be computed: the quality factor Q and the *R-over-Q*.

Inclusion of material finite conductivity: Q factor and 'R-over-Q'

1. From the History Tree window, under Model $\rightarrow$ Solids $\rightarrow$ vacuum $\rightarrow$ right click on *Cylinder1* $\rightarrow$ **Select** $\rightarrow$ **All Faces**.

2. Keeping all the faces selected, in the Project Manager window, under the HFSS Design, right click on Boundaries $\rightarrow$ **Assign** $\rightarrow$ **Finite Conductivity**.

3. Input the finite conductivity of Stainless-Steel, $\sigma = 1330000$ S/m -Fig. II.2.44.

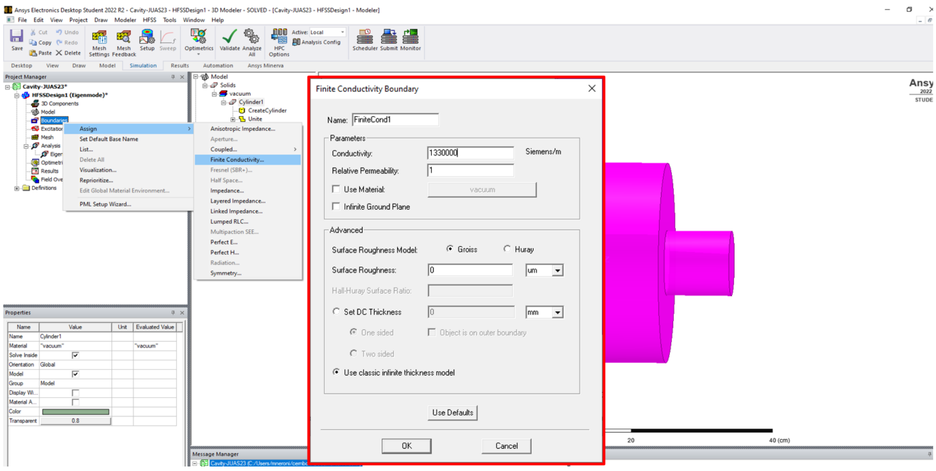


Fig. II.2.44: Assigning electrical conductivity of stainless-steel to the cavity walls.

4. The previous Solution has been invalidated by the boundary change, therefore the Analysis needs to be run again. From the Project Manager window, under the HFSS Design $\rightarrow$ **Analysis**, double-click on 'Eigenmode'.

5. In order to focus on the TM_{010} mode, set a minimum frequency higher than the first two eigen frequencies previously found: Minimum Frequency = 770 MHz.

6. We want to see just one mode: Number of Modes = 1. The rest of the settings remains unchanged.

7. Repeat the **Validation** process and run the simulation as in steps 9 and 10 as in Section II.2.5.3.

8. Find the value of the unloaded Q factor from the Project Manager window, under the HFSS Design $\rightarrow$ **Analysis**, right click on 'Eigenmode' and then select **Eigenmode Data** - Fig. II.2.45. The result is $Q_0 = 6509$.

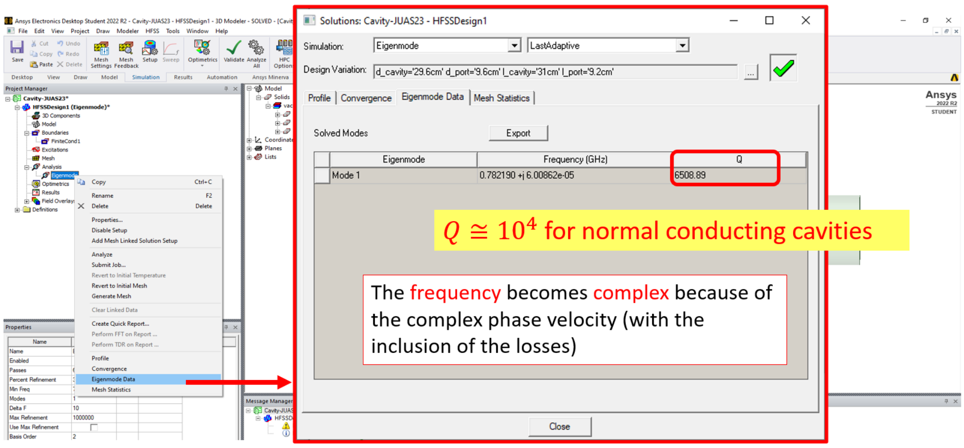


Fig. II.2.45: Result of the unloaded Q factor for the TM_{010} mode.

HFSS does not provide an automatic function to compute the value of R-over-Q. To obtain it, the integral of the electric field along the beam line needs to be computed, therefore a **BeamLine** must be defined. The complete formula has to be manually implemented in the **HFSS Fields Calculator**. The Linac definition of R-over-Q is used here:

$$\frac{R}{Q} = \frac{V_{\text{acc}}^2}{\omega U} = \frac{\left| \int_{\text{beamline}} E_z e^{\frac{j\omega z}{c}} dz \right|^2}{\omega U}. \tag{II.2.13}$$

1. From the Draw toolbar on top, select **Draw line**. The software will ask you if you want to create a non model object, you can select **Yes**.

2. Draw the line along the z-axis. The line should go from $z = -\frac{l_{\text{cavity}}}{2} - l_{\text{port}}$ to $z = \frac{l_{\text{cavity}}}{2} + l_{\text{port}}$. Hold 'Z' on the keyboard to draw the line along z.

3. After drawing the line, right-click on the working plane and select **Done**.

4. In the Project Manager window, right click on Field Overlays → **Calculator** (Fig. II.2.46).

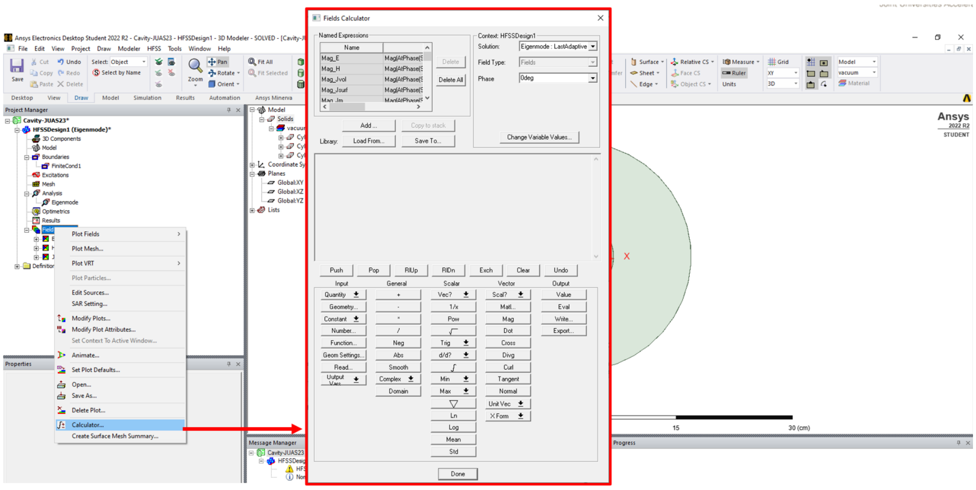


Fig. II.2.46: HFSS Fields Calculator.

The eq. II.2.13 has to be manually defined by making use of the buttons' palette at the bottom of the calculator. More information about the use of the Fields Calculator can be found in the "HFSS Fields Calculator Cookbook" [12].

5. Starting from the numerator, let's define the argument of the exponential in the integral. From the Input column, select **Output Vars** $\rightarrow$ **Freq** (Fig. II.2.47) - 1.

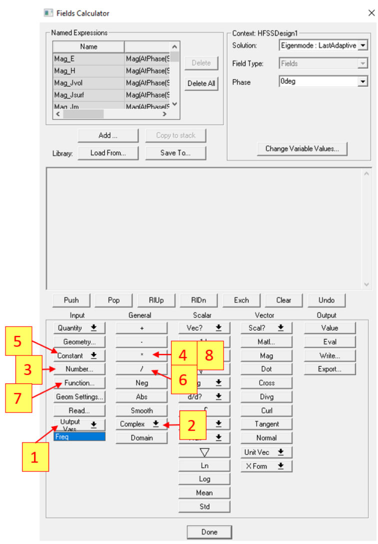


Fig. II.2.47: Definition of the exponential factor of the R-over-Q expression.

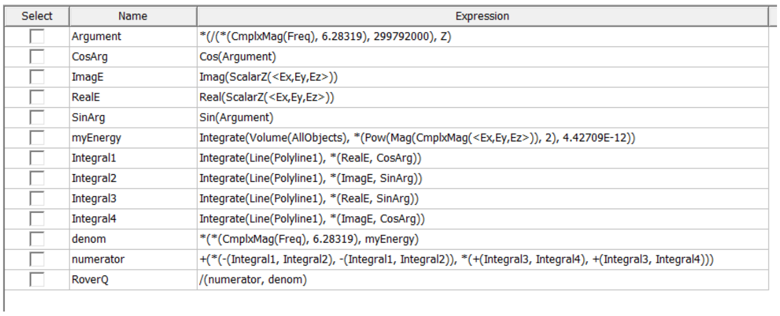
6. From the General column, click on **Complex** $\rightarrow$ **CmplxMag** - 2.

7. From the Input column, select **Number** $\rightarrow$ type 6.28319, equivalent to 2π - 3.

8. Under the General column, click on the multiplication operator "*" - 4.

9. From the Input column, select **Constant** $\rightarrow$ **c** - 5.

10. Under the General column, click on the division operator "/" - 6.
11. From the Input column, select **Function** → Scalar, **Z** - 7.
12. Under the General column, click on the multiplication operator "*" - 8.
13. Save the expression in the Named Expressions stack on the top left by clicking on "**Add**" and giving it a name.
14. Define all the expressions you need to calculate the R-over-Q according to the formula II.2.13, as it is shown in Fig. II.2.48. Note that you can use any saved Named Expressions at the top left by selecting it and clicking on **Copy to stack**. The expression will then appear in the Data Stack in the middle of the calculator.



Select	Name	Expression
	Argument	*(/(*(CmplxMag(Freq), 6.28319), 299792000), Z)
	CosArg	Cos(Argument)
	ImagE	Imag(ScalarZ(<Ex,Ey,Ez>))
	RealE	Real(ScalarZ(<Ex,Ey,Ez>))
	SinArg	Sin(Argument)
	myEnergy	Integrate(Volume(AllObjects), *(Pow(Mag(CmplxMag(<Ex,Ey,Ez>)), 2), 4.42709E-12))
	Integral1	Integrate(Line(Polyline1), *(RealE, CosArg))
	Integral2	Integrate(Line(Polyline1), *(ImagE, SinArg))
	Integral3	Integrate(Line(Polyline1), *(RealE, SinArg))
	Integral4	Integrate(Line(Polyline1), *(ImagE, CosArg))
	denom	*(*(CmplxMag(Freq), 6.28319), myEnergy)
	numerator	+(*(-(Integral1, Integral2), -(Integral1, Integral2)), *(+(Integral3, Integral4), +(Integral3, Integral4)))
	RoverQ	/(numerator, denom)

Fig. II.2.48: Definition of R-over-Q formula.

15. Select **RoverQ** from the Named Expression and click on **Copy to stack**. From Output column in the buttons' palette at the bottom, click on **Eval** to calculate the R-over-Q. The result will give a value of about RoverQ = 26, see Fig. II.2.49.

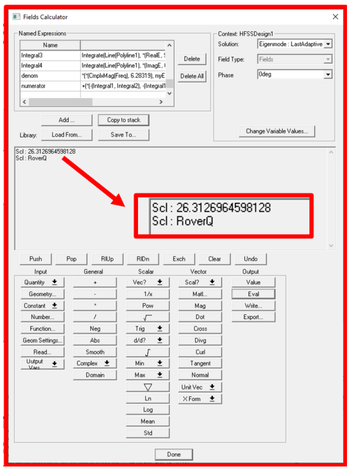


Fig. II.2.49: Evaluation of R-over-Q for the cylindrical cavity.

II.2.6 Introduction to the *Smith* chart

Exercise II.2.6.1: Reflections in the time-domain

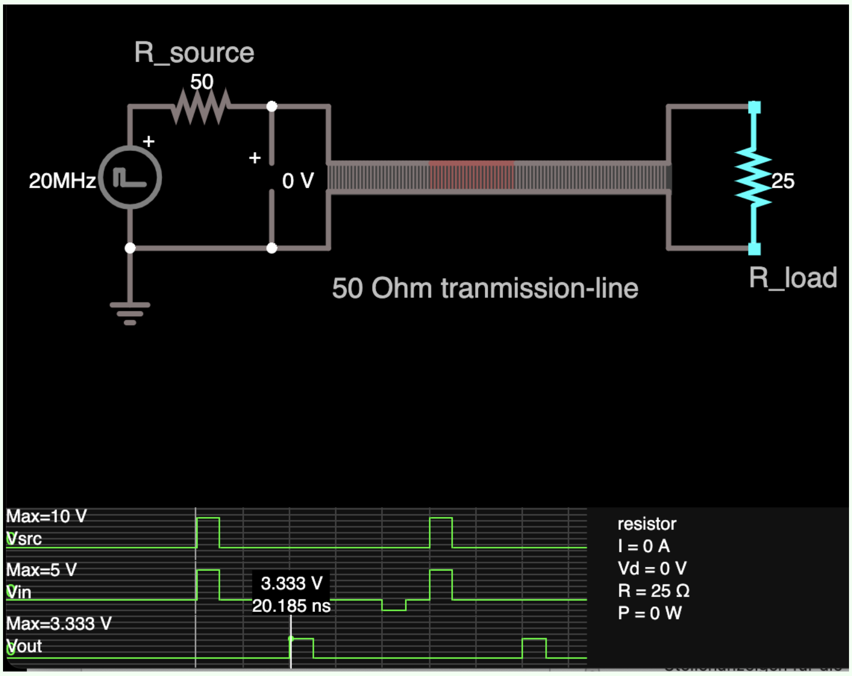


Fig. II.2.50: *CircuitJS1* demo for reflections in the time-domain.

Before starting with the *Smith* chart, let's have a little intuitive exercise on signal reflections – due to a mismatch of impedance's – in the time domain:

1. Go to Paul Falstad's circuit simulator page [13] and set up a simple transmission-line circuit as shown as in Fig. II.2.50. You can either directly modify the interactive webpage (mouse, right click), or download one of the Standalone (offline) versions and setup the circuit schematic. Here we use a $50\,\Omega$ transmission-line of $20\,\mathrm{ns}$ delay time. A circuit node list for this schematic `IdealTL_pulsed_Z050-RL.txt` can also be found on the JUAS InDiCo page, and can be imported to the *CiruitJS1* code under `Open File...`.

2. Select reasonable parameters for the circuit elements, e.g. `A/C Voltage Source:`
`Max Voltage=`$10\,\mathrm{V}$`, Waveform=Pulse, Frequency=`$20\,\mathrm{MHz}$`, Duty Cycle=10;`
`Resistor: R_source=`$50\,\Omega$`;`
`50 Ohm transmission-line: Delay=`$20\,\mathrm{ns}$`, Impedance=`$50\,\Omega$

3. Add a `Voltmeter/Scope Probe` at the beginning of the transmission-line and don't forget to add a `Ground`!

4. Run the circuit and change `R_load`, e.g. to $50\,\Omega$, $25\,\Omega$ and $100\,\Omega$, and observe the reflected signal on the transmission-line.

5. Inspect the values of the signal waveforms after a `STOP` by hovering with the mouse cursor.

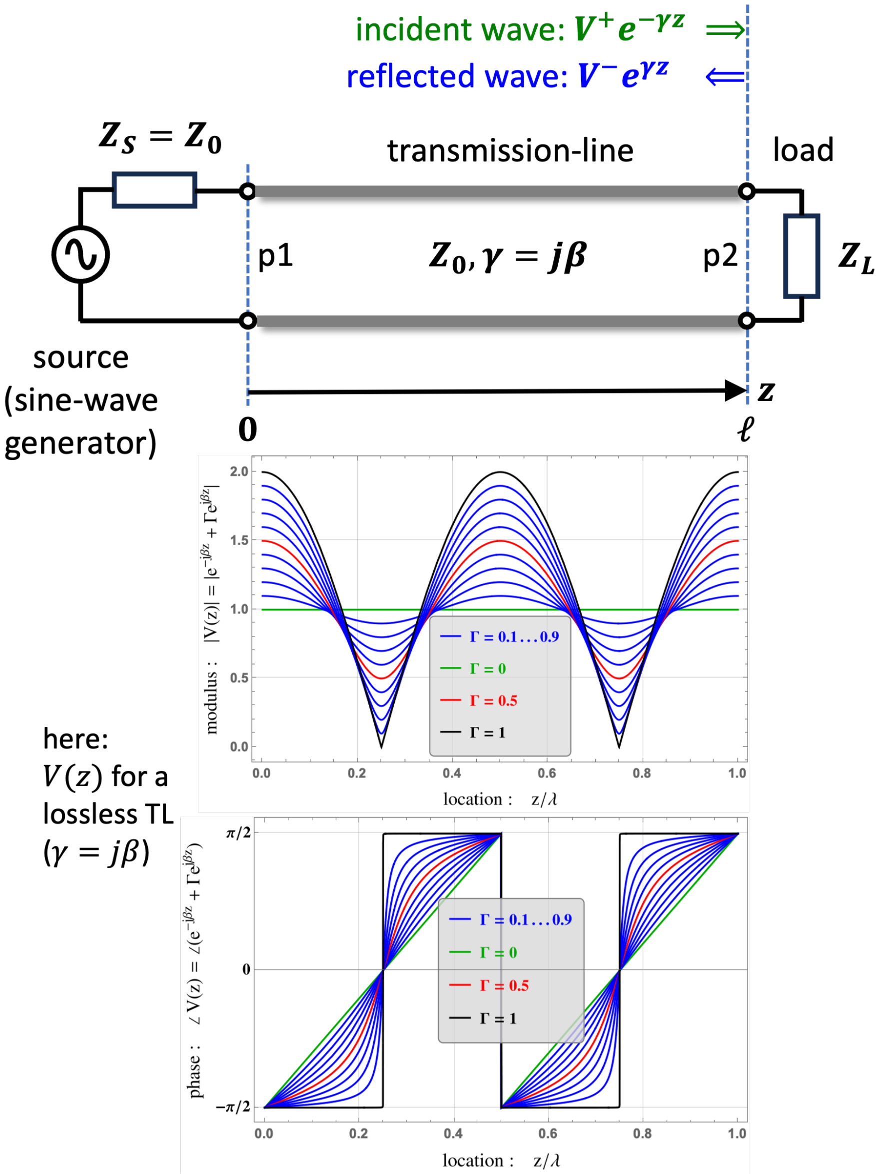


Fig. II.2.51: A loss-less transmission-line setup with a steady-state sine-wave source (upper part) and the corresponding voltage standing waves (lower part).

Figure II.2.51 (upper part) basically shows the same transmission-line schematic as the circuit simulation, Fig. II.2.50, but now with a sine-wave generator as source, indicating we are operating the circuit in steady-state at a frequency $f = \omega/(2\pi)$. In practice the source could also be a power amplifier or any other type of RF power source, and is characterised by its source impedance Z_S, which in most cases is resistive (real value only). The value of the load impedance Z_L in general is complex, and in many cases frequency dependent. In practice the load could be the input impedance of an amplifier or any other RF component, or e.g. the power coupling loop antenna of a cavity resonator. The source at port p1 and the load at port p2 are interconnected by a homogeneous transmission-line, e.g. a coaxial cable or a waveguide of physical length ℓ, having cross-section dimensions and made out of materials that define the characteristic impedance $Z_0 = E_\perp/H_\perp$, which in most cases can be approximated as real value $Z_0 = \sqrt{(R' + j\omega L')/(G' + j\omega C')} \approx \sqrt{L'/C'}$, and the complex propagation constant $\gamma = \alpha + j\beta$.

Example II.2.6.1: When do we apply RF principles?

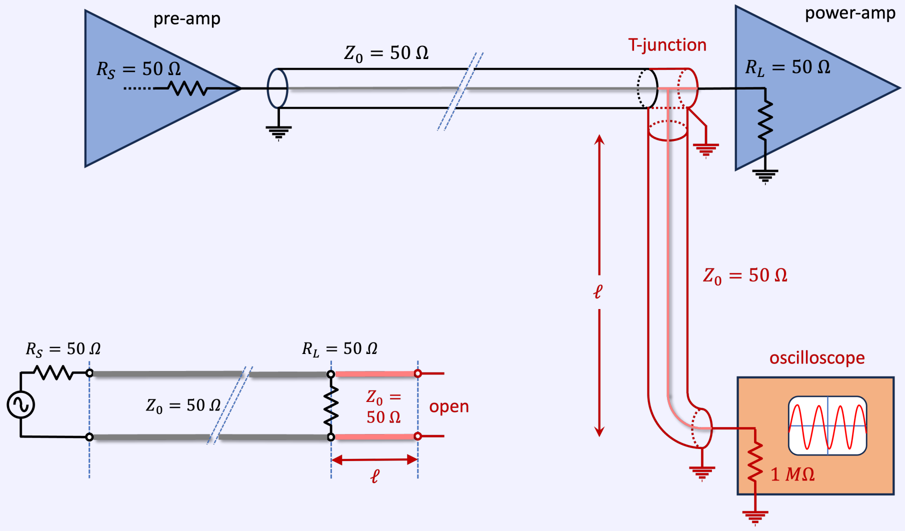


Fig. II.2.52: Setup of two RF amplifiers and an oscilloscope (upper right), and transmission-line equivalent circuit (lower left).

Figure II.2.52 shows the output of a RF pre-amplifier connected to the input of a RF power amplifier with a $50\,\Omega$ coaxial cable. The coaxial transmission-line is terminated at both ends in its characteristic impedance because $R_S = Z_0 = R_L$, therefore no reflections occur.
By adding a *T-junction* between the end of the coaxial cable and the power amplifier input, an *extra coaxial cable* of length ℓ is added, allowing one to observe the signal, e.g. with an oscilloscope. In this example we select a high input impedance ($1\,\mathrm{M\Omega}$) for the instrument to minimise additional power losses.

- Let's assume the extra coaxial cable length is $\ell = 0.5\,\mathrm{m}$ and is of type RG58 ($v_p \simeq 2/3\,c$). Operating the setup with a steady state, continuous wave (CW) sinusoidal signal of frequency $f = 10\,\mathrm{MHz}$, we find $\lambda_g = v_p/f = 20\,\mathrm{m}$, i.e. $40\times$ larger than the length of the cable. At $f = 10\,\mathrm{MHz}$ the extra cable of $\ell = 0.5\,\mathrm{m}$, which is basically *open* at the end, can be approximated as semi-lumped element and acts mainly as a capacitor of value $C \approx \ell/(Z_0 v_p) = 50\,\mathrm{pF}$, see also the equivalent circuit in the lower left of Fig. II.2.52. This additional capacitive effect will almost not alter characteristics of the setup, the signal transmission is basically the same and the reflections are below $10\,\%$.
- **But**, operating the setup at a higher frequency, e.g. $f = 100\,\mathrm{MHz}$ leads to $\lambda_g = v_p/f = 2\,\mathrm{m}$, which gives $\ell = \lambda_g/4$ for the piece of coaxial cable. Now this transmission-line acts as $\lambda/4$-transformer, and will transform the open ($1\,\mathrm{M\Omega}$) impedance at its end to a short at its beginning at the power amplifier input, thus short circuit the signal at R_L!

In the general case $Z_S \neq Z_0 \neq Z_L$ signal reflections will occur at both ports, p1 and p2, due to the impedance mismatches.

> As a *rule of thumb*, if the physical length of the transmission-line is short compared to the guided wavelength, e.g. $\ell \lesssim \lambda_g/10$, we often don't need to care about those reflection effects, as they are small and the impact on the signal transmission is minimum (see also Example II.2.6.1).

In the following however, we assume the length of the transmission-line to be large compared to the guide wavelength, $\ell > \lambda_g/10$, and we will only consider reflections at port p2, $Z_L \neq Z_0$, while the source at port p1 is matched, $Z_S = Z_0$ (see Fig. II.2.51).

II.2.6.1 The complex reflection coefficient Γ

In the steady-state regime of Fig. II.2.51 the E-fields of the forward travelling incident wave $\boldsymbol{E}^{inc}$ and the backward travelling reflected wave $\boldsymbol{E}^{refl}$ perform a superposition along the transmission-line, which lead to *standing voltage waves* in the form

$$v(z) = V^+ e^{-\gamma z} + V^- e^{\gamma z} \tag{II.2.14}$$

Assuming a lossless transmission line, $\gamma = j\beta$, and normalising $V(z) = v(z)/V^+$ follows

$$e^{-j\beta z} + \frac{V^-}{V^+} e^{j\beta z} = V(z,\Gamma) = e^{-j\beta z} + \Gamma e^{j\beta z} \tag{II.2.15}$$

with:

Definition II.2.6.1: Complex reflection coefficient Γ

$$\Gamma = \frac{V^-}{V^+} = \frac{\boldsymbol{E}^{refl}}{\boldsymbol{E}^{inc}} = \frac{b}{a} = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{II.2.16}$$

being the complex *reflection coefficient*.

The lower part in Fig. II.2.51 visualises Eq. (II.2.15), plotted in modulus and phase of $V(z,\Gamma)$ for $\Gamma = 0\ldots1$, along $z/\lambda = 0\ldots1$. Equation (II.2.16) gives the definition of the reflection coefficient Γ as ratio between the reflected wave b and the incident wave a, and in terms of the impedance mismatch between the reference impedance of the system Z_0 and a load impedance Z_L.
Often the modulus of the reflection coefficient, $|\Gamma|$, is expressed as logarithmic value in dB, which is called the *return loss*

$$RL\,[dB] = -20\log_{10}|\Gamma| = 10\log_{10}\frac{P^+}{P^-} \tag{II.2.17}$$

which is equivalent to the logarithmic ratio between the incident (P^+) and the reflected (P^-) power.

For completeness the *voltage standing wave ratio*, $VSWR$, needs to be mentioned

$$VSWR = \frac{|V_{\text{max}}|}{|V_{\text{min}}|} = \frac{|a|+|b|}{|a|-|b|} = \frac{1+|\Gamma|}{1-|\Gamma|} = \left|\frac{Z_L}{Z_0}\right| \tag{II.2.18}$$

which expresses the ratio between the maximum and minimum of $|V(z,\Gamma)|$, which are

$$|V_{\text{max}}| = |V^+| + |V^-|$$
$$|V_{\text{min}}| = |V^+| - |V^-|$$

In practice, reflections in RF systems are basically unavoidable, and in most cases Γ increases with frequency f. As a rule of thumb, reflections of $\Gamma < 0.1$ in a RF system, are viewed as acceptable, see also Table II.2.8.
It also should be noticed that Eq. (II.2.16) follows an equivalent circuit simplification of the electromagnetic behaviour of guided waves in transmission-lines. Consider two TEM transmission-lines of same characteristic impedance Z_0, but different cross-section dimensions. From an impedance point of view, following Eq. (II.2.16), there should be no reflections, but due to the discontinuity between the two transmission-lines the TEM field characteristic is locally distorted and reflections will occur.

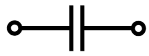
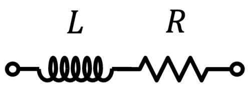
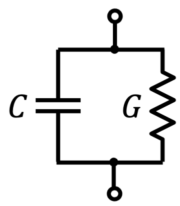
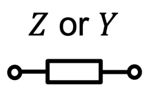
Table II.2.8: The reflection coefficient Γ and related metrics.

Γ	$VSWR$ $= \lvert Z_L/Z_0\rvert$	**Return loss** [dB] $-20\log_{10}\lvert\Gamma\rvert$	**Refl. power** $\lvert\Gamma\rvert^2$	**Inc. power** $1-\lvert\Gamma\rvert^2$
0.0	1.00	∞	0.00	1.00
0.1	1.22	20.0	0.01	0.99
0.2	1.50	14.0	0.04	0.96
0.3	1.87	10.5	0.09	0.91
0.4	2.33	8.0	0.16	0.84
0.5	3.00	6.0	0.35	0.75
0.6	4.00	4.4	0.36	0.64
0.7	5.67	3.1	0.49	0.51
0.8	9.00	1.9	0.64	0.36
0.9	19.00	0.9	0.81	0.19
1.0	∞	0.0	1.00	0.00

II.2.6.2 Reminder: the complex impedance plane

Many electrical, electronics or electromagnetic systems are built using circuit components. Instead of using EM field equations, their function are approximated by ideal lumped or distributed circuit elements, which are defined by the respective currents and voltages in the time or frequency domain. Table II.2.9 lists the most basic lumped circuit elements and their impedance and/or admittance in the frequency domain. Combining many of those circuit elements to a network enable the design of a specific subsystem

Table II.2.9: Symbols for the most basic linear, passive circuit elements.

Admittance	**Circuit symbol**	**Impedance**
conductor, with: **conductance $\boldsymbol{G}$** $[S]$		resistor, with: **resistance $\boldsymbol{R}$** $[\Omega]$
susceptance $\boldsymbol{B_L = 1/\omega L}$ $[S]$		inductor, with: **inductance $\boldsymbol{L}$** $[H]$ **reactance $\boldsymbol{X_L = \omega L}$** $[\Omega]$
capacitor, with: **capacitance $\boldsymbol{C}$** $[F]$ **susceptance $\boldsymbol{B_C = \omega C}$** $[S]$		**reactance $\boldsymbol{X_C = -1/\omega C}$** $[\Omega]$
		complex impedance example: $\boldsymbol{Z = R + j\omega L}$ $[\Omega]$
complex admittance example: $\boldsymbol{Y = G + j\omega C}$ $[S]$		
complex admittance Y		complex impedance Z

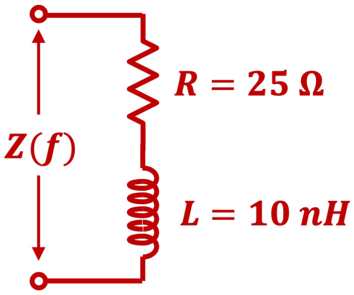
or more complex component with well defined characteristics, e.g., a RF filter, an amplifier, etc.

Any complex impedance Z or admittance Y can be based on a combination of two linear, passive components, a reactive (inductive of capacitive) element and a resistive (conductive) element. Resistance R, impedance Z and reactance X are inverse proportional to conductance G, admittance Y and susceptance B:

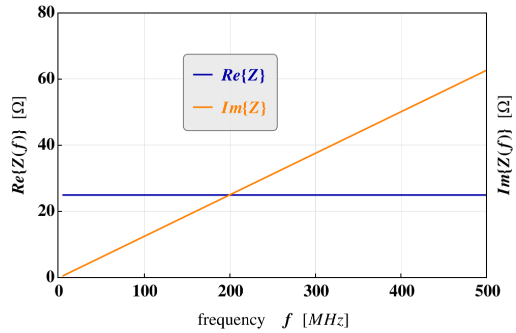
$$R = \frac{1}{G} \quad \text{(II.2.19)} \qquad Z = \frac{1}{Y} \quad \text{(II.2.20)} \qquad X = -\frac{1}{B} \quad \text{(II.2.21)}$$

The characteristic of a complex impedance (or admittance) can be visualised in different ways. Figure II.2.53 illustrates how a RL-series circuit (Fig. II.2.53a) behaves vs. frequency, which always requires two graphs, i.e. magnitude (Fig. II.2.53c) and phase (Fig. II.2.53d) or real / imaginary part (Fig. II.2.53b).

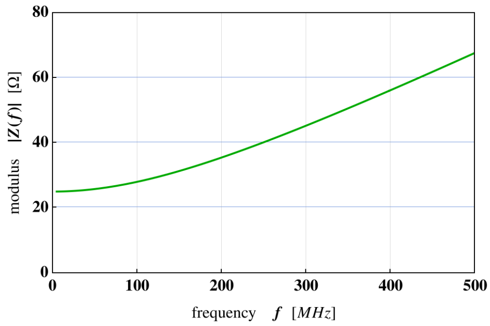
Alternatively the behaviour of $Z(f)$ can be visualised as single trace in the complex *impedance plane*, with the frequency f as parameter. Figure II.2.54 show the parametric plot of $Z(f)$ for the RL-series circuit example (Fig. II.2.53a) in the frequency range $5\,\text{MHz} < f < 500\,\text{MHz}$.



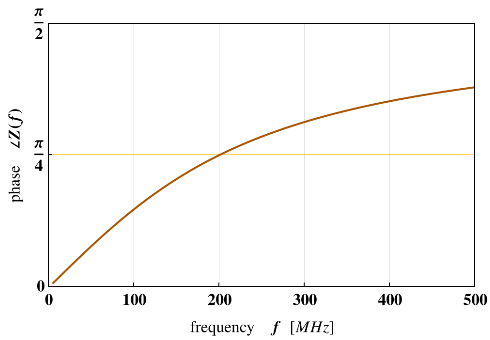
(a) Equivalent circuit of a complex impedance.



(b) $\mathrm{Re}\{Z\}$ and $\mathrm{Im}\{Z\}$ vs. f



(c) Modulus (magnitude) $|Z|$ vs. f.



(d) Argument (phase) $\angle Z$ vs. f.

Fig. II.2.53: Characteristic of a complex impedance Z vs. frequency f

The complex impedance (or admittance) plane is only defined for $\mathrm{Re}\{Z\} \geq 0$, i.e. in the positive "half-plane", as there is no negative resistance or conductance.

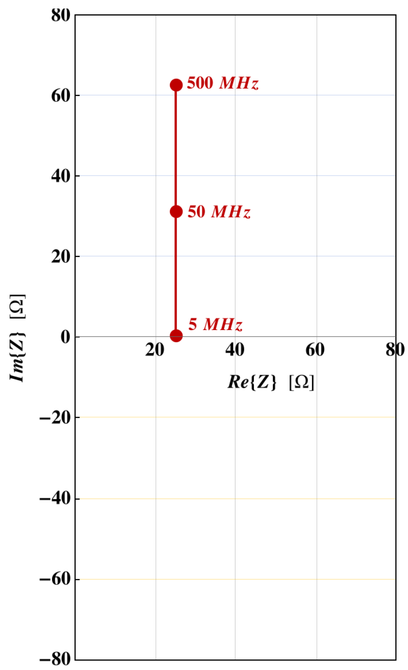


Fig. II.2.54: $Z(f)$ in the complex impedance plane.

II.2.6.3 The *Smith* chart

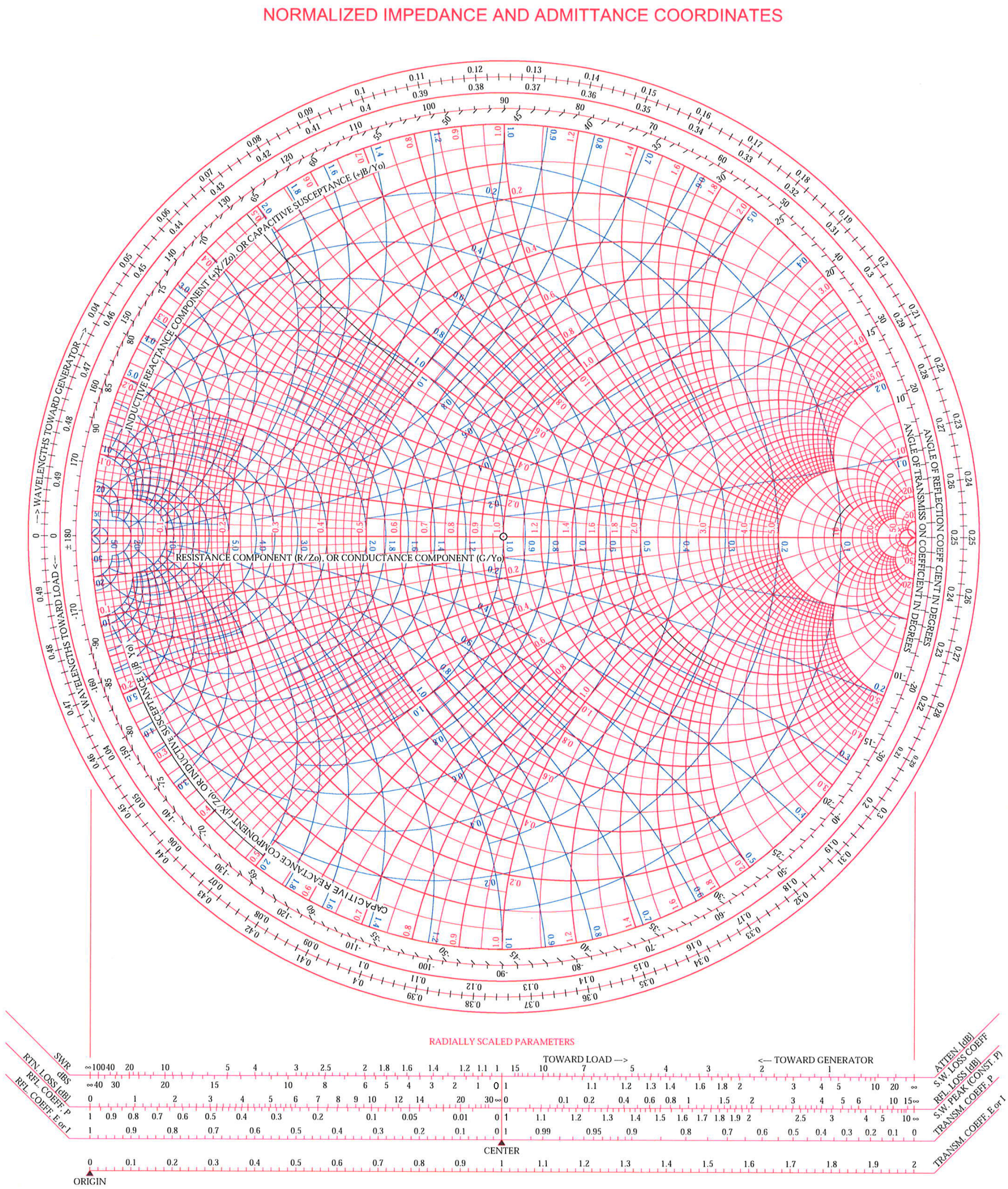


Fig. II.2.55: The *Smith* chart.

After this lengthy introduction we finally can head to the *Smith* chart. Fig. II.2.55 shows the traditional "paper" version of the *Smith* chart, the upper circular part is the *Smith* chart with overlaid *normalised* impedance (in red) and admittance (in blue) planes, the lower part are the "rulers".

Fig. II.2.56 tries to visually explain the concept of the *Smith* chart, which uses a *Möbius* transformation to map the complex impedance plane – actually only the "right" side for $\mathrm{Re}\{Z\} \geq 0$, in

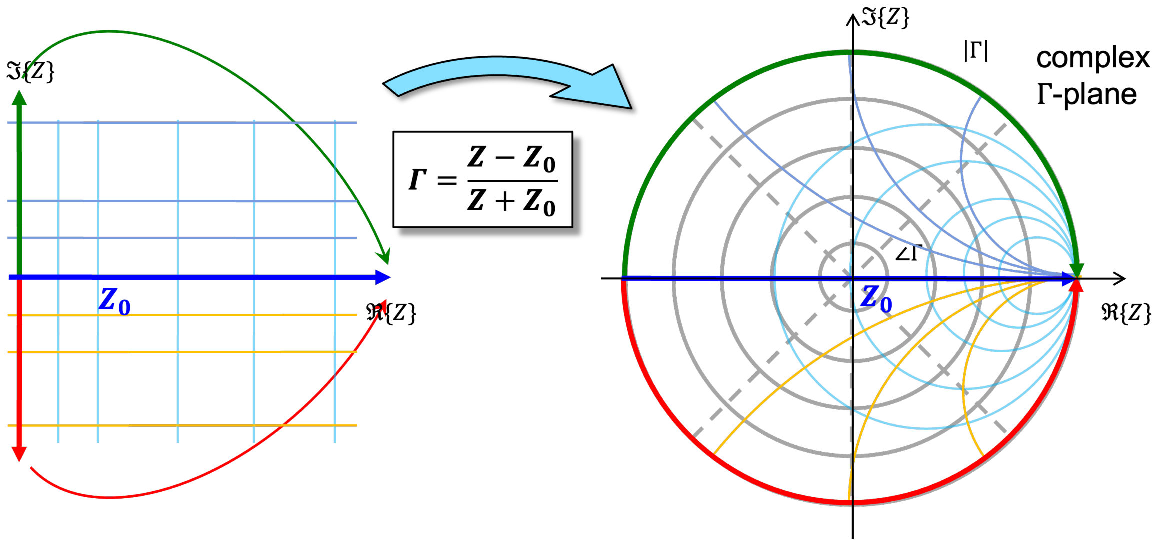


Fig. II.2.56: Mapping the complex impedance plane Z on the plane of the complex reflection coefficient Γ.

Cartesian coordinates – on the plane of the complex reflection coefficient Γ, in Polar coordinates, following Eq. (II.2.16):

$$\Gamma = \frac{Z - Z_0}{Z + Z_0} \tag{II.2.22}$$

Similar to the complex impedance plane Fig. II.2.54, the *Smith* chart is a parametric graph, with the complex impedance Z (or the complex admittance $Y = 1/Z$), linked to the complex reflection coefficient Γ, as variable and the frequency f as parameter. In the paper version of the *Smith* chart the impedance Z is normalised to the reference impedance Z_0:

$$z = \frac{Z}{Z_0} \tag{II.2.23}$$

again, typically to $Z_0 = 50\,\Omega$. For this normalized impedance z the transformation is then:

$$\Gamma = \frac{z-1}{z+1} \quad \Rightarrow \quad \frac{Z}{Z_0} = z = \frac{1+\Gamma}{1-\Gamma} \tag{II.2.24}$$

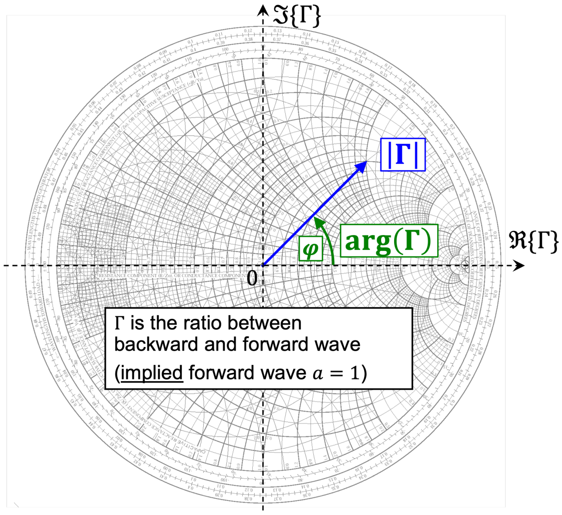
The paper version of the *Smith* chart, Fig. II.2.55, utilises the *normalised* complex impedance $z(f)$, Eq. (II.2.23). Software tools, network analyser displays, etc. typically visualise the un-normalised complex impedance $Z(f)$ in their *Smith* chart applications.

Test your knowledge!

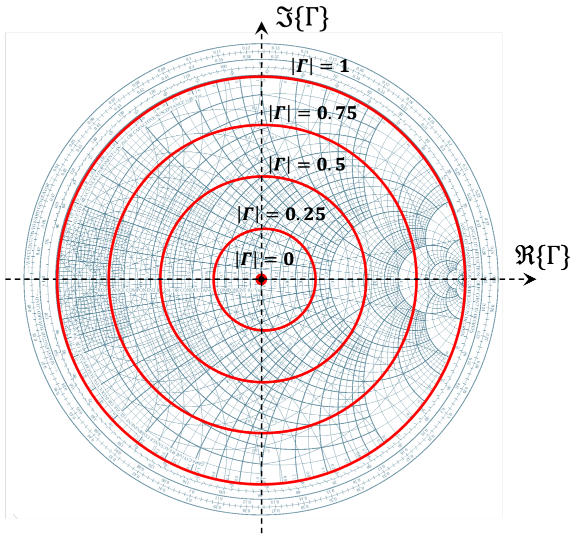
Fill the missing words:

The *Smith* chart transforms the

onto the complex Γ-plane (Reflection coefficient) within the unit circle.



(a) $\Gamma = |\Gamma|e^{j\varphi}$ is expressed in polar coordinates.



(b) $|\Gamma| = const.$ circles.

Fig. II.2.57: $\Gamma(f)$ in the *Smith* chart.

Figure II.2.57 illustrates how the complex reflection coefficient $\Gamma = |\Gamma|e^{j\varphi}$ is mapped into the *Smith* chart using polar coordinates and Fig. II.2.58 indicates "important" points and areas along with their values for Γ and the normalised impedance z.

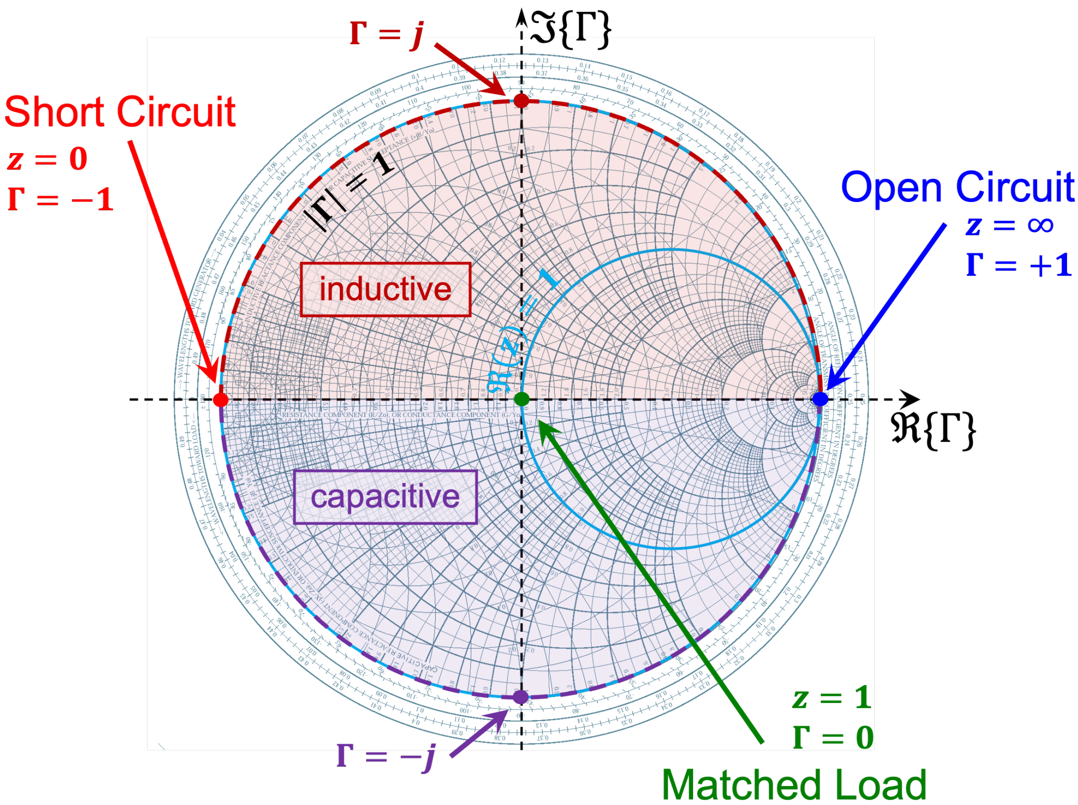


Fig. II.2.58: "Important" points and areas in the *Smith* chart.

Example II.2.6.2: A simple RL-series circuit in the *Smith* chart

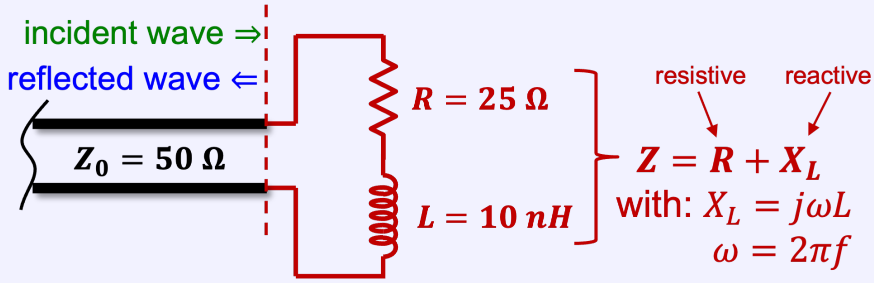
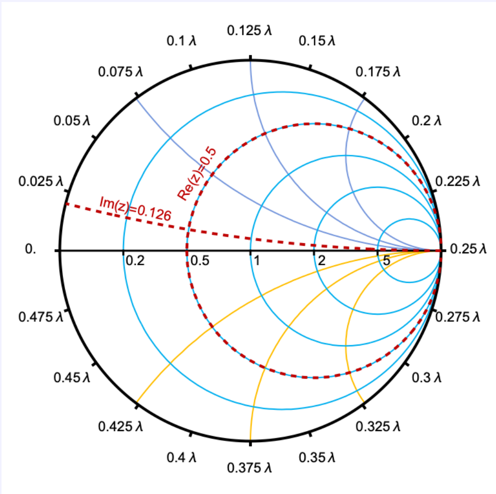


Fig. II.2.59: A $50\,\Omega$ transmission-line terminated with a lumped element RL-series circuit.

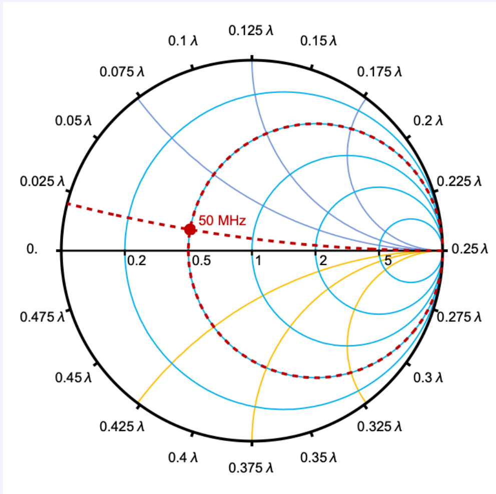
Consider a $50\,\Omega$ transmission-line terminated with the RL-series of Fig. II.2.53a, as shown in Fig. II.2.59. Let's find out how the graphical locus will look in the normalised *Smith* chart in the frequency range $50\,\text{MHz} < f < 500\,\text{MHz}$?!

– Calculate the resistive and reactive values of $Z = R + X_L$, with $R = 25\,\Omega$ and $X_L = j\omega L$ for $f = \omega/(2\pi) = 50\,\text{MHz}$. You should get: $Z \approxeq (25 + j6.28)\,\Omega$.

– Calculate the normalised load impedance $z = Z/Z_0$ for our reference impedance $Z_0 = 50\,\Omega$ given by the characteristic impedance of the transmission-line. You should get: $z \approxeq 0.5 + j0.126$



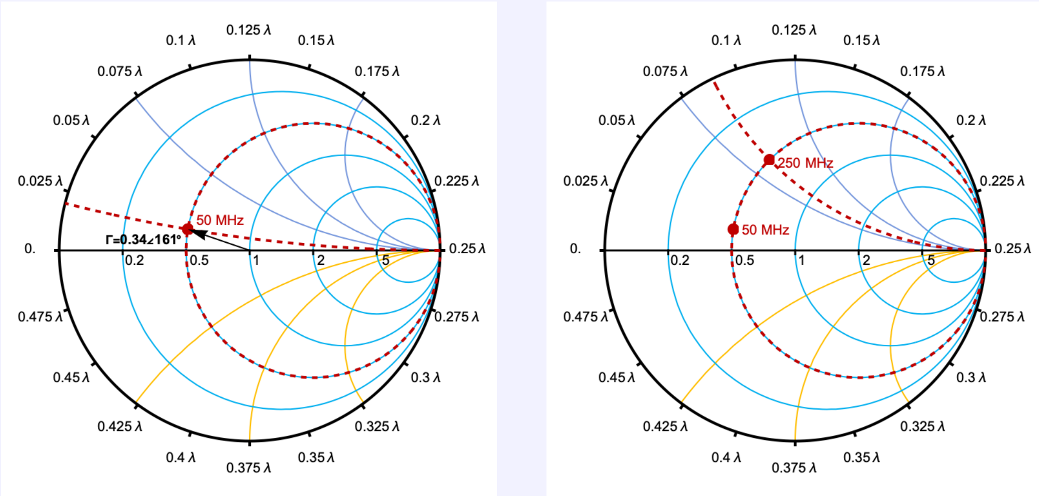
(a) Locate $z \approxeq 0.5 + j0.126\ \ldots$



(b) . . . and mark it as $50\,\text{MHz}$ point.

Fig. II.2.60: Locate the normalised load impedance $z(f = 50\,\text{MHz})$ in the *Smith* chart.

– Print out the *Smith* chart (Fig. II.2.55), locate $\text{Re}(z) = 0.5$ and $\text{Im}(z) = 0.126$ as illustrated in the simplified *Smith* chart in Fig. II.2.60a and mark the $50\,\text{MHz}$ impedance point, see Fig. II.2.60b.

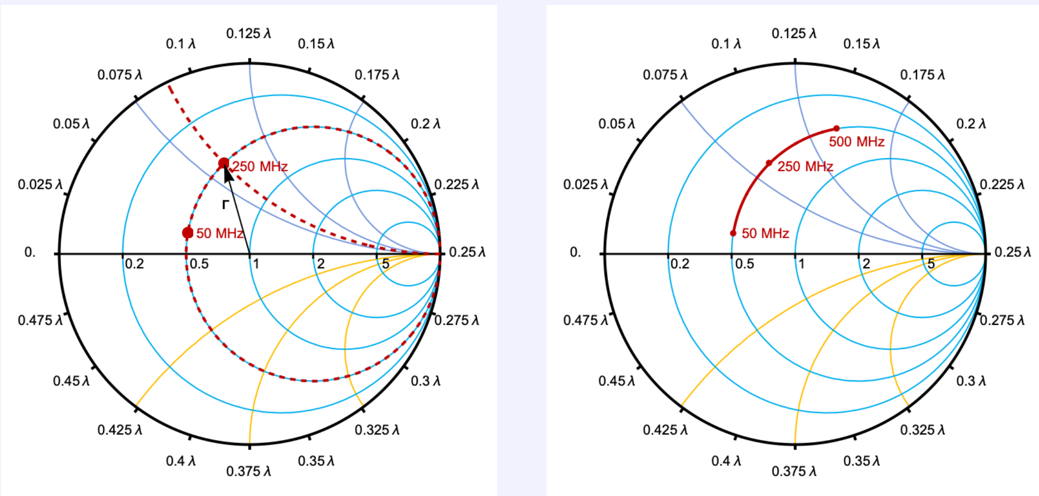


(a) Measure $|\Gamma|$ and $\angle\Gamma$ . . . **(b)** . . . and continue with the $250\,\mathrm{MHz}$ point.

Fig. II.2.61: Evaluate the reflection coefficient Γ.

– With help of a ruler and a pair of compasses evaluate $|\Gamma|$ and $\angle\Gamma$.
You find the angle of Γ by extending a straight line "origin (*zero*) – $z(50\,\mathrm{MHz})$" with a ruler to the outer rim of the *Smith* chart. On the 3rd outer circle read $\angle\Gamma \approxeq 161°$.
Take the length "origin (*zero*) – $z(50\,\mathrm{MHz})$" with the pair of compasses and measure the length equivalent $|\Gamma|$ value at the **RFL. COEFF, E or I** ruler below the *Smith* chart. You should read $|\Gamma| \approxeq 0.34$.

It is important to fully understand this procedure before moving on!

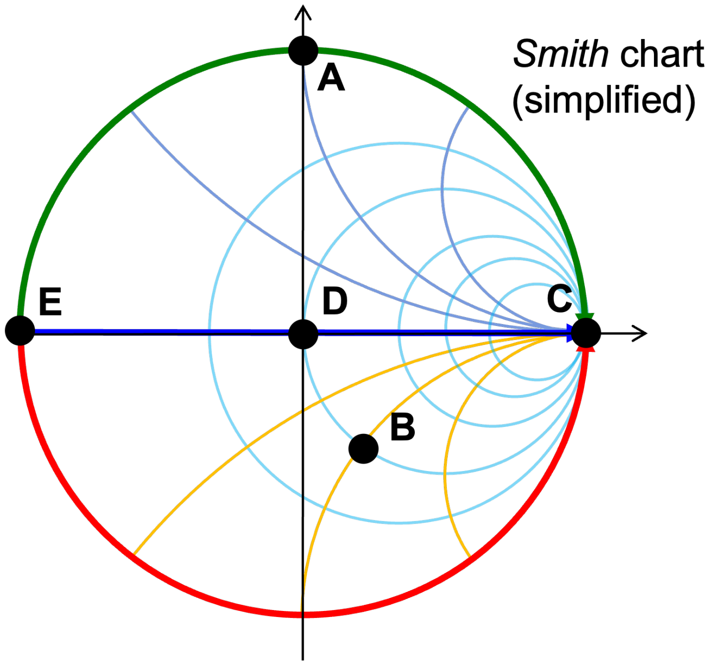


(a) Measuring $r = |\Gamma|$ at $f = 250\,\mathrm{MHz}$. **(b)** The final locus of $z(f)$.

Fig. II.2.62: Finalise the locus with points for z at 50, 250 and $500\,\mathrm{MHz}$.

– Repeat the procedure for $f = 250$ and $500\,\text{MHz}$. With the three points, $z(50\,\text{MHz})$, $z(250\,\text{MHz})$ and $z(500\,\text{MHz})$ marked in the *Smith* chart it is easy to connect them with the appropriate circle: You now have completed the locus of $z(f)$ in the *Smith* chart, Fig. II.2.62b.

Test your knowledge!



Fill the centre column with the correct answer, $1 \ldots 7$

Prompts		Possible Answers
A. Point A		1. $\Gamma = 1,\ z \to \infty$
B. Point B		2. $\Gamma = -j$
C. Point C		3. $\Gamma = 0,\ z = 1$, match
D. Point D		4. Point in the capacitive half plane
E. Point E		5. $\Gamma = +j$
		6. $\Gamma = -1,\ z = 0$
		7. Point in the inductive half plane

Exercise II.2.6.2: A *Smith* chart exercise

The following exercise helps you to better understand the *Smith* chart:

point #	P1	P2	P3	P4	P5
Z/Ω	∞		0		$100 + j100$
Γ		0		$0.7\angle -62°$	

1. Print out the *Smith* chart, Fig. II.2.55.
2. Mark points **P1 ...P5** of the table above into the *Smith* chart, assuming a reference impedance of $Z_0 = 50\,\Omega$.
3. Fill the missing Z and Γ in the empty fields of the table using the *Smith* chart!

This exercise was a part of an exam in 2016.

II.2.6.4 Impedance matching with the *Smith* chart

In accelerator RF, as well as in any RF engineering domain, *impedance matching* is one of the most common challenges, i.e. match a given complex load impedance Z_L (e.g. the input impedance of an amplifier, filter network, etc.) to the RF system reference impedance Z_0, which typically is $50\,\Omega$. In the past the *Smith* chart was the preferred impedance matching tool, but today it is replaced by dedicated software tools.

The visualisation freeware *Dellsperger Smith* [14] fills the gap between the traditional paper *Smith* chart (Fig. II.2.55) and commercial RF software tools for simple impedance matching problems and education. The software applies closed-form analytical expressions and the mapping between the complex impedance plane $Z(f)$ of the complex admittance plane $Y(f) = 1/Z(f)$ and the complex reflection coefficient $\Gamma(f)$, Eq. (II.2.22), for a given reference impedance Z_0.

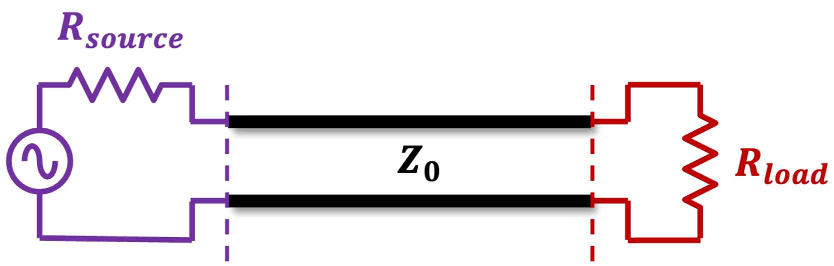
An important goal for the transfer of RF energy between a source, e.g. a transmitter, power source, amplifier, etc., and a load, like an RF cavity or accelerating structure, is to minimise unwanted reflections, thus maximising the power transfer. The maximum RF power is transferred from a transmission line with a reference characteristic impedance Z_0 to a load of complex impedance Z_L if

$$Z_L = Z = Z_0, \tag{II.2.25}$$

and from Eq. (II.2.22) follows $\Gamma = 0$, i.e. no reflections!

Unfortunately, in many practical situations, $Z_L = Z \neq Z_0$. Also, please note that in most RF systems, the reference impedance Z_0 is based on the characteristic impedance of coaxial cable transmission-lines, with a real value of $Z_0 = 50\,\Omega$. In most cases where $Z_L \neq 50\,\Omega$, an impedance-matching network has to be designed to minimise the unwanted reflections and maximise the power transfer. While there are many ways to solve an impedance-matching problem, the *Smith* chart has proven to be the most efficient, and the software-supported *Dellsperger Smith* software makes the impedance-matching particular handy.

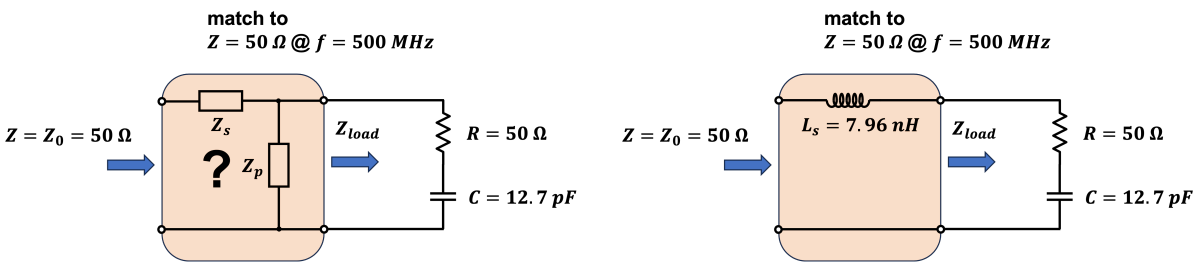
Test your knowledge!



When do no signal reflections occur at the end of the transmission-line?
(Mark all correct answers)

- ☐ $R_{source} = R_{load}$
- ☐ $R_{source} = Z_0$
- ☐ $Z_0 = R_{load}$
- ☐ $R_{source} = Z_0 = R_{load}$

II.2.6.4.1 A simple impedance-matching problem



(a) General lumped element matching network.
(b) Solution with an inductance.

Fig. II.2.63: A simple impedance matching problem.

Consider a RF system operating at a frequency $f = 500\,\text{MHz}$ on a load impedance

$$Z_{load} = R + \frac{1}{j\omega C} = 50\,\Omega + \frac{1}{j2\pi \times 500\,\text{MHz} \times 12.7\,\text{pF}} \simeq (50 - j25)\,\Omega \tag{II.2.26}$$

see also Fig. II.2.63. Following Eq. (II.2.22), this results in a reflection coefficient $\Gamma \simeq 0.242e^{-j76°}$. Figure II.2.63a indicates a general lumped element matching-network, Fig. II.2.63b the solution with

$Z_p \to \infty$ and an inductance $Z_s = L_s = 7.96\,\text{nH}$, which is the dual network of the capacitance C:

$$\omega L_s = \frac{1}{\omega C} \tag{II.2.27}$$

II.2.6.4.2 Impedance-matching exercises with Dellsperger Smith

The *Dellsperger Smith* software [14] is a freeware which runs only under the *MS Windows* operating system.

Perform the following steps and exercises:

Exercise II.2.6.3: *Dellsperger Smith* impedance matching exercise

1. Start the Smith V4.1 tool by *Fritz Dellsperger*.
2. Under *Tools* $\to$ *Settings*, turn off the blue *Y-plane* for this exercise. Leave all other settings at default.

   In this first exercise, use R, L and C lumped elements in *series* to match the load impedances at $f = 500\,\text{MHz}$ to $Z_0 = 50\,\Omega$. It requires only two components for each matching circuit.
3. Enter Z_L as first data point in the *Smith* chart by clicking on the “Keyboard” tab, which opens the “Data Point" panel. Here you may have to switch from “polar" to “cartesian" input coordinates to insert the “Re” and “Im" values of Z_L.
4. You will need only the *Z-plane* for this exercise, the *Y-plane* can be turned off in the “Settings” dialog.
5. Make use of “Undo”, eventually multiple times, if something goes wrong, even to return to the start.
6. Please complete the table:

Z_L	C Series	L Series	R Series
$(50 + j25)\,\Omega$			
$(50 - j25)\,\Omega$	–	$7.96\,\text{nH}$	–
$(4 + j21)\,\Omega$			
$(20 - j50)\,\Omega$			

7. Now use R, L and C lumped elements in *parallel* to match the load impedances at $f = 500\,\text{MHz}$ for $Z_0 = 50\,\Omega$. Again, it takes only two components for each matching circuit.

You will only need the *Z-plane* for this exercise!

Z_L	C Shunt	L Shunt	R Shunt
$(50 + j25)\,\Omega$			
$(50 - j25)\,\Omega$			
$(4 + j21)\,\Omega$			
$(20 - j50)\,\Omega$			

8. Match impedances at $f = 500\,\mathrm{MHz}$ to $Z_0 = 50\,\Omega$. Use only 2 *reactive* components, in series or in parallel, to create a *lossless* matching circuit.

   Hint: You will need both, the *Z-plane* and the *Y-plane* for this exercise!
   There are multiple solutions possible for this exercise!

Z_L	C Series	L Series	C Shunt	L Shunt
$(32 - j66)\,\Omega$				
$(13 - j9)\,\Omega$				
$(37 + j34)\,\Omega$				
$(78 + j78)\,\Omega$				

9. The input impedance $Z_L = (17 - j18)\,\Omega$ of an amplifier shall be matched to $Z_0 = 50\,\Omega$ at $f = 500\,\mathrm{MHz}$. Use only two coaxial lines with the characteristic impedances of $50\,\Omega$ and $25\,\Omega$ in a series configuration.

   What are the electrical length's of the two lines? (multiple solutions possible!)

Figure II.2.64 and Fig. II.2.65 show screenshots of the *Dellsperger Smith* software taken from Exercise II.2.6.3 #6 and #9, respectively.

Please note, the paper version of the *Smith* chart operates always with *normalised* impedances, Eq. (II.2.23), while the *Dellsperger Smith* software displays the absolute impedance Z, scaled by Z_0 given under *Tools* $\rightarrow$ *Settings* $\rightarrow$ *Default Z0*.

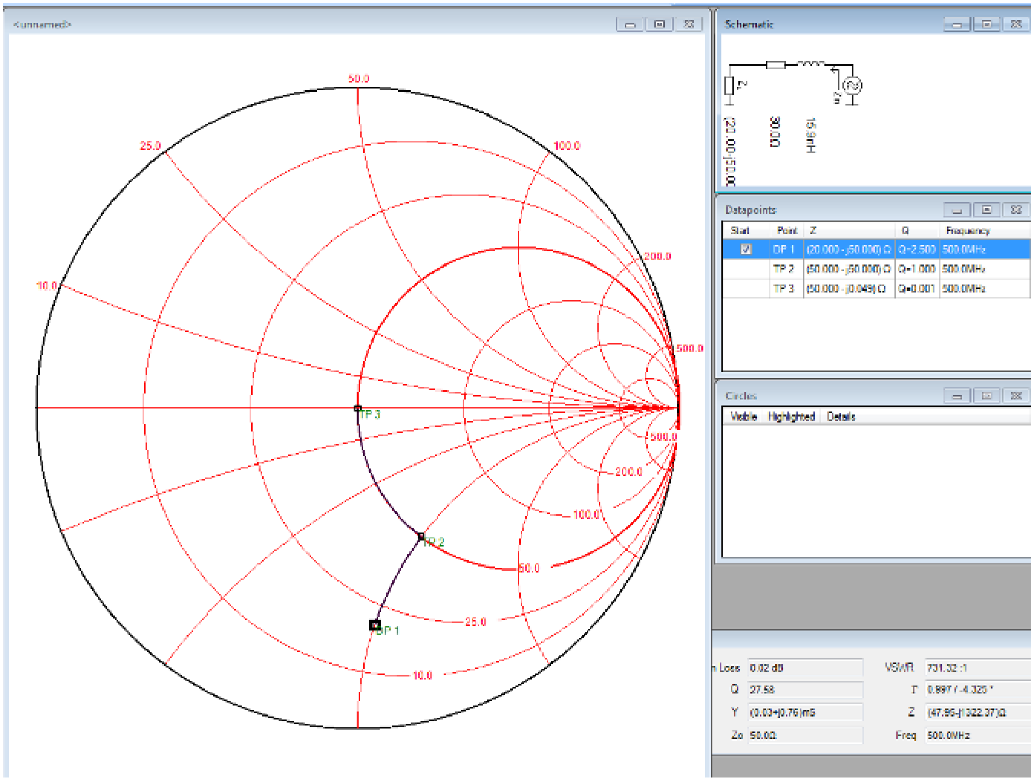


Fig. II.2.64: Screenshot impedance matching Exercise II.2.6.3 #6.

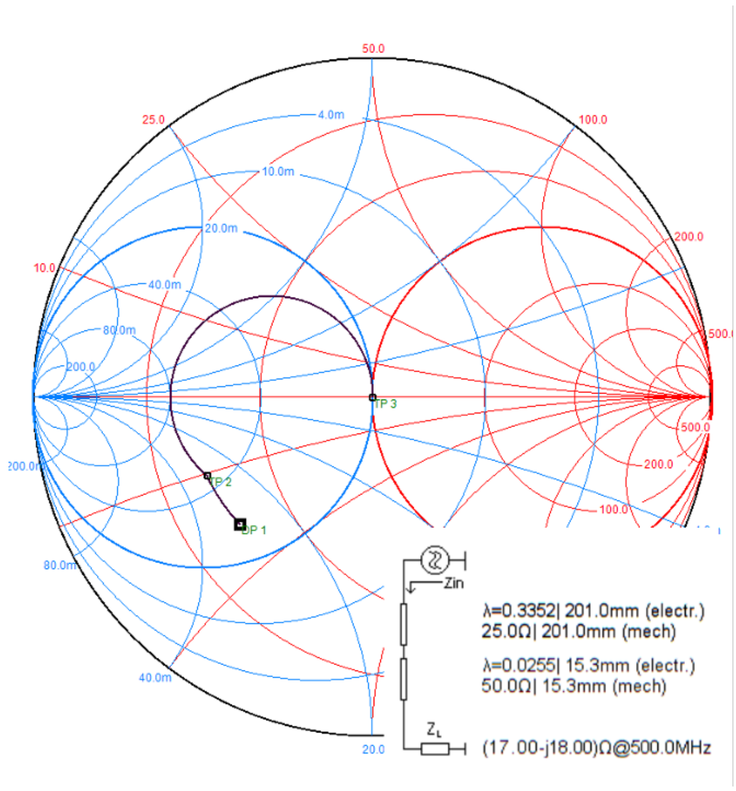


Fig. II.2.65: Screenshot impedance matching Exercise II.2.6.3 #9.

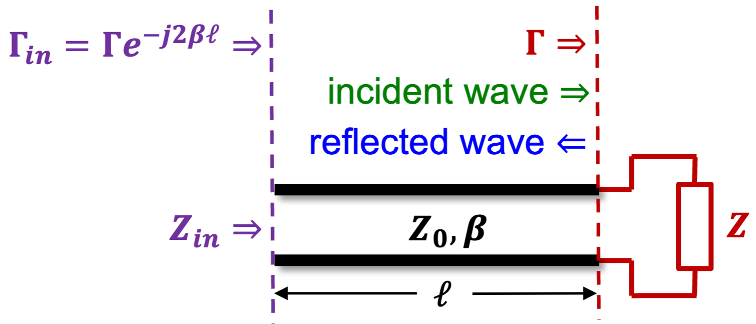


Fig. II.2.66: A transmission-line as impedance transformer.

II.2.6.4.3 *The transmission-line impedance transformer*

Clearly, Exercise II.2.6.3 #9 addresses the more advanced RF engineer, but still demonstrates the use of lossless transmission-lines operating as an impedance transformer. Fig. II.2.66 illustrates the concept, using a lossless transmission-line to add a phase delay of $2\beta\ell$ to an existing load impedance Z. Consequently, the reflection coefficient at the input of the transmission-line:

$$\Gamma_{\text{in}} = \Gamma e^{-j2\beta\ell} \tag{II.2.28}$$

$$\text{with: } \beta\ell = \theta \text{ phase delay (electrical length), and: } \beta = \frac{2\pi}{\lambda_g} = k \text{ wave number}$$

relates to the reflection coefficient Γ at the load end of the line. This results in a transformation of the load impedance Z to a new, different impedance Z_{in} at the input of the line. The *Smith* chart offers an effective, simple graphical way to calculate this transmission-line based impedance transformation.

Exercise II.2.6.4: Transmission-line impedance transformation

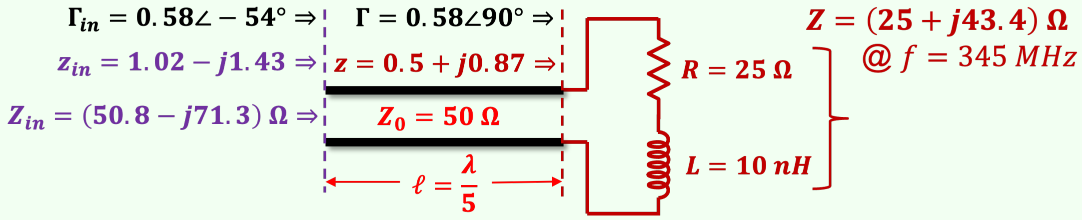


Fig. II.2.67: A $\lambda/5$ impedance transformation.

Figure II.2.67 shows our RL-series impedance from Example II.2.6.2, now operating at a frequency $f = 345\,\text{MHz}$ and extended with a $50\,\Omega$ transmission-line of length $\ell = \lambda/5$. While the results of the associated impedance transformation are already given, you should perform the exercise and verify the result, either using the paper *Smith* chart, Fig. II.2.55, or the *Dellsperger Smith* software.

Using the paper *Smith* chart requires to use the normalised load impedance $z = Z/Z_0 = 0.5 + j0.87$, with the *Dellsperger Smith* you can directly go with $Z = (25 + j43.4)\Omega$. Key element of the impedance transformation is to add the transmission-line of length $\ell = \lambda/5$ by rotating Γ

clockwise by $2\beta\ell = 4\pi/5 \equiv 144°$, i.e. from $0.125\,\lambda$ to $0.325\,\lambda$ in the *Smith* chart, as shown in Fig. II.2.68.

Complete the exercise by sketching the lumped element equivalent of Z_{in} and determining the values, and by calculating the physical length ℓ, assuming the transmission-line is a coaxial cable with PTFE (*Teflon*) as insulator (dielectric constant $\varepsilon_r = 2.1$).

