RF Measurement Concepts

P. Kowina[2]*, C. Völlinger*[1] *and M. Wendt*[1]
[1]CERN, Geneva, Switzerland
[2]GSI, Darmstadt, Germany

1 A note to the history of RF signal receiving and measurement techniques

In the early days of radio-frequency (RF) engineering the available instrumentation for measurements was rather limited. Besides elements acting on the heat developed by RF power (bi-metal contacts and resistors with a very high temperature coefficient) only point/contact diodes, and to some extent vacuum tubes, were available as signal detectors. For several decades the slotted measurement line, see Section 8.1, was the only commonly used instrument to measure impedances and complex reflection coefficients. Around 1960 the tedious work with these coaxial and waveguide measurement lines became considerably simplified with the availability of the vector network analyzer. At the same time the first sampling oscilloscopes with 1 GHz bandwidth arrived on the market. This was possible due to progress in solid-state (semiconductor) technology and advances in microwave elements (microstrip lines). Reliable, stable and easily controllable microwave sources are the backbone of spectrum and network analyzers, as well as sensitive (low-noise) receivers. The following sections focus on signal receiving devices such as spectrum analyzers. An overview of network analysis is given later in Section 5.

2 Basic definitions, elements and concepts

Before discussing key RF measurement devices, a brief overview of the most important components used in these devices and the related basic concepts are presented.

2.1 Decibel

Since the unit decibel (dB) is frequently used in RF engineering, a short introduction and definition of the terms are given. The decibel is a unit used to express relative differences between quantities, e.g. of signal power. It is expressed as the base-10 logarithm of the ratio of the powers between two signals:

$$P\ [\text{dB}] = 10 \cdot \log(P/P_0). \tag{1}$$

It is also common to express the signal amplitude in dB. Since power is proportional to the square of the signal amplitude, a voltage ratio in dB is expressed as:

$$V\ [\text{dB}] = 20 \cdot \log(V/V_0). \tag{2}$$

In Eqs. (1) and (2), P_0 and V_0 are the reference power and voltage, respectively. A given value in dB is the same for power ratios as for voltage ratios. It is important to note that there are no 'power dB' or 'voltage dB' as dB values always express a ratio. Conversely, the absolute power and voltage can be obtained from dB values by

$$P = P_0 \cdot 10^{\frac{P\,[\text{dB}]}{10}}, \tag{3}$$

$$V = V_0 \cdot 10^{\frac{V\,[\text{dB}]}{20}}. \tag{4}$$

The advantage using a logarithmic scale as unit of the measurement is twofold:

i) typical RF signal powers tends to span several orders of magnitude; and

Table 1: Overview of common dB values and their conversion into power and voltage ratios

	Power ratio	Voltage ratio
-20 dB	0.01	0.1
-10 dB	0.1	0.32
-6 dB	0.25	0.5
-3 dB	0.50	0.71
-1 dB	0.74	0.89
0 dB	1	1
1 dB	1.26	1.12
3 dB	2.00	1.41
6 dB	4	2
10 dB	10	3.16
20 dB	100	10
$n \cdot 10$ dB	10^n	$10^{n/2}$

ii) signal attenuation losses and gains can simply computed by subtraction and addition.

Table 1 helps to familiarize with signal ratios and the associated dB values.

Absolute levels are expressed using a specific reference value, these dB systems are not based on SI units. Strictly speaking, the reference value should be included in parentheses when giving a dB value, e.g. +3 dB (1 W) indicates 3 dB at $P_0 = 1$ W, thus 2 W. However, it is more common to add some typical reference values as letters after the unit, e.g. dBm defines dB using a reference level of $P_0 = 1$ mW. Thus, 0 dBm correspond to -30 dBW, where dBW indicates a reference level of $P_0 = 1$ W. Often a reference impedance of 50 Ω is assumed. Other common units are:

i) dBmV for small voltages with V_0 = 1 mV; and
ii) dBmV/m for the electric field strength radiated from an antenna with reference field strength $E_0 = 1$ mV/m.

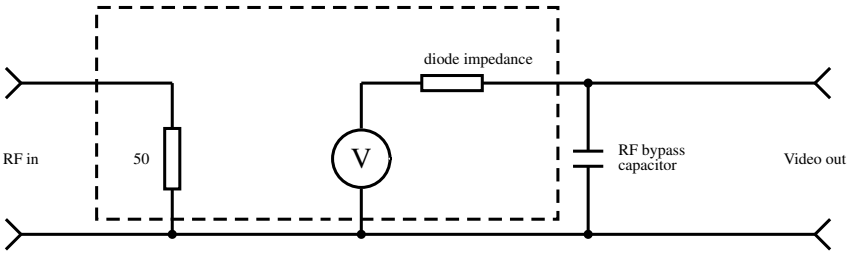


Fig. 1: Simplified equivalent circuit of a diode detector (w/o parasitic elements)

2.2 The RF diode

One of the most important elements, even today inside the most sophisticated RF measurement devices is the fast RF diode or *Schottky* diode. The basic metal–semiconductor junction has an intrinsically very

Fig. 2: A typical *Schottky* diode. The RF input of this detector diode is on the left and the video output on the right (courtesy *Agilent*).

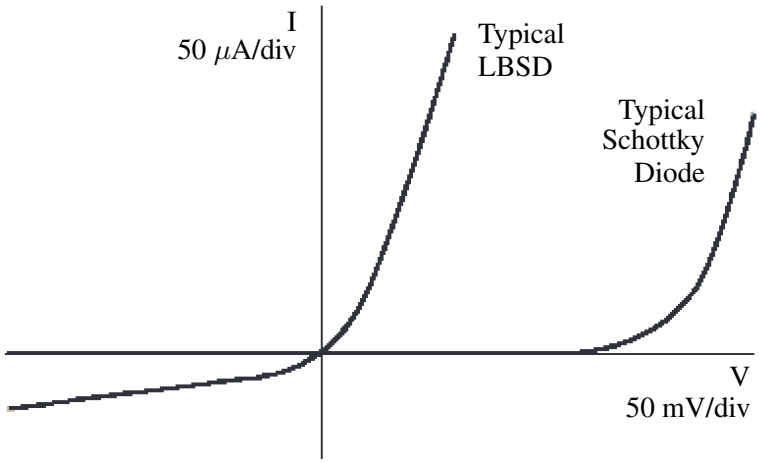


Fig. 3: Current as a function of voltage for different diode types (LBSD = low barrier *Schottky* diode)

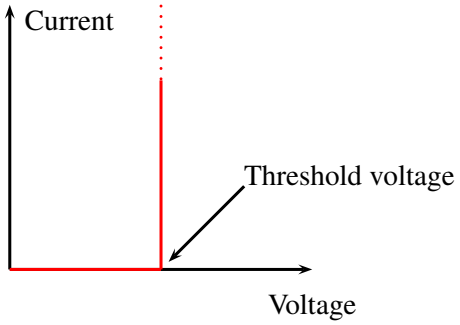


Fig. 4: The current–voltage relation of an ideal commutator with threshold voltage

fast switching time of well below a picosecond, provided that the geometric size and hence the junction capacitance of the diode has sufficiently small dimensions. However, the unavoidable, and voltage-dependent junction capacity will lead to limitations of the maximum operating frequency. The simplified equivalent circuit of such a diode is depicted in Fig. 1 and an example of a commonly used *Schottky* diode is shown in Fig. 2. of the most important properties of any diode is its IV-characteristic, which is the relation of the current passing the diode as a function of the applied voltage [1]. This relation is depicted graphically for two different types of diodes in Fig. 3. It shows, the diode is a non-ideal commutator (in contrary to that shown in Fig. 4) for small signals. Note that it is not possible to apply large signals, since this kind of diode would burn out. Although there exist versions with rather large power handling

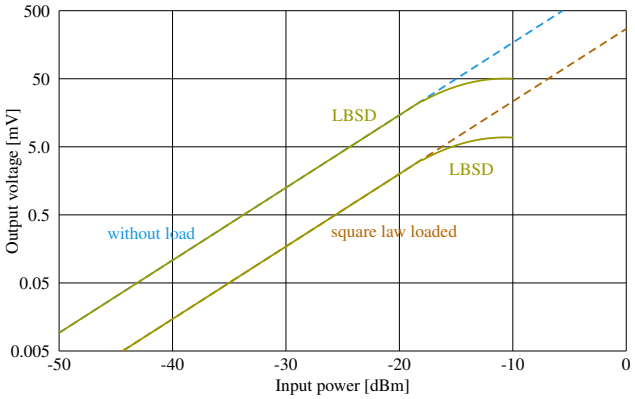


Fig. 5: Relation between input power and output voltage

capability of *Schottky* diodes, these can stand more than 9 kV and several tens of amperes, they are not suitable in microwave applications due to their large junction capacity. The region where the output voltage is proportional to the input power is called the square-law region (Fig. 5). In this region the input power is proportional to the square of the input voltage and the output signal is proportional to the input power, hence the name square-law region.

The transition between the linear region and the square-law region is typically between -10 and -20 dBm (Fig. 5). For a more detailed description, see [2].

There are some fundamental limitations when using diodes as detectors. The output signal of a diode (essentially DC or modulated DC if the RF is amplitude modulated) does not contain any phase information. In addition, the sensitivity of a diode limits the input level range to about -60 dB at best, which is not sufficient for many applications.

The minimum detectable power level of a RF diode is specified by the 'tangential sensitivity', which typically amounts to -50 to -55 dBm for 10 MHz video bandwidth at the detector output [3].

To overcome these limitations, a more sophisticated method to utilize the RF diode is required. This method is presented in the next section.

2.3 Mixer

To include the detection of very small RF signals a device with a linear response over a wide range of signal levels (from 0 dBm (= 1 mW) down to the thermal noise = -174 dBm/Hz = $4{\cdot}10^{-21}$ W/Hz) is highly preferred. A RF mixer provides these features by using one, two or four diodes in different configurations (Fig. 6). A mixer is essentially a frequency multiplier with a very high dynamic range, implementing in it's simplest form the function

$$f_1(t) \cdot f_2(t) \quad \text{with } f_1(t) = \text{RF signal} \quad \text{and } f_2(t) = \text{ local oscillator (LO) signal} \tag{5}$$

or more explicitly, for two sinusoidal signals with amplitudes a_i and frequencies f_i ($i = 1, 2$),

$$a_1 \cos(2\pi f_1 t + \varphi) \cdot a_2 \cos(2\pi f_2 t) = \frac{1}{2} a_1 a_2 \left[\cos((f_1 + f_2)t + \varphi) + \cos((f_1 - f_2)t + \varphi)\right]. \tag{6}$$

Thus, we obtain a response at the intermediate-frequency (IF) port as sum and difference frequencies of the local oscillator (LO $= f_1$) and RF ($= f_2$) signals. Examples of different mixer configurations are shown in Fig. 6, they all use diodes to multiply the two applied signals, RF and LO. These diodes operate like a switch, controlled by the frequency of the LO signal (Fig. 7). The response of a mixer in the time domain is depicted in Fig. 8.

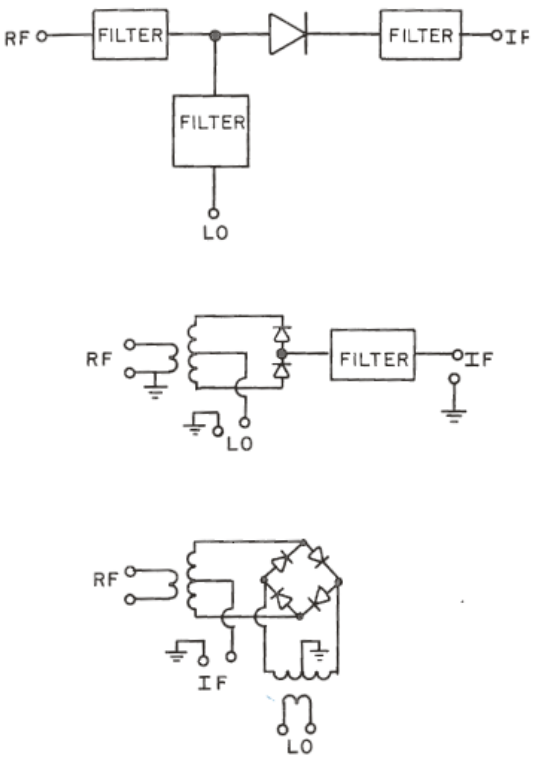


Fig. 6: Examples of different mixer configurations

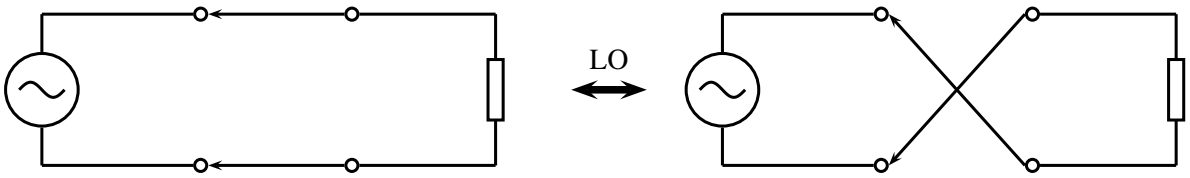


Fig. 7: Two circuit configurations interchanging with the frequency of the LO where the switches represent the diodes.

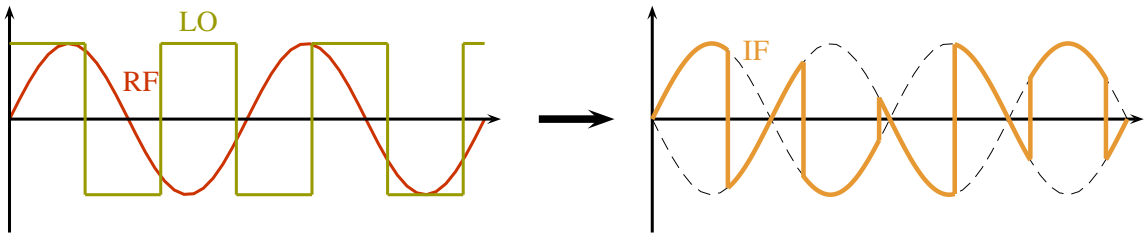


Fig. 8: Time-domain response of a mixer

The output signal is always in the “linear regime”, provided that the mixer is not saturated with respect to the RF input signal. Note, with respect to the LO signal the mixer has to be always in saturation to insure the diodes operate almost as an ideal switch. The phase of the RF signal is conserved in the output signal available at the IF output.

2.4 Amplifier

A linear amplifier, sometimes called “gain stage”, auguments the input signal by a factor which is usually indicated in decibels (dB). The ratio between the output and the input signals is called the transfer function and its magnitude – the voltage gain G – is measured in dB and given as

$$G[\text{dB}] = 20 \cdot \frac{V_{\text{RFout}}}{V_{\text{RFin}}} \quad \text{or} \quad \frac{V_{\text{RFout}}}{V_{\text{RFin}}} = 20 \cdot \log G[\text{lin}]. \tag{7}$$

The circuit symbol of an amplifier is shown in Fig. 9 together with its S-matrix.

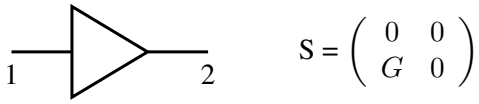


Fig. 9: Circuit symbol and S-matrix of an ideal amplifier

The bandwidth of an amplifier specifies the frequency range where it is usually operated, see Fig. 10. This frequency range is defined by the -3 dB points[1] of the magnitude response with respect to its maximum or nominal transmission gain, dividing the magnitude transfer function of the amplifier into a pass-band and a stop-band of equal transmitted power.

For an ideal amplifier the output signal would always be proportional to the input signal. However, a real amplifier is non-linear, typically for larger signals the transfer characteristic deviates from its linear properties, which is validated for small-signal amplification. When increasing the output power of an amplifier, a point is reached where due to the non-linearities the small-signal gain is reduced by 1 dB (Fig. 11). This output power level defines the so-called 1 dB compression point, which is an important measure of the output power capability, thus the dynamic range for the amplifier.

The transfer characteristic of an amplifier can be described in commonly used terms of RF engineering, i.e. the S-matrix, see Section 5. As implicitly contained in the S-matrix, both, amplitude and phase information of any spectral component are preserved when passing through an ideal amplifier. For a real amplifier the element $G = S_{21}$ (transmission from port 1 to port 2) is not a constant, but a complex function of frequency. Also the elements S_{11} and S_{22} are not zero.

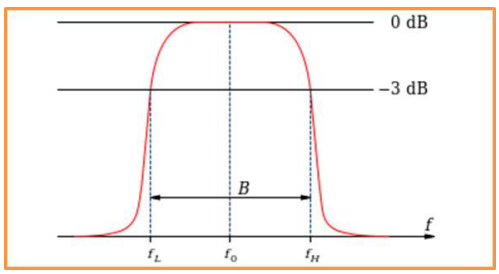


Fig. 10: Definition of the bandwidth

[1]The -3 dB points are the values left and right of a reference value, typically the local maximum of the amplifier transfer function, and are 3 dB below that reference.

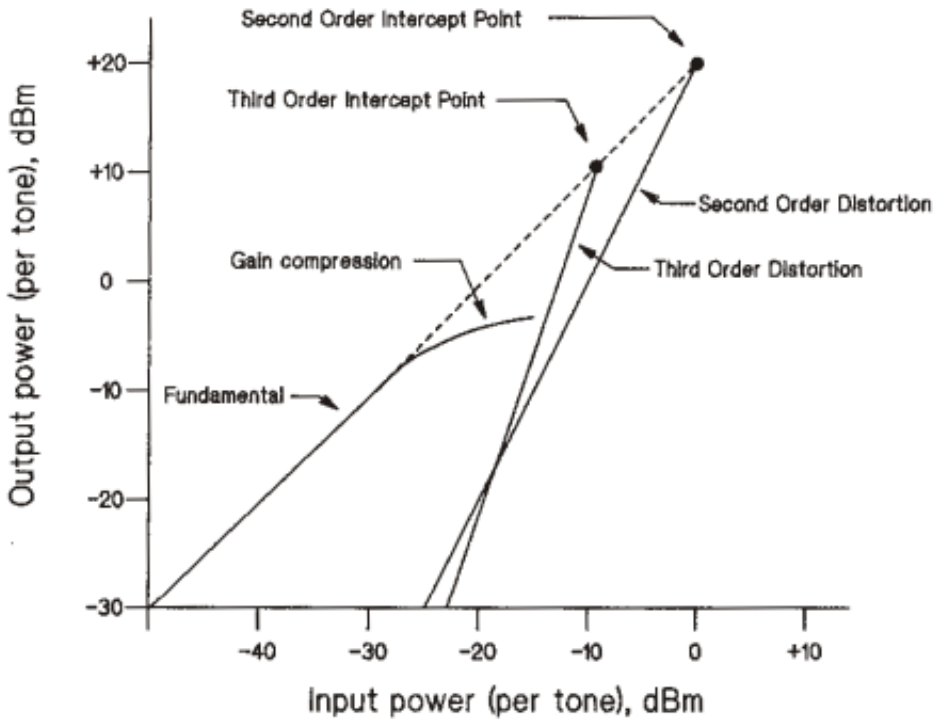


Fig. 11: Example for the 1 dB compression point [4]

2.5 Interception points of non-linear devices

Important characteristics of non-linear devices are the interception points. Here, only a brief overview is given, further information can be found in [4].

The most relevant interception points is the interception point of third order (IP3 point). Its importance derives from its straightforward determination, plotting the input versus the output power on a logarithmic scale (Fig. 11). The IP3 point is usually not measured directly, but is extrapolated from the data, measured at much lower power levels in order to avoid overload or damage of the device under test (DUT). Applying two signals $(f_1, f_2 > f_1)$ of closely spaced frequencies Δf simultaneously to the DUT, the intermodulation products appear at $+\Delta f$ above f_2 and $-\Delta f$ below f_1. This method is called the third-order intermodulation (TOI). An example of an automatized TOI measurement is shown in Fig. 12.

The transfer function of weakly non-linear devices can be approximated by a *Taylor* expansion. Using n higher order terms and plotting them together with an ideal linear device on a logarithmic scale results in two straight lines with different slopes ($x^n \overset{\log}{\rightarrow} n \cdot \log x$). Their intersection point is the intercept point of nth order. These points provide important information concerning the quality of non-linear devices.

In this context, the aforementioned 1 dB compression point of an amplifier is the intercept point of first order. For the method of measurements of the 1 dB compression point, see Section 7.4.

Similar characterization techniques can also be applied for mixers, which, with respect to the LO signal, cannot be considered as weakly non-linear devices.

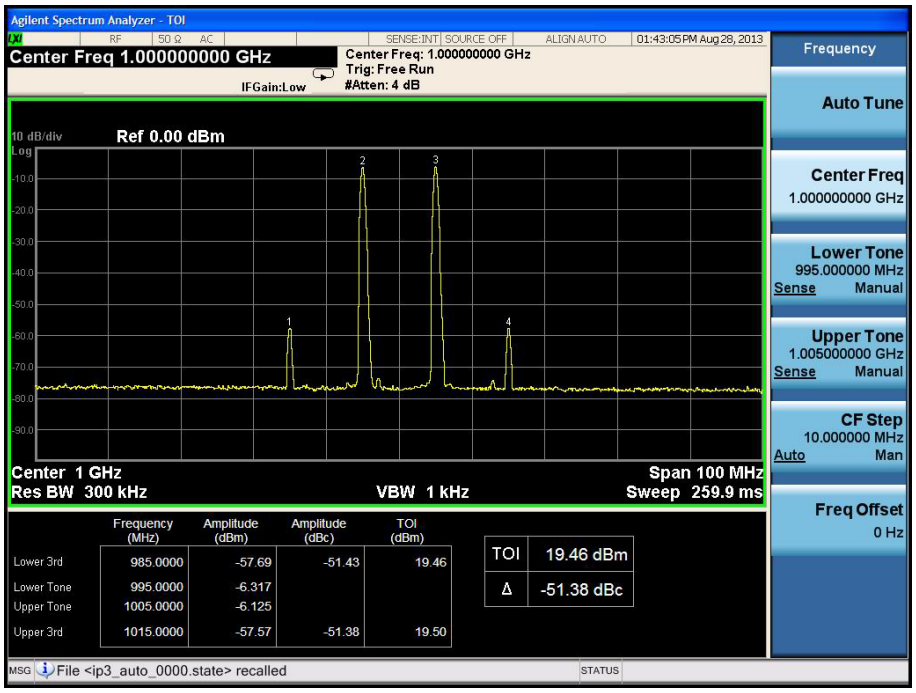


Fig. 12: An example of automatized TOI measurement

2.6 The superheterodyne concept

The word superheterodyne is composed of three parts: super (Latin: over), $\epsilon\tau\epsilon\rho\omega$ (hetero, Greek: different) and $\delta\upsilon\nu\alpha\mu\iota\sigma$ (dynamic, Greek: force), and can be translated as two forces superimposed[2]. Different abbreviations exist for the superheterodyne concept. In the USA it is often abbreviated by the simple word "heterodyne", and in Germany the shorter terms "super" or "superhet" are used.