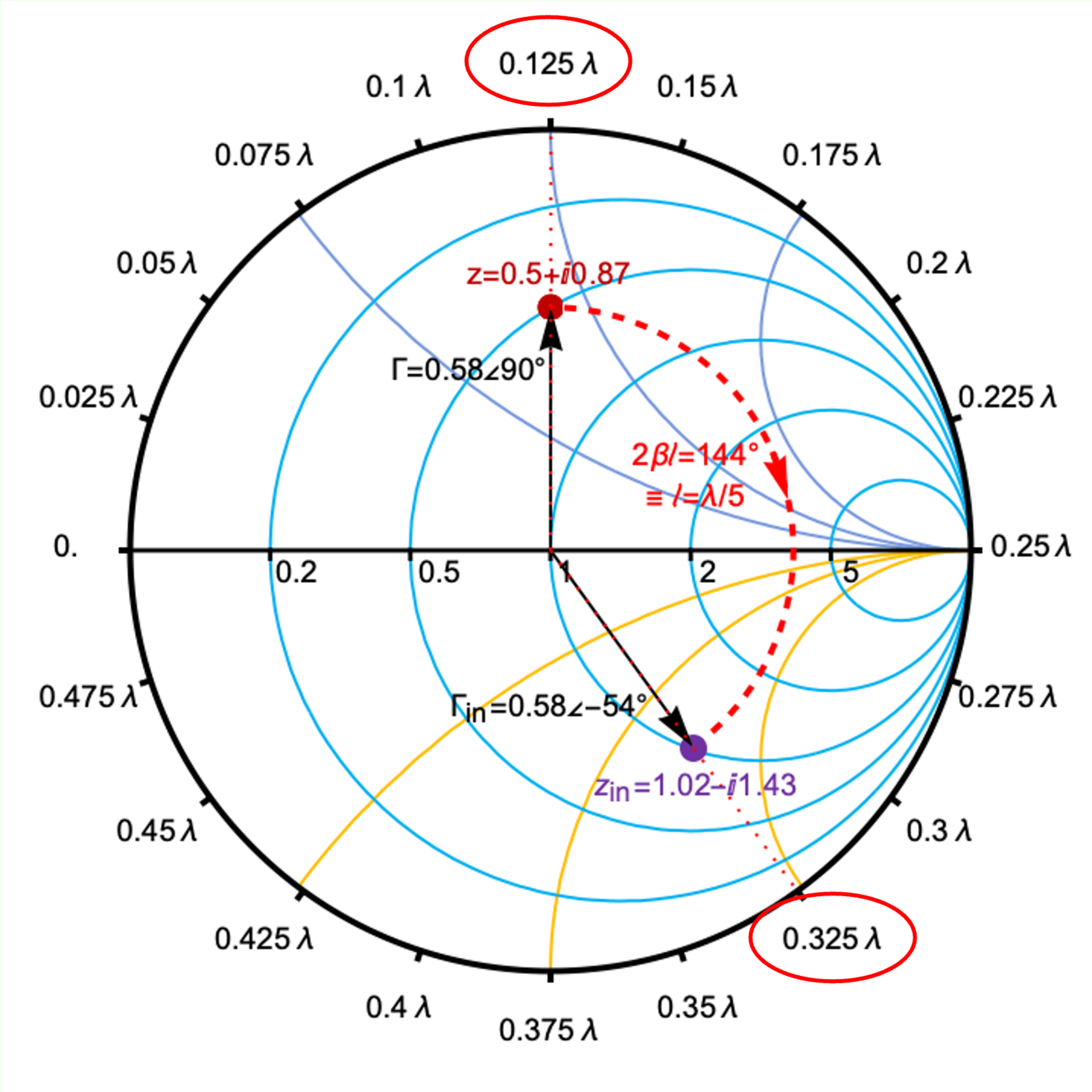


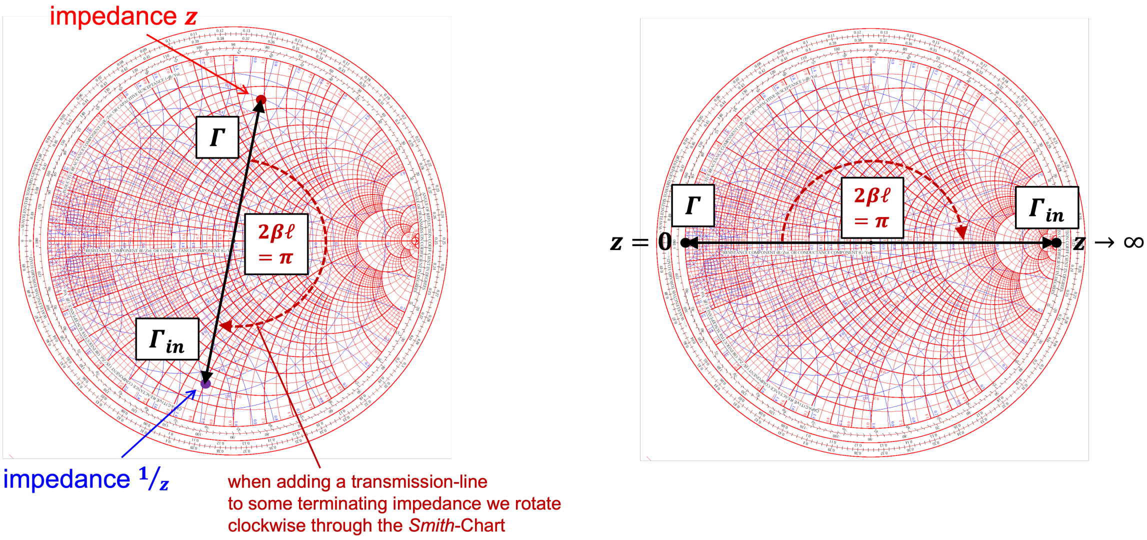
Fig. II.2.68: *Smith* chart of the $\lambda/5$ impedance transformation.

Of particular interest is a transmission-line of length

$$\ell = \frac{\lambda}{4} \equiv \beta\ell = \frac{\pi}{2}$$

which transforms a reflection coefficient Γ at the end of the line to its input as

$$\Gamma_{\text{in}} = \Gamma e^{-j2\beta\ell} = \Gamma e^{-j\pi} = -\Gamma$$



(a) General case.

(b) "Short"-to-"open" case (or vice versa).

Fig. II.2.69: $\lambda/4$ impedance inverter in the *Smith* chart.

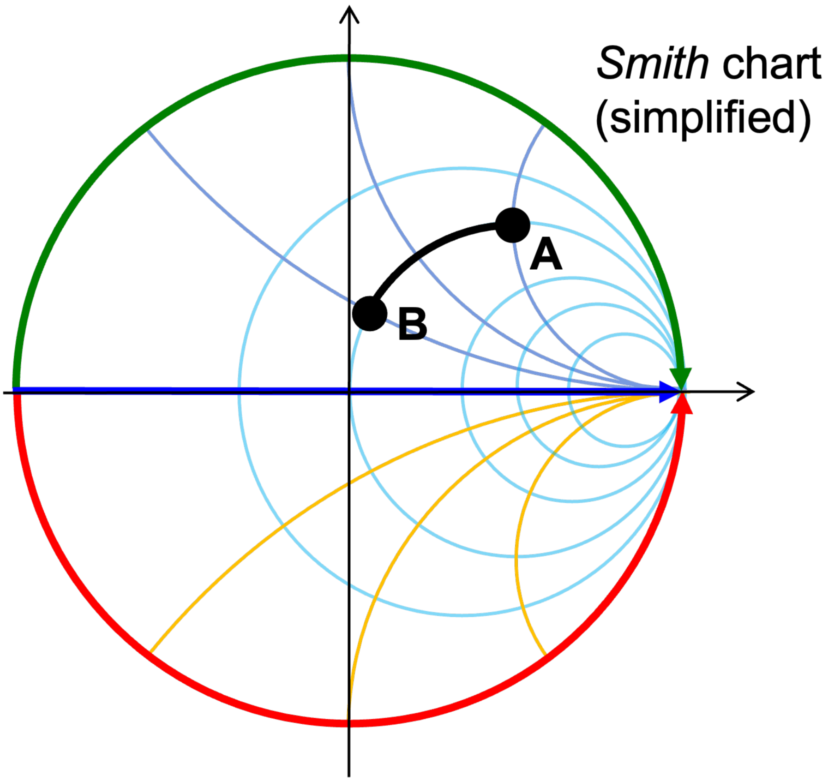
This results in the dimensionless, normalised impedance z at the end of the line to be inverted:

$$z_{\text{in}} = \frac{1}{z} \tag{II.2.29}$$

thus, the $\lambda/4$-line acts as (normalised) impedance inverter! (see also Fig. II.2.69a)

With $z = 0$, i.e. a short at the end of the transmission-line (Fig. II.2.69b) or $z \to \infty$ (open end of the transmission-line), the transmission-line effectively becomes a $\lambda/4$-resonator, also applied for acceleration of heavy ions or protons in linacs.

Test your knowledge!



Trace with marker points in the simplified *Smith* chart for an RL series impedance.
(Mark the correct answer)

- ☐ $f_B > f_A$
- ☐ $f_B < f_A$
- ☐ There is no frequency f related to Points A and B
- ☐ $f_B = f_A$

II.2.6.5 *Smith* chart summary

In this introduction to the *Smith* chart we touched only a few, general aspects, also because, today, the *Smith* chart has lost it's core function as impedance / reflection coefficient calculation tool. Many aspects of the *Smith* chart, like the various *rules* are replaced by software tools or conversion scripts. Still, the *Smith* chart not only helps to better understand the link between the reflection coefficient Γ and the complex impedance Z in a RF system with a reference impedance Z_0, it also remains relevant for the visualisation of complex impedances as parametric plot, e.g. the $S_{ii}(f)$ scattering parameter measured with a vector network analyzer (VNA). As a matter of fact, the *Smith* chart format measuring S_{11} with a VNA of an accelerating mode of a cavity resonator is the most efficient way to characterise its eigen-frequency and unloaded-Q.

II.2.7 Scattering parameters

II.2.7.1 Electrical networks

The behaviour of RF components, circuits and systems, like any other electrical or electronics circuit, system or phenomena can be described very well applying *Maxwell's* equations:

$$\begin{aligned} \nabla \cdot \mathbf{E} &= \frac{\rho}{\varepsilon}, \qquad & \nabla \times \mathbf{B} - \frac{\partial \mathbf{E}}{c^2 \partial t} &= \mu \mathbf{J}, \\ \nabla \times \mathbf{E} - \frac{\partial \mathbf{B}}{\partial t} &= 0, \qquad & \nabla \cdot \mathbf{B} &= 0. \end{aligned} \tag{II.2.30}$$

Equation (II.2.30) needs to be solved taking all the boundaries and materials into account, which can be accomplished numerically, e.g. for simple RF components like antennas, cavity resonators, etc. However, for more sophisticated, complex electrical circuits and systems a design and optimisation, e.g. of a telecommunication device like a cell-phone with hundreds of semiconductors and other electronics components based on *Maxwell's* equations quickly becomes unwieldy to impossible.

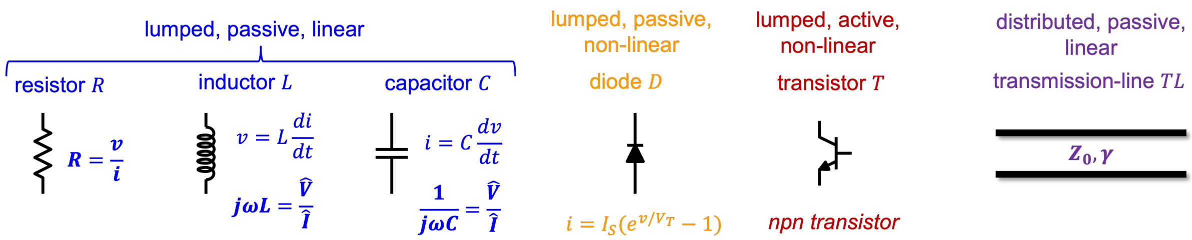
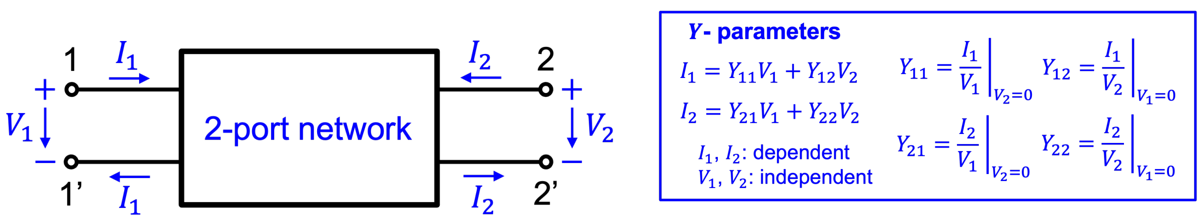


Fig. II.2.70: Examples of electrical circuit elements.

To circumvent this problem, a well-established method in electrical engineering is based on *circuit elements* forming an electrical network. Figure II.2.70 illustrates a few examples for those electrical circuit elements. Each circuit element is represented by an unique symbol and it's electrical characteristic and values are defined by the laws of *Ohm* and *Kirchhoff*. A meaningful linear or non-linear, passive or active function, e.g. an amplifier, a NAND-gate, a RF band-pass filter, etc., is then made from a combination of circuit elements forming the desired function as *electrical network*. Complex systems, described as electrical networks, can be divided into sub-networks, with each sub-network performing a well-defined function.



(a) 2-port network.

(b) 2-port Y-parameters.

Fig. II.2.71: 2-port Y-parameter network.

Functional blocks described by an electrical network have n *ports* and each port has two *terminals*. Linear networks are the most popular ones, Fig. II.2.71 shows a linear 2-port network which is defined by a matrix formalism based on the currents and voltages at the ports, here expressed in terms of the so-called *Y-parameters*, with the matrix coefficients being admittances defined by the dependent currents at the terminals. As Fig. II.2.71b shows, a linear 2-port network has four parameters, given as port voltages (V_1, V_2) and currents (I_1, I_2), here for the Y-parameter description the voltages are the independent, and the currents are the dependent parameters. Evidently there are other definitions for linear networks, known as Z-, h- and g-parameters.

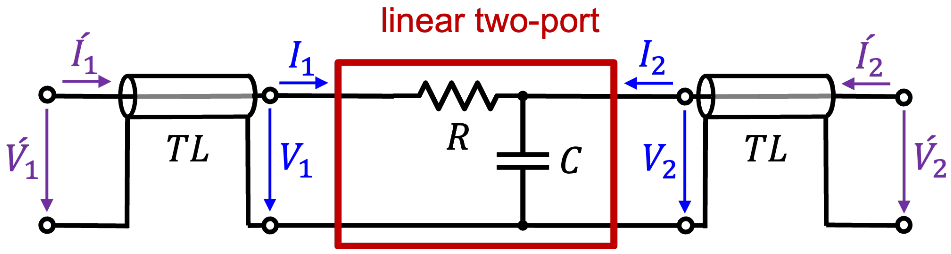


Fig. II.2.72: Measurement characterisation of a linear 2-port network.

In practice it is more convenient to express the parameters for linear networks in steady state, i.e. in the frequency-domain. This avoids using and solving differential equations, but again, is only applicable to time-invariant, linear systems! Figure II.2.72 shows a RC low-pass filter as example for a simple linear 2-port network, indicating V_1 and I_1 at port 1, and V_2 and I_2 at port 2. Based on the Y-parameter definition, Fig. II.2.71b, follows:

$$\text{Y-parameter:} \quad Y = \frac{1}{R}\begin{bmatrix} 1 & -1 \\ -1 & 1 + j\omega RC \end{bmatrix}$$

and similar for the Z-parameter follows:

$$\text{Z-parameter:} \quad Z = \frac{1}{j\omega C}\begin{bmatrix} 1 + j\omega RC & 1 \\ 1 & 1 \end{bmatrix}$$

Both descriptions, based on the port voltages and currents, are equivalent, however, the characterisation by a measurements of those voltages and currents at the ports is difficult and fails at high frequencies.

Any measurement instrument has to be connected via cables to the ports, in Fig. II.2.72 indicated as "TL", i.e. transmission-lines of a given physical length. As we have learned, the voltages and currents on a transmission-line are a **function of frequency (or time) AND space (location),** $\boldsymbol{V_i(\omega, z)}$, $\boldsymbol{I_i(\omega, z)}$, originating from the associated space/time varying EM-fields. This means, at higher frequencies $\acute{V}_i \neq V_i$ and $\acute{I}_i \neq I_i$, and therefore give errors when trying to characterise a linear network at RF frequencies based on port voltages and currents. Moreover, the circuit may become unstable or might be damaged when operating on a short or open end, following the definition of the network parameter. Parasitic effects, e.g. "stray"-capacitances and inductances, given by the measurement setup can further degrade the accuracy of the voltage or current measurement, and may make the characterisation of the network parameters impossible, in particular at microwave frequencies.

Fig. II.2.73: Transmission and reflections of optical waves at a car window (courtesy *Piotr Kowina*).

II.2.7.2 Power-waves and generalised S-parameters

Instead of defining the characteristics of an electrical linear network by the voltages and currents at its ports, a definition through incident and reflected waves at the ports proves to be more practical for linear RF networks. Figure II.2.73 shows analogous optical waves at a car window, with the incident optical waves (sun light) scattered at the surface of the car windows with partial reflections and transmission of the optical waves, defined by the refractive index of the window glass.

Figure II.2.74 illustrates the concept of incident a_i and reflected *power-waves* b_i at the ports $i \in 1 \ldots n$ of a linear RF network, referred as *device-under-test* (DUT) [15]. These power-waves will be scattered at the ports and inside the DUT RF network, thus the related parameters are called scattering or S-parameters.

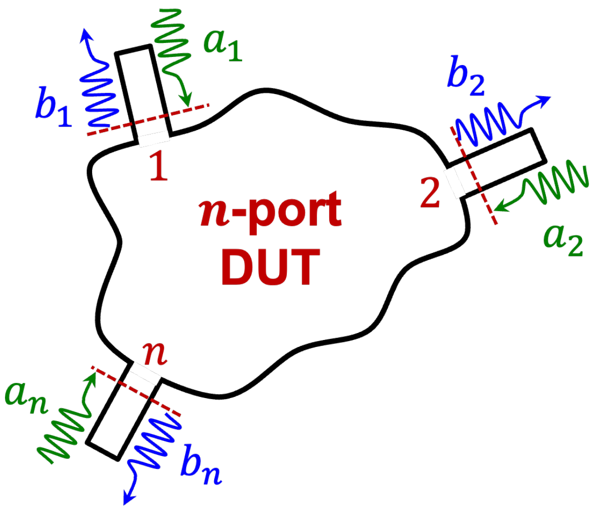


Fig. II.2.74: Power waves a the ports of a device-under-test (DUT).

Definition II.2.7.1: Generalised S-parameters

The generalised S-parameters for an arbitrary n-port microwave or RF network (DUT) are defined through a set of normalised complex voltage (power) waves:

$$\begin{aligned} \text{incident wave at port } i: \quad a_i &= \frac{V_i + Z_i I_i}{2\sqrt{\mathrm{Re}\{Z_i\}}} = \frac{V_i^{inc}}{\sqrt{\mathrm{Re}\{Z_i\}}} \\ \text{reflected wave at port } i: \quad b_i &= \frac{V_i - Z_i^* I_i}{2\sqrt{\mathrm{Re}\{Z_i\}}} = \frac{V_i^{refl}}{\sqrt{\mathrm{Re}\{Z_i\}}} \end{aligned} \tag{II.2.31}$$

with Z_i^* being the complex conjugate of Z_i. The incident a_i and reflected / transmitted power wave b_i at the i^{th}-port are given by the terminal voltage V_i and current I_i, and an arbitrary reference impedance Z_i.

Please note, the complex notation of Eq. (II.2.31) implies linear, time-invariant networks, and is described in the frequency-domain.

Each port of a DUT is defined by a *reference plane*, which is a well defined physical location, in Fig. II.2.74 indicated as dashed lines.

In practice, the characterisation of an RF network, the so-called device-under-test (DUT), is performed as a measurement using a *vector network analyzer* (VNA), or by a numerical analysis or optimisation, simulated on the computer. In case of the VNA, the DUT needs to be connected by transmission-lines, in most cases coaxial cables with a characteristic $Z_0 = 50\,\Omega$, to the ports. Therefore, we usually define the reference impedance of all DUT ports to be $Z_i = Z_0 = 50\,\Omega$.

Please note, some VNAs with a physical reference impedance of $Z_0 = 50\,\Omega$ allow a mathematical port impedance conversion to adapt to a port reference impedance $Z_0 \neq 50\,\Omega$.

II.2.7.3 1-port S-parameter

Before we continue with the definition of the most simple, the 1-port S-parameter, let's briefly summarise:

Electrical / electronic networks are

- electrical / electronic circuits with $1 \dots n$ ports.
- defined by the voltages $V_i(\omega)$ or $v_i(t)$ and currents $I_i(\omega)$ or $i_i(t)$ at their port terminals.
- characterised by circuit matrices, e.g. Z, Y, h, g, etc. based on V_i and I_i.

RF / microwave networks are

- RF or microwave circuits or subsystems, i.e. DUTs, like amplifiers, filters, transmission-lines, resonators, hybrids, circulators, etc., which may include distributed elements, having $1 \dots n$ ports.
- defined by incident a_i and reflected / transmitted power-waves b_i at the i^{th}-port, taking the physical location, i.e. *reference plane*, of the port into account.
- characterised by a matrix of scattering (S)-parameters based on the power-waves at the ports. The S-parameters are a function of frequency $f = \omega/2\pi$.
- The power-waves are normalised to a reference impedance, which in most cases is $Z_i = Z_0 = 50\,\Omega$.

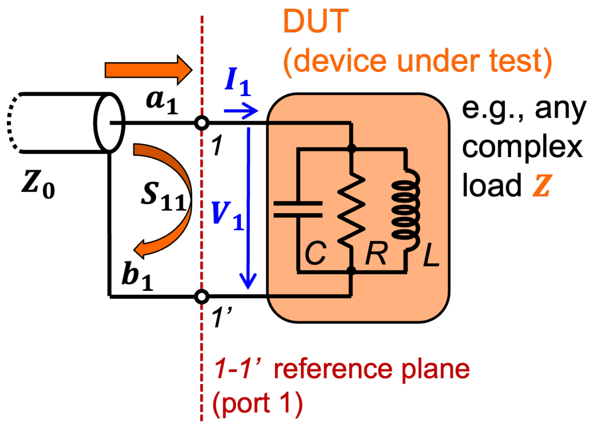


Fig. II.2.75: 1-port DUT.

Figure II.2.75 shows the most simple case, a 1-port DUT, here as example of a RLC parallel resonant circuit. It has only one port, with two terminals 1 and $1'$, defining the reference plane. Evidently, a 1-port RF network has only one S-parameter, S_{11}, defined as ratio of the reflected wave b_1 to the incident wave a_1:

$$S_{11} = \frac{b_1}{a_1} = \Gamma \tag{II.2.32}$$

and is nothing else than our reflection coefficient Γ. As Fig. II.2.75 suggests, the measurement equipment can be located at some physical distance from the DUT, and is connected with a transmission-line of characteristic impedance Z_0, usually a coaxial cable, only for some specific microwave applications a rectangular waveguide, to the reference plane of the port.

II.2.7.4 2-port S-parameters

2-port RF networks are the most popular, e.g. amplifiers, attenuators, filters, transmission-lines, isolators, etc., to name but some, are all 2-port DUTs. Figure II.2.76 shows a particular simple 2-port DUT, a complex impedance Z series circuit between the ports. Indicated is the setup for the definition of the

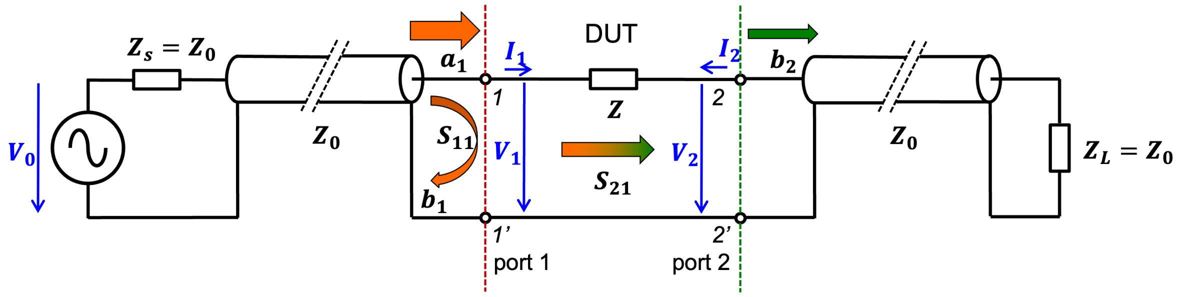


Fig. II.2.76: 2-port DUT: forward S-parameters.

forward S-parameters:

$$S_{11} = \left.\frac{b_1}{a_1}\right|_{a_2=0} \equiv \text{input reflection coefficient} \tag{II.2.33}$$

$$S_{21} = \left.\frac{b_2}{a_1}\right|_{a_2=0} \equiv \text{forward transmission gain} \tag{II.2.34}$$

with:

$$\text{independent parameters:} \quad a_1 = \frac{V_1^{inc}}{\sqrt{Z_0}} = \frac{V_1 + I_1 Z_0}{2\sqrt{Z_0}} \tag{II.2.35}$$

$$\text{dependent parameters:} \quad b_1 = \frac{V_1^{refl}}{\sqrt{Z_0}} = \frac{V_1 - I_1 Z_0}{2\sqrt{Z_0}} \tag{II.2.36}$$

$$b_2 = \frac{V_2^{refl}}{\sqrt{Z_0}} = \frac{V_2 - I_2 Z_0}{2\sqrt{Z_0}} \tag{II.2.37}$$

In a practical measurement setup the independent parameter a_1, Eq. (II.2.35), is supplied by the signal generator in the VNA, indicated as source voltage V_0 in Fig. II.2.76. The two dependent parameters, b_1, Eq. (II.2.36), and b_2, Eq. (II.2.37), are measured with the VNA, enabling to calculate S_{11}, Eq. (II.2.33), and S_{21}, Eq. (II.2.34).

IMPORTANT:
Please note $a_2 = 0$ in Eqs. (II.2.33) and (II.2.34). This means, no reflections should appear on port 2, otherwise the measurement would be uncontrolled. To insure this, port 2 needs to be terminated in its reference impedance, thus the load impedance at port 2 has to be: $Z_L = Z_0$, see also Fig. II.2.76. Of course, also the stimulus signal generator at port 1 is impedance matched, i.e. the source impedance is equal to the reference impedance, $Z_S = Z_0$, and therefore will absorb all reflected waves b_1 to ensure controlled measurement conditions without multiple reflections!
In addition:
ALWAYS terminate unused ports in their reference impedance, typically the characteristic impedance of the coaxial cables, e.g. Z_0 =50 Ω!

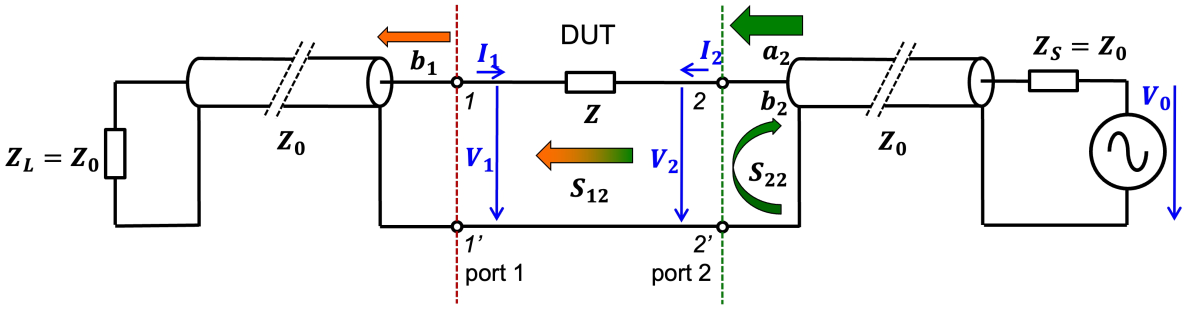


Fig. II.2.77: 2-port DUT: reverse S-parameters.

The definition of the *reverse* S-parameters follows a similar path and is illustrated in Fig. II.2.77:

$$S_{22} = \left.\frac{b_2}{a_2}\right|_{a_1=0} \equiv \text{output reflection coefficient} \tag{II.2.38}$$

$$S_{12} = \left.\frac{b_1}{a_2}\right|_{a_1=0} \equiv \text{reverse transmission gain} \tag{II.2.39}$$

with:

$$\text{independent parameters:} \quad a_2 = \frac{V_2^{inc}}{\sqrt{Z_0}} = \frac{V_2 + I_2 Z_0}{2\sqrt{Z_0}} \tag{II.2.40}$$

$$\text{dependent parameters:} \quad b_1 = \frac{V_1^{refl}}{\sqrt{Z_0}} = \frac{V_1 - I_1 Z_0}{2\sqrt{Z_0}} \tag{II.2.36}$$

$$b_2 = \frac{V_2^{refl}}{\sqrt{Z_0}} = \frac{V_2 - I_2 Z_0}{2\sqrt{Z_0}} \tag{II.2.37}$$

Of course, the dependent parameters b_1 and b_2 remain the same as for the forward S-parameters, Eqs. (II.2.36) and (II.2.37). Now the stimulus signal is supplied at port 2, indicated by V_0 in Fig. II.2.77, in practice the VNA will automatically switch to the correct port 2 if a S_{22} or S_{12} measurement is requested. In the case of the reverse S-parameters, port 1 has to be correctly terminated with the reference impedance, $Z_L = Z_0$, and again, the source impedance of the stimulus signal generator is matched to the reference impedance of port 2, $Z_S = Z_0$.

Most vector network analyzers are equipped with two ports. Still, it is possible with a 2-port VNA to perform the characterisation of DUTs with ports $n \geq 3$, again, unused ports need to be terminated in their reference impedance!

Combining Eqs. (II.2.33), (II.2.34), (II.2.38) and (II.2.39) leads to a system of linear equations for the 2-port DUT:

$$\begin{aligned} b_1 &= S_{11}a_1 + S_{12}a_2 \\ b_2 &= S_{21}a_1 + S_{22}a_2 \end{aligned} \tag{II.2.41}$$

with:

$$\left.\begin{aligned} S_{11} &= \left.\frac{b_1}{a_1}\right|_{a_2=0} \equiv \text{input reflection coefficient} \\ S_{22} &= \left.\frac{b_2}{a_2}\right|_{a_1=0} \equiv \text{output reflection coefficient} \end{aligned}\right\} \text{impedance measurements} \tag{II.2.42}$$

$$\left.\begin{aligned} S_{21} &= \left.\frac{b_2}{a_1}\right|_{a_2=0} \equiv \text{forward transmission gain (insertion loss)} \\ S_{12} &= \left.\frac{b_1}{a_2}\right|_{a_1=0} \equiv \text{reverse transmission gain (insertion loss)} \end{aligned}\right\} \begin{array}{l}\text{transmission} \\ \text{(insertion)} \\ \text{measurements}\end{array} \tag{II.2.43}$$

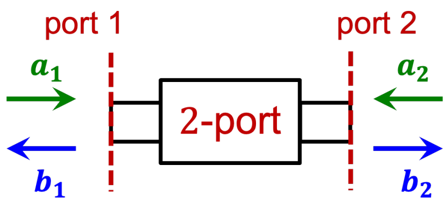
A load impedance $Z_L \neq Z_0$ at port 2, see Fig. II.2.76, will result in a reflection coefficient at the output (port 2):

$$\Gamma_{load} = \frac{Z_L - Z_0}{Z_L + Z_0}$$

which translates to a reflection coefficient at the input (port 1):

$$\Gamma_{\text{in}} = S_{11} + \frac{S_{21}\Gamma_{load}S_{12}}{1 - S_{22}\Gamma_{load}}$$

Test your knowledge!



Mark all correct answers for the S-parameters of a 2-port RF network

- ☐ a_1 and b_1 are independent parameters.
- ☐ $S_{11} = b_1/a_1$ ($a_2 = 0$) is the input reflection coefficient Γ_1.
- ☐ a_1 and a_2 are the incident waves at port 1 and port 2, respectively.
- ☐ b_1 and b_2 are the transmitted waves between port 1 and port 2, and vice versa.
- ☐ S_{21} and S_{12} are the forward and reverse transmission gains / losses.
- ☐ To characterise the S-parameters at port 2, port 1 needs to be terminated in its characteristic port impedance.

II.2.7.5 n-port S-parameters and the scattering matrix

The definitions in the previous two sections for 1- and 2-port S-parameters can be generalised to n-port DUTs. For the reflection coefficient, Eq. (II.2.32) for the 1-port, we get for the n-port network (DUT):

$$S_{ii} = \frac{b_i}{a_i} = \frac{\frac{V_i}{I_i} - Z_0}{\frac{V_i}{I_i} + Z_0} = \frac{Z_i - Z_0}{Z_i + Z_0} = \Gamma_i \tag{II.2.44}$$

with

$$Z_i = Z_0 \frac{1 + S_{ii}}{1 - S_{ii}} \tag{II.2.45}$$

and

$$Z_i = \frac{V_i}{I_i} \tag{II.2.46}$$

being the input impedance at the i^{th} port.
The electrical power reflected at the i^{th} port is:

$$|S_{ii}|^2 = \frac{\text{power reflected from port } i}{\text{power incident on port } i} \tag{II.2.47}$$

and the electrical power transmitted between ports is:

$$|S_{ij}|^2 = \text{power transmitted between ports } i \text{ and } j \tag{II.2.48}$$

Again, all ports are terminated in their reference impedance, usually $Z_i = Z_0$, i.e. the characteristic impedance of the coaxial cables, and also the source impedance is $Z_S = Z_i = Z_0$.

From the definition of power waves scattered at the ports of a linear n-port network, Eq. (II.2.31), we can further conclude:

$$\begin{aligned} \mathbf{a} &= \begin{bmatrix} a_1 & a_2 & a_3 & \dots & a_n \end{bmatrix} && \text{waves travelling } \textit{towards} \text{ the } n\text{-port} \\ \mathbf{b} &= \begin{bmatrix} b_1 & b_2 & b_3 & \dots & b_n \end{bmatrix} && \text{waves travelling } \textit{away} \text{ from the } n\text{-port} \end{aligned}$$

incident $\mathbf{a}$ and reflected waves $\mathbf{b}$ being vectors, and the relation between a_i and b_i ($i = 1 \dots n$) can be written as a system of n linear equations:

$$\begin{array}{llllll} \text{1-port:} & b_1 = S_{11}a_1 & +S_{12}a_2 & +S_{13}a_3 & +S_{14}a_4 & + \dots \\ \text{2-port:} & b_2 = S_{21}a_2 & +S_{22}a_2 & +S_{23}a_3 & +S_{24}a_4 & + \dots \\ \text{3-port:} & b_3 = S_{31}a_1 & +S_{32}a_2 & +S_{33}a_3 & +S_{34}a_4 & + \dots \\ \text{4-port:} & b_4 = S_{41}a_2 & +S_{42}a_2 & +S_{43}a_3 & +S_{44}a_4 & + \dots \end{array}$$

with a_i being the independent and b_i being the dependent variables. Following that, and generalising Eq. (II.2.41), the S-parameters of a linear n-port network can be simply expressed as *scattering (S)-matrix*

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1i} \\ S_{21} & S_{22} & \dots & S_{2i} \\ \vdots & & \ddots & \vdots \\ S_{i1} & S_{i2} & \dots & S_{ii} \end{bmatrix} \tag{II.2.49}$$

of the above described system of linear equations:

$$\mathbf{b} = \mathbf{S}\,\mathbf{a} \tag{II.2.50}$$

?

Test your knowledge!
Select all correct answers

- ☐ Y- and Z-parameters of electrical networks require a reference impedance Z_0.
- ☐ Scattering parameters of RF networks are based on normalised, complex power waves, incident and reflected at their ports.
- ☐ DUT stands for "Device Under Test", as acronym for a RF network to be characterised.
- ☐ S-parameters are only defined for a reference impedance of $Z_0 = 50\,\Omega$.
- ☐ Unused ports in a S-parameter measurement setup always need to be terminated in their characteristic port impedance.

II.2.7.6 Some properties of the S-matrix

II.2.7.6.1 Matched ports

The i^{th} port of a RF network is *matched* if

$$S_{ii} = 0 \tag{II.2.51}$$

i.e. no reflections occur from that port!

II.2.7.6.2 Reciprocal networks

A n-port network is *reciprocal* if

$$\mathbf{S}^T = \mathbf{S} \quad \Rightarrow \quad S_{ij} = S_{ji} \;\; \forall\, i, j \tag{II.2.52}$$

with $\mathbf{S}^T$ being the transposed matrix of $\mathbf{S}$. Reciprocal networks have a symmetric S-matrix. Most passive components are reciprocal, e.g. resistor, capacitor, inductor, transformer, etc.

Passive components using non-homogeneous materials like magnetised ferrites, plasma, etc. are non-reciprocal. Also most active components, like amplifiers, are non-reciprocal.

II.2.7.6.3 Symmetric networks

A n-port network is *symmetric* if

$$S_{ij} = S_{ji} \quad \wedge \quad S_{ii} = S_{jj} \tag{II.2.53}$$

i.e. it needs to be reciprocal and the reflection coefficients at the ports need to be identical (matrix and electrical symmetry).

Without proof: A symmetric network is always reciprocal.

II.2.7.6.4 Passive and lossless networks

A n-port is *passive* and *lossless* if its S-matrix is unitary

$$\mathbf{S}^{\dagger}\mathbf{S} = \mathbf{S}^{T}\mathbf{S}^{*} = \mathbf{I} \tag{II.2.54}$$

where $\mathbf{S}^{T} = (\mathbf{S}^{*})^{T}$ is the conjugate transposed (*Hermitian*) matrix and $\mathbf{I}$ is the identity matrix.

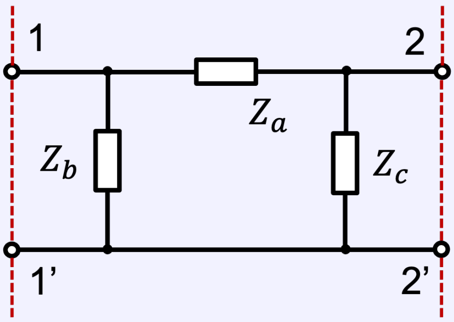
For a passive, lossless 2-port network follows:

$$(\mathbf{S}^{*})^{T}\,\mathbf{S} = \begin{bmatrix} S_{11}^{*} & S_{21}^{*} \\ S_{12}^{*} & S_{22}^{*} \end{bmatrix} \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

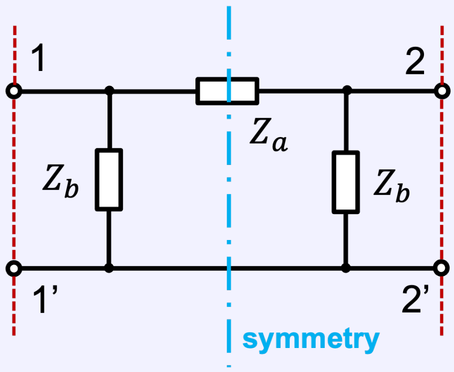
which lead to the following conditions:

$$\begin{gathered} \angle S_{11} - \angle S_{12} = \angle S_{21} - \angle S_{22} - \pi \\ |S_{11}| = |S_{22}| \\ |S_{12}| = |S_{21}| \\ |S_{11}| = \sqrt{1 - |S_{12}|^2} \end{gathered} \tag{II.2.55}$$

Example II.2.7.1: Examples of reciprocal and symmetric S-parameter networks



(a) general π-network.



(b) symmetric π-network.

Fig. II.2.78: The 2-port π-network.

Consider a 2-port π-network, Fig. II.2.78. The S-parameters for a general π-network, Fig. II.2.78a, can be derived from the "classical" network parameters, e.g. from the 2-port Y-parameters:

$$\begin{aligned} S11 &= \frac{(1 - Z_0Y_{11})(1 + Z_0Y_{22}) + Z_0^2Y_{12}Y_{21}}{\Delta} \\ S12 &= -2\frac{Z_0Y_{12}}{\Delta} \\ S21 &= -2\frac{Z_0Y_{21}}{\Delta} \\ S22 &= \frac{(1 + Z_0Y_{11})(1 - Z_0Y_{22}) + Z_0^2Y_{12}Y_{21}}{\Delta} \\ &\text{with: } \Delta = (1 + Z_0Y_{11})(1 + Z_0Y_{22}) - Z_0^2Y_{12}Y_{21} \end{aligned} \tag{II.2.56}$$

by replacing:

$$Y_{11} = \frac{1}{Z_a} + \frac{1}{Z_b}, \qquad Y_{12} = Y_{21} = -\frac{1}{Z_a}, \qquad Y_{22} = \frac{1}{Z_a} + \frac{1}{Z_c} \tag{II.2.57}$$

following Fig. II.2.78a. We find $S_{12} = S_{21}$, therefore the network is reciprocal, but $S_{11} \neq S_{21}$, thus it is not symmetric. Setting $Z_c = Z_b$, see Fig. II.2.78b, the S-matrix simplifies to:

$$\begin{aligned} \mathbf{S}_{(\pi,\text{ symm})} &= \frac{1}{\Delta}\begin{bmatrix} Z_aZ_b^2 - Z_0^2(Z_a + 2Z_b) & 2Z_0Z_b^2 \\ 2Z_0Z_b^2 & Z_aZ_b^2 - Z_0^2(Z_a + 2Z_b) \end{bmatrix} \\ &\text{with: } \Delta = (Z_a + Z_b)\left[Z_aZ_b + Z_0(Z_a + 2Z_b)\right] \end{aligned} \tag{II.2.58}$$

with $S_{12} = S_{21} \ \wedge \ S_{11} = S_{22}$, therefore this π-network is reciprocal and symmetric.

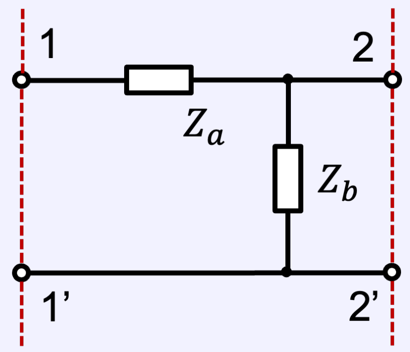


Fig. II.2.79: A simple voltage divider as 2-port network.

Setting in Fig. II.2.78a $Z_b \to \infty$ and $Z_c = Z_b$ leads to the simple voltage divider circuit as shown in Fig. II.2.79. From the S-parameters

$$\mathbf{S}_{(\text{div})} = \frac{1}{\Delta} \begin{bmatrix} Z_a(Z_0 + Z_b) - Z_0^2 & 2Z_0 Z_b \\ 2Z_0 Z_b & Z_a Z_b - Z_0(Z_0 + Z_a) \end{bmatrix} \qquad \text{(II.2.59)}$$
$$\text{with: } \Delta = Z_0(Z_0 + Z_a) + Z_b(2Z_0 + Z_a)$$

follows $S_{12} = S_{21} \ \wedge \ S_{11} \neq S_{22}$, i.e., this 2-port is reciprocal but not symmetric.

Exercise II.2.7.1: Passive and lossless S-matrices

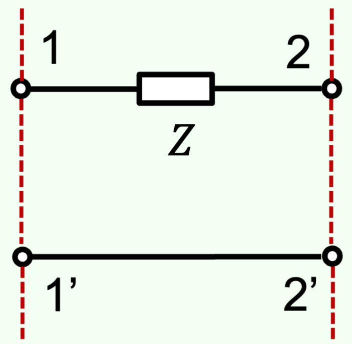


Fig. II.2.80: A simple series impedance 2-port network.

Figure II.2.80 shows a simple 2-port network for a series impedance Z. Analyse some properties of this network:

– Setup the S-matrix for the 2-port network of Fig. II.2.80

– Is the 2-port reciprocal and/or symmetric?

– For two numerical cases, $Z = R = 10\,\Omega$ and $Z = j\omega L = j10\,\Omega$, test the 2-port to be lossy / lossless using Eq. (II.2.55).

Test your knowledge!
Fill the centre column with the correct answer, $1 \dots 6$

Prompts		**Possible Answers**
A. matched		1. $S_{ii} = S_{jj}$
B. symmetric		2. $(\mathbf{S}^*)^T \mathbf{S} = \mathbf{I}$
C. reciprocal		3. $S_{ij} = S_{ji} \,\wedge\, S_{ii} = S_{jj}$
D. passive and lossless		4. $S_{ii} = 0$
		5. $\Gamma = +j$
		6. $S_{ij} = S_{ji}$

II.2.7.7 Examples of S-matrices

II.2.7.7.1 1-port

Any passive R, L, C, RC, RL, LC, or RLC lumped element circuit, or any combination of those elements leading to the single port network results in a 1-port S-matrix:

$$\mathbf{S} = S_{11} = \Gamma \tag{II.2.60}$$

Of course, the network may also include, or be solely made from, distributed elements (transmission-lines). "Special" cases of 1-port networks are:

- Ideal (matched) termination: $Z = Z_0 \;\Rightarrow\; S_{11} = 0$
- Ideal short: $Z = 0 \;\Rightarrow\; S_{11} = -1$
- Ideal open: $Z = \infty \;\Rightarrow\; S_{11} = +1$

While strictly speaking any simple cavity resonator is at least a 3-port, two waveguide beam-ports

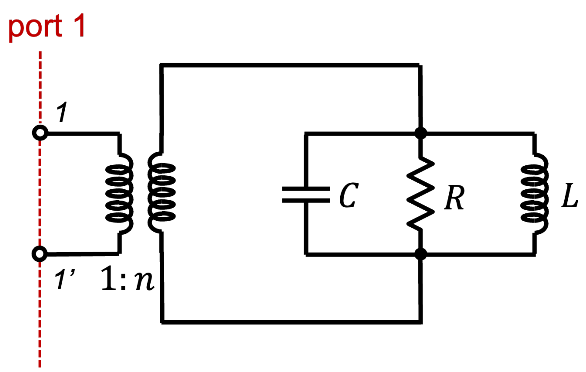


Fig. II.2.81: 1-port lumped element equivalent circuit of a cavity resonance.

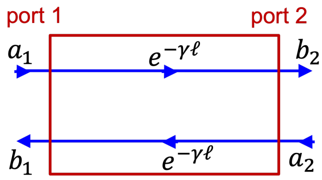


Fig. II.2.82: Signal flow graph (SFG) for an ideal, matched transmission-line.

which are part of the vacuum system, and a coaxial or waveguide port for the RF power coupler, it often can be treated as 1-port RF network. In that case we assume the eigenmode resonance of interest is trapped in the resonator with no or negligible fields near the beam-ports. We consider only a single RF coupler to characterise the 1-port S-parameter of a particular eigenmode, e.g. the TM010 accelerating mode of a cylindrical "pill-box" cavity. Figure II.2.81 shows the parallel RLC equivalent circuit for a cavity resonant mode, with an ideal transformer acting as coupling antenna.

II.2.7.7.2 2-port

2-port RF components and elements are the most popular, ranging from simple transmission-lines, attenuators, filters, etc. to complex RF power amplifiers.

$$\mathbf{S} = \begin{bmatrix} 0 & e^{-\gamma\ell} \\ e^{-\gamma\ell} & 0 \end{bmatrix} \quad \text{with:} \quad \begin{array}{l} \gamma = \alpha + j\beta \ : \ \text{propagation constant} \\ \alpha \ : \ \text{attenuation constant } [Np/m]^1 \\ \beta = 2\pi/\lambda \ : \ \text{phase constant } [rad/m] \end{array} \tag{II.2.61}$$

Equation (II.2.61) gives the S-parameter matrix for ideal, matched transmission-line, i.e. the characteristic impedance of the transmission-line Z_{TL} is equal the reference impedance Z_0 of the RF system, $Z_{TL} = Z_0$. Figure II.2.82 shows the corresponding *signal flow graph* (SFG), a graphical representation of the waves propagating through the S-matrix. For a lossless line of length $\ell = \lambda/4$ operating at $f = v_p/\lambda$ the S-matrix simplifies to

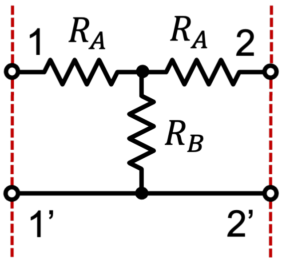
$$\mathbf{S} = \begin{bmatrix} 0 & -j \\ -j & 0 \end{bmatrix} \tag{II.2.62}$$

The 2-port S-matrix for an ideal *attenuator* is

$$\mathbf{S} = \begin{bmatrix} 0 & k \\ k & 0 \end{bmatrix} \quad \text{with:} \quad \begin{array}{c} k = V_2/V_1 = 10^{-(\Delta dB/20)} \ : \ \text{attenuation } k < 1,\ k \in \text{Re} \\ \Delta dB = 20\log_{10} V_1/V_2 \ : \ \text{attenuation in } [dB] \\ \alpha = -\ln k \ : \ \text{attenuation in } [Np] \end{array} \tag{II.2.63}$$

and is evidently independent of the frequency. Figure II.2.83 shows two variants of an RF attenuator, the T-network (Fig. II.2.83a) and the π-network (Fig. II.2.83b), and Fig. II.2.84 illustrates the signal flow graph, with the specific values for a 3 dB attenuator.

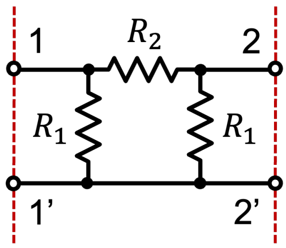
[1] $Np = Neper$, unitless ratio named after *John Napier* based on the natural logarithm base e: $Np = \ln(a/b)$. Compares to $dB = 20\log_{10}(a/b)$: $dB/Np = 20\log_{10}(a/b)/\ln(a/b) = 20/\ln(10) \approxeq 8.68588$



$$R_A = \frac{1-k}{1+k} Z_0$$

$$R_B = \frac{2k}{1-k^2} Z_0$$

(a) Attenuator T-network



$$R_1 = \frac{1-k}{1+k} Z_0$$

$$R_2 = \frac{1-k^2}{2k} Z_0$$

(b) Attenuator π-network.

Fig. II.2.83: 2-port RF attenuator circuits.

$$S_{21} = \frac{b_2}{a_1} = \frac{1}{\sqrt{2}}$$

$$S_{12} = \frac{b_1}{a_2} = \frac{1}{\sqrt{2}}$$

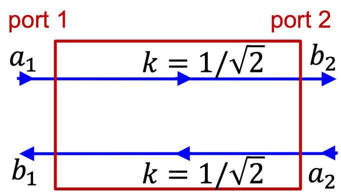


Fig. II.2.84: Signal flow graph for a 3 dB attenuator.

The S-matrix for an ideal amplifier, or *gain stage*, is:

$$\mathbf{S} = \begin{bmatrix} 0 & 0 \\ G_V & 0 \end{bmatrix} \quad \text{with:} \quad \begin{aligned} G_V &= V_{\text{out}}/V_{\text{in}} = 10^{g/20} && : \text{ voltage gain } |G_V| > 1 \\ g &= 20\log_{10} V_{\text{out}}/V_{\text{in}} && : \text{ gain in } [dB] \end{aligned} \tag{II.2.64}$$

The corresponding signal flow graph is shown in Fig. II.2.85a. While the S-parameters of an ideal amplifier are frequency independent, this is not the case for a "real-world" amplifier. Equation (II.2.65) shows the S-parameters for a E-pHEMT GaAs FET [2] at 10 GHz, not only the S-parameters are frequency dependent; neither the input, nor the output are matched and require an impedance-matching circuit.

$$\mathbf{S} = \begin{bmatrix} 0.78e^{j158^\circ} & 0.08e^{-j24^\circ} \\ 3.43e^{j42^\circ} & 0.3e^{-j159^\circ} \end{bmatrix} \quad \text{with:} \quad \begin{aligned} &\text{Datasheet } \textit{Avago} \text{ VMMK-1218:} \\ &f = 10\,\text{GHz},\ Z_0 = 50\,\Omega,\ T_A = 25\,°\text{C} \\ &V_{ds} = 2\,\text{V},\ I_{ds} = 20\,\text{mA} \end{aligned} \tag{II.2.65}$$

Amplifiers are non-reciprocal.

[2]HEMT = high electron mobility transistor (single atomic layer functional principle), GaAs = gallium arsenide (transistor material), FET = field-effect transistor (transistor type)

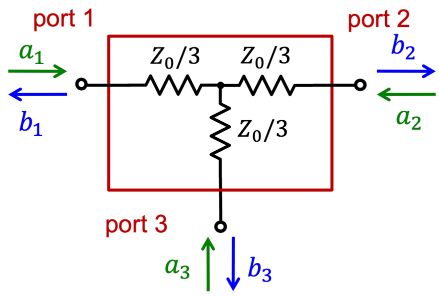


Fig. II.2.86: 3-port resistive power divider.

(a) Ideal amplifier.

(b) Microwave transistor operating at 10 GHz.

Fig. II.2.85: 2-port signal flow graphs for gain stages.

II.2.7.7.3 3-port

From the large variety of 3-port RF networks, Fig. II.2.86 shows the *resistive power divider* (or power splitter) as an example, with the S-matrix given as:

$$\mathbf{S} = \frac{1}{2}\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \qquad \begin{aligned} b_1 &= \frac{1}{2}(a_2 + a_3) \\ b_2 &= \frac{1}{2}(a_1 + a_3) \\ b_3 &= \frac{1}{2}(a_1 + a_2) \end{aligned} \tag{II.2.66}$$

The resistive power divider is frequency independent and the transfer-loss (and isolation) between ij ports is 6 dB. The resistive power divider circuit can be expanded to more ports.

There is a large variety of 3-port power dividers / splitters / combiners, notably the *Wilkinson* type. While many of those power divider networks offer better properties in terms of insertion loss and isolation compared to the resistive power divider, they are frequency dependent.

$$\mathbf{S} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad \begin{aligned} b_1 &= a_3 \\ b_2 &= a_1 \\ b_3 &= a_2 \end{aligned} \tag{II.2.67}$$

Equation (II.2.67) shows the S-matrix of an ideal *circulator* and Fig. II.2.87 the corresponding schematic. The incident power wave at any port is circulated, here shown clockwise, to the next port.

Terminating one port, e.g. port 3, with a matched load results in a 2-port network, called *isolator*, see also Fig. II.2.88:

$$\mathbf{S} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \qquad b_2 = a_1 \tag{II.2.68}$$

The most popular application of an isolator, or a circulator with an external load, is in "isolating" a power amplifier, e.g. triode, solid-state power amplifier, klystron, etc., from an reflective load, e.g. an accelerating cavity, to avoid damage of the amplifier by the reflected power. Both, isolator and circulator, are matched, but non-reciprocal devices, and in their real-world implementations are based on magnetised, anisotropic microwave ferrites and therefore frequency dependent.

II.2.7.7.4 4-port

Many of the RF couplers and hybrids are 4-port RF networks, and the *directional coupler* is the most popular one. Figure II.2.89a shows the operational schematic and Fig. II.2.89b a practical implementation of the most simple, single-stage directional coupler. As Fig. II.2.89b illustrates, this coupler is based on two TEM transmission-lines, which electromagnetically couple over a length $\ell = \lambda/4$, here implemented as coupled micro-strip lines on a printed circuit board (PCB). The directional coupler has a 2-fold symmetry, therefore inputs and outputs can be assigned in different ways to the ports without

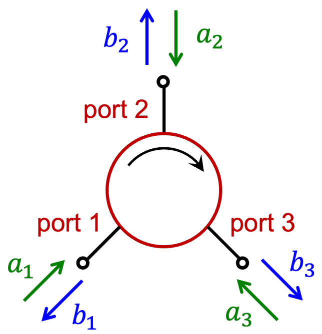


Fig. II.2.87: 3-port circulator.

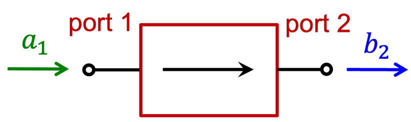
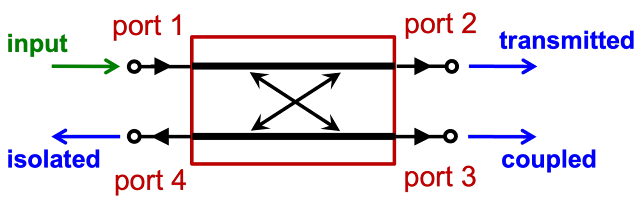
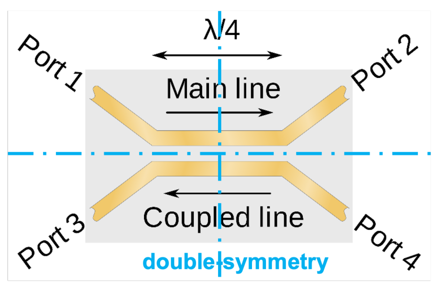


Fig. II.2.88: 2-port isolator.



(a) Schematic of a directional coupler.



(b) PCB micro-strip line implementation of a directional coupler.

Fig. II.2.89: The 4-port directional coupler.

altering the function.

$$\mathbf{S} = \begin{bmatrix} 0 & \tau & \kappa & 0 \\ \tau & 0 & 0 & \kappa \\ \kappa & 0 & 0 & \tau \\ 0 & \kappa & \tau & 0 \end{bmatrix} \qquad \begin{aligned} \tau &: \text{ transmission coefficient} \\ \kappa &: \text{ coupling coefficient} \end{aligned} \tag{II.2.69}$$

Equation (II.2.69) shows the S-matrix for an ideal directional coupler, i.e. infinite isolation and perfectly matched. In general, τ and κ have complex, frequency dependent values. The property of particular interest is the *directivity* of the coupler, a signal at the input port 1 of the main line (ports 1-2) results in a coupled signal at the coupled line port 3, but no signal on the coupled line port 4 (see Fig. II.2.89a). The same is true if port 2 is used as input, now the coupled power shows only on port 4 of the coupled line. In practice this means, the coupled ports 3 and 4 can distinguish between forward and backward travelling waves on the main line, ports 1 and 2. This made the directional coupler the preferred tool for the S-parameter test sets in early generation network analyzers, in accelerator RF applications high power directional couplers are connected between the RF power system and the input coupler of the acceleration cavity, to measure the beam loading by comparing forward and reflected power between the cavity and the RF power amplifier.

The schematic Fig. II.2.89a seems to be misleading wrt. Fig. II.2.89b. Please note, the coupled port of a pair of coupled TEM transmission-lines is always upstream with zero phase delay, the downstream port with $\pi/2$ phase delay is the isolated port – so-called *reverse* coupling – as indicated in Fig. II.2.89b. Waveguide implementation often leads to a *forward* coupling as indicated in the schematic Fig. II.2.89a.

Example II.2.7.2: The ideal, lossless TEM directional coupler

The coupled port 3 of a directional coupler based on coupled TEM transmission-lines is in-phase with the input signal at port 1, while the main-line output signal is in quadrature phase, i.e. delayed by $\pi/2$, see also Fig. II.2.89b. This leads to

$$\kappa = k, \quad \tau = -jt, \quad \text{with: } k,\, t \in \mathrm{Re} \tag{II.2.70}$$

in Eq. (II.2.69).

The condition for a lossless directional coupler is given in Eq. (II.2.54), from which follows:

$$\sum_{k=1}^{N} S_{ki} S_{kj}^{*} = 0 \;\forall\; i \neq j \quad \wedge \quad \sum_{k=1}^{N} S_{ki} S_{ki}^{*} = \sum_{k=1}^{N} |S_{ki}|^2 = 1 \;\forall\; i = j \tag{II.2.71}$$

Testing the S-matrix of the ideal TEM directional coupler, Eq. (II.2.69) with the condition Eq. (II.2.70), satisfies the first condition in Eq. (II.2.71):

$$S_{11}S_{12}^{*}+S_{21}S_{22}^{*}+S_{31}S_{32}^{*}+S_{41}S_{42}^{*}=0\cdot jt+(-jt)\cdot 0+k\cdot 0+0\cdot k=0$$

$$S_{11}S_{13}^*+S_{21}S_{23}^*+S_{31}S_{33}^*+S_{41}S_{43}^*=0\cdot k+(-jt)\cdot 0+k\cdot 0+0\cdot jt=0$$

$$S_{11}S_{14}^*+S_{21}S_{24}^*+S_{31}S_{34}^*+S_{41}S_{44}^*=0\cdot 0+(-jt)\cdot k+k\cdot jt+0\cdot 0=0$$

$$S_{12}S_{13}^*+S_{22}S_{23}^*+S_{32}S_{33}^*+S_{42}S_{43}^*=(-jt)\cdot k+0\cdot 0+0\cdot 0+k\cdot jt=0$$

$$S_{12}S_{14}^*+S_{22}S_{24}^*+S_{32}S_{34}^*+S_{42}S_{44}^*=(-jt)\cdot 0+0\cdot k+0\cdot jt+k\cdot 0=0$$

$$S_{13}S_{14}^*+S_{23}S_{24}^*+S_{33}S_{34}^*+S_{43}S_{44}^*=k\cdot 0+0\cdot k+0\cdot jt+(-jt)\cdot 0=0$$

while the second condition in Eq. (II.2.71)

$$|S_{11}|^2+|S_{21}|^2+|S_{31}|^2+|S_{41}|^2=0^2+t^2+k^2+0^2=1$$

$$|S_{12}|^2+|S_{22}|^2+|S_{32}|^2+|S_{42}|^2=t^2+0^2+0^2+k^2=1$$

$$|S_{13}|^2+|S_{23}|^2+|S_{33}|^2+|S_{43}|^2=k^2+0^2+0^2+t^2=1$$

$$|S_{14}|^2+|S_{24}|^2+|S_{34}|^2+|S_{44}|^2=0^2+k^2+t^2+0^2=1$$

leads to:

$$t^2+k^2=1 \quad \Rightarrow \quad t=\sqrt{1-k^2} \tag{II.2.72}$$

Finally, here the S-matrix for an ideal, lossless TEM directional coupler operating at the center frequency $f_c=\frac{v_p}{4\ell}$:

$$\mathbf{S}=\begin{bmatrix} 0 & -j\sqrt{1-k^2} & k & 0 \\ -j\sqrt{1-k^2} & 0 & 0 & k \\ k & 0 & 0 & -j\sqrt{1-k^2} \\ 0 & k & -j\sqrt{1-k^2} & 0 \end{bmatrix} \tag{II.2.73}$$

with k being the coupling coefficient:

$$0\leq k\leq \frac{1}{\sqrt{2}}$$

II.2.7.8 S-parameters in practice

In practice, the scattering parameters are a function of the frequency, $\mathbf{S}(f)$, but also some instruments or software applications can provide the time-domain equivalent by applying an inverse *Fourier* transform of the measured or calculated S-parameter data, sometimes referred as *synthetic pulse* time-domain analysis. There are basically only a few real-world, practical ways to acquire S-parameters:

- The characterisation of a RF component or subsystem by measurement, using a vector network analyzer (VNA). This is the most common case, with the instrument sweeping the DUT at a fixed stimulus RF power level over a range of frequencies $f_{\text{min}}<f<f_{\text{max}}$, usually in equidistant steps Δf.
- The S-parameters are supplied as data set by the manufacturer of the RF component or subsystem.
- Also, S-parameters are supplied by certain numerical simulation tools, like circuit analysing and

PCB software, e.g. *Keysight Pathwave* ADS, *Cadence* AWR Microwave Office and Allegro, Qucs, etc., or electromagnetic modeling software, e.g. *Dassault* CST Studio, *Ansys* HFSS, etc.

In all cases the S-parameter data is supplied by a standardised convention, as ASCII readable text file in the so-called *Touchstone* `SnP` format, which de-facto is the industry standard. In the *Touchstone* filename extension `filename.snp` the n indicates the number of ports of the S-parameter dataset.

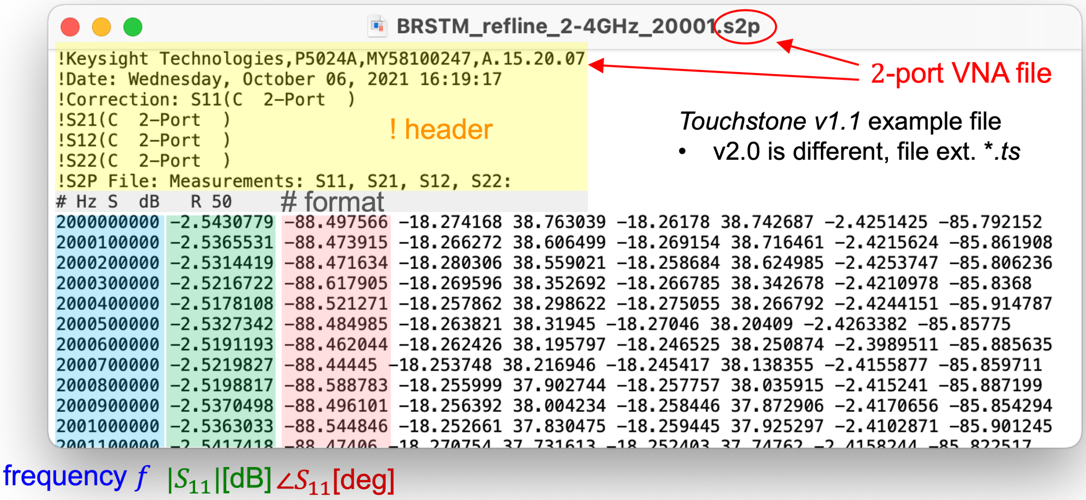


Fig. II.2.90: *Touchstone* `s2p` S-parameter file of a 2-port VNA measurement.

Figure II.2.90 shows the first couple of lines of a *Touchstone v1.1*[3] ASCII file, highlighting some of the key aspects:

- The file name extension specifies the number n of ports, that number is **not** equal to the number of columns in the file! Also please note, there might be some differences between `s1p`, `s2p`, `s3p` and `s4p` files wrt. the carriage return (CR) symbol.
- The comment header (!) includes some general information, e.g. type of instrument or software, date and time the analysis was performed, with or without correction and on which ports, and the S-parameter column order.
- The format line (#) defines the data format of the S-parameter, e.g. `mag[dB]/angle[deg]`, `mag/angle[rad]`, `real/imag`, the units for the frequency, e.g. `Hz`, `MHz`, `GHz`, and the reference impedance, e.g. 50 Ω.
- Following the # format line the actual S-parameter starts, with the first column being the frequency. For the given example Fig. II.2.90 the columns are
  `f[Hz] S11mag[dB] S11phase[deg] S21mag[dB] S21phase[deg]`
  `S12mag[dB] S12phase[deg] S22mag[dB] S22phase[deg].`
  Please note, the column delimiter may vary, e.g. space, comma, semicolon, etc.

[3] *Touchstone v2.0* uses a different format, has the file extension `*.ts` and is less common.

II.2.7.9 T-parameters

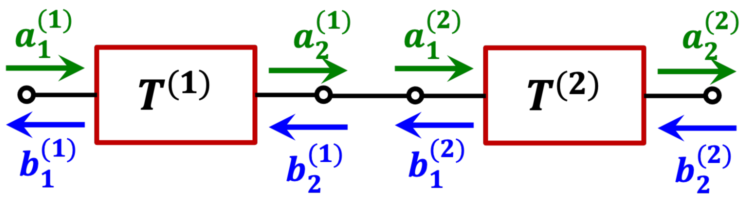


Fig. II.2.91: Cascading of two 2-port RF networks using T-parameters.

The S-parameters characterise linear, n-port RF networks by incident waves a_i travelling towards the ports and the reflected waves b_i travelling away from the ports, see Fig. II.2.74. The S-matrix expressed the dependent parameters – the reflected waves b_i – as linear function of the independent parameter – the incident waves a_i, Eq. (II.2.50). While this makes logical sense, it hinders a simple cascading of two or more 2-port networks defined by their S-matrices.

The *transfer* scattering parameters, *T-parameters*, circumvent this shortcoming and use a less intuitive definition of the power waves, expressing the waves at the (input) port 1 as function of the waves at the (output) port 2 for a 2-port RF network:

$$\begin{aligned} b_1 &= T_{11}a_2 + T_{12}b_2 \\ a_1 &= T_{21}a_2 + T_{22}b_2 \end{aligned} \qquad \text{with: } \mathbf{T} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \tag{II.2.74}$$

The T-parameter definition, Eq. (II.2.74), enables a cascading of 2-port networks by simple multiplication of the individual T-matrices, see also Fig. II.2.91:

$$\mathbf{T} = \mathbf{T}^{(1)}\mathbf{T}^{(2)} \dots \mathbf{T}^{(N)} \tag{II.2.75}$$

The conversion 2-port between S- and T-parameters is given as:

$$\begin{aligned} \mathbf{T} &= \frac{1}{S_{21}} \begin{bmatrix} -|\mathbf{S}| & S_{11} \\ -S_{22} & 1 \end{bmatrix} & \qquad \mathbf{S} &= \frac{1}{T_{22}} \begin{bmatrix} T_{12} & |\mathbf{T}| \\ 1 & -T_{21} \end{bmatrix} \\ &\text{with: } |\mathbf{S}| = S_{11}S_{22} - S_{12}S_{21} & &\text{with: } |\mathbf{T}| = T_{11}T_{22} - T_{12}T_{21} \end{aligned} \tag{II.2.76}$$

Please note:

- The multiplication of the cascading T-matrices in Eq. (II.2.75) is not commutative.
- The flow direction of the waves a_i, b_i, for T-parameters is different compared to the S-parameters, see Fig. II.2.91.
- There exists another definition for the scattering T-parameters, slightly different from Eq. (II.2.74).
- Sometimes the term *T-parameters* is used for other types of network parameters.

II.2.7.10 ABCD-parameters

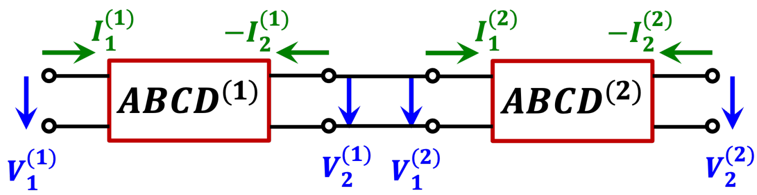


Fig. II.2.92: Cascading of two 2-port networks using ABCD-parameters.

Similar to the scattering T-parameters described in the previous paragraph, Section II.2.7.9, the *ABCD-parameters* – also referred as chain, cascade, or transmission parameters – are defined by the voltages and currents at the ports to enable a cascading of 2-port networks by matrix multiplication, see also Fig. II.2.92. Among the different definitions for ABCD 2-port parameters, this is the most popular one:

$$\begin{aligned} V_1 &= A\,V_2 - B\,I_2 \\ I_1 &= C\,V_2 - D\,I_2 \end{aligned} \qquad \text{with: } \mathbf{ABCD} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{II.2.77}$$

Please note the current arrow on port 2 being inverted, therefore defining the current at port 2 with a negative sign, $-I_2$, see Fig. II.2.92.

Like for the T-parameters, a complex system of two or more 2-port networks defined by their ABCD-parameters can be analysed by multiplication of their ABCD-matrices:

$$\mathbf{ACBD} = \mathbf{ABCD}^{(1)}\mathbf{ABCD}^{(2)}\ldots\mathbf{ABCD}^{(N)} \tag{II.2.78}$$

The conversion 2-port between S- and ABCD-parameters is given as:

$$\mathbf{ABCD} = \frac{1}{2S_{21}} \begin{bmatrix} (1+S_{11})(1-S_{22})+S_{12}S_{21} & [(1+S_{11})(1+S_{22})-S_{12}S_{21}]\,Z_0 \\ [(1-S_{11})(1-S_{22})-S_{12}S_{21}]\,/Z_0 & (1-S_{11})(1+S_{22})+S_{12}S_{21} \end{bmatrix}$$
$$\mathbf{S} = \frac{1}{A+B/Z_0+C\,Z_0+D} \begin{bmatrix} A+B/Z_0-C\,Z_0-D & 2(AD-BC) \\ 2 & -A+B/Z_0-C\,Z_0+D \end{bmatrix} \tag{II.2.79}$$

Example II.2.7.3: Stretched-wire measurement of the beam-coupling impedance

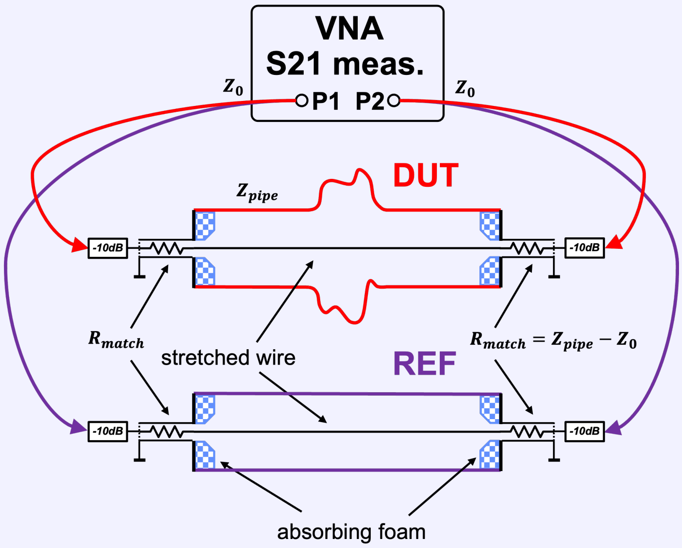


Fig. II.2.93: Schematic of a stretched-wire beam-coupling impedance measurement setup.

The characterisation of the longitudinal *beam-coupling impedance* $Z_{\|}$ of machine components is of vital importance in accelerator technology and beam dynamics. Among various ways to calculate or measure this impedance, the so-called "stretched-wire" measurement method plays a role to characterise the beam coupling-impedance in an effective way over a wide range of frequencies, here briefly mentioned in connection to the ABCD-parameters.
Figure II.2.93 illustrates the stretched-wire impedance measurement method for an arbitrary DUT, e.g. a beam pickup, kicker, bellow, etc., utilising a thin wire stretched through the symmetric centre of a beam pipe, thus composing a coaxial TEM transmission-line of characteristic impedance:

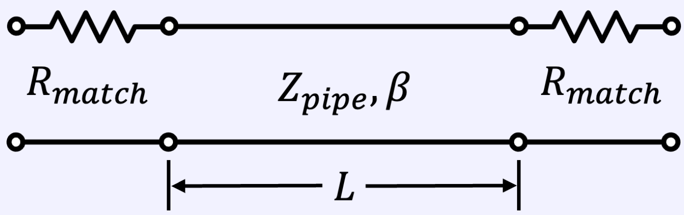
$$Z_{\text{pipe}} \approxeq 60 \ln \frac{D_{\text{pipe}}}{d_{\text{wire}}} \tag{II.2.80}$$

The method requires two S_{21} measurements, performed with a VNA in the frequency range on interest, $S_{21_{DUT}}$ on the device-under-test and $S_{21_{REF}}$ on an unperturbed reference line of same physical length L, leading to:

$$S_{21} = \frac{S_{21_{DUT}}}{S_{21_{REF}}} \tag{II.2.81}$$

The measurement results, Eq. (II.2.81), directly leads to longitudinal beam-coupling impedance:

$$Z_{\|} = -2\, Z_{\text{pipe}} \ln S_{21} \left(1 + j\frac{\ln S_{21}}{2\Theta}\right) \qquad \text{with: } \Theta = 2\pi\frac{L}{\lambda} \tag{II.2.82}$$



$$\begin{aligned} A &= 1 \\ B &= R_{match} \\ C &= 0 \\ D &= 1 \end{aligned} \qquad \begin{aligned} A &= \cos\Theta \\ B &= jZ_{\text{pipe}}\sin\Theta \\ C &= jY_{\text{pipe}}\sin\Theta \\ D &= \cos\Theta \end{aligned} \qquad \begin{aligned} A &= 1 \\ B &= R_{match} \\ C &= 0 \\ D &= 1 \end{aligned}$$

Fig. II.2.94: Schematic of the reference (REF) pipe with matching resistors R_{match} and the related ABCD-parameters for the individual 2-port networks.

As of the beam pipe and wire dimensions, $Z_{\text{pipe}} \gg Z_0$, typically Z_{pipe} is in the order of $200 \ldots 300\,\Omega$, while $Z_0 = 50\,\Omega$. To minimize uncontrolled reflections several impedance-matching improvements need to be applied, for the lower frequency range a lumped, low-inductance resistor of value

$$R_{match} = Z_{\text{pipe}} - Z_0$$

is used in series with Z_0 to provide the correct termination impedance to Z_{pipe}.
If a reference pipe is unavailable, e.g. because too complicated, expensive, cumbersome, etc., the $S_{21_{REF}}$ can be calculated based on cascading three 2-port networks based on the chain ABCD-parameters, see the equivalent schematic Fig. II.2.94. Following Eq. (II.2.78) we have to multiply:

$$\mathbf{ABCD}_{REF} = \begin{bmatrix} A_{REF} & B_{REF} \\ C_{REF} & D_{REF} \end{bmatrix} = \begin{bmatrix} 1 & R_{match} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\Theta & jZ_{\text{pipe}}\sin\Theta \\ jY_{\text{pipe}}\sin\Theta & \cos\Theta \end{bmatrix} \begin{bmatrix} 1 & R_{match} \\ 0 & 1 \end{bmatrix}$$

and then applying the $\mathbf{ABCD}$-to-$\mathbf{S}$ conversion of Eq. (II.2.79) to calculate

$$S_{21_{REF}} = \frac{2}{A_{REF} + B_{REF}/Z_0 + C_{REF}\, Z_0 + D_{REF}}$$

II.2.7.11 General n-port networks

A general n-port network may include different technologies, e.g. a combination of waveguides (rectangular, circular, elliptical) and TEM transmission-lines, such as coaxial lines, microstrip lines, etc., providing the ports. Figure II.2.95 illustrates on the left side a general RF network which shows 3 physical ports:

1. A TEM coaxial transmission-line
2. A rectangular waveguide
3. A circular waveguide

Using different port technologies results in different frequency ranges for the transmission of RF power into / out of the RF network. The frequency / mode chart in Fig. II.2.96 illustrates the situation for the example Fig. II.2.95.

At low frequencies only the coaxial TEM transmission-line transmits RF power, while both waveguides are used below the cut-off frequency of their fundamental modes. The physical 3-port network operates as a modal (or logical) 1-port network. This case is quite common in the analysis or measurement of resonant cavities, with a resonant frequency f_{res} (mostly the TM_{010} mode for standing wave cylindrical resonators) of the cavity well below the waveguide cut-off frequency of the beam pipe ports, i.e. $f_{TE_{11}}$ for the TE_{11} (H_{11}) mode for beam pipes with circular cross-section, $f_{res} < f_{TE_{11}}$.

At higher frequencies, in our example Fig. II.2.96, the fundamental H_{10} (TE_{10}) mode of the rectangular waveguide comes into play, and now we have a modal 2-port network.

Further increasing the operation frequency of the general RF network enables the H_{11} (TE_{11}) fundamental mode of the circular waveguide to propagate, see Fig. II.2.96. As this mode has two orthogonal polarisation, H_{11}^x, H_{11}^y (TE_{11}^x, TE_{11}^y), and we need to account for two additional modes, which now makes the physical 3-port network a modal 4-port!

At even higher frequency the rectangular waveguide in our example starts to propagate the higher-order H_{20} (TE_{20}) waveguide mode, in parallel to the fundamental H_{10} (TE_{10}) mode, thus adds another modal port. Operating the physical 3-port network at those frequencies results in a 5-port modal network, which is illustrated on the right side of Fig. II.2.95. Of course, in practice to operate a waveguide in a over-moded frequency regime is always questionable! Interestingly, there is no situation where this physical 3-port network actually has 3-ports!

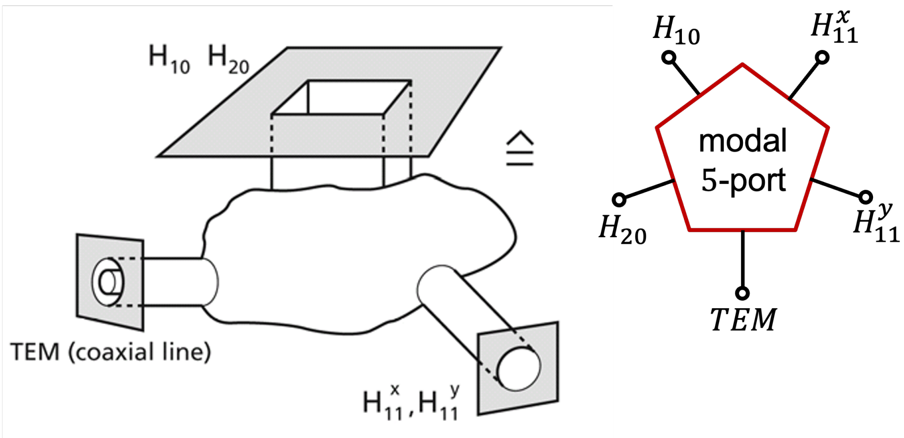


Fig. II.2.95: Example of a general n-port network with 3 physical, but 5 or more modal ports.

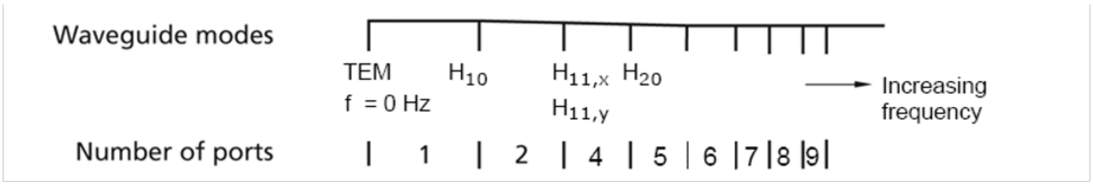


Fig. II.2.96: Frequency / mode chart for the general RF network, Fig. II.2.95.

The situation of general n-port networks often appears in numerical electromagnetic simulations. It is of utmost importance that the energy of **ALL** modes is absorbed at the ports, i.e. each EM-mode must be terminated by a distinct modal port! Please note, also TEM transmission-lines (coaxial lines) can propagate higher-order modes at high frequencies, which then need to be terminated as well!

II.2.7.12 Scattering parameter summary

The scattering S-parameters are used to characterise linear, time-invariant RF networks as function of the frequency. The inverse *Fourier* transformation allows to calculate the related time-domain behaviour of the system, however, some caveats have to be considered, e.g. $\sin(x)/x$ artefacts, some will be covered in the RF measurement techniques Section II.2.9. Today, most RF components and subsystems are described by their S-matrix, either by a measurement or by a numerical simulation, using the ASCII *Touchstone* SnP file format. It is of utmost importance to terminate unused ports in their characteristic impedance during the measurement or simulation, including "modal" ports caused my higher-order port modes or their polarisation.

In this introduction we could only cover some of the main aspects of S-parameters and their application. More advanced techniques and applications, e.g. on the signal flow graph (SFG), using S-parameters to analyse the stability of RF amplifiers, etc. are found in the literature.

Fig. II.2.97: A RF System-on-Chip (RF-SoC).

II.2.8 RF components, devices and subsystems

Any attempt for a comprehensive list of RF components, devices and subsystems in this chapter would fail, without speaking about a more detailed discussion. Instead, we try to cover a few selected RF devices which are typical and / or popular in accelerator RF systems, with the focus on their basic functional principle.

All of the examples of RF devices in this chapter are based on "classic" analog RF technologies, i.e. distributed and lumped elements, linear and / or non-linear networks and building blocks, having one, two or many ports. However, it should be noted than many of those RF functions can be replaced by equivalent, often better, performing digital functions. While the evolution in analogue RF technologies is slowing down, the advances on digital RF technologies are making substantial progress, at the time of writing, e.g. the RF System-on-Chip (RF-SoC), see Fig. II.2.97, seems to have the potential to replace many, if not all, the traditional components in a low-level RF (LLRF) system.

On the other hand, digital RF functions are always limited to low power levels, a few dBm's, therefore RF components and subsystems for higher power levels remain an analog domain.

II.2.8.1 RF components based on transmission-lines

While lumped elements, such as resistors, capacitors, inductors, semiconductors, etc. are used in general purpose, power, analog, digital and RF electronics, distributed elements, like transmission-lines and resonant structures, are typically found only in RF networks. The following examples of RF components are limited to TEM transmission-lines.

II.2.8.1.1 Low-loss/lossless transmission-line as phase-shifter

A reasonable short piece of a low-loss transmission-line can be assumed as *lossless*, i.e. with the attenuation constant $\alpha = 0$. For a *matched* transmission-line, Eq. (II.2.61) simplifies to:

$$\begin{aligned} S_{11} &= S_{22} = 0 \\ S_{21} &= S_{12} = e^{-j\beta\ell} \end{aligned} \tag{II.2.83}$$

In other words, a lossless or low-loss transmission line changes the phase of the output port wrt. the input port by $\theta = \beta\ell$, while the magnitude (amplitude) ratio stays at 1. Therefore, the most simple, intuitive application of a transmission-line is as *phase shifter* or *delay-line*. Figure II.2.98 shows some examples

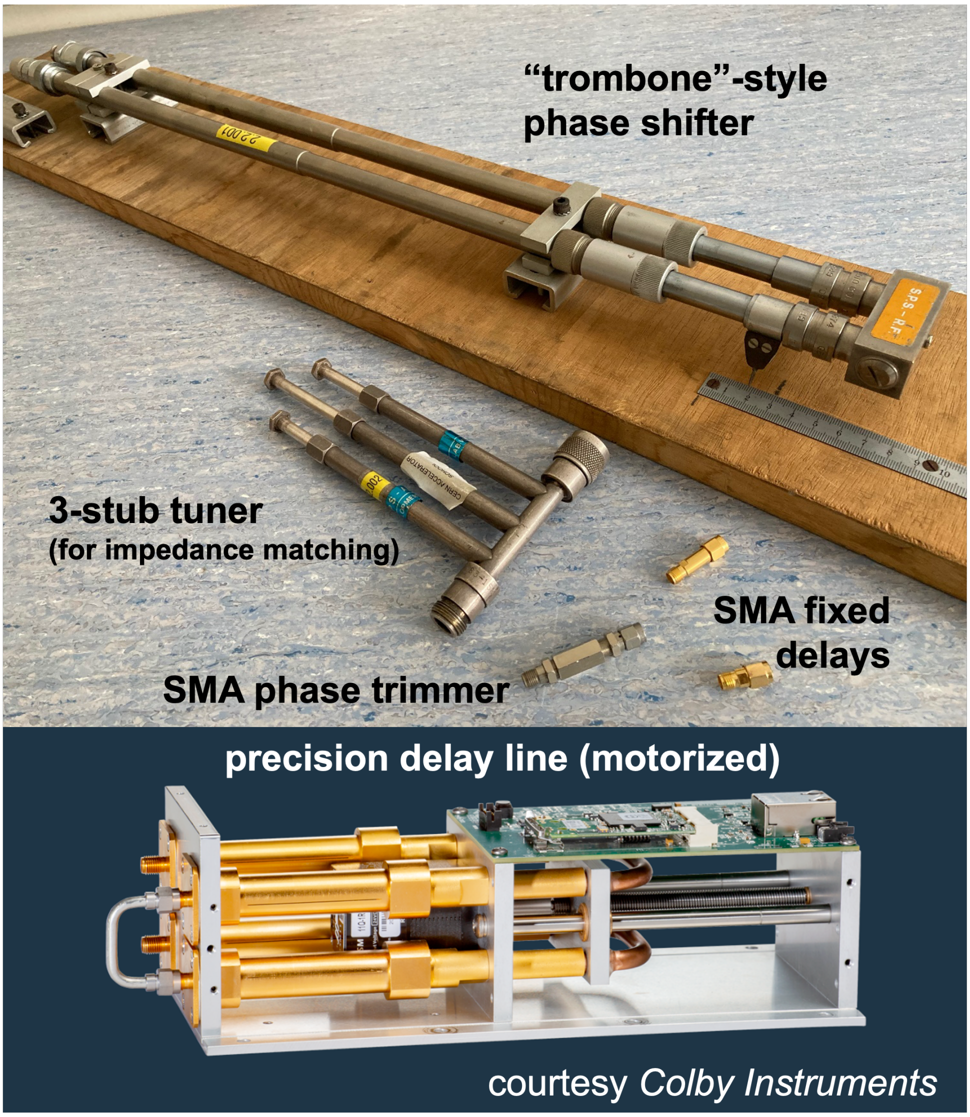


Fig. II.2.98: Coaxial TL components: Phase-shifters (delay-lines) and a 3-stub tuner.

of coaxial phase shifters / delay-lines, most of them with variable length $\ell_{\text{min}} < \Delta\ell < \ell_{\text{max}}$. We can use the variable length transmission-line to shift the phase by $\Delta\theta$ of a CW RF signal of frequency f:

$$\theta = \theta_{\text{min}} + \Delta\theta = \frac{2\pi f(\ell_{\text{min}} + \Delta\ell)}{v_p}$$

Most popular is the "trombone" style coaxial phase shifter. Sometimes they are motorised and used to remotely adjust the phase of medium or high-power RF wrt. to the accelerating structures (beam phase). At the JUAS RF hands-on training days we use the trombone shifter shown in Fig. II.2.98 to let the students verify $v_p \approx c$ for this "air" coaxial-line (air is the dielectric between the coaxial conductors). For small phase adjustments of low-level RF signals SMA-type "phase trimmers" are used, also little "barrel" coaxial elements are popular for a fixed phase delay.

Of course, the variable transmission-line can also be used as time-delay for broadband signals:

$$t_{delay} = t_{\text{min}} + \Delta t = \frac{\ell_{\text{min}} + \Delta\ell}{v_p}$$

As a matter of fact, any coaxial cable acts as time-delay, and at CERN many standard RF cables have a length with a specific signal delay value indicated on the cable.

> Please note, the time-delay of longer cables is frequency dependent (dispersion effects) also, due to the frequency dependent losses, the shape of the signal waveform is altered, i.e. longer cables cannot anymore be assumed as low-loss or lossless!

II.2.8.1.2 Transmission-line terminated with an arbitrary load

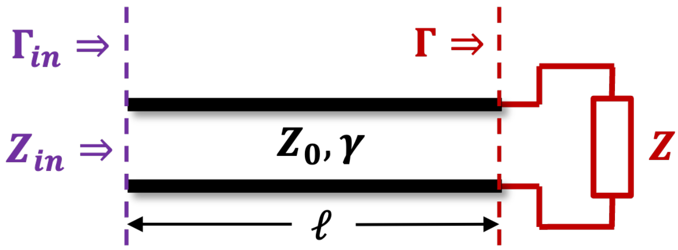


Fig. II.2.99: Transmission-line with arbitrary load.

A TEM transmission-line, defined by its characteristic impedance Z_0 and propagation constant $\gamma = \alpha + j\beta$ is terminated with an arbitrary, complex impedance Z, see also Fig. II.2.99. This load impedance Z is transformed through the transmission-line to a different impedance Z_{in}, seen at the input of the line:

$$\begin{aligned} Z_{\text{in}} &= Z_0 \frac{1+\Gamma_{\text{in}}}{1-\Gamma_{\text{in}}} \qquad \text{with: } \Gamma_{\text{in}} = \Gamma e^{-2\gamma\ell} \quad \text{and} \quad \Gamma = \frac{Z - Z_0}{Z + Z_0} \\ \Rightarrow \quad Z_{\text{in}} &= Z_0 \frac{Z\cosh\gamma\ell + Z_0\sinh\gamma\ell}{Z_0\cosh\gamma\ell + Z\sinh\gamma\ell} = Z_0 \frac{Z + Z_0\tanh\gamma\ell}{Z_0 + Z\tanh\gamma\ell} \end{aligned} \tag{II.2.84}$$

In many non-resonant applications, using short transmission-lines, the line can be approximated as *lossless*:

$$\begin{aligned} \alpha = 0 \quad &\Rightarrow \quad \gamma = j\beta \\ \Rightarrow \quad Z_{\text{in}} = Z_0 \frac{Z\cos\beta\ell + Z_0\sin\beta\ell}{Z_0\cos\beta\ell + Z\sin\beta\ell} &= Z_0 \frac{Z + Z_0\tan\beta\ell}{Z_0 + Z\tan\beta\ell} \end{aligned} \tag{II.2.85}$$

Popular applications for lossless transmission-lines are:

Quarter-wave transmission-line transformer:

$$\ell = \frac{\lambda}{4} \quad \Rightarrow \quad \theta = \beta\ell = \frac{\pi}{2} \quad \Rightarrow \quad Z_{\text{in}} = \frac{Z_0^2}{Z} \tag{II.2.86}$$

Terminated (matched) transmission-line:

$$Z = Z_0 \quad \Rightarrow \quad Z_{\text{in}} = Z_0 \tag{II.2.87}$$

Open transmission-line:

$$Z \to \infty \quad \Rightarrow \quad Z_{\text{in}} = -jZ_0 \cot \beta\ell \tag{II.2.88}$$

Short transmission-line:

$$Z = 0 \quad \Rightarrow \quad Z_{\text{in}} = jZ_0 \tan \beta\ell \tag{II.2.89}$$

Example II.2.8.1: Meaning of the *electrical length* θ of a transmission-line

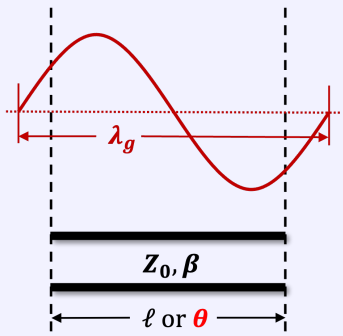


Fig. II.2.100: Physical length ℓ and electrical length θ of a lossless TEM transmission-line.

A lossless TEM transmission-line (TL) is fully characterised by its characteristic impedance Z_0 and its propagation constant $\gamma = j\beta$, were Z_0 is defined by the cross-section geometry and dimensions. The phase constant β, also called the *wave number* k, expresses the characteristic length of the TL in terms of an angle per unit length for a given guide wavelength λ_g:

$$\beta = \omega\sqrt{L'C'} = \frac{2\pi}{\lambda_g} = k \quad [\frac{rad}{m}] \tag{II.2.90}$$

For a TL of a specific physical length ℓ we then define the electrical length θ as:

$$\beta\ell = \theta = \frac{2\pi\ell}{\lambda_g} = \frac{2\pi f\ell}{v_p} = \frac{\omega\ell}{v_p} \tag{II.2.91}$$

In other words:
The electrical length θ expresses the length of a TL in terms of a phase angle for the guide wavelength λ_g at the operating frequency f, e.g. for $\theta = 2\pi$ the physical length ℓ of the TL is such that one full oscillation of the frequency $f = v_p/\lambda_g$ fits on the line. Figure II.2.100 tries to illustrate this link between the physical and the electrical length of a TL.

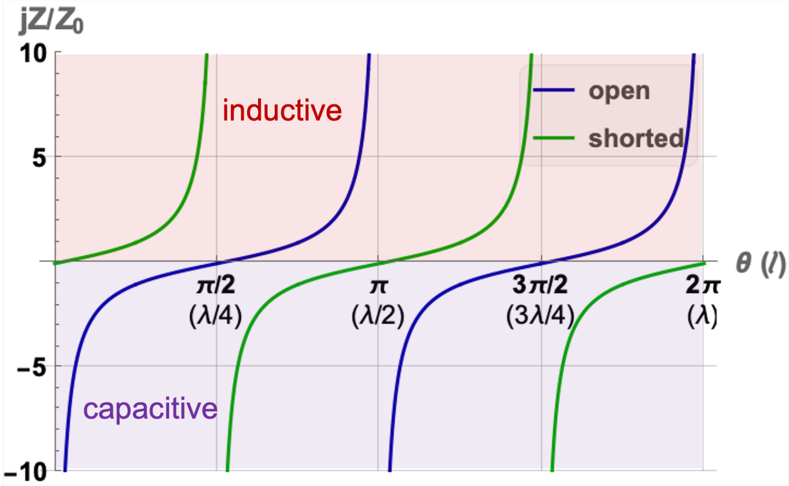


Fig. II.2.101: Normalized input impedance of a transmission-line, terminated with an open or a short.

Table II.2.10: Open or shorted lossless transmission-line, acting "inductive" or "capacitive" at the input.

physical length ℓ **electrical length** θ	$0 < \ell < \lambda_g/4$ $0 < \theta < \pi/2$	$\lambda_g/4 < \ell < \lambda_g/2$ $\pi/2 < \theta < \pi$	$\lambda_g/2 < \ell < 3\lambda_g/4$ $\pi < \theta < 3\pi/2$	$3\lambda_g/4 < \ell < \lambda_g$ $3\pi/2 < \theta < 2\pi$
lossless TL, open	"capacitive"	"inductive"	"capacitive"	"inductive"
lossless TL, shorted	"inductive"	"capacitive"	"inductive"	"capacitive"

II.2.8.1.3 Open / shorted transmission-line

Following Eq. (II.2.88) and Eq. (II.2.89), the input impedance Z_{in} of a lossless transmission-line, see also Fig. II.2.101, is either "inductive" or "capacitive", depending on the length of the line. Table II.2.10 summarises the facts for short, lossless TEM transmission-lines, i.e. a lossless transmission-line with open ($Z \to \infty$) or short ($Z = 0$) termination acts a lumped reactive element, capacitor or inductor, at the operating frequency.

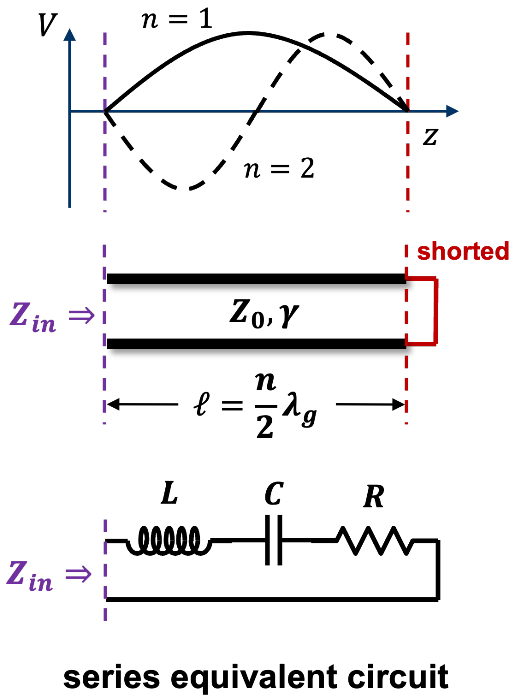
Please remember:

– A "capactive" element has the form: $Z_C = \dfrac{1}{j\omega C} = -j\dfrac{1}{\omega C}$

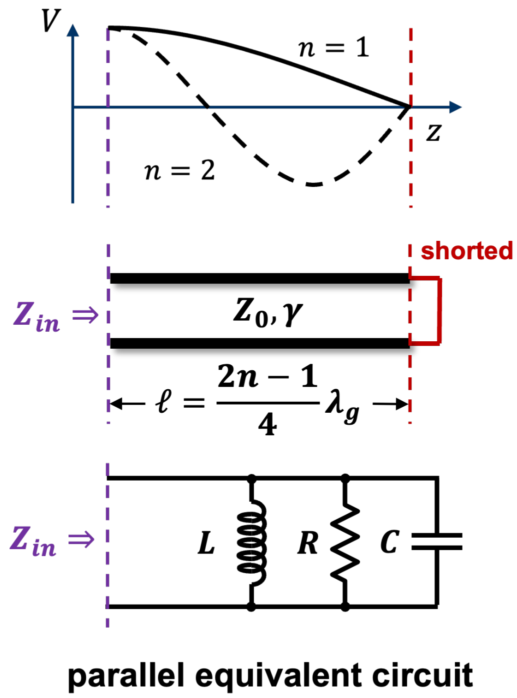
– An "inductive" element has the form: $Z_L = j\omega L$

II.2.8.1.4 Transmission-line resonators

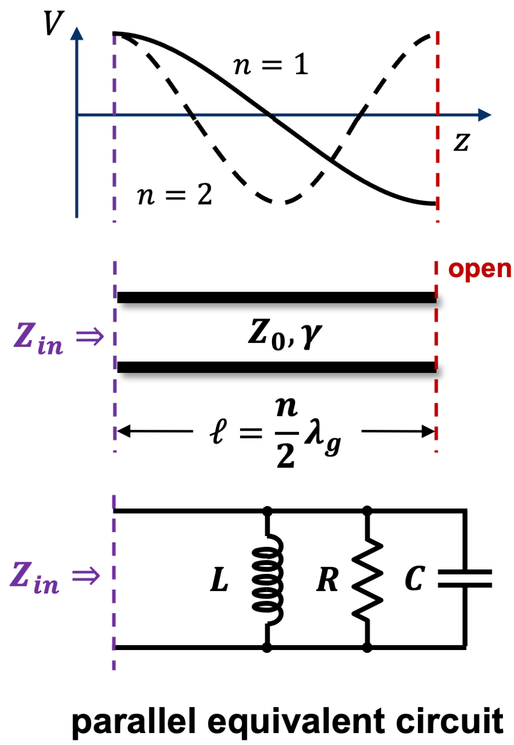
A short section of a TEM transmission-line can also be used as a *resonator*, for use in a RF circuit, as well as for the application to accelerate charged particles. To operate as resonator, the boundary conditions at both ends of the transmission-line need to be appropriate wrt. the guide wavelength. That fact results in four cases, *half-wave*, $n\lambda/2$, and *quarter-wave*, $(2n-1)\lambda/4$, resonators with shorted or open end



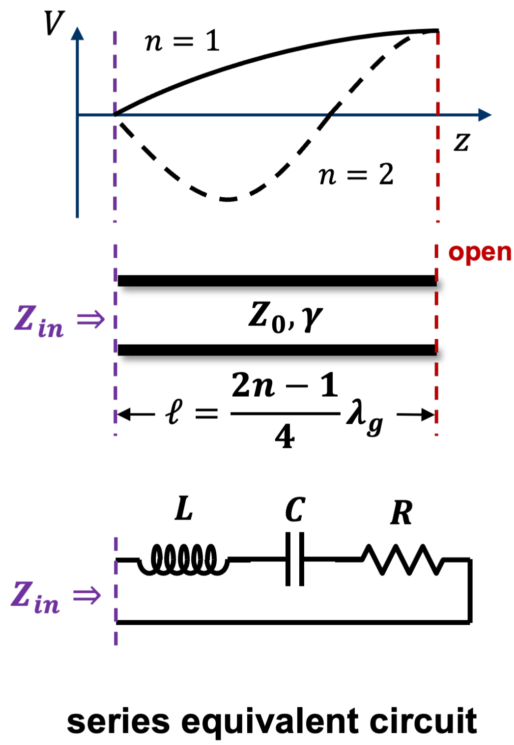
(a) half-wave, shorted.



(b) quarter-wave, shorted.



(c) half-wave, open.



(d) quarter-wave, shorted.

Fig. II.2.102: Transmission-line resonators.

termination, as summarised in Fig. II.2.102.
The upper part of each case illustrates the voltage standing waves along the transmission-line given by the shorted or open termination condition at the end of the line for the fundamental mode, $n = 1$, and the dashed trace for the first higher-order mode, $n = 2$. Below, the middle part, shows the shorted or open TEM transmission-line itself and in the lower, bottom part the appropriate lumped element equivalent circuit is given.

For the analysis, we need to take the losses into account, but we assume them to be small:

$$\alpha\ell \ll 1 \quad \Rightarrow \quad \tanh\alpha\ell \approxeq \alpha\ell \tag{II.2.92}$$

Furthermore we assume the transmission-line resonator is operated at or near the resonance:

$$\omega = \omega_0 + \Delta\omega \tag{II.2.93}$$

with ω_0 being the resonant angular frequency and $\Delta\omega$ being a small deviation. The analysis reveals the parameters for the lumped elements of the equivalent circuit, as shown in Fig. II.2.102, shown here as example for shorted half-wave resonator:

II.2.8.1.4.1 Half-wave resonator, shorted end

Input impedance, see also Eq. (II.2.84) with $Z = 0$

$$Z = 0 \quad \Rightarrow \quad Z_{\text{in}}^{s} = Z_0 \tanh\left(\alpha + j\beta\right)\ell = Z_0 \frac{\tanh\alpha\ell + j\tan\beta\ell}{1 + j\tanh\alpha\ell\tan\beta\ell} \tag{II.2.94}$$

With:

$$\ell = \frac{\lambda}{2} = \pi\frac{v_p}{\omega_0} \quad \Rightarrow \quad \beta\ell = \pi + \pi\frac{\Delta\omega}{\omega_0} \quad \Rightarrow \quad \tan\beta\ell = \tan\left(\pi + \pi\frac{\Delta\omega}{\omega_0}\right) = \tan\left(\pi\frac{\Delta\omega}{\omega_0}\right) \approxeq \pi\frac{\Delta\omega}{\omega_0}$$

follows for the input impedance of the shorted half-wave TL resonator:

$$Z_{\text{in}}^{2s} \approxeq Z_0 \frac{\alpha\ell + j\pi\frac{\Delta\omega}{\omega_0}}{1 + j\underbrace{\alpha\ell\pi\frac{\Delta\omega}{\omega_0}}_{\approxeq 0}} \approxeq Z_0\left(\alpha\ell + j\pi\frac{\Delta\omega}{\omega_0}\right) \tag{II.2.95}$$

and has the form:

$$Z_{\text{in}}^{2s} = R + j2\Delta\omega L = R + j\left(\omega L - \frac{1}{\omega C}\right) \tag{II.2.96}$$

which leads to the values for the lumped elements of the series equivalent circuit (Fig. II.2.102a):

$$R = Z_0\alpha\ell; \qquad L = \frac{\pi}{2}\frac{Z_0}{\omega_0}; \qquad C = \frac{1}{\omega_0^2 L}$$

The unloaded Q-factor for this resonator follows as:

$$Q_0 = \frac{\omega_0 L}{C} = \frac{\pi}{2\alpha\ell} = \frac{\beta}{2\alpha}$$

Please note, due to the periodic behaviour of the TL resonator, the resonances appear at:

$$\ell = \frac{n}{2}\lambda_g, \quad \text{with: } n = 1, 2, 3, \ldots$$

Superconducting half-wave resonators (SC-HWR) became popular as accelerating structures for low-β heavy ion linacs. Figure II.2.103 shows an example of a 322.5 MHz SC-HWR for the *Facility of Rare Isotope Beams* (FRIB). As the section view indicates, the coaxial TL resonator is shorted at both ends and has its maximum E-field in the centre of the structure. Therefore, the beam-ports and a hole in the centre conductor are located in this symmetry plane, allowing the beam to pass and be accelerated by the E-field peaking at the two gaps.

The following three paragraphs summarise the formalism for the other TL resonator cases shown in Figs. II.2.102b to II.2.102d for completeness.

II.2.8.1.4.2 Quarter-wave resonator, shorted end

Input impedance:

$$Z = 0 \quad \Rightarrow \quad Z_{\text{in}}^{s} = Z_0 \tanh\left(\alpha + j\beta\right)\ell = Z_0 \frac{1 - j\tanh\alpha\ell\cot\beta\ell}{\tanh\alpha\ell - j\cot\beta\ell} \tag{II.2.97}$$

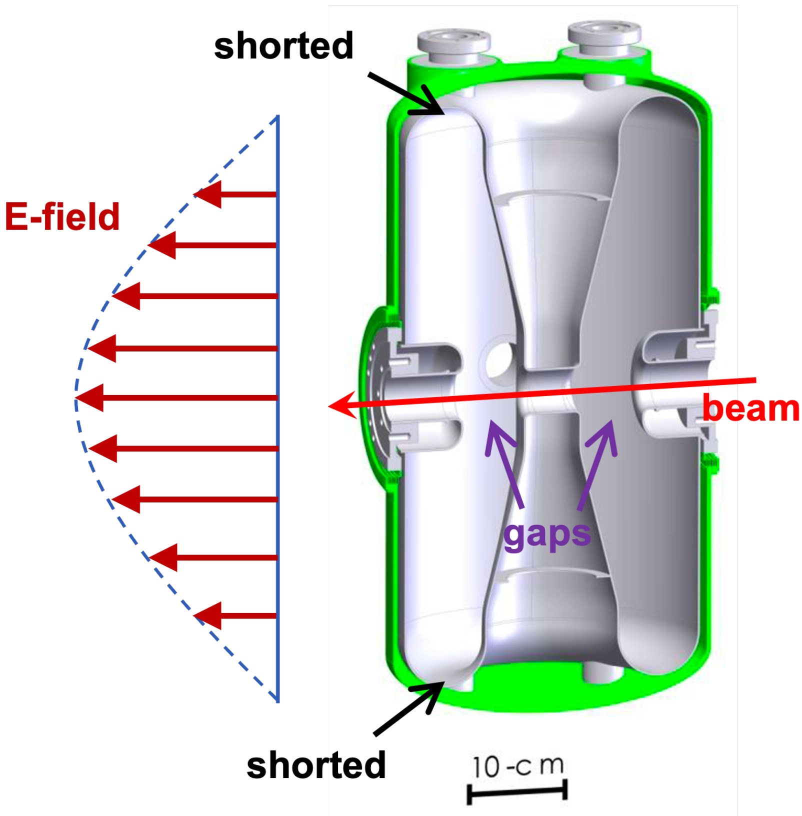


Fig. II.2.103: Sectional view of a 322.5 MHz SRF half-wave resonator for FRIB (courtesy *J. P. Holzbauer*). The relative E-field strength is sketched left of the cavity for illustration.

With:

$$\ell = \frac{\lambda}{4} = \frac{\pi}{2}\frac{v_p}{\omega_0} \quad \Rightarrow \quad \beta\ell = \frac{\pi}{2} + \frac{\pi}{2}\frac{\Delta\omega}{\omega_0} \quad \Rightarrow \quad \cot\beta\ell = -\tan\left(\frac{\pi}{2}\frac{\Delta\omega}{\omega_0}\right) \approxeq -\frac{\pi}{2}\frac{\Delta\omega}{\omega_0}$$

follows for the input impedance of the shorted quarter-wave TL resonator:

$$Z_{\text{in}}^{4s} \approxeq Z_0 \frac{1 + j\,\overbrace{\alpha\ell\frac{\pi}{2}\frac{\Delta\omega}{\omega_0}}^{\approxeq 0}}{\alpha\ell + j\frac{\pi}{2}\frac{\Delta\omega}{\omega_0}} \approxeq \frac{Z_0}{\alpha\ell + j\frac{\pi}{2}\frac{\Delta\omega}{\omega_0}} \tag{II.2.98}$$

It has the form:

$$Z_{\text{in}}^{4s} = \frac{1}{1/R + j2\Delta\omega C} = \frac{1}{\frac{1}{R} - j\frac{1-\omega^2 LC}{\omega L}} \tag{II.2.99}$$

which leads to the values for the lumped elements of the parallel equivalent circuit (Fig. II.2.102b):

$$R = \frac{Z_0}{\alpha\ell}; \qquad C = \frac{\pi}{4}\frac{1}{\omega_0 Z_0}; \qquad L = \frac{1}{\omega_0^2 C}$$

The unloaded Q-factor for this resonator follows as:

$$Q_0 = \omega_0 RC = \frac{\pi}{4\alpha\ell} = \frac{\beta}{2\alpha}$$

The resonances appear at:

$$\ell = \frac{2n-1}{4}\lambda_g, \quad \text{with: } n = 1, 2, 3, \ldots$$

II.2.8.1.4.3 Half-wave resonator, open end

Input impedance:

$$Z = 0 \quad \Rightarrow \quad Z_{\text{in}}^{o} = Z_0 \coth\left(\alpha + j\beta\right)\ell = Z_0 \frac{1 + j\tanh\alpha\ell\tan\beta\ell}{\tanh\alpha\ell + j\tan\beta\ell} \tag{II.2.100}$$

With:

$$\ell = \frac{\lambda}{2} = \pi\frac{v_p}{\omega_0} \quad \Rightarrow \quad \beta\ell = \pi + \pi\frac{\Delta\omega}{\omega_0} \quad \Rightarrow \quad \tan\beta\ell = \tan\left(\pi + \pi\frac{\Delta\omega}{\omega_0}\right) = \tan\left(\pi\frac{\Delta\omega}{\omega_0}\right) \approxeq \pi\frac{\Delta\omega}{\omega_0}$$

follows for the input impedance of the open half-wave TL resonator:

$$Z_{\text{in}}^{2o} \approxeq Z_0 \frac{1 + j\,\overbrace{\alpha\ell\pi\frac{\Delta\omega}{\omega_0}}^{\approxeq 0}}{\alpha\ell + j\pi\frac{\Delta\omega}{\omega_0}} \approxeq \frac{Z_0}{\alpha\ell + j\pi\frac{\Delta\omega}{\omega_0}} \tag{II.2.101}$$

It has the form:

$$Z_{\text{in}}^{2o} = \frac{1}{1/R + j2\Delta\omega C} = \frac{1}{\frac{1}{R} - j\frac{1-\omega^2 LC}{\omega L}} \tag{II.2.102}$$

which leads to the values for the lumped elements of the parallel equivalent circuit (Fig. II.2.102c):

$$R = \frac{Z_0}{\alpha\ell}; \qquad C = \frac{\pi}{2}\frac{1}{\omega_0 Z_0}; \qquad L = \frac{1}{\omega_0^2 C}$$

The unloaded Q-factor for this resonator follows as:

$$Q_0 = \omega_0 RC = \frac{\pi}{2\alpha\ell} = \frac{\beta}{2\alpha}$$

The resonances appear at:

$$\ell = \frac{n}{2}\lambda_g, \quad \text{with: } n = 1, 2, 3, \ldots$$

II.2.8.1.4.4 Quarter-wave resonator, open end

Input impedance:

$$Z = 0 \quad \Rightarrow \quad Z_{\text{in}}^{o} = Z_0 \coth\left(\alpha + j\beta\right)\ell = Z_0 \frac{\tanh\alpha\ell - j\cot\beta\ell}{1 - j\tanh\alpha\ell\cot\beta\ell} \tag{II.2.103}$$

With:

$$\ell = \frac{\lambda}{4} = \frac{\pi}{2}\frac{v_p}{\omega_0} \quad \Rightarrow \quad \beta\ell = \frac{\pi}{2} + \frac{\pi}{2}\frac{\Delta\omega}{\omega_0} \quad \Rightarrow \quad \cot\beta\ell = -\tan\left(\frac{\pi}{2}\frac{\Delta\omega}{\omega_0}\right) \approxeq -\frac{\pi}{2}\frac{\Delta\omega}{\omega_0}$$

follows for the input impedance of the open quarter-wave TL resonator:

$$Z_{\text{in}}^{4o} \approxeq Z_0 \frac{\alpha\ell + j\frac{\pi}{2}\frac{\Delta\omega}{\omega_0}}{1 + j\underbrace{\alpha\ell\frac{\pi}{2}\frac{\Delta\omega}{\omega_0}}_{\approxeq 0}} \approxeq Z_0\left(\alpha\ell + j\frac{\pi}{2}\frac{\Delta\omega}{\omega_0}\right) \tag{II.2.104}$$

It has the form:

$$Z_{\text{in}}^{4o} = R + j2\Delta\omega L = R + j\left(\omega L - \frac{1}{\omega C}\right) \tag{II.2.105}$$

which leads to the values for the lumped elements of the series equivalent circuit (Fig. II.2.102d):

$$R = Z_0\alpha\ell; \qquad L = \frac{\pi}{4}\frac{1}{\omega_0 Z_0}; \qquad C = \frac{1}{\omega_0^2 L}$$

The unloaded Q-factor for this resonator follows as:

$$Q_0 = \frac{\omega_0 L}{C} = \frac{\pi}{4\alpha\ell} = \frac{\beta}{2\alpha}$$

The resonances appear at:

$$\ell = \frac{2n-1}{4}\lambda_g, \quad \text{with: } n = 1, 2, 3, \ldots$$

Figure II.2.104 shows an example of a so-called *interdigital-line* band-pass filter, which is based on coupled, quarter-wave resonators. The resonators are implemented as $\lambda/4$ air-stripline elements, which keeps the losses low, and have an adjustable capacitive loading at the open ends, which allows one to fine-tune the resonant frequency. The filter approximates a *Bessel* filter function, i.e. a constant phase vs.

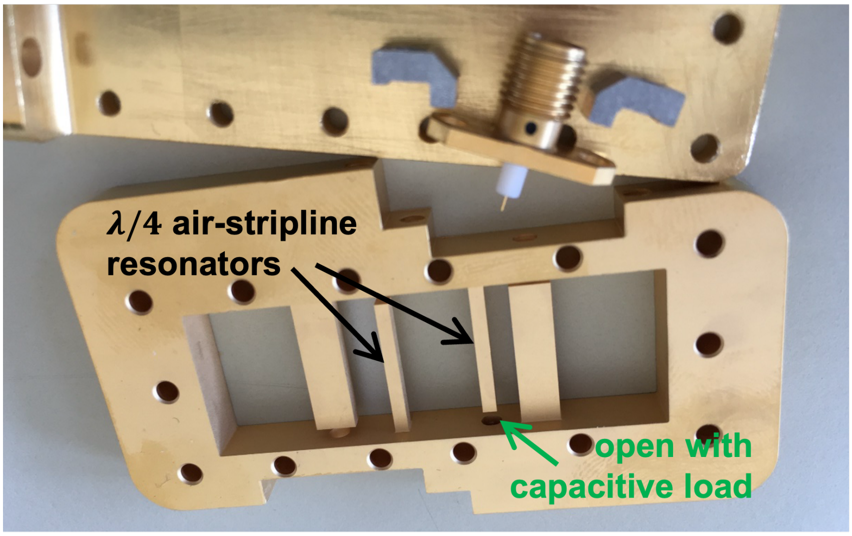


Fig. II.2.104: 4.8 GHz band-pass filter based on coupled quarter-wave resonators.

frequency, and is used for the LHC *Schottky* monitor. As the resonators exhibit higher-order eigenmodes, RF filters like this one develop so-called *spurious passbands* at multiples of the fundamental resonant frequency, at ≈14.4 GHz, etc.

More details on the theory of coupled resonators and RF/microwave filters can be found in textbooks like [3, 16, 17].

II.2.8.1.5 TEM Transmission-lines as semi-lumped elements

Industry deliver high-quality passive lumped circuit elements, which approximate the ideal lumped elements, see also Table II.2.9, e.g. a coil-like *inductor* with the inductance L, a single or multi-layer

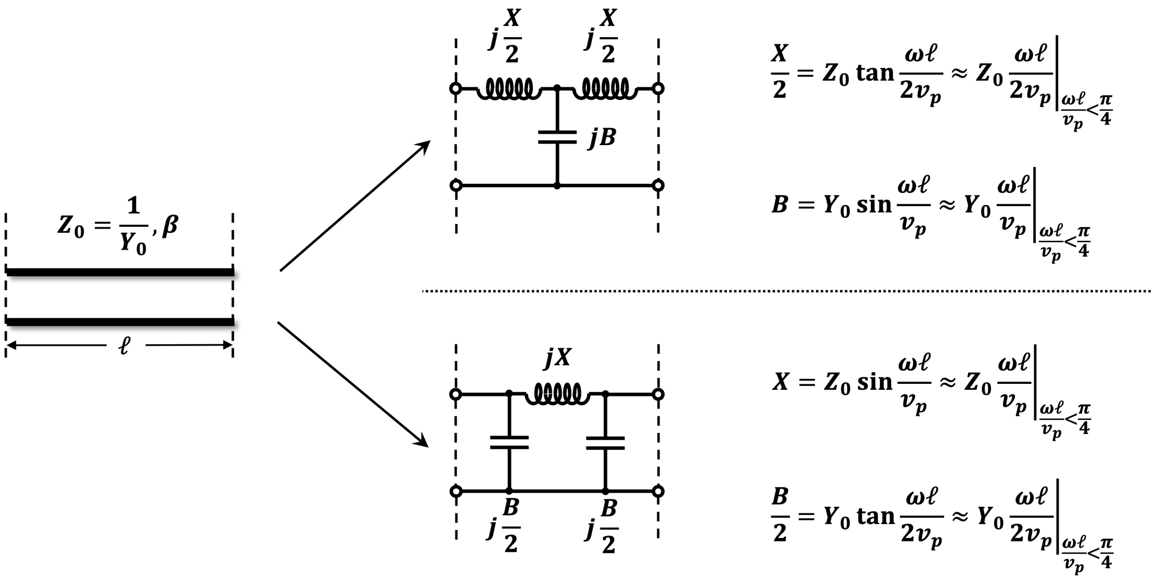


Fig. II.2.105: Equivalent lumped element models for short TEM transmission-lines, $\theta < \pi/4$.

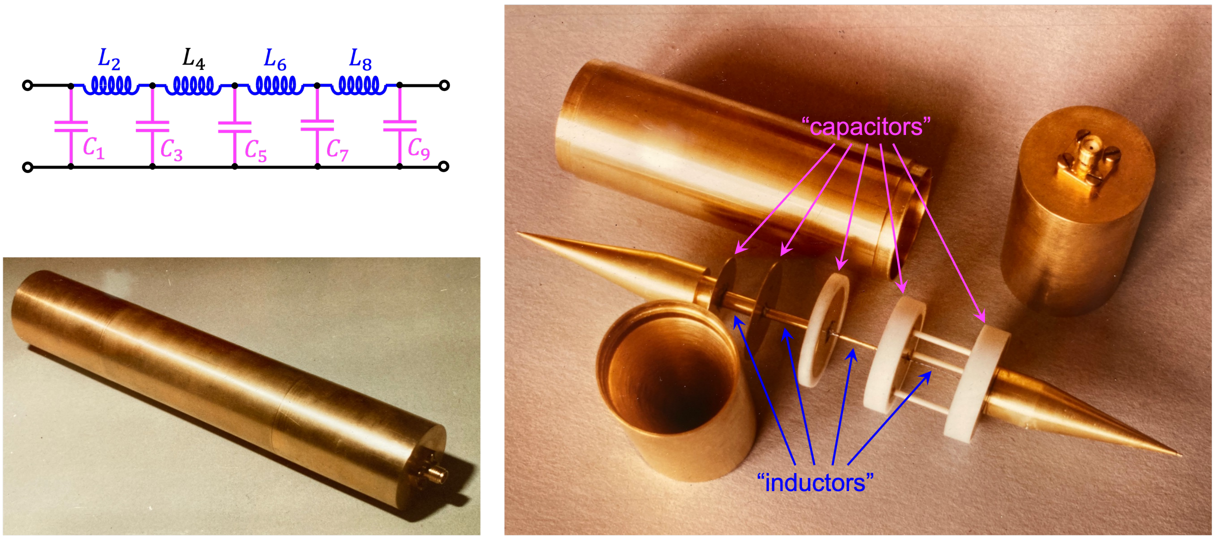


Fig. II.2.106: A 9-element “tubular” low-pass filter made of short TEM transmission-line sections.

parallel-plate *capacitor* with capacitance C, etc. At higher frequencies however, the behaviour deviates from the ideal lumped element due to uncontrolled parasitic effects, e.g. at some higher frequency an inductor becomes a resonator and at even higher frequency may become capacitive!

A short piece of a TEM transmission-line, satisfying $\theta < \pi/4$, can approximate a lumped inductive or capacitive elements with well-defined parasitic behaviour, as shown in Fig. II.2.105. As the equivalent circuits in Fig. II.2.105 suggests, the low-pass LC ladder filter network is a perfect application using short pieces of TEM transmission-lines as semi-lumped circuit elements to realise a RF low-pass. Figure II.2.106 shows as example a 9-element *Gaussian* low-pass filter with a $\sim$500 MHz 3 dB cut-off frequency. Here the capacitors are approximated by very short coaxial TL sections of low-Z_0 (large diameter of the coaxial inner conductor) and the inductors by somewhat longer coaxial TL sections of high-Z_0 (small diameter of the coaxial inner conductor), see also the right-hand side of Fig. II.2.106. For a successful implementation of this so-called “tubular” low-pass filter the capacitive effect of the discontinuities between the TL sections of low and high characteristic impedance have to be taken into account.

II.2.8.1.6 Power divider / combiner

Combining / adding two or more RF signals or dividing / splitting a single RF line in two, or more lines, of course without reflections, is a common RF engineering task. The power levels range from low-level RF (LLRF) signal applications, $\ll$0 dBm, to the high-power RF output from klystrons or inductive output tubes (IOT) to feed, e.g. multiple accelerating structures from a single power source. The resistive power divider, Fig. II.2.86, has limited use because of its high insertion loss, 6 dB, and no isolation between the ports, Eq. (II.2.66).

II.2.8.1.6.1 Wilkinson power divider / combiner

The *Wilkinson* power divider / combiner is popular to add / split lower level RF signals, and is typically implemented in a stripline or micro-strip planar transmission-line technology, see Fig. II.2.107. It utilises

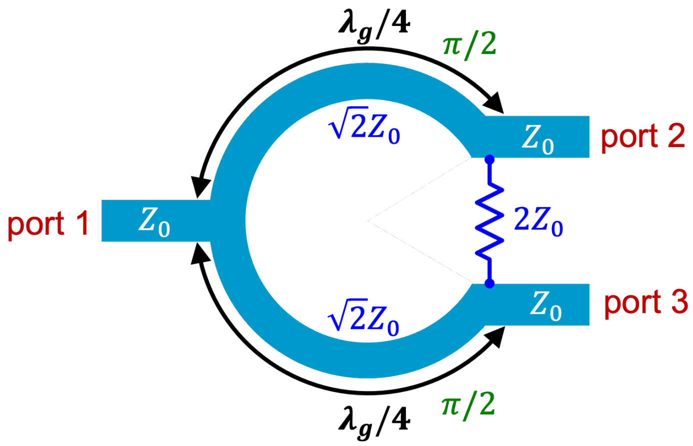


Fig. II.2.107: Stripline / micro-strip planar schematic of the *Wilkinson* power divider / combiner.

two $\lambda/4$ impedance transformers of $\sqrt{2}Z_0$, see also Eq. (II.2.86), to match the impedances of the two output ports 2 and 3, which are in parallel to the input port 1. A resistor of value $2Z_0$ between ports 2 and 3 is required for matching and isolation of the output ports, it will not dissipate any power in case of equal level signals at those output ports. At the operating frequency $f = 4v_g/\lambda_g$ the S-matrix of the ideal 3-port *Wilkinson* power divider / combiner is:

$$\mathbf{S} = \frac{-j}{\sqrt{2}} \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \tag{II.2.106}$$

The 3-port network is matched, reciprocal and symmetric, but is not lossless. Compared to the 3-port 6 dB resistive power divider, Eq. (II.2.66), the coupling loss of the *Wilkinson* power divider is:

$$|S_{21}| = |S_{31}| = |S_{12}| = |S_{13}| = \frac{1}{\sqrt{2}} \equiv 20 \log_{10} \frac{1}{\sqrt{2}} = -3\,\mathrm{dB}$$

This comes at the expense of a limited bandwidth spanning one octave, however, the bandwidth can be increased using multi-stage stepped impedance transformers. Other variants of the *Wilkinson* power divider / combiner include multiple ports and different characteristic impedances at the ports.

II.2.8.1.6.2 Waveguide T-junction power divider

Dividing / splitting the RF power is particular relevant for the distribution of the high-power RF in an accelerator RF system, which is typically supplied by waveguides for RF frequencies >1 GHz. A single RF power source, may feed two or more cavities or acceleration structures by splitting the RF power.

Figure II.2.108 (left and center) shows the two waveguide power divider options:

E-plane T The sidearm (port 3) couples in series to the parallel E-field, and the RF power signals out of

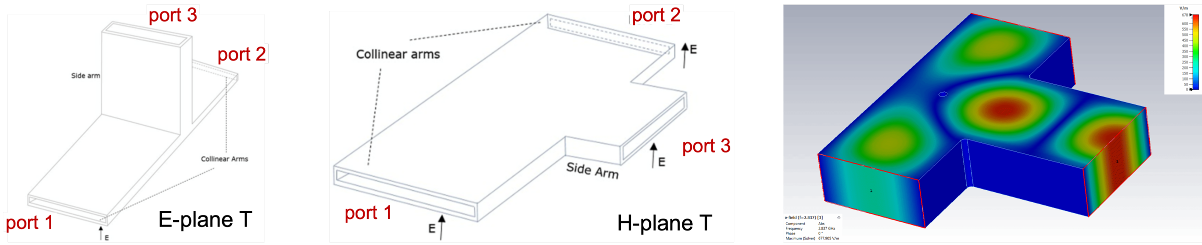


Fig. II.2.108: Waveguide T-junction, left: *E-plane T*, center: *H-plane T*, right: $|E|$-field of a S-Band waveguide H-plane T-junction.

the collinear output ports 1 and 2 have opposite phase. The S-matrix follows as:

$$\mathbf{S_E} = \frac{1}{2}\begin{bmatrix} 1 & 1 & \sqrt{2} \\ 1 & 1 & -\sqrt{2} \\ \sqrt{2} & -\sqrt{2} & 0 \end{bmatrix} \tag{II.2.107}$$

H-plane T The sidearm (port 3) couples in parallel to the parallel H-field, and the RF power signals out of the collinear output ports 1 and 2 are in phase. The S-matrix follows as:

$$\mathbf{S_H} = \frac{1}{2}\begin{bmatrix} 1 & -1 & \sqrt{2} \\ -1 & 1 & \sqrt{2} \\ \sqrt{2} & \sqrt{2} & 0 \end{bmatrix} \tag{II.2.108}$$

Similar the TEM transmission-line *Wilkinson* power divider, the waveguide T-junctions have $3\,\mathrm{dB}$ coupling losses. But, the coupled ports 1 and 2 are neither impedance matched, nor isolated, only the input port 3 is impedance matched.

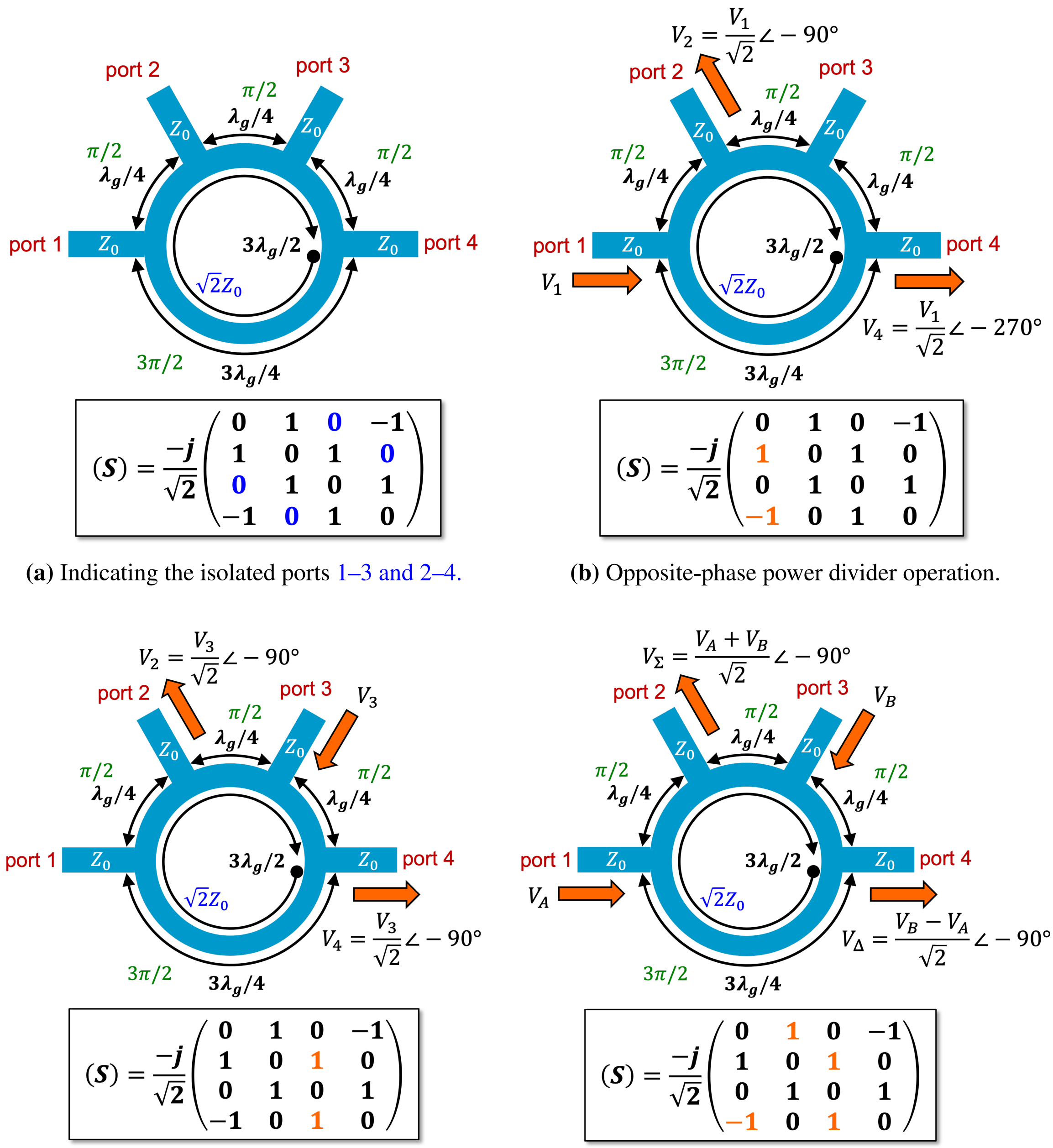
Attention: The nomenclature for the S-parameter ports of the *Wilkinson* power divider is different from that of the waveguide T-junctions:

Wilkinson: input: port 1, output: ports 2 and 3

T-junction: input: port 3, output: ports 1 and 2

II.2.8.1.7 RF coupling structures

RF coupling structures, or simple RF *couplers* typically are passive 4-port[4] RF transmission-line (TL) networks, with two or more TLs, or TL ports coupled to each other in various ways, e.g. by quarter-wave impedance transformers or electromagnetically. In the following some of the most popular examples of RF couplers, often used in accelerator RF systems are shown.



(a) Indicating the isolated ports 1–3 and 2–4.

(b) Opposite-phase power divider operation.

(c) In-phase power divider operation.

(d) Δ-Σ operation.

Fig. II.2.109: 4-port 180° “rat-race” hybrid coupler in planar TEM stripline or micro-strip technology.

II.2.8.1.7.1 180° hybrid coupler

Figure II.2.109 shows the schematic of the 180°, 3 dB hybrid coupler, also called “ring” or “rat-race” coupler, realised in a planar TEM stripline or micro-strip technology. It utilises $n\lambda/4$ impedance transformers arranged in a ring to couple the four ports together, which results in a passive, lossless, matched,

[4] Couplers may have more than four ports. 3-port couplers usually have the 4[th]-port internally terminated.

reciprocal and anti-symmetric 4-port network:

$$\mathbf{S} = \frac{-j}{\sqrt{2}} \begin{bmatrix} 0 & 1 & 0 & -1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ -1 & 0 & 1 & 0 \end{bmatrix} \tag{II.2.109}$$

Ports 1–3 and 2–4 are isolated to each other, as indicated in the S-matrix of Fig. II.2.109a. Among a variety of applications, Fig. II.2.109b indicates the operation as 3 dB power divider with port 1 as input and ports 2 and 4 as outputs. In this configuration the output signals have opposite, 180° phase to each other. Instead, using port 3 as input divides the signal between ports 2 and 4, but now those output signals are in-phase, 0° to each other, see Fig. II.2.109c.

Feeding two RF signals, A and B, of same frequency and different amplitude simultaneously into ports 1 and 3 enables the 180° hybrid coupler to operate as Δ-Σ hybrid; at port 2 the sum and at port 4 the difference of the amplitudes of the applied input signals will appear, see the details in Fig. II.2.109d.

The 180° hybrid coupler is a narrow-band device, however, using multi-stage impedance transformers the bandwidth can be increased.

II.2.8.1.7.2 Waveguide "magic-T"

The "magic-T" is a 180°, 3 dB coupler in waveguide technology, see the photo in Fig. II.2.110. As Fig. II.2.110 suggests, it is a combination of a E-plane T and a H-plane T, presented in Paragraph II.2.8.1.6.2, in a single waveguide coupling element. As a result, the 4-port S-matrix of the ideal waveguide magic-T

$$\mathbf{S} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{bmatrix} \tag{II.2.110}$$

Fig. II.2.110: Waveguide "magic-T".

is almost identical to that of the ideal TEM 180° hyrbid ring coupler, Eq. (II.2.109).

> The port labels of the waveguide magic-T differ from the TEM planar hybrid ring coupler!
>
> **Hybrid ring coupler:** Inputs: ports 1 and 3, outputs: ports 2 and 4
>
> **Waveguide magic-T:** Inputs: ports 1 and 2, outputs: ports 3 and 4
>
> The hybrid ring coupler adds a 90° phase shift to all S-parameters, the magic-T does not.

Similar to the TEM planar hybrid ring coupler of the previous Paragraph II.2.8.1.7.1, the magic-T waveguide coupler offers several functions depending on the supplied RF signals, Fig. II.2.111. Also here we have isolated ports, the collinear ports 1–2 are isolated to each other, as well as the ports 3–4, see Fig. II.2.111a. Figure II.2.111b shows the operation as in-phase power divider and Fig. II.2.111c as opposite-phase power divider. Applying in-phase RF signals of same frequency and different amplitude simultaneously at ports 1 and 2 results in the Δ-Σ arithmetic's, with the Σ-signal at port 3: $E_\Sigma \propto E_1 + E_2$, and the Δ-signal at port 4: $E_\Delta \propto E_1 - E_2$.

II.2.8.1.7.3 90° quadrature hybrid coupler

Similar to the other couplers described in this section, also the 90° *quadrature* hybrid coupler has a coupling loss of 3 dB. The 4-port S-matrix of the ideal quadrature hybrid at the operation frequency:

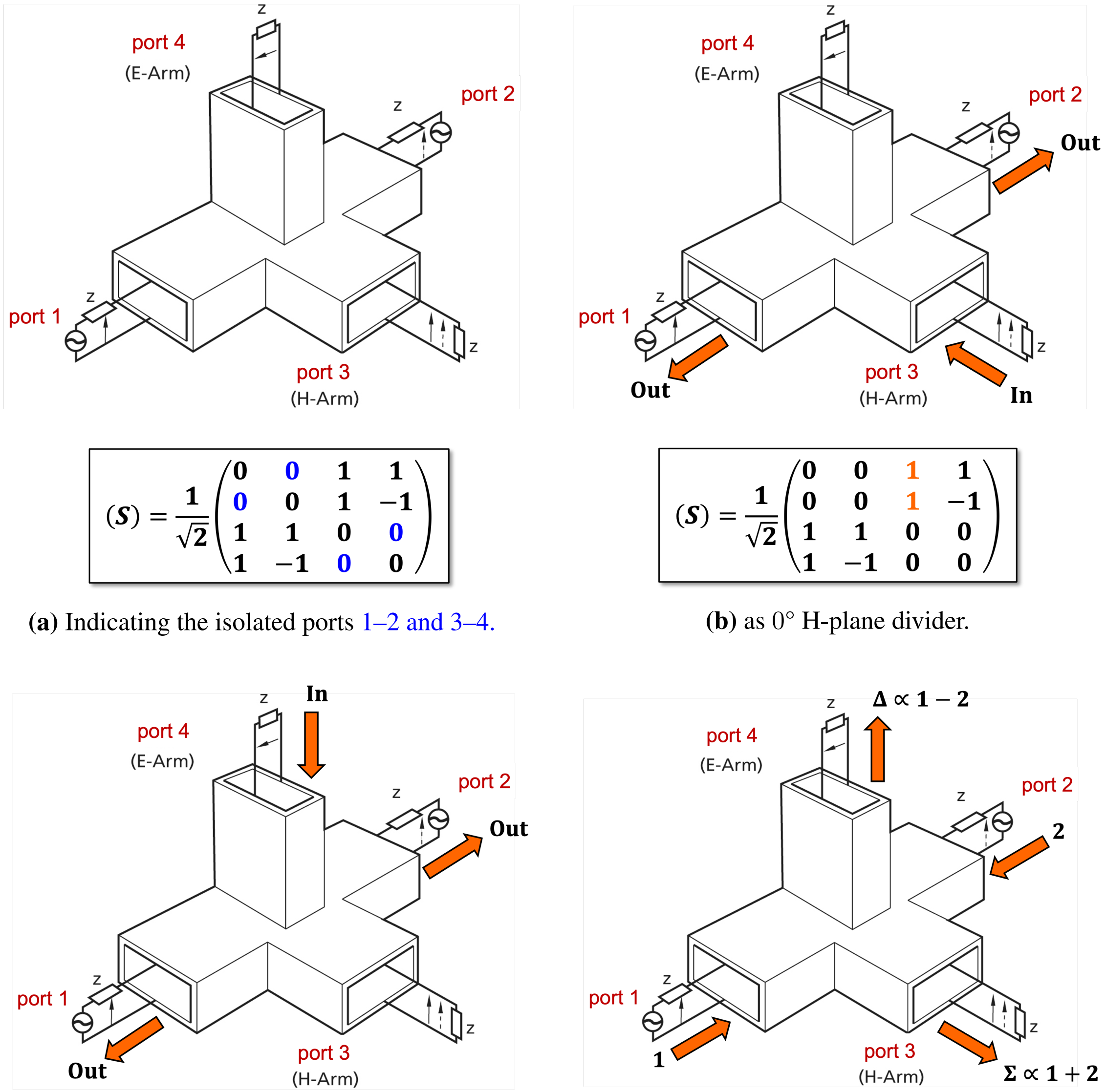
$$\mathbf{S} = \frac{-1}{\sqrt{2}} \begin{bmatrix} 0 & j & 1 & 0 \\ j & 0 & 0 & 1 \\ 1 & 0 & 0 & j \\ 0 & 1 & j & 0 \end{bmatrix} \tag{II.2.111}$$

shows, this components is passive, lossless, has matched ports and is reciprocal, as well as symmetric. Figure II.2.112 shows layout and operation of the quadrature hybrid coupler, here illustrated in TEM planar stipline or micro-strip technology. The four ports are coupled by two-plus-two quarter-wave transmission-lines, one pair acts as an impedance transformer, the other pair just as a 90° phase shifter. Ports 1–4 and 2–3 are isolated from each other, as shown in Fig. II.2.112a.

Figure II.2.112b shows the operation of the quadrature hybrid as power divider, please note, the phase argument of the signals at the output ports 2 and 3 always have 90° phase difference! This behaviour gives the name: *quadrature* hybrid, as the output signal are in quadrature to each other. The quadrature hybrid is very popular in I-Q (in-phase, quadrature-phase) RF applications, which is de-facto the standard in most RF signal processing applications, e.g. in accelerator LLRF and beam instrumentation systems, as well as more general in the telecommunication industry.

Figure II.2.112 shows the operation of the quadrature hybrid if RF signals of same frequency and different amplitudes are simultaneously supplied at the isolated ports 1 and 4. This results in signals at the outputs 2 and 3 which have the same amplitude, but their phase difference

$$\tan^{-1}\left(\frac{V_A}{V_B}\right) - \tan^{-1}\left(\frac{V_B}{V_A}\right) = 2\tan^{-1}\left(\frac{V_A}{V_B}\right) - \frac{\pi}{2}$$



$$(S) = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix}$$

(a) Indicating the isolated ports 1–2 and 3–4.

$$(S) = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix}$$

(b) as 0° H-plane divider.

$$(S) = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix}$$

(c) as 180° E-plane divider.

$$(S) = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix}$$

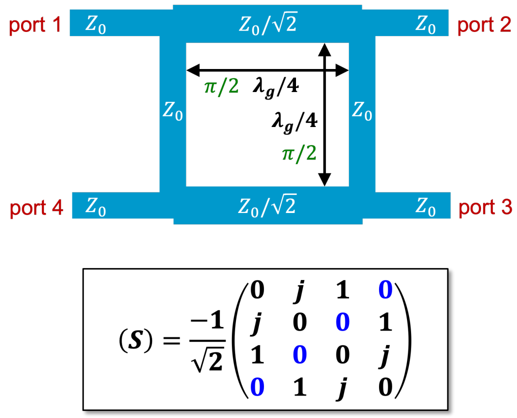
(d) as Δ-Σ coupler.

Fig. II.2.111: Operation of the "magic-T".

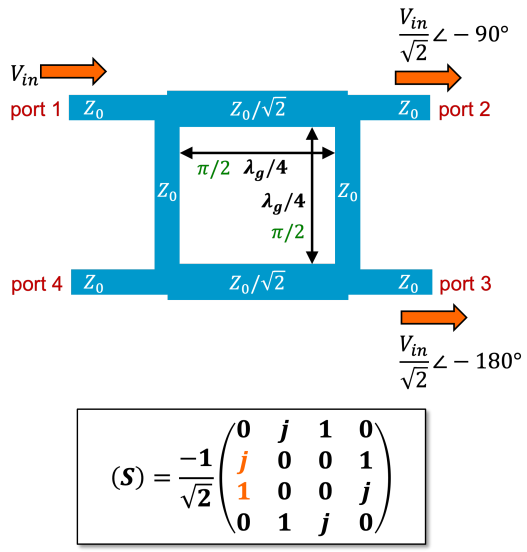
is a function of the amplitude ratio V_A/V_B of the input signals. Thus, the quadrature hybrid operates as amplitude-ratio to phase difference converter. Similar to the Δ-Σ operation of the 180° hybrid, see Paragraph II.2.8.1.7.1, this RF arithmetic feature is used in some beam instrumentation applications, e.g. for the processing of beam position monitor signals.

II.2.8.1.7.4 Directional coupler

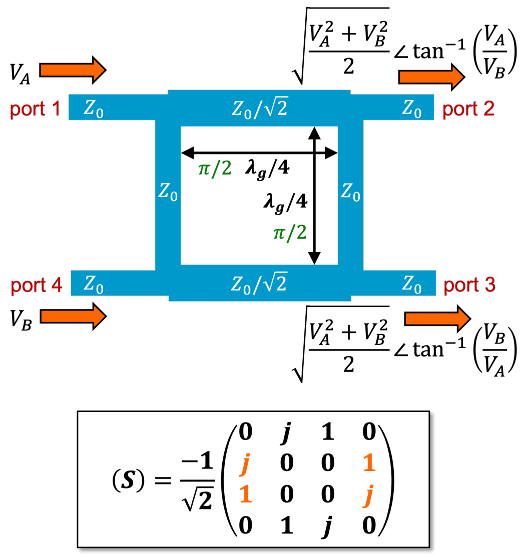
The S-matrix and some key characteristics of an ideal *directional coupler* were already mentioned in Section II.2.7.7.4. This 4-port transmission-line component is based on two transmission-lines which are electromagnetically coupled over some distance, which is the physical coupling length $\ell = \lambda_g/4$. The most simple coupling configuration are two symmetrically arranged TEM transmission-line in planar



(a) Indicating the isolated ports 1–4 and 2–3.



(b) I-Q signal divider operation.



(c) Amplitude-ratio to phase-difference converter.

Fig. II.2.112: 4-port 90° quadrature hybrid coupler in planar TEM stripline or micro-strip technology.

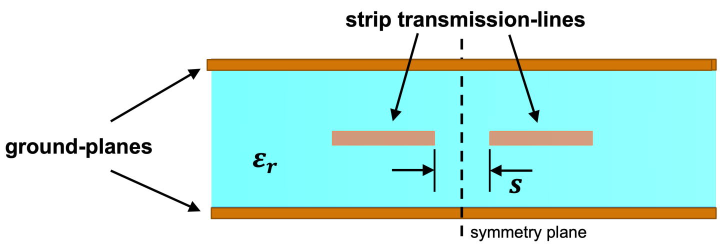


Fig. II.2.113: Cross-section geometry of a symmetric pair of coupled strip transmission-lines.

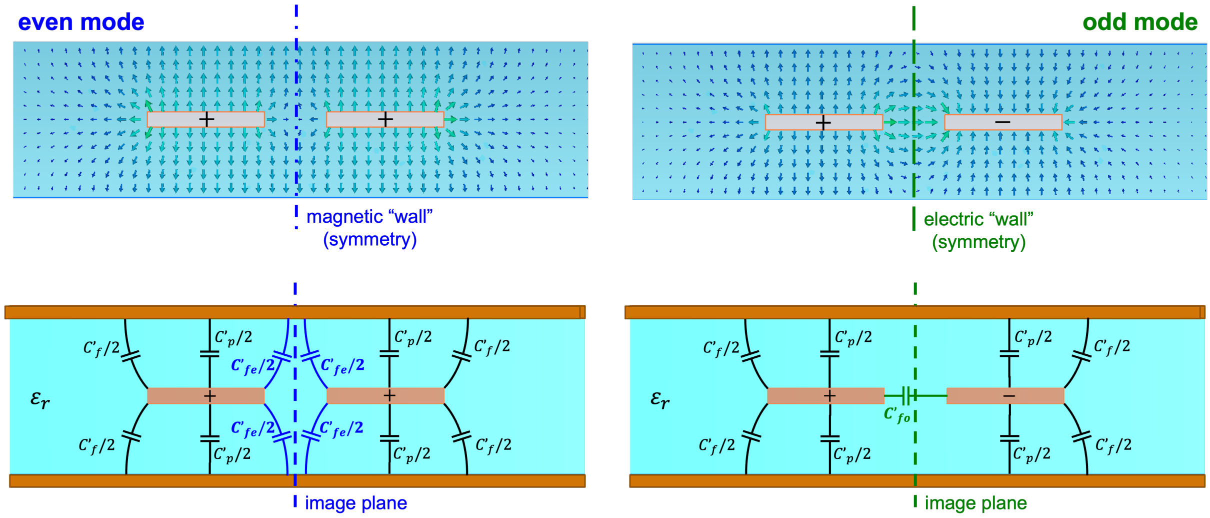


Fig. II.2.114: *Even* (left) and *odd* (right) modes of coupled striplines;
Upper: E-field, lower: equivalent schematics.

stripline or micro-strip technology, as shown in Fig. II.2.89b.

For the understanding of the coupling mechanism an electrostatic analysis of the 2-dimensional (2D) cross-section geometry is sufficient. Figure II.2.113 illustrates the cross-section of two symmetrically arranged strip transmission-lines (striplines), i.e. flat conductors between ground-planes embedded in a dielectric substrate ε_r. The two conductors are spaced by some distance, space s, from each other. The upper part of Fig. II.2.114 shows the E-field as result of an electrostatic analysis of this conductor configuration. On the left side of Fig. II.2.114 both stripline conductors are excited with the same positive (+) potential wrt. ground, called *even-mode* excitation. The right side instead shows the E-field in *odd-mode* excitation, the left conductor has a + potential, while the right conductor has a − potential (of same value) wrt. ground.

The lower part of Fig. II.2.114 shows the equivalent circuit schematic of the coupled striplines for the two cases, with C'_p covering the parallel plate capacitance and C'_f the fringe field capacitance (both per unit length). In the even-mode case C'_f between the conductors becomes C'_{fe}, while in the odd-mode case the fringe capacitance between the conductors C'_{fo} dominates. This leads to:

$$\begin{array}{ll} \text{even-mode:} & \text{odd-mode:} \\ C'_{0e} = C'_p + C'_f + C'_{fe} & C'_{0eo} = C'_p + C'_f + C'_{fo} \\ Z_{0e} = \dfrac{\sqrt{\varepsilon_r}}{cC'_{0e}} & Z_{0o} = \dfrac{\sqrt{\varepsilon_r}}{cC'_{0o}} \end{array} \tag{II.2.112}$$

with Z_{0e} being the even-mode characteristic impedance and Z_{0o} the odd-mode characteristic impedance of the coupled TL configuration. The coupling coefficient or coupling factor k is defined from those characteristic impedance's as:

$$k = \frac{Z_{0e}/Z_{0o} - 1}{Z_{0e}/Z_{0o} + 1} \tag{II.2.113}$$

Example II.2.8.2: Differential signal transmission with coupled transmission-lines

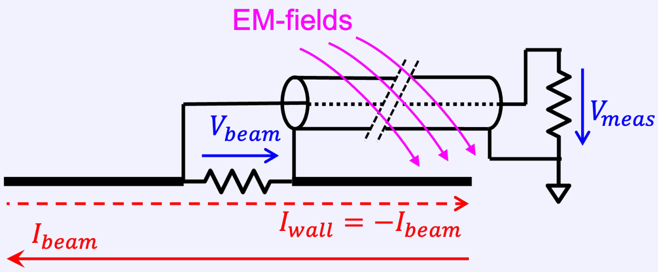
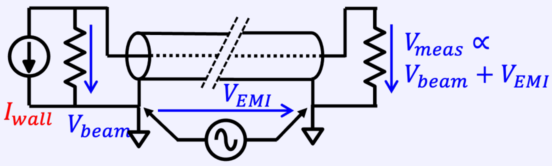


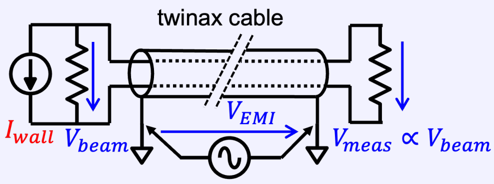
Fig. II.2.115: Beam current measurement in an electromagnetically "polluted" environment.

Coupled transmission-lines have many applications beyond directional couplers, one of them being *differential* signal transmission. Examples range from audio engineering to gigabit serial data signal transmission in computational hardware.