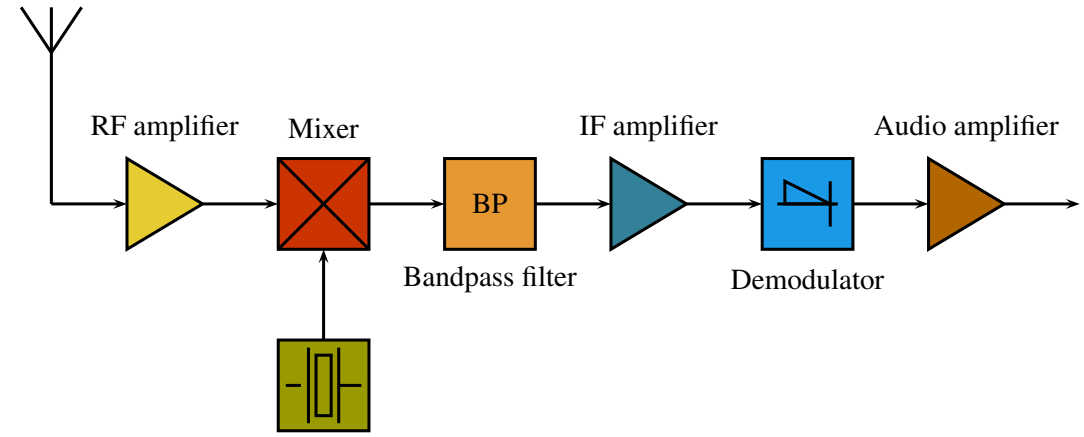


Fig. 13: Schematic drawing of a superheterodyne radio receiver

A "weak" incident (RF) signal is subjected to non-linear superposition (i.e. mixing or multiplication) with a "strong'' sine wave signal from a LO. At the mixer output sum and difference frequencies of the RF and LO signals appear. The LO signal can be tuned such that this IF output signal is always

[2]The direct translation (roughly) would be: another force becomes superimposed.

of same frequency, or stays within a very narrow frequency band. Therefore, a fixed-frequency bandpass with excellent transfer characteristics can be used, which is cheaper and easier to realize than a variable bandpass of the same performance. Also, gain-stages (amplifiers) operating at a lower IF frequency are of better quality and/or are more affordable. A well-known application of this principle is any simple radio receiver (Fig. 13).

3 Spectrum analyser

RF spectrum analyzers can be found in virtually every control room of a modern particle accelerator. They are used for many aspects of beam diagnostics including Schottky signal acquisition and observation of RF signals. A spectrum analyzer is in principle very similar to a common superheterodyne broadcast receiver, except with respect to the choice of functions, change of parameters, and in general a more sophisticated, high quality design. It sweeps automatically through a specified frequency range, which corresponds to an automatic turning of the tuning knob on a radio. The signal is then displayed in the amplitude/frequency plane. Originally, these kind of measurement instruments were setup manually and used a cathode ray tube (CRT) as display. Nowadays, with the availability of low-cost, powerful digital electronics for control and signal processing, basically every instrument can be remotely controlled. A microprocessor permits fast and reliable settings of the instrument, and an analog-digital-converter (ADC) in connection with digital signal processing hardware performs the acquisition and pre-processing of the measured signal values. The digital data processing enables extensive data treatment for error correction, complex calibration routines and self tests, which are a great improvement for RF signal measurements. However, the user of such sophisticated systems may not always be aware of the basic analogue signal path and processing, before the signals are digitized and prepared for user interaction. The basics of these analogue sections is discussed as follows.

In general, we distiguish two types of spectrum analyzers:

- the scalar spectrum analyzer (SA) and
- the vector spectrum analyzer (VSA).

The SA provides only information of the amplitude of the applied signal, while the VSA provides information of the phase as well.

3.1 Scalar spectrum analyzer

A common oscilloscope displays a signal in the amplitude-vs.-time format (time domain). The SA follows a different approach and displays the RF signal in the frequency domain.

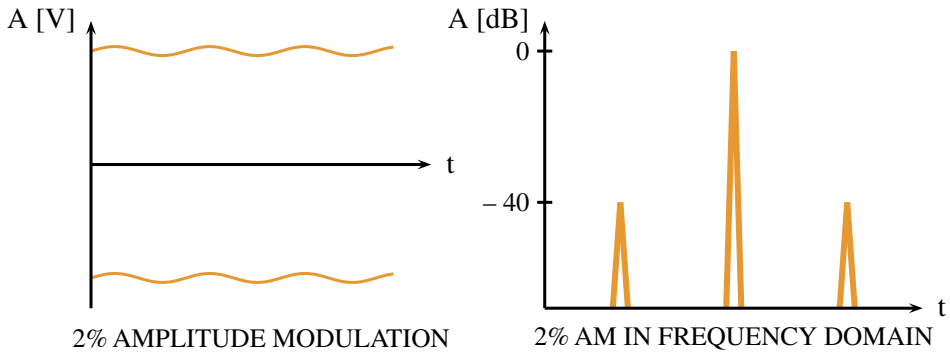


Fig. 14: Example of amplitude modulation in time and frequency domains

One of the major advantages of the frequency-domain visualization lies in the higher sensitivity to perturbations of periodic signals. For example, a 2% distortion of a sine-wave signal is already difficult

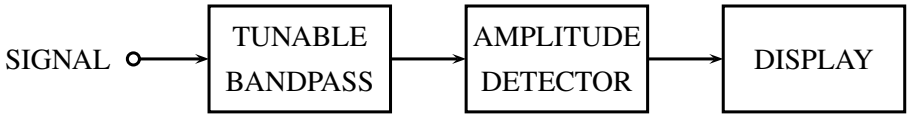


Fig. 15: A tunable bandpass as a simple spectrum analyser (SA)

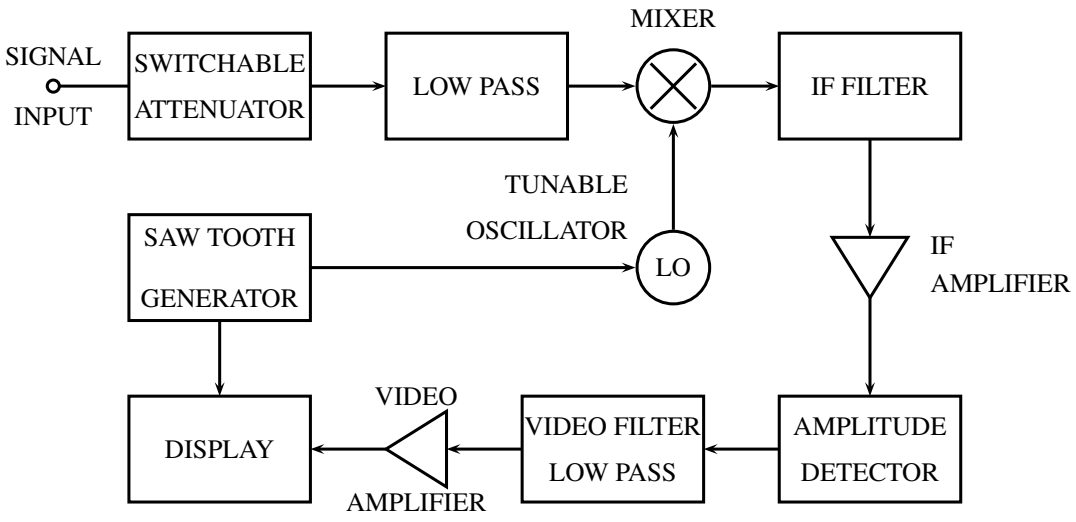


Fig. 16: Block diagram of a spectrum analyzer

to be observed on a the time domain display, but in the frequency domain on a logarithmic magnitude scale the related "harmonics" (Fig. 14) are clearly visible (here -40 dB below the main spectral line). A very faint amplitude modulation (AM) of 10^{-12} (power) on some sinusoidal signals would be completely invisible on a time domain trace, but can be displayed as two side harmonics 120 dB below the carrier in the frequency domain [5].

In the following we consider only "classical" SAs, based on a swept tuned band-pass filter analysis (Fig. 15), or utilizing the heterodyne receiver principle (Fig. 16).

The simplest form of a swept frequency spectrum analyzer is based on a tunable bandpass. This may be a classical lumped element LC circuit or a YIG filter (YIG = yttrium iron garnet) for frequencies >1 GHz. The LC filter exhibits poor tuning, stability and resolution. YIG filters are used in the microwave range (as preselectors) and for YIG oscillators. Their tuning range is about one decade, with Q values exceeding 1000.

For superior performance, the superheterodyne principle is applied basically in all commercial spectrum analyzers (Fig. 13). As already mentioned, the non-linear element (four-diode mixer or double-balanced mixer) delivers mixing products, like

$$f_{\text{signal}} = f_{\text{RF}} = f_{\text{LO}} \pm f_{\text{IF}}. \tag{8}$$

Assuming an input frequency range f_{RF} from 0 to 1 GHz for the spectrum analyzer shown in Fig. 16 and f_{LO} ranging between 2 and 3 GHz, results in a frequency chart as shown in Fig. 17.

Obviously, for a wide range of input frequencies, while rejecting any image response, requires a sufficiently high IF. A similar situation occurs for AM- and FM-broadcast receivers (AM-IF = 455 kHz, FM-IF = 10.7 MHz). But, for a high IF (e.g. 2 GHz) a stable, narrowband IF filter is very challenging,

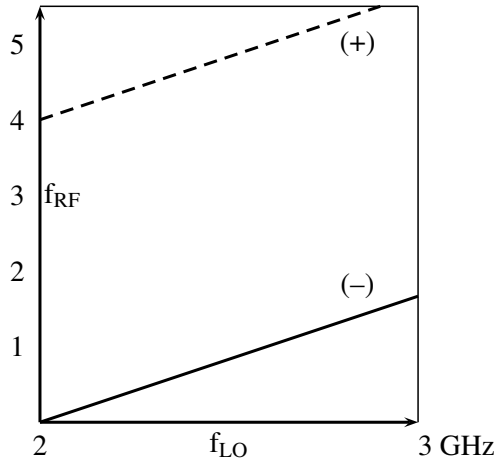


Fig. 17: Frequency chart of the SA of Fig. 16, f_{IF} = 2 GHz

therefore most SAs and high-quality receivers use more than a single IF. Certain SAs have four different LOs, some fixed, some tunable. To perform a large tuning range, the first, and for fine tuning (e.g. 20 kHz range), the third LO are variable.

Multiple mixing stages may also be necessary when downconverting to a lower IF (required when using high-Q quartz filters) to ensure a good image response suppression of the mixers.

It can be demonstrated that the frequency of the n^{th} LO must be higher than the (say) 80 dB bandwidth (BW) of the $(n-1)^{\text{th}}$ IF band-pass filter. A disadvantage of multiple mixing is the possible generation of intermodulation lines if amplitude levels in the conversion chain are not carefully controlled.

The requirements of a modern SA with respect to frequency generation and mixing are

- high resolution,
- high stability (drift and phase noise),
- wide tuning range,
- no ambiguities

and, with respect to the amplitude response

- large dynamic range (>100 dB),
- calibrated, stable amplitude response,
- low internal distortions.

It is important to notice that the bandwidth Δf of the IF band-pass filter is linked to sweep rate (or step width and rate when using a synthesizer):

$$\frac{\mathrm{d}f}{\mathrm{d}t} < (\Delta f)^2. \tag{9}$$

In other words, the signal frequency has to remain stable within $\Delta T = 1/\Delta f$ for a given IF bandwidth Δf, which ensures steady-state conditions of the selected IF filter.

On many instruments the proper relation between Δf and the optimum sweep rate is selected automatically, but it can always be altered manually (setting of the resolution bandwidth).

Caution is advised when applying, but not necessarily displaying, two or more strong (> 10 dBm) signals to the input. Third-order intermodulation products may appear (generated at the first mixer or amplifier) and could lead to misinterpretation of the signals to be analyzed.

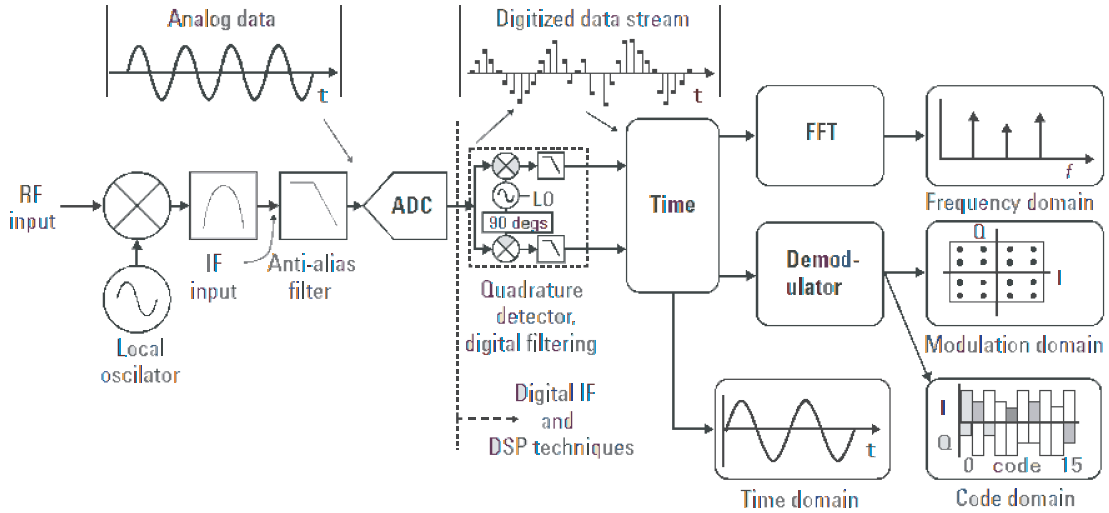


Fig. 18: Block diagram of a vector spectrum analyser

Spectrum analyzers usually have a rather poor noise figure of 20–40 dB, as they often do not use pre-amplifiers in front of the first mixer (dynamic range, linearity). But, with a good pre-amplifier, the noise figure can be reduced to almost that of the pre-amplifier. This configuration permits amplifier noise-figure measurements with a reasonable resolution of about 0.5 dB. The input of the amplifier to be tested is connected to the hot and cold terminations, and the two corresponding traces on the SA display are evaluated [6–10].

3.2 Vector spectrum and fast Fourier transform analyzer

The modern vector spectrum analyser (VSA) is essentially a combination of a two-channel digital oscilloscope and a fast *Fourier* transformation (FFT) based spectrum display. The incoming signal is down-converted, band-pass (BP) filtered, and passed to an analog-to-digital converter (ADC) (generalized Nyquist for BP signals; $f_{\text{sample}} = 2 \cdot \text{BW}$). Fig. 18 shows a typical, simplified schematic of a modern VSA.

The digitized signal is split into I (in-phase) and Q (quadrature, 90 degree offset) components with respect to the phase of some reference oscillator. Without this reference, the term “vector” would be meaningless for a spectral component.

One of the great advantages of a VSA, it easily allows to separate AM and FM components.

An example of vector spectrum analyzer display and performance is given in Figs. 19 and 20. Both figures were obtained during measurements of the electron cloud in the CERN Super Proton Synchrotron (SPS).

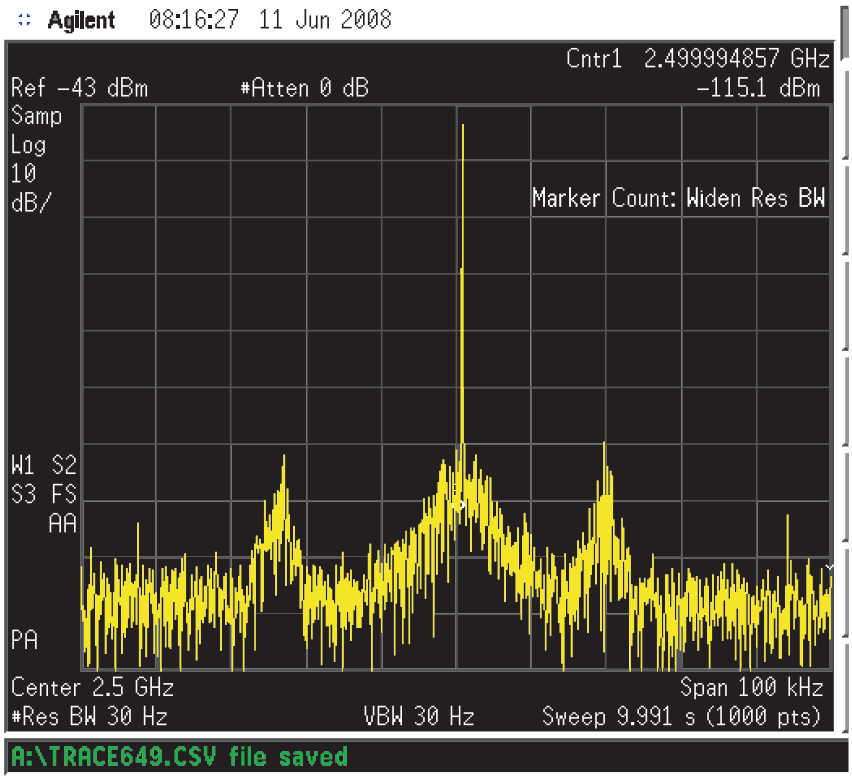


Fig. 19: Single-sweep FFT display similar to a very slow scan on a swept spectrum analyser

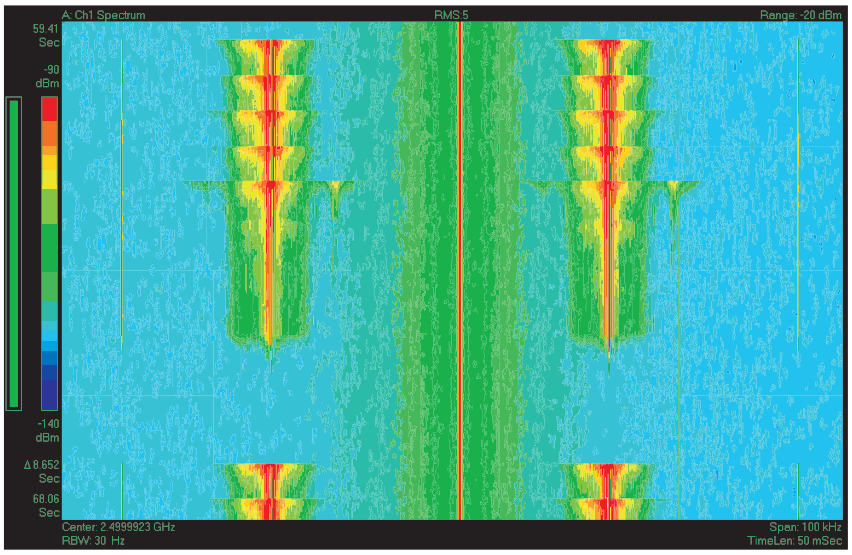


Fig. 20: Spectrogram display containing about 200 traces as shown on the left-hand side in colour coding. Time runs from top to bottom.

4 Noise basics

The concept of "noise" was originally studied for audible sound caused by statistical variations of the air pressure with a wide flat spectrum (white noise). It is now 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 characterize 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:

$$\frac{N_\mathrm{L}}{\Delta f} = hf\left(\mathrm{e}^{hf/kT} - 1\right)^{-1}, \tag{10}$$

where N_L is the noise power delivered to the load, $h = 6.625 \cdot 10^{-34}$ J s the *Planck* constant and $k = 1.38056 \cdot 10^{-23}$ J/K *Boltzmann's* constant.