Figure II.2.115 shows an example that is quite typical in particle accelerators, the measurement of the beam intensity, i.e. the usually rather low beam current, in presence of an environment that is "polluted" by external electromagnetic fields of high intensities (EMI: electromagnetic interference), which can originate from pulsed sources like kicker or septa power supplies, or other RF or pulsed high power installations with limited shielding. A long, *single-ended* coaxial cable is used to transfer the signal from the accelerator tunnel to the read-out system. But due to the non-zero impedance (resistance) of the outer cable shield conductor an unwanted error signal V_{EMI} adds to the beam pickup signal V_{beam} that is to be measured, see Fig. II.2.116a. A differential signal transmission, e.g. using a so-called "twinax" or similar 2-conductor plus shield cable can substantially reduce this so-called *common-mode* interference, see Fig. II.2.116b.



(a) Single-ended coaxial cable.



(b) Differential "twinax" shielded cable.

Fig. II.2.116: Signal transmission in presence of EMI.

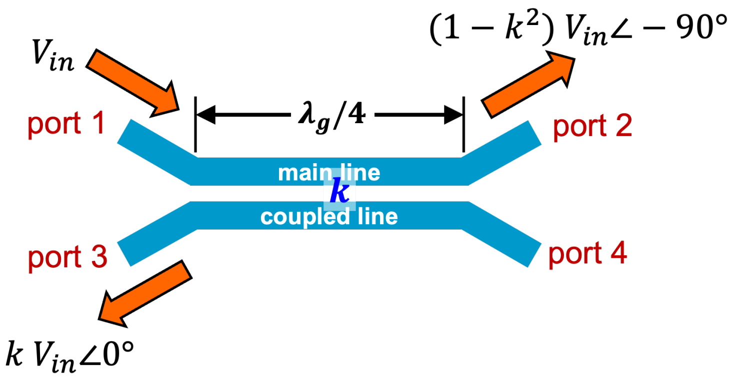


Fig. II.2.117: Single-section directional coupler operating at the center frequency.

Please note:
The differential-mode termination impedance is:

$$Z_{diff} = 2Z_{0o}$$

In the same way a common-mode termination impedance would calculate:

$$Z_{common} = \frac{Z_{0e}}{2}$$

but this type of signal transmission is rare.

The S-matrix for an ideal TEM directional coupler, Eq. (II.2.73), was given for the operation at the center frequency f_c, which follows from the physical length $\ell = \lambda_g/4$ as:

$$\theta = \beta\ell = \frac{\pi}{2} \quad \Rightarrow \quad f_c = \frac{c}{4\ell\sqrt{\varepsilon_r}}$$

Figure II.2.117 illustrates the operation using the *main line* port 1 as input, port 2 as main line output, and the *coupled line* port 3 as coupled output. Ports 1–4 and 2–3 are isolated from each other.

As the electrical length is a function of the frequency

$$\theta(f) = \frac{2\pi f \ell\sqrt{\varepsilon_r}}{c}$$

see also Eq. (II.2.91), so are the S-parameters of the ideal TEM directional coupler:

$$\begin{aligned} S_{ii} &= 0 \quad \text{(all ports are matched)} \\ S_{31}(\theta) = S_{13}(\theta) = S_{42}(\theta) = S_{24}(\theta) &= \frac{jk\sin\theta}{\sqrt{1-k^2}\cos\theta + j\sin\theta} \\ S_{21}(\theta) = S_{12}(\theta) = S_{43}(\theta) = S_{34}(\theta) &= \frac{\sqrt{1-k^2}}{\sqrt{1-k^2}\cos\theta + j\sin\theta} \\ S_{41} = S_{14} = S_{32} = S_{23} &= 0 \quad \text{(isolated ports)} \end{aligned} \tag{II.2.114}$$

Figure II.2.118 shows $|S_{31}|/k = f(\theta)$, each "lobe" has an octave bandwidth, which can be increased by using multiple coupling sections of same length and slightly varying coupling coefficient as illustrated in Fig. II.2.119 for $n = 3$ coupling sections.

The directional coupler is a very popular RF component, many variants exist beside the discussed coupler based on coupled TEM transmission-lines, e.g. waveguide directional couplers, broadband, multi-section directional couplers, dual directional couplers with two coupled lines, directional couplers with transmission-lines of different characteristic impedances, etc.

TEM directional couplers are *backward* couplers, waveguide directional couplers are based on slot or hole coupling, and exists as *forward* and backward couplers. Figure II.2.120a illustrates the forward slot-coupling mechanism, here the backward port is internally terminated. Figure II.2.120b shows a photo of a commercial waveguide directional coupler, here both ports of the coupled waveguide are accessible.

Directional couplers are also used in particle accelerators as beam diagnostics or manipulation devices, Fig. II.2.121 shows some examples. Usually the beam itself acts as main line, and two or more coupling lines are arranged symmetrically for beam pickup or manipulation.

Example II.2.8.3: Monitoring of forward and reflected power on an accelerating structure

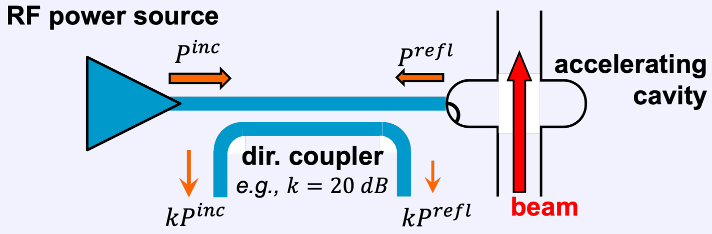


Fig. II.2.122: Monitoring of forward and reflected power.

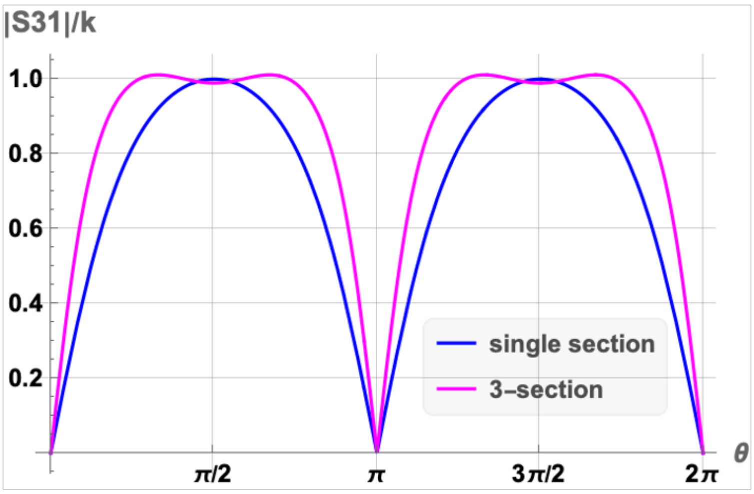


Fig. II.2.118: $|S_{31}|/k$ of the ideal TEM directional coupler as function of the electrical length θ.

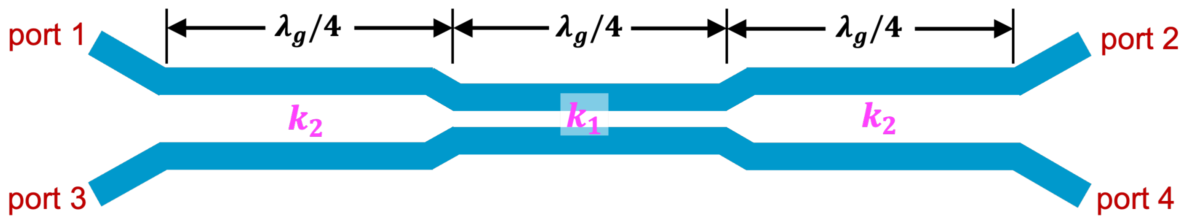
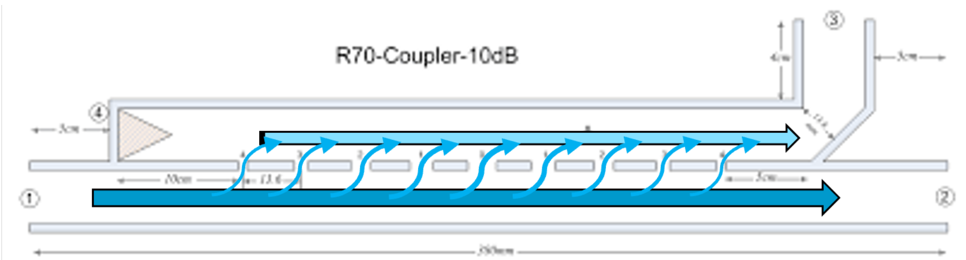


Fig. II.2.119: TEM directional coupler with $n = 3$ coupling sections.

As already explained in section Section II.2.7.7.4, the directional coupler has the unique feature to distinguish between forward and backward travelling waves on the main line via the ports of the coupled line. This leads to an important application in accelerator RF engineering, illustrated in Fig. II.2.122:
RF power is supplied from a high power source, e.g. a klystron, IOT, triode or semiconductor power amplifier to an accelerating structure (resonant cavity or travelling wave structure) via

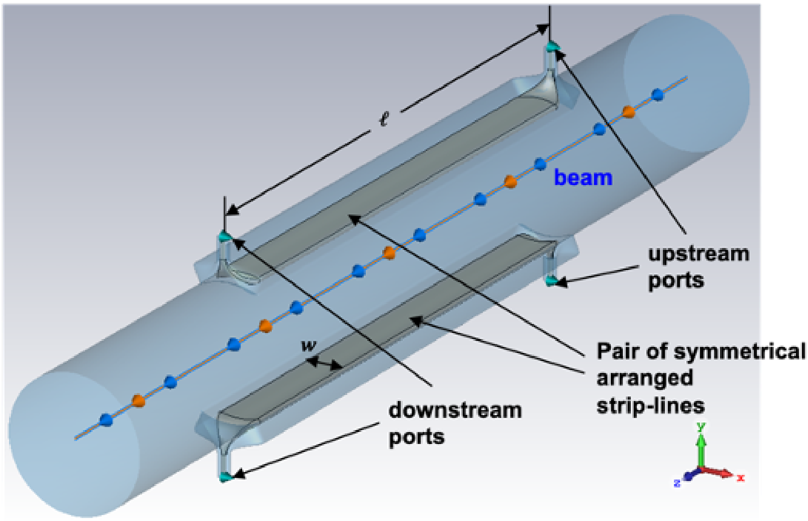


(a) WG forward slot-coupling.

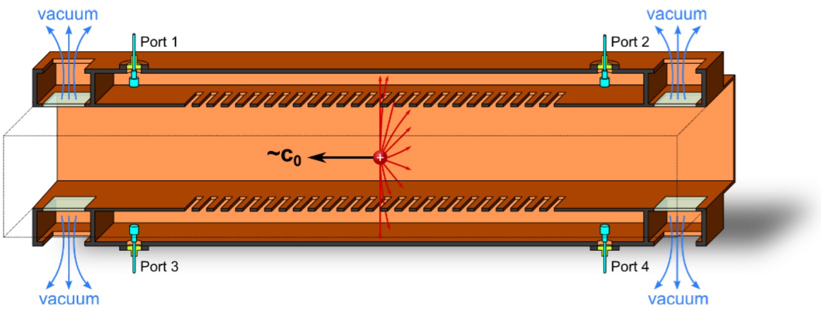
(b) Commercial WG directional coupler.

Fig. II.2.120: Waveguide directional coupler.

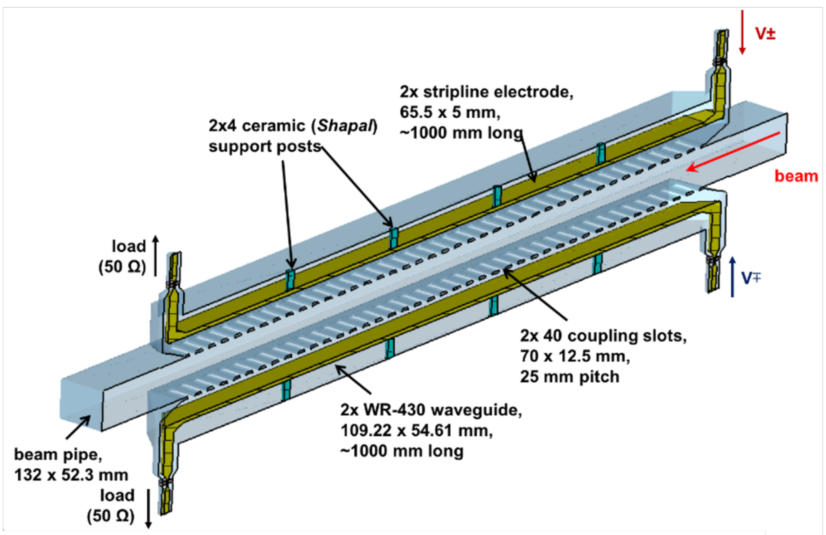
a coaxial or waveguide transmission-line. At times the beam is present, most of the supplied



(a) Stripline beam position monitor (BPM).



(b) LHC *Schottky* beam pickup (CERN).



(c) SPS intra-bunch broadband stripline kicker (CERN).

Fig. II.2.121: Directional couplers as beam pickups and kickers.

RF power will be absorbed by the beam, only little power is reflected depending on the precise tuning of the system to e.g. prevent the *Robinson* instability. However, at times the beam is not present, for injection / extraction kicker gaps, or beam formats required by the users or the stable operation of the machine, all RF power is reflected! Therefore, the continuous monitoring of forward (incident) and reflected RF power on the main RF supply line is mandatory, and is often included in a feedback system. As of the high power on the RF supply line, a low coupling coefficient of the directional coupler is in most cases sufficient, typical are values $k \leq -20\,\mathrm{dB}$[a], such that the insertion loss on the main line is negligible, see also Table II.2.11.

Table II.2.11: Directional coupler main line insertion loss vs. coupling factor, as of Eq. (II.2.73).

coupling factor $[dB]$	**main line insertion loss** $[dB]$
3	3.00
6	1.25
10	0.46
20	0.044
30	0.004

For a coupling factor of $3\,\mathrm{dB}$, precisely: $-3\,\mathrm{dB} \equiv 1/\sqrt{2}$, also the main line loss is $3\,\mathrm{dB}$, see Table II.2.11. In this case the S-matrix of the directional coupler, Eq. (II.2.73) is similar to that of the quadrature hybrid, Eq. (II.2.111), and thus allows similar functionality.
Adding a $\lambda/4$ transmission-line to one of the ports of a $3\,\mathrm{dB}$ directional coupler leads to a S-matrix similar to that of the 180° hybrid, Eq. (II.2.109).

[a] Usually the sign is omitted when describing the coupling in dB values, and is defined as *coupling loss*

Most of the 4-port couplers discussed in this section are symmetric. This allows to exchange the ports used for input and output signals in various ways without compromising the function.

II.2.8.2 The circulator

The circulator, and the isolator as well, have already been introduced in Section II.2.7.7.3, for the ideal circulator as example of a 3-port S-matrix, Eq. (II.2.67). As mentioned, the circulator is passive and matched, but non-reciprocal as it "circulates" the incident RF power signal from one port to the next port, but not vice versa.

Figure II.2.123 shows the practical realisation of the circulator, Fig. II.2.123a the waveguide version, often used in high power applications, and Fig. II.2.123b the stripline version which often is exploited with coaxial connectors and used for low power applications. As the operational principle of the circulator is based on the anisotropic and non-reciprocal properties of magnetised microwave ferrite materials, sometimes between a pair of biasing permanent magnets, it has a given operating frequency range based on geometric and technology details.

Please note, circulators don't exist for DC or very low frequencies. Typical operating frequencies start at $\gtrsim 100\,\mathrm{MHz}$. Four-port versions of the circulator also exist, some realised as high isolation, dual circulator based on two 3-port isolators.

Example II.2.8.4: Protection of the power output stage from reflected RF power

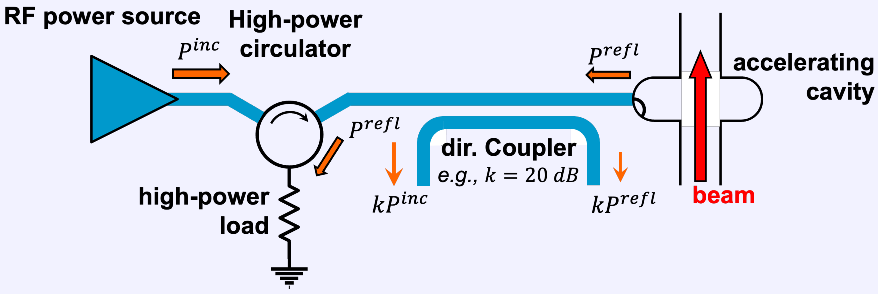
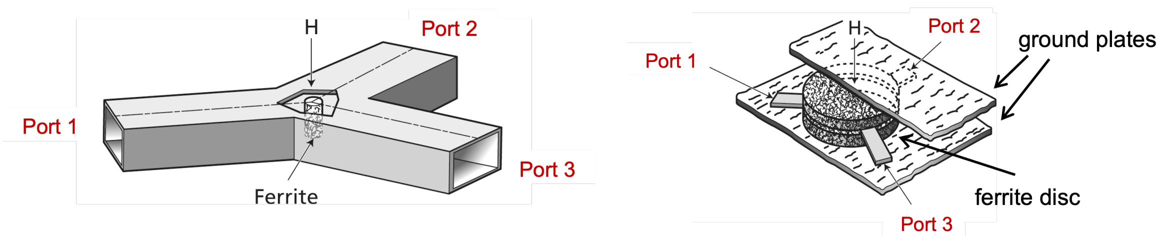


Fig. II.2.124: RF power output stage protection.

There are several popular applications for the circulator, e.g. as *duplexer* in radar systems, however, in accelerator RF the circulator is mostly utilised as *isolator*, i.e. one port of the circulator is terminated with a dummy load.

Figure II.2.124 shows a typical application to protect the output stage of the high power RF amplifier from the power that might be reflected from the accelerating cavity in absence of beam loading. It is basically an extension of Fig. II.2.122 by a high power isolator, which in practice is made of separate components, a high power circulator (often in waveguide technology) and a high power dummy load, usually water cooled. In this way no, or only very little RF power can reach the output of the RF power amplifier, it's sensitive output stage is protected against reflected RF power!



(a) Waveguide circulator.

(b) Stripline circulator.

Fig. II.2.123: The circulator.

To reduce the CO_2 footprint of accelerators, R&D is underway to reuse the energy of the otherwise unused reflected RF power.

II.2.8.3 RF filters

There are many types of filters: coffee filters, oil filters, air filters,. . . , and then there are electrical / electronic filter networks. Here we introduce the *RF filter* as a 2-port, passive electrical network, see Fig. II.2.125, which in most cases is symmetric and reciprocal, often lossless, and usually is not perfectly matched at the ports. A passive electrical or RF filter can be as simple as a RC two-component circuit shown in Fig. II.2.72, still, the theory behind it is rather complex, among the many articles and books here just two references for further studies [18, 19].

Electrical and RF filters are usually defined by their behaviour in the frequency domain and categorised in *low-pass*, *high-pass*, *band-pass* and *band-stop* filters, see the $|S_{21}|(\omega/\omega_c)$ frequency domain responses in Fig. II.2.126. There are also other types of filters, e.g. *all-pass* filters used as time-delay network, *diplexers* or *multiplexers* with three or more ports, or filters which are defined in the time-domain to approximate a specific TD response waveform.

As the RF filters are usually characterised in the frequency-domain, by their $S_{21}(\omega)$ transfer function, this "automatically" results in a specific time-domain behaviour, e.g. the response to a δ-*Dirac* or step function. Please remember the link between time- and frequency domain behaviour of linear systems.

In Fig. II.2.126 ω_c specifies the so-called *cut-off frequency*, which divides *passband* and *stopband* of the filter, i.e. the frequency range in which the filter passes or stops the RF signal between input and output port. Usually $\omega_c \equiv \omega_{3dB}$, the 3 dB cut-off frequency, is defined as

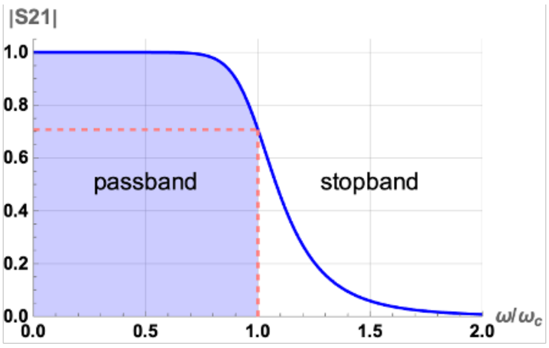
Definition II.2.8.1: 3 dB bandwidth

$$|S_{21}|(\omega = \omega_{3dB}) = \frac{1}{\sqrt{2}} \qquad \text{(voltage definition)} \tag{II.2.115}$$

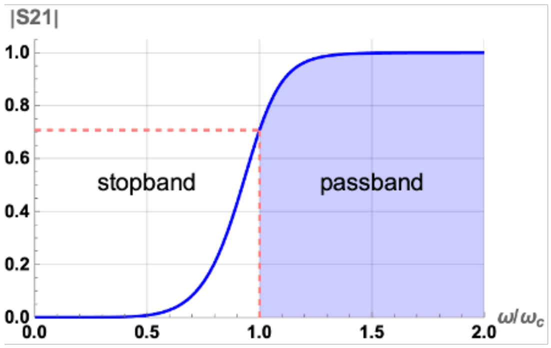
$$|S_{21}|^2(\omega = \omega_{3dB}) = \frac{1}{2} \qquad \text{(power definition)} \tag{II.2.116}$$

All four filter categories are based on a so-called *low-pass prototype* filter, which is a well defined

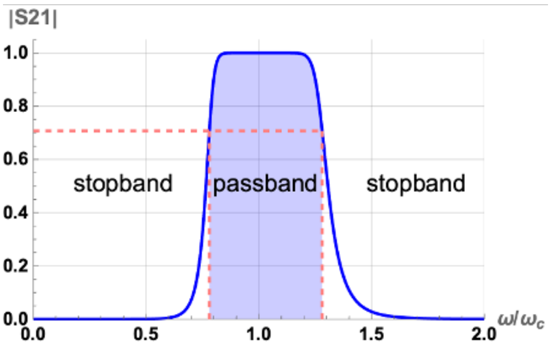
Fig. II.2.125: A 2-port, passive RF filter network.



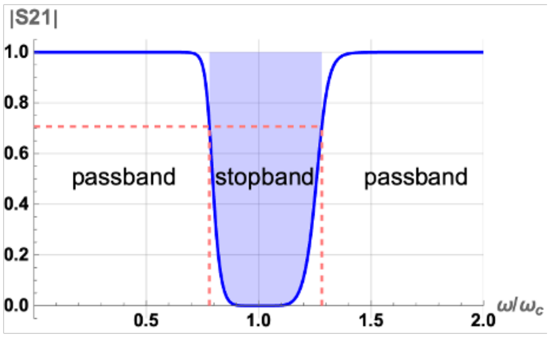
(a) Low-pass filter (LPF).



(b) High-pass filter (HPF).



(c) Band-pass filter (BPF).



(d) Band-stop filter (BSF).

Fig. II.2.126: Filter categories.

transfer function $S_{21}(\omega)$ for $\omega_c = 1$ and $R_L = 1\,\Omega$ that also specifies the characteristics of the transition region between passband and stopband. Popular LP-prototype designs are e.g., *Chebyshev*, *Butterworth*, *Bessel-Thomson*, *Gaussian*, etc., approximating a specific, analytical defined goal.

The circuit operating as *RF filter* between ports 1 and 2 in Fig. II.2.125 is in most cases realised of linear, passive components, e.g. lumped elements (capacitors, inductors, resistors) and / or distributed elements (transmission-lines, resonators), thus, RF filters are passive. If the filter circuit is made only of reactive lumped components, ideal capacitors and inductors, we have a lossless 2-port network, Eq. (II.2.55) applies, and all power from the RF source is then absorbed in either the source resistor R_S and / or the load resistor R_L, or to be more precise, in a ratio between them defined by $S_{21}(\omega)$. Evidently, the ports of a lossless filter are not matched. However, there exist a lossy *absorptive* filter design with matched ports, and also a combination of e.g. a low-pass and a high-pass filter into a so-called 3-port *diplexer* allows the matching of its input port.

Exercise II.2.8.1: RF filter analysis

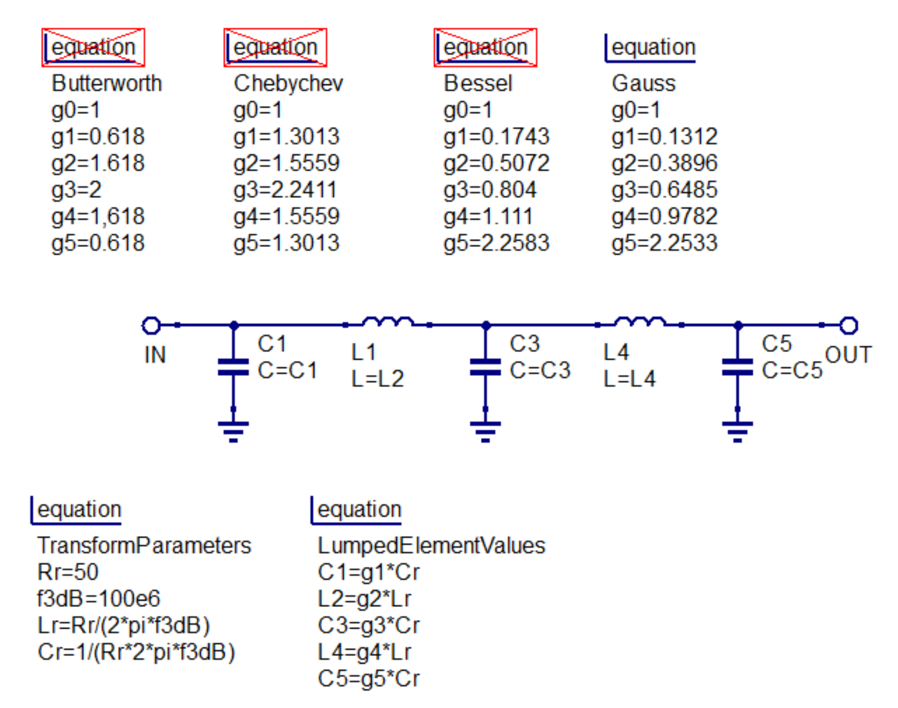
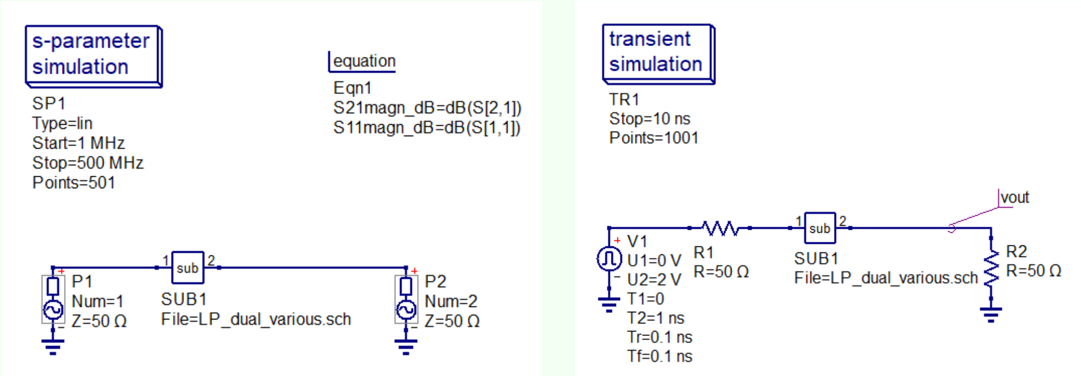


Fig. II.2.127: Low-pass filter sub-circuit used in *Qucs*.

While the synthesis of a new type of RF filter can be complicated, the analysis of known filters with help of a numerical circuit analysis software is rather trivial and educational! For this exercise we use the *QucsStudio* freeware [20], which has evolved from the original Qucs circuit simulator.

Both, *QucsStudio* and the original *Qucs* are sufficient for this exercise. *Qucs* is not anymore maintained, but runs on a variety of computer platforms, while *QusStudio* is maintained but requires *MS Windows*.



(a) 2-port S-parameter analysis. **(b)** Impulse response analysis.

Fig. II.2.128: Filter analysis with *Qucs* in frequency- and time-domain.

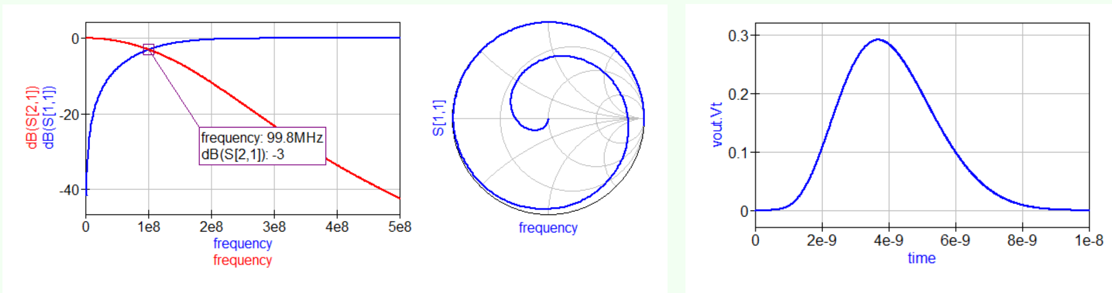
1. Download QucsStudio (or Qucs) and start the software with `start.bat`.

2. Familiarise yourself with the tool, you will find many examples and tutorials on the internet. Please note, your *Qucs* files will be stored in the `.qucs` directory under your username directory. Also please note, the file formats of *QucsStudio* and *Qucs* are different! (incompatible!)

3. Create a `New` project and fill the schematic window (`untitled`) with a `s-parameter simulation` similar to Fig. II.2.128a. Replace the `SUB1` sub-circuit block by a simple RC or RL low-pass circuit from the `Components` $\rightarrow$ `lumped components` palette.

4. `Simulate` the `s-parameter simulation` with reasonable values of the circuit components and display the results, e.g. with a `Cartesian` plot of `dB(S[1,1])`.

5. Once you are more familiar with the *Qucs* software, arrange a sub-circuit block with the 5$^{\text{th}}$-order low-pass ladder network as shown in Fig. II.2.127 and analyse the filter circuit in the frequency-domain (`s-parameter simulation`, Fig. II.2.128a) and in the time-domain (`transient simulation`, Fig. II.2.128b).



(a) Frequency-domain. **(b)** Time-domain.

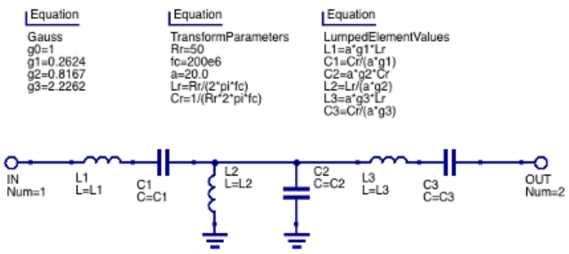
Fig. II.2.129: Filter analysis results.

6. Fig. II.2.129 shows the results for a *Gaussian* low-pass filter approximation, as it was selected in the schematic, Fig. II.2.127.

If you dig deeper in the *QucsStudio* simulator, you will find under `Tools` → `Filter synthesis` more sophisticated filter tools, however, advanced tools for RF filters, like optimisation based on distributed elements, filter matching, etc., are only available on commercial software.

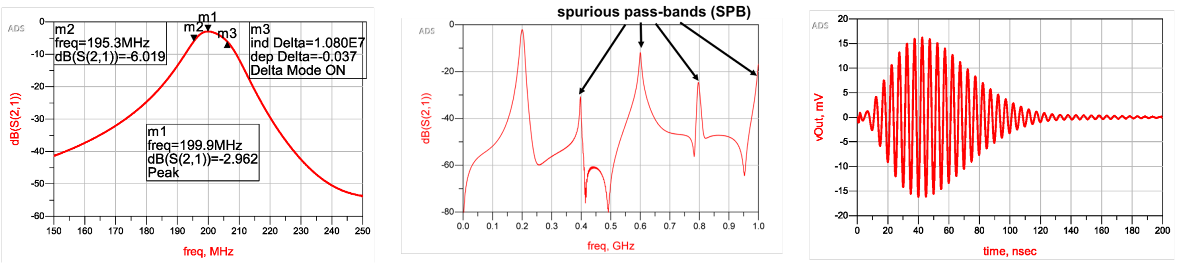
Industry offers a wide selection of RF filters, with a suitable solution for most applications. However, our field of accelerator RF engineering sometimes may require a more exotic filter characteristic and therefore an in-house developed RF filter. This requires filter and network theory, with the starting point at the requirements, e.g. cut-off or center frequency, bandwidth, filter characteristic, number of sections, etc. Based on an existing or a new established lumped element LP prototype, a LPproto → LPF, HPF, BPF or BSF transformation has to be performed, followed by the often non-trivial implementation of real-world components.

Figure II.2.130 shows the design of a RF band-pass filter for beam instrumentation purposes, operating at 200 MHz with a bandwidth of 10 MHz. Figure II.2.130a shows the equivalent circuit of the 3rd-order *Gaussian* filter approximation and Fig. II.2.130b the practical realisation as so-called “hair-pin” stripline filter based on coupled $\lambda/4$ resonators. The $|S_{21}|(f)$ measurement, Fig. II.2.130c left, shows a satisfactory behaviour around the center frequency of 200 MHz, however, Fig. II.2.130c center illustrates the fundamental problem of RF filters based on transmission-lines or resonators:
The filter exhibits unwanted higher-order, so-called “spurious” pass-bands.
Still, the time-domain performance, Fig. II.2.130c right, relevant for this application, fulfilled the re-



(a) 3rd-order lumped element equivalent network.

(b) Coupled resonator stripline realisation.



(c) Measured filter performance.

Fig. II.2.130: A 200 MHz *Gaussian* RF band-pass filter with 10 MHz bandwidth.

quirements. Of course, in practice a low-pass filter was added to suppress the spurious pass-bands.

II.2.8.4 RF amplifiers

The ideal *RF amplifier*, see also Fig. II.2.131, was briefly mentioned in Section II.2.7.7.2, with Eq. (II.2.64) being the S-matrix. In practice, the *voltage gain* of an amplifier is given as a scalar value, the phase is of no relevance:

$$|G_V| = |S_{21}| = \frac{|V_{\text{out}}|}{|V_{\text{in}}|} = 10^{\frac{g}{20}} \tag{II.2.117}$$

and in most cases the gain is expressed in the logarithmic dB scale

$$g = 20 \log_{10} \frac{|V_{\text{out}}|}{|V_{\text{in}}|} = 10 \log_{10} \frac{P_{\text{out}}}{P_{\text{in}}} \tag{II.2.118}$$

with

$$\frac{P_{\text{out}}}{P_{\text{in}}} = G_P = |G_V|^2 \tag{II.2.119}$$

being the *power gain* of the matched ideal amplifier, which is also called *transducer gain*.

Equations (II.2.64) and (II.2.117) to (II.2.119) are defined for an ideal amplifier with matched ports ($S_{11} = S_{22} = 0$)! *Power gain* and other definitions are different for unmatched *gain stages* with matching sections at the input and output.

As simple as the ideal amplifier is from a mathematical point of view, the real world RF amplifier is the most complicated RF component, and for particle accelerators also a very important component! Before going through a few examples of RF power amplifiers typically used for accelerators, let us briefly cover the most important *figures of merit* of the general RF amplifier.

II.2.8.4.1 3 dB bandwidth (BW)

The gain g of RF amplifiers, sometimes also called *gain stages*, is frequency dependent, in other words, any RF amplifier is designed for a given frequency range, and most RF amplifiers have zero gain at DC (direct current, 0 Hz). Therefore, the RF amplifier has a band-pass filter like characteristic, see Fig. II.2.132, with the 3 dB bandwidth defined by the higher f_{HI} and lower f_{LO} 3 dB cut-off frequencies

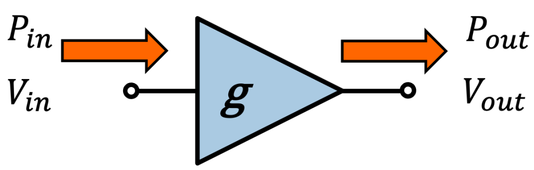


Fig. II.2.131: Schematic symbol for an amplifier.

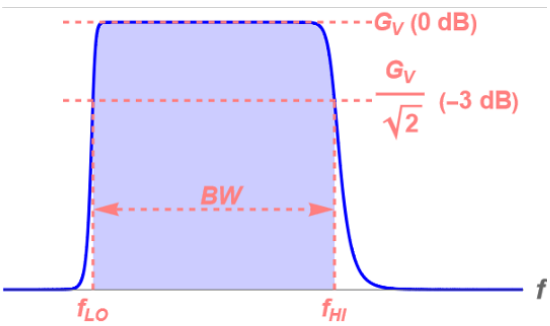


Fig. II.2.132: Amplifier bandwidth definition.

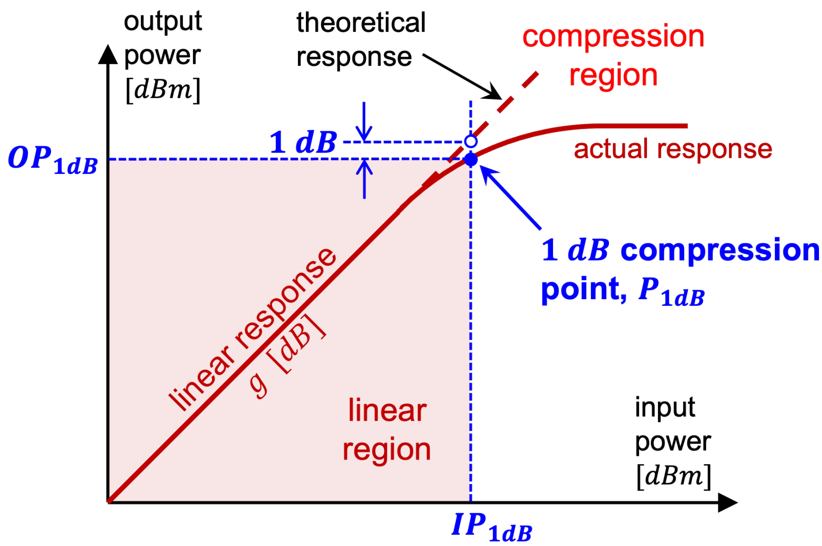


Fig. II.2.133: Output vs. input power and 1 dB compression point of an amplifier.

based on Eq. (II.2.115):

$$BW = f_{BW} = f_{HI} - f_{LO} \tag{II.2.120}$$

In most cases in accelerators, the RF amplifiers don't require to be broadband, as the RF signal to be amplified is a continuous-wave (CW) signal or a narrow-band signal, perhaps with some modulation. Exceptions exist, e.g. amplifiers for pulsed RF or beam instrumentation applications. In the latter case the transient behaviour of the amplifier needs to be taken into account, please note, the band-pass characteristic of industry broadband RF amplifiers typically is of *Butterworth* or *Chebyshev* response, thus, the group-delay varies substantially at the band edges (near f_{LO} and f_{HI}) which may result in unwanted ringing (oscillations) in the time-domain transient response.

> A good rule of thumb in selecting the appropriate bandwidth for an RF amplifier application: Just as much bandwidth as necessary!
> Unnecessary large bandwidth results in additional noise and / or the amplification of unwanted signal content.

II.2.8.4.2 Linearity

The ideal amplifier is a linear element

$$P_{\text{out}} = G_P P_{\text{in}} \quad \text{or} \quad P_{\text{out}} = 10^{\frac{g}{10}} P_{\text{in}} \quad \forall\, P_{\text{in}} \tag{II.2.121}$$

as indicated by the dashed *theoretical response* trace in Fig. II.2.133.

II.2.8.4.2.1 1 dB *compression point* ($P_{1\,\text{dB}}$)

Unfortunately, the ideal amplifier does not exist, as a matter of fact, any amplifier exhibits a non-linear behaviour, in particular at higher excitation levels at the input. Figure II.2.133 displays the *actual re-*

sponse (solid trace) of the output power P_{out} vs. the input power P_{in} for a typical RF amplifier, indicating the deviation from the *theoretical response* (dashed trace) of an ideal amplifier at higher input power levels.

Definition II.2.8.2: 1 dB compression point (P_{1dB})

The *1 dB compression point* of an amplifier, also called P_{1dB}, indicates the power level at which the output power differs by $1\,\text{dB}$ from that of an ideal amplifier, see also Fig. II.2.133

The *gain elements* used in a RF amplifier, e.g. semiconductors (transistors, FETs, etc.) or sometimes vacuum tubes, have an intrinsic non-linear characteristic, but fortunately their non-linear behaviour can be linearised to almost perfectly by an appropriate feedback circuit topology. Therefore, the dominant non-linear effect of an amplifier is related to the internal supply rail voltage levels, the maximum output voltage – and the equivalent output power into a load resistor – is limited by that rail voltage. If we drive the amplifier beyond, e.g. with a perfect sinusoidal signal, the amplifier will *saturate* and the output waveform "clipped", the pure sinusoidal waveform deforms and the amplifier generates unwanted higher-order harmonics at the output (compression region in Fig. II.2.133).

As Fig. II.2.133 indicates, the $1\,\text{dB}$ compression point is defined for the output power of the amplifier, $P_{1dB} = OP_{1dB}$, i.e. the *output 1 dB compression point*. However, sometimes RF amplifier manufacturers specify the $1\,\text{dB}$ compression for the equivalent input power, IP_{1dB}, which is related to OP_{1dB}:

$$P_{1dB}\,[dBm] = OP_{1dB}\,[dBm] = IP_{1dB}\,[dBm] + g\,[dB] \tag{II.2.122}$$

Example II.2.8.5: dBm is not dB: $dBm \neq dB$

The SI unit for power is *Watt* $[W = VI]$, however, in RF engineering often the preferred unit for power is dBm, as Fig. II.2.133 and Eq. (II.2.122) indicate. dBm is a power unit utilising an absolute logarithmic scale similar to the dB scale, but with a reference of $P_{\text{ref}} = 1\,\text{mW}$:

$$P\,[dBm] = 10\log_{10}\frac{P\,[W]}{P_{\text{ref}}} \quad \text{with: } P_{\text{ref}} = 1\,\text{mW} \tag{II.2.123}$$

Equation (II.2.123) simplifies the power calculation in a chain of RF components with gain (amplifiers) and insertion loss (attenuators, transmission-lines, and other passive elements) to a simple addition of their $dB(m)$-values, see as example Eq. (II.2.122). Some popular values reflecting Eq. (II.2.123) are listed in Table II.2.12, along with the equivalent RMS voltage for a sinusoidal signal into a load resistor of value $R_l = 50\,\Omega$.
In practice the maximum output power level for a low or medium power RF amplifier is usually expressed in dBm, in terms of the $1\,\text{dB}$ compression point, however, the output level for high power RF amplifiers is usually expressed in W.

Table II.2.12: RF power and voltage into $R_l = 50\,\Omega$.

$P\,[dBm]$	$P\,[W]$	$V\,[RMS]$ for $R_l = 50\,\Omega$
$+\ 30\,\mathrm{dBm}$	$1\ \ \mathrm{W}$	$7.07\ \ \mathrm{V}$
$+\ 20\,\mathrm{dBm}$	$100\,\mathrm{mW}$	$2.24\ \ \mathrm{V}$
$\mathbf{0\,dBm}$	$\mathbf{1\,mW}$	$\mathbf{224\,mV}$
$-\ 20\,\mathrm{dBm}$	$10\,\mu\mathrm{W}$	$22.4\,\mathrm{mV}$
$-120\,\mathrm{dBm}$	$1\,\mathrm{fW}$	$224\,\mathrm{nV}$
$-174\,\mathrm{dBm}$	$4 \times 10^{-21}\,\mathrm{W}$	$446\,\mathrm{pV}$

noise power in a bandwidth $BW = 1\,\mathrm{Hz}$ at room temperature

II.2.8.4.2.2 3^{rd}-order intercept (TOI) point, $IP3$

The non-linear compression region of an RF amplifier results in so-called *intermodulation* (IM) effects for two-tone or other more complex signals beyond a pure sinusoidal signal excitation. This leads to another specification for the linearity of a RF amplifier, the 3^{rd}*-order intercept (TOI) point*, and it is de-facto the most common, simple, safe and reproducible way to quantify the *linearity* of a device. Instead of driving the amplifier into the highly non-linear saturation region – as e.g. required to characterise the $1\,\mathrm{dB}$ compression point – and expressing the non-linearities, generated by the amplifier for a pure sinusoidal input signal excitation as *Taylor* series expansion for the higher-order harmonics at the output, a two-tone excitation measurement method is applied, keeping the amplifier in the quasi-linear regime and still characterising it's non-linear behaviour.

Again, linearity means the output signal of the device is directly proportional to the input signal, see Eq. (II.2.121). The higher the TOI point, the better the linearity of the amplifier and the lower the level of the so-called *intermodulation distortion* it may generate. As explained, most active devices (RF amplifiers, gain blocks, etc.) are typically linear only over a certain input power range, but if a certain power level is exceeded, the device becomes non-linear. What happens then?

Operating in the non-linear compression region (see Fig. II.2.133) will create signal *distortions*, for a pure sine wave signal at the input there will be additional unwanted *harmonics* and *intermodulation (IM) products* (also called intermodulation distortion or IMD).

Harmonics are copies of a sinusoidal signal appearing at multiples of the fundamental frequency (f_1 is the fundamental or 1st harmonic, $2f_1$ is the 2nd harmonic, $3f_1$ is the 3rd harmonic, etc.). Typically the amplitude of the harmonics decreases as their order increases.

Intermodulation products occur when at least two (or more) signals (tones) *mix* in a non-linear device. Mixing produces new, additional signals as sum and difference of the original frequencies. Not only the input tone signals of slightly different frequencies f_L and f_U can mix with each other, but they can also mix with the higher order harmonics, such as $2f_L$, $2f_U$ etc. Hence, we have ad-

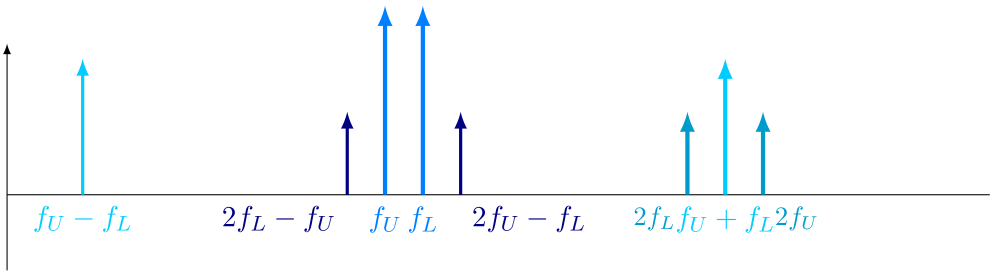


Fig. II.2.134: Two-tone (f_L, f_U) intermodulation products due to amplifier non-linearities.

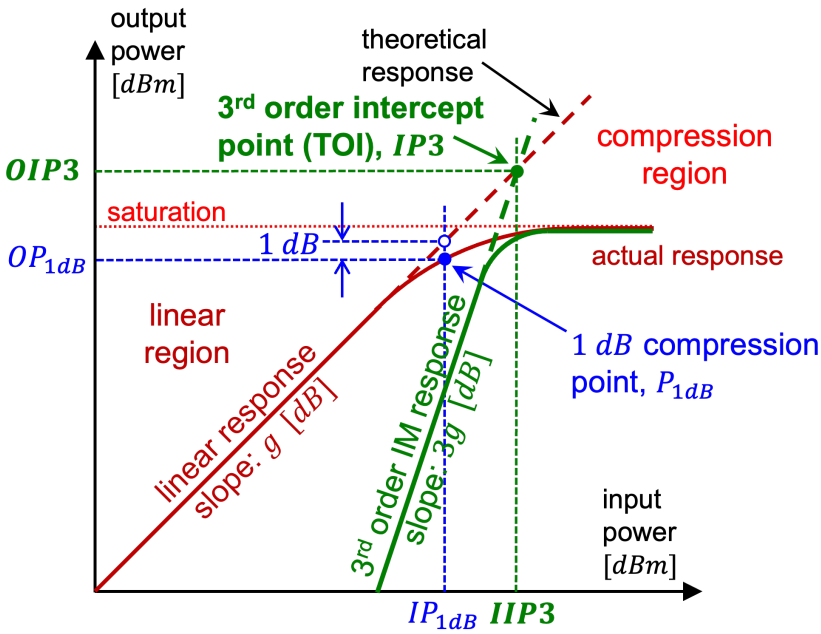


Fig. II.2.135: Output vs. input power and 3rd-order intercept (TOI) point of an amplifier.

ditional products at: $2f_L + f_U$, $2f_L - f_U$, $2f_U + f_L$, $2f_U - f_L$, ... (see Fig. II.2.134). The order i of the harmonics and intermodulation product is the sum of their (unsigned) coefficients.

In most cases the intermodulation products are easy to handle, higher order harmonics typically have very low amplitudes and can usually be ignored or filtered, and even higher-frequency products often fall outside the amplifier bandwidth.

The typical safe way to reject these products is through, e.g. band-pass filtering, however, this becomes particular difficult if the unwanted mixing products are very close to the operating frequency band of the fundamental signals. This is the case for the 3rd-order products $2f_L - f_U$ and $2f_U - f_L$, which are difficult to reject by filtering due to their proximity to the two fundamentals f_L and f_U, see Fig. II.2.134.

Figure II.2.135, as an extension of Fig. II.2.133, illustrates the TOI point concept graphically. In this double-logarithmic output vs. input power graph we have the ideal and the actual response of an amplifier, dashed and solid dark-red traces, which expresses the gain g of the amplifier. The green trace

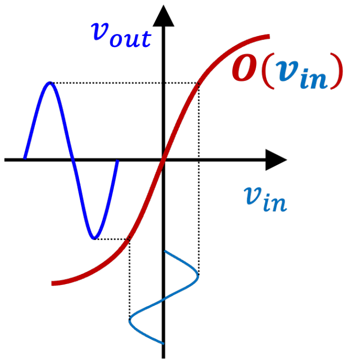


Fig. II.2.136: Symmetric voltage transfer function $\mathcal{O}(v_{\text{in}})$ of a typical RF amplifier.

shows the response of the amplifier for the 3rd-order intermodulation harmonic if the imperfect amplifier is supplied with a two-tone signal f_L, f_U.

Please note, the response to the 3rd-order IM harmonic is different from the response to the fundamental tone, it shows up at higher input power levels and has a steeper gradient of $3g$. This leads to the definition of the 3rd-order intercept point:

Definition II.2.8.3: 3rd order intercept point ($IP3$)

The 3rd-order intercept (TOI) point, also called $IP3$, is the crossing of the extrapolated ideal linear gain response g of an amplifier in a $V_{\text{out}}\,[dBm] = f(V_{\text{in}}\,[dBm])$ graph with the also extrapolated 3rd-order intermodulation response to a two-tone excitation signal, see Fig. II.2.135. The TOI point is a purely mathematical concept, the actual IP3 value lies in the compression region, beyond the saturation of the amplifier.

In practice it is sufficient to characterise the slopes (the gradients) of the linear response and of the 3rd-order IM response to locate the IP3, and as Fig. II.2.135 demonstrated, this can be accomplished with modest signal levels.

Manufactures of RF amplifiers specify the IP3 value, often along with the $P_{1dB} value$, and again attention has to be paid if this value is related to the input power $IIP3$, or to the output power $OIP3$, see Fig. II.2.135.

II.2.8.4.2.3 Relation between TOI point and $1\,\text{dB}$ *compression point*

As explained in the previous paragraph, the 3rd-order intercept point is based on the 3rd-order IM nonlinearities of an amplifier caused by two sinusoidal RF signals

$$v_{\text{in}}(t) = \hat{a}\cos\omega t = \hat{a}\cos 2\pi f t \tag{II.2.124}$$

of same amplitude $\hat{a}$ and slightly different frequency:

$$f \in f_L,\, f_U \quad \text{with: } \Delta f = \frac{f_U - f_L}{2} \ll \frac{f_U + f_L}{2}$$

resulting in the 3rd-order IM products:

$$\begin{aligned} f_{U3rd} &= \frac{f_U + f_L}{2} + 3\Delta f \\ f_{L3rd} &= \frac{f_U + f_L}{2} - 3\Delta f \end{aligned}$$

While non-linear at large signals, the voltage transfer function $\mathcal{O}(v_{\text{in}})$ of most RF amplifiers is symmetric:

$$\mathcal{O}(-v_{\text{in}}) = -\mathcal{O}(v_{\text{in}}),$$

as depicted in Fig. II.2.136, and of course is limited due to the internal supply voltages. The non-linear behaviour of $\mathcal{O}(v_{\text{in}})$ can be expressed as *Taylor* series expansion

$$\mathcal{O}(v_{\text{in}}) = G_V v_{\text{in}} - D_3 v_{\text{in}}^3 + \ldots$$

with G_V being the linear voltage gain and D_3 accounting for the 3rd-order distortions. If we further make use of the *Taylor* series approximation:

$$\cos^3(\omega t) \approxeq \frac{4}{3}\cos(\omega t) + \frac{1}{4}\cos(3\omega t)$$

we find that the output signal of the amplifier follows approximately

$$\mathcal{O}\left(v_{\text{in}}(t)\right) \approxeq \left(G_V - \frac{3}{4} D_3 \hat{a}^2\right) \hat{a}\cos(\omega t) - \frac{1}{4} D_3 \hat{a}^3 \cos(3\omega t) \tag{II.2.125}$$

The 3rd-order intercept point is found were the linear component G_V and the non-linear component D_3 of the fundamental harmonic $\cos(\omega t)$ in Eq. (II.2.125) are equal:

$$\hat{a}^2 = \frac{4G_V}{3D_3} \tag{II.2.126}$$

The $1\,\text{dB}$ compression point is located $1\,\text{dB} \equiv 10^{1/20} \approxeq 1.122$ below the linear (ideal) amplifier characteristic:

$$1.122\left(G_V - \frac{4}{3} D_3 \hat{a}^2\right)\hat{a} = G_V \hat{a} \quad \Rightarrow \quad \hat{a}^2 = \frac{10^{1/20} - 1}{10^{1/20}} \frac{4G_V}{3D_3} \approxeq 0.10875 \frac{4G_V}{3D_3} \tag{II.2.127}$$

Comparing Eq. (II.2.126) and Eq. (II.2.127), we find $10\log_{10}(0.10875) \approxeq -9.6357\,\text{dB}$, thus the $1\,\text{dB}$ compression point is approximately $9.6\,\text{dB}$ below the TOI point.

II.2.8.4.3 Noise in RF amplifiers

The *noise figure* (NF) characterises the **degradation** of the *signal-to-noise ratio* (SNR) of an electronic device, component, (sub)system, etc., here our low-noise RF amplifier. As a signal passes through a system or device, t he noise figure tells us the relative amount of noise that is being **added** to the signal as it travels through the system or device.

By evaluating the noise figure of e.g. an amplifier, we can calculate the *sensitivity* from its bandwidth. The value of the noise figure value is a key parameter when handling low-level signals and enables to quantify the added noise of the amplifier, network or system.

In an ideal world, an amplifier with nothing connected at the input, should have no signal at the output. However, in the real world we have noise!
Different types of noise exist, but the most dominant is thermal noise. In a real-world application we expect the thermal noise at the input will be amplified by the gain of the amplifier, appearing at the output.

II.2.8.4.3.1 Noise basics

Before introducing *noise factor* and *noise figure* of an amplifier, let's briefly review what *noise* in electronics and RF systems mean:

The concept of "noise" was originally studied for audible sound caused by statistical variations of the air pressure over a wide, flat frequency spectrum (white noise). It is also used for electrical signals, with the "noise floor" determining the lower limit of the signal transmission. Typical noise sources are: *Brownian* movement of charges (thermal noise), variations of the number of charges involved in the conduction (flicker noise) and quantum effects (*Schottky* noise, shot noise). Thermal noise is only emitted by structures with electromagnetic losses, which, by reciprocity, also absorb power. Pure reactances do not emit noise (emissivity = 0).

Different categories of noise have been defined:

white which has a flat spectrum,

pink being low-pass filtered and

blue being high-pass filtered.

In addition to the spectral distribution, the amplitude density distribution is also required in order to characterise a stochastic signal. For signals generated by superposition of many independent sources, the amplitude density has a *Gaussian* distribution. The *noise power density* delivered to a load by a black body is given by *Planck's* formula:

$$\frac{P_N}{\Delta f} = hf \left(\mathrm{e}^{hf/k_B T} - 1 \right)^{-1}, \tag{II.2.128}$$

where P_N is the noise power delivered to a load, $h = 6.625 \times 10^{-34}$ Js the *Planck* constant, and $k_B = 1.380\,56 \times 10^{-23}$ J/K the *Boltzmann* constant.

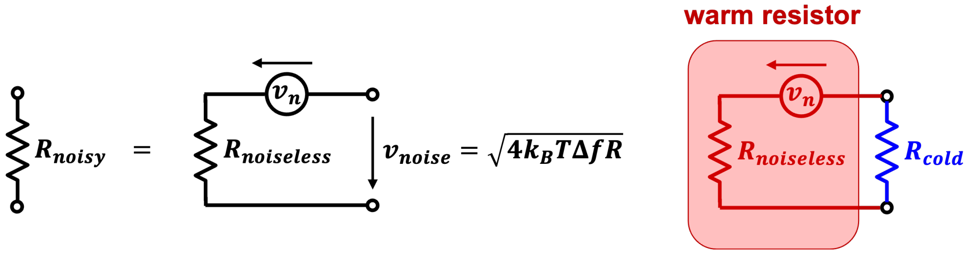


Fig. II.2.137: Equivalent circuit of a noisy resistor terminated by a noiseless load.

Equation (II.2.128) indicates a constant noise power density up to about $f = 120\,\text{GHz}$ (at $T = 290\,\text{K}$) with $1\,\%$ error. Beyond that, the power density decays and there is no "ultraviolet catastrophe", i.e. the total integrated noise power is finite.

The radiated power density of a black body is given as

$$W_\text{r}(f,T) = \frac{hf^3}{c^2\left[\text{e}^{hf/k_BT}-1\right]}. \tag{II.2.129}$$

For $hf \ll k_BT$ the *Rayleigh–Jeans* approximation of Eq. (II.2.128) holds, and:

$$P_N \approxeq k_BT\Delta f, \tag{II.2.130}$$

is the *noise power* delivered to a matched load of a given frequency range of bandwidth $B = \Delta f$, T is called the *noise temperature*, which does not have to be equal to the physical temperature of the device. The noise voltage $v(t)$ of a resistor R with no load is given as

$$\overline{v^2(t)} = 4k_BTR\Delta f \tag{II.2.131}$$

and the short-circuit current $i(t)$ by

$$\overline{i^2(t)} = 4\frac{k_BT\Delta f}{R} = 4k_BTG\Delta f, \tag{II.2.132}$$

where $v(t)$ and $i(t)$ are stochastic signals, and G is $1/R$. The linear averages $\overline{v(t)}, \overline{i(t)}$ vanish, of importance are the quadratic averages $\overline{v^2(t)}, \overline{i^2(t)}$, which do not vanish. The available power, which is independent of R, is given by (see also Fig. II.2.137):

$$\frac{\overline{v^2(t)}}{4R} = k_BT\Delta f. \tag{II.2.133}$$

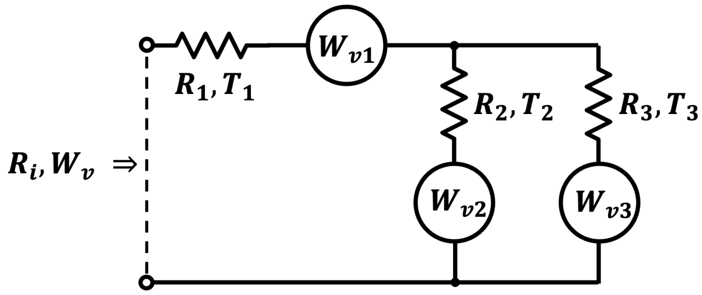


Fig. II.2.138: Noisy 1-port with $j = 3$ resistors of different temperatures [21, 22].

from which the *spectral density function* is defined as [21][5]

$$\begin{aligned} W_\mathrm{v}(f) &= 4k_BTR, \\ W_\mathrm{i}(f) &= 4k_BTG, \\ \overline{v^2(t)} &= \int_{f_1}^{f_2} W_\mathrm{v}(f)\mathrm{d}f. \end{aligned} \tag{II.2.134}$$

A noisy resistor may be composed of several components (resistive network). Typically, it is made from a carbon grain structure, which has a homogeneous temperature. But if we consider a network of resistors with different temperatures, and hence with an inhomogeneous temperature distribution (Fig. II.2.138), the spectral density function becomes

$$W_\mathrm{v} = \sum_j W_{\mathrm{v}j} = 4k_BT_\mathrm{n}R_\mathrm{i}, \tag{II.2.135}$$

where $W_{\mathrm{v}j}$ are the individual noise sources (see also Fig. II.2.139), T_n is the total noise temperature, and R_i the total input impedance. For simplicity we assume all $W_{\mathrm{v}j}$ are uncorrelated.

The relative contribution (β_j) of a lossy element to the total noise temperature T_n is equal to the relative dissipated power multiplied by its temperature T_j:

$$T_\mathrm{n} = \beta_1T_1 + \beta_2T_2 + \beta_3T_3 + \cdots = \sum_j \beta_jT_j \tag{II.2.136}$$

[5] All citations in this paragraph are text books in German language, however, similar books and articles exist in English.

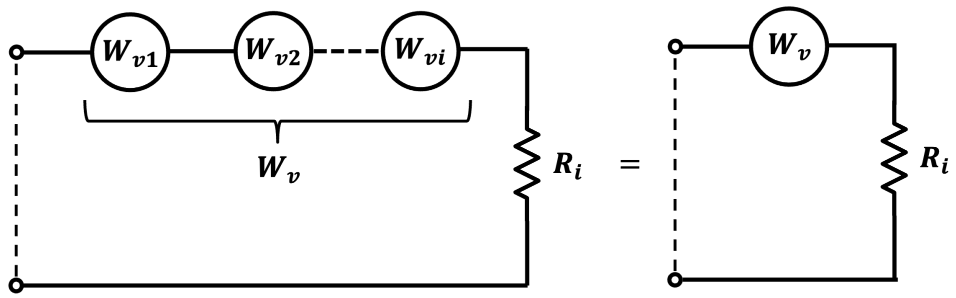


Fig. II.2.139: Equivalent sources for the circuit of Fig. II.2.138.

Please note, in Eq. (II.2.135) β_j are the coefficients related to the fractional part of the input power dissipated in each resistor R_j.

So far, only pure resistors have been considered. Looking at complex impedances, it is evident, losses occur only from dissipation in $\mathrm{Re}(Z)$. The available noise power is independent of the magnitude of $\mathrm{Re}(Z)$ with $\mathrm{Re}(Z) > 0$. For Figs. II.2.138 and II.2.139, Eq. (II.2.135) still applies, except R_i is replaced by $\mathrm{Re}(Z_\mathrm{i})$. However, in complex impedance networks the spectral power density W_v becomes frequency dependent [22].

Example II.2.8.6: Noise temperature

A good example is the noise temperature of a satellite receiver, which is nothing else than a directional antenna. The noise temperature of free space amounts roughly to 3 K. The losses in the atmosphere, which is an air layer of $10\ldots20\,\mathrm{km}$ height, causes a noise temperature at the antenna output of approximately $10\ldots50\,\mathrm{K}$. This is well below our reference ambient temperature of $T_0 = 290\,\mathrm{K}$.
The rules mentioned above apply to passive structures. A forward-biased (by an external power supply) *Schottky* diode has a noise temperature of about $T_0/2{+}10\,\%$. As the biased *Schottky* diode is not in it's thermodynamic equilibrium, only half of the carriers contribute to the noise [21]. Still, it represents a real $50\,\Omega$ resistor when properly forward biased.
For transistors, in particular field-effect transistors (FETs), the physical mechanisms are somewhat more complicated. Noise temperatures of 50 K have been observed for a FET operating at a physical (device) temperature of 290 K.

II.2.8.4.3.2 Noise factor and noise figure

The *Noise factor* or *noise figure* is a figure of merit that applies for RF amplifiers operating near the noise floor, i.e. amplifying RF signals of very low levels. These types of amplifiers are called *low-noise amplifiers* (LNA).

As explained in the previous paragraph, any resistive source impedance R at the input of an RF amplifier will contribute with some noise power $N_P = N_i$, Eqs. (II.2.130) and (II.2.133), to the signal power at the input S_i to be amplified. An ideal amplifier would not alter the *signal-to-noise ratio* (SNR),

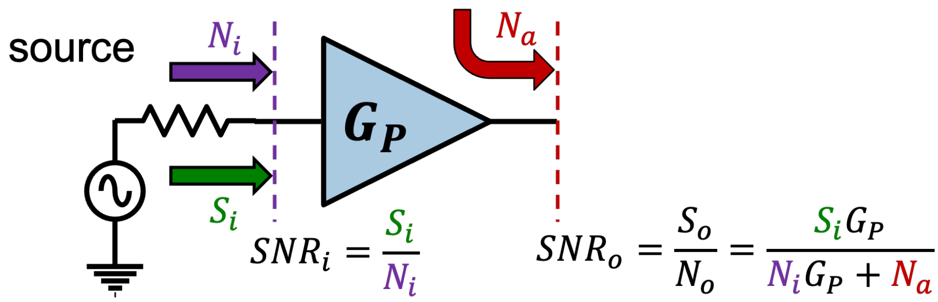


Fig. II.2.140: Noise power N_r added by an amplifier.

$SNR_i = SNR_o$:

$$\frac{SNR_i}{SNR_o} = \frac{S_i/N_i}{S_o/N_o} = 1$$

Definition II.2.8.4: Noise factor F

However, even a very good real-world LNR will add some noise power N_a, which leads to the definition of the noise factor (see also Fig. II.2.140):

$$F = \frac{SNR_i}{SNR_o} = \frac{S_i/N_i}{S_o/N_o} = 1 + \underbrace{\frac{N_a}{N_i G_P}}_{\text{added noise}} \tag{II.2.137}$$

which is always greater than 1 ($F > 1$).

Noise factor F and noise temperature T are related as:

$$F = 1 + \frac{T}{T_0} \tag{II.2.138}$$

with T_0 being the physical temperature of the amplifier or device.

Cascading (daisy chaining) RF amplifiers or devices, each with a given power gain G_{Pi} and noise factor N_i leads to an overall noise factor

$$F_{tot} = F_1 + \frac{F_2 - 1}{G_{P1}} + \frac{F_3 - 1}{G_{P1}G_{P2}} + \cdots + \frac{F_n}{G_{P1}G_{P1}\ldots G_{Pn-1}} \tag{II.2.139}$$

Evidently, the overall noise factor F_{tot} is dominated by the noise factor F_1 of the 1st element in the chain.

Definition II.2.8.5: Noise figure NF

The definition of the noise figure NF is based on the noise factor, Eq. (II.2.137), and is simply the logarithmic ratio of input and output SNR:

$$NF = 10\log_{10}(F) = 10\log_{10}\left(\frac{SNR_i}{SNR_o}\right) \tag{II.2.140}$$

The noise figure of an attenuator, Eq. (II.2.63), is equal to it's attenuation value, $NF = \Delta dB$, assuming it is operating at $T = T_0$. Following Eq. (II.2.139), it becomes clear that, minimising the noise contribution, also minimises attenuation and insertion losses from passive devices, e.g. cables, filters, attenuators, hybrids, located in front of the 1st LNA gain stage!

II.2.8.4.4 RF power amplifiers

Going back to the principle schematic of an accelerator RF system, Fig. II.2.1, the high power RF system has to bridge the gap between the output signal generated by the LLRF system, which typically is in the mW-regime, $0\ldots+20\,\mathrm{dBm}$, and the high power RF drive signal for the accelerating structure, as required by the beam, typically in the kW or even MW-regime, e.g. $+60\ldots+90\,\mathrm{dBm}$, please remember Eq. (II.2.123). As for the large gain required, $g>60\,\mathrm{dB}$, the RF power amplifier (PA) of an accelerator is sometimes separated into a drive stage and a high power stage. The drive stage is often an off-the-shelf commercial RF power amplifier, $P_{\text{out}}=1\ldots10\,\mathrm{kW}$, while the high power stage in most cases is a custom tailored or in-house developed RF power stage, optimised for a particular accelerator and it's beam conditions.

In the following we focus on the RF power amplifier output stage. The *ideal* RF power amplifier should:

- have a perfect **linear gain**, $g=10\log_{10}(P_{\text{out}}/P_{\text{in}})\ [dB]$, **no saturation** and **infinite power**.
- have **no,** or only **minimum delay**.
- be perfectly **impedance matched** at input and output, i.e. no reflections $S_{11}=S_{22}=0$.
- **not add any noise**.
- be **unconditionally stable** and **resistant against reverse power**, $S_{12}=0$
- be **radiation hard**.

Furthermore, the ideal power amplifier should be **efficient** and convert all AC wall plug power into RF power at the output.

Definition II.2.8.6: Power amplifier efficiency (PAE)

$$PAE=\frac{P_{\text{out}}-P_{\text{in}}}{P_{ACwall}} \tag{II.2.141}$$

Beside operating frequency f_{RF} and / or the bandwidth BW, a major question we are facing for the specification of a RF power amplifier, e.g. for a new accelerator is: How much power is required?!

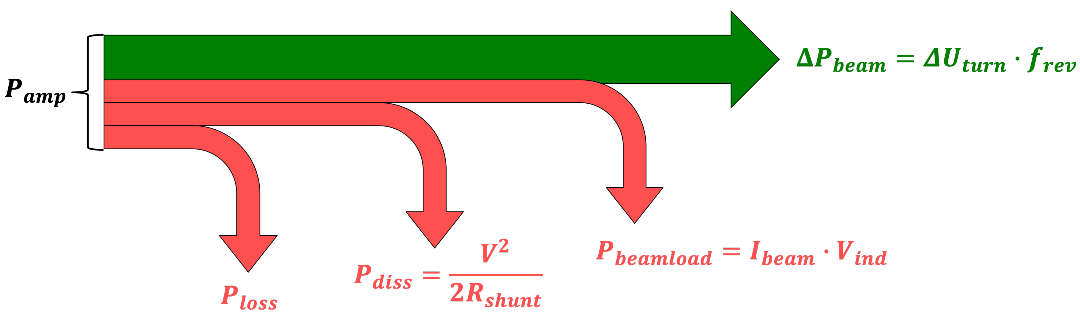


Fig. II.2.141: RF power requirement for a storage ring.

Figure II.2.141 illustrates the RF power requirement in a simplified graph for a storage ring accelerator. The amplifier has to deliver sufficient RF power:

- to accelerate the beam $\rightarrow$ *wanted*
  For a storage ring this equivalent to the power loss the beam experiences, e.g. due to synchrotron radiation.
  $$\Delta P_{\text{beam}} = \Delta U_{turn} f_{rev}$$
  For a synchrotron or linac this is the power required to increase the beam momentum, thus the beam energy.

- to compensate the beam induced voltage V_{ind} $\rightarrow$ P_{refl} $\rightarrow$ *heat*
  This is the power required to compensate the *beam loading*, e.g. generated in the accelerating mode of a standing wave cavity, which appears as reflected power at the main coupler and ultimately is transformed into heat in the dummy load of the circulator.
  $$P_{\text{beamload}} = I_{\text{beam}} V_{\text{ind}}$$
  Also the power losses due to the *beam coupling impedance* of passive elements like beam monitors, kickers, vacuum pumps, bellows, etc. need to be taken into account.

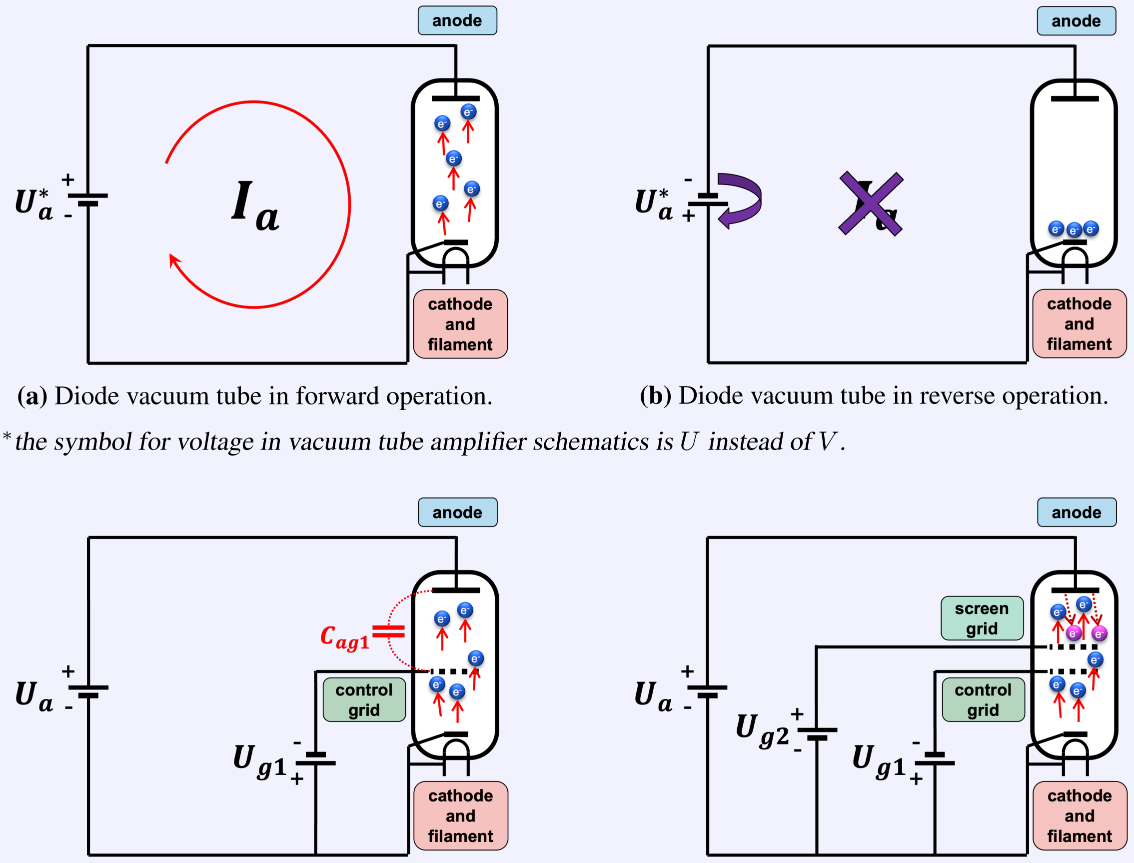
- to compensate the resistive wall losses of the cavity $\rightarrow$ *heat*
  This power loss is defined by the unloaded Q-factor Q_0 of the accelerating mode, e.g. for a normal conducting standing-wave mode cavity,
  $$P_{\text{diss}} = \frac{V^2}{2R_{\text{shunt}}} \quad \text{with: } R_{\text{shunt}} = \frac{R}{Q} Q_0,$$
  and basically is related to the electric conductivity (ρ or κ) of it's surface wall material within the skin-depth δ.

- to compensate the electrical losses in the RF distribution system $\rightarrow$ *heat*
  This is the sum of all *insertion losses* of the RF distribution system located between the output of the power amplifier and the main RF power coupler input of the accelerating structure. It is mainly the losses of the high power transmission-lines (coaxial lines or waveguides), but also includes the losses of the directional coupler and the circulator. Figure II.2.124 illustrates a very simple arrangement of a single power amplifier driving a single accelerating cavity, however, often a single power stage has to drive multiple accelerating cavities or vice versa, in that case the losses of the power splitters / combiners need to be included.

II.2.8.4.4.1 The tetrode-based RF power amplifier

While vacuum tubes no longer play a role in today's consumer and commercial electronics systems, they are still an important asset as high RF power source for particle accelerators. Let's walk through some basics in a simple, graphically oriented way:

Example II.2.8.7: Basics of a grid tube power amplifier



(a) Diode vacuum tube in forward operation. **(b)** Diode vacuum tube in reverse operation.

** the symbol for voltage in vacuum tube amplifier schematics is U instead of V.*

(c) Triode vacuum tube operation. **(d)** Tetrode vacuum tube operation.

Fig. II.2.142: From a diode to a tetrode.

A *vacuum tube* with a *cathode* based on coated metals, e.g. carbites, borides, nitrides, etc., and heated by a filament enables the thermionic emission of electrons, forming a cloud of free electrons near the cathode. Adding another electrode, the *anode*, and supplying a potential difference U_a between cathode and anode results in an anode current I_a, see Fig. II.2.142a. Reversing the supply voltage U_a stops the anode current, $I_a = 0$, see Fig. II.2.142b. Thus, this configuration is a *diode*.

Adding a *control grid*, see Fig. II.2.142c, transforms the passive vacuum diode to an active element, the *triode*, with the control grid voltage U_{g1} proportionally modulating the anode current $I_a \propto U_{g1}$. The most important parameter of a vacuum tube is the *transconductance* g_m, defined as

$$g_m = \frac{\Delta I_a}{\Delta U_{g1}}$$

At higher modulation frequencies on U_{g1} the parasitic capacitance C_{ag1} between anode a and control grid $g1$ limits the operation and tends to cause uncontrolled oscillations.

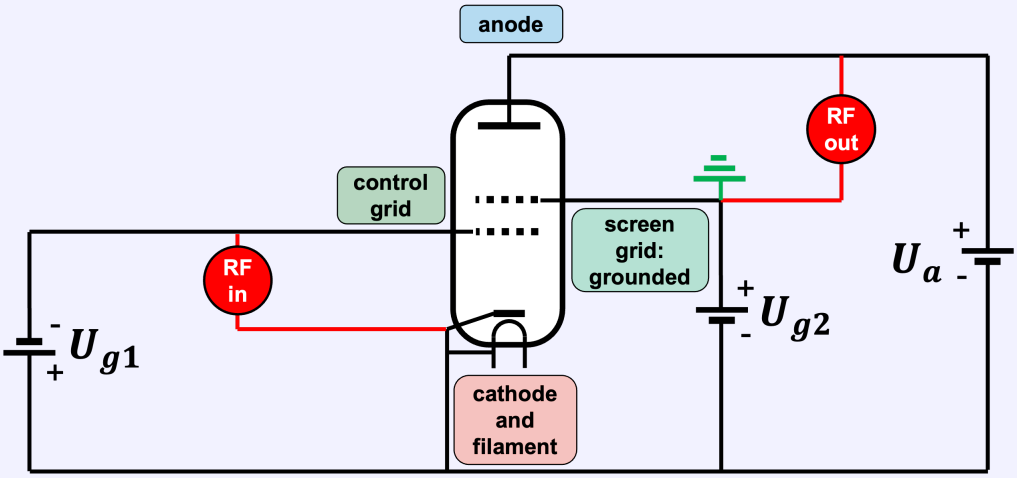


Fig. II.2.143: RF power requirement for a storage ring.

Adding another grid, the *screen grid* (sometimes called lower anode) between anode and control grid, biased with a fixed positive potential U_{g2}, decouples anode a and control grid $g1$, see Fig. II.2.142d. This *tetrode* configuration is most popular for high power RF vacuum tube amplifiers and typically gives a higher gain compared to a triode. However, some limitations are linked to the secondaries, generated at the anode from the incoming I_a electrons, which are accelerated towards the screen grid. This requires a special surface treatment of the anode to reduce the emission of secondaries.

Figure II.2.143 shows the highly simplified schematics of a *grounded grid* tetrode 200 MHz RF power amplifier, as it is still in operation since 1976 at the CERN Super Proton Synchrotron (SPS), based on *RS2004* tetrodes, see also Fig. II.2.144. Here the screen grid is on ground potential and the cathode is elevated to a negative potential.

Fig. II.2.144: CERN SPS 200 MHz, 1 MW transmitter station with eight *RS2004* tetrodes.

> Tetrode vacuum tube RF power amplifiers are still a good choice for high RF power output, $>100\,\mathrm{kW}$ CW per tube, on a small footprint, radiation tolerant(!), and therefore able to be located in very close proximity to the accelerating structures inside the accelerator tunnel. Unfortunately the know-how of this technology is vanishing!

RF power amplifiers based on tetrodes typically have a narrow-band operating frequency range, often tuned using an internal resonator to the specific RF frequency of interest. Tetrode-based RF amplifiers can reach an output power of $1\,\mathrm{MW}$ and more, the operating frequency is usually below $200\,\mathrm{MHz}$.

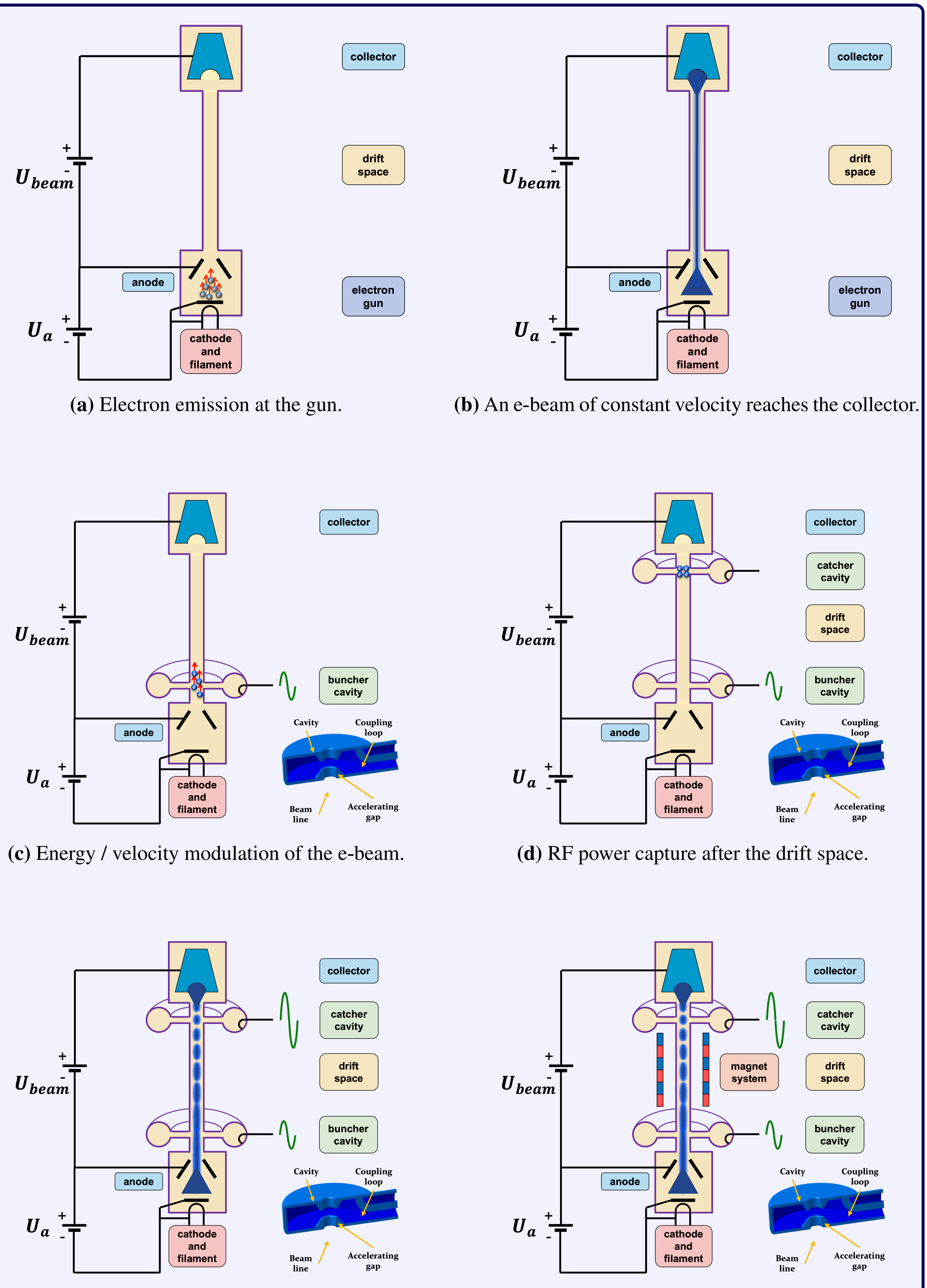
II.2.8.4.4.2 A high output power RF source: The klystron

For high RF output power at high frequencies the linear beam type, the so-called *klystron*, is the only option. The klystron is not a typical RF amplifier, instead it is a kind of *mini linear accelerator* in itself, perhaps the most complicated piece of accelerator RF technology! Here we cover only the operational principle in a non-mathematical way:

Example II.2.8.8: Basics of a linear beam tube

Figure II.2.145 explains the operating principle of a klystron graphically in several steps:
At the lower part in, see Fig. II.2.145a, we have the *electron gun* of the klystron, which is similar to a vacuum diode, see also Fig. II.2.142a. Here we illustrate the emission of the electrons from the heated cathode and accelerated by U_a towards the anode.
The anode of the klystron has a "hole", allowing the free electrons to pass through and be further accelerated by the U_{beam} voltage, through a *drift space*, towards the *collector*, which is on the higher potential than the anode. This results in a DC beam current, with the electrons moving at constant velocity between gun and collector, illustrated as dark blue stroke in Fig. II.2.145b.
The beam voltage U_{beam} is in the order of some $100\,\mathrm{kV}$, therefore the electrons travel with non-relativistic velocity. A single cell *buncher cavity* is located at the beginning of the drift space, see Fig. II.2.145c, and supplied with a RF drive power for a velocity modulation of the electrons passing the cavity gap.
Due to the time-varying E-field the leading electrons in the gap of the buncher cavity receive less force compared to the trailing electrons, such that in the following drift space the leading electrons travel with lower velocity and fall behind, while the trailing electrons travel with higher velocity and catch-up. This results in a bunching process along the drift, such that the maximum (optimal) bunching happens at it's end, where one or more *catcher cavities* are located, see Fig. II.2.145d.
Here the **kinetic energy** of the bunched electrons is **converted into RF power**.
Figure II.2.145e illustrates the bunching process of the electron beam along the drift space. Please note, buncher and catcher cavities operate at the same resonant frequency. To compensate for the *space-charge* of the electron beam a *solenoidal magnet* is wrapped around the drift space, see Fig. II.2.145f.

> Notably, the klystron offers a relatively high gain on the order of $60\,\mathrm{dB}$, and also, to improve the efficiency *multi-beam klystrons* (MBK) have been developed.



(a) Electron emission at the gun.

(b) An e-beam of constant velocity reaches the collector.

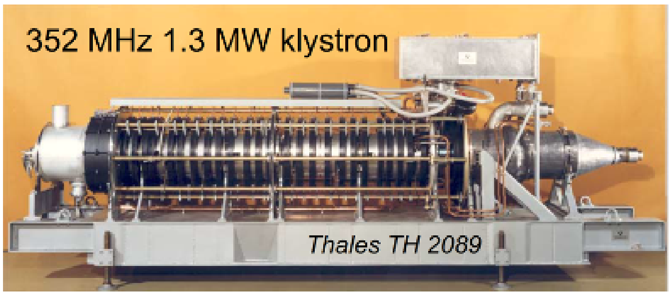
(c) Energy / velocity modulation of the e-beam.

(d) RF power capture after the drift space.

(e) Illustration of the e-beam bunching.

(f) Solenoidal magnet system to compensate space charge.

Fig. II.2.145: Operating principle of a klystron.



(a) 352 MHz, 1.3 MW klystron for CW operation, used at CERN LEP (until 2000).

(b) 12 GHz klystron for pulsed operation (CLIC), 50 MW during 1.5 μs pulse-length.

Fig. II.2.146: Two examples of a klystron.

Figure II.2.146 illustrates two examples of a klystron:
Figure II.2.146a shows the 352 MHZ CW klystron used at the CERN LEP ring accelerator until year 2000. LEP used many of those klystrons, a total of approximately 50 MW RF power was installed.
Figure II.2.146b shows the 12 GHz klystron developed for the CERN Compact LInear Collider (CLIC), operating in pulse mode.

Beside the klystron itself, there is a substantial set of auxiliary systems required to complete a klystron-based RF transmitter, e.g. high-voltage (HV) power supply, water cooling (mainly for the collector), radiation shielding, technical interlock and control systems, etc. For klystrons operating in pulse mode, a HV pulse modulator with a pulse-forming network (PFN) is required, moreover, sometimes a pulse compressor, e.g. a SLED (SLAC Energy Doubler) structure, composed out of two resonant cavities and a 3 dB, 90° waveguide hybrid coupler, is used as an efficient way to provide higher RF power.

Two-beam linear collider
Please note the *two-beam acceleration* principle, which is a spin-off of the klystron, proposed by CERN for a Compact LInear Collider (CLIC) [23]. Instead of individual klystrons feeding the acceleration structures via a waveguide distribution system, a 2$^{\mathrm{nd}}$, low-energy linear accelerator with a high current *drive beam* is located parallel to the high-energy linear accelerator, feeding the acceleration structures with the extracted RF power via very short waveguides. In a simplistic view, the two-beam acceleration principle acts as transformer, transforming the RF energy from a high-current, low-energy beam into a high-energy, low-current beam.

II.2.8.4.4.3 Solid-state RF power amplifiers

Tetrodes, klystrons and other vacuum tube based RF power sources have some common disadvantages, e.g.

- Ageing effects of the cathode limit the lifetime.
- A failure of a single element in the RF power source usually trips the beam stored in a ring accelerator.

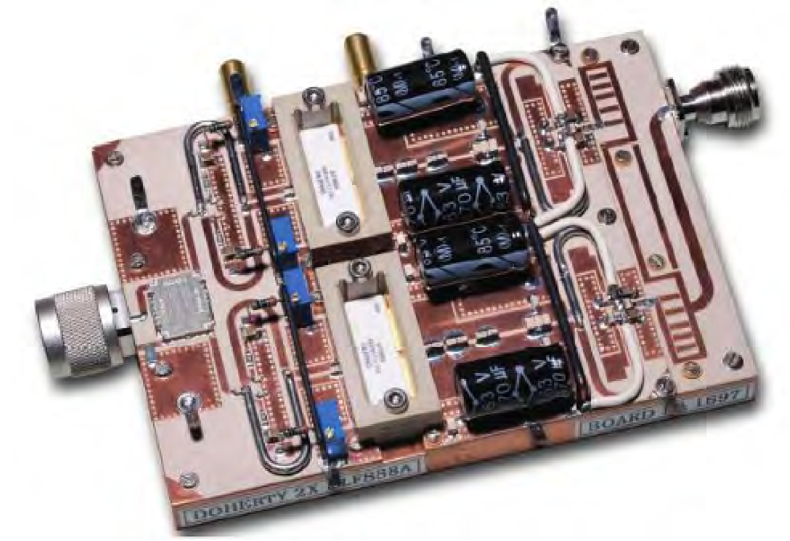


Fig. II.2.147: *Ampleon* AN11325 2-way *Doherty* RF amplifier module using *BFL888A* LDMOS transistors.

Recent developments in *solid state* technologies enable the use of *semiconductor*-based RF power amplifiers for accelerator applications. Instead of a single, or a few high-power RF gain elements, as typical for a tetrode, klystron, IOT or similar high-power RF sources, this *transistor*-based RF amplifier approach is based on small, compact gain modules, each delivering only moderate RF power, but with many of them combined, to provide the total required RF power. Figure II.2.147 pictures a typical industry RF power module for 200 W CW output power in the UHF frequency range. The solid-state RF power gain stage is based on a *push-pull* amplifier. Here we illustrate the principle of operation:

Example II.2.8.9: Basics of solid-state RF amplifiers

Solid-state RF amplifiers are based on the transistor as active element, and while there are different types of transistors, with different characteristics and operating principles, here we explain the basic principle of the RF power gain stage based on the traditional *bipolar junction transistor* (BJT), again, in a very simplistic, graphically supported way:

BJT's are manufactured in two flavours:
As NPN transistor, with the *basis* semiconductor area, made of e.g. silicon (Si), being doped with a material of positive (P) majority carriers, while the *emitter* and *collector* areas are doped with a negative (N) majority carrier material, and as PNP transistors with the doping of those areas in a complementary way, N-doping for the basis and P-doping for emitter and collector. As a result of this doping a BJT without or with little DC bias at the basis – in so-called *class B* or *class AB* operation – amplifies only positive (NPN) or negative (PNP) signals:

NPN BJT are active for positive signals at the input, $V_{\text{in}} > 0$,
and stays passive for negative input signals ($V_{\text{in}} < 0$).

PNP BJT are active for negative signals at the input, $V_{\text{in}} < 0$, and stays passive for positive input signals ($V_{\text{in}} > 0$).

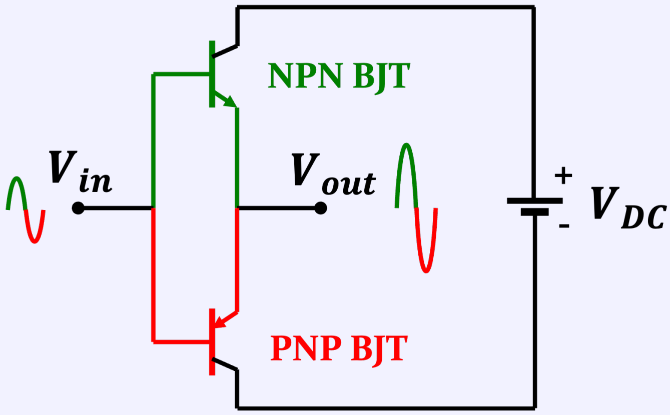


Fig. II.2.148: Operation principle of a NPN / PNP BJT pair in push-pull amplifier configuration.

Combining a NPN and a PNP BJT as illustrated in the very simplified schematic Fig. II.2.148 results in the push-pull amplifier configuration, which has a low bias current running through the transistors and therefore a high efficiency. A biasing network (not shown in Fig. II.2.148) is required to optimise for efficiency and low distortions, also impedance matching networks at in- and output are required.

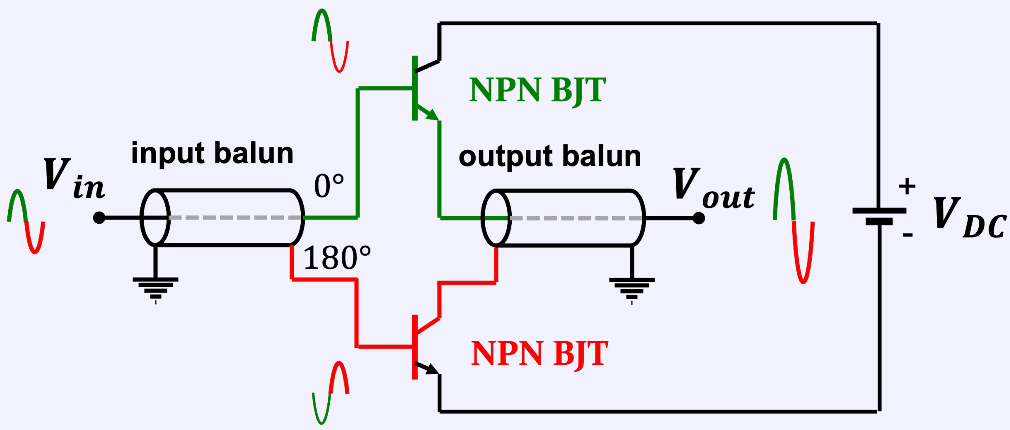


Fig. II.2.149: Operation principle of two NPN BJT's as push-pull amplifier.

The scheme of Fig. II.2.149 requires two complementary BJT's, NPN and PNP, with perfect symmetrical characteristics! Unfortunately industry so far is unable to deliver this ambitious goal. In practice, the characteristics of the NPN BJT's are far superior to their PNP counterparts. Figure II.2.149 shows the push-pull amplifier solution based on two identical NPN BJT's. For it's operation, a signal inverter is required at in- and output, here illustrated in form of a passive *bal-un*, i.e. *bal*anced-to-*un*balance network, here shown in the form as short piece of coaxial transmission-line. The balun provides a non-inverted (0°) and a inverted (180°) signal for the basis's of the two NPN BJT's, so that they operate in push-pull mode, i.e. only one transistor at a time is active per half-wave (see Fig. II.2.149). A similar balun is required to combine the two symmetric output waveforms into a single asymmetric (unbalanced) output signal.

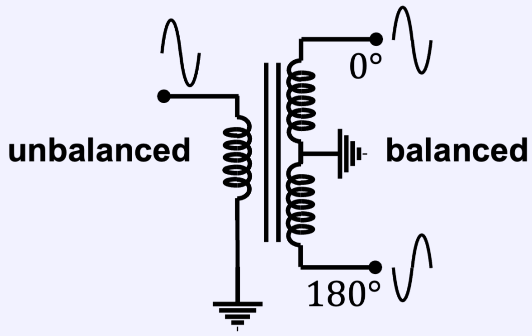
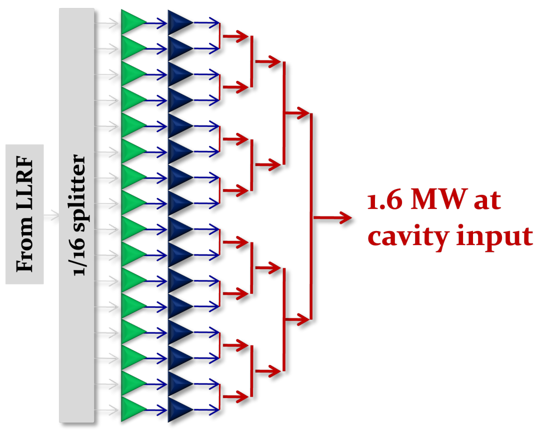


Fig. II.2.150: Balun: A balanced-to-unbalanced transformer (or vice versa).

There is a large variety of balun implementations, the coaxial-line realisation shown in Fig. II.2.149 is only one example. Circuit layout and implementation details depend a lot on the operating frequency and required RF power rating. For completeness and as reference, Fig. II.2.150 shows the balun in form of an ideal transformer.
The balun-based push-pull RF amplifier approach substantially simplifies the complementary symmetric transistor technology obstacle, now the amplifier module can be realised with two identical transistor of the same technology. Still, the design of an appropriate balun, particular for the output remains a challenge!

At the time of writing, LDMOS Si and GaN field-effect transistors (FET) are the preferred choice for most RF power amplifier modules. Their operation principle differs from a BJT, however, the balun-based push-pull mode operation remains. Typically, an individual RF power module achieves a few 100 W output power, therefore attention is given on a non-reflective power combiner with minimum insertion loss to combine the output power of many modules with correct timing, i.e. phase balance. Figure II.2.151 illustrates a solid-state RF power amplifier installation for the Super Proton Synchrotron

(a) Hall with 2×16 solid-state RF amplifier towers.



(b) 16-to-1 power combination schema.

Fig. II.2.151: 200 MHz, 2×1.6 MW solid-state RF amplifier installation at the SPS (CERN).

(SPS) at CERN. It requires an entire hall, see Fig. II.2.151a to accommodate one (of two) 200 MHz, 1.6 MW peak power RF amplifiers, each consisting of 2 × 16 towers, and each tower holding 80 RF amplifier modules (a total of 5120 modules). Figure II.2.151b shows a simplified schematic of the power combination for 16 PA towers, of course, here the "devil is in the details".

RF power amplifier: Tube or solid-state?

Table II.2.13: Tube vs. solid-state RF power amplifiers.

Prefer tube amplifier, when	Prefer solid-state amplifier, when
Amplifier(s) must be installed in the accelerator tunnel	Amplifier(s) can be located in a non-radioactive environment
Expecting substantial "spikes" from beam induced voltage	Circulator(s) can be installed to protect the amplifier
Large output power of a single device is required, without combiners	Time delay due to unavoidable combiner stages is of little relevance
Not much real-estate is available	Sufficient space can be made available
High peak power in pulse mode	Continuous wave (CW) operation
Amplifier(s) must be compact and/or in close proximity to the cavity	Amplifier(s) can be located far away from the cavity

Table II.2.13 gives some guidelines for choosing tube vs. solid-state RF power amplifiers. However, these are not hard limits and need to be re-evaluated on the specific case!

The double-logarithmic graph of Fig. II.2.152 illustrates power output capability vs. operating frequency for different RF power sources, including some not discussed in this paragraph (IOTs, CCTWTs). Please note, the blue-dashed trace for solid-state modules is enhanced 32×, meaning for a combination of 32 solid-state based RF power modules.

II.2.8.5 Summary of RF components, devices and subsystems

The topics covered in this section are neither complete nor comprehensive. From the almost infinite variety of RF components, devices and subsystems only a few could be introduced and briefly discussed. The selection made was based on a personal, subjective choice, with some focus on transmission-line based RF components. However, the topics covered here do play a major role in accelerator RF and related fields. Even as digital signal processing, see Fig. II.2.97, increasingly takes over the classical domain of low-level RF signal processing and manipulation, traditional RF components remain relevant, important and continue to be required, e.g. for high-power RF applications, for the signal conditioning of very low-level RF signals, for the processing of RF signals which cover an instantaneous, high dynamic range, and for many other accelerator RF applications.

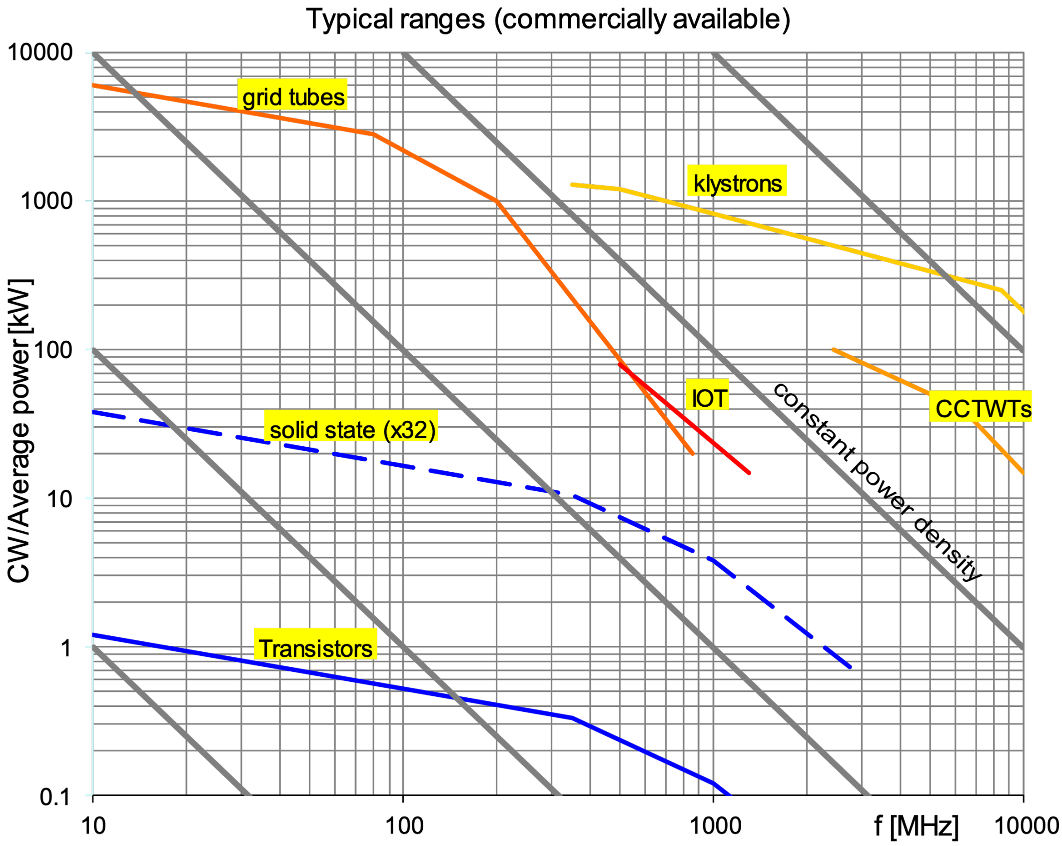


Fig. II.2.152: Power capability of commercially available RF amplifier type (as of 2020).

II.2.9 RF measurement techniques

The design, engineering and realisation of most technical hardware systems or scientific apparatuses follow some systematic procedures, e.g. to divide large, complex structures in smaller pieces, and to study and analyse the individual building blocks, first by a simplified formalism using analytical equations, and then followed by improved numerical approximations, analysis and optimisation. In the final design stage, a prototype is manufactured and it's performance, characteristics and limitations are tested with the help of *measurements*. This is also true, but not limited, to accelerator RF systems. The measurement characterisation is the ultimate test and verification if the design goals and specifications of a RF apparatus, component or subsystem are met.

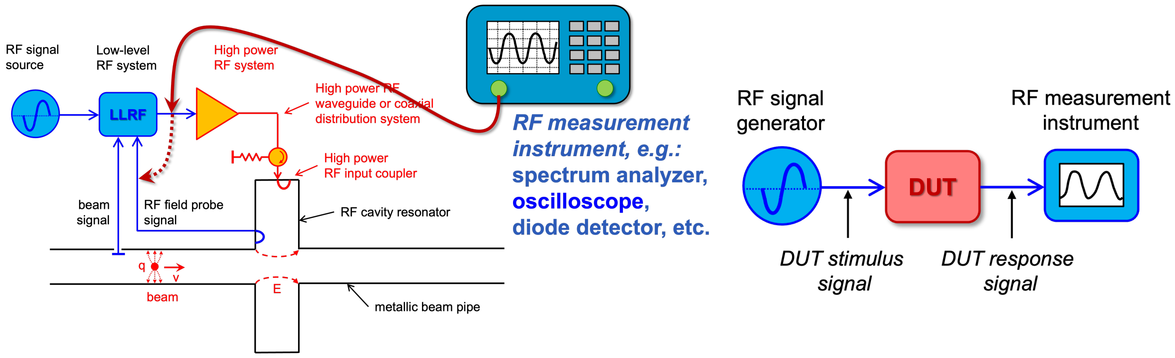
> Performing reproducible and precise RF measurements is an art! It requires many repeats of a variety of measurement exercises and experiments, and over the years sufficient skills and experience maybe collected to correctly setup a RF measurement and get meaningful results. This cannot be trained by just reading this section, or by watching online videos, but requires practical exercises, as given, e.g. in the RF measurement training part during the JUAS Practical Days, or at the Advanced CERN Accelerator School (CAS) course, or at a dedicated training course at the US Particle Accelerator School (USPAS).

II.2.9.1 Overview

RF measurements are performed using dedicated, often highly specialised instruments, usually provided by the industry of RF measurement equipment. Still, in most cases RF measurements are a subset of electrical measurements, our RF instruments are basically electrical measurement equipment and therefore we have to follow the underlying electrical principles and procedures to successfully execute our RF measurements.

We can separate two basic forms of RF measurements, usually performed in any accelerator institute or laboratory, see also Fig. II.2.153

– The RF measurements are performed in situ, i.e. *in the field*, at the running machine and / or the active RF system (Fig. II.2.153a). We measure RF signals somewhere within the RF system or



(a) In situ at the accelerator RF system. **(b)** On the test bench.

Fig. II.2.153: RF measurements.

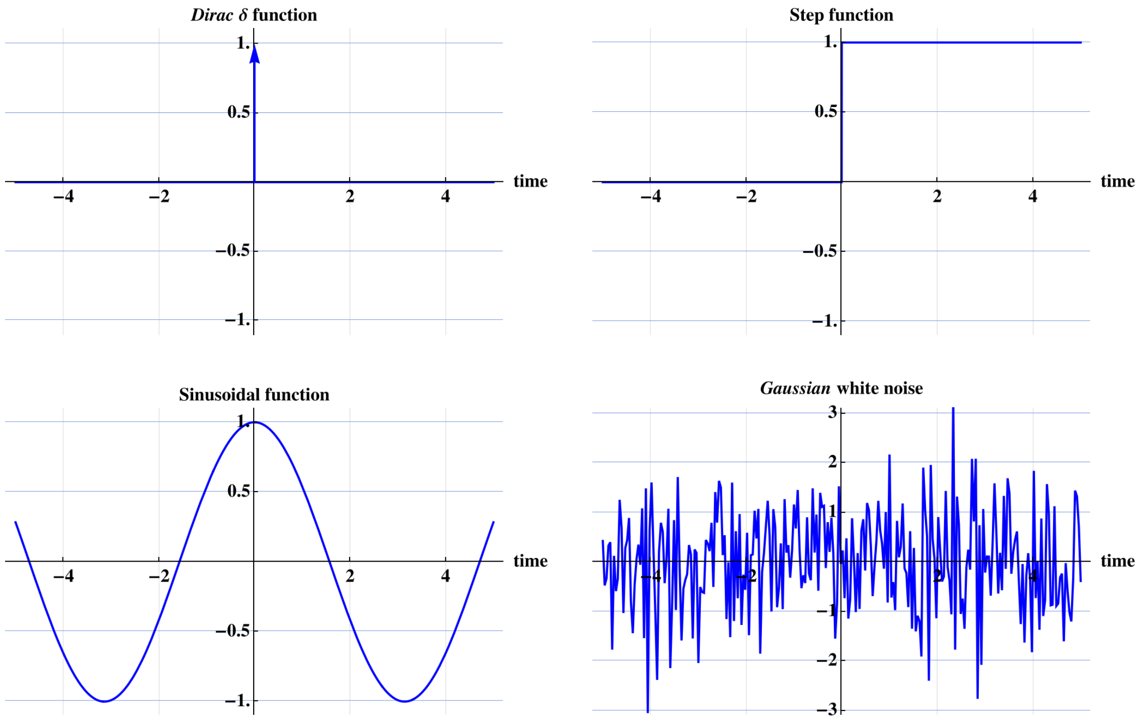


Fig. II.2.154: Popular test functions used as stimulus signal.

follow more closely the beam induced signals from antennas in the accelerating structures or from beam pickups. This requires the RF instrument being able to detect and analyse those signals, like the spectrum analyser, the oscilloscope, or the diode detector.

– We test and characterise a RF component or subsystem in the laboratory room, i.e. on a *test bench* (Fig. II.2.153b). Unless the RF *Device Under Test* (DUT) is a RF source generator, this requires a *RF generator* to supply a *stimulus signal*, and again an RF measurement instrument to detect and analyse the signal response from the DUT.

For the characterisation of RF and electrical systems we use dedicated *test functions* as stimulus signal to execute our DUT measurements in the laboratory, see Fig. II.2.154. For most RF measurements the $\cos$ or $\sin$ functions are the preferred test signals.

Moreover, most RF components and subsystems to be analysed and tested are characterised by their S-parameters, see Section II.2.7, and the measurement is performed by a *vector network analyser* (VNA), as illustrated in Fig. II.2.155. As explained later in this section, the VNA combines the RF measurement instrument and the RF generator for the sinusoidal stimulus signal, as shown in Fig. II.2.153b, into a single instrument. The “standard” VNA has two ports, but VNAs with four or many more ports are also available to simultaneously characterise multi-port DUTs. The inverse *Fourier* transform option of the VNA enables the mathematically equivalent of a *Dirac* δ or step stimulus function, to analyse the time-domain impulse or step response of linear DUTs, of course with some limitations.

Here a brief overview of the most popular instruments used for measurements on RF systems and components:

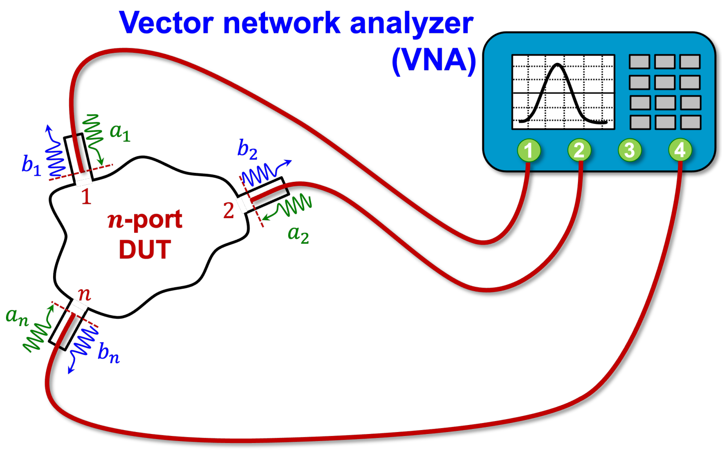


Fig. II.2.155: S-parameter measurement with a vector network analyser (VNA).

Oscilloscope The *oscilloscope* displays the waveform of a voltage signal as function of time, i.e. it allows to observe and analyse one or more signals as *time-domain* (TD) waveform. The signal to be displayed can be of periodic nature, or a transient or a burst signal with an arbitrary waveform. The oscilloscope can be used in the field, e.g. to analyse beam-related signals from a beam pickup or cavity antenna, but also on the test bench, in connection with a pulse or arbitrary waveform generator. Most oscilloscopes have two, four or more inputs, enabling the display of different signals simultaneously, and also a external trigger input to synchronise the displayed signal to other signal sources.

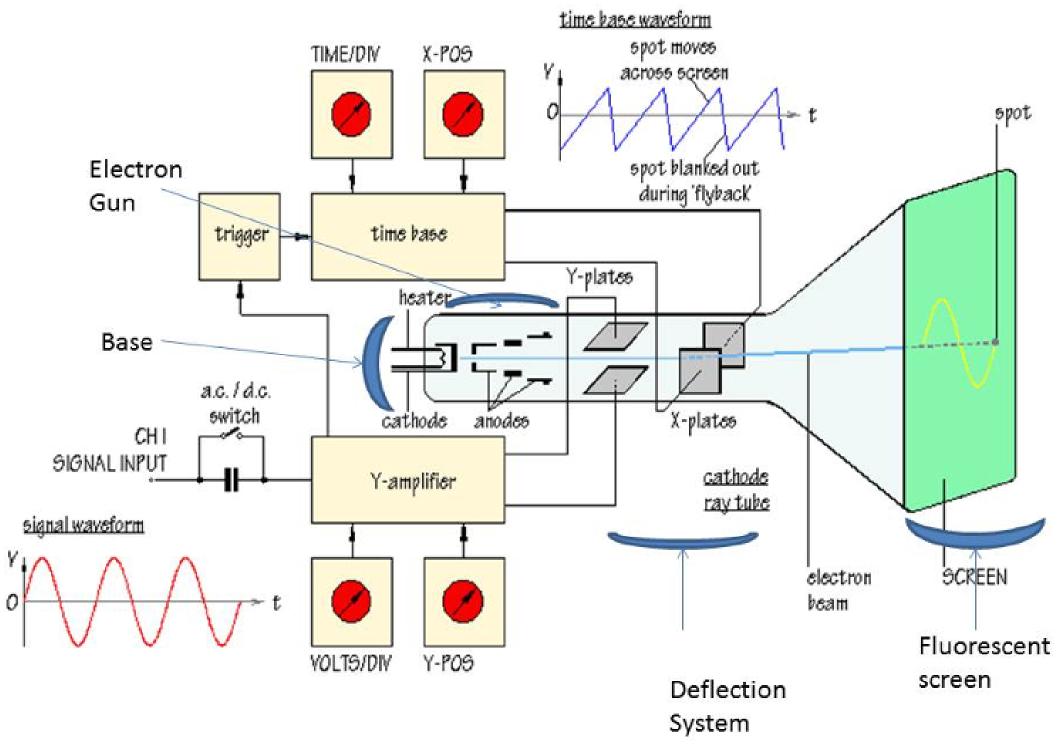


Fig. II.2.156: Schematic of a traditional CRT-based oscilloscope.

Today, the traditional form of the oscilloscope based on a cathode-ray tube (CRT) exists only in the museum, see the schematic Fig. II.2.156 for "historic" reasons. All "modern" oscilloscopes are based on analog-to-digital converters (ADC), and most of the internal signal processing is performed in the digital domain. Still, look and feel of many oscilloscopes of today resemble the traditional analog CRT oscilloscope.

While the oscilloscope is a very versatile instrument, able to display any signal directly as time-domain waveform, thus in an intuitive format, it's use in RF engineering has some limitations. This is mainly because of the limited dynamic range, typically 40 to $60\,\mathrm{dB}$, and the upper frequency limit. The dynamic range of oscilloscopes operating to high frequencies, e.g. $> 40\,\mathrm{GHz}$, have an even lower dynamic range, while being costly, which substantially limits their use, e.g. to evaluate non-linear effects in RF systems.

RF diode detector / power meter A zero-bias *Schottky* diode is often used as *RF diode detector* or *RF power meter*. It provides the rectified *video* output signal which is proportional to the power of the RF signal at the input, thus, provides the information of the power of an high frequency RF signal. The RF power meter is a passive, simple and robust instrument, and operates over a wide range of frequencies, few MHz to many GHz, and has a large dynamic range. Of course, the phase information of the RF input signal is lost by the detection, still this very simple RF power meter is very useful, e.g. for the monitoring of the forward / reflected power in the feed-line of an accelerating structure.

Vector signal / spectrum analyser The *spectrum analyser* (SA) is used to display a RF signal in a "*frequency-domain* (FD) like" fashion. It utilises a narrow-band receiver to sweep across a given range of frequencies to detect and display the RF signal power as function of the frequency.

> This requires the input signal to be *time invariant* during the SA sweep period!

Unlike the oscilloscope, the spectrum analyser has a very large dynamic range, $140\,\mathrm{dB}$ and more, and can be used in the field to observe and analyse the signals in the RF system, the signals from beam pickups, or to located unwanted EM fields or EMI signals in an accelerator using an appropriate antenna. On the laboratory test bench the spectrum analyser is used, e.g. for noise measurements (requires a noise source) and for intermodulation measurements (TOI) of amplifiers (requires two RR signal generators).

The traditional analog spectrum analyser today is replaced by a *vector signal / spectrum analyser* (VSA), sometimes called *FFT analyzer*, which comes in a hybrid format, with an analog RF up / down converter front-end, followed by a digital signal processing section. These blocks of the VSA combine all the functions of the traditional analog spectrum analyser with some of a digital oscilloscope, e.g. using a digital signal processor (DSP) to perform a fast-*Fourier* transform (FFT) to analyse the down-converted input signal. While the traditional spectrum analyser requires the input signal to be periodic, the VSA can also analyse non-periodic and transient signals, and can provide phase information. Typical applications of the SA / VSA are the observation of tune sideband signals, bunched-beam spectrum measurements, or the transient behaviour of phase-locked loops in the low-level RF (LLRF) system.

- The digital oscilloscope directly digitises the input signal using a fast sampling, high bandwidth ADC. It has a limited dynamic range and a very wide bandwidth, up to many GHz.
- The input signal of the VSA is *down-converted* to an intermediate frequency (IF) band, which then is digitised by an ADC with high resolution. The VSA has a large dynamic range, the bandwidth is given by the IF-band and the performance of the ADC, and is typically in the $100\,\mathrm{MHz}$ regime.

Slotted coaxial (or waveguide) transmission-line The *slotted line* was used in the early days of RF engineering to measure the VSWR, and therefore to evaluate an arbitrary, complex load impedance Z_L terminating the slotted transmission-line of known characteristic impedance Z_0. Today this measurement is performed with a VNA, however, for illustration of standing waves along a transmission-line an exercise with the slotted line still gives a very valuable introduction into RF measurements.

Vector network analyser The *vector network analyser* (VNA) combines the functions of a vector signal / spectrum analyser, a RF source generator and a S-parameter test set, i.e. a set of broadband directional couplers. It excites a port of a device-under-test (DUT), a RF component or subsystem (e.g. RF network, filter, antenna, amplifier, etc.), with a steady-state, sinusoidal waveform of a given frequency (at a constant amplitude), and measures the DUT response in magnitude and phase, thus, determining the S-parameters.

Similar to the spectrum analyser, the VNA measurement covers a selectable frequency range, and the instrument measures the S-parameters in predefined frequency steps. Again, the DUT has to be **time-invariant** during the frequency sweep.

The VNA allows the S-parameter characterisation of passive and active RF components as function of the frequency, and by applying an inverse *Fourier* transformation, also enables time-domain response measurements and *time-domain reflectometry* (TDR). Beside the frequency sweep, the VNA also has a power sweep function, which allows to characterise the $1\,\mathrm{dB}$ compression point of RF amplifiers. Four-port VNAs have additional functions, e.g. they allow to combine two physical ports into a virtual port, and therefore a mix of single-ended and differential port DUT characterisation.

The value of a modern VNA in the RF laboratory cannot be overstated! It de-facto is the most versatile and comprehensive RF measurement tool!

Today, basically all RF measurement instruments are equipped with coaxial input ports, with a characteristic impedance of $50\,\Omega$. To ensure a mode-free TEM operation to $100\,\mathrm{GHz}$ and beyond, related high performance coaxial test cables and connectors are utilised. Waveguide measurements are done via appropriate WG-to-coaxial adapters.

II.2.9.2 RF signal modulation

The concept of signal *modulation* is linked to RF engineering, as well as to RF measurement techniques, since the early days of wireless telecommunication. Among the many modulation methods we briefly discuss the two most common modulation methods:

II.2.9.2.1 Amplitude modulation

In electronics, telecommunications, and mechanics, modulation means varying some aspect of a sinusoidal, continuous wave (`CW`) *carrier signal* with an information-bearing waveform, to then transmit it in the radio frequency regime.

Why do we modulate signals? There are at least two reasons:

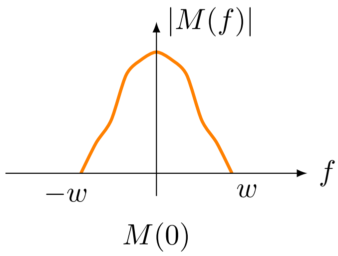
To allow the simultaneous transmission of two or more information, i.e. message signals, by translating them to different frequencies, and to take advantage of a wireless transmission on a RF carrier frequency, utilising the greater efficiency and smaller size of antennas operating at high frequencies.

II.2.9.2.1.1 Theoretical background

In *amplitude modulation* (`AM`), the amplitude of the wave carrier, which has a much higher frequency than the message signal, is varied proportionally to the message signal. Although the message signal is the one we want to transmit, what actually is transmitted is the carrier signal, which “carries” the desired message signal. This simply means that the message signal is imprinted on the envelope of the carrier signal. This carrier signal is called the *modulated signal*, while the information carrying signal, is referred to as the *modulating signal*. From an accelerator physics point of view, amplitude modulation resembles to *betatron* oscillations.

Let $m(t)$ be the modulating signal, a *baseband* signal with a total bandwidth w and a spectrum that satisfies:

$$|M(f)| = 0, \quad |f| \geq w$$



and let the modulated (carrier) signal be:

$$c(t) = A_c \cos 2\pi f_c t,$$

where A_c is the amplitude and f_c is the frequency of the carrier signal.

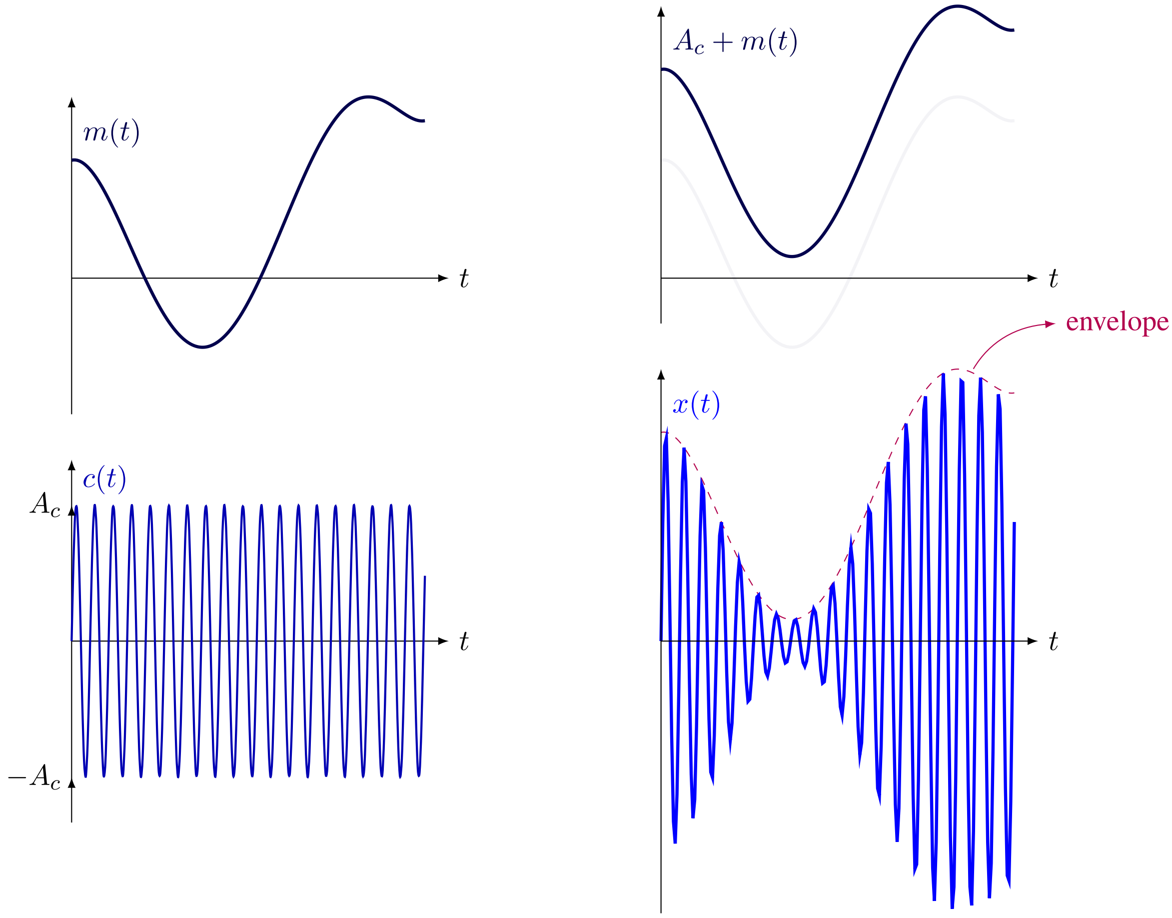


Fig. II.2.157: Carrier and baseband signals before and after amplitude modulation, following Eq. (II.2.142)

In this so-called *Double Side Band-Amplitude Modulation-Total Carrier* (`DSB-AM-TC`)[6], the amplitude of the modulated signal changes linearly with the amplitude of the modulating signal, while the addition of a strong carrier component facilitates demodulation. If f_m is the highest frequency components in the modulating signal m(t), the AM-modulated signal has the form (see also Fig. II.2.157):

$$x(t) = [A_c + m(t)] \cos 2\pi f_c t, \quad \text{with} \quad f_c \gg f_m \tag{II.2.142}$$

The modulation index μ is an important parameter, and is defined as ratio between the absolute minimum value of the baseband information signal (assuming that it does not contain a DC component) and the peak value of the carrier signal:

$$\mu = \frac{|min(m(t))|}{A_c}$$

[6]DSB-AM-TC is the standard AM method that produces *sidebands* on each side of the carrier frequency. Single-sideband modulation uses bandpass filters to eliminate one of the sidebands and possibly the carrier signal, which improves the ratio of message power to total transmission power, reduces power handling requirements of line repeaters, and permits better bandwidth utilisation of the transmission medium. Other forms of amplitude modulations can be:
Vestigial Side Band (`VSB-AM`), *Double Side Band-Amplitude Modulation-Suppressed Carrier* (`DSB-AM-SC`), *Single Side Band-Amplitude Modulation* (`SSB-AM`), or variations of these. Here, we focus only on the DSB-AM-TC, for simplicity.

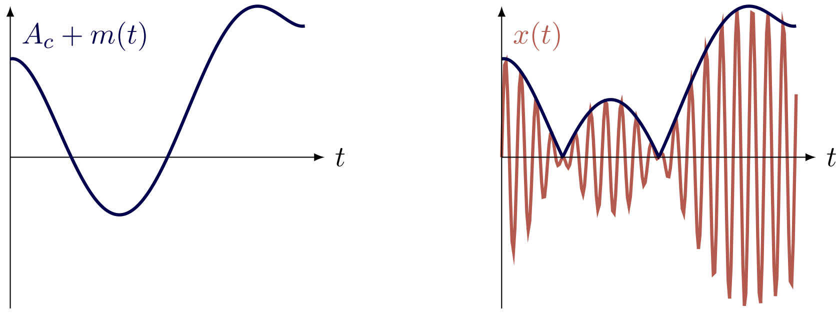


Fig. II.2.158: The effect of "overmodulation", $\mu \geq 1$, in the modulated carrier signal $x(t)$.

If $A_c + m(t) < 0$, then *overmodulation* occurs. This phenomenon leads to a distortion of the carrier envelope, which does follow any more according to the information signal and therefore does not allow a correct demodulation in the receiver, in case an asynchronous demodulation method is applied.

Therefore simply follows: **Avoid overmodulation!**
$\mu < 1$ must hold and in the following $\mu < 1$ is always assumed.

In this case, asynchronous demodulation can be applied, which is considerably easier to implement by using a simple envelope detector.
The effect of overmodulation in the modulated carrier signal is shown in Fig. II.2.158. A standard asynchronous demodulator would follow the wrong indicated envelope $\neq m(t)$. It would require a more complicated synchronous demodulator to still detect $m(t)$ correctly.

Note: An equivalent formula for the modulation index can be deduced:

$$\frac{x_{\text{max}}}{x_{\text{min}}} = \frac{A_c(1+\mu)}{A_c(1-\mu)} \quad \Longrightarrow \quad \mu = \frac{x_{\text{max}} - x_{\text{min}}}{x_{\text{max}} + x_{\text{min}}}$$

where $x = x(t) = [A_c + m(t)] \cos 2\pi f_c t$ is the AM-modulated signal.

In the frequency domain representation, the amplitude modulation produces a signal with the power concentrated at the carrier frequency and two adjacent sidebands. Each sideband is equal in bandwidth to that of the modulating signal, and is a mirror image of the other.

Finally, at the receiving station, the message signal is extracted from the modulated carrier by a process called *demodulation*.

To calculate the frequency spectrum, in its general form, we apply the *Fourier* transform to II.2.142:

$$\begin{aligned} X(f) = \mathcal{F}[x(t)] &= \mathcal{F}[(A_c + m(t)) \cos 2\pi f_c t] \\ &= \mathcal{F}[m(t) \cos 2\pi f_c t] + \mathcal{F}[A_c \cos 2\pi f_c t] \\ &= \frac{1}{2}[M(f - f_c) + M(f + f_c)] + \frac{1}{2} A_c[\delta(f - f_c) + \delta(f + f_c)] \end{aligned} \tag{II.2.143}$$

with $M(f) = \mathcal{F}[m(t)]$. Figure II.2.159 illustrates the frequency spectrum of the amplitude modulated carrier signal, shown is the modulus $|X(f)|$, i.e. the magnitude spectrum. The total bandwidth of each sideband is $2w$, which is related to negative frequencies of the baseband signal. The total bandwidth of the modulated signal is $2(f_c + w)$.

II.2.9.2.1.2 A practical AM measurement exercise

For the analysis of modulated signals it is very useful to observe the signal in a twofold manner:

- On the **oscilloscope**, as time-domain waveform.
- On a **spectrum analyser**, in terms of it's frequency spectrum.

The spectrum analyser displays a signal, equivalent to the *Fourier* transform of successive segments of the signal. Any changes of the baseband message signal or the modulation index will also be reflected in the spectral content and displayed by the spectrum analyser.

In the following we present a test bench measurement example, analysing the amplitude modulation with a simple sine-wave signal. Below we refer to the following RF measurement equipment, however, similar instruments can serve as well:

1. A RF signal generator, `Agilent technologies, E8257D`, to generate the desired signals, and in our case it also serves to internally modulate the carrier CW signal.
2. An oscilloscope, `LeCroy, WaveRunner`, to observe the signals as time domain waveform.
3. A vector network analyser (VNA), utilising the build-in spectrum analyser functionality, `Keysight, P5024A`, to analyse the signal by its *Fourier* components, and to observe the spectral content of the signal.
4. A power divider, `Suhner & Suhner, 4901.01.B`, to split the signal from the RF generator to simultaneously observe the signal on the oscilloscope and the spectrum analyser.
5. Various cables and connectors, wherever needed. Preferably, use coaxial cables with BNC connectors. Minimum required: 2 cables.

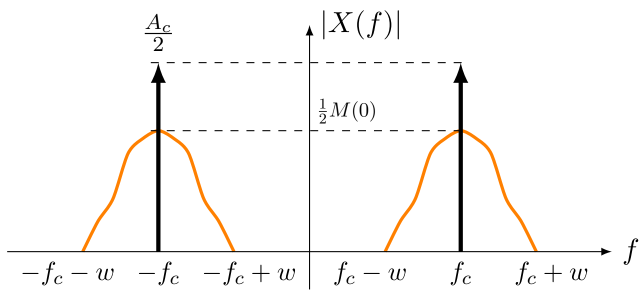


Fig. II.2.159: Frequency spectrum (magnitude) of the amplitude modulated signal, following Eq. (II.2.143)

Lab Procedure II.2.9.1: AM Measurement setup

Consider a carrier signal of $f_c = 100\,\mathrm{MHz}$ and an amplitude of $6\,\mathrm{dBm}$, and a modulating signal of $f_m = 10\,\mathrm{kHz}$. Set these frequency and amplitude values of the modulated signal on the RF signal generator, and press the `AM button`.

1. Set the `AM depth softkey` to $10\,\%$, the `AM Rate softkey` to $10\,\mathrm{kHz}$ and the `AM Waveform softkey` to `Sine`. Make sure that the `AM softkey` is `On`. Leave all other softkeys as default (unchanged). Make sure that you press `MOD ON` and `RF ON`.
2. Set the correct timebase on the oscilloscope, in order to be able to see the modulating signal properly, e.g. $10\,\mu$s/division should be fine.
3. Set the VNA to spectrum analyser mode. Go to `Response` $>$ `Meas` $>$ `Meas Class` and set it to spectrum analyser. Then select `A`. Set the center frequency to $1\,\mathrm{MHz}$ and the span approximately 3 times the modulating frequency ($10\,\mathrm{kHz}$ for now, but later you should change it accordingly). Set the `RBW` to $500\,\mathrm{Hz}$ or so, to ensure a reasonable resolution.

Let's now try to change the modulation index, given as percent value (i.e. the amount of carrier amplitude), so we can observe the effects in both domains.

Lab Procedure II.2.9.2: AM Measurement parameters

Vary `AM Depth` in a range $25\,\%$ to $90\,\%$ and observe the differences between $25\,\%$, $50\,\%$ and $90\,\%$.

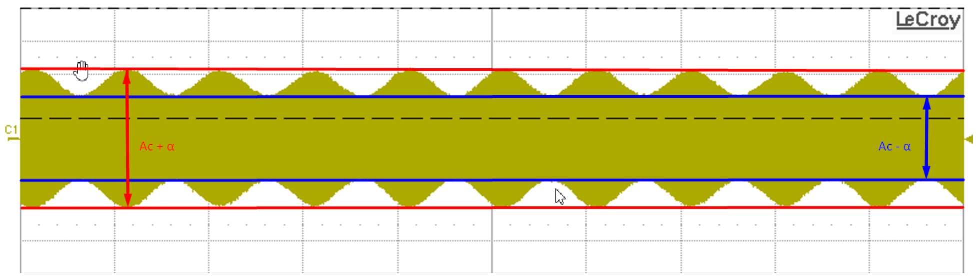


Fig. II.2.160: $25\,\%$ modulation depth of a carrier with a sine wave as modulating signal for $f_m = 10\,\mathrm{kHz}$ and $f_c = 100\,\mathrm{MHz}$, observed as time domain waveform on the oscilloscope.

The measurement results from the oscilloscope, Figs. II.2.160 to II.2.162, indicate, as the modulation depth increases, the message signal becomes better visible as envelope of the carrier signal in the time-domain view on the oscilloscope. On the spectrum analyser, observing the signals in a "frequency-domain like" fashion on a logarithmic dBm scale for the magnitude, the two sidebands are clearly visible for all settings of the modulation depth, see Fig. II.2.163.

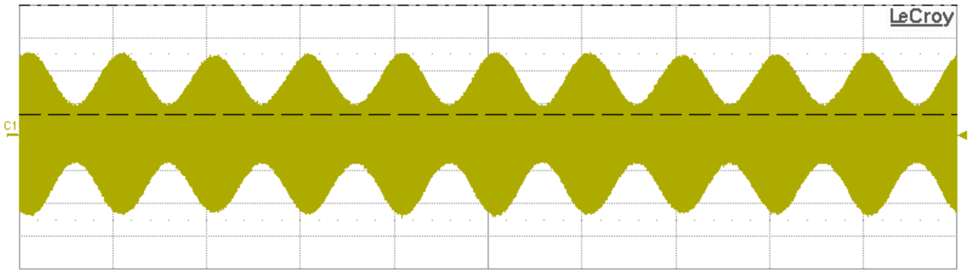


Fig. II.2.161: $50\,\%$ modulation depth of a carrier with a sine wave as modulating signal for $f_m = 10\,\text{kHz}$ and $f_c = 100\,\text{MHz}$, observed as time domain waveform on the oscilloscope.

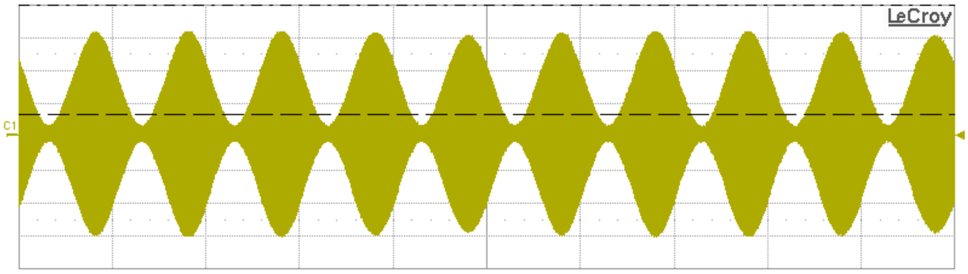
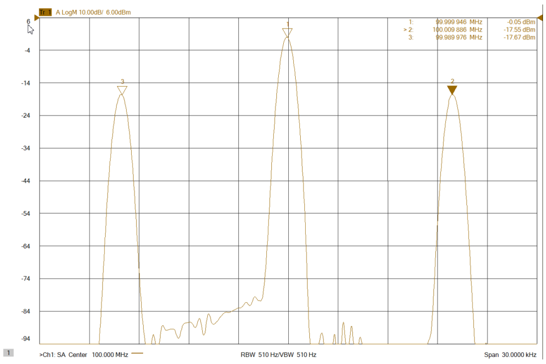
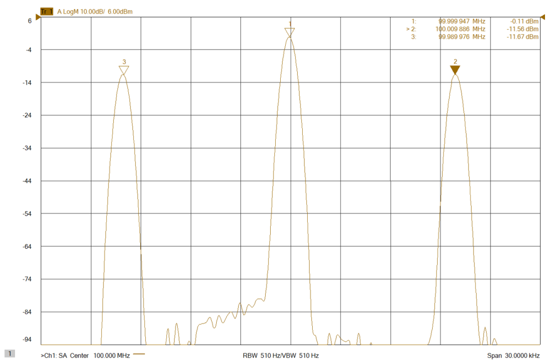


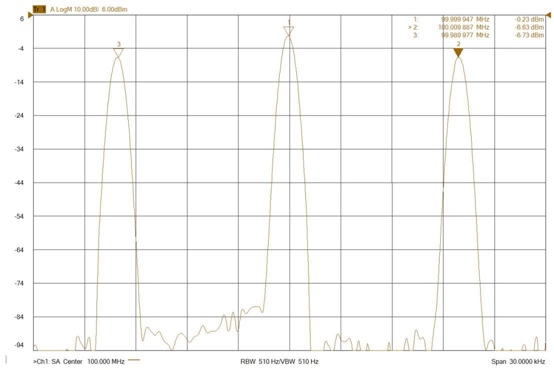
Fig. II.2.162: $90\,\%$ modulation depth of a carrier with a sine wave as modulating signal for $f_m = 10\,\text{kHz}$ and $f_c = 100\,\text{MHz}$, observed as time domain waveform on the oscilloscope.



(a) $25\,\%$ modulation depth.



(b) $50\,\%$ modulation depth.



(c) $90\,\%$ modulation depth.

Fig. II.2.163: AM Modulation of a CW carrier with a sine wave as modulating signal for $f_m = 10\,\text{kHz}$ and $f_c = 100\,\text{MHz}$, observed as frequency spectrum with a spectrum analyser.

Exercise II.2.9.1: Arithmetic exercise for AM with a sine wave as modulating signal

To evaluate the amplitude values of the carrier and sideband signals, set the `AM Depth` to 25%. Set a marker to f_c, another marker to $f_c + f_m$ and a 3rd marker to $f_c - f_m$, as shown in Fig. II.2.163. Consider the modulation depth as $\mu \cdot 100\%$, with a sine wave as modulating signal.
If $m(t) = \alpha \cos(2\pi f_m t)$, the modulating signal becomes:

$$\begin{aligned} x(t) &= [A_c + \alpha \cos(2\pi f_m t)] \cos(2\pi f_c t) \\ &= A_c \cos(2\pi f_c t) + \frac{\alpha}{2} \cos(2\pi(f_c - f_m)t) + \frac{\alpha}{2} \cos(2\pi(f_c + f_m)t) \end{aligned}$$

or, using the modulation index μ,

$$x(t) = A_c[\cos(2\pi f_c t) + \frac{\mu}{2} \cos(2\pi(f_c - f_m)t) + \frac{\mu}{2} \cos(2\pi(f_c + f_m)t) \tag{II.2.144}$$

For the frequency domain follows:

$$\begin{aligned} X(f) = \frac{A_c}{2}[\delta(f - f_c) + \delta(f + f_c)] &+ \frac{\alpha}{4}[\delta(f - f_c + f_m) + \delta(f + f_c + f_m)] \\ &+ \frac{\alpha}{4}[\delta(f - f_c - f_m) + \delta(f + f_c + f_m)] \\ = \frac{A_c}{2}\Big[\delta(f - f_c) + \delta(f + f_c) &+ \frac{\mu}{2}[\delta(f - f_c + f_m) + \delta(f + f_c + f_m)] \\ &+ \frac{\mu}{2}[\delta(f - f_c - f_m) + \delta(f + f_c + f_m)]\Big] \end{aligned} \tag{II.2.145}$$

Figure II.2.164 illustrates Eqs. (II.2.144) and (II.2.145).

To demonstrate the arithmetic example, lets take $\mu = 0.25$ or 25% as modulation depth, for the amplitude of the modulated signal of equivalent 6 dBm power. As of the 6 dB-power divider on the output of the RF signal generator, the signal power that reaches both, the oscilloscope and the spectrum analyser, will be $\approx$0 dBm, assuming the cables are sufficiently good and add no significant insertion losses.
Then, using the oscilloscope, we measure: $A_c + \alpha = 396\,\mathrm{mV}$ and $A_c + \alpha = 242.5\,\mathrm{mV}$, as indicated in Fig. II.2.160 by the coloured lines. From that follows $\alpha = 152.5\,\mathrm{mV}$, and we find: $A_{c_{RMS}} = 319.25\,\mathrm{mV}/\sqrt{2} = 225.74\,\mathrm{mV}$. Hence, $\mu = a/A_c = 0.24$, which is very close to the desired 25% modulation depth.

In a 50Ω system with sinusoidal waveforms:

$$V_{\text{dBm}} = 20\log_{10}(\frac{V}{V_{\text{ref}}})$$

with: $V_{\text{ref}} = \sqrt{P_{\text{ref}} Z_0} = 223.6\,\mathrm{mV}(RMS)$, $Z_0 = 50\,\Omega$.
Since $V = A_{c_{RMS}}$, we have $V_{\text{dBm}} = 20\log_{10}(\frac{225.74\,\mathrm{mV}}{223.6\,\mathrm{mV}}) \approx 0\,\mathrm{dBm}$.
Therefore, the amplitude of the delta-function related to the carrier is 0 dBm.

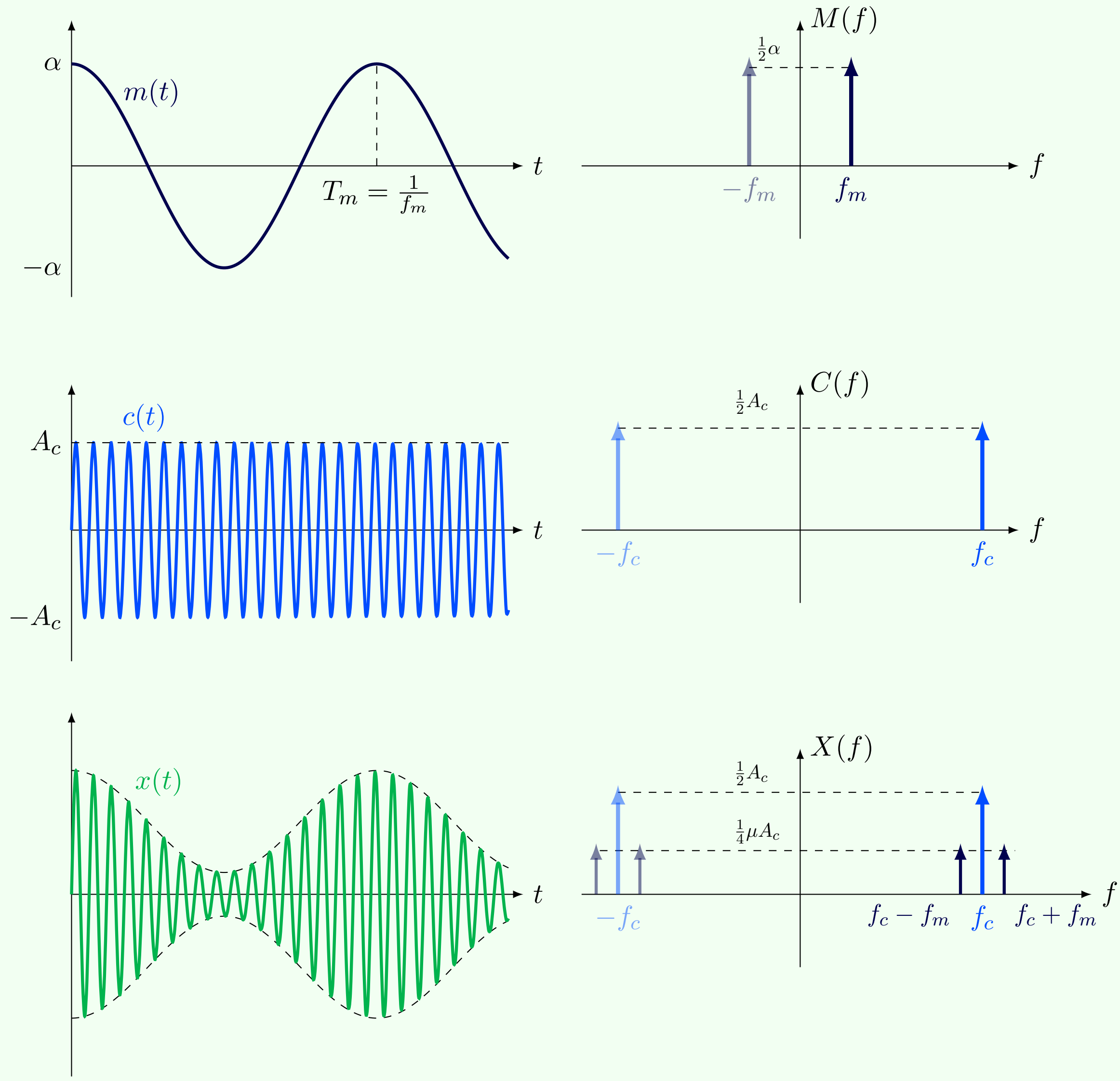


Fig. II.2.164: Time and frequency domain view of a sine-wave modulated carrier signal

Note: Given the *Fourier* transform of a cosine, we would expect the amplitude of the carrier to be $-3\,\mathrm{dBm}$. However, the spectrum analyser processes only "positive" frequencies, while the *Fourier* transform returns a symmetric result in form of negative and positive frequencies. Therefore, the power received by the spectrum analyser is not separated in positive and negative frequency bands, but is just detected, processed and displayed as a positive frequency band, Thus, the amplitude is twice the value of the *Fourier* transform, in our case equal to $0\,\mathrm{dBm}$.

For the amplitude of the sidebands, relatively to the carrier amplitude, we find:

$$\left(\frac{\alpha/4}{A_c/2}\right)_{\mathrm{dB}} = \left(\frac{\mu A_c/4}{A_c/2}\right)_{\mathrm{dB}} = -20\log_{10}\frac{\mu}{2} = -20\log_{10}\frac{0.24}{2} \approx -18\,\mathrm{dB}$$

which is confirmed by our measurement shown in Fig. II.2.163a.

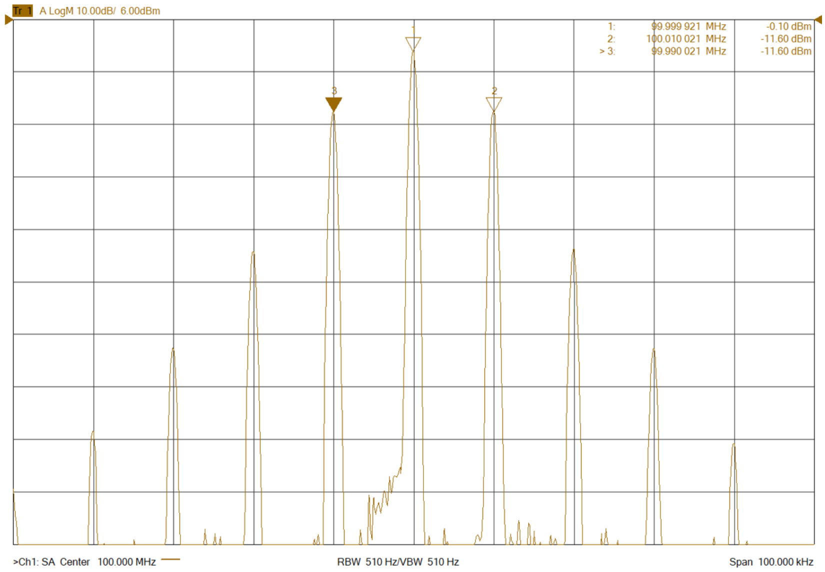


Fig. II.2.165: 50% modulation depth with a sine wave as modulating signal for $f_m = 10\,\mathrm{kHz}$ and $f_c = 100\,\mathrm{MHz}$, displayed over a larger frequency range.

Lab Procedure II.2.9.3: AM Measurement parameters

Set the `AM depth` to 50%. Set the frequency range of the spectrum analyser to $10 \times f_m$, keep the central frequency at f_c.

The large dynamic range, displayed on a logarithmic scale, enables for more detailed signal analyses, e.g. for modulation *distortions*.

With the broader frequency span setting, as given in Lab Procedure II.2.9.3, clearly many more sidebands are observable in the measured spectrum than the two given by the theory, see also Fig. II.2.165. We observe distortions in form of 2$^{\mathrm{nd}}$, 3$^{\mathrm{rd}}$ and 4$^{\mathrm{th}}$ harmonics in the spectrum, while those distortions stay unnoticed in the time-domain view of the oscilloscope (Fig. II.2.161). The second harmonic sidebands at $f_c \pm 2f_m$ are $40\,\mathrm{dB}$ below the carrier. However, the distortion is characterised relative to the primary sidebands, here a $28\,\mathrm{dB}$ difference between the primary and 2$^{\mathrm{nd}}$ harmonic sidebands represents a distortion of $4\,\%$.

Lab Procedure II.2.9.4: AM Measurement parameters

Now set the `AM depth` to $25\,\%$, set the `AM Rate` to $500\,\mathrm{kHz}$ and keep the `AM Waveform` to `Sine`. Make sure that the `AM` softkey is set to `On`.