Equation (10) indicates a constant noise power density up to about 120 GHz (at 290 K) with 1% error. Beyond, 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

$$W_\mathrm{r}(f,T) = \frac{hf^3}{c^2\left[\mathrm{e}^{hf/kT} - 1\right]}. \tag{11}$$

For $hf \ll kT$ the *Rayleigh–Jeans* approximation of Eq. (10) holds:

$$N_\mathrm{L} = kT\Delta f, \tag{12}$$

where in this case N_L is the power delivered to a matched load. The noise voltage $v(t)$ of a resistor R with no load is given as

$$\overline{v^2(t)} = 4kTR\Delta f \tag{13}$$

and the short-circuit current $i(t)$ by

$$\overline{i^2(t)} = 4\frac{kT\Delta f}{R} = 4kTG\Delta f, \tag{14}$$

where $v(t)$ and $i(t)$ are stochastic signals, and G is $1/R$. The linear averages $\overline{v(t)}, \overline{i(t)}$ vanishes, important are the quadratic averages $\overline{v^2(t)}, \overline{i^2(t)}$. The available power (which is independent of R) is given by (see also Fig. 21)

$$\frac{\overline{v^2(t)}}{4R} = kT\Delta f. \tag{15}$$

from which the spectral density function is defined as [6]

$$\begin{aligned} W_\mathrm{v}(f) &= 4kTR, \\ W_\mathrm{i}(f) &= 4kTG, \\ \overline{v^2(t)} &= \int_{f_1}^{f_2} W_\mathrm{v}(f)\mathrm{d}f. \end{aligned} \tag{16}$$

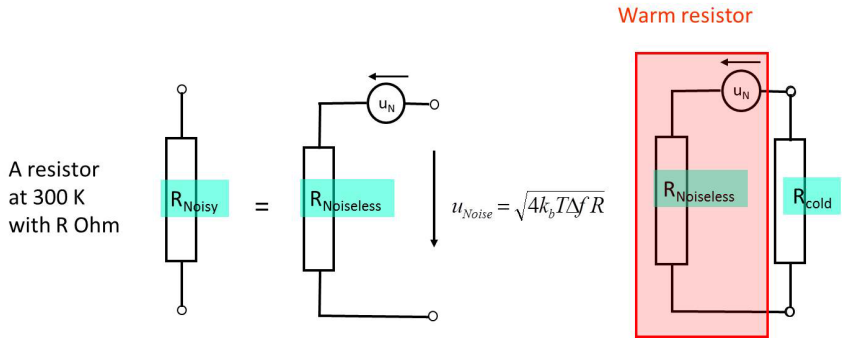


Fig. 21: Equivalent circuit of a noisy resistor terminated by a noiseless load

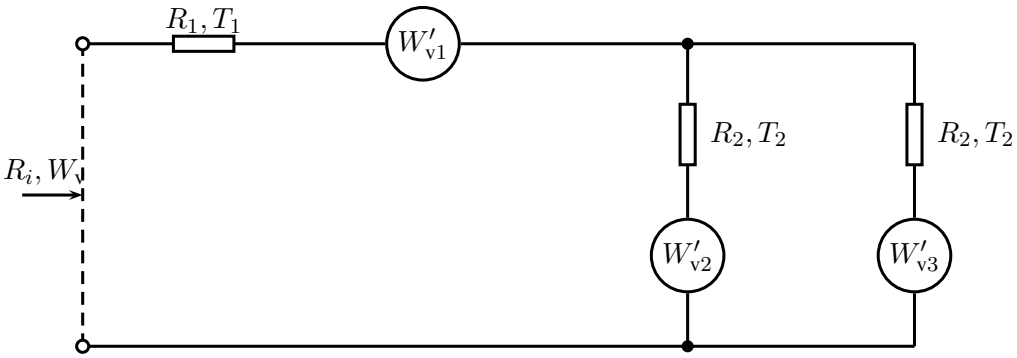


Fig. 22: Noisy one-port with resistors of different temperatures [6, 11]

A noisy resistor may be composed of many elements (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. 22), the spectral density function becomes

$$W_\mathrm{v} = \sum_j W_{\mathrm{v}j} = 4kT_\mathrm{n}R_\mathrm{i}, \tag{17}$$

where $W_{\mathrm{v}j}$ are the individual noise sources (Fig. 23), T_n is the total noise temperature, R_i the total input impedance, and β_j are coefficients indicating the fractional part of the input power dissipated in the resistor R_j. For simplicity it is assumed that all $W_{\mathrm{v}j}$ are uncorrelated.

The relative contribution (β_j) of a lossy element to the total noise temperature is equal to the relative dissipated power multiplied by its temperature:

$$T_\mathrm{n} = \beta_1 T_1 + \beta_2 T_2 + \beta_3 T_3 + \cdots = \sum_j \beta_j T_j \tag{18}$$

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 to 20 km height, causes a noise temperature at the antenna output of about 10 to 50 K. This is well below our room temperature of 290 K.

So far, only pure resistors have been considered. Looking at complex impedances, it is evident, losses occur only from dissipation in Re(Z). The available noise power is independent of the magnitude

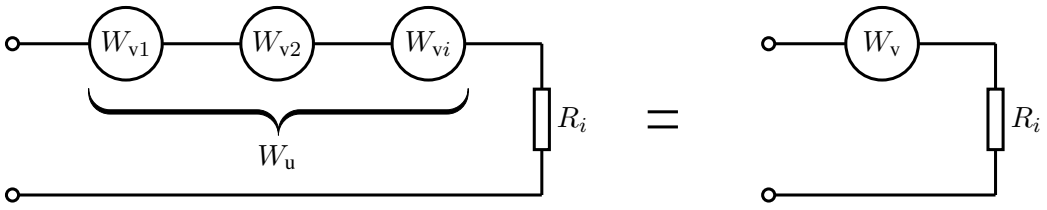


Fig. 23: Equivalent sources for the circuit of Fig. 22

of Re(Z) with Re(Z) > 0. For Figs. 22 and 23, Eq. (17) still applies, except R_i is replaced by Re(Z_i). However, in complex impedance networks the spectral power density W_v becomes frequency dependent [11].

The rules mentioned above apply to passive structures. A forward-biased Schottky diode (external power supply) has a noise temperature of about $T_0/2$ + 10%. A biased Schottky diode is not in thermodynamic equilibrium and only half of the carriers contribute to the noise [6]. But, it represents a real 50 Ω 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 at 290 K physical temperature.

4.1 Noise-figure measurements with the spectrum analyzer

Consider an ideal (noiseless) amplifier, terminated at its input (and output) with a load at 290 K with an available power gain (G_a). At the output we measure [7, 12]:

$$P_\mathrm{a} = kT_0 \Delta f G_\mathrm{a}. \tag{19}$$

For $T_0 = 290$ K (sometimes 300 K), we obtain kT_0 = −174 dBm/Hz (−dBm = decibel below 1 mW). At the input we determine for a given signal S_i a certain signal-to-noise ratio $S_\mathrm{i}/N_\mathrm{i}$, and at the output $S_\mathrm{o}/N_\mathrm{o}$, from what the noise factor F is defined as:

$$F = \frac{S_\mathrm{i}/N_\mathrm{i}}{S_\mathrm{o}/N_\mathrm{o}} \tag{20}$$

and its logarithmic equivalent NF follows as:

$$NF = 10 \log F \tag{21}$$

An ideal amplifier has $F = 1$ or $NF = 0$ dB. The noise temperature of this amplifier is 0 K, and signal and noise levels at the output are linearly increased by the gain. A real amplifier adds some noise, which leads to a decrease in $S_\mathrm{o}/N_\mathrm{o}$ due to the added noise N_a:

$$F = \frac{N_\mathrm{a} + N_\mathrm{i} G_\mathrm{a}}{N_\mathrm{i} G_\mathrm{a}} = \frac{N_\mathrm{a} + kT_0 \Delta f G_\mathrm{a}}{kT_0 \Delta f G_\mathrm{a}}. \tag{22}$$

For a linear system the minimum noise factor amounts to $F_\mathrm{min} = 1$ or $NF_\mathrm{min} = 0$ dB, however, for non-linear systems one may experience a noise factor $F < 1$.

Noise factor and noise temperature are related by

$$T_\mathrm{e} = \frac{N_\mathrm{a}}{k \Delta f G_\mathrm{a}} = T_0 (F - 1) \tag{23}$$

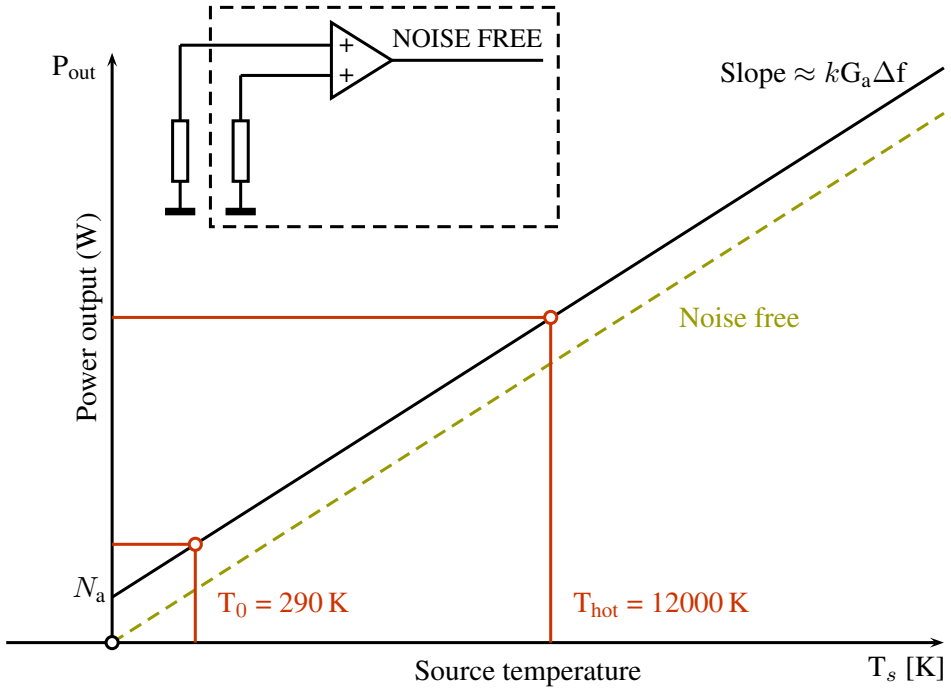


Fig. 24: Relation between source noise temperature T_s and output power P_out for an ideal (noise-free) and a real amplifier [7, 12].

with T_e being the equivalent temperature of a source impedance into a perfect, noise-free device that would produce the same added noise N_a [12].

The so-called 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, e.g. 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):

$$ENR_\mathrm{dB} = 10 \log \left(\frac{T_\mathrm{h} - T_\mathrm{c}}{T_0} \right) \tag{24}$$

For most noise figure calculations the linear form is more useful:

$$ENR = 10^{\frac{ENR_\mathrm{dB}}{10}} \tag{25}$$

The noise source is connected to the amplifier or DUT to be analyzed, providing noise "on" (N_2) and "off" (N_1) conditions. The ratio of these noise powers is called the Y*-factor*:

$$Y = \frac{N_2}{N_1} \tag{26}$$

Y-factor and ENR can be used to determine the noise slope of the DUT, as illustrated in Fig. 24. 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:

$$N_\mathrm{a} = kT_0 \Delta f G_1 \left(\frac{ENR}{Y - 1} - 1 \right) \tag{27}$$

The SA, operating in automatized *noise figure mode*, 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 – the total *system noise factor*

$$F_{\text{sys}} = \frac{ENR}{Y - 1} \tag{28}$$

which includes noise contributions from all parts of the system. In case the "cold" noise temperature $T_{\text{c}} \neq T_0 = 290$ K, Eq. 28 becomes

$$F_{\text{sys}} = \frac{ENR - Y(T_{\text{c}}/T_0 - 1)}{Y - 1} \tag{29}$$

For low ENR noise sources, $T_{\text{h}} < 10\,T_{\text{c}}$, an alternative equation holds:

$$F_{\text{sys}} = \frac{ENR(T_{\text{c}}/T_0)}{Y - 1} \tag{30}$$

If Y is close to 1, i.e. $F_{\text{sys}} \gg ENR$, the system noise factor "masks" the noise generated by the noise source, making an accurate measurement difficult or impossible. Therefore the Y-factor method is limited to noise figure measurements with $NF \approx 10$ dB below the ENR of the noise source.

The literature explains a variety of other noise figure measurement methods [6,8–10,13], including the "3 dB" method [12] for the measurement of high noise figure devices, where the Y-factor method is limited.

The noise figure of a cascade of amplifiers is given as [6, 7, 11–13]

$$F_{\text{total}} = F_1 + \frac{F_2 - 1}{G_{\text{a1}}} + \frac{F_3 - 1}{G_{\text{a1}} G_{\text{a2}}} + \cdots . \tag{31}$$

As Eq. (31) shows, the first amplifier in a cascade has a dominant effect on the total (system) noise figure, provided G_{a1} is not too small and F_2 not too large. In order to select the best amplifier from a number of different units to be cascaded, the noise measure M

$$M = \frac{F - 1}{1 - (1/G_{\text{a}})}. \tag{32}$$

helps to select the optimal unit:
The amplifier with the smallest M should be selected as first unit in the cascade [12].

5 Introduction to network analysis and S-parameters

One of the most common measurement tasks in the field of RF engineering is the analysis of circuits and electrical networks. Such networks can be a simple one-port (two-pole), containing only a few passive components (resistors, inductors and capacitors) or they may be complex units, consisting of passive, active and/or non-linear components with several input and output ports.

A vector network analyzer (VNA) is one of the most versatile and valuable pieces of measurement equipment used in a RF laboratory or particle accelerator control room. The network analysis is performed by exciting the device under test (DUT) with a well-defined input signal in terms of frequency and amplitude, and recording the response of the network, for each frequency step as complex value of the reflection and/or transmission coefficients. These are the coefficients of the scattering parameters (S-Parameter), the properties to characterize a DUT at RF and microwave frequencies. The best commercially available network analyzers can cover a frequency range of ten (and more) orders of magnitude (from a few Hz to many GHz), with a resolution down to 0.1 Hz.

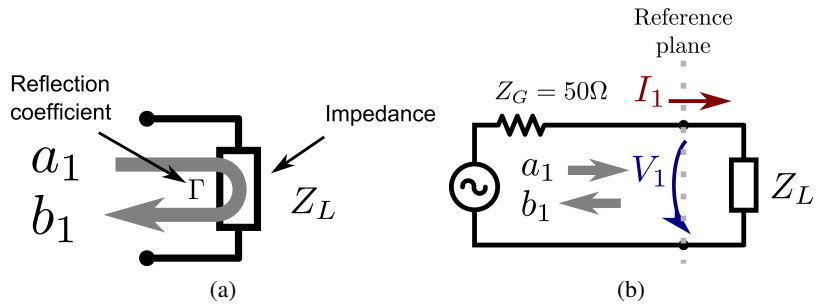


Fig. 25: Wave quantities of a one-port (with two poles) and impedance Z_L: (a) incident (a_1) and reflected (b_1) wave; (b) relation of a_1 and b_1 to V_1 and I_1.

In the following sections, scalar and vector network analyzers are introduced and measurement techniques for the determination of S-parameters of networks are discussed. S-parameters are basically defined only for linear networks. In the real world, many DUTs are at least weakly non-linear (e.g. mixers, or active elements such as amplifiers). For the analysis of these devices certain approximations or extensions of the definitions are required [15].

Another interesting application is the determination of the beam transfer function (BTF), where the DUT is a circulating particle beam in an accelerator.

5.1 One-port networks

In RF engineering, *wave quantities* are preferred in favor currents or voltages for the characterization of RF circuits. We can distinguish between incident (a) and reflected waves (b). The incident wave travels from a source to the DUT – the reflected wave travels in the opposite direction. This terminology is preferred, because in RF engineering the linear geometrical dimensions of a circuit often are larger than 10% of the corresponding free-space wavelength. Wave functions are defined in time and *spacial* coordinates, and for this fact are preferred to voltages and currents, which typically are only defined in time. This also requires the definition of a reference plane, i.e. the physical location in space to which the measurement refers. Without this reference plane, e.g. the phase of the reflection coefficient would be undefined, which would make vectorial measurements impossible. Of course, a mathematically correct description of the DUT in terms of voltages and currents still holds, and also will return correct results, but working with wave quantities turns out to be much more convenient in practice. Both network discription methods – if correctly applied – have no fundamental limitation, e.g. S-parameters can be used at very low frequencies and voltage and current descriptions can also be used at very high frequencies. Both methods are fully equivalent, for any frequency; the results are mutually convertible. This fact is expressed by conversion rules, namely S-parameters can be converted into impedances and vice versa.

The interface of the DUT to the outside world is utilized by one or more *pole pairs*, which are commonly referred as *ports*. A device with one pair of poles (as in Fig. 25a) is defined as one-port, where one incident (a_1) and one reflected (b_1) wave can propagate simultaneously. The index of the wave quantities represents the number of the port.

The wave quantities can be determined from the voltage and current at the port. They are related to each other

$$a_1 = \frac{V_1 + I_1 Z_0}{2\sqrt{Z_0}}, \qquad b_1 = \frac{V_1 - I_1 Z_0}{2\sqrt{Z_0}}, \tag{33}$$

where V_1 and I_1 represent the voltage and current respectively at the port as depicted in Fig. 25b. Z_0 is an arbitrary reference impedance (often, but not necessarily always, the characteristic impedance $Z_0 = Z_\mathrm{G} = 50\ \Omega$ of the system).

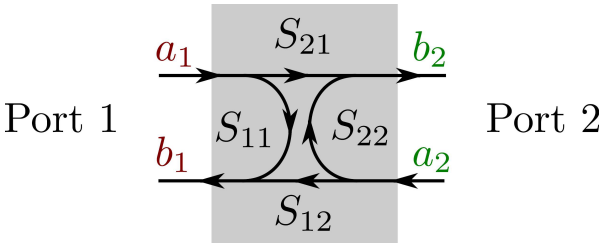


Fig. 26: All possible S-parameters of a two-port network

The wave quantities have the dimension of $\sqrt{W}$ (see [14]). This normalization is important for the conservation of energy. The power traveling towards the DUT is calculated by $P_{\text{inc}} = |a|^2$, the reflected power by $|b|^2$. It is important to note that this definition is mainly used in the USA – in European notation, the incident power is usually calculated by $P_{\text{inc}} = 0.5|a|^2$. These conventions have no impact on the calculation of S-parameters and only need to be considered when the absolute power is of interest.

The reflection coefficient Γ represents the ratio between the incident wave and the reflected wave of a specific port. It is defined as

$$\Gamma = \frac{b_1}{a_1}. \tag{34}$$

By substitution with Eq. (33), we can find a relation between the complex (load) impedance Z_L of a one-port and its complex reflection coefficient Γ:

$$\Gamma = \frac{Z_L - Z_0}{Z_L + Z_0}. \tag{35}$$

5.2 Two-port networks

For electrical networks with two ports (e.g. attenuators, amplifiers) we find more quantities to be measured. Besides the reflection coefficients for each port, the transmission in forward and reverse directions also needs to be characterized. We now require the definition of the scattering parameters (S-parameters) for two ports. The idea is to describe how the incident energy on one port is scattered by the network and exits through the other ports. All possible signal paths through a two-port are shown in Fig. 26. A two-port has four complex, frequency-dependent scattering parameters:

$$S_{11} = \frac{b_1}{a_1}, \qquad S_{12} = \frac{b_1}{a_2}, \qquad S_{21} = \frac{b_2}{a_1}, \qquad S_{22} = \frac{b_2}{a_2}. \tag{36}$$

Here S_{11} and S_{22} are equal to the reflection coefficients Γ of their respective ports – but *only* under the condition that the corresponding other port is terminated in its characteristic impedance. S_{21} and S_{12} are the forward and reverse transmission coefficients, respectively. The first index of the S-parameter defines at which port the outgoing wave is observed, the second index defines at which port the network is excited. This leads to the counterintuitive appearing situation, that for forward transmission the corresponding S-parameter is S_{21}, not S_{12}. The S-parameters are measured following exactly the same definition. The internal source of the network analyzer excites an incident wave on port one, namely a_1. Now b_1 and b_2, the outgoing waves from the DUT, are measured, which allows the determination of S_{11} and S_{21} (provided that port one and port two are terminated with their characteristic impedances).

It is very important to *always* terminate all ports of the DUT with their respective characteristic impedances. In many situations this is Z_0, but there are cases where the characteristic impedance is different between port one and port two, e.g. a transformer with a turns ratio of two, leading to an impedance transformation by a factor of four. In this case the characteristic impedance would be for port one 50 Ω and for port two 12.5 Ω.

The termination prevents unwanted reflections and ensures the DUT is only excited by a single incident wave. For practical S-parameter measurements this implies that any port of the DUT needs to be connected to a matched load corresponding to the characteristic impedance of this port. This rule includes in particular the port connected to the VNA output port, or in other words, the generator impedance has also to match the impedance of the DUT. For example, the analysis of a DUT with 25 Ω characteristic impedance is not simply straightforward on a 50 Ω network analyzer, unless special care is taken.. But, permitting a modern VNA, applying a special calibration procedure allows the modification of the characteristic impedance of each VNA port to any value (within a reasonable range from $> 5\ \Omega$ to $< 500\ \Omega$), and in this way to adapt to the requirements of the DUT. However, the situation of the termination of ports becomes more complicated for the characterization of beam elements, like beam pickups, kickers, and accelerating structures, where strictly speaking the beam (waveguide) ports also need to be terminated in their characteristic impedance. Often simple solutions can be applied, like microwave absorbing foam, to avoid unwanted reflections from open beam ports.

The S-parameters are an intrinsic property of the DUT and not a function of the incident power used for the measurement (condition of linearity). Obviously, the S-parameters measured shall be independent of the instrumentation used to perform the measurement.

Once all n^2 S-parameters for a given n-port network are measured, the properties of this network can be described by a set of linear equations. For incident waves a_1 and a_2 of arbitrary phase and magnitude on a two-port, the outgoing or scattered waves b_1 and b_2 can be determined

$$\begin{aligned} b_1 &= S_{11}a_1 + S_{12}a_2, \\ b_2 &= S_{21}a_1 + S_{22}a_2. \end{aligned} \tag{37}$$

These equations can be written in matrix format, for convenience:

$$\vec{b} = \mathbf{S}\,\vec{a} \tag{38}$$

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}. \tag{39}$$

The S-matrix is a linear model of the DUT. Its diagonal elements represent the reflection coefficients of each port. The remaining elements characterize all possible signal transmission paths between the ports. S-parameters are in general complex and a function of frequency. The set of linear equations given by the S-matrix must be solved for a single frequency at a time. S-parameters are typically acquired over a certain frequency range (span) for a number N of discrete, equidistant frequency steps. With N data points, the system of equations has to be solved N times. A discussion of the general properties of the S-matrix can be found in [14].