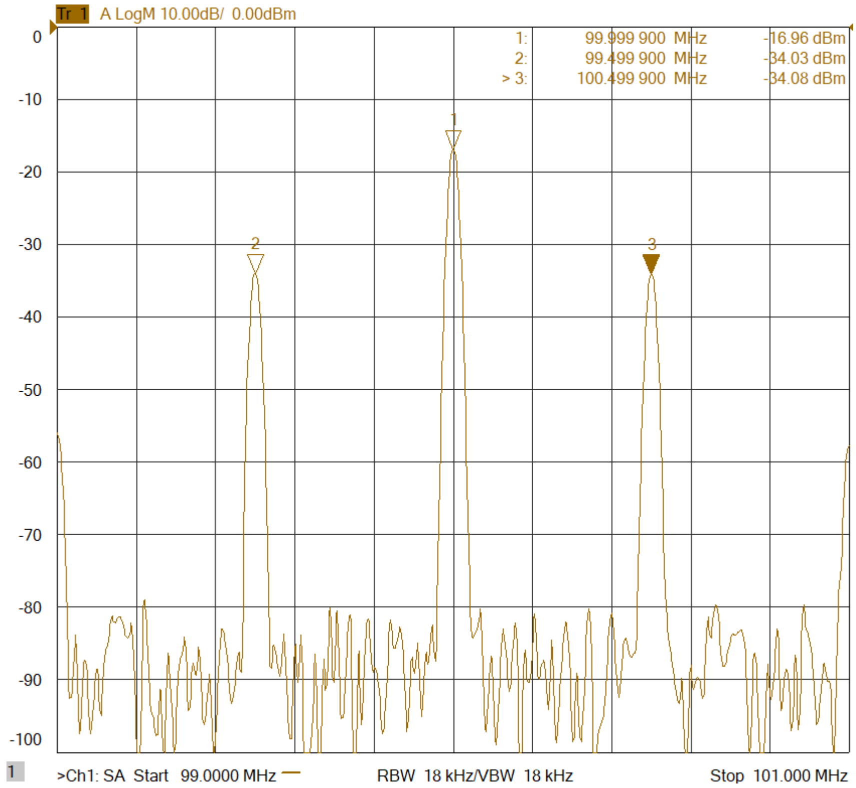


Fig. II.2.166: $25\,\%$ modulation depth of a carrier with a sine wave as modulating signal for $f_m = 500\,\mathrm{kHz}$ and $f_c = 100\,\mathrm{MHz}$, observed as frequency spectrum, applying a wider span on the spectrum analyser.

Changing the frequency f_m of the modulating signal, here a $50\times$ increase, changes the bandwidth of the modulated signal. Therefore, the frequency axis on the spectrum analyser (frequency span, see Fig. II.2.166) and the time axis on the oscilloscope need to be adjusted accordingly, demonstrating the inverse behaviour of time and frequency, $t = 1/f$.

> The sinusoidal waveform as message signal is the most simple case. Experiment with other waveforms for the message signal, e.g. triangular or rectangular signals. Try to compare the measured amplitude values of the sideband harmonics with the theory, following Eq. (II.2.143).

II.2.9.2.2 Frequency modulation

Frequency modulation (FM) and *phase modulation* (PM) are variants of the *angle modulation*. FM is more popular compared to PM, and therefore we here cover only FM, and only a few fundamental aspects.

II.2.9.2.2.1 Theoretical background

Angle modulation modifies the *frequency* or the *phase* of carrier signal, while keeping the amplitude fixed:

$$A_c \cos\left(\underbrace{2\pi f_c t + \vartheta}_{\theta(t)} \right)$$

As its name implies, the baseband information signal is imprinted on the angle $\theta(t)$ of the carrier signal.

This modulation can be *phase modulation* (PM) or *frequency modulation* (FM). Although it requires a more expensive and complex electronics, its advantage over, e.g. AM, is a better quality for the information to be transmitted, because it is not as much affected by noise and has a lower distortion.

Frequency modulation (FM) is most commonly used for radio and television broadcast, but also in telemetry, radar, seismic prospecting and monitoring systems, two-way radio systems, sound synthesis, magnetic tape-recording systems and some video-transmission systems. From an accelerator physics point of view, frequency modulation resembles to *synchrotron* oscillations.

FM signals can be generated either as *direct frequency modulation*, e.g. supplying the message signal directly into a voltage-controlled oscillator, or as *indirect frequency modulation*, i.e. by integrating the message signal to generate a phase-modulated signal, which then is used to modulate a crystal-controlled oscillator, followed by a frequency multiplier to produce the FM signal.

Consider the carrier signal:

$$x(t) = A_c \cos \underbrace{2\pi f_c t}_{\theta(t)} \tag{II.2.146}$$

with the frequency given as:

$$f_c = \frac{\theta(t)}{2\pi t} = \frac{\omega_c}{2\pi}$$

which is the number of cosine cycles per unit of time:

$$f_c = \frac{\left[\theta(t+dt) - \theta(t)\right]/2\pi}{dt}$$

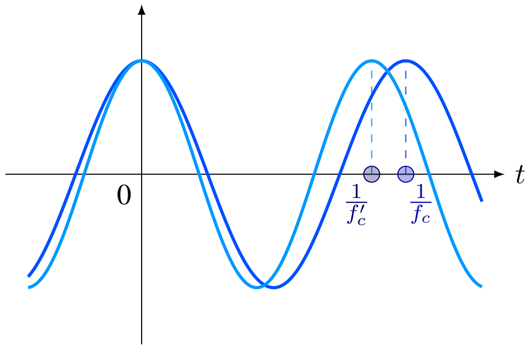
For the carrier signal $x(t) = A_c \cos\left[\theta_i(t)\right]$ at an *instantaneous phase* $\theta_i(t)$ we define the *mean frequency*:

$$f_{\Delta t}(t) = \frac{\theta_i(t+\Delta t) - \theta_i(t)}{2\pi \Delta t}$$

and at the limit $\Delta t \to 0$, we define the *instantaneous frequency*:

$$f_i(t) = \lim_{\Delta t \to 0} f_{\Delta t}(t) = \lim_{\Delta t \to 0} \frac{\theta_i(t+\Delta t) - \theta_i(t)}{2\pi \Delta t} \quad \Longrightarrow \quad f_i(t) = \frac{1}{2\pi} \frac{d\theta_{i(t)}}{dt} \tag{II.2.147}$$

Thus, we have a modulation signal with either constant (PM) or variable frequency (FM). Practically, a change in frequency means, the cosine function will not cross the horizontal time axis at equidistant points:



Definition II.2.9.1: Frequency modulation - FM

We define the **frequency modulation** as:

$$\boxed{f_i(t) = f_c + k_f m(t)} \tag{II.2.148}$$

where k_f is a constant, namely the *frequency sensitivity*, measured in Hz/V.

To calculate the signal, we use $f_i(t) = \frac{d\theta_i(t)}{2\pi dt}$:

$$\begin{aligned} \frac{1}{2\pi}\frac{d\theta_i(t)}{dt} &= f_c + k_f m(t) \qquad \Longrightarrow \\ \theta_i(t) &= 2\pi f_c t + 2\pi k_f \int_0^t m(\tau)d\tau \end{aligned}$$

Hence, our frequency modulated signal follows as:

$$\boxed{x(t) = A_c \cos\left[2\pi f_c t + 2\pi k_f \int_0^t m(\tau)d\tau\right]} \tag{II.2.149}$$

If we just consider a sinusoidal modulating message signal of frequency f_m:

$$m(t) = A_m \cos(2\pi f_m t)$$

the instantaneous frequency follows as:

$$\begin{aligned} f_i(t) &= f_c + k_f A_m \cos(2\pi f_m t) \\ &= f_c + \Delta f \cos(2\pi f_m t) \qquad \text{with } \Delta f = k_f A_m \end{aligned}$$

Definition II.2.9.2: Maximum frequency deviation

We define:

$$\Delta f = k_f A_m \tag{II.2.150}$$

as the **maximum frequency deviation**, which depends only on the amplitude of the modulating signal.

We recall the definition of the instantaneous frequency:

$$\frac{1}{2\pi}\frac{d\theta_i(t)}{dt} = f_i(t) \qquad \Longrightarrow \qquad \theta_i = 2\pi \int_0^t f_i(t)dt$$

Hence, for a sinusoidal signal:

$$\theta_i(t) = 2\pi f_c t + \frac{\Delta f}{f_m} \sin 2\pi f_m t$$

follows the modulated FM signal as:

$$x(t) = A_c \cos\left[2\pi f_c t + \frac{\Delta f}{f_m} \sin 2\pi f_m t\right]$$

Definition II.2.9.3: Modulation index β

We define the **modulation index β** in FM as:

$$\beta = \frac{\Delta f}{f_m} = \frac{A_m k_f}{f_m} \tag{II.2.151}$$

which expresses the maximum **frequency deviation** in the modulated carrier signal.

The final expression for a `FM` signal follows as:

$$x(t) = A_c \cos(2\pi f_c t + \beta \sin 2\pi f_m t) \tag{II.2.152}$$

Typical modulation indices range $1 \leq \beta \leq 20$. Keep in mind that β affects the required bandwidth of the modulated signal. For larger values of β the modulation is called *wideband* FM, while for smaller β, e.g. $\beta < 0.3$, we have *narrowband* FM `(NB-FM)`.

To evaluate the spectral content of the FM signal, we express Eq. (II.2.152) in its complex exponential form:

$$\begin{aligned} x(t) &= \mathrm{Re}\left[A_c \exp(j2\pi f_c t + j\beta \sin 2\pi f_m t)\right] \\ &= \mathrm{Re}\left[A_c \cdot \exp(j\beta \sin 2\pi f_m t) \cdot \exp(j2\pi f_c t)\right] \\ &= \mathrm{Re}\left[\tilde{x}(t) \exp(j2\pi f_c t)\right] \end{aligned}$$

with $\tilde{x}(t)$ given as the *complex envelope function*:

$$\tilde{x}(t) = A_c \exp\left[j\beta \sin(2\pi f_m t)\right]$$

The complex envelope function has a time period of $T_m = 1/f_m$ and can be rewritten in form of a *Fourier* expansion:

$$\tilde{x}(t) = \sum_{n=-\infty}^{\infty} c_n \exp(j2\pi n f_m t)$$

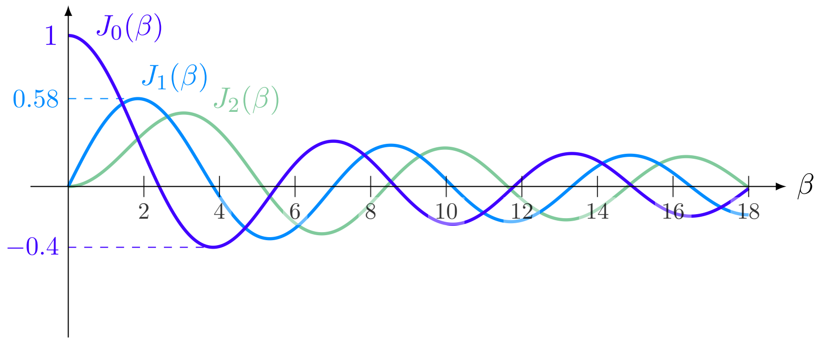


Fig. II.2.167: *Bessel* functions of first kind.

with the complex *Fourier* coefficient:

$$c_n = f_m \int_{-\frac{1}{2}T_m}^{\frac{1}{2}T_m} \tilde{x}(t) \cdot \exp(-j2\pi n f_m t) dt$$

$$c_n = f_m A_c \int_{-\frac{1}{2}T_m}^{\frac{1}{2}T_m} \exp\left[j\beta \sin(2\pi f_m t) - j2\pi n f_m t\right] dt$$

Substituting $x = 2\pi f_m t$ results in:

$$c_n = \frac{A_c}{2\pi} \int_{-\pi}^{\pi} \exp\left[j(\beta \sin x - nx)\right] dx \tag{II.2.153}$$

The above integral has no closed-form analytical solution, but follows the *Bessel* functions of first kind:

Definition II.2.9.4: Bessel functions of first kind

The Bessel function of the first kind has an argument β and is defined as follows:

$$J_n(\beta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp\left[j(\beta \sin x - nx)\right] dx \tag{II.2.154}$$

Numerical values of this function can be obtained from tables, diagrams, calculators or computer software, see also Fig. II.2.167.

Substituting Eq. (II.2.154) in Eq. (II.2.153) gives:

$$c_n = A_c J_n(\beta)$$

and for the complex envelope function follows:

$$\tilde{x}(t) = A_c \sum_{n=-\infty}^{\infty} J_n(\beta) \exp(j2\pi n f_m t)$$

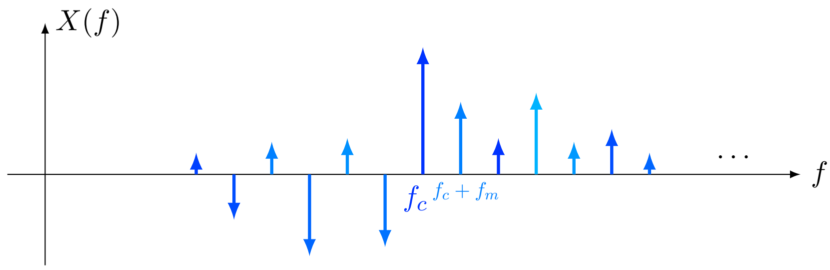


Fig. II.2.168: Spectral harmonics of a FM signal with a sinusoidal message signal.

Now we can express Eq. (II.2.152) in form of a *Fourier* series expansion:

$$x(t) = A_c \operatorname{Re}\left[\sum_{n=-\infty}^{\infty} J_n(\beta) \exp\left[j2\pi(f_c + nf_m t)\right]\right]$$

$$x(t) = A_c \sum_{n=-\infty}^{\infty} J_n(\beta) \cos\left[2\pi(f_c + nf_m)t\right] \tag{II.2.155}$$

with J_n being the *Bessel* function (of first kind).

From Eq. (II.2.155) follows the frequency domain equivalent:

$$X(f) = \frac{A_c}{2} \sum_{n=-\infty}^{\infty} J_n(\beta) \left[\delta(f - f_c - nf_m) + \delta(f + f_c + nf_m)\right] \tag{II.2.156}$$

as an infinite sum of sideband spectral harmonics $\pm nf_m$, see also Fig. II.2.168. Due to the *Bessel* function characteristic the amplitude of those harmonics decrease with increasing n, i.e. as we move away from the carrier frequency f_c. We also notice that the FM spectrum extends to infinity, which at a first look seems problematic.

Recalling Definition II.2.9.3 of the modulation index, we notice how a change in amplitude A_m and/or frequency f_m of the modulation (baseband) signal is reflected in the frequency spectrum. For example, if we vary β, while keeping the modulation signal frequency f_m constant, thus by varying the amplitude A_m, we notice the effect of the *Bessel* function on the total bandwidth of $X(f)$. More harmonics appear as we increase β by increasing A_m, see Fig. II.2.169. However, in practice we can limit the required bandwidth to a harmonic number $n_{\max}$ given by the frequency deviation Δf, which appears symmetric to the carrier frequency f_c. This is expressed by an empirical rule-of-thumb based on *Carson*'s rule for the transmitting bandwidth B_T:

$$\begin{aligned} B_T &\simeq 2\Delta f + 2f_m \\ &= 2\beta f_m + 2f_m \\ &= 2(\beta + 1)f_m \end{aligned} \tag{II.2.157}$$

Let's compare this bandwidth to the AM counterpart, which is $2w = B_T = 2f_m$ for a sinusoidal modulation signal. The FM signal, especially for high values of the modulation index, e.g. $\beta = 5$, requires a higher bandwidth, here $B_T = 12f_m$, and therefore is considered as a broadband signal.

In general, for a given bandwidth B_T, we have to ensure that $n_{\max}$ harmonic frequencies are covered, where:

$$B_T = 2n_{\max} f_m \tag{II.2.158}$$

and $n_{\max}$ or B_T are given following *Carson*'s rule.

A different way to evaluate the required bandwidth for the FM spectrum is based on the level of the harmonics, and define the bandwidth as threshold for $> 1\,\%$ of A_c. Alternatively, we could include frequencies until we reach $98\,\%$ of the strength of the transmitted signal.

The above analysis was performed for a purely sinusoidal waveform of the modulating signal, and cannot be simply used for a signal with arbitrary waveform of bandwidth w. Instead, we still may consider a sinusoidal signal of frequency f_s, and get correct results if we consider the most adverse case, i.e. the highest frequency content, see also Fig. II.2.170.

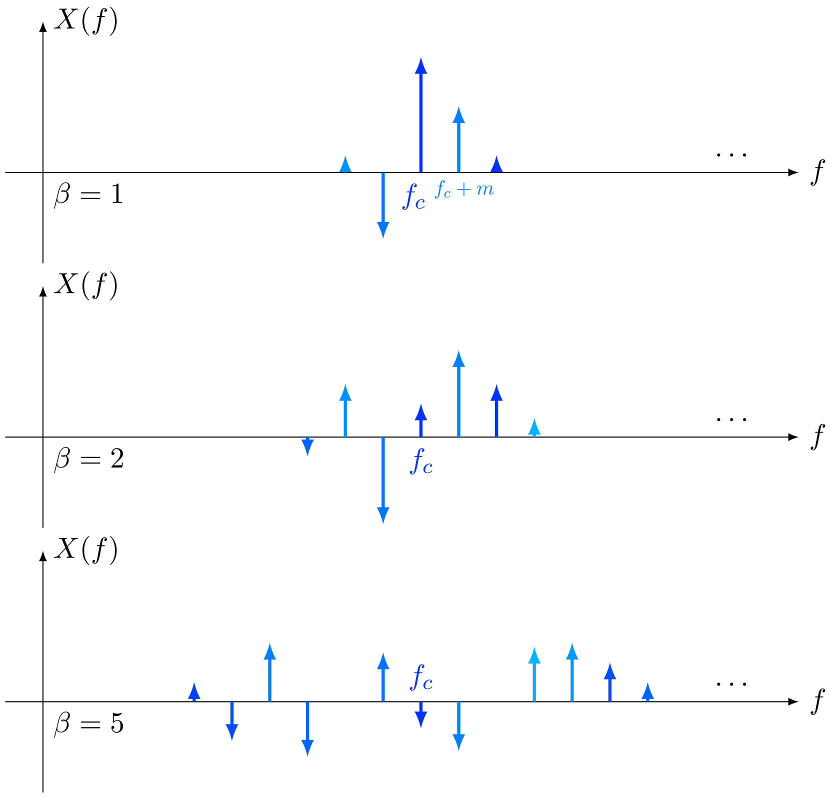


Fig. II.2.169: FM spectrum for different modulation indices of the sinusoidal modulation signal.

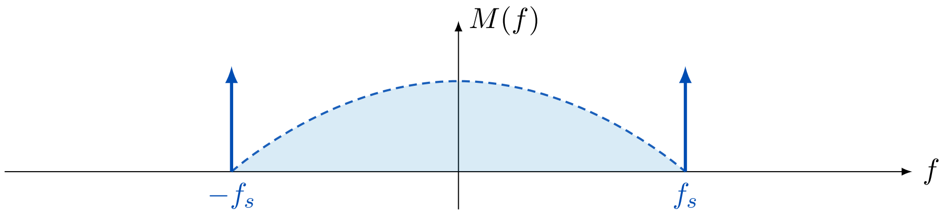


Fig. II.2.170: FM spectrum for arbitrary baseband signal waveforms.

Definition II.2.9.5: FM signal analysis for arbitrary waveforms of the modulation signal

For this analysis, we reassign the variable names as follows:

$$f_m \leftrightarrow f_s, \quad \beta \leftrightarrow D \quad \text{is the frequency deviation ratio}$$

which results in:

$$\boxed{D = \frac{\Delta f}{f_s}} \quad \underset{(Carson\text{'s rule})}{\boxed{B_T = 2(D+1)f_s}}$$

Note: A typical value in FM radio is $\Delta f = 75\,\mathrm{kHz}$.

II.2.9.2.2.2 FM measurements

Measurements exercises for FM are similar to the measurement example given for AM. Here we assume the same measurement equipment and basic setup, however, for better visualisation in this script we use different carrier and modulation frequencies compared to the AM measurement exercise.

Lab Procedure II.2.9.5: FM Measurement setup

Let's exercise with a carrier signal of $f_c = 10\,\mathrm{MHz}$, at an amplitude of $6\,\mathrm{dBm}$, and a modulating signal of $f_m = 1\,\mathrm{MHz}$. Set these frequency and amplitude values of the modulated signal on the RF signal generator, and press the `FM/ΦM` button.

1. Set the `FM Rate` softkey to $1\,\mathrm{MHz}$, the `FM Dev` softkey to $2\,\mathrm{MHz}$, and the `Waveform` softkey to `Sine`. Make sure that the `FM` softkey is `On`. Leave the all other softkeys as default. Make sure you press `MOD ON` and `RF ON`.

2. Set an appropriate values for the timebase on the oscilloscope to see the modulating signal properly, e.g. $1\,\mu$s/division should be fine.

3. Set the VNA to spectrum analyzer mode. Go to `Response` > `Meas` > `Meas Class` and set it to spectrum analyzer. Then select `A`. Set the center frequency tp $10\,\mathrm{MHz}$ and the span approximately $10\times$ the modulating frequency ($1\,\mathrm{MHz}$ for now, but later you should change it accordingly). Set the `RBW` to $500\,\mathrm{Hz}$ or similar, to ensure a reasonable resolution.

Now observe the FM-modulated signal on the oscilloscope, see Fig. II.2.171. It should look like the yellow waveform, which is the modulated signal. In red shown, the sinusoidal modulating signal.

Lab Procedure II.2.9.6: FM Measurement parameters

For the measurement of the spectral harmonics with the spectrum analyser, set the carrier frequency to $50\,\mathrm{MHz}$ and the `FM Deviation` to $50\,\mathrm{kHz}$. Set the modulating frequency f_m to the following values:

- $f_m = 500\,\mathrm{kHz} \implies \quad \beta = 0.1$
- $f_m = 50\,\mathrm{kHz} = \Delta f \implies \quad \beta = 1$
- $f_m = 10\,\mathrm{kHz} \implies \quad \beta = 5$
- $f_m = 5\,\mathrm{kHz} \implies \quad \beta = 10$
- $f_m = 3.33\,\mathrm{kHz} \implies \quad \beta = 15$
- Finally set f_m as low as possible, e.g. $< 1\,\mathrm{Hz}$, $\implies \quad \beta \to \infty$

Figure II.2.172 summarises the measured FM spectras for the parameters given in Lab Procedure II.2.9.6. For the narrowband FM case, $\beta = 0.1$, Fig. II.2.172a, the frequency span setting of the spectrum analyser, $150\,\mathrm{kHz}$, was too narrow, the FM sideband harmonics of $f_m = 500\,\mathrm{kHz}$ lie outside the visible range, and we see only the $f_c = 50\,\mathrm{Mhz}$ carrier signal. Therefore, an example for a "bad" measurement, the `SPAN` should be changed to e.g. $3\,\mathrm{MHz}$. All other FM frequency spectrum measurements shown in Fig. II.2.172 make sense, and illustrate the metrics of the underlying *Bessel* function behaviour.

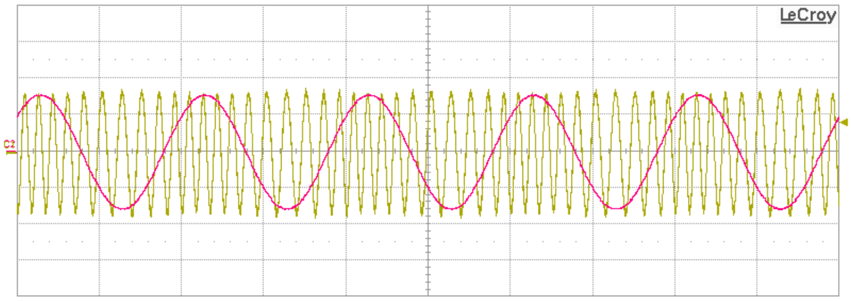
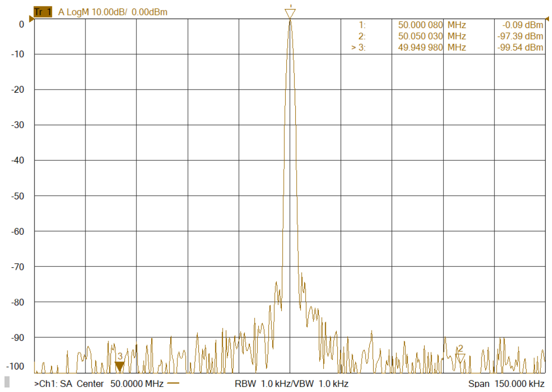
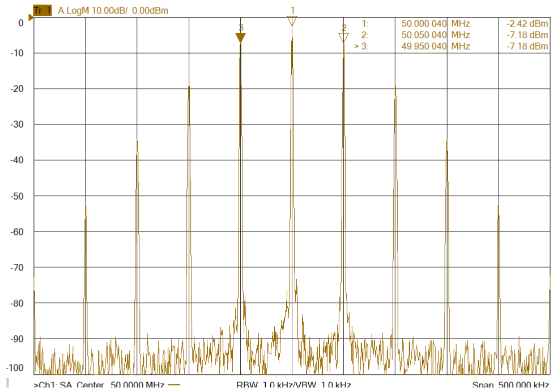


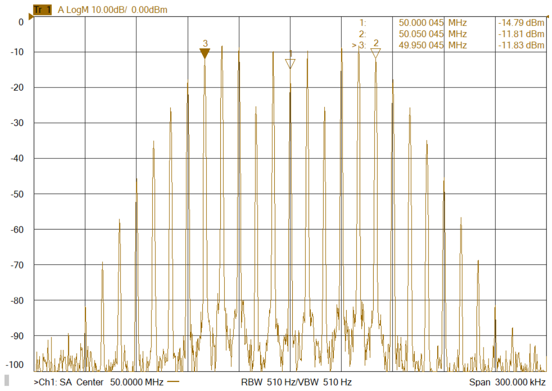
Fig. II.2.171: FM with a sinusoidal waveform as modulating signal at $f_m = 1\,\mathrm{MHz}$ and $f_c = 10\,\mathrm{MHz}$, with $\beta = 2$, observed as time domain waveform on the oscilloscope.



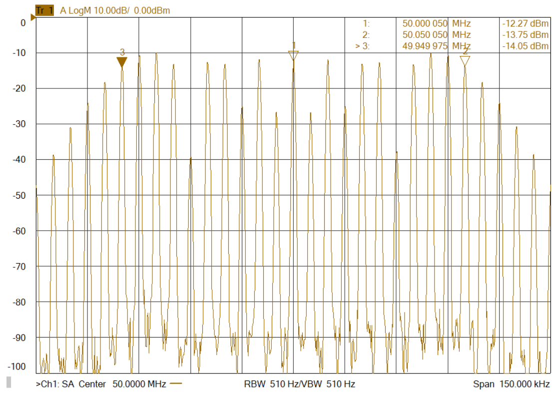
(a) $f_m = 500\,\text{kHz} \quad \Longrightarrow \quad \beta = 0.1.$



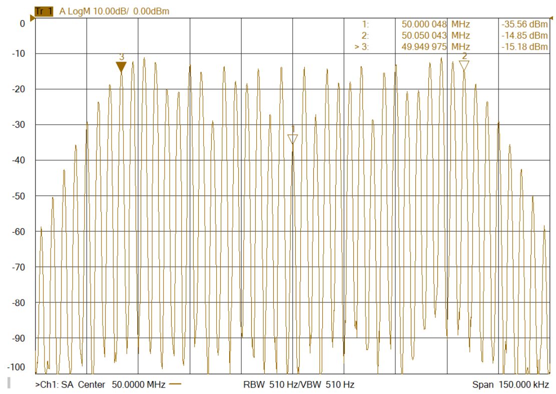
(b) $f_m = 50\,\text{kHz} = \Delta f \quad \Longrightarrow \quad \beta = 1.$



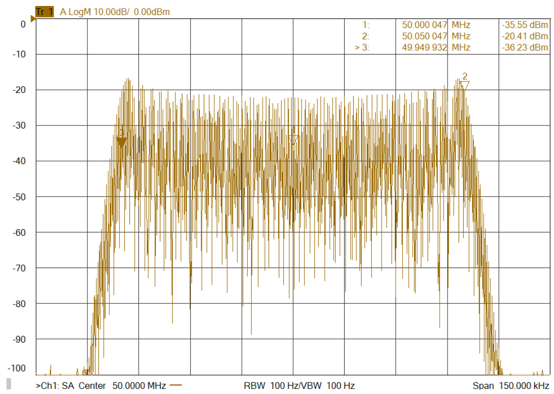
(c) $f_m = 10\,\text{kHz} \quad \Longrightarrow \quad \beta = 5.$



(d) $f_m = 5\,\text{kHz} \quad \Longrightarrow \quad \beta = 10.$



(e) $f_m = 3.33\,\text{kHz} \quad \Longrightarrow \quad \beta = 15.$



(f) $f_m \to 0 \quad \Longrightarrow \quad \beta \to \infty.$

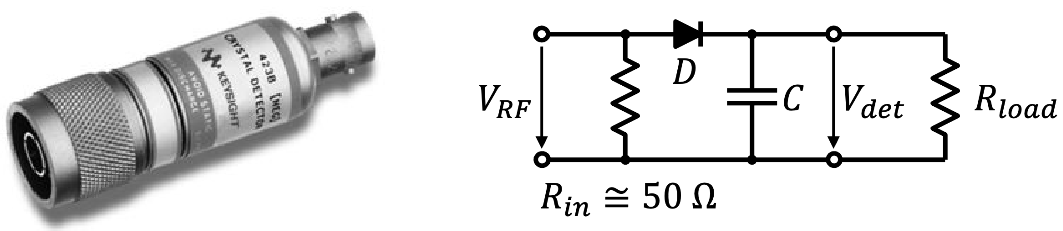
Fig. II.2.172: FM modulation of a CW carrier with as sine wave as modulation signal, for $f_c = 50\,\text{MHz}$ and $\Delta f = 50\,\text{kHz}$, observed as frequency spectrum with a spectrum analyser.

For the extreme case of a very low modulation frequency, $f_m < 1\,\text{Hz}$, Fig. II.2.172f, it is worth to alter the `SWEEP TIME` setting of the spectrum analyser, which then will automatically result in change of the `RBW` and `VBW` bandwidths. The unsynchronised internal sweep of the spectrum analyser and the time period $T_m = 1/f_m$ of the modulating sine wave signal are bouncing with each other, and remember that the spectrum analyser assumes a time invariant behaviour of the signal during the sweep period!

II.2.9.3 RF measurement instruments – some fundamentals and a brief overview

Modern RF measurement instruments, like the vector signal / spectrum analyser or the vector network analyser, are complicated, and still evolving, measurement systems. Therefore, the following introduction is focused on a few concepts and principles utilised in those instruments, also some historical notes are given.

II.2.9.3.1 The RF diode detector



(a) A coaxial RF diode detector. **(b)** Schematics of the RF diode detector.

Fig. II.2.173: RF diode detector.

The *RF diode detector*, or just short *RF detector*, is the most basic RF measurement instrument, see Fig. II.2.173, details beyond the scope in this section can be found in [24]. This simple, passive 2-port element acts as broadband RF *envelope detector*, similar to a *AM demodulator*. Figure II.2.174 illustrates the principle of operation:
A RF signal waveform supplied at the input of the detector (see also Fig. II.2.173b) is *rectified* by a diode (D), here only the positive part of the input signal (blue trace), $V_{RF} > 0$, is passed, and then low-pass filtered such that the envelope of the RF signal appears at the output, V_{det} in Fig. II.2.173b, illustrated as red trace. If the RF signal amplitude is constant, or varies only slowly over time, the detected RF signal envelope can be measured by connecting a digital voltmeter (DVM) at the detector output, to visualise faster, high-frequency baseband signal content, an oscilloscope can be used.

Key element of the RF detector is the *diode*. We briefly introduced this passive, non-linear circuit element in the discussion of the tetrode RF amplifier, see Example II.2.8.7. However, for the RF diode detector the so-called *Schottky*[7] diode is preferred, which is based on a semiconductor-metal *Schottky*-barrier junction. Compared to the more popular p-n junction diode, the *Schottky* diode has two favourable features, making it more beneficial for RF signal detection:

Reverse recovery time , i.e. the time it takes to switch from the conducting forward region to the non-conduction reverse region, which is much lower compared to pn semiconductor diodes.

Forward bias voltage , i.e. the bias voltage required to achieve the conducting state. The forward bias voltage for a pn diode is typically $600 \ldots 700\,\mathrm{mV}$, while the *Schottky* barrier diode starts already to conduct at $150 \ldots 450\,\mathrm{mV}$. As of this low bias voltage, some manufactures call them "zero-bias" *Schottky* diode.

[7] named after *Walter H. Schottky*

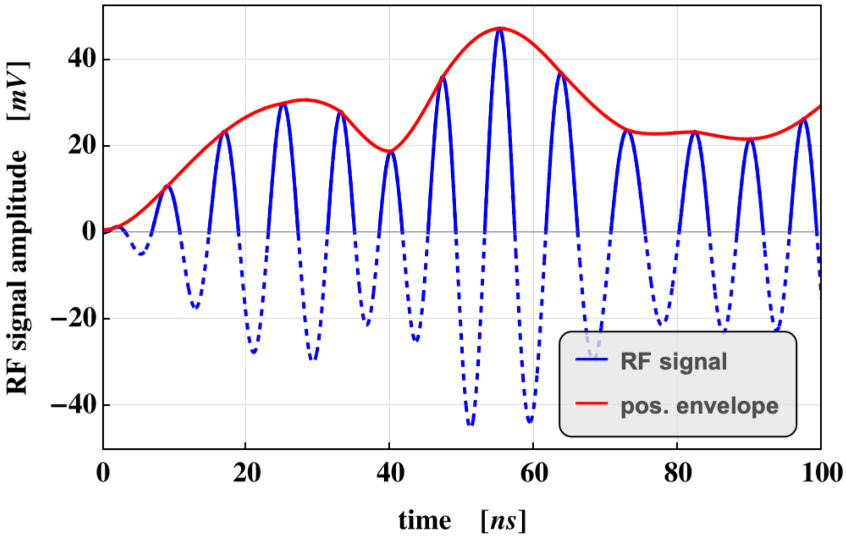
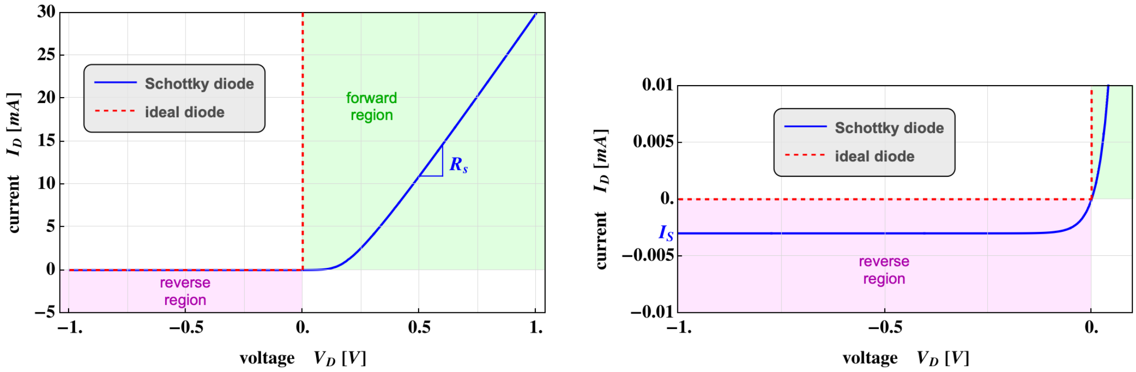


Fig. II.2.174: Envelope detection of a RF signal.



(a) Diode current vs. applied voltage.

(b) IV characteristic in the reverse region.

Fig. II.2.175: IV characteristics of a "zero-bias" *Schottky* barrier diode, compared to an ideal diode switch.

The IV-characteristic of the *Schottky* diode follows

$$I_D = I_S \left(e^{\frac{V_D - I_D R_S}{n V_T}} - 1 \right) \tag{II.2.159}$$

with V_D being the voltage applied to the diode, I_D the current through the diode, I_S the reverse-bias saturation current, $V_T = kT/q$ the thermal voltage ($V_T \approx 26\,\mathrm{mV}$ at room temperature), n the ideality or quality factor, and R_S the series resistance. Figure II.2.175 illustrates Eq. (II.2.159) for $I_S = 3\,\mu\mathrm{A}$, $R_S = 25\,\Omega$ and $n = 1.06$, in comparison to an ideal diode switch.

Please note the rather high values of R_S and I_S of the *Schottky* diode, which are negligible for a traditional pn-junction diode.

A self-biasing circuit and some impedance matching elements are added to the *Schottky* diode to complete the RF diode detector circuit, Fig. II.2.173b shows a simplified schematic of a positive

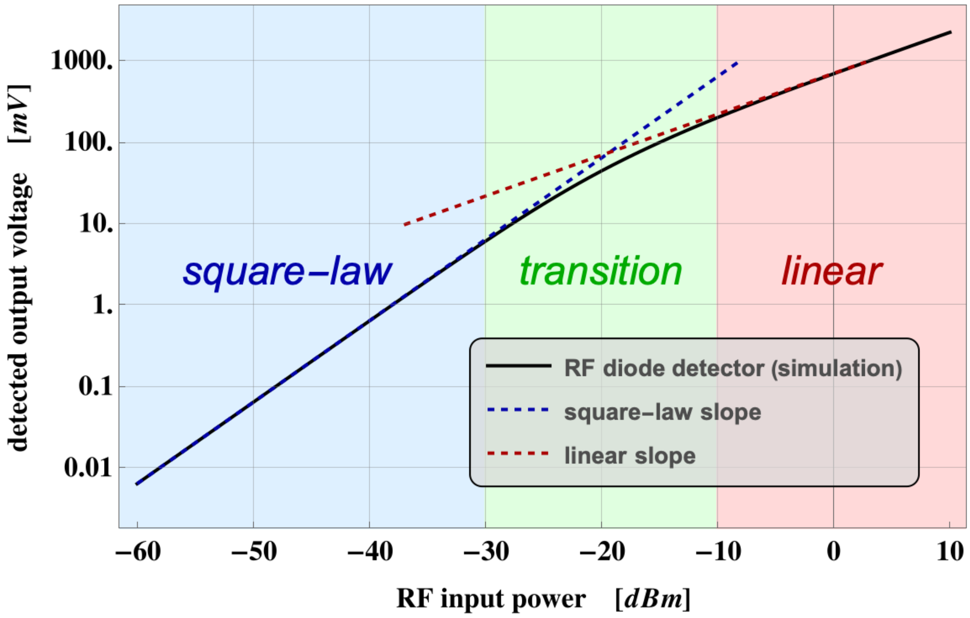


Fig. II.2.176: RF diode detector characteristic: DC output voltage vs. RF input power.

operating detector circuit. For most applications a coaxial implementation of a broadband RF detector is suitable, see Fig. II.2.173a, here shown with a N-type coaxial connector for the RF input and a BNC connector for the detected DC output signal.

Figure II.2.176 shows the DC output voltage vs. RF input power characteristic of a typical RF diode detector, here based of a simulation following [25] for a *HP* HSMS-286x series *Schottky* diode detector[8]. Please note the double-logarithmic scales, the RF input power is given in logarithmic dBm values and the detected output voltage axes is presented in a logarithmic scale! The detected output voltage V_{out} follows the RF input signal power P_{in} by

$$V_{\text{out}} = K\left(\sqrt{P_{\text{in}}}\right)^x \tag{II.2.160}$$

For low power RF input signals, $\lesssim -30\,\mathrm{dBm}$, $x \approx 2$, i.e. the detected output voltage is proportional to the input signal power

$$V_{\text{out}} = \gamma P_{\text{in}} \tag{II.2.161}$$

and we are operating the detector in the so-called *square-law* region, indicated as blue, dashed slope in Fig. II.2.176, with $\gamma = K$ being the diode voltage sensitivity (in $\mathrm{mV/mW}$). For high power RF input signals, $\gtrsim -10\,\mathrm{dBm}$, the diode impedance varies with the power level, and the slope is related to the diode capacitance, operating frequency and load resistance, $x \lessapprox 1$ in Eq. (II.2.160). This is called the *linear* operating region of the RF diode detector, indicated as red, dashed slope in Fig. II.2.176. In between those two regions, within a range of approximately $-30\,\mathrm{dBm} < P_{RF} < -10\,\mathrm{dBm}$, the diode detector operates in the *transition* region.

The usable *dynamic range* of the RF diode detector is usually referred to the square-law region, the upper end is sometimes defined as a $0.5\,\mathrm{dB}$ compression of the voltage sensitivity γ, the lower end is

[8] at CW RF input frequency $1.8\,\mathrm{GHz}$, operating temperature 25° C, load resistance $47\,\mathrm{k\Omega}$, and a bias voltage of $380\,\mathrm{mV}$.

related to the signal-to-noise ratio (SNR) of the detected output signal measured by the following video amplifier, and is given by the *tangential signal sensitivity* (TSS). The TSS value depends on a variety of factors, e.g. RF frequency, video bandwidth (of the detected output signal), DC bias current of the diode, noise figure of the video amplifier, etc., and is defined in different ways, e.g.

$$P_{TSS} = \frac{3.23 \times 10^{-10}\sqrt{BW\,F\,R_v}}{\gamma} \quad @\,300^{\circ}\ K$$

for $SNR = 8\,\text{dB}$, with R_v being the video resistance of the diode, F the noise factor of the video amplifier and BW the video bandwidth. Another definition is based on the minimum pulsed RF input signal that causes an output signal that is equal, or *tangential* to the output noise peaks. Typical values for TSS are $P_{TSS} \approx -50 \ldots -60\,\text{dBm}$, which leads to a dynamic range in the square-law region of $\sim 40\,\text{dB}$, which can be extended with some biasing "tricks" [26].
On the other hand, by careful measurement and correction of the non-linearities of $Vout = f(P_{\text{in}})$, also accounting for the ambient temperature, load impedance, operating frequency, etc., e.g. using a look-up table or fitting function, the diode detector can be used beyond the square-law region, which results in a large dynamic range of $\sim 60\,\text{dB}$, or more.

As of the time of writing, RF detectors based on *Schottky* diodes are offered for a variety of broad frequency ranges up $100\,\text{GHz}$ and beyond, while starting at a few MHz. Beside the RF detector implemented in a coaxial housing, see Fig. II.2.173a, RF detectors are also available for the different types of rectangular waveguides. Of course, the measured RF power as quasi DC voltage at the output of the detector is always a scalar value, there is no information about the RF phase! However, for many applications in accelerator RF this is sufficient, and still a very important and useful information, e.g. the monitoring of the forward and reflected power at a directional coupler near the cavity feed point with a pair of RF diode detectors, followed by a digitizer system, see also Fig. II.2.122.

An impedance matching circuit helps to match the RF input of the diode detector to, e.g. $50\,\Omega$ for coaxial detectors. However, the rectified waveform caused by the switching diode generates a wide range of higher order harmonics, and a major portion of this harmonic RF power is reflected. Therefore, a circulator / isolator located in front of the RF detector can reduce this unwanted harmonic power penetrating into the RF system.

II.2.9.3.2 The slotted transmission-line

The maximum transfer of RF power along a RF system is one of the most important goals in accelerator RF, and in RF engineering in general. This requires a good matching of the characteristic impedance of the transmission-lines used – typically $50\,\Omega$ for coaxial-lines – throughout all interconnections of the installation, leading to minimum reflections and best RF power transmission. This requires a measurement of the complex impedances at the ports of the various components, e.g. amplifiers, couplers, filters, isolators, etc., used in the RF system.

In practice, the measurement of a complex impedance Z_i at port i is performed as measurement of the reflection coefficient $\Gamma_i = S_{ii}$ of this unknown load impedance $Z_L = Z_i$ in a system of a well-known reference impedance Z_0, Eqs. (II.2.16) and (II.2.44). In the "early days" of RF engineering, sophisticated measurement instruments like a vector network analyzer were not available, and this type of measurement was realised with the so-called *slotted line* setup, see Fig. II.2.177. The slotted line measurement utilises the fact of standing waves appearing along a transmission-line of characteristic impedance Z_0 if it is terminated with an arbitrary load $Z_L \neq Z_0$, Fig. II.2.51. The slotted line illustrated in Fig. II.2.177 is a straight, air-dielectric $50\,\Omega$ coaxial line, having a thin slot along the entire length, enabling a little pin antenna to probe the electric field along this almost lossless TEM transmission-line. Similar slotted lines exist also in form of waveguide measurement setups.

Figure II.2.178 illustrates the details of the slotted line $VSWR$ measurement setup, which requires a matched RF signal source providing a CW sinusoidal RF stimulus signal of a selected frequency f to the input of the slotted line, a RF detector as discussion in Section II.2.9.3.1, and a digital (or analogue) voltmeter (DVM) to measure the detected output voltage. The E-field pin probe can be moved along the line – a precision scale helps to mark the exact location along the longitudinal z-axis – and the detected output voltage $|V(z)|$ will follow the standing wave E-field pattern given by an unknown DUT load impedance Z_X. Following Section II.2.6.1 the $VSWR$ is found from the maxima $E_{\text{max}} \propto |V_{\text{max}}|$ and the minima $E_{\text{min}} \propto |V_{\text{min}}|$ of the measured standing wave E-field signal along the line:

$$VSWR = \frac{|V_{\text{max}}|}{|V_{\text{min}}|} \tag{II.2.162}$$

As the slotted line is air-dielectric, $v_p \approxeq c$, the spacing of two consecutive minima $\Delta_{\text{min}} = \lambda/2$ can be used to verify the frequency of the stimulus signal

$$f = \frac{c}{2\Delta_{\text{min}}}$$

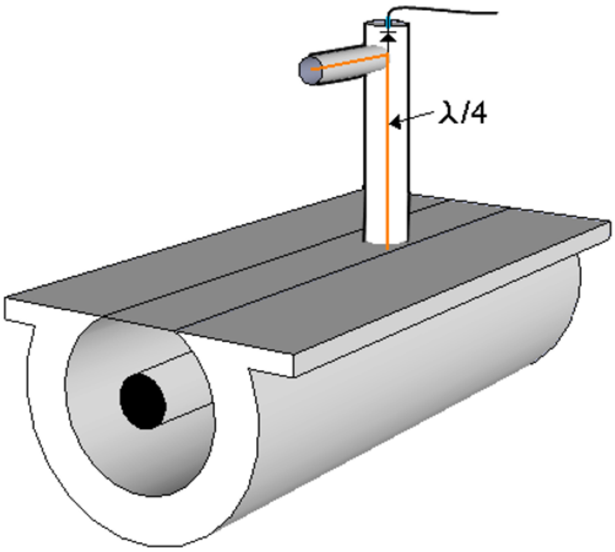


Fig. II.2.177: A slotted coaxial transmission-line for $VSWR$ measurements.

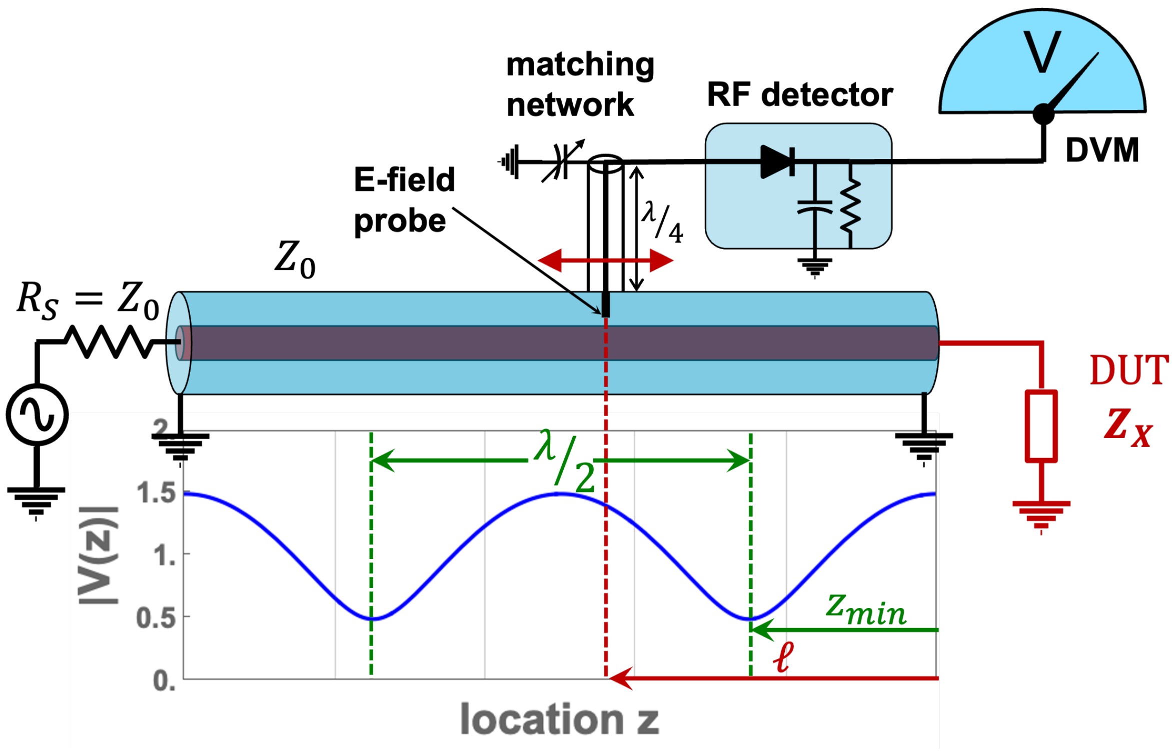


Fig. II.2.178: Impedance measurement through a VSWR characterisation along a slotted line.

From Eq. (II.2.162) it is straightforward to calculate the modulus of the reflection coefficient

$$|\Gamma| = \frac{VSWR - 1}{VSWR + 1}, \tag{II.2.163}$$

see also Eq. (II.2.18). However, to evaluate the phase is a bit more tricky. It requires to measure the distance z_{min} between the last minima on the line, i.e. the first minima from the end of the line, and the actual end of the line at the DUT, see Fig. II.2.178, taking the length of connectors, etc. to the DUT correctly into account. Now we can calculate the argument of the reflection coefficient

$$\angle\Gamma = \frac{4\pi}{\lambda} z_{\text{min}} - \pi \tag{II.2.164}$$

With Eq. (II.2.163) and Eq. (II.2.164) we finally find the value of the unknown load impedance

$$Z_X = Z_0 \frac{1 - \Gamma}{1 + \Gamma} \tag{II.2.165}$$

with Z_0 being the known characteristic impedance of the slotted line.

To successfully measure the E-field with a little pin probe, the low input impedance of the diode detector needs to be transformed, a $\lambda/4$ transmission-line transformer, Eq. (II.2.86) is used to transform towards a high impedance at the pin antenna see Fig. II.2.178. A variable capacitor or similar element is still required to tweak the matching as the $\lambda/4$ TL-transformation is frequency dependent.

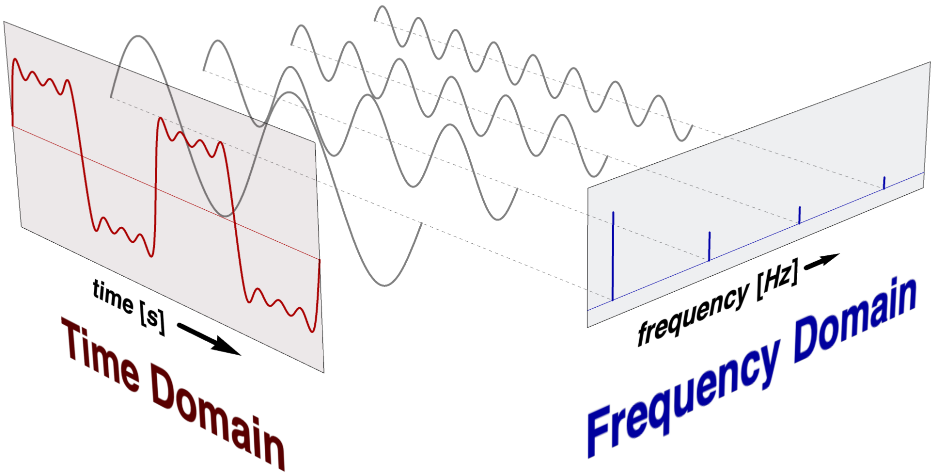


Fig. II.2.179: Waveform of a periodic signal in time and frequency domain.

The slotted line measurement is low-cost, but extremely time consuming and error prone! Please note, also the non-linear behaviour of the RF diode detector, Eq. (II.2.160), has to be taken into account, and note, we have to measure the RF voltage of the standing wave maxima and minima, not the RF power!
Today, the slotted line measurement is obsolete, it would take hours, or even days to characterise an unknown load impedance over a larger frequency range with many points, which would take a modern VNA less than a second. Still, the educational benefit of the slotted line is tremendous, and we always introduce this type of RF measurement at the hands-on RF training during the JUAS Practical Days!

II.2.9.3.3 The spectrum analyzer

Spectrum analyzer (SA), *vector signal analyzer* (VSA) and *FFT-analyzer* are different names for basically the same type of instrument, to detect, analyse and monitor RF signals by it's frequency content, using slightly different technologies. In the following we simply stick with *spectrum analyzer* (SA), but try to explain some of the still evolving technology principles.

The spectrum analyzer is the counterpart to the oscilloscope, while the oscilloscope presents the signal waveform as a function of time, the spectrum analyzer retrieves the spectral content of the signal to present it as function of the frequency – for positive frequencies only – Fig. II.2.179 illustrates the principle for a periodic signal waveform. As the frequency domain is a purely mathematical concept, the signal to be measured has to be time-invariant during the time it takes the spectrum analyzer to sweep the entire frequency axis. The traditional analogue RF spectrum analyzer requires the signal waveform to be periodic, today most spectrum anlyzers have a digital back-end for the signal processing (vector signal analyzer, FFT-analyzer) and have not this limitation.

II.2.9.3.3.1 The superheterodyne principle

To understand the internal function and operation of a spectrum analyser (and as well the vector network analyzer, discussed in the next section), we have to understand how a *radio* works! Even as radios and TVs have been replaced by other technologies, some of their internals are still applied in modern telecommunication products, as well as in RF instruments like the spectrum and network analyzer.

Figure II.2.180 shows the block schematic of a traditional, simple radio receiver. The weak signal from the antenna is amplified by a low-noise RF gain stage, followed by a selective band-pass filter, tunable to receive the desired radio station. Another RF amplifier may be required in front of the demodulator, which separates the modulated audio content from the RF carrier. Finally we need an audio amplifier and a speaker. BUT, this simple radio receiver cannot be build, it is not realisable!
So, what is wrong here?!

It is the tunable band-pass filter! As a matter of fact, even with modern, most sophisticated technologies a band-pass filter, tunable in the range $87 \ldots 108\,\mathrm{MHz}$, with a bandwidth of a few $10\,\mathrm{kHz}$ cannot be made. Moreover, also the demodulation at the high carrier frequency is difficult.
Figure II.2.181 shows the way out of the dilemma, it is called the *superhetorodyne* receiver, also called *superhet*, or simply *super*. Instead of trying to design and build RF components at high frequencies, which is difficult, expensive and sometimes impossible, we realise those RF components at a favourable, fixed frequency, called the *intermediate frequency* (IF). In Fig. II.2.181 this IF stage is indicated in purple, and consist of a band-pass filter and a high-gain IF amplifier, e.g. operating at a IF of $10.7\,\mathrm{MHz}$.

In front of the IF-stage is the heart of the superheterodyne receiver, the *down-converter*, which consist of a frequency *mixer* and a tunable *local oscillator* (LO), both shown in red (Fig. II.2.181). The mixer is a non-linear RF 3-port and acts as a frequency-converter, the RF and LO signal frequencies at its inputs are "mixed", delivering a signal at the IF output of the mixer that contains difference and sum frequencies, $f_{IF} = f_{RF} \pm f_{LO}$. The LO is a tunable RF sine-wave generator, it has to be tuned to a frequency that is the offset of the IF frequency, e.g. above the RF frequency of the desired radio station.

The superheterodyne receiver, Fig. II.2.181, looks more complicated than the simple receiver in Fig. II.2.180. However, it was, and still is the most successful RF receiver principle, it is also converted into the "digital domain" and has many applications beyond RF engineering. Figure II.2.181 illustrates a *single-stage* superhet, for the operation at higher frequencies double and triple superheterodyne receivers have been developed, which than have two or three down-converter and IF stages. An *automatic gain control* (AGC) feedback-loop is used to keep the demodulates signal level constant by automatically con-

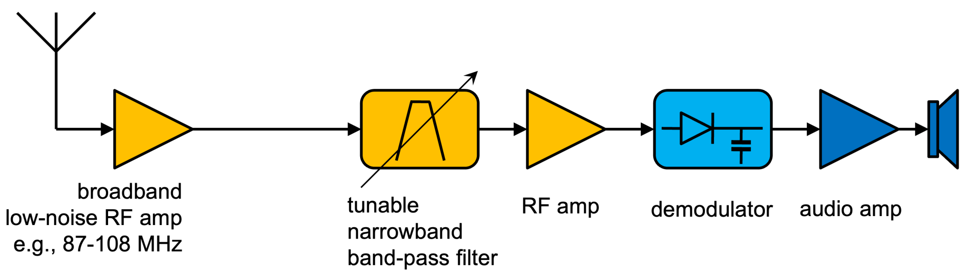


Fig. II.2.180: A "too simple" radio receiver.

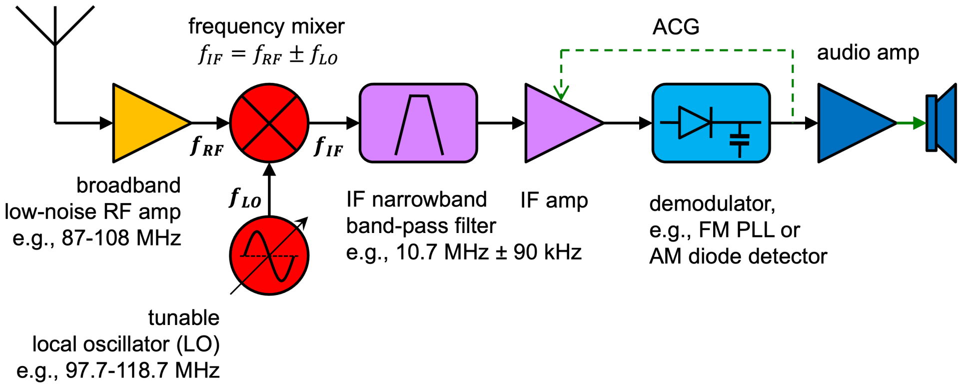
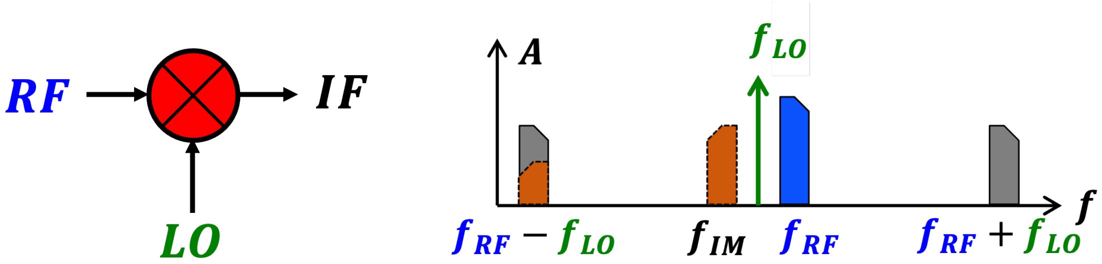


Fig. II.2.181: A superheterodyne radio receiver.



(a) Symbol of the mixer.

(b) Frequency conversion with the mixer.

Fig. II.2.182: The mixer operating as frequency converter.

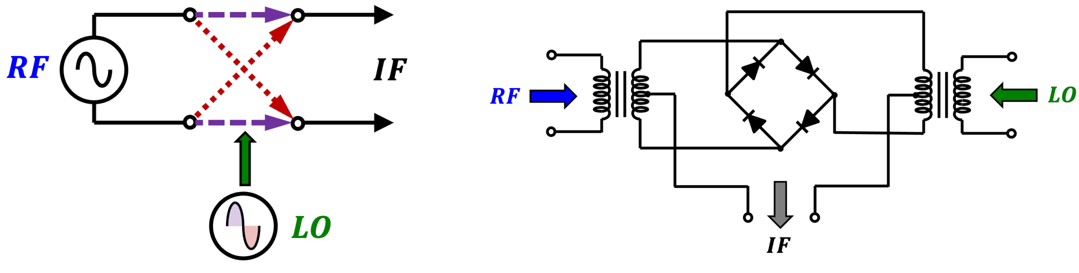
trolling the gain of the IF stage(s), indicated as dashed feedback signal path in Fig. II.2.181. preventing saturation effects on the following gain stages.

> To cover a single large frequency range, as required for the spectrum analyzer, a common technique is to first *up-convert* to a higher frequency and then down-convert in one or several IF steps.
> Also please note, the IF stage may consist of several gain stages and (band-pass) filters.

II.2.9.3.3.2 The mixer

Strictly speaking, the mixer is a RF component and could have already been introduced in the previous section. However, the mixer, operating as frequency converter (see also Fig. II.2.182), is mission critical in the RF up- and down-converters of the superheterodyne receiver sections in spectrum and network analysers, and therefore discussed in this section.

The frequency down-conversion with a mixer in a superheterodyne receiver circuit enables the detection and characterisation of RF signals over a very large dynamic range, from almost thermal noise level ($\equiv -174\,\mathrm{dBm/Hz}$) up to $\approx 10\,\mathrm{dBm}$. The mixer acts as a frequency multiplier, ***RF*** is the RF input signal, ***LO*** is a sinusoidal input signal from a local oscillator, and ***IF*** is the down-converted intermediate



(a) Operational principle.

(b) The double-balanced mixer.

Fig. II.2.183: Schematics of the frequency mixer.

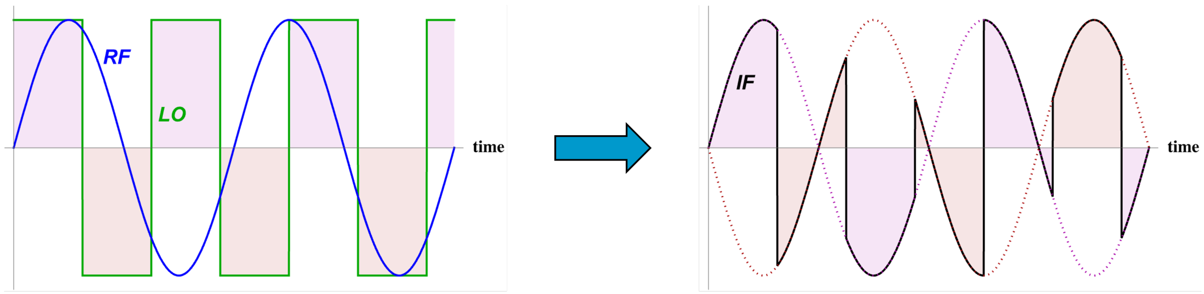


Fig. II.2.184: Time-domain signals of a mixer for $f_{LO} = 1.6 f_{RF}$.

frequency signal at the output, see Fig. II.2.182a. For sinusoidal signals at the inputs of the mixer we get:

$$
\begin{aligned}
v_{RF}(t) &= A_{RF} \sin\left(\omega_{RF} t + \varphi_{RF}\right) \\
v_{LO}(t) &= A_{LO} \sin\left(\omega_{LO} t + \varphi_{LO}\right) \\
v_{IF}(t) &= v_{RF}(t) \times v_{LO}(t) \\
v_{IF}(t) &= \frac{1}{2} A_{RF} A_{LO} \left\{\sin\left[(\omega_{RF} - \omega_{LO})t + (\varphi_{RF} - \varphi_{LO})\right]\right. \qquad \text{down-conversion} \\
&\qquad \left. + \sin\left[(\omega_{RF} + \omega_{LO})t + (\varphi_{RF} + \varphi_{LO})\right]\right\} \qquad \text{up-conversion}
\end{aligned}
\tag{II.2.166}
$$

Figure II.2.182b illustrates the frequency conversion for a frequency band at the RF input, (shown in blue). The mixer will always produce the difference $(f_{RF} - f_{LO})$ and the sum frequencies $(f_{RF} + f_{LO})$ between f_{RF} and f_{LO}, the selection of down- or up-conversion is performed by the band-pass (plus eventual low- or high-pass) filter in the IF stage.

From an operational point of view, the mixer is basically a cross-bar switch controlled by f_{LO}, see Fig. II.2.183a. Every half LO-cycle, $1/(2 f_{LO})$, the sign of RF input signal is flipped at the IF-output, which in practice requires the LO control signal to always be larger then the RF input signal, $A_{LO} > A_{RF}$. Therefore, in a time-domain signal visualisation, the LO signal can be assumed as rectangular switch control waveform, see Fig. II.2.184.

Industry offers a large variety of passive (diode-based) and active (transistor-based) RF mixers, as chips, with coaxial connectors and for waveguides. High-frequency mixers – mm-wavelength regime – with waveguide RF input and coaxial IF output can be attractive to extend the frequency range of existing measurement equipment, spectrum and/or network analysers. Figure II.2.183b shows as example the

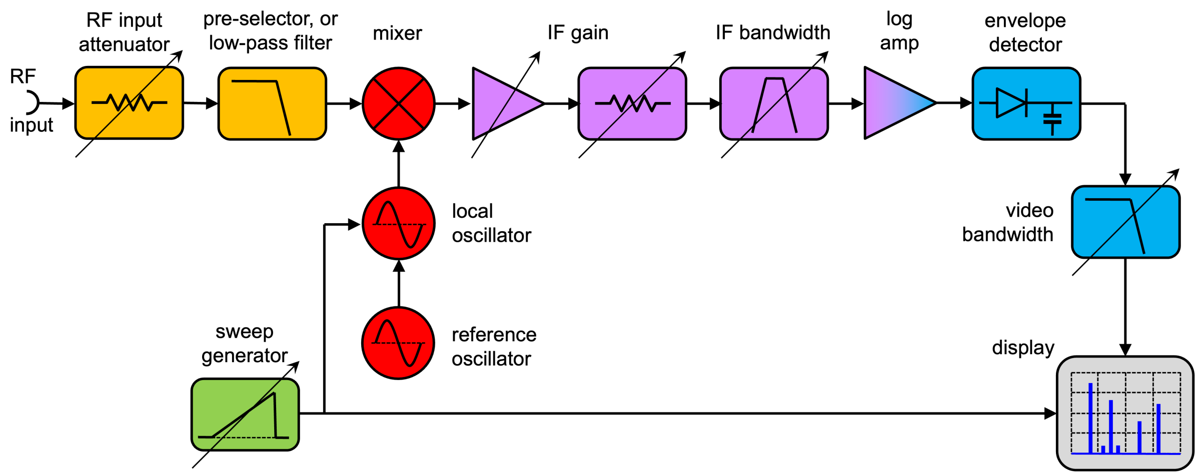


Fig. II.2.185: Simplified block schematic of a "traditional" spectrum analyzer.

schematic of the popular *double-balanced* mixer, which is based on four diodes in a bridge configuration, and the RF- and LO-signals supplied via baluns. Form the LO-signal view, the diodes are always switched fully in saturation, while from the RF-signal perspective the diodes stay in the square-law regime. This symmetric circuit arrangement offers a good LO-to-RF signal isolation and the rejection of other unwanted spurious signals in the IF-output.

Please note the difference of an ideal mixer and its "real-world" implementation:

$$\begin{aligned} f_{IF} &= f_{RF} \pm f_{LO} && \text{ideal mixer} \\ f_{IF} &= m f_{RF} \pm n f_{LO} && \text{"real-world" mixer} \end{aligned} \tag{II.2.167}$$

Beside that, most mixers also produce an "image", i.e. a *side-band* of the RF-signal, indicated in orange in Fig. II.2.182b, which then unfortunately gets down-converted to the same IF baseband. The careful section of LO and IF frequencies, the use of appropriate filters and eventually using so-called *image-rejection* mixers can address this problem.

II.2.9.3.3.3 The "traditional" spectrum analyzer

After this lengthy introduction we finally come to the spectrum analyzer. Figure II.2.185 shows a simplified block schematic of the "traditional" version of this measurement instrument, as it was build unit the early 2000s, almost entirely based on RF and analogue components, but microprocessor controlled. It is not too difficult to recognise the superheterodyne receiver sections, but, there are a few additional blocks, e.g. a *sweep generator*, a *reference oscillator*, a *logarithmic amplifier*, and a *display* (CRT or digital), among others.

The sweep generator is used to define the `Start` and the `Stop` frequencies by controlling the local oscillator, and – directly or indirectly – the `Sweep Time` of the trace cursor moving horizontally across the display. The vertical axis of the display is controlled by the down-converted and demodulated RF input signal, which in combination with the horizontal sweep ramp results in a frequency-domain representation of the signal waveform at the RF input. A few remarks to the operational relevant circuit blocks:

sweep generator The *graticule* of the SA display is divided in 10×10 `Divisions`, see Fig. II.2.186a. As mentioned, the saw-tooth waveform of the sweep generator is controlled by the user setting to the desired frequency range, i.e. by selecting a `Start` and a `Stop` frequency, or alternatively as `Center` and `Span` frequencies for the horizontal axis of the graticule. The *sweep-rate* $df/dt = SPAN/ST$ of the saw-tooth ramp is usually automatically selected and defines the `Sweep Time` (ST) based on the `Span` setting.

IF gain The vertical axis is for most cases scaled in logarithmic dBm units and the scaling is provided by the user in terms of, e.g. $\mathrm{dB/div}$. The user also has to set a `Reference` level (Ref) in dBm to the vertical scale, e.g. to the center division (`5 Div`) or to the upper division (`10 Div`). These settings of the vertical scale will modify the IF gain accordingly, but will also modify other offset and gain settings not shown in Fig. II.2.185.

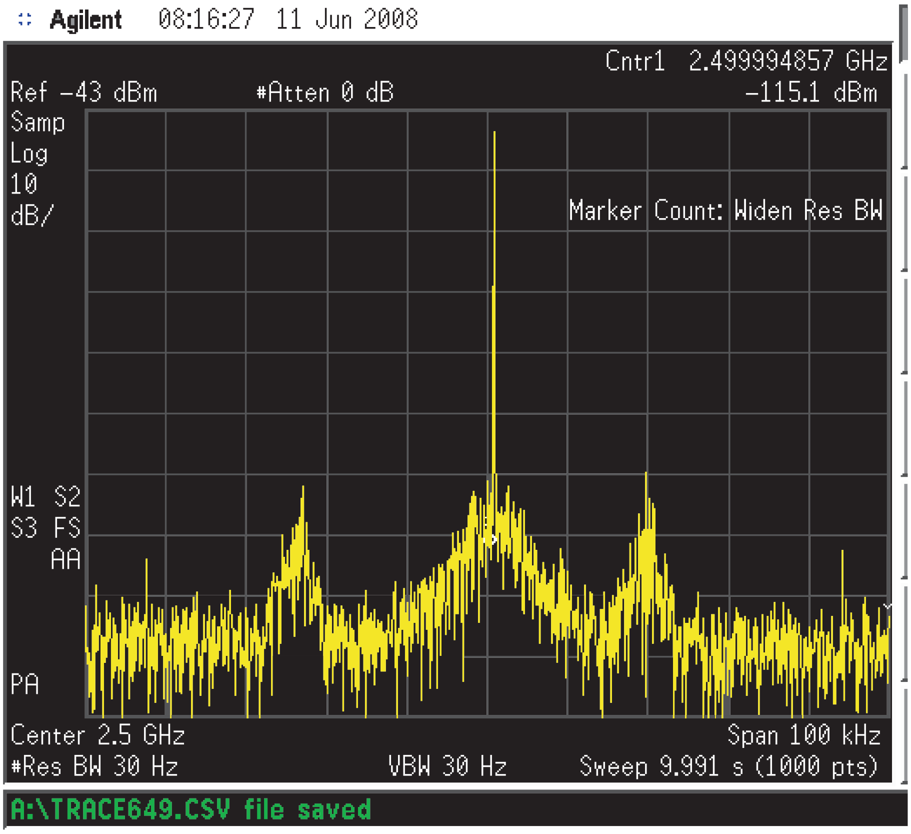
IF bandwidth The *Fourier* transform of a pure sine-wave is a *Dirac* δ-function, located at the frequency $f = \omega/(2\pi)$. The measurement and display of a perfect sine-wave signal with the spectrum analyzer would however require the bandwidth of the IF band-pass filter to be infinite narrow. In practice the user has to select a `Resolution Bandwidth` (RBW), which is the *IF bandwidth*, to a reasonable value, typically in a range between a few Hz and a few MHz. The more narrow the bandwidth of the *Gaussian* IF band-pass filter, the higher is the resolution, resulting in a better SNR. However, selecting a narrow IF filter bandwidth takes the filter more time to settle to a steady-state, thus the sweep-rate has to decrease accordingly, and a frequency sweep between `Start` and `Stop` frequency will take longer. The SA automatically follows the *coupling* rule

$$\frac{df}{dt} = \frac{SPAN}{ST} < \frac{RBW^2}{k} \tag{II.2.168}$$

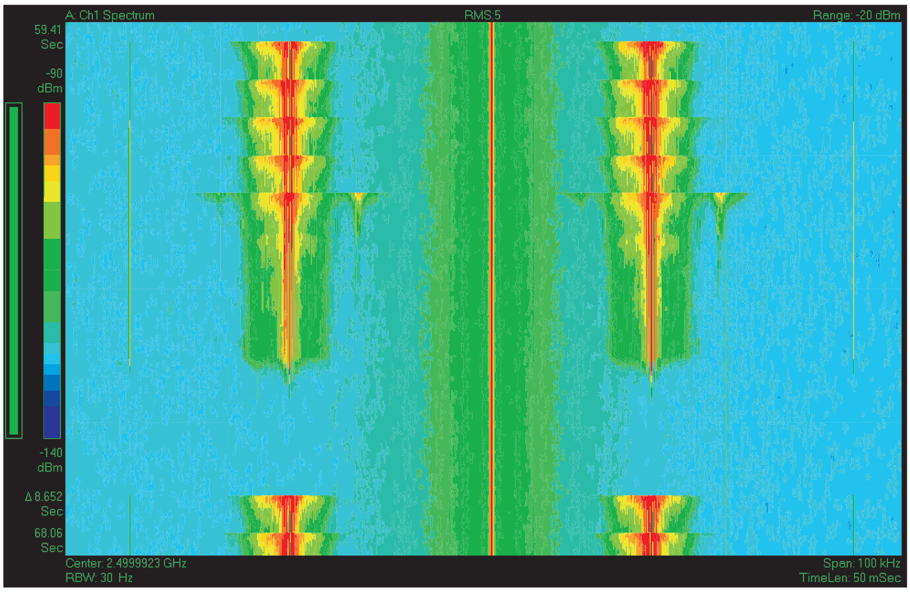
for a given IF filter bandwidth RBW, ensuring a steady-state condition, i.e. a stable signal amplitude within $\Delta T = 1/RBW$. $k \approx 2 \ldots 3$ in Eq. (II.2.168) accounts for the synchronously-tuned, *Gaussian*-like IF filter characteristic.

video bandwidth (VBW) is the bandwidth of the low-pass video filter of the demodulated signal. Usually the SA automatically sets $VBW = RBW$ as part of the automatic coupling of $SPAN$, ST, RBW, and VBW. Decreasing VBW has the effect similar to a signal averaging over several frequency sweeps, it improves the SNR but increases the sweep time.

[9]courtesy *F. Caspers*, CERN.



(a) Spectrum analyzer display with typical settings for a high resolution bandwidth.



(b) Spectrogram display for 200 measurement traces, arranged in a colour-coded contour graph. Time runs vertically from top to bottom.

Fig. II.2.186: Spectrum analyzer measurements of electron cloud studies in the SPS (CERN)[9]

The measurement of low-level RF signals, e.g. $< -120\,\mathrm{dBm}$, requires a low *noise-floor*, thus a low RBW value, which automatically leads to a long, eventually very long sweep time! Therefore, it is advisable to reduce the `Span` as much as possible, and of course to set the RF input attenuator to $0\,\mathrm{dB}$. Furthermore, the rather poor noise performance of the SA can be substantially improved with a build-in pre-amplifier (optional, not shown in Fig. II.2.185) or a dedicated external LNA.

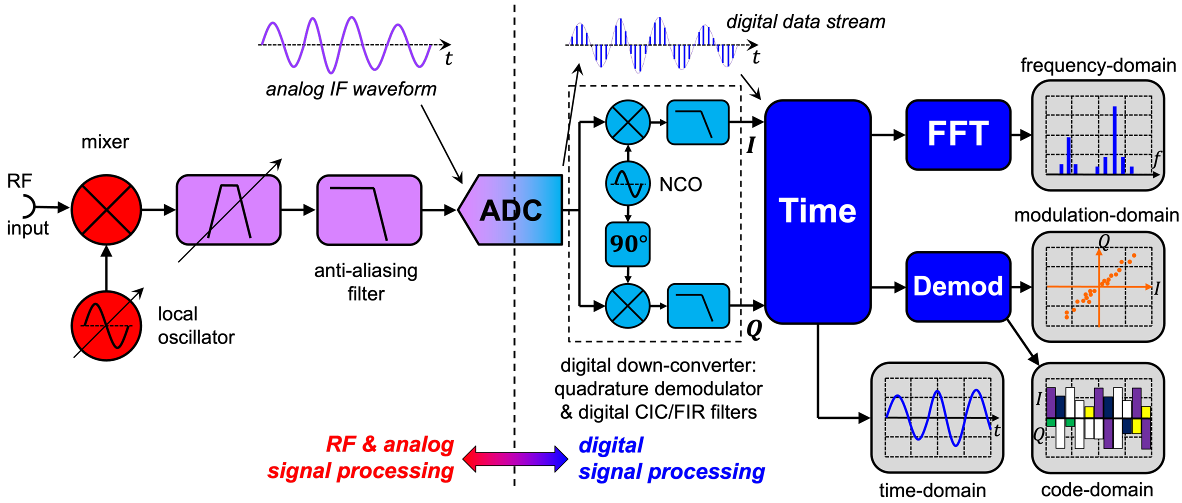


Fig. II.2.187: Simplified block schematic of a "modern" vector signal (FFT) analyzer.

The "traditional" analogue RF SA, Fig. II.2.185, utilises a logarithmic amplifier in the IF stage, which acts as a dynamic compressor, a large IF input range, $> 120\,\mathrm{dB}$, is mapped on a small range of a few volts for the demodulated video signal. In practice, logarithmic amplifier and envelope detector (demodulator) are interleaved on a single chip.

The tunable local oscillator is based on a phase-lock loop (PLL) RF generator and needs to be stabilised by a reference oscillator.

II.2.9.3.3.4 Vector signal (FFT) analyzer

Today, all spectrum analysers are vector signal or FFT analysers, see Fig. II.2.187, which digitise the IF signal and apply the digital signal processing techniques used in telecommunication systems. Key element is the *analog-to-digital converter* (ADC), which provides a digital replica of the IF signal to a *digital down-converter* (DDC), which in may cases is integrated on the ADC chip itself. The DDC acts as baseband demodulator with two orthogonal outputs, an *in-phase* I-channel, equivalent to the real part of a complex number, and a *quadrature-phase* Q-channel, equivalent to the imaginary part of a complex number. The *numerically controlled oscillator* (NCO) is tuned to the IF frequency and phase-locked to the reference oscillator of the system (sorry, some of these details are not depicted in Fig. II.2.187), and therefore enabling the frequency analysis of the RF signal waveform – not limited to a scalar, like for the "traditional" analogue RF spectrum analyzer – in a vector format, plus the measurement of non-periodic and transient RF waveforms. The performance of the vector signal analysis is mostly given by the ADC characteristics, i.e. the maximum IF bandwidth is less then half the ADC clock rate (*Nyquist-Shannon* theorem), and the spurious-free dynamic range is linked to the *effective number of bits* (ENOB) of the ADC.

Figure II.2.186 shows electron cloud measurements performed long time ago with an older generation of a vector signal spectrum analyzer, still, gives an impression of the many display options. The "modern" vector signal analyser, Fig. II.2.188 is / can be equipped with an almost infinite number options and features, mostly tailored to the telecommunications industry. However, noise figure and IP3 measurements, phase-noise and clock-jitter analysis are some of the very useful options also for accelerator RF engineering. Another feature of the vector signal analyzer is the fact to better understand AM and FM modulation sidebands in complicated RF signals as the sign information of each harmonic is available.

The block schematics discussed in this section, Figs. II.2.185 and II.2.187, present a very simplistic picture of the spectrum analyser technology. In most cases, the spectrum analyzer utilises two to four frequency converters, with the first one being an up-converter. The layout of frequency ranges, and LO and IF frequencies has to be optimised to avoid unwanted image and spurious frequencies in the detection system of the SA. More details on this, and on the technology of spectrum analyzers in general can be found in literature from industry [27].

Lab Procedure II.2.9.7: TOI measurement with the spectrum analyzer

The measurement of the 3rd intercept point (TOI or IP3) of a RF amplifier, see also paragraph II.2.8.4.2.2, is straightforward, however, beside the spectrum analyzer it requires some additional equipment, see also Fig. II.2.189:

1. A spectrum analyzer, here a `Anritsu MS2692A`, 50 Hz – 26.5 GHz was used to observe the spectral content of the signals Fig. II.2.188.

2. Two RF signal generators (ideally identical one), e.g. `Agilent E8257D`, 250 kHz – 20 GHz, to generate the two sinusoidal RF tone signals, Fig. II.2.189a.

3. A RF power combiner, to combine the two signals from the generators. Please ensure the

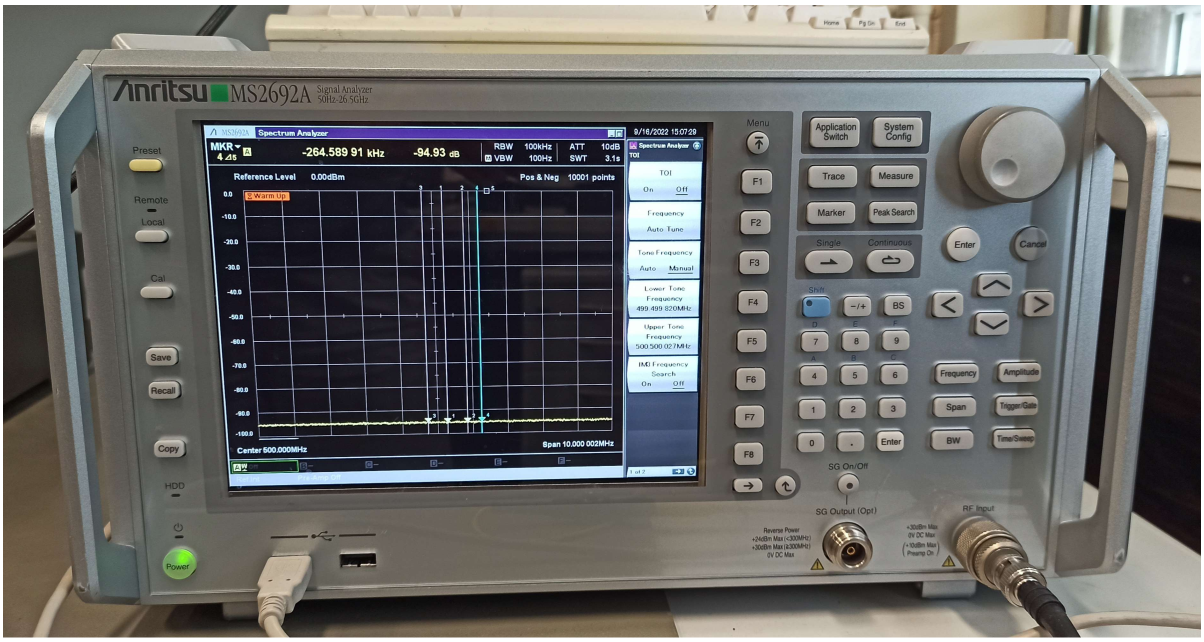


Fig. II.2.188: An older, but still "modern" vector signal analyzer (*Anritsu*).

power combiner operates up to the frequency range of interest, here 1 GHz, Fig. II.2.189b.

4. Three RF attenuators, $2\times$ 10 dB (or similar) attenuation, and one of 4 dB (or similar) attenuation, Fig. II.2.189b.

(a) Two RF signal generators.

(b) Power combiner and attenuators.

Fig. II.2.189: Additional equipment for a TOI (IP3) measurement.

Preferably use BNC cables for the following setup:

1. Connect the two **same value** attenuators to ports 1 and 2 of the power combiner, they are essential to improve the isolation between the input ports of the power combiner. It is recommended to use attenuators of 10 dB attenuation, or more, except a power combiner with very high isolation (> 40 dB) is available.

2. Connect the 4 dB (or similar) attenuator to the 3rd sum port (S). This attenuator is optional, but recommended to improve the return loss of the sum port of the power combiner.

3. Connect the outputs of the signal generators to the two ports (1 and 2) of the combiner, to the two attenuators.

4. Connect the 3rd (S) sum port to the input of the amplifier.

5. Connect the output of the amplifier to the input of the spectrum analyzer.

6. Set the frequency of the 1st RF signal generator to $499.5\,\mathrm{MHz}$ and to $500.5\,\mathrm{MHz}$ on the 2nd generator.

7. Set the amplitude of both the signal generators to $-20\,\mathrm{dBm}$, or less.

8. On the spectrum analyzer, from `Measure` menu choose the `TOI` automatic measurement.

9. Press `TOI:On` and press `Frequency Auto Tune`. You should see four harmonic signal peaks, if you see more, you should check again your setup. In the latter case, the sources may be not properly isolated from each other.

10. Verify if the two main harmonic peaks in the spectrum are of equal amplitude. If not, correct the amplitude of one signal accordingly until they are equal.

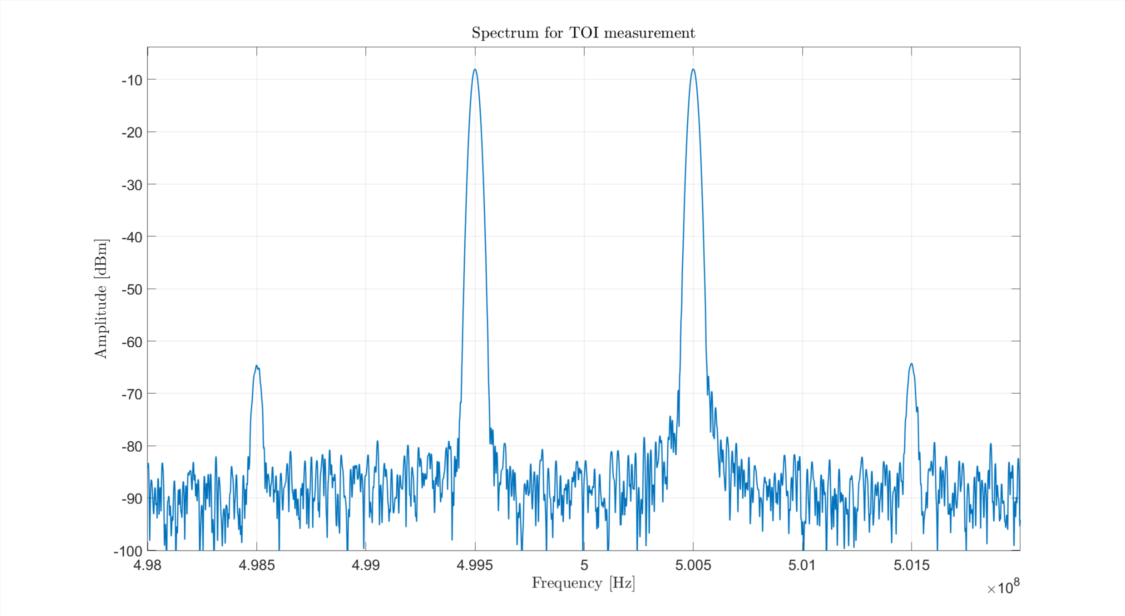


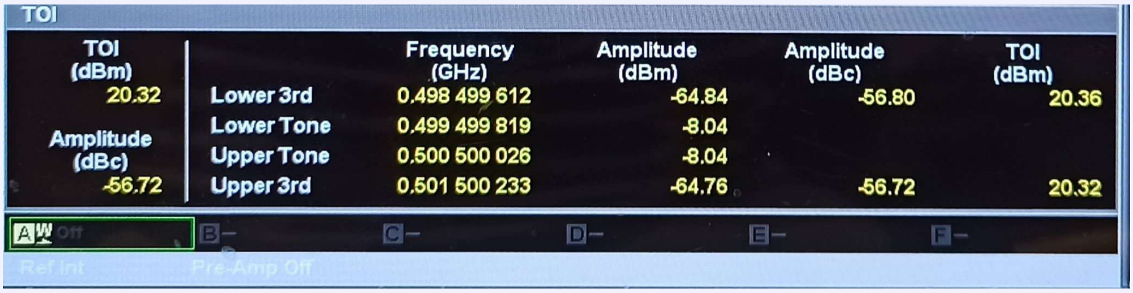
Fig. II.2.190: Spectrum for the TOI measurement

Figure II.2.190 shows the measured spectrum of the two tones and the 3rd-order intermodulation harmonics. By locating markers to the peaks of the harmonics we measure the level of the fundamental tones as $\approx -8.04\,\mathrm{dBm}$, and for the intermodulation tones as $\approx -64.8\,\mathrm{dBm}$. Hence, their difference is $56.25\,\mathrm{dBm}$.

Now the TOI can be calculated:

$$(TOI)_{\mathrm{dBm}} = P_{tone} + \frac{P_{\Delta}}{2} = -8.04 + \frac{56.76}{2} = -8.04 + 28.34 = 20.34dBm$$

which is very close to the value given by the automated process of the signal analyzer (Fig. II.2.191).



	Frequency (GHz)	Amplitude (dBm)	Amplitude (dBc)	TOI (dBm)
Lower 3rd	0.498 499 612	-64.84	-56.80	20.36
Lower Tone	0.499 499 819	-8.04		
Upper Tone	0.500 500 026	-8.04		
Upper 3rd	0.501 500 233	-64.76	-56.72	20.32

Fig. II.2.191: Automated TOI measurement values from the signal analyzer.

Lab Procedure II.2.9.8: Noise figure measurement with the spectrum analyzer

Noise in RF amplifiers was discussed in Section II.2.8.4.3, including giving definitions for the noise factor F, Definition II.2.8.4, and the noise figure NF, Definition II.2.8.5.

Consider an ideal (noiseless) amplifier, terminated at its input (and output) with a source (and load) resistor, operating at 290 K ambient (room) temperature with an available power gain G_P. At the output we measure:

$$P_a = k_B T_0 \Delta f G_P$$

For $T_0 = 290\,\mathrm{K}$ [a] we obtain $k_B T_0 = -174\,\mathrm{dBm/Hz}$. At the input we determine a given signal with a certain signal-to-noise ratio S_i/N_i, and at the output a SNR of S_o/N_o, which results in the definitions of noise factor, Eq. (II.2.137), and noise figure Eq. (II.2.140).

NF indicates the performance of a device due to its internal components and physical limits. Of course, we want the noise figure value to be as low as possible because this indicates a better noise performance of the device. A low noise figure means, the device adds very little noise, and a high noise figure means it adds a lot of noise.

An ideal amplifier has $F = 1$ or $NF = 0\,\mathrm{dB}$. The *noise temperature* of this amplifier is 0 K, and the signal and noise levels at the output will linearly increase with the gain. A real-world amplifier will adds some noise on its own, which leads to a decrease in S_o/N_o due to the added noise N_a, see also Eq. (II.2.138):

$$F = \frac{N_a + N_i G_a}{N_i G_P} = \frac{N_a + kT\Delta f G_a}{k_B T_0 \Delta f G_P} \tag{II.2.169}$$

The noise factor F and the noise temperature T are related as:

$$T = T_e = \frac{N_a}{k_B \Delta f G_P} = T_0(F-1) \tag{II.2.170}$$

with T_e being the equivalent temperature of a source impedance into a perfect, noise-free device, e.g. a RF amplifier with power gain G_P, that would produce the same added noise N_a, please compare with Eqs. (II.2.130) and (II.2.138).

The *Y-factor method* is a popular way to measure the noise figure. It is based on a switchable noise source with two calibrated values N_1 and N_2 for the noise temperature, T_c and T_h, corresponding to “cold” and “hot”. Usually a dedicated *noise diode* is used as noise source, switched between non-bias and bias operation to provide the two noise temperatures. The calibrated noise level is defined as *excess noise ratio* (ENR):

$$ENR_{\text{dB}} = 10 \log \frac{T_h - T_c}{T_0}$$

or in linear form:

$$ENR = 10^{\frac{ENR_{\text{dB}}}{10}}$$

The noise source is connected to the amplifier or device-under-test (DUT)to be analysed, providing noise “on” (N_2) and “off” (N_1) conditions. The ratio of these noise powers is called the Y-factor, and it is essentially the ratio of the noise power at the output of the DUT with and without the added noise:

$$Y = \frac{N_2}{N_1}$$

Y-factor and ENR can be used to determine the noise slope of the DUT. The calibrated ENR of the noise source represents a reference level for the input noise, which allows the calculation of the internal (added) noise N_a of the DUT:

$$N_a = k_B T_0 \Delta f G_P \left(\frac{ENR}{Y - 1} - 1 \right)$$

Most spectrum analyzers (SA) have the option for operating in an automatised noise figure measurement mode, which automatically controls the noise diode, i.e. switching between “hot” (on) and “cold” (off) states, acquiring the DUT output signal, and computes – based on the calibrated ENR of the noise diode – the total system noise figure:

$$F = \frac{ENR}{Y - 1} \tag{II.2.171}$$

which includes noise contributions from all parts of the system. In case the “cold” noise temperature $Tc \neq T_0 = 290K$:

$$F = \frac{ENR - Y(\frac{T_c}{T_0} - 1)}{Y - 1}$$

To perform noise figure and gain measurements with the SA, a noise source is needed, which adds a well-defined and ideally white noise to the DUT, the excess noise ratio (ENR), see Fig. II.2.192.

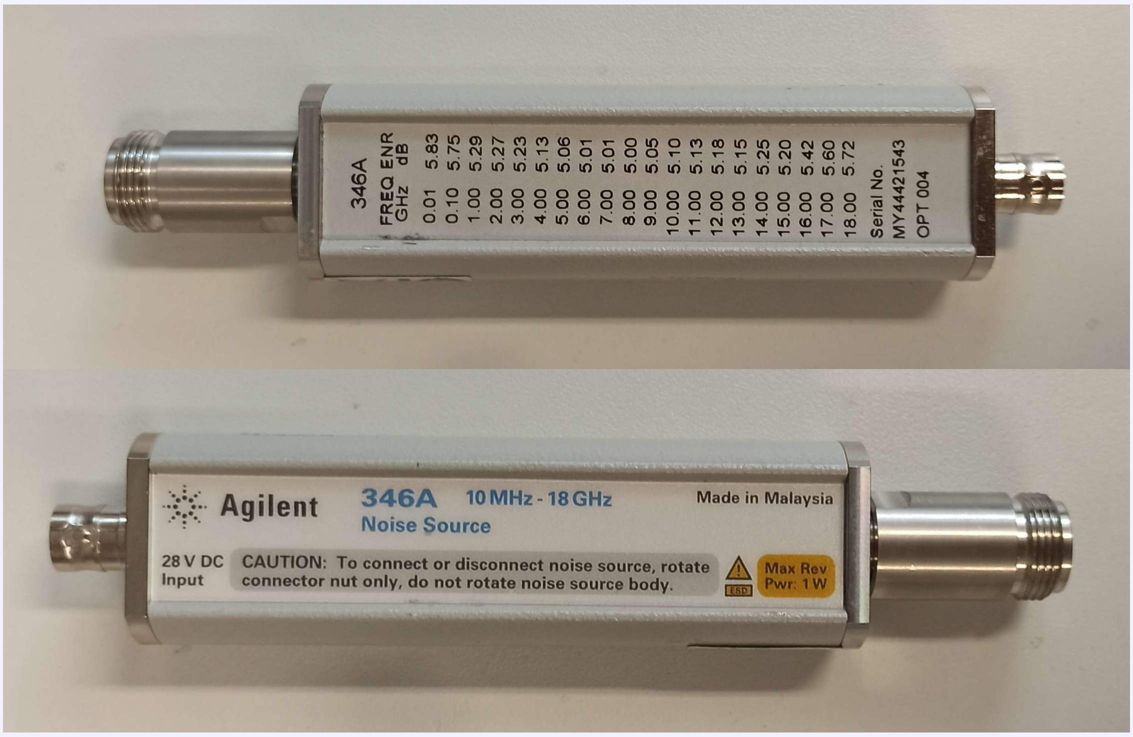


Fig. II.2.192: A noise source, `Agilent 346A`, 10 MHz $-$ 18 GHz.

The output of the noise source has a minor frequency and temperature dependency, and the user has to upload the ENR tables into the analyzer. Typically, this data is printed a table on the back of the noise source (Fig. II.2.192). In our case, we will assume that value to be constant, ≈ 5.4 dB, as our frequency range of interested is $\approx 0 - 2$ GHz.
Preferably use BNC cables for the following setup:

1. Connect the `noise source control output` on the back of the spectrum analyzer with a cable to the input of the `Agilent 346A noise source`, and then connect its output to the input of the amplifier.

2. Connect the output of the amplifier to the input of the spectrum analyzer.

3. On the SA from the `Application Switch` select the automatic `Noise Figure` measurement.

4. Go to `Sweep Setting` and set `Start` to 10 MHz, or similar (it doesn't really matter as long as it is a reasonable low frequency) and `Stop` to 2 GHz.
   Set `Sweep Points`, e.g. to 21. (For the graphs presented here 51 points were used, but that would take a lot more time, 21 measurement points are sufficient).

5. From `Frequency Mode` select `Sweep`.

6. Go to `Common Setting` > `ENR` > `Noise Source Select` > `NC346A` to match the one we use.

7. `ENR mode` > `spot`, `spot ENR` > 5.3 dB, to approximately match the ENR from the table on the back side of our noise diode.

8. Go to `Common setting` > `cal setup` > set `min ATT` to 0 dBm, if it is not already set.

9. As you notice, the `CAL Status` indicated the instruments is uncalibrated. To fix that, press `Calibration Now`, and after waiting some seconds you should see `OK` on the `CAL Status`.

Figure II.2.193 shows the measurement results, in the upper graph the noise figure vs. frequency, in the lower graph the gain vs. frequency. With increasing frequency, the noise figure increases, hence it degrades. The same applies for the gain, however, the specified operational region of the amplifier ends at 1 GHz, and in this region the noise figure is always below 4 dB, and the gain is > 30 dB, as indicated in the amplifier specifications. Overall, the amplifier seems to operate satisfactorily in its specified frequency range.

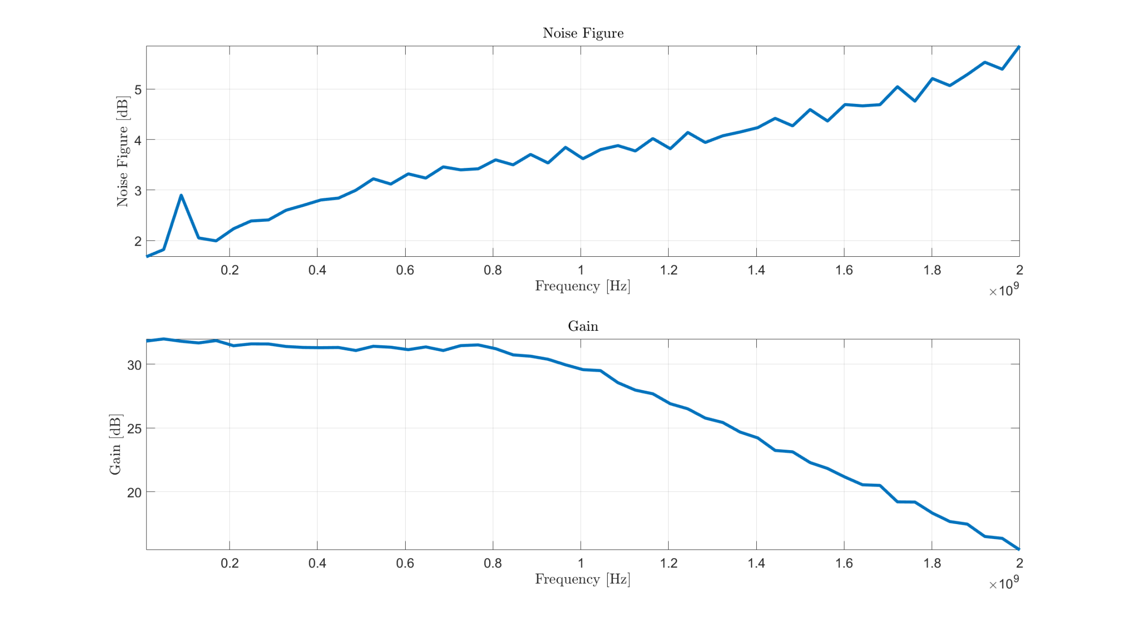


Fig. II.2.193: Noise figure and gain vs. frequency measured by the spectrum analyzer.

[a]It is important that the device (e.g. the amplifier) stays at constant room temperature of 290 K. This temperature plays a role because the noise contribution in electronics is mainly caused by thermal agitation of electrons, the *thermal noise*. Sometimes T_0 is referred to 300 K, in any case, this T_0 value should be kept constant throughout the measurement procedure, as it is important to achieve correct results!

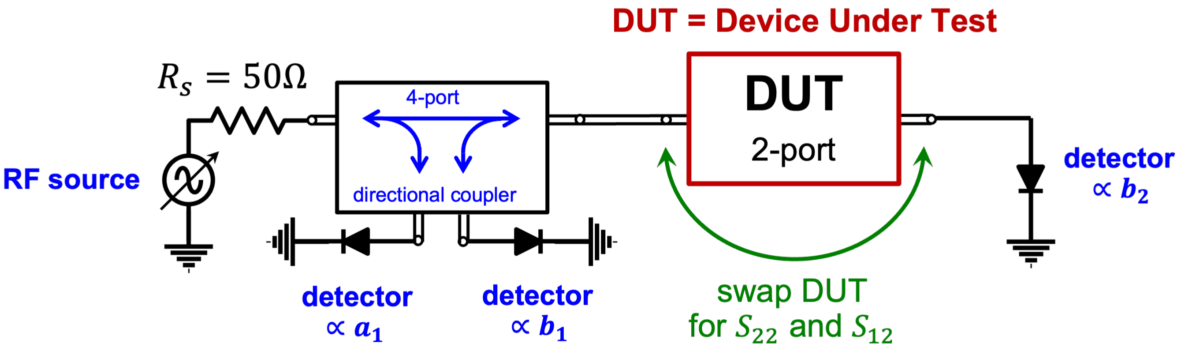


Fig. II.2.194: A simple method to measure S-parameters.

II.2.9.3.4 The vector network analyzer

As mentioned in Section II.2.9.1, the RF *vector network analyzer* (VNA) is a laboratory instrument to characterise RF components, elements and subsystems, primary by measuring their S-parameters, see also Fig. II.2.155. As the name "vector" network analyzer suggests, there have also been scalar network analyzers, however, today all RF network analyzers are vector network analyzers and the related acronym "VNA" is used for the naming.

II.2.9.3.4.1 How to measure S-parameters?

The scattering parameters (S-parameters), and their advantages to describe the characteristics of RF components and subsystems have been discussed in Section II.2.7. A method to measure the reflection coefficient $\Gamma = S_{ii}$ at a given port ii of the DUT, based on a $VSWR$ measurement utilising a slotted transmission-line, was introduced in Section II.2.9.3.2. However, this method appears to be too time-consuming and error prone, and it cannot measure the transmission coefficients S_{ij}.

Figure II.2.194 shows a simple measurement setup based on a directional coupler, see also paragraph II.2.8.1.7.4, to measure $S_{ii}(f)$ and $S_{ij}(f)$, here illustrated for a 2-port DUT S-parameter characterisation. Beside a directional coupler with well known parameters, we also need three well characterised RF diode detectors, as discussed in Section II.2.9.3.1, and a RF source which delivers a sinusoidal waveform of adjustable frequency, e.g. a RF sweep generator or synthesizer. The RF diode detectors must have a perfect match to the DUT, resp. directional coupler port impedances, i.e. a resistance of $50\,\Omega$ at the RF input, to avoid uncontrolled reflections in the measurement setup ($a_2 = 0$). Of course, a control and data acquisition (DAQ) system is also required, but not shown in Fig. II.2.194.

The measurement of the 2-port DUT S-parameters is then performed in two steps:

1. Set the RF source generator to the desired frequency, or frequency range f, and acquire $a_1(f)$ and $b_1(f)$ at the directional coupler and $b_2(f)$ at the output port of the DUT. From $a_1(f)$ and $b_1(f)$ follows

$$S_{11}(f) = \left.\frac{b_1(f)}{a_1(f)}\right|_{a_2=0}$$

and $a_1(f)$ and $b_2(f)$ gives

$$S_{21}(f) = \left.\frac{b_2(f)}{a_1(f)}\right|_{a_2=0}$$

2. Swap the ports of the DUT and repeat the procedure, now we are acquiring $a_2(f)$, $b_2(f)$ and $b_1(f)$, while $a_1 = 0$, which results in $S_{22}(f)$ and $S_{12}(f)$.

For DUTs with more than two ports, more port “swapping” is required, while ensuring the unused ports are always perfectly terminated in their port impedance!

Unfortunately the RF diode detectors do not measure signals $\propto a_i$ or $\propto b_i$, they do measure only the rectified RF envelope, thus the scalar values of the RF signals $\propto |a_i|$ and $\propto |b_i|$, as explained in Section II.2.9.3.1. Therefore, our simple S-parameter setup, Fig. II.2.194, is a **scalar** “network analyzer” (SNA), measuring $|S_{ii}(f)|$ and $|S_{ij}(f)|$!

II.2.9.3.4.2 VNA principle of operation

The vector network analyzer (VNA) is based on the superheterodyne receiver principle and share some hardware aspects of the vector signal analyzer, see the previous Section II.2.9.3.3. Figure II.2.195 shows a simplified block schematic of a 2-port VNA, illustrating the key elements of the instrument[10]. The DUT is connected via high-quality $50\,\Omega$ coaxial cables to the two ports. The RF- and LO-source generators are phase-locked to a reference oscillator (not shown) and provide the RF and LO signals for the four down-converters, enabling a simultaneous measurement of the four power waves, a_1, b_1, a_2 and b_2, as a function of the frequency f of the RF source stimulus signal at the two ports of the DUT. A cross-bar switch toggles to feed the RF source signal to the couplers of port 1 and port 2, providing the incident a_i power wave. The IF sections (1-of-4 shown) resemble similar technologies like in the vector signal analyzer, Fig. II.2.187, and provide a vector representation of the demodulated signals.

[10]Different block configuration schemes exist, depending of VNA type and manufacturer.

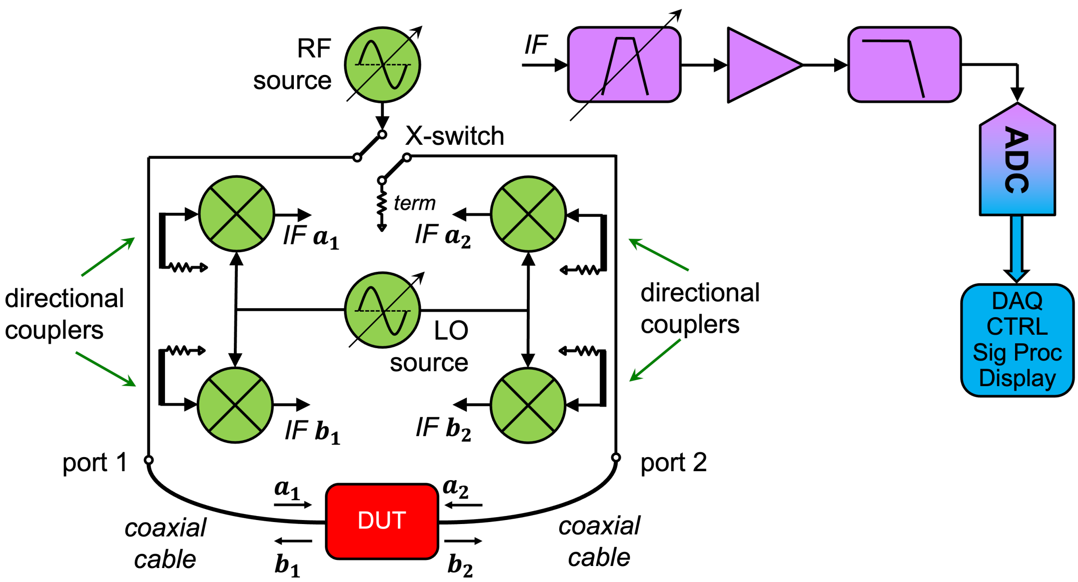


Fig. II.2.195: Simplified block schematic of a vector network analyser.

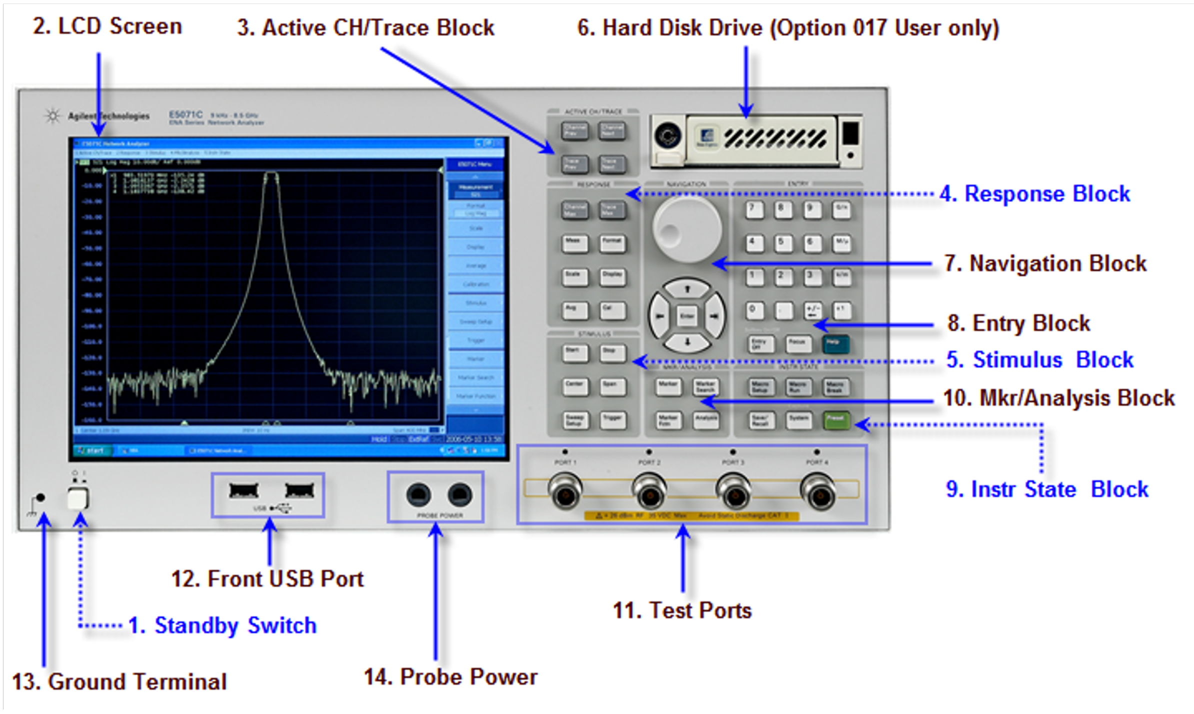


Fig. II.2.196: Front-panel with ports, display and input buttons of a typical VNA.

The separation of forward a_i and backward b_i travelling to the ports of the DUT can be performed by directional couplers, as indicated in Fig. II.2.195. However, some VNAs span a very large frequency range, starting at a few kHz or MHz, and spanning up to $100\,\mathrm{GHz}$ and more! It is difficult to impossible to make a directional coupler for such a wide frequency range, therefore, resistive bridge couplers or other proprietary techniques are used to facilitate that challenge.

Figure II.2.196 shows the front-panel of a typical VNA, here with four test ports using N-type coaxial connectors. The look and feel of the VNA varies a bit between the manufacturers, but it is similar to driving a car, the differences are manageable and in some aspects the operation of a VNA is similar to that of a spectrum analyzer. Usually there are a few main *menus* to setup and operate the VNA, each can have one or more levels of sub-menus to go into the details. The *stimulus* buttons allows to setup the frequency range, RF power, etc., and with the *measurement* buttons we can select one or more S-parameters to be displayed in the format of interest, e.g. S_{11} in a *Smith*-chart format and / or S_{21} in magnitude and phase. Other menus allow to set markers, automatic measurements, the display, etc.

Beside the typical VNA shown in Fig. II.2.196, there exist a class of very compact, so-called *USB-VNAs* with similar, or even more advanced performance. These instruments have no front-panel and are entirely operated via a laptop computer.

II.2.9.3.4.3 VNA calibration

The hardware the even most advanced, expensive “ultra-modern” VNA is not perfect, e.g. the internal source is not perfectly matched to $50\,\Omega$ port impedance – over the entire frequency range –, its internal directional couplers or bridge networks have a finite directivity, since there exists no ideal, infinite directivity in practice, and finally, the coaxial cables between VNA and DUT ports have frequency-dependent attenuation (dispersion) effects.

This calls for a calibration to correct all these unwanted, but systematic errors, to guarantee a precise S-parameter characterisation of the DUT itself, independent of most systematic imperfections of the measurement instrument and cables. There are several *calibration procedures* to eliminate some, or even all of the mentioned deficiencies, and while their names differ a bit among the VNA manufacturers, the methods are basically the same. The simplest calibration method is called the *response correction*, and is applied for transmission coefficients S_{ij}. It basically is a S_{ij} (or S_{ji}) transmission measurement of a quasi “zero length” ideal transmission-line, often referred as “thru” calibration, by connecting the two cable ends of the two-port VNA with each other. For the given setting of the VNA, i.e. `Start` / `Stop` frequency, # of frequency `Points`, `Resolution Bandwidth`, `Power` level, etc., magnitude and phase are acquired and stored as calibration *reference data* file $S_{ij\text{ref}}$ in the non-volatile memory of the VNA for each frequency point. Now, a DUT can be connected between the cable ports, with the connectors serving as *reference planes* of the calibrated system setup, i.e. the VNA plus the cables. The VNA is then measuring in *corrected* calibrated mode and continuously performs:

$$S_{ij\text{corr}} = \frac{S_{ij\text{meas}}}{S_{ij\text{ref}}}$$

> Strictly speaking, a “zero-length” *thru* transmission-line is only given for coaxial VNA cables with different genders at the DUT side, i.e. female and male connectors. The use of any coaxial gender adapter will add a small section of uncalibrated, but basically lossless transmission-line, which in many cases can be ignored, except for measurements that require knowledge of the absolute phase $\angle S_{ij}$.

However, this simple calibration procedure eliminates only a few of the twelve systematic errors being present in a 2-port VNA measurement setup, basically the frequency response errors *transmission tracking* of the setup. Other errors, such as the impedance mismatch of the RF source, the impact of the finite directivity in the separation of forward / backward travelling waves, and the impedance mismatch of the test port cables (*reflection tracking*), are still present. Therefore, a more sophisticated, and widely popular calibration technique is applied to the reflection measurements S_{ii}, called the *vector correction*, and is also known as the `Short-Open-Load` (SOL) calibration technique.
This technique corrects for three independent error sources mentioned above and needs to be applied to each VNA test port.

For the vector correction, the VNA applies an internal, mathematical error model, the error adapter, as illustrated in Fig. II.2.197. Without error adapter the reflection coefficient $\Gamma_{iM} = S_{iiM}$ measured by the VNA at DUT port i and the true reflection coefficient of the DUT, $\Gamma_{iDUT} = S_{iiDUT}$, are different,

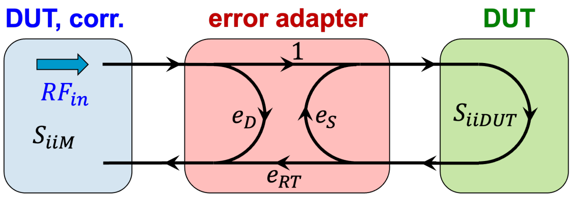


Fig. II.2.197: Error model of the VNA.

Fig. II.2.198: Interpretation of VNA error terms

error term	interpretation
e_D	directivity
e_S	source match
e_{RT}	reflection tracking

$S_{iiM} \neq S_{iiDUT}$. The error adapter is a matrix with three parameters, e_D, e_S, and e_{RT} to correct for the three systematic error mentioned above, see also Fig. II.2.198,:

$$S_{iiM} = e_D + e_{RT} \left(\frac{S_{iiDUT}}{1 - e_S S_{iiDUT}} \right) \tag{II.2.172}$$

The three, frequency dependent unknowns, e_D, e_S, and e_{RT}, in Eq. (II.2.172) are found by a measurement, i.e. the *calibration* procedure, using three *standards*, a short, an open and a load, to be located at the physical reference plane at the end of the VNA cable instead of the VNA port. In other words, we need to carry out three independent calibration measurements for each frequency point, to solve the three coupled equations hidden in Eq. (II.2.172) for three complex unknowns, e_D, e_S, and e_{RT}.

The unknowns of the error network are determined applying a calibration measurement with three different, but known, calibration DUTs. These calibration DUTs do not need to be perfect, but, the electromagnetic properties need to be known to a great precision. The tabulated complex, frequency-dependent S-parameters of the calibration standards are provided by the manufacturer of the calibration hardware (they are often referred as *calibration kit*), and are stored in the VNA memory as calibration kit reference data. Usually the calibration DUTs represent a `Short` circuit, an `Open` circuit, and an impedance matched `Load`, i.e. a 50 Ω termination resistor, enabling the VNA to determine the frequency-dependent error model. However, the error model becomes invalid if different test cables are used, or if the VNA settings are modified, and would require a re-calibration under those circumstances. Now the VNA continuously applies the error correction, Eq. (II.2.172), during the DUT measurement, and the *reference plane* is mathematically “moved” to the end of the test cables. Only the DUT network “behind” the reference plane is taken into account for the measurement.

The impact of the VNA calibration is shown in Fig. II.2.199, which presents a $S_{11} f$ measurement of a high-quality 50 Ω termination, with and without SOL calibration (vector correction). For an ideal termination, no reflection should be present, i.e. $S_{11} = 0 \equiv -\infty$ dB. In this example the calibration of the VNA improves the measurement quality by approximately 20 dB! In case of an ideal short or open, i.e. a total reflection of the incident RF wave a_1, $S_{11} = 1 \equiv 0$ dB), the non-calibrated S_{11} response typically displays a residual with values of a fraction of a dB, up to a few dBs below the 0 dB line; after calibration these errors typically reduces to few milli-decibels.

From the VNA calibration discussion so far we can follow:

Response calibration removes the frequency response errors. Requires a “zero-length” lossless TL `Thru` (T) standard and strictly also an [Isolation] standard. (However, the *isolation calibration* can often be omitted.

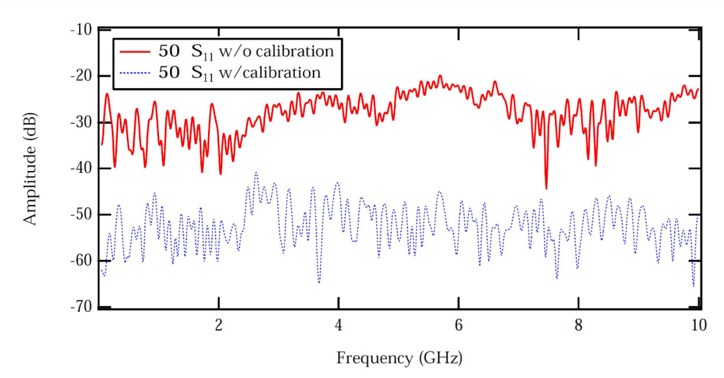


Fig. II.2.199: $S_{11}(f)$ return loss for an almost ideal terminated port, with and without SOL calibration.

1-port calibration removes directivity, source match, and reflection tracking errors by a vector correction method that uses a texttttShort-Open-Load (SOL) calibration.

full n-port calibration removes all systematic errors of the VNA setup by combining response and 1-port calibration, and expanding the method to n-ports. E.g. a *full 2-port calibration* corrects for twelve systematic errors – directivity, source and load match, reflection and transmission tracking, and crosstalk – therefore a 12×12 complex error matrix has to be established for each measured frequency point.

For measurements on devices with popular RF coaxial connectors, e.g. SMA ($3.5\,\mathrm{mm}$) or N-type, calibration standards such as a load termination, an open and a short circuit are available as calibration kits, see the example Fig. II.2.200a. As mentioned, to successfully perform the calibration procedure for the reflection coefficient, the tabulated values, representing the electromagnetic properties of the calibration standards, have to be present in the VNA. Obviously, those tables provided by the manufacturer of the calibration kit do not have an infinite frequency resolution. The instrument applies an interpolation

(a) Mechanical with `Short`, `Open`, `Load`, and accessories. **(b)** 4-ch electronic (USB).

Fig. II.2.200: VNA calibration kits for $3.5\,\mathrm{mm}$ (SMA) connectors.

procedure if the selected frequency points don't match with the tabulated values of the calibration kit.

The calibration technique described so far is a well established industry standard for RF and microwave VNA measurements. However, it has a substantial disadvantage for the user: it is tedious and time consuming, in particular for the calibration of a multiport VNA setup that requires three, four or more ports.
Already for the full 2-port calibration requires eight different connections of a calibration standard to satisfy the 12-term error model. The manual procedure of connection and de-connection of the calibration standards and the correct interaction with the VNA calibration procedure is time consuming and error prone. The situation becomes even worse when performing a full four-port calibration, requiring 32 connections and de-connections of standards! For this reason, the *electronic calibration kit* method is available and now very popular. For this procedure, each port is connected via the measurement cable to the electronic calibration box (shown in Fig. II.2.200b), which holds the different calibration standards, and switches them automatically controlled by the VNA. This method enables to perform a full four-port calibration in less than a minute.

Any VNA measurement result without proper calibration of the setup is questionable!

II.2.9.3.4.4 Options and features

The vector network analyzer comes with a set of standard options and features, which can be expanded with even more options to be purchased.

Standard features typical, but incomplete list:

- One or more S-parameters can be visualised simultaneously on the display, in the same, or in separate graphs.
- There is a large variety of formats available, in *Cartesian* coordinates (magnitude and phase, real and imaginary part, group delay, etc.), but also in polar coordinates or as *Smith*-chart display, for S_{ii} measurements.
- The data can be displayed in linear or logarithmic ($\log_{10}$) scaling.
- VNA settings include `Start` and `Stop`, or `Center` and `Span` frequencies, `IF Bandwidth`, number of frequency `Points`, port `Power`, to name a few.
- Instead of a frequency sweep the VNA can perform a *power sweep* at a given frequency, e.g. to simplify the $1\,\mathrm{dB}$ compression point measurement of a RF amplifier.
- Frequency sweep points are usually placed equidistant along the linear scaled frequency axis, but can also be spread logarithmic, or in segments of different spacing.
- Interpolation techniques are applied whenever possible to minimise re-calibration efforts.
- Marker and marker functions enable simple measurements, e.g. $3\,\mathrm{dB}$ bandwidth of a band-pass filter, markers on the *Smith*-chart provide additional information like R and L or C values of a series / parallel circuit equivalent impedance.
- A variety of trace data transformations, most popular is $Z \Leftrightarrow \Gamma$.

- A variety of advanced calibration and data manipulation procedures, like adding offsets or factors to the measured data. Popular is the *port extension* that allow the calibrated reference plane to be "moved" mathematically, defined by a piece of transmission-line, e.g. to account for inaccessible TL hardware.
- A variety of VNA settings, display (picture capture) and data file formats. Most popular is the `SnP` *Touchstone* file format for the S-parameters, see also Section II.2.7.8.

> Please note, the VNA always performs the measurement of all S-parameters simultaneously, as defined by the calibration, regardless of what measurement and format is on the display. Therefore, the `SnP` data file includes all S-parameters in form of complex, calibrated values vs. frequency, and not only what is on the VNA display!

Hardware options a few examples:

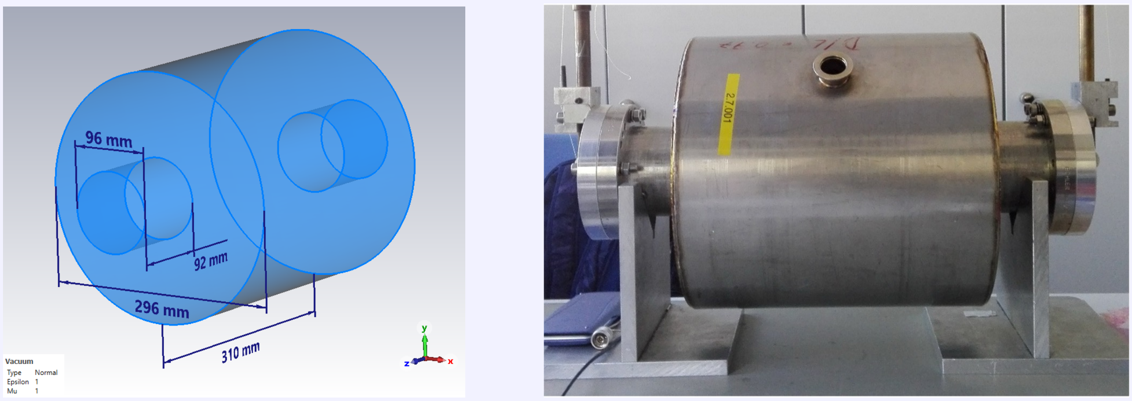
- Four, ore more more test ports. The 4-port VNA is quite popular and offers an entire set of additional features, based of *virtual ports*. Virtual ports enable, e.g. balanced-to-unbalanced (balun) transformations (and vice versa) and therefore the characterisation of coupled TL properties, e.g. Z_{0e}, Z_0o, see also paragraph II.2.8.1.7.4.
- Additional 2$^{\text{nd}}$ RF source that enables 3-port measurements on mixers and up-/down-converters.
- Integrated spectrum analyzer functionality.
- Automatic calibration system down to DC.

Software options It is impossible to list the many software options! Like for the spectrum analyzer, most are tailored to the telecommunications industry. Of more general interest are the *noise figure* measurement option and the options related to the *time domain* transformation, discussed in the next paragraph.

Example II.2.9.1: Characterisation and measurements of a simple "pill-box" cavity

The characterisation of eigen-modes of of a resonant cavity by a VNA-based measurement is one of the most fundamental task in accelerator RF engineering, and will be presented in this example in some detail.

> Here we cover only *normal conducting* cavities, *superconducting* accelerating structures are very different "animals" and require highly tailored measurement methods. Furthermore, we focus on a simple, single-cell cylindrical "pill-box" cavity, and here on the characterisation of it's $TM010$ eigen-mode by VNA measurements.



(a) 3D geometry. **(b)** Photo of the cavity.

Fig. II.2.201: A simple "pillbox" cavity.

Analytical calculations and theory

Table II.2.14: Dimensions of the cylindrical pill-box cavity with beam-ports.

pill-box cavity dimensions		
cavity radius	a	148 mm
cavity height	h	310 mm
beam-port radius	a'	48 mm
beam-port extension	h'	92 mm

Figure II.2.207 shows he geometry with dimension and a photo of the cavity, the dimensions are summarised in Table II.2.14.

Applying a modal expansion to the boundary problem, here as ideal cylinder **without** the *beam-ports*, the transverse magnetic modes propagate at cut-off frequencies given by Eq. (II.2.173), while the frequencies of the transverse electric modes follow Eq. (II.2.174):

$$f_{TMnml} = \frac{c}{2\pi}\sqrt{\left(\frac{p_{nm}}{a}\right)^2 + \left(\frac{\pi l}{h}\right)^2} \tag{II.2.173}$$

$$f_{TEnml} = \frac{c}{2\pi}\sqrt{\left(\frac{p'_{nm}}{a}\right)^2 + \left(\frac{\pi l}{h}\right)^2} \tag{II.2.174}$$

where p_{nm} and p'_{nm} are the m roots of the *Bessel* function $J_n(x)$ and the derivative *Bessel* function $J'_n(x)$, respectively. Table II.2.15 shows the first two roots of these functions, which are of interest for our cavity, in our case these two lowest modes are the $TE111$ and the $TM010$ eigen-modes. Theoretically, their cut-off frequencies are 766 MHz and 776 MHz, but the exact values may vary due to inaccuracies of the dimensions and geometric variations (like the beam-ports).

Table II.2.15: Some roots of the *Bessel* functions.

m	***Bessel* Function**		**Derivative *Bessel* Function**	
	$J_0(x)$	$J_1(x)$	$J'_0(x)$	$J'_1(x)$
1	2.4048	3.8317	3.8317	1.8412
2	5.5201	7.0156	7.0156	5.3314

The mode of particular interest is the $TM010$ mode, as it is the one which has an E-field along the z-axis, which enables to accelerate the charged particles of the beam.
Another important parameter which characterises the "quality" of the eigen-mode, and can be used to compare cavities of same geometry but made out of different materials, is the *quality factor*, or Q-factor. It is defined by the losses of the cavity compared to the stored energy,

$$Q = 2\pi \frac{\text{energy stored}}{\text{energy dissipated per cycle}} = \omega_{res} \frac{\text{energy stored}}{\text{energy loss/second}} = \frac{\omega_{res} U}{P_d} = \frac{f_{res}}{f_{BW}} \tag{II.2.175}$$

and can be calculated by measuring the bandwidth relative to its central frequency, as shown in Eq. (II.2.175), where f_{res} is the resonant frequency of the particular eigen-mode to be analysed, and $f_{BW} = f_{+3\,\mathrm{dB}} - f_{-3\,\mathrm{dB}}$ is it's $3\,\mathrm{dB}$ bandwidth, see also Definition II.2.8.1. The Q-factor depends on the materials of the cavity walls (conductive losses), and on losses due to fillings or open boundaries (dielectric and radiation losses), but is also impacted by an external network, like in our case, where a measurement setup is attached. The unloaded Q-factor, Q_0, considers only out of the intrinsic losses, while the loaded-Q, Q_L, takes into account also the external loading effects, and this is the value that is actually measured!

$$\frac{1}{Q_L} = \frac{1}{Q_0} + \frac{1}{Q_{ext}} \tag{II.2.176}$$

They are related as shown in Eq. (II.2.176), where Q_{ext} reflects the Q-factor due to the external network, e.g. the RF generator with its source impedance and the coupling circuit. For an ideal cylindrical cavity (without beam-ports), Q_0 can be calculated analytically

$$Q_0 = \frac{a}{\delta}\left(1 + \frac{a}{h}\right)^{-1} \tag{II.2.177}$$

$$\text{with:} \quad \delta = \sqrt{\frac{2}{\omega_{res}\sigma\mu}} \tag{II.2.178}$$

following Eq. (II.2.177) – using the skin depth, δ, as defined by Eq. (II.2.178) – where ω_0 is the angular resonant frequency of the eigen-mode, σ is the conductivity and μ is the magnetic permeability. For the cavity under test, considering a conductivity of stainless steel of $\sigma = 1.3 \times 10^6\,\mathrm{S/m}$ and a permeability of $\mu = \mu_r\mu_0$, with $\mu_0 = 4\pi 1 \times 10^{-7}\,\mathrm{H/m}$, and $\mu_r = 1$ for air or vacuum, the analytical calculated Q-factor is $Q_0 = 6321.8$.

The last important figure of merit is the R/Q, called R-over-Q. It describes how efficient the cavity transfers its stored energy to the beam and it is defined as ratio between the shunt impedance, R, and the unloaded Q-factor Q_0, as defined in Eq. (II.2.177), with V_{acc} being the accelerating voltage and U the stored energy:

$$\left(\frac{R}{Q}\right) = \frac{V_{acc}^2/P_d}{2\omega_{res}U/P_d} = \frac{V_{acc}^2}{2\omega_{res}U} \tag{II.2.179}$$

For the $TM010$-mode of the cylindrical cavity follows:

$$\left(\frac{R}{Q}\right) = \frac{4\eta_0}{\chi_{01}^3 \pi J_1^2(\chi_{01})} \frac{\sin^2\left(\frac{\chi_{01}}{2}\frac{h}{a}\right)}{\frac{h}{a}} \approx 128 \frac{\sin^2\left(1.2024\frac{h}{a}\right)}{\frac{h}{a}} \tag{II.2.180}$$

Analytically, for a cylindrical cavity operating at the TM010 mode, the value of R/Q can be determined using Eq. (II.2.180), with η_0 being the characteristic impedance of free space ($\eta_0 = 120\pi\,\Omega$), and χ_{01} the 1$^{\text{st}}$ root of the 0$^{\text{th}}$-order *Bessel* function. For the given dimensions of our cavity follows $R/Q = 10.4\,\Omega$.

Table II.2.16: Analytically derived parameters of our cylindrical cavity without beam-ports.

parameter	**value**
f_{TE111}	766 MHz
f_{TM010}	776 MHz
Q_0 ($TM010$)	6321.8
R/Q ($TM010$)	10.4 Ω

Table II.2.16 summarises the parameters acquired analytically for our idealised cylindrical cavity. Please note $f_{TM010} > f_{TE111}$, which strictly speaking makes it **not** a pill-box cavity, as this requires f_{TM010} to be the lowest mode!

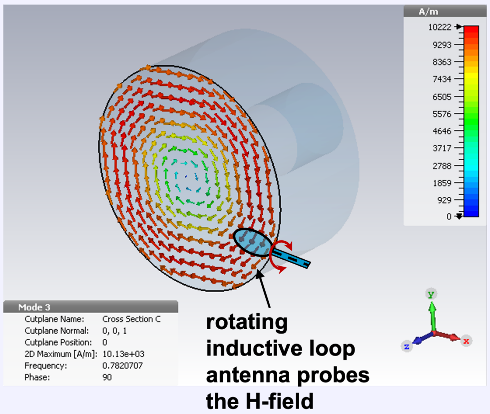
In summary, please note the three important parameters of the $TM010$ accelerating mode to be characterised:

$\boldsymbol{f_{res}}$ The resonant frequency f_{res} of the eigen-mode, here the $TM010$-mode, given by the size of the cavity.

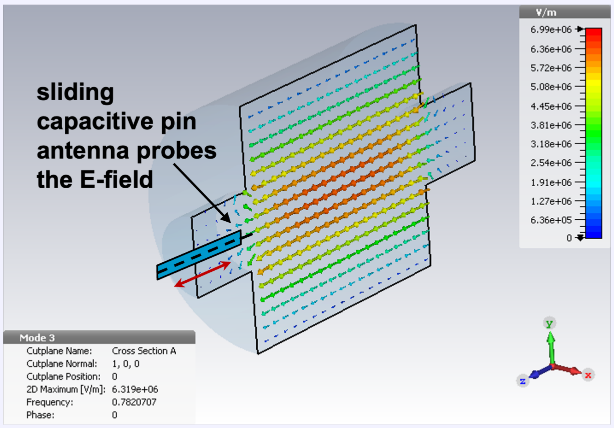
$\boldsymbol{Q_0}$ The unloaded Q-factor Q_0 of the eigen-mode, given by the material properties of the cavity.

$\boldsymbol{R/Q}$ The equivalent characteristic impedance R/Q of the eigen-mode, given by geometric shape of the cavity.

VNA measurement setup



(a) Inductive loop antenna (H-field coupling).



(b) Capacitive pin antenna (E-field coupling).

Fig. II.2.202: Coupling to the $TM010$-mode.

A resonant cavity for particle acceleration needs to have one or more *coupling ports* to feed the RF energy and for measurement purposes. Our cylindrical "demo" cavity (Fig. II.2.201b) is equipped with several coupling ports at different locations for measurement purposes. Figure II.2.207 illustrates the two coupling options to the $TM010$-mode, i.e. to the magnetic field using an inductive operating *loop-antenna* located at the cavity rim (Fig. II.2.202a), or to the electric field using a capacitive operating *pin-antenna*, e.g. slide in through the beam-port (Fig. II.2.202b). Figure II.2.203 shows some probe antennas use in our VNA measurement setup, on the left three loop-antennas of different size, on the right a small pin-antenna. In practice the loop-antenna is preferred as it does not interfere with the beam-ports and by using an appropriate, simple mechanical setup, the loop can be rotated, changing the effective cross-area to the $TM010$-mode E-field, and therefore modifying the coupling to the mode.

Fig. II.2.203: Probe antennas for the pill-box cavity characterisation.

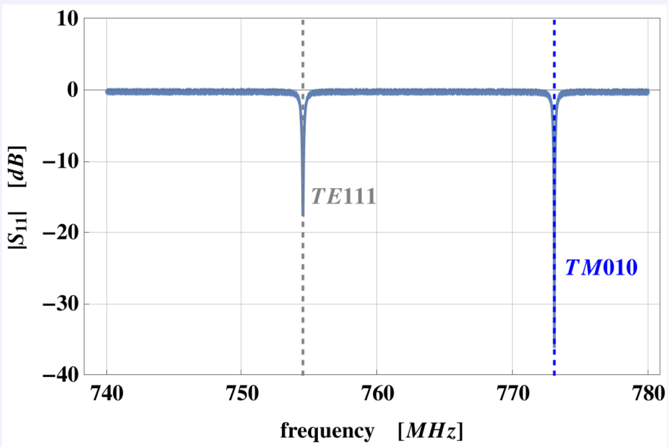


Fig. II.2.204: $|S_{11}|(f)$ VNA measurement of the first two modes of the cylindrical pill-box cavity of Fig. II.2.201a.

We basically have two options for the VNA measurement characterisation of the $TM010$ mode of the pill-box cavity:

- A S_{11} reflection measurement using a single probe antenna.
- A S_{21} (or S_{12}) transmission measurement, which requires the installation and tuning of two probe antennas.

For the characterisation of the mode frequency $f_{res} = f_{TM010}$ and the related unloaded Q-factor, Q_0, the S_{11} reflection coefficient measurement is more meaningful and simple, and therefore the preferred choice. For the measurement of the R/Q both methods will be discussed.

After setting up all the measurement hardware, i.e. the VNA, the pill-box cavity with the loop-antenna installed and rotated such that it's area is perpendicular to the $TM010$-mode H-field (in the xy-plane) for high coupling, and the cabling, an initial verification of the first two eigen-modes is performed by a quick $|S_{11}|(f)$ measurement in the range $740\,\text{MHz} < f < 780\,\text{MHz}$, see Fig. II.2.204. For this initial measurement no VNA calibration or "fancy" setup parameters are required, instead, we need to verify which of the two resonances displayed on the trace belongs to the $TM010$-mode? Therefore, a little *perturbation* is required, e.g. by inserting a metallic rod through the beam-port along the z-axis into the cavity. The metallic object will couple to the longitudinal E-field of the $TM010$-mode, but not to the $TE111$-mode, as the latter one has no longitudinal E-field components on or near the z-axis. As a result, the $TM010$ resonance on the $|S_{11}(f)|$ VNA trace will change when inserting the rod perturbation, as a wider resonance, also shifting its resonance frequency, while the $TE111$ stays unchanged. For this measurement procedure, the *trace memory* of the VNA is a very helpful tool to better identify changes of the setup during the measurement. We now have identified the $TM010$ eigen-mode, which has a resonance frequency of $f_{res} = 773.18\,\text{MHz}$, see also Fig. II.2.204.

Measurement of f_{res} and Q_0

With the $TM010$-mode now clearly identified, we can "zoom-in" by setting the VNA center frequency `Center` to $f_{res} = 773.18\,\text{MHz}$ and to a reasonable narrow `Span`, e.g. to $4\,\text{MHz}$, see

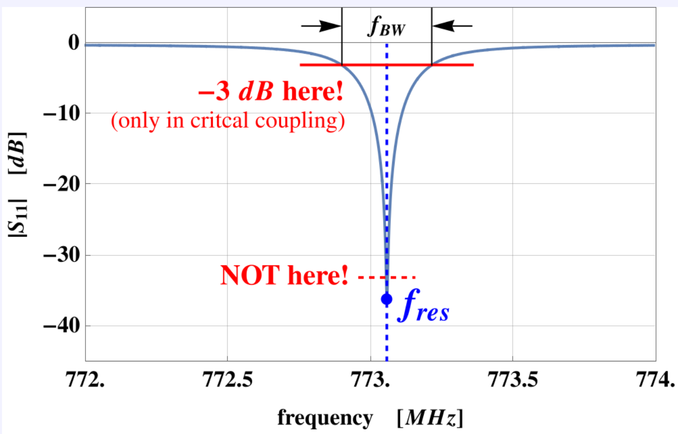


Fig. II.2.205: $|S_{11}(f)|$ of the $TM010$-mode.

Fig. II.2.205. It also is advisable to increase the default value of the number of frequency `Points` for the measurement, e.g. to 8001 points or an even higher value, if the instrument permits.

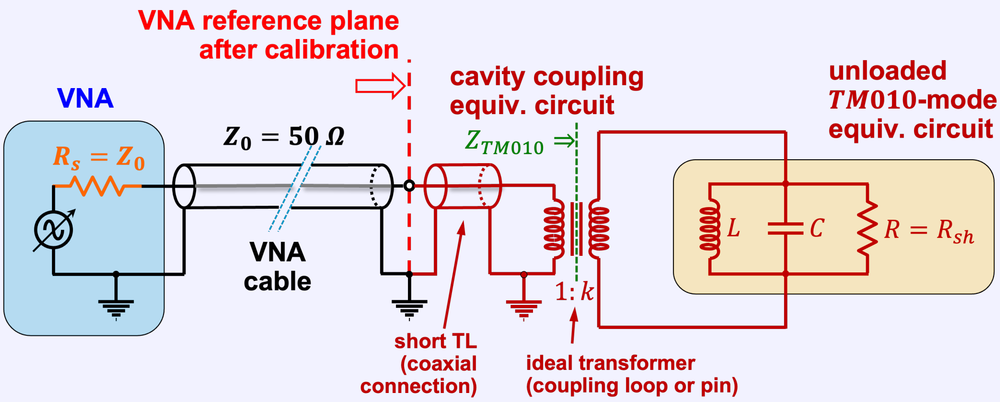


Fig. II.2.206: Equivalent circuit of the $S_{11}(f)$ $TM010$-mode VNA measurement setup.

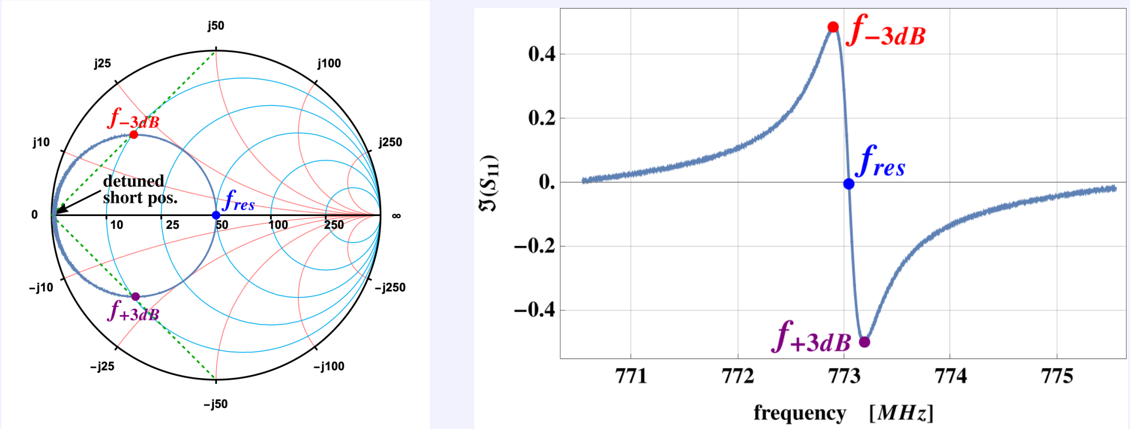
For the required VNA 1-port calibration, as well as for the following measurement procedure, the equivalent circuit of the setup, with the $TM010$ resonance as RLC parallel circuit, helps to understand the details, see Fig. II.2.206. The goal: the measurement of the *unperturbed*, intrinsic $TM010$-mode parameters f_{res} and Q_0, highlighted as the unloaded RLC parallel equivalent circuit part in Fig. II.2.206. Evidently, the coupling loop-antenna connected to the VNA port alters that picture, the cavity is always *loaded* by source impedance R_s of the connected setup, here the VNA. The probe antenna is represented in the equivalent circuit as *ideal transformer*, thus transforming the source impedance $R_s k^2 = R_{sh}$ to the shunt impedance of the resonance. To complete the picture, a small piece of transmission-line is added into the overall equivalent circuit Fig. II.2.206, representing the coaxial connection of the loop-antenna, which in our setup can be up to a few cm long. Also indicated in Fig. II.2.206 is the *reference plane* for the VNA calibration at the end of the VNA cable, which connects to the coaxial connector of the loop-antenna.

With the cavity loaded by the attached VNA source impedance, the most simple and elegant way to retrieve the unloaded-Q, Q_0, is to bring the resonance into *critical coupling*, i.e. arranging the coupling coefficient k of the probe antenna such that

$$Q_0 = Q_{ext} \quad \Rightarrow \quad Q_0 = 2Q_L \qquad \text{(critical coupling)} \tag{II.2.181}$$

This means, at the resonant frequency, f_{res}, the resonator as load to the VNA is impedance matched, and we have maximum power transfer. The straightforward way to achieve the critical coupling condition Eq. (II.2.181) is to display and observe the measured $S_{11}(f)$ while slowly rotating the coupling loop-antenna until the displayed circle at f_{res} precisely hits the $Z = 50\,\Omega$ match point in the center of the *Smith* chart, see also Fig. II.2.207a.

Now the setup measures the $TM010$-mode in critical coupling and e.g. taking f_{BW} from the modulus $|S_{11}(f)|$ displayed on the VNA screen – with help of the VNA trace markers – and using Eqs. (II.2.175) and (II.2.181) returns $Q_L = Q_0/2$, see also Fig. II.2.205.



(a) $S_{11}(f)$ in *Smith* chart format.

(b) Imaginary part $\mathrm{Im}[S_{11}(f)]$.

Fig. II.2.207: $TM010$-mode measured in critical coupling.

Still, the S_{11} measurement is not perfect! As we did not account for the small piece of transmission-line of the probe antenna connector, the parametric impedance circle of our $S_{11}(f)$ measurement displayed in the *Smith* chart format is not located on the left side of the chart, symmetrically wrt. the real axis, at the so-called *detuned short position*, see also Fig. II.2.207a. Instead, the circle is somehow at a "rotated" location on the chart, which requires to be corrected. As we cannot include the probe antenna connector into the VNA calibration, we need to apply a *port extension* or an *electrical delay*, i.e. a mathematical compensation for this small piece of uncalibrated, lossless transmission-line of some $\sim 100\,\mathrm{ps}$ delay. We can estimate the delay-time at put the values into the VNA as `Port Extension` or `Electrical Delay`, and then verify the trace of the imaginary part $\mathrm{Im}[S_{11}(f)]$ over a large frequency `Span` being perfectly horizontal, it should have no slope. This procedure probably requires an extra VNA calibration step, as we have to change the frequency `Span` setting. Now, going back to the *Smith* chart format on the VNA, with the previous frequency `Span` and VNA calibration setting, the measured circle should be at or near the detuned open position. A fine adjustment using the VNA *phase offset* adds or subtracts some $\Delta\angle S_{11}$ to the measured $\angle S_{11}(f)$ argument, thus rotates the circle in the *Smith* chart.

It may take some iterations to set and fine-tune all parameters of the VNA and adjust the loop-antenna correctly, but finally the VNA display should look like Fig. II.2.207. For convenience, we can display the $S_{11}(f)$ *Smith* chart format and the $\mathrm{Im}\,S_{11}(f)$ imaginary format simultaneously. With help of the VNA trace `Markers` and `Marker Search` functions it is simple to extract f_{BW} and calculate Q_0. From the $\mathrm{Im}\,S_{11}(f)$ display we can set *minimum* and *maximum* `Markers` on the trace to located $f_{+3\,dB}$ and $f_{+3\,dB}$, and a 3rd `Marker` to the zero-crossing to locate f_{res}, see

Fig. II.2.207b. We then calculate:

$$Q_L = \frac{f_{res}}{f_{+3\,dB} - f_{-3\,dB}}$$

$$Q_0 = 2\frac{\sqrt{f_{+3\,dB}f_{-3\,dB}}}{f_{+3\,dB} - f_{-3\,dB}} \approx \frac{f_{+3\,dB} + f_{-3\,dB}}{f_{+3\,dB} - f_{-3\,dB}}$$

In a similar fashion we can use `Markers` on the $S_{11}f$ *Smith* chart format, which can be set to display the values e.g. frequency, and real and imaginary part of $Z(f)$, which are updated as the marker is moved along the measured S_{11} circle, i.e. the frequency points of the VNA measurement. To get f_{BW} for Q_L we need to located two `Markers` to the minimum and maximum value of ImS_{11}, plus a a 3rd `Marker` has to be set to $Z = (50 + j0)\,\Omega \;\equiv\; \Gamma = 0$ for the match-point at f_{res}, see Fig. II.2.207a.

> As mentioned before, a sufficient number of measurement points is important to ensure the markers can be set to the exact locations!

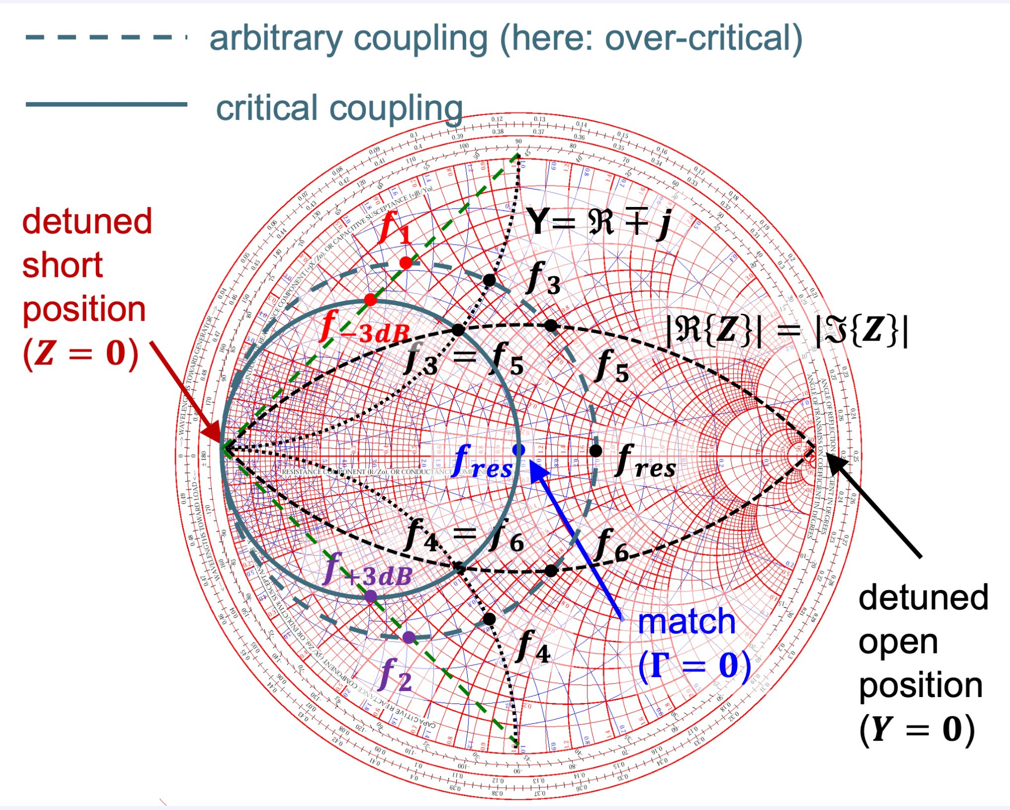


Fig. II.2.208: *Smith* chart for a critical and an over-critical coupled resonance.