6 Scalar network analysis

A scalar network analyzer measures only the amplitude, i.e. the magnitude of a – reflected or transmitted – signal, the phase is not available. Consequently, only the absolute value (the magnitude) of the complex S-parameters can be obtained. Today scalar network analyzers are basically obsolete, however, some key components and circuits are also found in VNAs, making this instrument a methodical way to introduce the concept of network analysis.

A simple network analysis set-up, as it was used more than 50 years ago, is shown in Fig. 27. The measurement is performed in two steps, in the first step (Fig. 27, left) without the DUT to measure the power of the incident signal (V_1). Then the DUT is inserted (Fig. 27, right), and V_2 is measured. Following the magnitude of the transmission coefficient is calculated:

$$|S_{21}| \propto \frac{V_2}{V_1}. \tag{40}$$

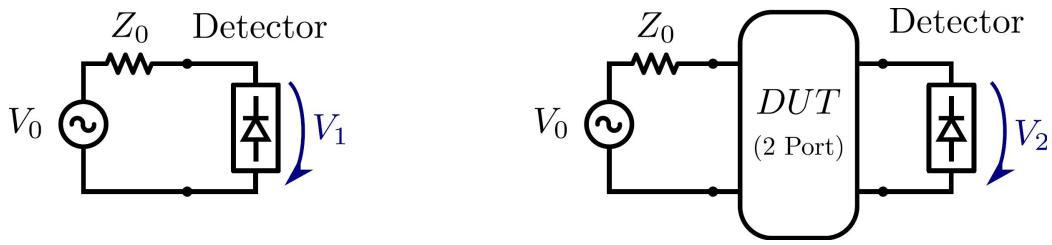


Fig. 27: A simple measurement set-up for the scalar transmission coefficient ($|S_{21}|$)

To obtain the results in decibels, a logarithmic amplifier was connected to the output of the detector. It has a logarithmic transfer function ($V_{\text{out}} = \log V_{\text{in}}$) and permits the display of a large dynamic range on a dB scale. Furthermore, mathematical operations like multiplication or division, e.g. required for normalization in Eq. (40), transforms simply into into an addition or subtraction, handled by operational amplifiers.

As detector any kind of device converting the input RF signal into a DC voltage is applicable, assuming its transfer function is "reasonable"[3] proportional to the RF power. There are basically three possibilities to achieve this:

Rectifier A fast *Schottky* diode and a low-pass filter are used to convert the input RF signal to a DC voltage. Operating the diode in its square-law region ($P_{\text{in}} < -10$ dBm) results in an output voltage proportional to the RF power; see Section 2.2.
Advantages: cheap, fast response (depending on $f_{\max}$ of the output filter).
Limitations: Commercially available RF power meters, based on *Schottky* diodes, can operate from -60 dBm (limited by tangential sensitivity) up to about +30 dBm (damage level). The non-linearity of the output signal versus input power is compensated by electronic means (look-up table). Coaxial RF *Schottky* detectors are usually limited to maximum frequencies of approximately 100 GHz, essentially determined by the coaxial connector technology available. Usually an input matching network is required to match the input impedance of the *Schottky* diode to $Z_0 = 50\ \Omega$.

Thermal measurement Several types of detectors based on heating effects are available for the measurement of RF power. In a bolometer (thermistor or barretter), the high temperature coefficient of the thermal conductivity of certain metals or metal alloys is exploited. The temperature change ΔT of dissipated heat of the RF input signal is measured utilizing a DC-based temperature measurement, while applying a correction of the non-linearities. Barretters utilize the positive temperature coefficient of metals like tungsten and platinum. Thermistors consist of a metal oxide with a strong negative temperature coefficient. Another class of RF power meters based on heating is the thermo-element, which takes advantage of the thermo-electrical coefficient of a junction between two different metals. A well-known example is the Sb-Bi junction, which has a temperature coefficient of about 10^{-4} V/K, which is one of the highest values available for this kind of detector. Even larger values can be achieved using semiconductor–metal junctions, where thermoelectric coefficients of 250 μV/K have been achieved. For further details, see [16].

Mixer Multiplying two sinusoidal signals with different frequencies results in signals of sum and difference frequencies at the multiplier's output; see Section 2.3. Technically this frequency mixing principle allows to convert a range of high-frequency signals to a much lower intermediate frequency (IF) band. Now the RF power measurement is performed in simpler ways at this IF.

[3] With the term "reasonable" we point out the fact, that many detectors have a non-linear relation between input power and output voltage.

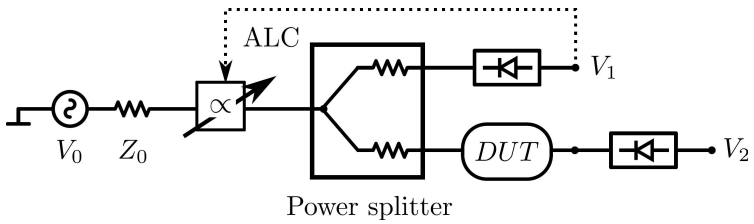


Fig. 28: Simplified circuit diagram of a typical automatic gain control

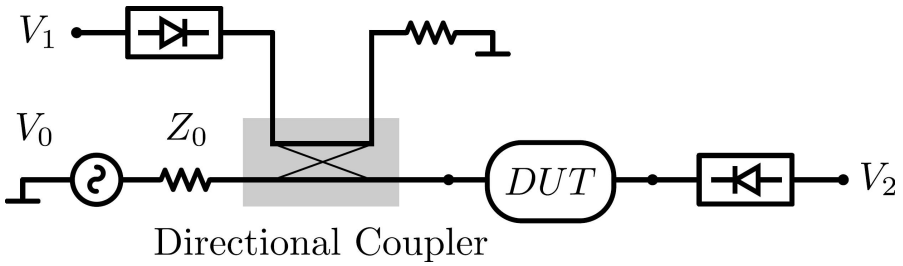


Fig. 29: Feedback loop of a typical automatic gain control (AGC)

6.1 Automatic Gain Control (AGC)

Often RF measurements are performed over a wide range of frequencies, requiring the signal strength, i.e. the amplitude V_0 of the source to be constant. This is usually achieved by an active feedback loop (*levelling*), keeping V_0 constant, independent of the operation frequency. Any feedback loop requires a process variable which has to be detected and controlled to a well defined set point, here the output signal level V_1. For the automatic gain control (AGC) loop in a NA e.g. a resistive power divider can be used to provide this reference signal, while keeping inputs and outputs matched to $Z_0 = 50\ \Omega$ (Fig. 28). For this example, the test signal arriving at the DUT is reduced by 6 dB due to the insertion loss of the resistive power divider. However, the AGC feedback loop ensures the stimulus signal applied to the DUT has always a constant, well defined power level over a wide frequency range.

For the characterization of linear DUTs, only the ratio V_2/V_1 is of interest, which is independent of the absolute value of V_0. In this case the S-parameter measurements do not require an AGC loop of the RF generator, but in practice the gain control has many advantages, in particular for measurements on weakly non-linear elements, such as amplifiers.

6.2 Directional couplers

Replacing the resistive power divider by a directional coupler reduces the insertion loss substantially, the principle is outlined in Fig. 29. V_1 is an attenuated replica – defined by the coupling factor – of the forward-traveling wave, which is only used for as reference for the gain control. Typically, directional couplers with a coupling coefficient of -20 dB are used for the purpose, they offer a transmission attenuation in the main branch of less than 0.3 dB. In contrast to the resistive power splitter, the transmission-line based directional coupler has a limited frequency range, and therefore other issues.

Modern network analyzers (both scalar and vectorial versions) measure the forward-transmission, as well as the reflection coefficient of a DUT simultaneously, without the need to manually re-connect DUT ports. Each port of the instrument is equipped with a dual directional coupler, providing simultaneously replicas of the incident and reflected waves from the DUT, see Fig. 30. These directional couplers, in combination with some required switches and attenuators are commonly called *test set*. In the early days, network analyzers consisted of separate building blocks, like S-parameter test set, frequency generator, display and controller unit. All these elements had to be connected by many external cables. Modern instruments have all those building blocks integrated in a single unit, including advanced

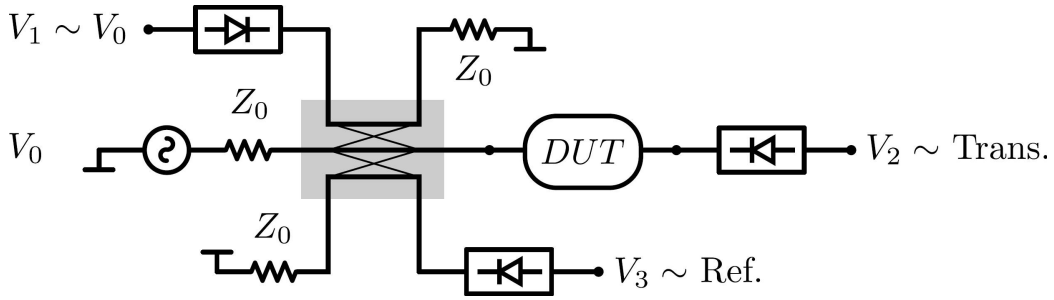


Fig. 30: Dual directional coupler in a network analyzer

computer controls with digital data acquisition and post-processing.

Based on Fig. 30, the reflection and transmission coefficients are defined as

$$|S_{11}| \propto \frac{V_3}{V_1}, \qquad |S_{21}| \propto \frac{V_2}{V_1}. \tag{41}$$

From the ratio of the reflected wave to the incident wave (S_{11}), valuable quantities like standing wave ratio (SWR), reflection coefficient, impedance, admittance as well as return loss of the DUT are determined. From the ratio of the transmitted wave to the incident wave (S_{21}), gain resp. insertion loss, the transmission coefficient, the insertion phase, and group delay of the DUT can be characterized.

7 Vector measurements

A vector network analyzer (VNA) is able to measure the magnitude *and phase* of a complex S-parameter. There are different hardware configurations which implement this kind of RF instrument, e.g. six-port reflectometers, certain RF bridge methods, or superheterodyne RF network analyzers. Here only the latter will be introduced.

7.1 The modern vector network analyzer

A modern VNA contains a RF generator which produces the signal stimulating the DUT. This signal is usually generated by a synthesizer-type oscillator and is adjustable in very fine steps over a large frequency range, in a programmable manner. Since all modern VNAs operate with analog and/or digital downconverters (mixing), the generation of a tracking LO frequency is also necessary. This tracking LO is typically generated by PLL circuits and represents essentially a second oscillator following the main frequency with a specified frequency offset.

The observation (IF) band signal is typically processed digitally, allowing bandwidth settings over a wide range, e.g. 1 Hz to 20 MHz and more. In all stages of the signal path the vectorial nature of the signal is preserved, both phase and magnitude are processed, in the digital domain usually as I-Q (in-phase – quadrature-phase) data, equivalent to real and imaginary parts. Details on the internal signal processing of a VNA are found in [17, 18]. Note, similar to the spectrum analyzer, the sweep time and resolution bandwidth cannot be adjusted independently. A modern four-port vector network analyzer is shown in Fig. 31.

Although complete network analysis of any N-port can be performed with a two-port VNA, a four-port unit is extremely convenient for many measurement tasks. It permits a quick analysis, e.g. of a directional coupler or a three-port circulator without the need for swapping cables, it also introduces virtual ports of balanced nature, and many other valuable features.

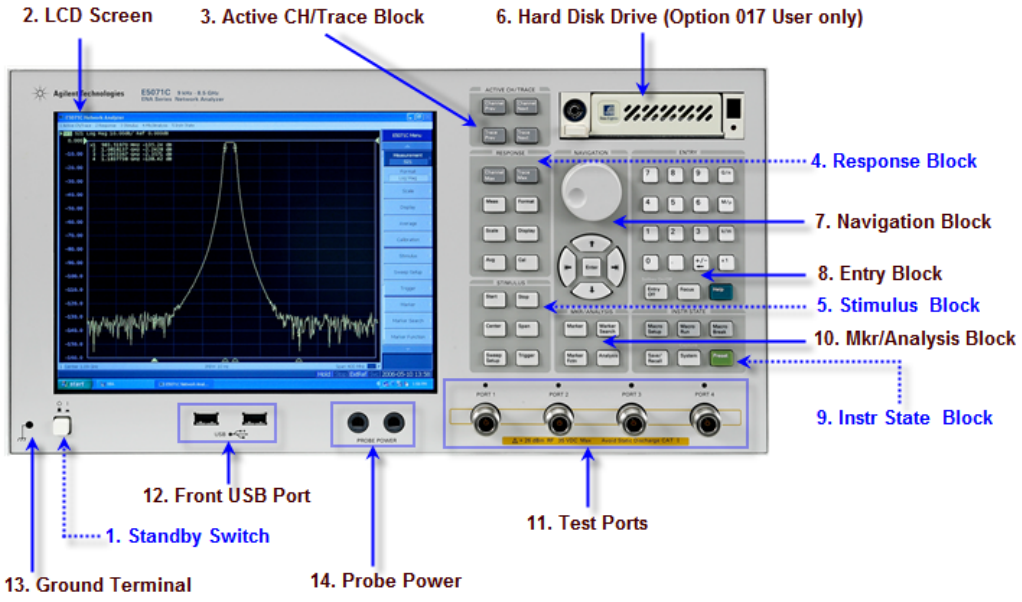


Fig. 31: A modern four-port VNA

7.2 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 (fast) *Fourier* transformation[4] and vice versa, assuming the entire frequency vector data (magnitude and phase, or real and imaginary) is present. This is the basis of the synthetic pulse technique, available on many modern VNAs. It was commercially introduced by Hewlett-Packard in the 1980s for network analyzer applications.

It renders the VNA even more versatile, allowing to display the impulse (*Gaussian*) and/or step response of the DUT, and to perform time-domain reflectometry (TDR) measurements. Typical applications of this measurement techniques are:

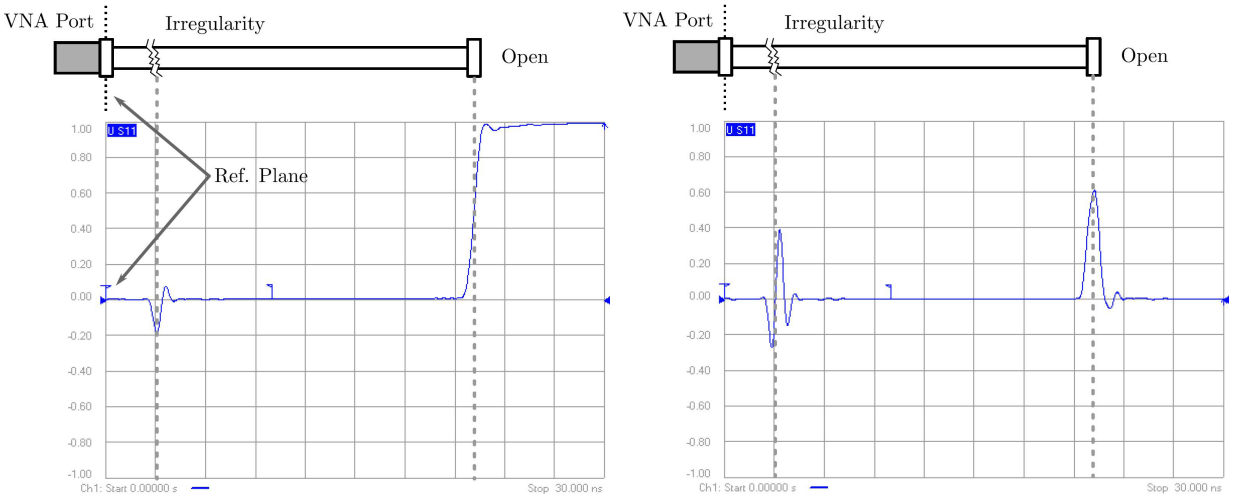
1. Localizing and evaluating discontinuities (faults) in transmission lines.
2. Separating the scattering properties of sections of complicated RF networks by time-domain gating.
3. Echo cancellation (in multipath environments).
4. The synthetic pulse time-domain reflectometry can be very useful in trouble-shooting, e.g. of the accelerator beam-pipe. By using waveguide modes it was successfully used to detect an obstacle in the LHC beam-pipe.

The only constraint of the applicability of the synthetic pulse measurement technique, the DUT has to be a *linear* and *time-invariant* (LTI) system.

A measurement example is shown in Fig. 32. A transmission line with a given length and some perturbation is connected to a calibrated VNA. The real part of the *Fourier*-transformed reflection coefficient ($S_{11}(\omega)$) is plotted versus time. The VNA permits the display of either, the synthetic step (Fig. 32a) or the impulse response (Fig. 32b). The step is simply obtained by (numerical) integration of the impulse response data.

The incident synthetic pulse is scattered from the discontinuity, but also from the open end of the

[4]More precisely: by a discrete *Fourier* transformation (DFT). The fast *Fourier* transformation (FFT) is just an optimized form of th eDFT, exploiting the symmetry of 2^n data samples, thus saving computation time. However, both algorithms will produce the same result for the same input data.



(a) (b)

Fig. 32: Synthetic pulse measurement with a VNA: (a) step response; (b) impulse response.
The measured frequency data is converted by an inverse discrete Fourier transformation (iDFT) to the time domain. Now the synthetic impulse response of a transmission-line, here a coaxial cable, is displayed over time. The reflections of the incident pulse on any irregularity or discontinuity, as well as the end of the cable are clearly identified. By measuring the time delay between the the reference plane and the location of the irregularity, or end of the cable (displayed as pulse or step in the reflection coefficient) the electrical length of the cable can be calculated.

cable. The travel time for the pulse can be read on the horizontal axis on the time-domain display. In this example we measure a delay of $t_d = 22$ ns until the open end of the cable becomes visible. This time accounts for the impulse traveling towards the open end *and back*; thus, the factor $1/2$ has to be taken into account when calculating physical length l of the transmission-line:

$$l = \frac{c}{\sqrt{\varepsilon_r}} \cdot \frac{1}{2} t_d. \tag{42}$$

In this example the relative dielectric constant of the insulation in the coaxial cable is $\epsilon_r = 2.3$ (PTFE Teflon), which returns a cable length of $l = 2.2$ m. The same method can be applied for obtaining the position of any irregularity or discontinuity (deformation, bad connector) along the cable. Nearly all VNAs with time-domain option permit the designation of the velocity factor ($1/\sqrt{\varepsilon_r}$ for a homogeneously filled transmission line) and thus convert travel time or electrical length to physical distance on the display.

Note that the step response shown in Fig. 32a returns the local reflection factor versus time. Along the cable it amounts to $\Gamma = 0$, except for the position of the irregularity, indicating a well-matched 50 Ω transmission line. At the end we notice a positive step to $\Gamma = 1$, indicating an open circuit (see Table 2).

The reflected pulse in the impulse response trace (Fig. 32b), related to the open end of the cable does not reach unit amplitude due to fact of cable attenuation of the transmission line used for this example – a semi-rigid coaxial cable approximately 2 m length. The amplitude of this reflection from the open end indicates the attenuation over twice the electrical length of the cable at the equivalent center frequency ($f_{\text{max}} = 3$ GHz, $f_{\text{centre}} = 1.5$ GHz) of the measurement.

For practical applications of the synthetic pulse technique, certain basic properties of the discrete Fourier transform should be kept in mind, they are summarized in Table 3. For example, a long cable needs to be tested. This requires a long time window to ensure all multiple refections have decayed to zero, which needs attention to ensure a sufficient narrow frequency sampling. The time interval Δt is

Table 2: Important values of the reflection coefficient

DUT	Z_L	Γ
Open circuit	∞	+1
Short circuit	0	–1
Matched load	Z_0	0
Load	$Z_0/2$	–1/3
Load	$2Z_0$	1/3

Table 3: Important characteristics of the FFT

Time domain		Frequency domain
T_{max} (time span)	$\leftrightarrow$	Δf (frequency resolution)
Δt (time resolution)	$\leftrightarrow$	f_{max} (frequency span)

related to $1/\Delta f$, and this reciprocal relation may cause issues if settings are kept in "automatic" mode.

On the other hand, if a bad connector or cable damage needs to be located along a transmission-line, a high resolution in time is required. Thus, the VNA has to measure over a wide frequency span (f_{max}). Obviously, we would like to often use both, a high frequency span and a close spacing of the samples in the frequency domain, but there are practical limitations: namely, the number of data points available. Usually in modern instruments the number of data points available amounts to 60000 and, depending on the application, compromises have to be accepted.

Performing time-domain measurements with the vector network analyzer calls for two basic modes, the "low-pass", or the "band-pass" mode to be selected.

7.2.1 Low-pass mode

In low-pass mode the basic discrete *Fourier* transformation algorithm is applied. This returns certain constraints on the frequency-domain measurement data of the DUT (Fig. 33a). The iDFT demands a start frequency to always be 0 Hz (DC), and data is acquired in equidistant frequency steps between start and stop frequency. Since most VNAs cannot measure at very low frequencies, the data points from DC to the minimum operation frequency of the VNA are extrapolated mathematically. Data points for negative frequencies are derived from the measured samples on the corresponding positive frequencies by complex conjugation. Compared to the bandpass mode, this effectively doubles the number of data points available for the calculation of the time trace. For this particular symmetry, the discrete *Fourier* transformation returns a purely real-valued time trace. A practical time domain reflectometry (TDR) measurement routine is setup as follows:

1. The DUT is connected, the port and type of measurement are selected (transmission or reflection).
2. The frequency range of interest and the number of data points are entered (this relates to the time domain by Table 3)
3. After pushing the soft key, "set frequency low-pass"[5], the instrument choses the exact sampling frequencies.
4. Once the sampling points are defined, the VNA has to be calibrated (open, short, load for reflection measurements).