Sometimes the coupling is fixed, it cannot be changed by rotating or sliding a probe antenna or is fixed and defined by a measurement using the main coupler. In that case the VNA S_{11}

measurement results in an arbitrary coupling:

$$
\begin{aligned}
O_0 &> Q_{ext} \qquad \text{over-critical coupling} \\
O_0 &< Q_{ext} \qquad \text{under-critical coupling}
\end{aligned}
$$

Figure II.2.208 illustrates the situation for an *over-critical* coupled mode (dashed circle) in comparison to critical coupling of the mode (continuous circle), always assuming the VNA setup is calibrated, with the measured circle in the *Smith* chart aligned to the detuned short position by applying appropriate adjustments for electrical length or port extension and phase offset. Thanks to the various `Marker` display value options of the VNA, we can not only calculate Q_0 for an arbitrary coupled mode from the *Smith* chart format, but also Q_L and Q_{ext}. The trace `Markers` of the VNA allow to display – beside the frequency f itself or other parametric values – the complex load impedance $\mathrm{Re}(Z) + j\,\mathrm{Im}(Z)$ (or admittance $\mathrm{Re}(Y) + j\,\mathrm{Im}(Y)$), transformed by k^2, or the complex reflection coefficient$Re(\Gamma) + j\,\mathrm{Im}(\Gamma)$. One of the three markers should always be set onto the real axis of the *Smith* chart, its frequency value returns f_{res}. Two additional markers set to f_n and f_m are used to defined a 3 dB-bandwidth $f_{BW} = f_m - f_n$, which translates with f_{res} into a Q-value, Eq. (II.2.175), for the loaded, unloaded or external Q-factor depending on the marker location as shown in Fig. II.2.208 and summarised in Table II.2.17.

Table II.2.17: Q-value calculation from the marker values in the S_{11} *Smith* chart format.

marker $f_{n,m}$	**set marker value to**	**Q-value calculation**
$f_{1,2}$, $(f_{-3\,dB}, f_{+3\,dB})$	$\lvert\mathrm{Im}(S_{11})\rvert = \lvert\mathrm{Im}(\Gamma)\rvert = max.$	$Q_L = \frac{f_{res}}{f_2 - f_1}$
$f_{3,4}$	$Y = (\mathrm{Re} \pm j)/Z_0$ [a]	$Q_{ext} = \frac{f_{res}}{f_4 - f_3}$
$f_{5,6}$	$\lvert\mathrm{Re}(Z)\rvert = \lvert\mathrm{Im}(Z)\rvert$	$Q_0 = \frac{f_{res}}{f_6 - f_5}$

[a] Re = any real part value.

The values found from the VNA measurements, $f_{res} = 773.058$ MHz and $Q_0 = 4900$, differ from the theoretical values due to the beam- and probe-ports, some mechanical imperfections and tolerances, as well as unknowns of the exact material properties.

Measurement of the R/Q

The R/Q is a geometric parameter of the cavity and was introduced earlier in this exercise.

$$
V_{acc} = \left| \int dz \mathbf{E}(z) \cos\left(\frac{\omega z}{\beta c}\right) \right| \tag{II.2.182}
$$

The accelerating voltage V_{acc} in Eq. (II.2.179) is given as integral of the longitudinal electric field components $E_{\|}$ for a given eigen-mode - here the $TM010$-mode – along the path of the beam through the cavity, Eq. (II.2.182), i.e. on the z-axis ($x = y = 0$), which usually is the symmetry

axis of the resonant structure. In Eq. (II.2.182) the factor $\cos\left(\frac{\omega z}{\beta c}\right)$ is related to the *transit time* it takes a particle of velocity βc to pass the cavity gap.
The measurement of $E_{\|}(z)$, and therefore of $V_{acc} \Rightarrow R/Q$, is based on a *perturbation method* that slides a small object through the cavity, perturbing the EM-field resonance which can be observed with the VNA as frequency shift $\Delta f = f - f_{res}$ [b].

$$\frac{f - f_{res}}{f_{res}} = \frac{\Delta f}{f_{res}} = \frac{1}{U}\left[\mu_0\left(k_{\|}^{H}|H_{\|}|^2 + k_{\perp}^{H}|H_{\perp}|^2\right) - \varepsilon_0\left(k_{\|}^{E}|H_{\|}|^2 + k_{\perp}^{E}|H_{\perp}|^2\right)\right] \tag{II.2.183}$$

Equation (II.2.183) gives the relative frequency shift, knows as *Slater*'s perturbation theorem, with $k_{\|}^{H}$, $k_{\perp}^{H}$, $k_{\|}^{E}$ and $k_{\perp}^{E}$ being the coefficients proportional to the magnetic or electric "polarisability" of the object, acting on the related EM-field component. For our analysis, an elongated metallic object is preferred, such that it perturbs only the dominant longitudinal $E_{\|}$-field of the $TM010$-mode on the z axis, which simplifies Eq. (II.2.183) to:

$$\mathbf{E}(z) = E_{\|}(z) = \sqrt{U\frac{\Delta f(z)}{f_{res}}\frac{-1}{k_{\|}^{E}\varepsilon_0}} \tag{II.2.184}$$

with the electric field $E_{\|}$ being normalised to the root of the stored energy, U. The coefficient $k_{\|}^{E}$ in Eq. (II.2.184) depends on the geometry of the perturbing object. For practical reasons we use a syringe needle which can be approximated as ellipsoid of half length l and radius r, following Eq. (II.2.185), see also [28].

$$k_{\|}^{E} = \frac{\pi}{3}l^3\left[\sinh^{-1}\left(\frac{2}{3\pi}\frac{l}{a}\right)\right]^{-1} \tag{II.2.185}$$

Again, please note in Eq. (II.2.185)) l refers to the half of the length of the object!

The syringe needle used in our measurements had a total length of $10\,\mathrm{mm}$, therefore $l = 5\,\mathrm{mm}$ and a radius $a = 0.59\,\mathrm{mm}$, therefore follows $k_{\|}^{E} \approx 9.65 \times 10^{-8}\,\mathrm{m}^3$.

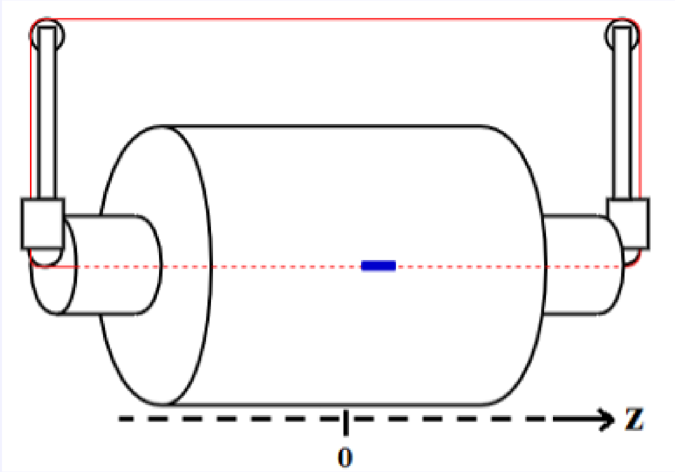


Fig. II.2.209: Schematic of the bead-pull measurement,
indicating the wire in red and the perturbation needle in blue.

Figures II.2.209 and II.2.210 illustrate our simple, manual bead-pull measurement setup. The syringe needle as perturbation object was fixed on a thin, non-metallic fishing line, and a ruler was used to locate its position in the cavity. We opted to close the beam-ports with metal-flanges, just with a little hole for the wire, which present a better defined boundary.
There are two bead-pull methods to measure the frequency shift:

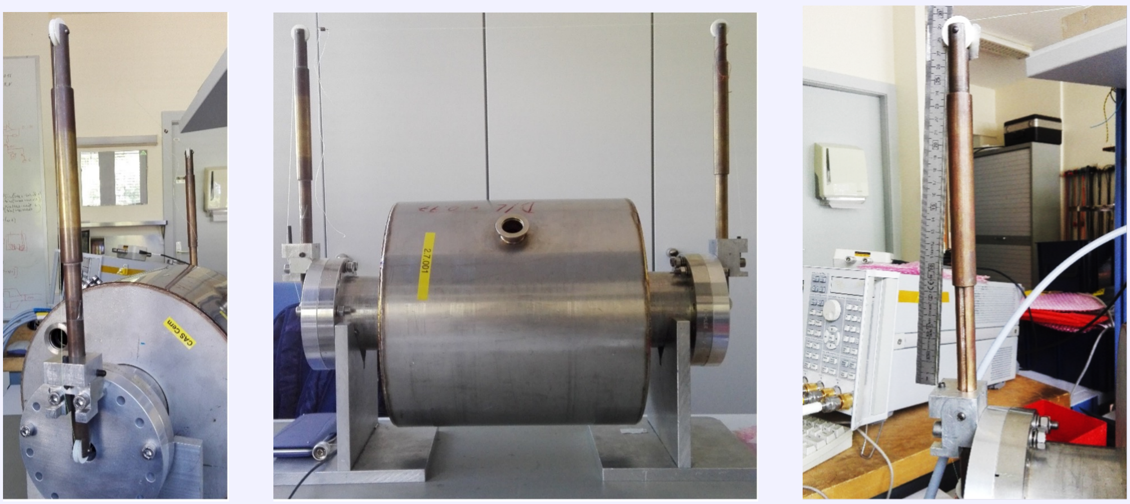
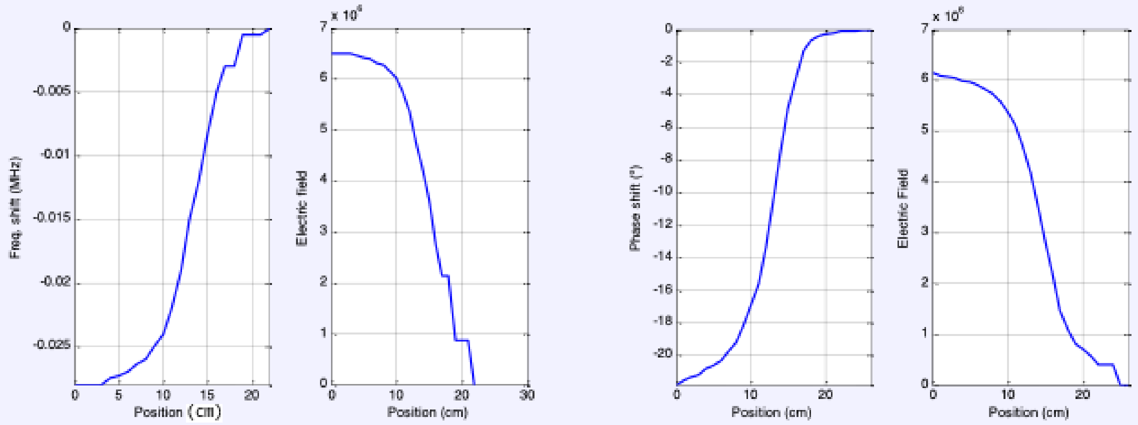


Fig. II.2.210: Bead-pull measurement setup with the pill-box cavity (center),
and detail of the wire pulling structure (left) and the ruler to located the bead-pull position (right).

Direct frequency shift measurement using the S_{11} reflection coefficient Using a single magnetic loop probe, first the resonant frequency has to be measured without the perturbing object. Then, the object is pulled through the cavity, and the frequency shift is monitored, e.g., in steps of $1\,\mathrm{cm}$. Because of the symmetry of the cavity, it is sufficient to probe only half of the cavity. The results are shown in Fig. II.2.211a. From the integration of the electrical field, which follows from Eq. (II.2.182), a value $R/Q = 15.5\,\Omega$ was obtained.



(a) S_{11} frequency shift measurement. **(b)** S_{21} phase shift measurement.

Fig. II.2.211: Measured frequency shift (left) and calculated electric field (right).

Phase shift measurement based on the S_{21} transmission coefficient The second method is an indirect measurement, were the frequency shift is obtained by a phase shift of a transmitted signal, i.e. a S_{21} measurement precisely at the resonance frequency f_{res}:

$$\frac{f(z) - f_{res}}{f_{res}} = \frac{1}{2Q_L} \tan \phi(z) \tag{II.2.186}$$

with $\tan \phi(z) = \angle S_{21} @ f_{res}$ being the measured S_{21} phase as the bead is pulled through the structure along the z-axis. Therefore, two magnetic loop probes are used.

It is of utmost importance to precisely excite the $TM010$-mode at the resonance frequency (center frequency f_{res}), as all following measurement steps depend on it. Additionally, both probes should be tuned for a very weak coupling, approximately $-0.5\,\mathrm{dB}$ peak amplitude for $|S_{11}|$ and $|S_{22}|$.

Once the exact value of the $TM010$ resonance frequency is determined, the center frequency of the VNA is set to f_{res}, and the VNA is configured for a *zero-span* frequency ($0\,\mathrm{Hz}$) to excite and analyse only at that particular frequency. With the perturbation needle at the out of the cavity position, a *reference phase* of S_{21} is measured and, if necessary, a phase offset may added to set it to 0 degrees. Similar to the direct frequency shift measurement method, the syringe needle is pulled step-wise through the cavity, now monitoring the phase in steps of $1\,\mathrm{cm}$. The results are shown in Fig. II.2.211b, in this case $R/Q = 16.5\,\Omega$ is obtained, a similar value to the previous method.

It has been observed that phase drifts over time, most likely due to temperature variations, may tarnish the measurement result. To minimise this influence, the bead-pull phase measurement has to be performed quickly.

In practice, the $\angle S_{21}$ phase shift method is preferred. Utilising a computer controlled setup with motorised pulling mechanics of the bead, several sets of $\angle S21$ phase measurements can be performed, with some small offset in frequency above and below f_{res} in a range

$$0.5 < \frac{\Omega_+ - \Omega_-}{2} < 5 \qquad \text{where} \qquad \Omega = Q_L \left(\frac{f}{f_{res}} - \frac{f_{res}}{f} \right) \tag{II.2.187}$$

Measuring $|S21(f_\pm)|$, evaluating

$$|S21(f_+)| - |S21(f_-)| = g(\tilde{f_{res}}) \tag{II.2.188}$$

and minimising $||S21(f_+)| - |S21(f_-)||$ through several bead-pull sweeps by changing f_+ and f_- until $\tilde{f_{res}}$ does not vary ensures the measurement is performed exactly at $min.\tilde{f_{res}} \simeq f_{res}$. Moreover, synchronising the VNA `Zero Span` sweep and setting the `Sweep Time` to the value of the pull time of the bead through the structure, will already give a qualitatively display $E_{||}(z)$ on the VNA $\angle S_{21}$ trace.

[b] Please note, Δf is defined and used in different ways throughout this chapter!

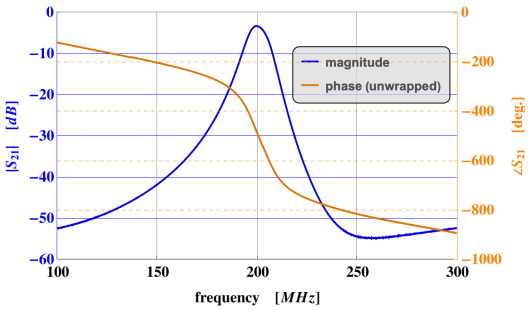
II.2.9.3.4.5 Time-domain transformation (synthetic pulse technique)

For any linear system, the frequency-domain information (data) can be converted to the time-domain by an inverse *Fourier* transformation[11] and vice versa, assuming the vector data (magnitude and phase, or real and imaginary) is present over the entire frequency range of interest. This is the basis of the *time-domain* measurement, also referred as *synthetic pulse technique*, today available as option on most VNAs. It was commercially introduced by *Hewlett-Packard* in the 1980s for their network analyzer applications.

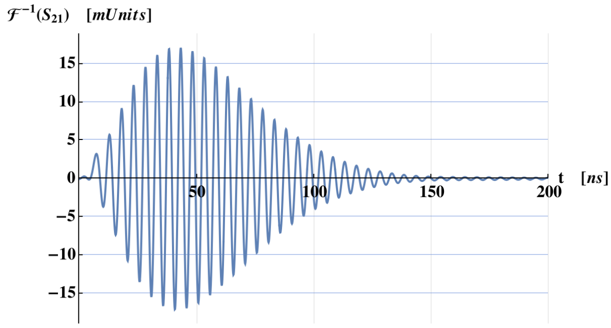
It renders the VNA even more versatile, allowing to display the impulse and/or step response of the DUT, and to perform time-domain reflectometry (TDR) measurements. Typical applications of this measurement techniques are:

- Localising and evaluating discontinuities (faults) in transmission lines.
- Separating the scattering properties of sections of complicated RF networks by time-domain gating.
- Echo cancellation, e.g. in *multipath* environments.
- The synthetic pulse time-domain reflectometry (TDR) is very useful in the prototyping phase of TL-based beam pickups and kickers, as well as for trouble-shooting of accelerator installations,

[11] More precisely: by an inverse discrete *Fourier* transformation (iDFT). The fast *Fourier* transformation (FFT), and its inverse variant iFFT, is an optimised form of the DFT, originally exploiting the symmetry of 2^n data samples, therefore saving computation time. Today the 2^n factorisation requirement is obsolete, and of course, both algorithms will produce the same result for the same input data.



(a) Frequency domain characteristic.



(b) Time domain impulse response.

Fig. II.2.212: S_{21} transmission measurement of a band-pass filter.

e.g. faults in coaxial signal cables and/or connections, and even issues in the accelerator beam-pipe. By using waveguide modes it was successfully used to detect an obstacle in the LHC beam-pipe at CERN.

The important constraint of the applicability of the time domain transformation / synthetic pulse measurement technique, the DUT has to be a *linear* system, and has to be *time-invariant* during the measurement (sweep) time, i.e. it has to be a LTI system.

Figure II.2.212 illustrates the principle of the time domain transformation technique of the VNA, here for a S_{21} transmission measurement of the band-pass filter shown in Fig. II.2.130. On the left side shown, Fig. II.2.212a, the “usual” display of $S_{21}(f)$, formatted in magnitude (dB) and phase (deg , unwrapped), here presented from the stored `SnP` data file. On the right side, Fig. II.2.212b, the same VNA measurement, but now displayed using the build-in time domain `Transformation`, i.e. the inverse discrete *Fourier* transform (iDFT) of the data, $\mathcal{F}^{-1}(S_{21})$, again presented from the stored VNA data file (as `csv` time-series).

Please note, the time domain transformation is implemented differently among the various manufacturers of VNAs. For example, *Kesight* (also *Agilent* and *Hewlett-Packard*) VNAs require the $\mathbf{S}(f)$ measurement to be displayed in `Real` format before activating the `Transformation`, given the fact the time domain data are real values.

Like the spectrum analyzer, the VNA acquires the complex S-parameters only for positive frequencies, $f > 0$. For the iDFT the VNA makes use of the conjugate symmetry $\mathbf{S}(f) = \mathbf{S}^*(-f)$, mandatory for a real signal in the time-domain, $\mathcal{F}^{-1}(\mathbf{S}) \in \mathrm{Re}$. The VNA applies the transformation

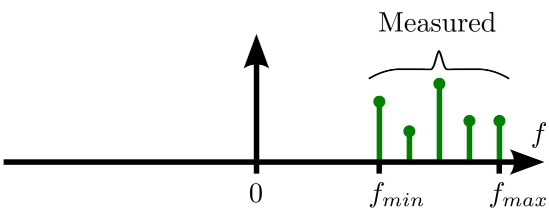
$$s(t) = \mathcal{F}^{-1}[\mathbf{S}(f)] = \int_0^{+\infty} \mathrm{Re}[\mathbf{S}(f)]\cos(2\pi f t) - \mathrm{Im}[\mathbf{S}(f)]\sin(2\pi f t)dt$$

of course as discrete variant of the inverse *Fourier* transform.

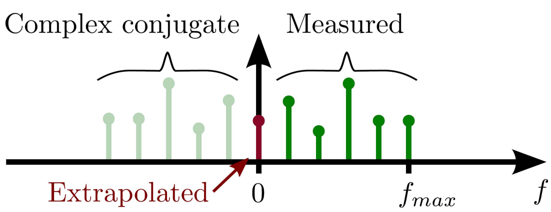
There are a few, but important details linked to the VNA time-domain transformation:

Band-pass mode The VNA samples $\mathbf{S}(f)$ with equidistant steps Δf in a frequency range $f_{\text{min}} < f < f_{\text{max}}$, see also Fig. II.2.213a, a DUT with values of $|\mathbf{S}f|$ reasonable small at the band-ends, f_{min} and f_{max}, i.e. covering the entire pass-band and beyond of band-pass like system. The band-pass mode is the *default* `Transformation` setting in most VNAs, and is applicable for analysing any band-pass like system, as it was performed for the example shown in Fig. II.2.212. Please note the low values of $|S_{21}| < -50\,\text{dB} \approxeq 0.0032$ at the band-ends $f_{\text{min}} = 100\,\text{MHz}$ and $f_{\text{max}} = 300\,\text{MHz}$. In most cases the band-pass mode don't require any "special care".

Low-pass mode The VNA is a RF measurement instrument that does not operate at 0 Hz, direct-current (DC), instead, the lowest operating frequency is typically a few kHz or MHz. This is a problem for analysis of low-pass like systems to be transformed into the time domain, and requires the user to set the `Transformation` of the instrument in `Low-pass` mode. For given settings of `Start` and `Stop` frequency, and the number of `Points`, which defines the equidistant sampling interval $df = (\texttt{Start} - \texttt{Stop})/(\texttt{Points} - 1)$ the VNA will slightly adjust the `Start` and `Stop` frequency settings automatically, if needed, to ensure an extrapolation of the equidistant frequency sampling will exactly fall on $f = 0\,\text{Hz}$, see Fig. II.2.213b. The VNA will then take some of the low frequency samples of $\mathbf{S}(f)$ and its conjugate complex $\mathbf{S}^*(-f)$ to extrapolate the value $\mathbf{S}(f = 0\,\text{Hz})$. A wrong setting of these parameters is usually sensed by the instrument and results in a *warning*, if ignored, the time-domain response may show some unphysical artefacts, like a "crawling" baseline.

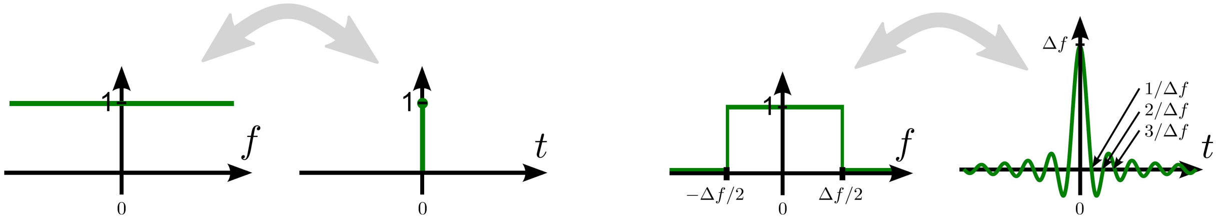


(a) Band-pass mode.



(b) Low-pass mode.

Fig. II.2.213: Sampling of the frequency-domain S-parameter measurement $\mathbf{S}(f)$.



(a) Infinite band. **(b)** Finite band.

Fig. II.2.214: Signal transformation between time and frequency-domain.

> Industry offers electronic calibration kits that perform down to DC to simplify the calibration of the VNA for broadband time-domain measurements of low-pass systems.

Windowing An ideal *Dirac* δ-pulse in the time-domain transforms into an infinite spectrum in the frequency-domain, and vice verse, it has an *infinite* band, see Fig. II.2.214a. Due to the nature of the VNA, measuring to an upper frequency limit with a RF stimulus of constant power, we always have a signal response with a *finite* band in the frequency domain, which for e.g. $S_{ij} = 1$ has a rectangular distribution and transforms into a *sinc*-function signal waveform in the time-domain, and vice versa (see Fig. II.2.214b):

$$\begin{array}{rcl} \text{Frequency domain} & \Longleftrightarrow & \text{Time domain} \\ \mathrm{rect}\left(\dfrac{f}{\Delta f}\right) & \Longleftrightarrow & \dfrac{\sin(\pi\Delta f t)}{\pi t} = \Delta f \cdot \mathrm{sinc}\,(\pi\Delta f t)\,. \end{array} \tag{II.2.189}$$

i.e. into an non-physical, non-causal time-domain signal waveform. Any VNA sweeps to a perhaps very high, but limited `Stop` frequency ($\equiv \Delta f/2$), which still may not be a problem, as long as the system under analysis converges $|\mathbf{S}_{DUT}(f = \Delta f/2)| \to 0$ (or to a very low value) at the `Stop` frequency $\Delta f/2$. But if $|\mathbf{S}_{DUT}(f = \Delta f/2)| \neq 0$ the "*sinc-function* windowing effect" appears in the time-domain signal, see Fig. II.2.214b, as the transformation is always the multiplication of $\mathbf{S}_{DUT}(f)$ and the VNA measurement *window*:

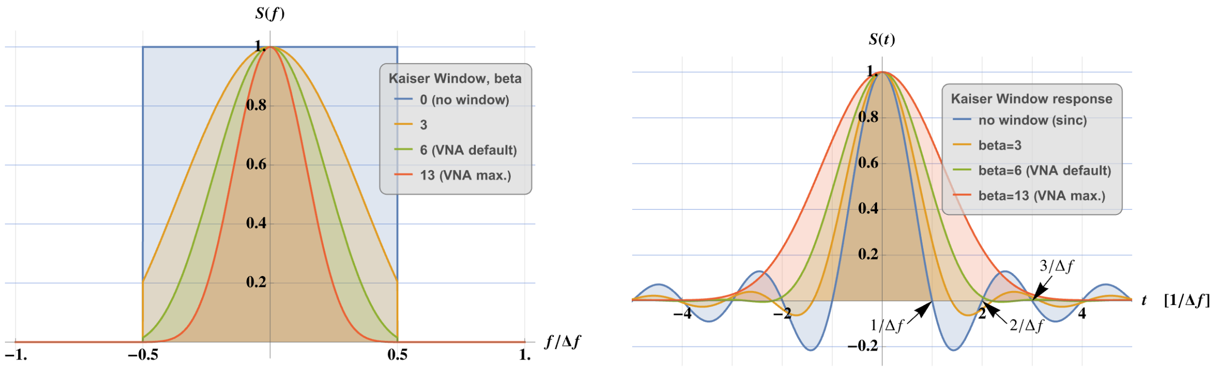
$$s(t) = \mathcal{F}^{-1}\left[\mathbf{S}_{DUT}(f) \times \mathrm{rect}\left(\frac{f}{\Delta f}\right)\right]$$

with $\Delta f = 2 f_{\mathtt{Stop}}$[12]. This unwanted effect can appear when measuring high-pass or all-pass systems, e.g. made from low-loss coaxial transmission lines, waveguides, or cavity resonators.

To mitigate the windowing effect, the VNA has one or more windowing low-pass weighting functions build in. Popular is the *Kaiser* window function, also called the *Kaiser-Bessel* window

$$w_{kb}(f,\beta) = \begin{cases} \frac{I_0(\beta\sqrt{1-2f}\sqrt{1+2f})}{I_0(\beta)} & \text{for } -1/2 < f < 1/2 \\ 0 & \text{otherwise} \end{cases}$$

[12] To keep things reasonable simple, we limit this discussion to the analysis of low-pass systems, with $f_{\mathtt{Start}}$ being near $f = 0$.



(a) in the frequency-domain. **(b)** in the time-domain.

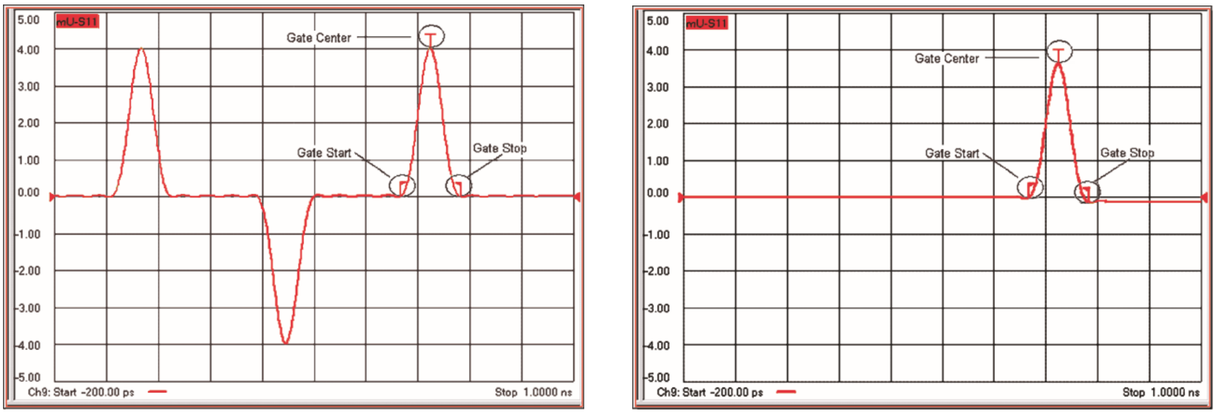
Fig. II.2.215: The *Kaiser* window function.

Figure II.2.215 illustrates the *Kaiser* window multiplication in the frequency-domain (Fig. II.2.215a) and the result in the time-domain (Fig. II.2.215b), here normalised to $S(t = 0) = 1$. In most cases the *default* setting of the *Kaiser* window parameter β, `beta=6`, turns out to be the best compromise between a good time resolution and low $\sin(x)/x$ ripple effects, see Fig. II.2.215b. Instead, setting `beta=0`, effectively disables the *Kaiser* window, and improves the time resolution at the cost of a large $\sin(x)/x$ ripple, while setting β to a high value, e.g. `beta=13`, typically the maximum value in many *Keysight* VNAs, minimises the ripple at the cost of a low time resolution.

***Dirac* δ or step function response** The time-domain transformation of the VNA can be applied to any of measured S-parameter, S_{ii} or S_{ij}, and results in the time-domain *impulse response*, equivalent to the response of the system to a *Dirac* δ-function in the time-domain, see the S_{21} example Fig. II.2.212b. In some cases the response to a *unit step* function $u(t)$, also called *Heaviside* function $H(t)$ is of interest, which is performed in the VNA by multiplying the measured $\mathbf{S}_{DUT}(f)$ with the *Fourier* transform of the Heaviside function $\mathcal{F}[H(t)] = \delta(f)/2 - j/(2\pi f)$, numerically. The step response is of particular interest for reflection coefficient measurements, $\Gamma_i = S_{ii}$, of transmission-line arrangements, this type of analysis is equivalent to the *time-domain reflectometry* (TDR), see also Exercise II.2.6.1. Basically all VNAs equipped with the time-domain option allow the used to select the *Dirac* δ or the step response.

Gating The *gating* feature allows to set `Start` and `Stop` gate markers along the time-domain trace to set a *gate*, which will zero the measurement between the gate markers (*notch gate*) or outside the gate markers (*band-pass gate*). Figure II.2.216 illustrates the use of gate markers, in Fig. II.2.216a the `Start` and `Stop` gate markers are set, also a `Center` marker is displayed, but the gate is still `Off`. In Fig. II.2.216b the gate is `On`, here activated as *band-pass gate*, “zeroing” the parts of the measurement outside of the gate.

The time-domain gating function enables the VNA to selectively “remove” parts of the measurement, e.g. of a transmission-line circuit, as time and physical location (space) are connected, see also the example on time-domain reflectometry. A typical application could be to remove the effect of a vacuum feedthrough to better analyse the actual circuit behind, e.g. a stripline kicker or beam



(a) Gate off. **(b)** Gate on.

Fig. II.2.216: Applying a *band-pass gate* in the time-domain.

monitor, or to de-embed a RF PCB circuit from the effects of the coaxial-to-stripline connectors. Figure II.2.217 shows the use of the gating function for a simple example, a S_{11} measurement of a terminated cable. Figure II.2.217 (a) displays $|S11(f)|$ and Fig. II.2.217 (b) the transformed time-domain impulse response $S_{11}(t)$, clearly visible are the reflections caused by the impedance discontinuities of the cable connectors at both ends. In Fig. II.2.217 (c) a gate is applied to notch-out the effect of the first connector, and Fig. II.2.217 (d) illustrates the result in the frequency domain by turning of the time-domain transformation while the gate is active.

The gating function is not a "brick-wall" as the gate markers on the time=domain trace may imply. Instead, the gating is transformed back into the frequency-domain and applied via a *Kaiser* or similar window function. As the gating is a non-linear operation, is may generate artefacts in the signal which were not in the original measured response! Clearly, the gating is an advanced function of the VNA and requires special training. As a rule of thumb, gating into resonant waveforms will give unpredictable results, instead, it is a good practice to always set the gate markers to locations of the time trace that have an asymptotically "zero" value, i.e. please do not cut into a signal trace different from zero!

Example II.2.9.2: Time-domain reflectometry with the VNA

The *time-domain reflectometry* (TDR) measurement was introduced to broadband sampling oscilloscopes in the 1960s, it basically is a measurement of the reflection coefficient Γ in the time-domain, similar to the principle discussed in Exercise II.2.6.1. Instead of a short pulse, for the TDR the stimulus function is a very long rectangular pulse (μs regime) with a fast rise time (ps regime), while only the first part of the pulse is used, approximating a unit step function.
The applications of TDR measurements in accelerator RF and beam instrumentation are numerous, in the laboratory to analyse and optimise all kind of transmission-line systems, for quality control of industry components and in the R&D phase of TL-based components and systems, as

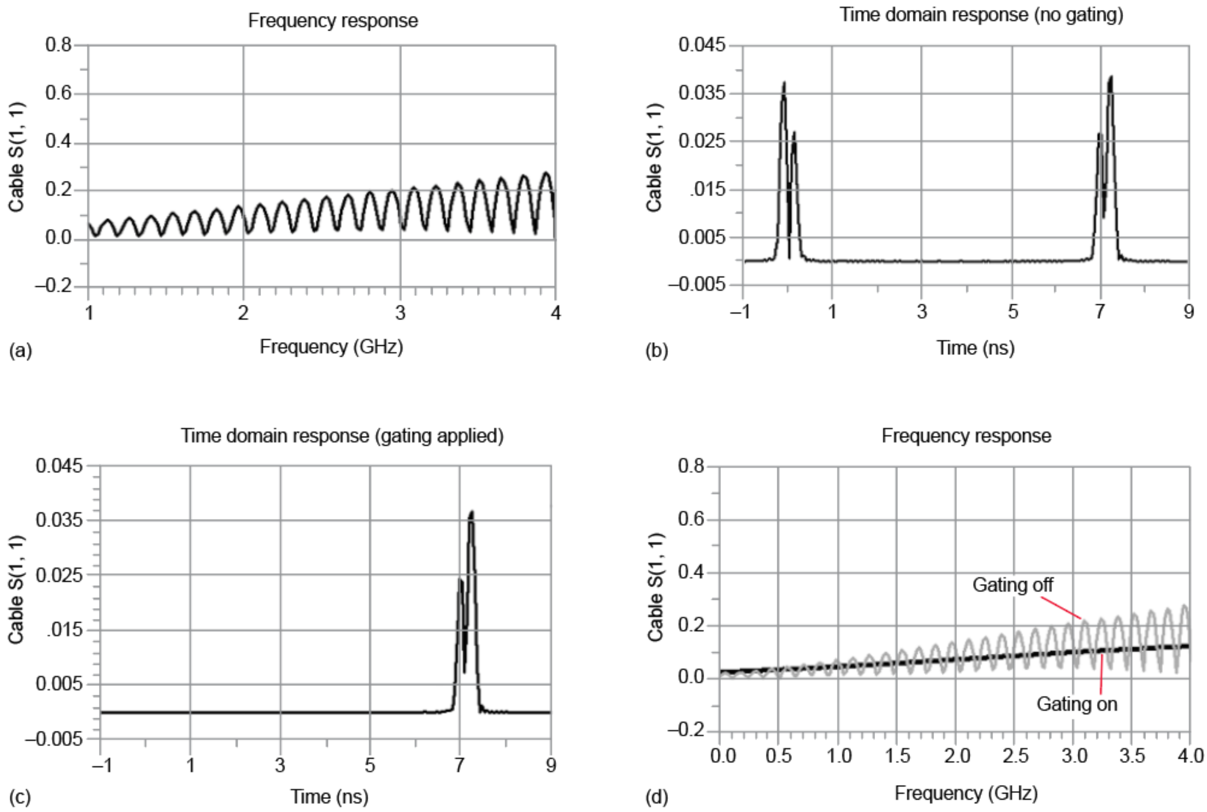
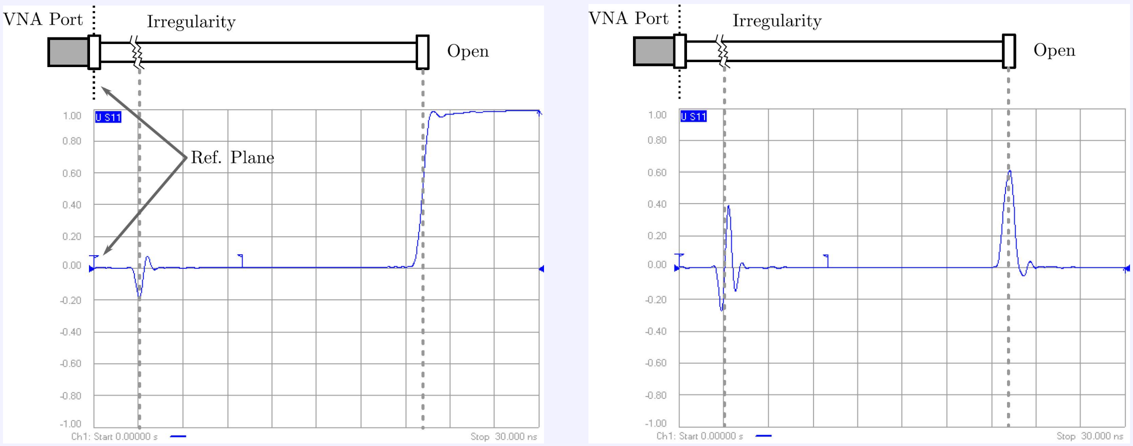


Fig. II.2.217: Operational sequence of time-domain gating:
(a) $|S_{11}(f)|$ frequency response of a cable. (b) $S_{11}(t)$ time-domain response. (c) gating applied to "remove" the first discontinuity. (d) $|S_{11}(f)|$ frequency response with gate `On` and `Off`.

well as in the field to trace errors on cables, connections, feedthroughs and TL-based components and subsystems. The use of the synthetic pulse technique, i.e. the time-domain transformation feature of the VNA, connected with gating and other signal processing features gives the TDR measurement further benefits, however, at the cost to be limited to LTI systems.

Figure II.2.218 illustrates the VNA-based TDR measurement on a transmission-line, i.e. on a $50\,\Omega$ coaxial cable with a open end, for both, the unit step response (Fig. II.2.218a) and the *Dirac* δ impulse response. The TDR measurement follows the procedure already been discussed:



(a) Using the `Step` stimulus function. **(b)** Using the `Impulse` stimulus function.

Fig. II.2.218: VNA TDR measurement on an open transmission-line (cable).

– Select `Start` and `Stop` frequency to a wide range, preferable to the minimum and maximum frequency range of the VNA to give the best possible time resolution.

– Select reasonable values for the number of `Points`, stimulus `Power`, `IF Bandwidth`, etc.

– Select the `Low-pass mode` for the time-domain transformation, and allow the VNA to adjust `Start` and `Stop` frequency, or select the frequency values and number of points following the discussed guidelines.

– Perform a 1-port calibration.

– Select a S_{11} measurement, displayed in `Real` format for *Keysight* VNAs.

– Select the `Step` stimulus function response (preferred), or the `Impulse` stimulus function response, and turn on the `Time-Domain Transform`.

– Adjust `Start` and `Stop` time, and the vertical scale accordingly for a good visualisation of the time-domain result.

For the TDR measurement the `Step` function result seems more intuitive, the vertical axis directly gives the reflection coefficient Γ, now as *real* value in the range $-1 < \Gamma < +1$. In the example, (Fig. II.2.218a, the cable has a fault ("Irregularity"), which is displayed as reflection of $\Gamma \approx -0.2$. The VNA has the transformation Eq. (II.2.165) build-in, which allows the vertical axis, and any markers accordingly, to be displayed in $Z[\Omega]$ units. The time axis t is linked to the physical location z as:

$$z = \frac{c}{\sqrt{\varepsilon_r}} \frac{1}{2} t$$

with ε_r being the *dielectric constant* of the homogeneous filling of cable insulation (for non-homogeneous TL dielectrics ε_{eff} has to be evaluated), e.g. $\varepsilon_r = 2.1$ for *PTFE Teflon*. The factor $1/2$ accounts for the incident and reflected pulse passing twice the transmission-line. Assuming a PTFE-filled coaxial cable, we find the length to be approximately $t \approx 22\,\mathrm{ns} \equiv \ell \approx 2.28\,\mathrm{m}$, and the location of the "irregularity" is approximately at $0.31\,\mathrm{m}$. A *velocity factor* can be set in the VNA, such that the marker value on the trace directly displays the Γ or Z value as function of the physical location z.

Table II.2.18: Some examples of values for the reflection coefficient wrt. the load impedance.

load impedance Z_L	reflection coefficient Γ
∞ (open circuit)	+1
0 (short circuit)	-1
Z_0 (matched load)	0
$Z_0/2$	-1/3
$2Z_0$	1/3

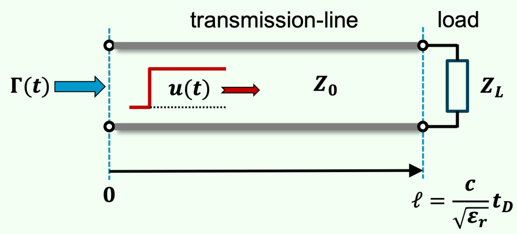
Please note the reflection coefficient $\Gamma(t)$, transformed into the time-domain, always is a *real* number, Table II.2.18 shows some example values, however, they are still in-line with Definition II.2.6.1.

The TDR procedure described in this example follows a manual setup of the VNA, utilising the time-domain transformation based on equidistant samples (frequency `Points`) of the S_{11} measurement in connection with a unit step function. Some VNAs provide more sophisticated TDR measurement implementations, *advanced TDR*, with a mostly automatised setup, also using a segmented sampling approach with fewer, wider spaced samples at low frequencies and more, dense spaced samples at higher frequencies, e.g. to improve the S/N. Display and look-and-feel of the advanced TDR resembles the classic TDR, while displaying the measurement results simultaneously in time and frequency-domain. 4-channel VNA enable even more sophisticated *differential TDR* measurements on coupled transmission-lines by combining two physical, unbalanced ports into a single *virtual*, balanced port.

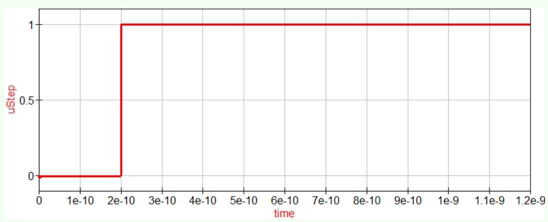
Evidently, the synthetic pulse time-domain technique gives the VNA an additional range of measurement applications, however, in some cases it cannot replace a traditional pulse generator / oscilloscope or TDR measurement setup. The VNA time-domain transformation always requires the DUT to be linear and time-invariant (LTI system) throughout the sweep time, which makes the actual measurement, including the internal calculations, rather slow compared to a traditional time-domain measurement setup. Some time-domain measurements, like the response of a RF feedback system, or a RF oscillator during startup, RF amplifiers near saturation, etc., with the DUT not in a steady-state, time-invariant status, or being non-linear, cannot be performed using the VNA synthetic pulse time-domain technique, still, the range of VNA time-domain applications remains very large. Moreover, the dynamic range of a VNA, typically $> 100\,\mathrm{dB}$ is superior compared to oscilloscopes, in particular when operating in a high frequency regime $> 20\,\mathrm{GHz}$, which leads to a substantial better signal-to-noise ratio.

Similar, or even more then the spectrum / signal analyzer, the vector network analyzer is a complex RF measurement system, and thanks to the combination of RF and signal processing technologies, can perform almost any RF measurement. This is particular true for the latest generation of 4-port (or more) high-end USB VNAs with integrated spectrum analyzer hardware, 2$^{\text{nd}}$ RF source and all software options including the advanced TDR functionality, capable to operate up to $50\,\mathrm{GHz}$ and beyond. To learn and understand more about the VNAs, please have a look at the documentation and seminars from the vendors [29–32].

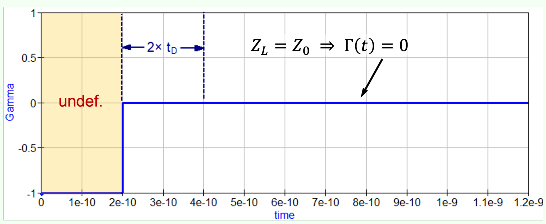
Exercise II.2.9.2: The reflection coefficient Γ in the time-domain



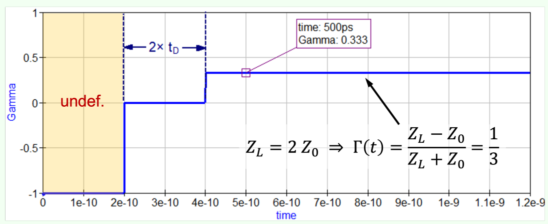
(a) TDR measurement principle.



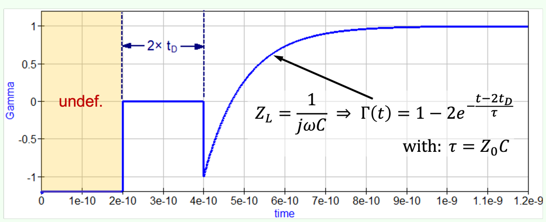
(b) The unit step (Heaviside) stimulus function.



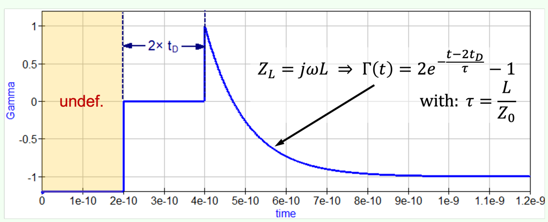
(c) $Z_L = 50\,\Omega$, no reflections!



(d) $Z_L = 100\,\Omega$, 33 % reflections.



(e) Capacitive load $Z_L = 2\,\mathrm{pF}$.



(f) Inductive load $Z_L = 5\,\mathrm{nH}$.

Fig. II.2.219: Examples of the time-domain reflection coefficient $\Gamma(t)$ for different load impedances Z_L.

Figure II.2.219 illustrates the behaviour of the reflection coefficient $\Gamma(t)$ for different load impedances Z_L, as measured by a TDR using a unit step pulse as stimulus function. Instead of an actual TDR measurement, the examples were generated by a *QucsStudio* simulation, see Fig. II.2.220. For simplicity, a `transient simulation` was used, with a `ideal voltage pulse source` as unit step stimulus function, and delaying the step pulse by $0.2\,\mathrm{ns}$ for better illustration. This means, the first $0.2\,\mathrm{ns}$ of the time traces in Fig. II.2.219 are undefined. Please notice, the time-delay due to a $30\,\mathrm{mm}$ long piece of is $t_D = \ell/c \approx 0.1\,\mathrm{ns}$, as the signal velocity for this `ideal transmission line` model uses $v = c\ (\varepsilon = 1)$.
Instead of the `transient simulation`, we also could chose the `s-parameter simulation` or `AC simulation` in *QucsStudio*, and make use of the build-in `Freq2Time` inverse discrete *Fourier* transform, as well as form the build-in *Kaiser* window function, to adapt the TDR simulation to the actual VNA TDR functionality.

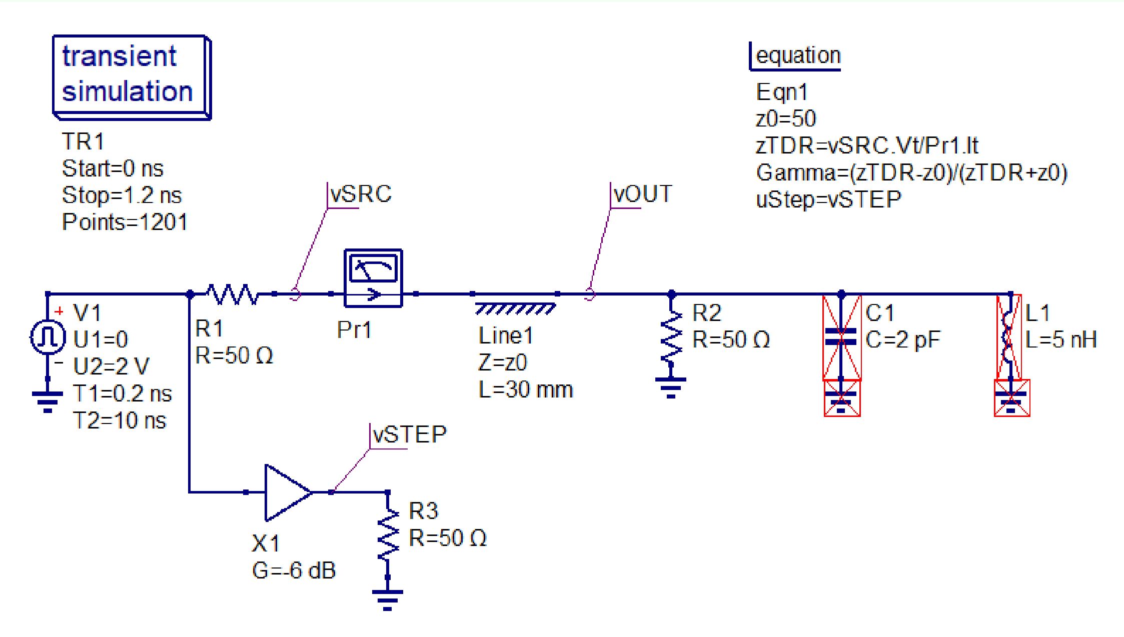


Fig. II.2.220: *QucsStudio* schematic for a simple TDR simulation.

II.2.9.4 Summary RF measurement techniques

While this chapter gives some theoretical background and the fundamental concepts of RF measurements, it cannot replace experience and know-how of those techniques, which only can be a accrued by performing, and continuously exercising hands-on RF measurements in the laboratory and in the field! Again, please note a systematic work order in RF engineering, and beyond:

- Start with the conceptual design, and divide large and complex systems into smaller subsystems and components.
- Design, analyse and understand those components or subsystems by approximation with analytical mathematical descriptions, which allows modifications and changes of parameters in a quick way.

- Model the design numerically, with more accuracy to the details, as basis for a prototype.
- Test and measure the prototype in the laboratory, and compare the results with the analytical and numerical analysis.
- Complete the entire system following the above steps and then test and measure the system with beam in the accelerator.

Numerical software tools and simulation suites are an excellent help in the R&D phase of accelerator RF components and systems, but they cannot replace the analysis and characterisation of the performance and the limitations of those systems by hands-on RF measurements.

In this chapter we could only give a brief introduction to the two most important RF measurement instruments, the spectrum / signal analyzer and the vector network analyzer, plus a few notes on “historic” and other RF measurement techniques. More information is found from the manufacturers, but the most important lessons will be leaned by execution of your own, hands-on RF measurements.

A Decibel [dB] ("dee-bee"), or not to be . . .

Bel is a logarithmic unit to express large ratios between values, popular is the tenth fraction, *deci-Bel*:

$$1\,\mathrm{dB} = \frac{1}{10}\mathrm{B} = 0.1\,\mathrm{B}\ (\mathrm{Bel})$$

While used in many engineering disciplines, the dB used in electrical and RF engineering usually expresses large radios between two electrical power values

$$P_{\mathrm{dB}} = 10\log_{10}\left(\frac{P_1}{P_2}\right)$$

or between two voltages or currents:

$$V_{\mathrm{dB}} = 20\log_{10}\left(\frac{V_1}{V_2}\right) \qquad I_{\mathrm{dB}} = 20\log_{10}\left(\frac{I_1}{I_2}\right)$$

with:

$$\frac{P_1}{P_2} = 10^{\left(\frac{P_{\mathrm{dB}}}{10}\right)} \qquad \frac{V_1}{V_2} = 10^{\left(\frac{V_{\mathrm{dB}}}{20}\right)} \qquad \frac{I_1}{I_2} = 10^{\left(\frac{I_{\mathrm{dB}}}{20}\right)}$$

Some important values:

dB ratio	$\boldsymbol{P_1/P_2}$	$\boldsymbol{V_1/V_2}$
$n\times$ dB	10^n	$10^{n/2}$
40 dB	10000	100
20 dB	100	10
10 dB	10	~3.16
6 dB	~4	~2
3 dB	~2	~1.41
0 dB	**1**	**1**
−3 dB	~0.5	~0.71
−20 dB	0.01	0.1

Please note, the 3 dB ratio (half power) is common specification for the bandwidth!

B “dB” is not “dBm”

dBm is defined as a logarithmic power unit, based on dB (deci-Bel) and a reference power of $P_{\mathrm{ref}} = 1\,\mathrm{mW}$

$$P_{\mathrm{dBm}} = 10\log_{10}\left(\frac{P}{P_{\mathrm{ref}}}\right), \qquad \text{with: } P = P_{\mathrm{ref}} 10^{\left(\frac{P_{\mathrm{dBm}}}{10}\right)}$$

dBm may also be used as logarithmic voltage unit, e.g. for the popular $Z_0 = 50\,\Omega$ impedance we calculate $V_{\mathrm{ref}} = \sqrt{Z_0 P_{\mathrm{ref}}} = \sqrt{0.05}\,\mathrm{V} \approxeq 0.2236\,\mathrm{V}$ (RMS).

$$V_{\mathrm{dBm}} = 20\log_{10}\left(\frac{V}{V_{\mathrm{ref}}}\right), \qquad \text{with: } V = V_{\mathrm{ref}} 10^{\left(\frac{V_{\mathrm{dBm}}}{20}\right)}$$

The use of dBm as logarithmic voltage (or current) unit **strictly requires** the waveform to be **sinusoidal!**

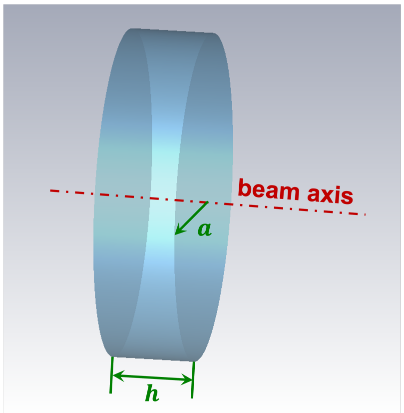
Some important values:

dBm	***P***	***V*** **(RMS)**
$90\,\mathrm{dBm}$	$1\,\mathrm{MW}$	$7.07\,\mathrm{kV}$
$60\,\mathrm{dBm}$	$1\,\mathrm{kW}$	$223.6\,\mathrm{V}$
$30\,\mathrm{dBm}$	$1\,\mathrm{W}$	$7.07\,\mathrm{V}$
$20\,\mathrm{dBm}$	$100\,\mathrm{mW}$	$2.24\,\mathrm{V}$
$10\,\mathrm{dBm}$	$10\,\mathrm{mW}$	$707\,\mathrm{mV}$
$6\,\mathrm{dBm}$	$4.0\,\mathrm{mW}$	$446\,\mathrm{mV}$
$\mathbf{0\,dBm}$	$\mathbf{1.0\,mW}$	$\mathbf{224\,mV}$
$-20\,\mathrm{dBm}$	$10\,\mu\mathrm{W}$	$22.4\,\mathrm{mV}$
$-60\,\mathrm{dBm}$	$1.0\,\mathrm{nW}$	$224\,\mu\mathrm{V}$
$-120\,\mathrm{dBm}$	$1.0\,\mathrm{fW}$	$224\,\mathrm{nV}$
$-174\,\mathrm{dBm}$	$4 \times 10^{-21}\,\mathrm{W}$	$0.446\,\mathrm{nV}$

Please note, $-174\,\mathrm{dBm}$ is the equivalent noise power in a bandwidth $BW = 1\,\mathrm{Hz}$ at room temperature.

C Exercises

1. An "empty" (air), cylindrical "pill-box" cavity (ideal cylinder, no beam-ports) is made of aluminium, which has a conductivity of $\sigma_{AL} = 3.8 \times 10^7\,\mathrm{S/m}$.
   The eigen-mode, used for the acceleration of charge particles should have a resonant frequency of $f_{res} = 500\,\mathrm{MHz}$.



   (a) What type of eigen-mode is used to accelerate charged particles?

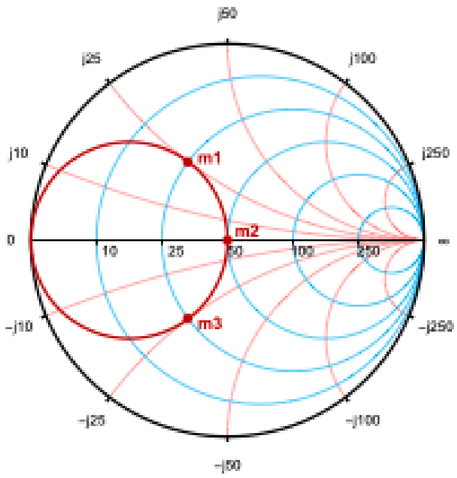
   (b) Calculate the diameter of the cavity.

   (c) Why is it a good idea to keep the height h of the cylindrical cavity always smaller than the diameter $2a$: $h < 2a$?

   (d) For this cavity we chose the height h to be half of the radius a. What is the dimension of h?

   (e) What is the Q-factor of the cavity?
   (Hint: Start by calculating the skin depth for aluminium at the resonant frequency)

2. The fundamental, accelerating mode of an "empty" (air), cylindrical "pillbox" prototype cavity (ideal cylinder, no beam ports) was characterised by S-parameter measurement with a vector network analyser (VNA). The three markers distributed on the *Smith* chart display – see the graph on the right – of the S_{11} measurement read:



   m1: 999.96 MHz
   m2: 1000.00 MHz
   m3: 1000.04 MHz

   A bead-pull perturbation measurement determined the "geometric factor" of the cavity: $R/Q = 90\,\Omega$

   (a) What is the resonant frequency of the cavity?

   (b) What is the $\pm 3\,\mathrm{dB}$ bandwidth of the cavity, and what kind of coupling is performed?

   (c) Calculate the unloaded Q-factor of the cavity.

   (d) Sketch the equivalent lumped circuit for the accelerating mode of this cavity.

   (e) Compute the element values of the equivalent circuit.

   (f) The cavity is fed from a RF power amplifier with a source impedance of $R_s = 50\,\Omega$. What is required transformer ratio of the coupling loop to match the cavity shunt impedance to the generator source impedance?

(g) Calculate the necessary RF power for a gap voltage of $1\,\mathrm{MV}$ peak.

Bonus question

(h) Compute the transit-time factor of the pillbox resonator.

i. Compute the radius a of the cavity.

ii. Compute the height h of the cavity.
(Hint: You may use the approximation for small arguments $\sin(x) \approxeq x$ in the analytical expression of R/Q.)

iii. Compute the transit-time factor T, assuming the gap length g is equal to the height h of the cavity, for a beam travelling with a velocity $v = 0.9c$.

3. Given are the S-matrices for five ideal RF components:

$$\mathbf{S_A} = \begin{bmatrix} 0 & 0 \\ 20 & 0 \end{bmatrix} \quad \mathbf{S_B} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \mathbf{S_C} = \begin{bmatrix} 0 & -j0.995 & 0.1 & 0 \\ -j0.995 & 0 & 0 & 0.1 \\ 0.1 & 0 & 0 & -j0.995 \\ 0 & 0.1 & -j0.995 & 0 \end{bmatrix}$$

$$\mathbf{S_D} = \frac{1}{2}\begin{bmatrix} 1 & -1 & \sqrt{2} \\ -1 & 1 & \sqrt{2} \\ \sqrt{2} & \sqrt{2} & 0 \end{bmatrix} \quad \mathbf{S_E} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$$

(a) Assign the S-matrices $\mathbf{S_A} \ldots \mathbf{S_E}$ to the components:

component	transmission-line	power splitter	directional-coupler	amplifier	circulator
S-matrix					

(b) What are the properties of the S-matrices in terms of impedance-matched, reciprocal and symmetry? Please complete the table:

	$\mathbf{S_A}$	$\mathbf{S_B}$	$\mathbf{S_C}$	$\mathbf{S_D}$	$\mathbf{S_E}$
matched	X				
reciprocal					
symmetric					

(c) What is the gain in dB of the amplifier?

(d) What is the length of the transmission-line, expressed in wavelengths?

(e) What is the coupling coefficient κ in dB of the directional-coupler?

D Solutions

D.1 In-text questions

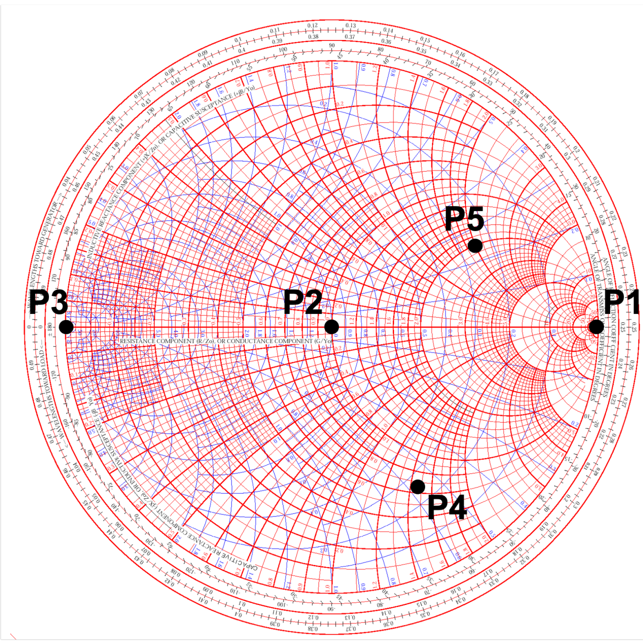
Section II.2.6.3

The *Smith* chart transforms the complex impedance plane
onto the complex Γ-plane (Reflection coefficient) within the unit circle.

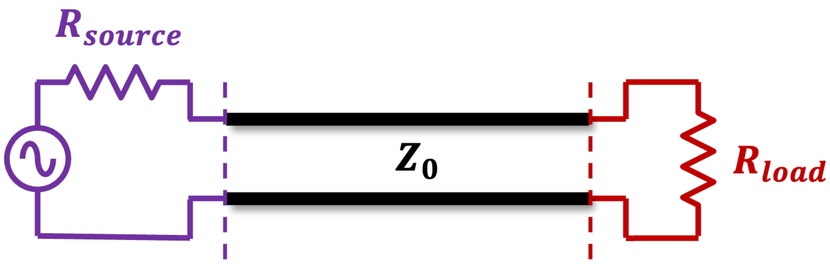
Prompts		Possible Answers
A. Point A	A5	1. $\Gamma = 1,\ z \to \infty$
B. Point B	B4	2. $\Gamma = -j$
C. Point C	C1	3. $\Gamma = 0,\ z = 1$, match
D. Point D	D3	4. Point in the capactive half plane
E. Point E	E6	5. $\Gamma = +j$
		6. $\Gamma = -1,\ z = 0$
		7. Point in the inductive half plane

Exercise II.2.6.2

point #	P1	P2	P3	P4	P5
Z/Ω	∞	50	0	$31 - j74$	$100 + j100$
$\boldsymbol{\Gamma}$	$1\angle 0° \ (+1)$	0	$1\angle 180° \ (-1)$	$0.7\angle -62°$	$0.62\angle 30°$



Section II.2.6.4



When do no signal reflections occur at the end of the transmission-line?

- ☐ $R_{source} = R_{load}$
- ☐ $R_{source} = Z_0$
- ☒ $Z_0 = R_{load}$
- ☒ $R_{source} = Z_0 = R_{load}$

Exercise II.2.6.3 #6

Z_L	C Series	L Series	R Series
$(50 + j25)\,\Omega$	12.7 pF	–	–
$(50 - j25)\,\Omega$	–	7.96 nH	–
$(4 + j21)\,\Omega$	15.2 pF	–	46 Ω
$(20 - j50)\,\Omega$	–	15.9 nH	30 Ω

Exercise II.2.6.3 #7

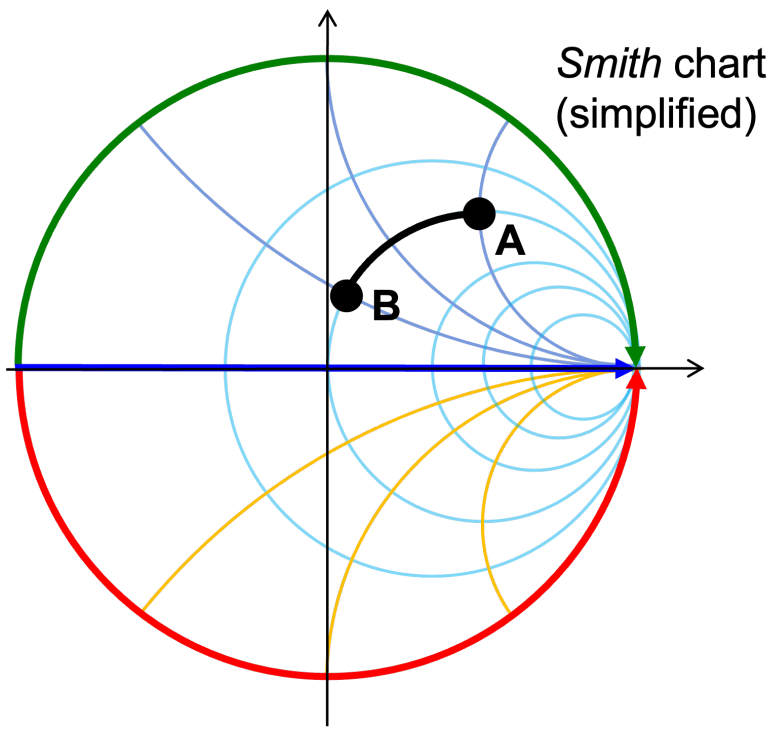
Z_L	C Shunt	L Shunt	R Shunt
$(50 + j25)\,\Omega$	2.55 pF	–	251.5 Ω
$(50 - j25)\,\Omega$	–	39.8 nH	250.8 Ω
$(4 + j21)\,\Omega$	14.6 pF	–	89.3 Ω
$(20 - j50)\,\Omega$	–	18.5 nH	76.8 Ω

Exercise II.2.6.3 #8

Z_L	C Series	L Series	C Shunt	L Shunt
$(32 - j66)\,\Omega$	–	$24.5\,\mathrm{nH}$	–	$102.1\,\mathrm{nH}\ *$
$(13 - j9)\,\Omega$	$24.5\,\mathrm{pF}\ *$	–	–	$9.44\,\mathrm{nH}$
$(37 + j34)\,\Omega$	$26.3\,\mathrm{pF}\ *$	–	$3.77\,\mathrm{pF}$	–
$(78 + j78)\,\Omega$	$4.38\,\mathrm{pF}$	–	–	$108.1\,\mathrm{nH}\ *$

The first element is marked with a $*$.
Other solutions are possible!

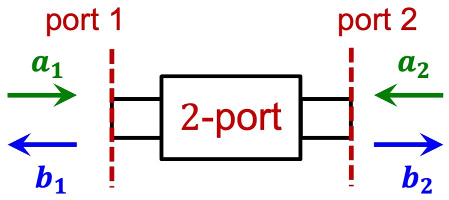
Exercise II.2.6.3 #9: see Fig. II.2.65.



Trace with marker points in the simplified *Smith* chart for an RL series impedance.
(Mark the correct answer)

- ☐ $f_B > f_A$
- ☒ $f_B < f_A$
- ☐ There is no frequency f related to Points A and B
- ☐ $f_B = f_A$

Section II.2.7.4



Mark all correct answers for the S-parameters of a 2-port RF network

- ☐ a_1 and b_1 are independent parameters.
- ☒ $S_{11} = b_1/a_1$ ($a_2 = 0$) is the input reflection coefficient Γ_1.
- ☒ a_1 and a_2 are the incident waves at port 1 and port 2, respectively.
- ☐ b_1 and b_2 are the transmitted waves between port 1 and port 2, and vice versa.
- ☒ S_{21} and S_{12} are the forward and reverse transmission gains / losses.
- ☒ To characterise the S-parameters at port 2, port 1 needs to be terminated in its characteristic port impedance.

Section II.2.7.5

Select all correct answers

- ☐ Y- and Z-parameters of electrical networks require a reference impedance Z_0.
- ☒ Scattering parameters of RF networks are based on normalised, complex power waves, incident and reflected at their ports.
- ☒ DUT stands for "Device Under Test", as acronym for a RF network to be characterised.
- ☐ S-parameters are only defined for a reference impedance of $Z_0 = 50\,\Omega$.
- ☒ Unused ports in a S-parameter measurement setup always need to be terminated in their characteristic port impedance.

Section II.2.7.6

Fill the centre column with the correct answer, $1 \dots 6$

Prompts		Possible Answers
A. matched	A4	1. $S_{ii} = S_{jj}$
B. symmetric	B3	2. $(\mathbf{S}^*)^T \mathbf{S} = \mathbf{I}$
C. reciprocal	C6	3. $S_{ij} = S_{ji} \ \wedge \ S_{ii} = S_{jj}$
D. passive and lossless	D2	4. $S_{ii} = 0$
		5. $\Gamma = +j$
		6. $S_{ij} = S_{ji}$

D.2 Solutions to the exercises, Section C

1. (a) TM010 ($\equiv$ E010)

   (b) diameter $2a = 0.459\,\mathrm{m}$

   (c) To keep the frequency of the accelerating TM010 mode below that of the unwanted higher-order TE111 mode.

   (d) $h = a/2 = 0.1147\,\mathrm{m}$

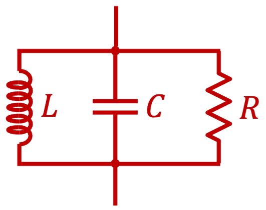
   (e) $\delta = 3.65\,\mu\mathrm{m}$, $Q = 20650$

2. (a) $f_{res} = f_{m2} = 1000.00\,\mathrm{MHz}$

   (b) $f_{BW3dB} = f_{m3} - f_{m1} = 80\,\mathrm{kHz}$. The cavity is in critical coupling.

   (c) $Q_0 = 2Q_L = 25000$

   (d)



   (e) $R = 2.25\,\mathrm{M\Omega}$, $L = 14.324\,\mathrm{nH}$, $C = 1.7684\,\mathrm{pF}$

   (f) $k = 212.1$

   (g) $P = 222.2\,\mathrm{kW}$

   (h) i. $a = 0.1147\,\mathrm{m}$

   ii. $h/a \approxeq 0.4865 \ \Rightarrow \ h \approxeq 55.82\,\mathrm{mm}$

   iii. $\lambda_{TM010} = 0.2998\,\mathrm{m} \ \Rightarrow \ T = 0.931$

3. (a)

component	transmission-line	power splitter	directional-coupler	amplifier	circulator
S-matrix	$\mathbf{S_E}$	$\mathbf{S_D}$	$\mathbf{S_C}$	$\mathbf{S_A}$	$\mathbf{S_B}$

(b)

	$\mathbf{S_A}$	$\mathbf{S_B}$	$\mathbf{S_C}$	$\mathbf{S_D}$	$\mathbf{S_E}$
matched	X	X	X		X
reciprocal			X	X	X
symmetric			X		X

(c) $g = 20\log_{10} G = 20\log_{10}|S_{21}| = 26\,\text{dB}$

(d) $S_{21} = S_{12} = -1 = \exp(-j\pi) \;\Rightarrow\; \theta = \beta\ell = \pi \;\Rightarrow\; \ell = \lambda/2$

(e) $S_{31} = S_{13} = S_{24} = S_{42} = k \;\Rightarrow\; \kappa = 20\log_{10} k = -20\,\text{dB}$

References

[1] Karle S. Packard, *The Origin of Waveguides: A Case of Multiple Rediscovery*, IEEE Transactions on Microwave Theory and Techniques, Vol. MTT-32, no. 9, September 1984

[2] Zhang K. , Dejie L. *Electromagnetic Theory for Microwaves and Optoelectronics*, 2nd ed. (Springer, Berlin Heidelberg, 2008), doi:10.1007/978-3-540-74296-8.

[3] D.M. Pozar, *Microwave engineering*, 4th ed. (Wiley, Hoboken, NJ, 2012), Internet Archive.

[4] H.H. Meinke, F.W. Gundlach, *Taschenbuch der Hochfrequenztechnik*, 3rd ed. (Springer, Heidelberg, 1968), doi:10.1007/978-3-662-13082-7.

[5] Zinke, O. and Brunswig, H., *Hochfrequenztechnik I*, 5th Ed. (Springer, Heidelberg 1995), doi:10.1007/978-3-662-00777-8

[6] Barrett, R. M. and Barnes M. H., *Microwave printed circuits*, Radio TV News, vol. 46, 1951

[7] Bhat, B. and Koul, S. K., *Stripline-like Transmission Lines for Microwave Integrated Circuits*, 2007, New Age International Publishers.

[8] Hoffmann, R. K., *Integrierte Mikrowellenschaltungen*, (Springer Verlag, 1983).

[9] Ansys, Inc., https://www.ansys.com/, last visited 3 July 2024.

[10] M. Wendt, "RF Training: Numerical Analysis", CERN Accelerator School Advanced, Sevrier, France, November 2022.

[11] S. Udongwo *et al.*, "Numerical Analysis of RF Problems", CAS course on "RF for Accelerators", Berlin, Germany, June 2023.

[12] ANSYS, Inc., "HFSS Fields Calculator Cookbook", Release 17.0, December 2015. `http://www.ansys.com`

[13] P. Falstad, simple circuit simulator in Javascript, http://falstad.com/circuit/

[14] F. Dellsperger, *Smith*-chart tool, http://www.fritz.dellsperger.net/smith.html.

[15] K. Kurokawa,"Power waves and the scattering matrix", IEEE Transactions on Microwave Theory and Techniques, vol.13, iss.2, pp. 194–202, March 1965

[16] G. L. Matthaei, L. Young and E. M. T. Jones, "Design of Microwave Filters, Impedance-matching Networks, and Coupling Structures", Stanford Research Institute, Menlo Park, CA, January 1963

[17] Jia-Sheng Hong, M. J. Lancaster, "Microstrip Filters for RF/Microwave Applications", 1st Edition, John Whiley & Sons, Inc., 2001, ISBN 0-471-38877-7 (Hardback) or ISBN 0-471-22161-9 (Electronic)

[18] R. Saal and E. Ulbrich,"On the Design of Filters by Synthesis", IRE Transactions on Circuit Theory, vol.6, iss.4, pp. 284–327, December 1958

[19] Anatol I. Zverev, "Handbook of Filter Synthesis", John Whiley & Sons, Inc., 1967, ISBN 978-0-471-74942-4

[20] M. Markgraf, QucsStudio – A Free and Powerful Circuit Simulator, https://qucsstudio.de.

[21] B. Schiek and H.J. Sieveris, "Rauschen in Hochfrequenzschaltungen", Hüthig, Heidelberg, 1984, ISBN 3-7785-2007-5.

[22] O. Zinke and H. Brunswig, “Lehrbuch der Hochfrequenztechnik, Zweiter Band”, Springer, Berlin, 1974, ISBN 3-540-06245-9.

[23] Ph.N. Burrows *et al.* (eds.). The Compact Linear e^+e^- Collider (CLIC) – 2018 Summary Report, CERN-2018-005-M (CERN, Geneva, 2018), doi:10.23731/CYRM-2018-002.

[24] *Agilent Technologies* Application Note 923: “Schottky Barrier Diode Video Detectors” 1999 Agilent Technologies inc., 5954-2079 (11/99)

[25] *Hewlett Packard* Application Note 1156: “Diode Detector Simulation using Hewlett-Packard EESOF ADS Software” 1998 Hewlett Packard Co., 5968-1885E (10/98)

[26] *Agilent Technologies* Application Note 956-5: “Dynamic Range Extension of Schottky Detectors” 1999 Agilent Technologies Inc., 5952-8335 (11/99)

[27] *Keysight Technologies* Application Note 150: “Spectrum Analysis Basics” 1971 - 2016 Keysight Technologies Inc., 5952-0292 (Nov. 2016)

[28] F. Caspers, G. Dôme, Precise perturbation measurements of resonant cavities higher order mode identification, CERN SPS/85-46 (ARF), Geneva, Switzerland, November 1985

[29] *Keysight Technologies* Application Note “Understanding the Fundamental Principles of Vector Network Analysis” 2022 Keysight Technologies Inc., 5965-7707EN (August 10, 2022)

[30] *Rohde & Schwarz* Primer “Fundamental of Vector Network Analysis” Version 1.1

[31] *Keysight Technologies* “Network Analysis Basics” 2004 - 2014 Keysight Technologies Inc., 5965-7917E (July 31, 2014)

[32] *Keysight Technologies* Application Note “Time Domain Analysis Using a Network Analyzer” 2017 - 2024 Keysight Technologies Inc., 5989-5723N (January 24, 2024)