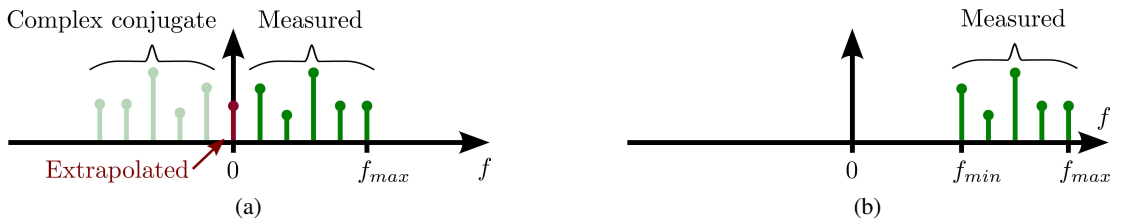


Fig. 33: Sampling of frequency points for the different operating modes: (a) low-pass mode; (b) bandpass mode.

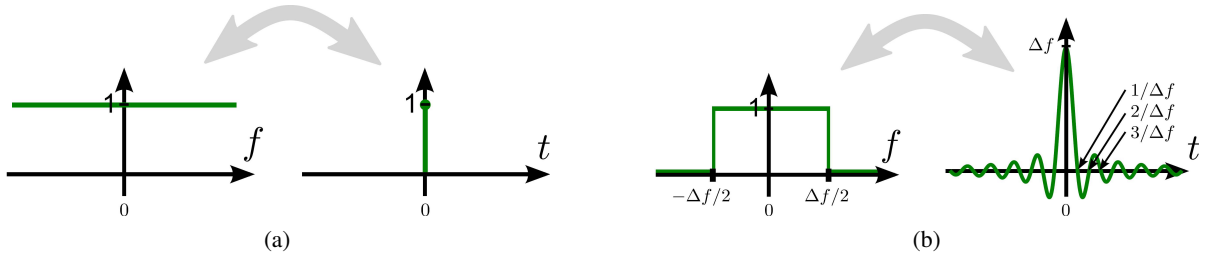


Fig. 34: (a) Infinite frequency span. (b) Limited frequency span. The limited frequency span Δf [6]of the VNA leads to "distortions" of the time-domain synthetic pulse measurement. The ideal response is convoluted with a *sinc* function, which characteristics depend on Δf.

In the low-pass mode, the trace appearing on the screen for time domain reflectometry (TDR) or time domain transmission (TDT) is basically equivalent to what a real-time TDR or sampling oscilloscope display; see Section 7.2.6.

7.2.2 Band-pass mode

In band-pass mode (Fig. 33b) the spectral lines (frequency-domain data points) no longer need to be equidistant, and extrapolated down to DC, they just need to cover the frequency range of interest, e.g. from $f_{\text{min}} = 1.2$ GHz to $f_{\text{max}} = 1.5$ GHz. The start and stop frequencies of the VNA can be chosen arbitrarily, which returns a high degree of flexibility and is especially suited for the measurement of devices having a limited frequency range (example: waveguide-mode reflectometry).

The bandpass mode is the equivalent to a narrowband TDR (and also time-domain transmission TDT) using the synthetic pulse technique. It permits the display of the impulse response only, since no extrapolated information on a DC component is available. The measurement clearly identifies position and size of perturbations along a transmission line, including waveguides. Their characterization in terms of capacitive, inductive or resistive properties is possible, but not straightforward [19]. Details on the general properties and mathematical backgrounds of the low-pass and bandpass modes are found in [18, 20].

7.2.3 Windowing

As the VNA always samples a limited frequency spectrum, starting at f_{min} and stopping at f_{max}, the acquired spectrum is clipped by a rectangular envelope. Performing the iDFT, rectangular windowing artifacts show up in the time-domain data, as compared in Fig. 34.

[5]This soft key may appear with slightly different naming, depending on the definitions of the manufacturer.
[6]not to be confused with the previous definition of Δf for the equidistant frequency samples

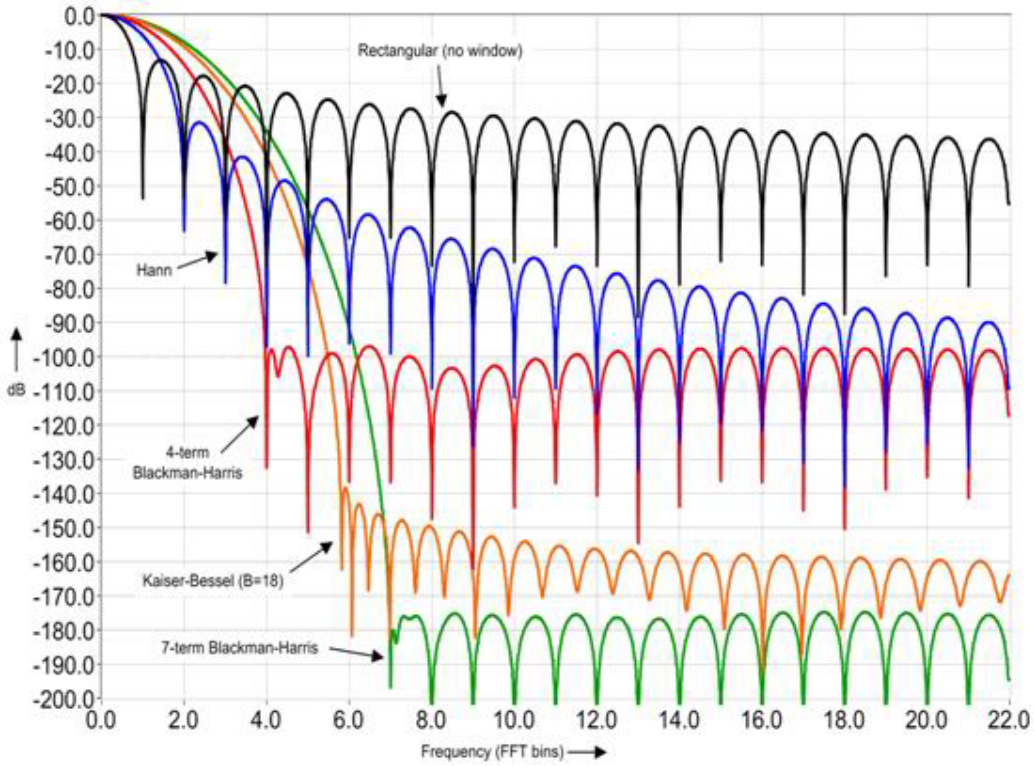


Fig. 35: Typical window functions to suppress strong sidelobes

An infinite spectrum of constant density (shown in Fig. 34a) leads to a Dirac-pulse function in the time domain. The Dirac pulse contains by definition all frequency components of equal power. In Fig. 34b, the spectrum is limited, for example, by the maximum operation frequency of the VNA, or by some user settings. This can be expressed by multiplication of the ideal spectrum with a rectangular function. The iDFT of a rectangular function of width Δf leads to a $sinc$ function (sometimes denoted as si function) in the time domain. This relation is shown in Eq. (43) and graphically in Fig. 34.

$$\begin{aligned} \text{Frequency domain} \quad &\Longleftrightarrow \quad \text{Time domain} \\ \operatorname{rect}\left(\frac{f}{\Delta f}\right) \quad &\Longleftrightarrow \quad \frac{\sin\left(\Delta f \pi t\right)}{\pi t} = \Delta f \cdot \operatorname{sinc}\left(\Delta f \pi t\right). \end{aligned} \tag{43}$$

To mitigate the effect of rectangular clipping of the spectrum in the time domain result, various weighting functions are available. They smoothly filter (reduce) the amplitude of the spectrum around $f_{\min}$ and $f_{\max}$ in band-pass and low-pass mode. This helps to reduce the strong sidelobes (ringing) in the time domain. However, the price to be paid is a reduced pass-band, thus limiting the time resolution and the ability to distinguish between two closely spaced impulses. The user has to select a reasonable trade off between the window weighting functions, depending on the requirements of the particular measurement. The effect of some window functions on main and sidelobes is shown in the frequency domain(!) on a logarithmic scale in Fig. 35.

7.2.4 Gating

The gating option of the VNA allows to eliminate or select parts of the time-domain signal, provided they are reasonably well separated in the time-domain trace.

For example, the already mentioned cable, connecting to the VNA port, is assumed to have an internal irregularity at a certain position. By suitable selection of a time-domain gate (highlighted in

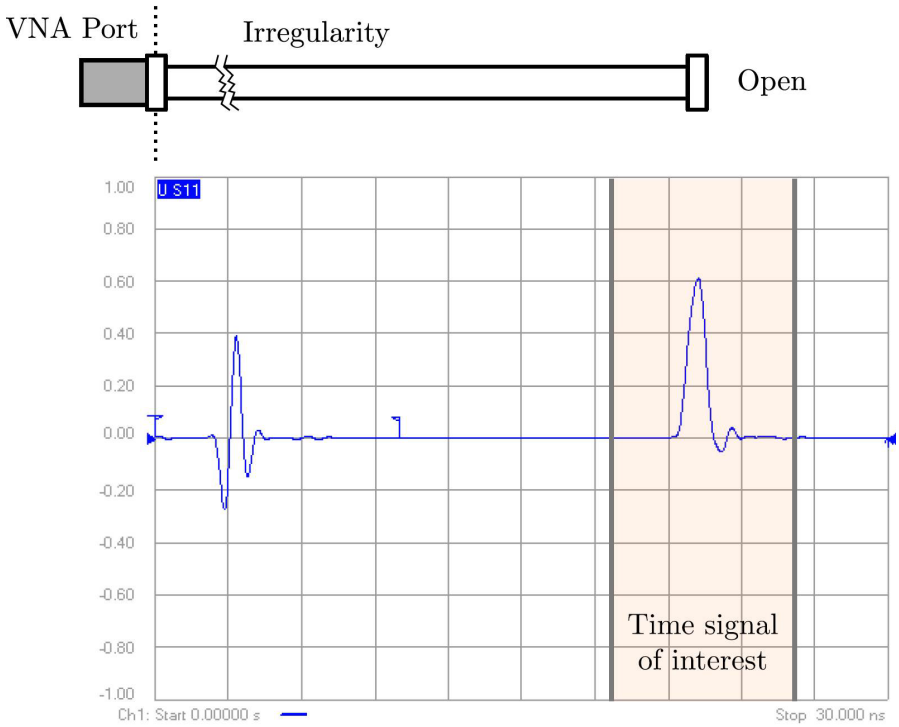


Fig. 36: Only the signal in a certain time window is of interest. After selection, the FFT of this window will be calculated. Here the real values of the synthetic impulse response are shown on a linear scale.

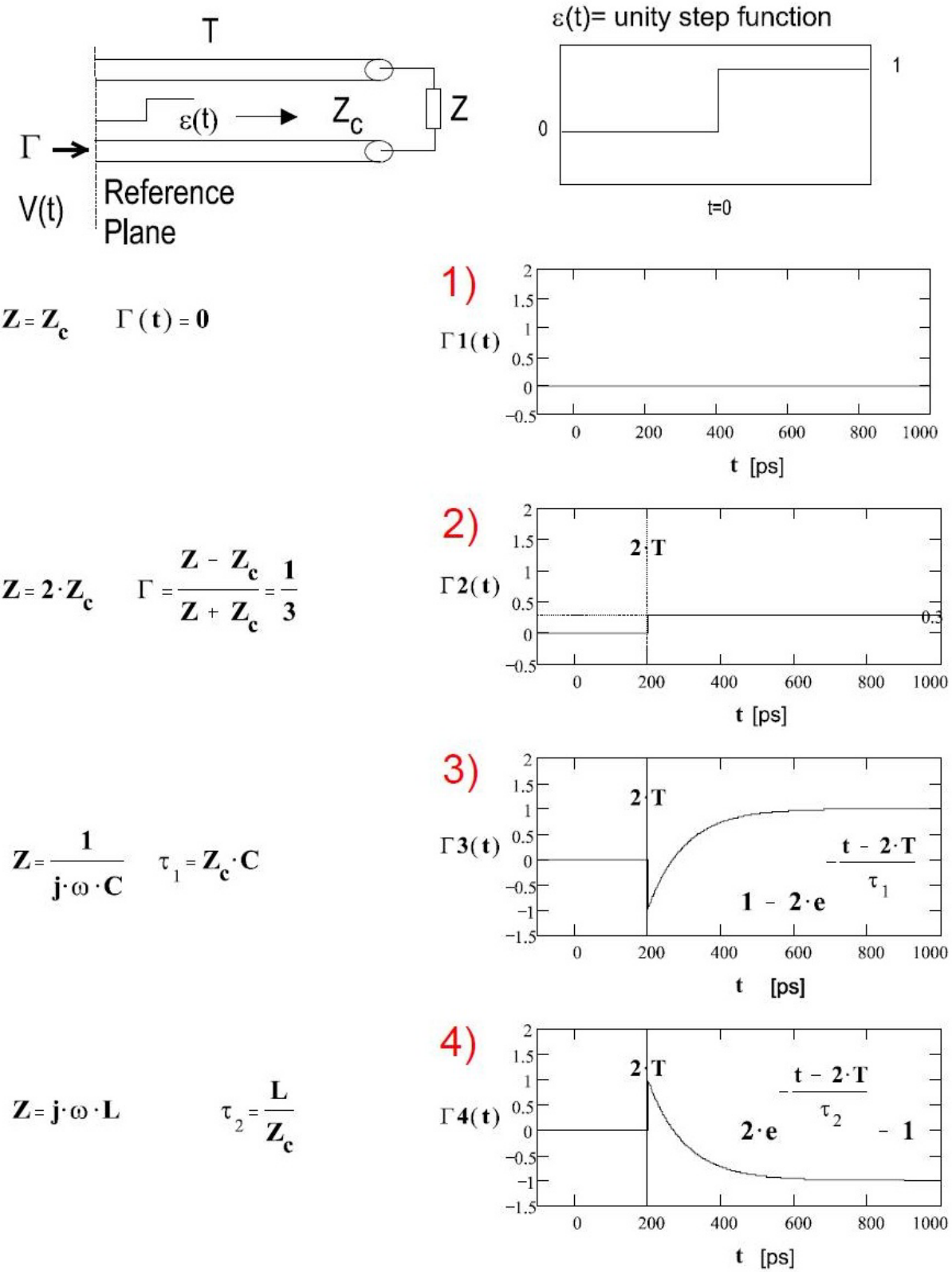
Fig. 36 from $t \approx 18$ ns to $t \approx 26$ ns), the desired portion of the time domain trace (here, the total reflection at the open cable end) can be separated from the the rest of the trace (set to zero). This allows an analysis, e.g. by transformation back to the frequency domain, of the interesting part of the circuit without influence of multiple reflections and perturbations from discontinuities, etc. (de-embedding). For transmission measurements, usually the *first* arriving pulse in the time domain is selected, thus suppressing the effect of all following reflections and related signals. For reflection measurements, the first, but also following pulse response in the time-domain trace may be selected.

The implemented time-domain gating function is not a "brick wall", but a soft switch applying a weighting function similar to the iDFT window function. As it is a *non-linear* operation, it may generate additional frequency components which were not present in the original signal. As general practical guide line, the gate should not cut into a signal trace different from zero.

7.2.5 Examples of synthetic pulse time-domain measurements

A collection of measurement examples of simple DUTs are shown in Fig. 37. For all cases depicted, the VNA is set up in step response operation. The traces from top to bottom show:

1. Matched load ($Z = Z_C$). As Γ is equal to zero, the response is zero everywhere.
2. Moderate (resistive) mismatch ($Z = 2Z_C$, e.g. 100 Ω in a 50 Ω system). During the first 200 ps the trace displays the well impedance-matched cable, following the reflection coefficient jumps to a positive, constand value due to the impedance mismatch.
3. Capacitor. The TDR displays the capacitive load for a moment as a short circuit, and resumes with an exponential function, as the capacitor is charged. The final state is equivalent to an open circuit, as expected.
4. Inductor. In the TDR the inductive load appears at $t = 200$ ps as an open circuit, followed by an exponential decay function. The steady state results in a short circuit, as the inductor is fully



$$Z = Z_c \qquad \Gamma(t) = 0$$

$$Z = 2 \cdot Z_c \qquad \Gamma = \frac{Z - Z_c}{Z + Z_c} = \frac{1}{3}$$

$$Z = \frac{1}{j \cdot \omega \cdot C} \qquad \tau_1 = Z_c \cdot C$$

$$Z = j \cdot \omega \cdot L \qquad \tau_2 = \frac{L}{Z_c}$$

Fig. 37: Examples of an arbitrary impedance, measured in TDR

conducting.

7.2.6 Comparison to true time-domain measurements

There is a wide range of applications for the discussed synthetic pulse time-domain technique. A VNA in time-domain low-pass step mode has a very similar range of applications as a TDR sampling oscilloscope. However, the synthetic pulse method is limited to strictly linear systems, therefore the analysis of transient or non-linear systems, e.g. settling response of a microwave oscillator after power up would not give very meaningful results. In other words, for highly non-linear and time-varying DUTs true time-domain measurements, based on pulse generators and oscilloscopes are still indispensable, e.g. an air traffic radar system, where we have linear but time-varying conditions.

The dynamic range of a typical sampling oscilloscope is limited to about 60 to 80 dB with a maximum input signal of 1 V and a noise floor around 0.1 to 1 mV (typical broadband oscilloscope). The dynamic range of the VNA is > 100 dB, allowing similar maximum input levels of approximately +10 dBm (some VNAs allow +20 dBm). Both instruments are using basically the same kind of detector, either a balanced mixer (four diodes) or a sampling head (two, four or six diodes), but the essential difference lies in the noise floor and the average signal power arriving at the receiver input. In case of the VNA the measurement is based on a continuous-wave (CW) signal with bandwidth of a few Hz, and thus can obtain with appropriate filtering a very good signal-to-noise ratio[7].

A traditional sampling oscilloscope acquires the data during a short time with a rather low repetition rate (typically around 100 kHz up to a few MHz), with all the thermal noise power spread over the entire frequency range (typically 20–50 GHz bandwidth). With this low average signal power (around a microwatt) the signal spectral density is orders of magnitude lower compared to the VNA measurement procedure (it acquires signals continuously), which explains the large difference in dynamic range (even without gain switching).

A more detailed discussion about time-domain reflectometry with vector network analysers can be found in [20].

7.3 Calibration methods

The hardware of even an "ultra-modern" VNA is not perfect, e.g. the internal source is not perfectly impedance matched to $50\ \Omega$ (over the entire frequency range), its internal directional couplers 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 compensate all these unwanted effects, to guarantee a precise, instrument independent analysis of the DUT. There are several calibration procedures to eliminate some, or all of the mentioned deficiencies. The easiest is called the "response calibration", typically applied for transmission, rarely for reflection measurements. It basically is a S_{21} (or S_{12}) transmission measurement of a quasi "zero length" ideal transmission-line, by connecting the two cable ends of the two-port VNA with each other. For the given VNA setting, i.e. start / stop frequency, # of freq. points, resolution bandwidth, power level, etc., magnitude and phase are acquired and stored as $S_{21\text{reference}}$ in the non-volatile memory 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 (VNA plus cables). In calibrated mode the VNA performs:

$$S_{21\text{DUTcal.}} = \frac{S_{21\text{DUTmeas.}}}{S_{21\text{reference}}}, \tag{44}$$

However, this simple calibration procedure eliminates essentially the frequency-dependent losses and phase-transfer functions of the test cables only. But, the mismatch between cable and generator, and

[7] Remember the thermal noise is proportional to measurement bandwidth. Its density at room temperature is –174 dBm/Hz.

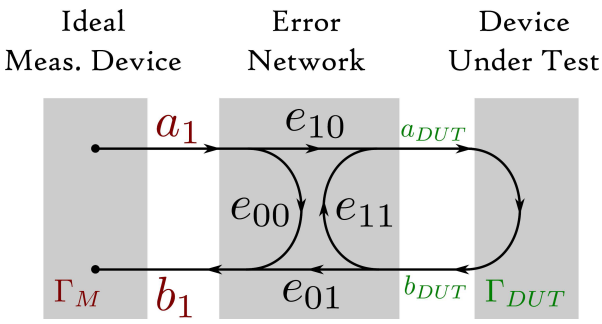


Fig. 38: Error model of a VNA. The parameters e_{xx} of the error network are determined by the calibration procedure and used to determine the true (corrected) result (Γ_{DUT}) based on the measured result (Γ_{M}).

Table 4: Interpretation of VNA error terms

Error term	Interpretation
e_{10}	Reflection tracking
e_{00}	Directivity
e_{11}	Test-port match

the impact of the finite directivity are still present. A more sophisticated, and widely popular calibration technique for the reflection measurements needs to be performed: the open, short and match technique. This technique covers the three independent error sources mentioned above: finite directivity, generator mismatch and the transfer function of the cables.

The VNA applies an internal error model, shown in Fig. 38. The measured raw data acquired by the instrument (Γ_{M}) is distorted by certain systematic errors. These errors are modeled via four parameters: e_{10}, e_{00}, e_{01}, e_{11}, based on the error network model of Fig. 38. e_{nn} are in general complex and frequency dependent parameters, furthermore $e_{10} = e_{01}$. The error parameters are extracted and stored when performing a suitable calibration method, i.e. open, short, match, such that the true value of the DUT (Γ_{DUT}) is calculated and presented accordingly. In simple terms, we need to carry out three independent measurements for each frequency point, to solve three coupled equations with three complex unknowns. These error terms represent the above-mentioned effects as listed in Table 4.

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, only the electromagnetic properties need to be known with 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 an open circuit, a short circuit and a matched load (termination), enabling the VNA to determine the frequency-dependent error model. This is altered 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 during the DUT measurement, and the *reference plane* is "moved" to the end of the test cables. Only the DUT networks "behind" the reference plane are taken into account for the measurement.

The impact of the VNA calibration is demonstrated in Fig. 39, which presents a S_{11} measurement of a high-quality 50 Ω termination, with and without VNA calibration. 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 20 dB! In case of a short (total reflection, $S_{11} = 1 \equiv 0$ dB), a

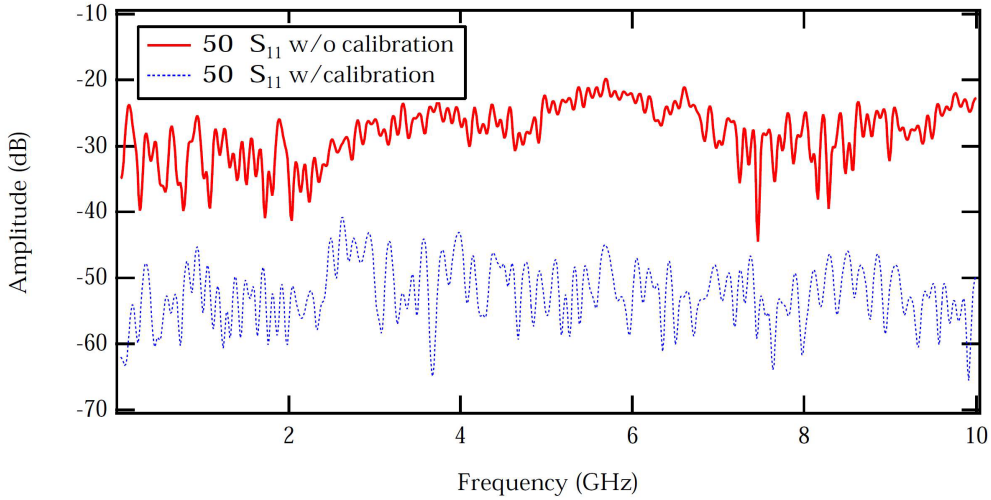


Fig. 39: S_{11} measurement of a 50 Ω termination with and without calibration. The calibration provides 20 dB improvement for this frequency range.

(a) (b)

Fig. 40: Typical calibration kits for a VNA: (a) manual (open, short, match); (b) electronic

non-calibrated S_{11} response typically displays a residual with values of a fraction of a dB, up to a few dB below the 0 dB line (same for an open); after calibration these error reduces to a few millidecibels.

So far, we have covered the "response calibration" and the "complete one-port calibration". To perform completely error-corrected transmission measurements, the "full two-port calibration" procedure has to be applied. Therefore, the error model is expanded to include the errors from the receiving port, requiring a calibration of each port based on the just discussed "complete one-port calibration" method. Also, for transmission, we need two standards, i.e. the "response calibration" and the "isolation calibration", however, latter often may be omitted. In summary, the "full two-port calibration" consists out of a "complete one-port calibration" procedure for each port, which requires open, short and match standards, plus the "response calibration" and eventually the "isolation calibration". In total eight calibration measurements have to be performed to bring the VNA into the desired *CAL* status.

For measurements on devices with standard coaxial connectors, e.g. SMA or N-type, calibration standards such as a termination, an open and a short circuit are available (shown in Fig. 40a). As mentioned, to successfully perform the calibration procedure for the reflection coefficient, the tabulated

values, representing the electromagnetic properties of the calibration standards, has to be present in the VNA. Obviously, the tabulated parameters of the calibration kit does not have an infinite frequency resolution. The instrument applies an interpolation procedure if the selected frequency points are not exactly at 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 if a calibration of a multiport VNA is required.
Already for the full two-port calibration requires eight calibration measurements to satisfy the eight-term error model. The manual procedure of connection and de-connection of the calibration standards is time consuming, boring, and prone to errors. The situation becomes even worse when performing a full four-port calibration (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. 40b), 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. Again, like for the manual calibration method, the standards do not need to be perfect, but well known, reproducible (switching) and stable. More details are found in [17, 18].

7.4 1 dB compression point measurement

A single tone sine-wave source is connected to the input of an amplifier and its amplitude level is gradually increased versus time. Monitoring the output of this amplifier, we notice a proportional dependence between input and output powers for small signal levels. This proportionality is referred as the linear gain factor. For higher input signal levels, this relationship does not hold any more, since the amplifier is not a perfectly linear system, and suffers from "saturation" effects. A fraction of the output power will appear at other frequencies, which are higher order harmonics of the input signal. Typically the second and third harmonics are dominant, and the related signal distortion is referred as harmonic distortion. In parallel, we observe a *compression* of the gain for the fundamental signal. The actual gain falls off below the small-signal gain response (Fig. 41). If this deviation amounts to 1 dB, we have reached the "1 dB compression point". Typically the industry refers to the output power, when specifying the 1 dB compression point for their RF products.

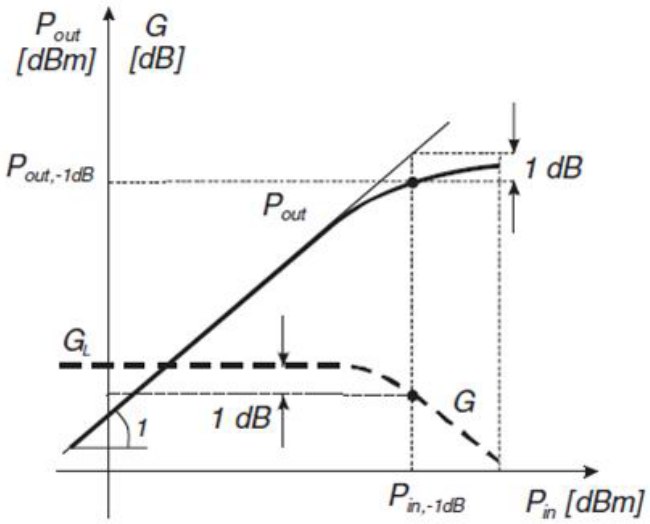


Fig. 41: Definition of the 1 dB compression point for an amplifier: input vs. output power at the point where the power level falls below 1 dB from its (linearly) predicted value.

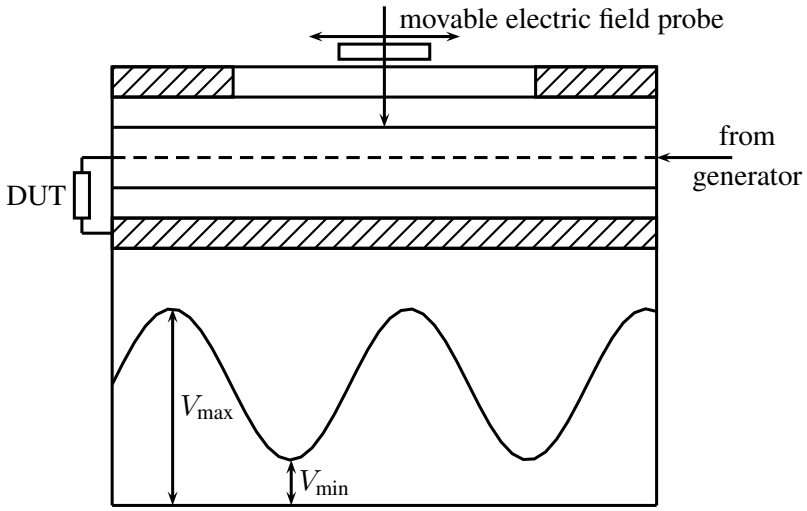


Fig. 42: Schematic view of a measurement set-up used to determine the reflection coefficient as well as the voltage standing wave ratio of a device under test (DUT) [21].

The 1 dB compression point is an important figure of merit, used to characterize the linearity of a RF system, in particular the performance of small-signal and power amplifiers. It can be comfortably measured with most VNAs in CW mode, i.e. choosing a single frequency and performing a power sweep. In power sweep mode, the instrument displays a trace similar as shown in Fig. 41.

8 Introduction to the Smith chart

Even with today's availability of computer-aided simulation and circuit simulation software suites, the *Smith* chart is still a very valuable and important tool that facilitates an interpretation of the (half) complex impedance plane with respect to the S-parameters, and the related calculations and measurements. This section gives a brief overview of the concept, and more importantly, of how to use the chart. Its definition, as well as an introduction of how to navigate on the chart are illustrated. Some typical examples illustrate the broad range of applications of the *Smith* chart.

8.1 Voltage standing wave ratio ($VSWR$)

With modern RF measurement equipment available today it is rather easy to precisely measure the reflection factor Γ, even for complicated networks. In the "good old days" though, this was performed by measuring the electrical field strength[8] along a slotted coaxial line, which has a longitudinal slit to allow a small field probe to be slided to any location along the line (Fig. 42). This electric field probe, protruding into the field region of the coaxial line near the outer conductor, picked up an E-field signal, which was displayed on a microvoltmeter after rectification via a microwave diode. While moving the probe, field maxima and minima, as well as their position and spacing where recorded. From this information the reflection factor Γ and the voltage standing wave ratio ($VSWR$ or SWR) were determined:

– Γ is defined as the ratio of the electrical field strength E of the reflected wave versus the forward-traveling wave:

$$\Gamma = \frac{E \text{ of reflected wave}}{E \text{ of forward-traveling wave}}. \tag{45}$$

[8]The electrical field strength was used, since its measurement was considerably easier than that of the magnetic field.

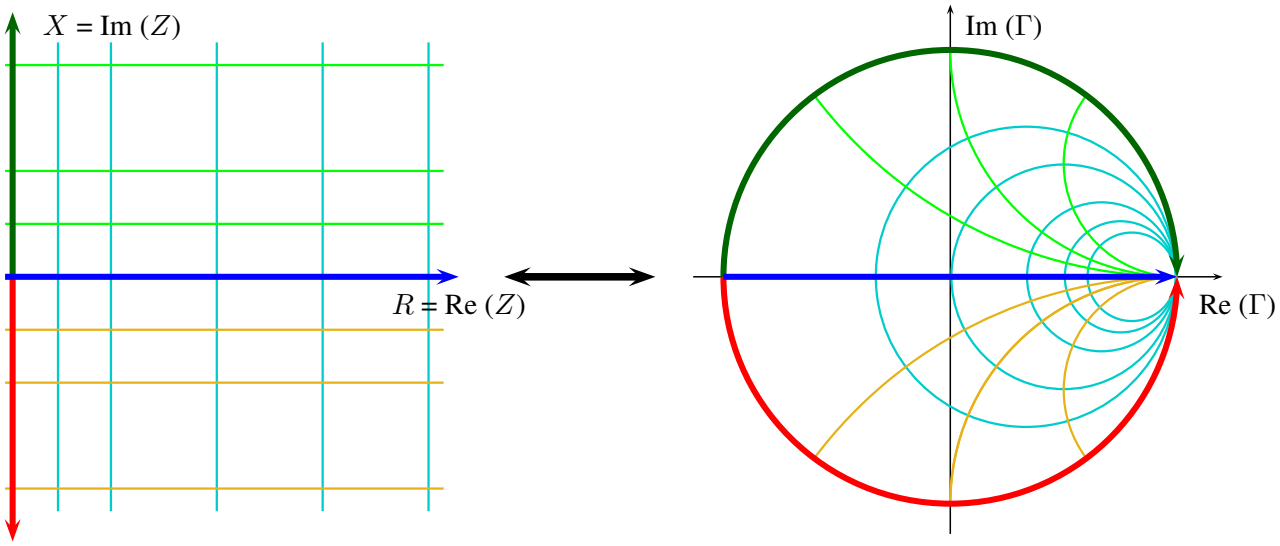


Fig. 43: Illustration of the *Moebius* transformation from the complex impedance plane to the Γ plane, commonly known as *Smith* chart.

– The $VSWR$ is defined as the ratio of maximum to minimum measured voltages:

$$VSWR = \frac{V_{\text{max}}}{V_{\text{min}}} = \frac{1 + |\Gamma|}{1 - |\Gamma|}. \tag{46}$$

Although today these measurements are far easier to conduct, the definitions of the aforementioned quantities are still valid. On top, their importance has not diminished in the field of microwave engineering, both reflection coefficient as well as $VSWR$ are still a vital part of the everyday life of a microwave engineer performing simulations or measurements.

8.2 Definition of the *Smith* chart

The *Smith* chart [22] provides a graphical representation of Γ that permits the determination of quantities like the $VSWR$, or the impedance of a device under test (DUT). It uses the bilinear *Moebius* transformation, projecting the complex impedance plane on the complex Γ plane:

$$\Gamma = \frac{Z - Z_0}{Z + Z_0} \quad \text{with} \quad Z = R + \mathrm{j}\, X. \tag{47}$$

As shown in Fig. 43, the half–plane with positive real part of impedance Z is mapped to the interior of the unit circle of the Γ plane.

8.2.1 Properties of the transformation

In general, this transformation has two main properties:

– generalized circles are transformed to generalized circles (note that a straight line is nothing else than a circle with infinite radius and is therefore mapped as circle to the *Smith* chart);
– angles are preserved locally.

Figure 44 illustrates how certain basic shapes transform between impedance and Γ planes.

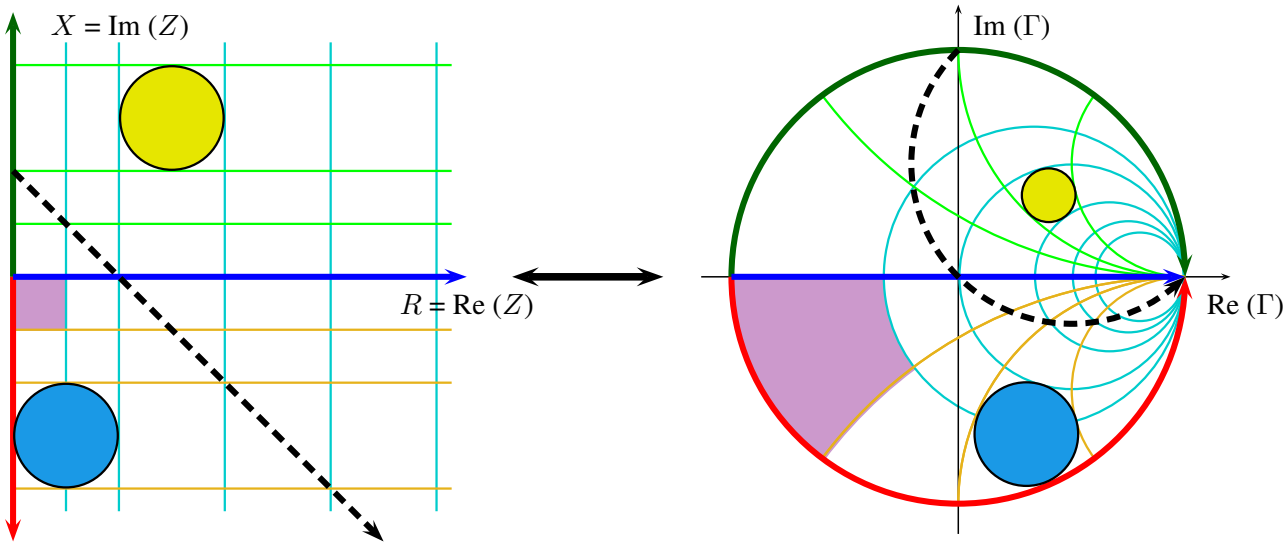


Fig. 44: Illustration of the transformation of basic shapes from the Z to the Γ plane.

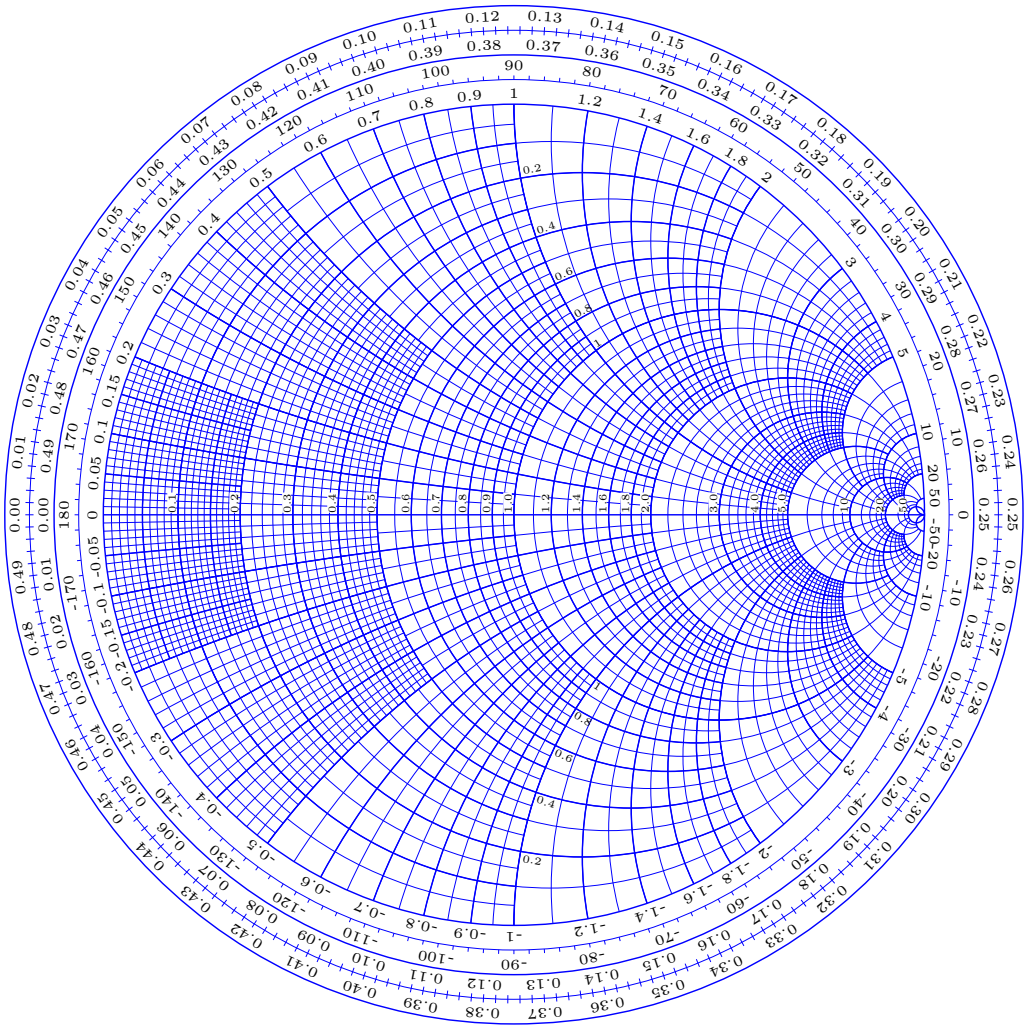


Fig. 45: Example of a typical *Smith* chart

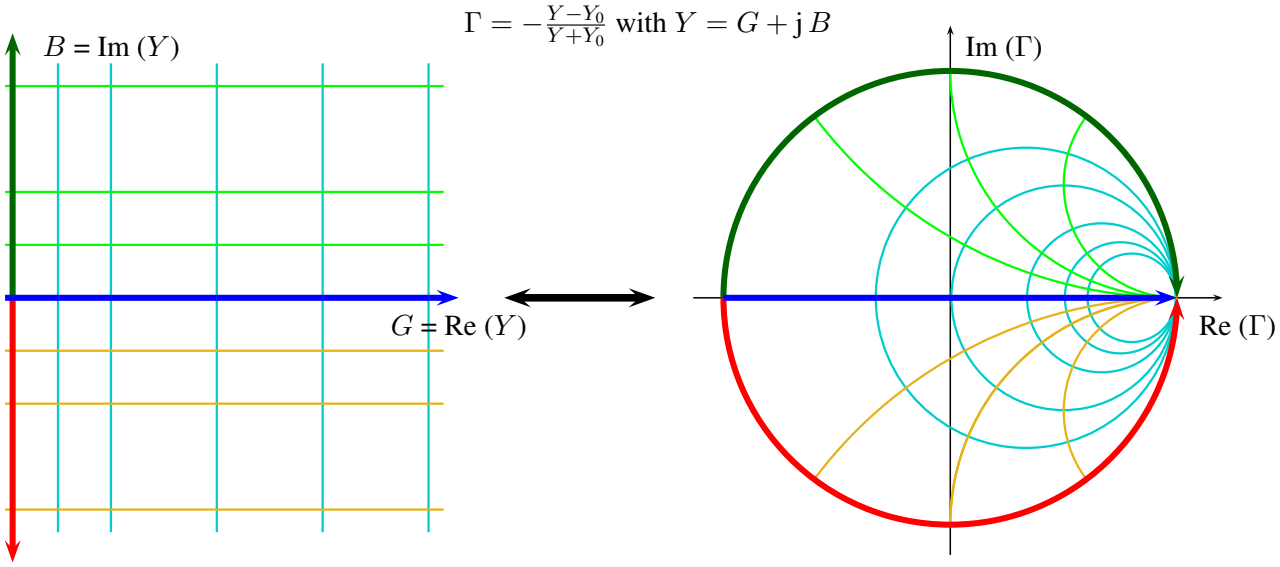


Fig. 46: Mapping of the admittance plane into the Γ plane

8.2.2 Normalization

The Smith chart is usually normalized to a reference impedance Z_0 (= real):

$$z = \frac{Z}{Z_0}. \tag{48}$$

This simplifies the transformation:

$$\Gamma = \frac{z-1}{z+1} \quad \Leftrightarrow \quad z = \frac{1+\Gamma}{1-\Gamma}. \tag{49}$$

Although $Z_0 = 50\ \Omega$ is the most common reference impedance (typical characteristic impedance of coaxial cables) and many applications use this normalization, any other real, positive value is valid. *Therefore, it is crucial to check the normalization assumed, before using any chart.*

Being unfamiliar, the *Smith* charts appears confusing at a first look, with a fine grid from the Z-plane mapped to a dense grid of many circles on the chart (Fig. 45).

8.2.3 Admittance plane

The *Moebius* transformation which generates the *Smith* chart also provides a mapping of the complex admittance plane ($Y = 1/Z$, or normalized $y = 1/z$) into the same chart:

$$\Gamma = -\frac{y-1}{y+1} = -\frac{Y-Y_0}{Y+Y_0} = -\frac{1/Z - 1/Z_0}{1/Z + 1/Z_0} = \frac{Z-Z_0}{Z+Z_0} = \frac{z-1}{z+1}. \tag{50}$$

Using this transformation results in the same chart, but mirrored at the center of the *Smith* chart (Fig. 46). Often both mappings, the admittance and the impedance plane are combined into one chart, which then looks even more overwhelming. For reasons of simplicity all illustrations in this article use only the mapping from the impedance to the Γ plane.

8.3 Navigation in the *Smith* chart

The representation of circuit elements in the *Smith* chart is discussed in this section, starting with some important points inside the chart. The following examples of circuit elements illustrate their representation in the chart.

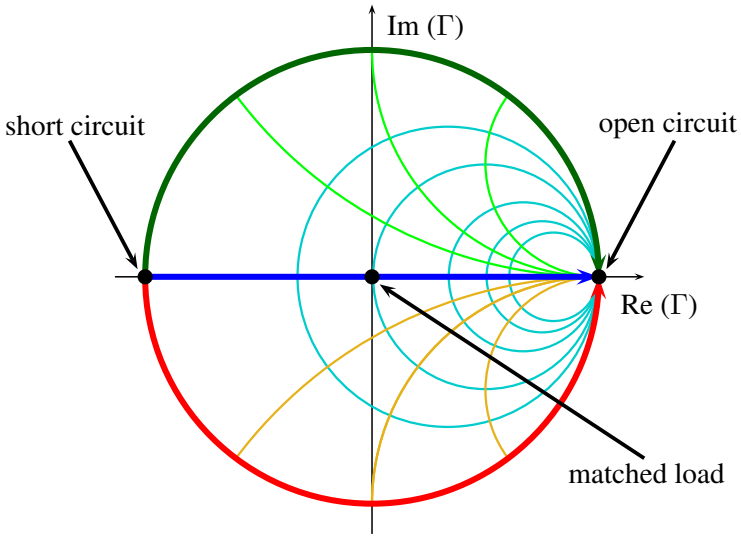


Fig. 47: Important points in the Smith chart

8.3.1 Important points

There are three important points in the chart:

1. Open circuit with $\Gamma = 1, z \to \infty$.
2. Short circuit with $\Gamma = -1, z = 0$.
3. Matched load with $\Gamma = 0, z = 1$.

They all are located along the real axis at the beginning and the end, which are also on the outer circle (imaginary axis), and at the center of the *Smith* chart (Fig. 47). The upper half of the chart is "inductive", since it corresponds to the positive imaginary part of the impedance. The lower half is "capacitive", as it is corresponding to the negative imaginary part of the impedance.

Concentric circles around the center represent constant reflection factors (Fig. 48). Their radius is directly proportional to the magnitude of Γ; therefore, a radius of 0.5 corresponds to reflection of 3 dB (half of the signal is reflected), whereas the outermost circle (radius = 1) represents total reflection. Evidently, matching problems are clearly visualized in the Smith chart, since a mismatch will lead to a reflection coefficient larger than 0, see Eq. (51).

$$\text{Power into the load = forward power - reflected power: } P = \frac{1}{2}\left(|a|^2 - |b|^2\right) = \frac{|a|^2}{2}\left(1 - |\Gamma|^2\right). \quad (51)$$

In Eq. (51) the European notation is used[9]: power $= |a|^2/2$. Furthermore it should be noted, $(1 - |\Gamma|^2)$ corresponds to the losses due to the impedance mismatch.

Even though here we limit to the mapping of the impedance plane to the Γ plane, The admittance is simple to determine, since

$$\Gamma(\frac{1}{z}) = \frac{1/z - 1}{1/z + 1} = \frac{1 - z}{1 + z} = \left(\frac{z - 1}{z + 1}\right) \text{ or } \Gamma(\frac{1}{z}) = -\Gamma(z). \quad (52)$$

In the *Smith* chart this fact is visualized as a 180° rotation of the vector of a given impedance (Fig. 49).

[9]The commonly used notation in the USA: power = $|a|^2$. These conventions have no impact on the S-parameters, but they are relevant for absolute power calculations. Since this is rarely used in context with *Smith* chart gymnastics, the actual power definition used is not critical.

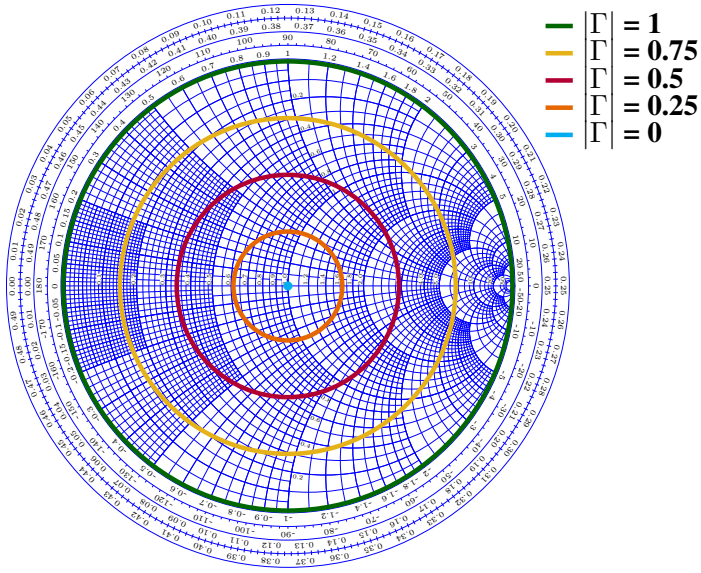


Fig. 48: Illustration of circles representing a constant reflection factor

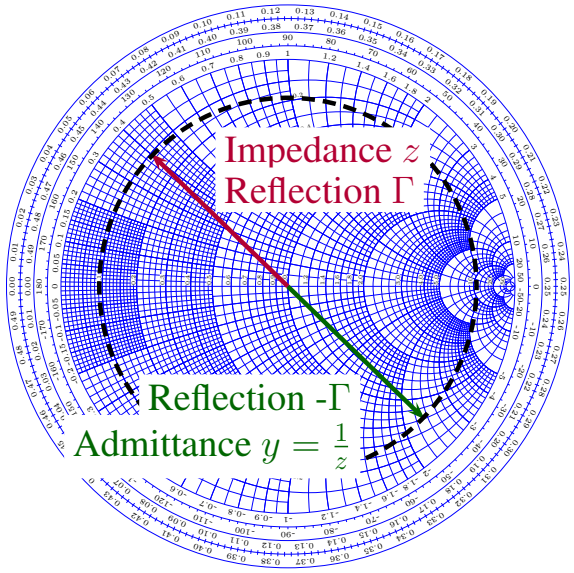


Fig. 49: Conversion of an impedance to the corresponding admittance in the *Smith* chart

8.3.2 Impedance of simple, passive lumped element circuits

Consider a simple passive circuit: a lumped, reactive element (inductance L, or capacitance C) of arbitrary value connected in series to an resistance R. The corresponding signature of this circuit in the *Smith* chart, varying the inductance resp. capacitance, is a circle. For a given type of impedance, the trace of this circle follows a clockwise (inductance), or anticlockwise (capacitance) movement (Fig. 50). If a lumped, reactive element is connected in parallel to R, the pattern is basically the same, but rotated by 180° (Fig. 51). It is equivalent to the discussed admittance mapping. Summarizing both cases, results in a simple rule for the navigation in the *Smith* chart:

Reactive elements connected in series follow the trajectory of a circle in the impedance plane. Inductances move clockwise, capacitances move anticlockwise when increasing their value. Reactive elements connected in parallel follow a circular trajectory in the admittance plane, clockwise for capacitances,

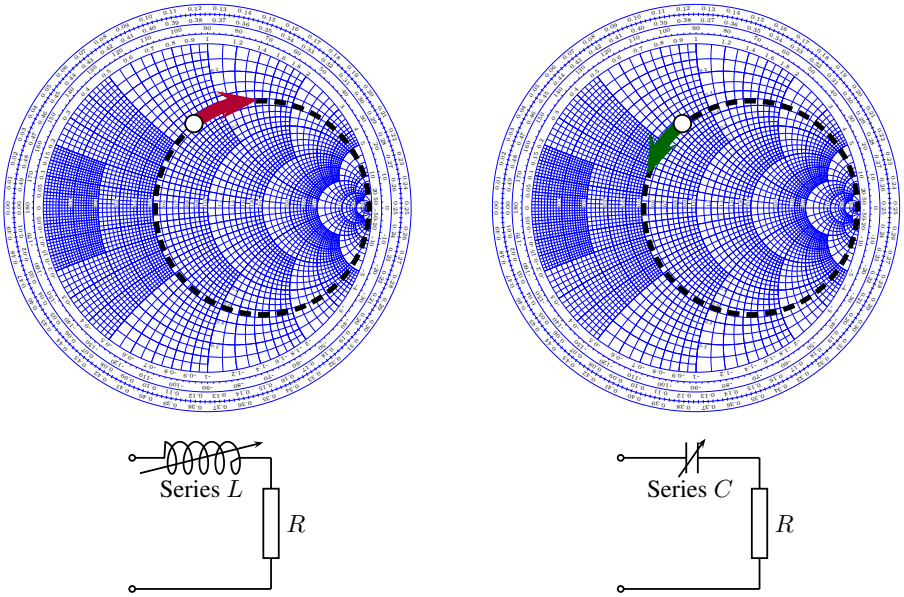


Fig. 50: Circular traces of reactances with varying value connected in series to a fixed impedance

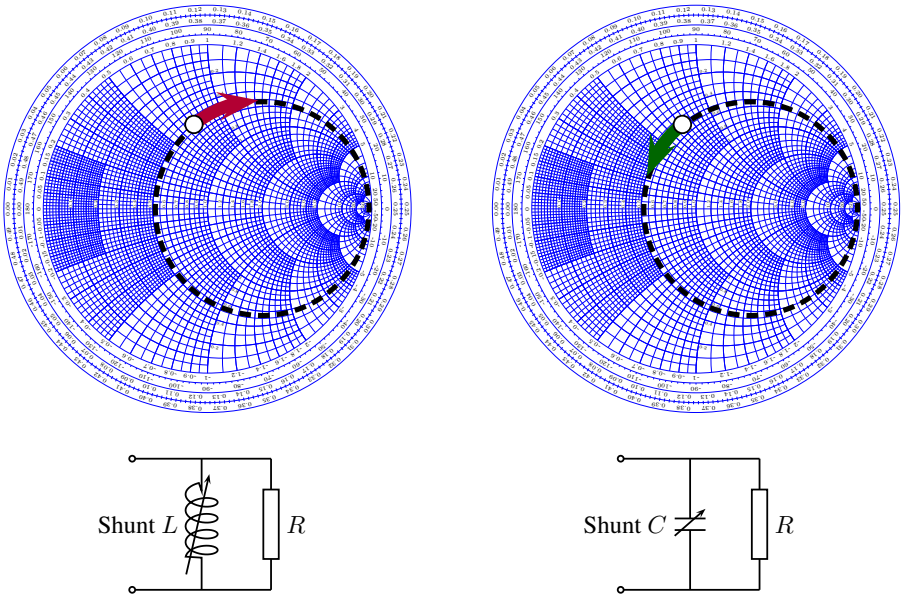


Fig. 51: Circular traces of reactances with varying value connected in parallel to a fixed impedance

anticlockwise for inductances.

This rule is illustrated in Fig. 52.

8.3.3 Impedance transformation using a transmission-line

The S-matrix of an ideal, lossless transmission-line of physical length l is given by

$$S = \begin{bmatrix} 0 & e^{-j\beta l} \\ e^{-j\beta l} & 0 \end{bmatrix}, \tag{53}$$

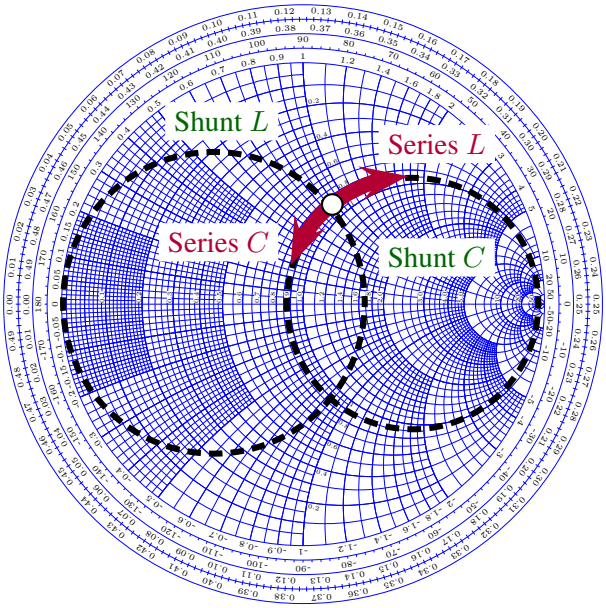


Fig. 52: Navigation in the *Smith* chart when connecting reactive elements.

where $\beta = 2\pi/\lambda$ is the propagation coefficient at the wavelength λ ($\lambda = \lambda_0$ for $\epsilon_\mathrm{r} = 1$).

The lossless transmission-line changes only the phase between its ports. Adding a short piece of, e.g. of coaxial cable in front of a load impedance, will turn the corresponding circle of Z_load clockwise, which is effectively a transformation of the reflection factor Γ_load (without line) to the new reflection factor $\Gamma_\mathrm{in} = \Gamma_\mathrm{load} e^{-j2\beta l}$. Graphically speaking, the vector corresponding to Γ_in is rotated clockwise by an angle of $2\beta l$ (Fig. 53).

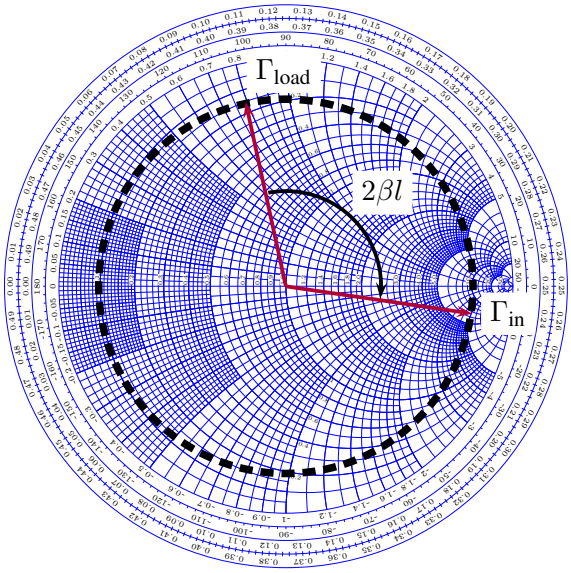


Fig. 53: Adding a lossless transmission-line of physical length l to an impedance Z_load

The input impedance of a lossless transmission-line of characteristic impedance Z_0, terminated with Z_load is given by:

$$Z_\mathrm{in} = Z_0 \frac{Z_\mathrm{load} + jZ_0 \tan(\beta l)}{Z_0 + jZ_\mathrm{load} \tan(\beta l)} \tag{54}$$

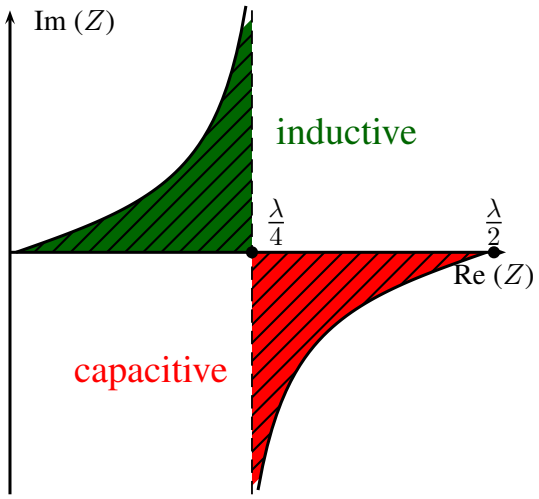


Fig. 54: Impedance of a transmission line as a function of its length l

and the corresponding reflection coefficient follows as mentioned:

$$\Gamma_{\text{in}} = \Gamma_{\text{load}} e^{-j2\beta l} \tag{55}$$

Depending on the values of β, Z_0, Z_{load}, and l, the input impedance will be quite different from the load impedance Z_{load}. Special cases are:

- $l = \lambda/2$: $Z_{\text{in}} = Z_{\text{load}}$
- $l = \lambda/4$: $Z_{\text{in}} = Z_0^2/Z_{\text{load}}$ (impedance transformer)
- $Z_{\text{load}} = Z_0$: $Z_{\text{in}} = Z_0$ (matched termination)
- $Z_{\text{load}} = jX_{\text{load}}$: $Z_{\text{in}} = jX_{\text{in}}$ (reactive load $\Rightarrow$ reactive input impedance)
- $l \ll \lambda$: $Z_{\text{in}} = Z_{\text{load}}$ (basically no line present)

Terminating a transmission-line with a short circuit, $Z_{\text{load}} = 0$, simplifies Eq. 54 to

$$Z_{\text{in}} = jZ_0 \tan(\beta l) \tag{56}$$

which results in an "inductive" or "capacitive" impedance behavior at the input, depending on the length of the line (see Fig. 54).

Adding a transmission-line of length $\lambda/4$ interestingly results in a change of Γ by a factor -1:

$$\Gamma_{\text{in}} = \Gamma_{\text{load}} e^{-j2\beta l} = \Gamma_{\text{load}} e^{-j2(\frac{2\pi}{\lambda})l} \overset{l=\frac{\lambda}{4}}{=} \Gamma_{\text{load}} e^{-j\pi} = -\Gamma_{\text{load}}. \tag{57}$$

Again, this is equivalent to inverting an impedance z to its admittance $1/z$, or the clockwise rotation of the impedance vector by 180°. Especially when starting with a short circuit ($Z_{\text{load}} = 0 \Rightarrow -1$ in the *Smith* chart), adding a transmission line of length $\lambda/4$ transforms it into an open circuit ($+1$ in the *Smith* chart), and vice versa.

8.3.4 Two-port examples

The general form of Eq. 55 returns the input reflection coefficient Γ_{in} for a 2-port network terminated with Z_{load}, i.e. a reflection coefficient Γ_{out} at the output port:

$$\Gamma_{\text{in}} = S_{11} + \frac{S_{12}S_{21}\Gamma_{\text{load}}}{1 - S_{22}\Gamma_{\text{load}}}. \tag{58}$$

Lets evaluate some examples, defined by their S-matrix, which map their impedance to particular characteristic lines and circles on the *Smith* chart. For this illustration, a very simplified *Smith* chart, consisting just of the outermost circle (imaginary axis) and the real axis is used.

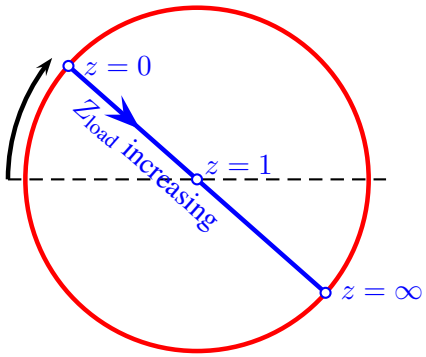


Fig. 55: Rotation of the real axis, therefore the reference plane of the *Smith* chart when adding a transmission-line

8.3.4.1 Transmission-line of length $\lambda/16$

The S-matrix of a $\lambda/16$ transmission-line is

$$\mathrm{S} = \begin{bmatrix} 0 & \mathrm{e}^{-\mathrm{j}\frac{\pi}{8}} \\ \mathrm{e}^{-\mathrm{j}\frac{\pi}{8}} & 0 \end{bmatrix} \tag{59}$$

has a input reflection coefficient of

$$\Gamma_{\text{in}} = \Gamma_{\text{load}} \mathrm{e}^{-\mathrm{j}\frac{\pi}{4}} \tag{60}$$

This corresponds to a rotation of the real axis of the *Smith chart* by an angle of 45° (Fig. 55) and hence a change of the reference plane of the chart (Fig. 55). Consider, for example, a transmission-line terminated by a short and hence $\Gamma_{\text{load}} = -1$. The resulting reflection coefficient is then equal to $\Gamma_{\text{in}} = \mathrm{e}^{-\mathrm{j}\frac{\pi}{4}}$.

8.3.4.2 3 dB attenuator

The S-matrix of a 3 dB attenuator is given by

$$\mathrm{S} = \begin{bmatrix} 0 & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 0 \end{bmatrix}. \tag{61}$$

The resulting reflection coefficient is

$$\Gamma_{\text{in}} = \frac{\Gamma_{\text{load}}}{2} \tag{62}$$

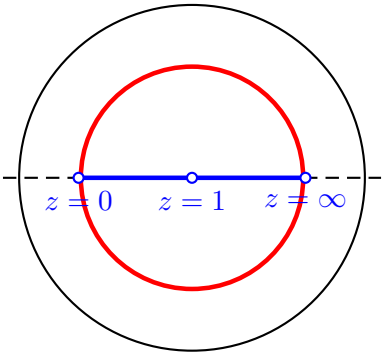


Fig. 56: Effect of an attenuator in the *Smith* chart

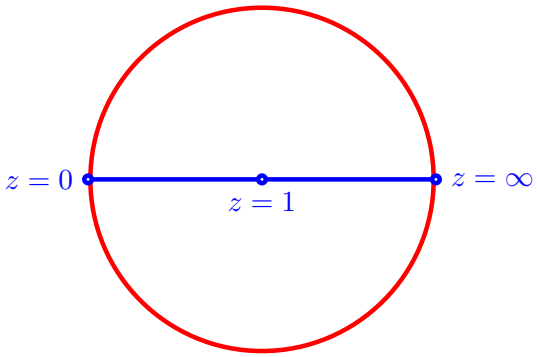


Fig. 57: A load resistor of variable value in the simplified *Smith* chart. Since the impedance has a real part only, the trace remains on the real axis of the Γ plane.

In the Smith chart, the connection of such an attenuator causes the outermost circle to shrink to a radius of 0.5, see Fig. 56[10].

8.3.5 Resistive load

Fig. 57 illustrates how the real axis is passed, if a resistive load changes its value $0 < z < \infty$.

8.4 Examples for applications of the Smith chart

In this section two examples of typical RF problems demonstrate how the *Smith* chart greatly facilitates their solutions.

8.4.1 A step in the characteristic impedance

Consider a junction between two infinitely short cables, an incoming with a characteristic impedance of $Z_1 = 50\Omega$, the outgoing with $Z_2 = 75\Omega$ (Fig. 58). Both ports are matched in their characteristic impedance.

The incident waves are denoted with a_i $(i = 1, 2)$, the reflecting waves with b_i. The reflection coefficient at port 1 follows as

$$\Gamma_1 = \frac{Z_2 - Z_1}{Z_2 + Z_1} = \frac{75 - 50}{75 + 50} = +0.2. \tag{63}$$

[10] An attenuation of 3 dB corresponds to a reduction by a factor 2 in power.

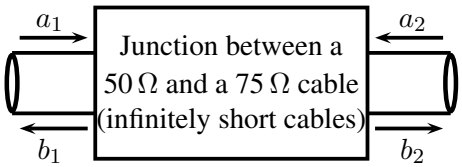


Fig. 58: Junction between two coaxial cables, one with with $Z_1 = 50\Omega$, the other with $Z_2 = 75\Omega$ characteristic impedance. Infinitely short cables are assumed – only the junction is considered.

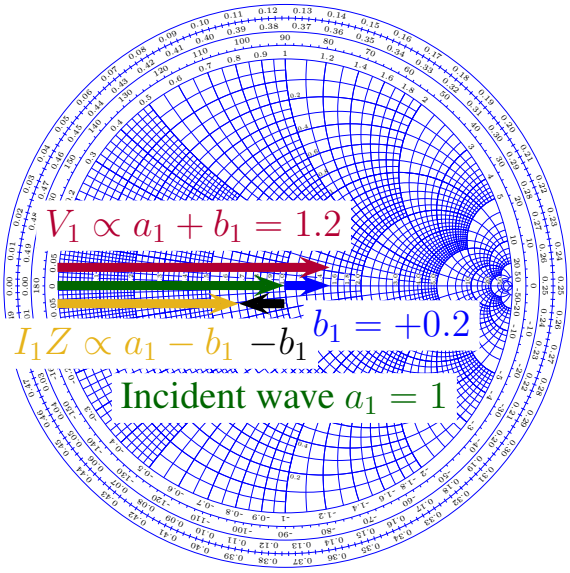


Fig. 59: Visualization of the two-port formed by the two cables of different characteristic impedances

Thus, the voltage of the reflected wave at port 1 is 20% of the incident wave ($b_1 = a_1 \cdot 0.2$), and the reflected power at port 1 is $\Gamma_1^2 = 0.04 \equiv 4\%$. From conservation of energy, the transmitted power has to be 96%, i.e. $b_2^2 = 1 - \Gamma_1^2 = 0.96$.

The voltage transmission coefficient in this particular case computes $t = 1 + \Gamma$, and the output voltage of the transmitted wave at port 2 is *higher* than the voltage of the incident wave at port 1: $V_{\text{transmitted}} = V_{\text{incident}} + V_{\text{reflected}} = 1 + 0.2 = 1.2$. Also, note that this structure is not symmetric ($S_{11} = +0.2 \neq S_{22} = -0.2$), but reciprocal ($S_{21} = S_{12} = \sqrt{1 - \Gamma_1^2}$). As all impedances are real, the corresponding vectors show up in the *Smith* chart on the real axis (Fig. 59).

8.4.2 Quality (Q) factor of a cavity

The second example shows the calculation of the quality factor of a cavity resonator with help of the *Smith* chart.

A cavity at or near to one of its eigenmode resonances can be approximated by a parallel RLC equivalent circuit (Fig. 60). The resonance condition is given as

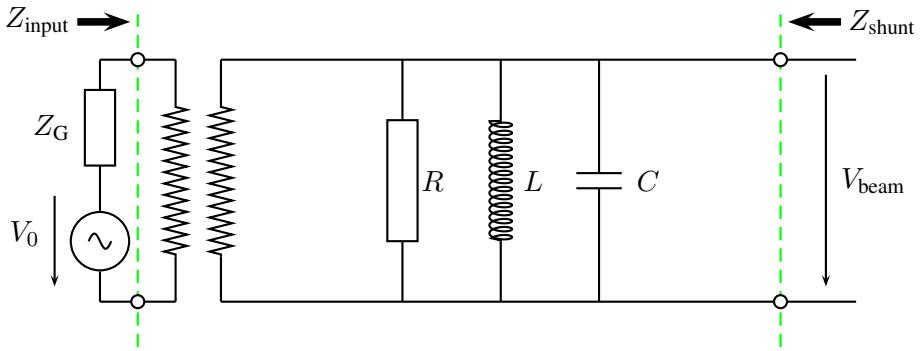


Fig. 60: Equivalent circuit of a cavity near resonance. The transformer describes the coupling of the cavity (typically $Z_{\text{shunt}} \approx 1$ MΩ, as seen by the beam) to the generator (often $Z_G = 50\ \Omega$).

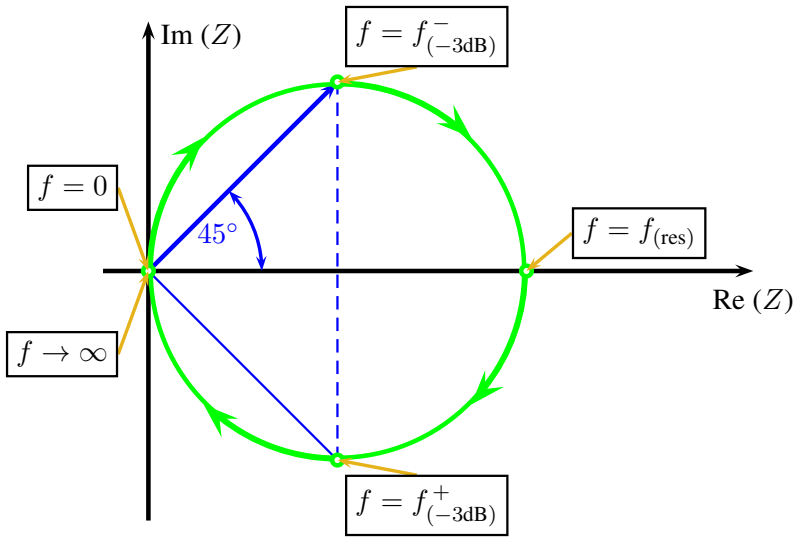


Fig. 61: Schematic drawing of the 3 dB bandwidth in the impedance plane

$$\omega L = \frac{1}{\omega C} \tag{64}$$

from which the resonance frequency follows

$$\omega_{\mathrm{res}} = \frac{1}{\sqrt{LC}} \qquad \text{or} \qquad f_{\mathrm{res}} = \frac{1}{2\pi}\frac{1}{\sqrt{LC}}. \tag{65}$$

The impedance Z of the cavity equivalent circuit is simply

$$Z(\omega) = \frac{1}{\frac{1}{R} + \mathrm{j}\omega C + \frac{1}{\mathrm{j}\omega L}}. \tag{66}$$

The 3 dB bandwidth Δf refers to the points where Re(Z) = Im(Z), which correspond to two vectors with an argument of 45° (Fig. 61) and an impedance of $|Z_{(-3\ \mathrm{dB})}| = 0.707R = R/\sqrt{2}$.

In general, the quality factor Q of a resonant circuit is defined as the ratio of the stored energy W over the energy dissipated P in one oscillation cycle:

$$Q = \frac{\omega W}{P}. \tag{67}$$

However, the Q factor for a resonance can also be calculated using the 3 dB bandwidth and the resonance frequency:

$$Q = \frac{f_{\mathrm{res}}}{\Delta f}. \tag{68}$$

For a cavity, three different quality factors are defined:

- Q_0 (unloaded Q): Q factor of the unperturbed system, i.e. the stand-alone cavity;
- Q_{L} (loaded Q): Q factor of the cavity when connected to a generator and/or measurement circuits;
- Q_{ext} (external Q): Q factor that describes the degeneration of Q_0 due to the generator and/or diagnostic impedances.

All these Q factors are linked via a simple relation:

$$\frac{1}{Q_{\mathrm{L}}} = \frac{1}{Q_0} + \frac{1}{Q_{\mathrm{ext}}}. \tag{69}$$

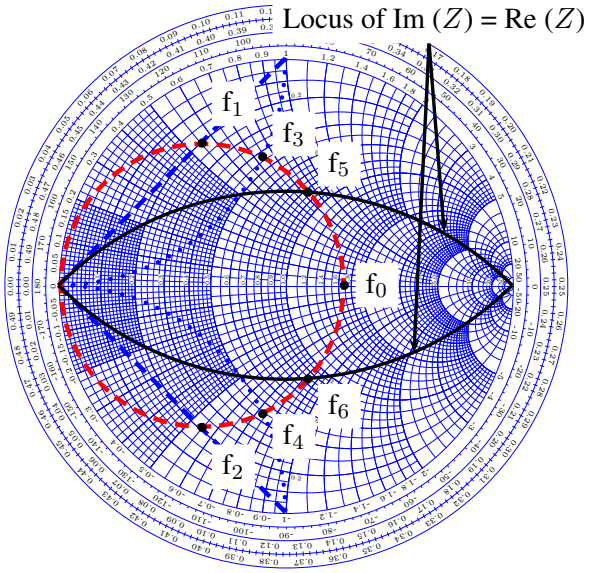


Fig. 62: Evaluation of the different Q factors of a resonant cavity with help of the *Smith* chart

The coupling coefficient β is then defined as

$$\beta = \frac{Q_0}{Q_{\text{ext}}}. \tag{70}$$

This coupling coefficient has not to be confused with the propagation coefficient of transmission lines, which is also denoted as β.

In the *Smith* chart, a resonant circuit shows up as a circle (Fig. 62, dashed, red circle shown in the "detuned short" position). The larger the circle, the stronger is the coupling. Three types of coupling are distinguished, depending on the range of $beta$ (= size of the circle, assuming the circle is in the "detuned short" position):

- Undercritical coupling ($0 < \beta < 1$): the radius of the resonance circle is smaller than 0.25. Hence, the center of the chart ($\Gamma = 0$) lies outside the circle.
- Critical coupling ($\beta = 1$): the radius of the resonance circle is exactly 0.25. Hence, the circle crosses $\Gamma = 0$ at the resonance frequency f_{res}.
- Overcritical coupling ($1 < \beta < \infty$): the radius of the resonance circle is larger than 0.25. Hence, the center of the chart lies inside the circle.

In practice, the circle may be rotated around the origin due to the transmission lines between the resonant circuit and the measurement device.

From the different marked frequency points in Fig. 62 the 3 dB bandwidth, and thus the quality factors Q_0, Q_{L} and Q_{ext} are determined as follows:

- The unloaded Q is determined from f_5 and f_6. The condition for these points is Re(Z) = Im(Z), with the resonance circle in the "detuned short" position.
- The loaded Q is determined from f_1 and f_2. The condition to find these points is $|\text{Im}(S_{11})| \rightarrow$ max. in "detuned short" position.
- The external Q is calculated from f_3 and f_4. The condition to determine these points is $Z = \pm$j in "open short" position, which is equivalent to $Y = \pm$j in "detuned short" position

To determine the points f_1 to f_6 with a network analyzer, the following steps are applicable:

- f_1 and f_2: set the marker format to Re(S_{11}) + j Im(S_{11}) and determine the two points where Im(S_{11}) = max.
- f_3 and f_4: set the marker format to Z and find the two points where $Z = \pm$ j.
- f_5 and f_6: set the marker format to Z and locate the two points where Re(Z) = Im(Z).

9 Summary

Some fundamental concepts on RF devices, instruments, and singal processing techniques have been presented in this introduction to RF measurement concepts. Advantages of various measurement methods using spectrum and network analyzers were presented. In the last section the definition of the *Smith* chart, and its usage were illustrated with several examples. This article supports the practical part of the CAS intermediate-level RF course, and serves as background information.

10 Hand-on experiments

The following hands-on experiments are foreseen:

10.1 Spectrum analyzer test stand 1:

- Measurements of several types of modulation (AM, FM and PM) in time and frequency domain.
- Superposition of AM and FM spectra (unequal carrier sidebands).
- Concept of a spectrum analyzer: the superheterodyne method. Practice different settings (video bandwidth, resolution bandwidth etc). Advantage of FFT spectrum analyzers.

10.2 Spectrum analyzer test stand 2:

- Measurement of the TOI point of some amplifiers (intermodulation tests).
- Concept of noise-figure and noise-temperature measurements, testing a noise diode, the basics of thermal noise.
- Electromagnetic compatibility (EMC) measurements (e.g. analyze your cell-phone spectrum).
- General concepts of non-linear distortions and application of vector spectrum analyzers, spectrogram mode.
- Measurement of the RF characteristic of a microwave detector diode (output voltage versus input power ... transition between regimes output voltage proportional to input power and output voltage proportional to input voltage).

10.3 Spectrum analyser test stand 3:

- Concept of noise-figure and noise-temperature measurements, testing a noise diode, the basics of thermal noise.
- Noise-figure measurements on amplifiers and also attenuators.
- The concept and meaning of excess noise ratio (ENR) numbers.
- Noise temperature of the fluorescent tubes in the room using a satellite receiver.

10.4 Network analyser test stand 1:

- Calibration of the vector network analyzer.
- Navigation in the Smith chart.
- Application of the triple-stub tuner for matching.
- Measurements of the light velocity using a trombone (constant-impedance adjustable coaxial line) in the frequency domain.
- N-port (N = 1–4) S-parameter measurements for different reciprocal and non-reciprocal RF components.
- Self-made RF components: calculate, build and test your own attenuator (and then take it home).

10.5 Network analyzer test stand 2:

- Measurements of the light velocity using a trombone (constant-impedance adjustable coaxial line) in the time domain.
- Two-port measurements for active RF components (amplifiers).
- A 1 dB compression point (power sweep).
- Beam transfer impedance measurements with the wire (button pick-up, stripline pick-up).

10.6 Network analyzer test stand 3:

- Measurements of the characteristic cavity features (Smith-chart analysis).
- Cavity perturbation measurements (bead pull).
- Perturbation measurements using rectangular waveguides.
- Standing wave ratio (SWR) measurements using a waveguide measurement line and movable probe.

References

[1] G.D. Vendelin, A.M. Pavio and U.L. Rohde, *Microwave Circuit Design Using Linear and Nonlinear Techniques*, second ed. (Wiley-Interscience, New Jersey, 2005), ISBN-10 0-471-41479-4.

[2] F. Caspers, Proc. CERN Accelerator School, RF Engineering for Particle Accelerators, Oxford, UK, 1991, p.181.

[3] M. Thumm, W. Wiesbeck and S. Kern, *Hochfrequenzmesstechnik* (Teubner, Stuttgart/Leipzig, 1998), ISBN 3-519-16360-8.

[4] R.A. Witte, *Spectrum and Network Measurements* (Prentice-Hall, New Jersey , 1991), ISBN 0-13-826959-9.

[5] W.O. Schleifer, *Hochfrequenz und Mikrowellenmesstechnik in der Praxis* (Hüthig, Heidelberg, 1981), ISBN 3-7785-0675-7.

[6] B. Schiek and H.J. Sieveris, *Rauschen im Hochfrequenzschaltungen* (Hüthig, Heidelberg, 1984), ISBN 3-7785-2007-5.

[7] P.C.L. Yip, *High Frequency Circuit Design and Measurement* (Chapman and Hall, London, 1990), ISBN 0-412-34160-3.

[8] G. Evans and C.W. McLeisch, *RF-Radiometer Handbook* (Artech, Dedham, 1977), ISBN 0-89006-055-X.

[9] F.R. Connor, *Noise* (Edward Arnold, London, 1973), ISBN 0-7131-3306-6.

[10] F. Landstorfer and H. Graf, *Rauschprobleme der Nachrichtentechnik* (Oldenbourg, München, 1981), ISBN 3-486-24681-X.

[11] O. Zinke and H. Brunswig, *Lehrbuch der Hochfrequenztechnik, Zweiter Band* (Springer, Berlin, 1974), ISBN 3-540-06245-9.

[12] Agilent Technologies, Inc., *Fundamentals of RF and microwave noise figure measurements*, Agilent Application Note 57-1, 2010.

[13] B. Schiek, *Messysteme der Hochfrequenztechnik* (Hüthig, Heidelberg, 1984), ISBN 3-7785-1045-2.

[14] F. Caspers, RF engineering basic concepts: S-parameters, CAS Proc., 2010, CERN Yellow Report CERN-2011-007, pp. 67-93.

[15] J. Verspecht and D. Root, *Polyharmonic Distortion Modeling*, IEEE Microwave Magazine, Vol. 7, Issue 3, June 2006, pp. 44-57.

[16] M. Thumm, W. Wiesbeck and S. Kern, *Hochfrequenzmesstechnik, Verfahren und Messsysteme* (Teubner, Stuttgart, 1998), ISBN 978-3519163602.

[17] M. Hiebel, *Fundamentals of Vector Network Analysis* (Rohde & Schwarz, München, 2007), ISBN 3939837067.

[18] Agilent Technologies, Inc., *Understanding the fundamental principles of vector network analysis*, Agilent Application Note AN 1287-1, 2000.

[19] Anritsu Company, *Time domain measurements using vector network analyzers*, Anritsu Application Note No. 11410-00206, R , 2009.

[20] Agilent Technologies, Inc., *Time domain analysis using a network analyzer*, Agilent Application Note 1287-12, 2012.

[21] H. Meinke and F.-W. Gundlach, *Taschenbuch der Hochfrequenztechnik* (Springer, Berlin, 1992).

[22] P. Smith, *Electronic Applications of the Smith Chart* (Noble Publishing, Atlanta, 2000), ISBN 1-884932-39-8.

11 Appendix: Meaning of the rulers below the Smith chart

How to use the rulers that are often plotted below the Smith chart?

A commonly used set of rulers is usually found below the Smith chart, see Fig. 63. There are four rulers, some with an upper and lower part, to quickly estimate and compare some important properties in terms of modulus values. For the following discussion lets split the upper three rulers at the line marked *CENTER* to a left and right part, each to be discussed separately. These rulers start at the *CENTER*, referring to the center of the Smith chart, and end at the left or right boundary, referring to the circular boundary of the Smith chart. The 4th ruler at the bottom is different, it starts at the left boundary *ORIGIN* and ends at the right boundary.

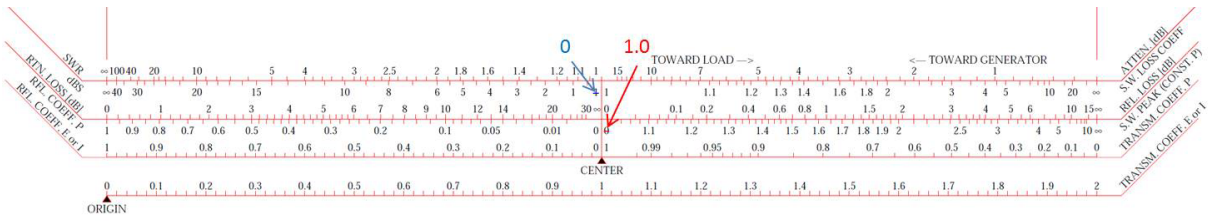


Fig. 63: Example for a set of rulers that can be found underneath the Smith chart (please note corrections in respect to the RF-course printouts)

First ruler, left/upper part in Fig. 64 is marked as *SWR* which mean actually VSWR, i.e. voltage standing wave ratio. It ranges between one – for the matched case (center of the Smith chart) and infinity – for total reflection (boundary of the Smith chart), respectively. The upper part is in linear scale, the

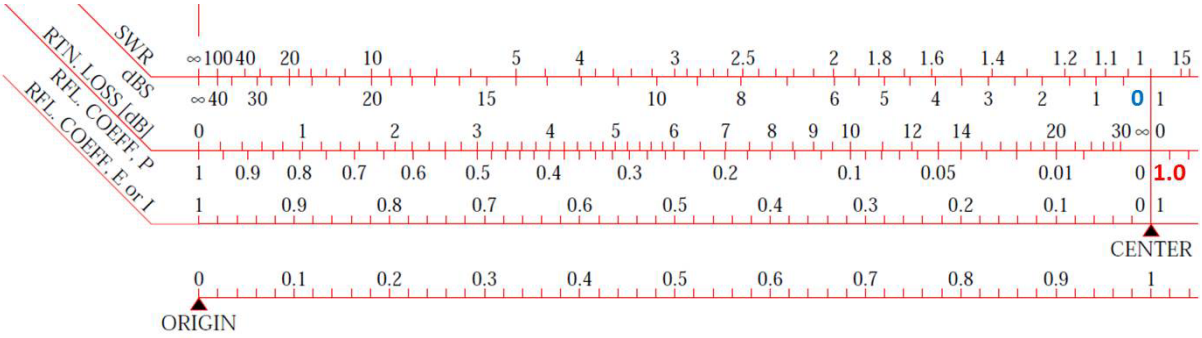


Fig. 64: Left part of the rulers usually plotted underneath the Smith Chart

lower part of this ruler is in dB, noted as dBS (dB referred to Standing Wave Ratio). Example: SWR = 10 corresponds to 20 dBS, SWR = 100 corresponds to 40 dBS [voltage ratios, not power ratios].

Second ruler, left/upper part, marked as *RTN.LOSS* i.e. return loss in dB. This indicates the amount of reflected wave expressed in dB. Thus, in the center of SC nothing is reflected and the return loss is infinite. At the boundary we have full reflection, thus return loss is 0 dB. The lower part of the scale denoted as *RFL.COEFF.P* is a reflection coefficient in terms of POWER (proportional $|\Gamma|^2$). If there is no reflected power for the matched case locus is in the center of the Smith chart (SC). On the contrary, if normalized reflected power is equal to 1 locus is at the boundary.

Third ruler, left, marked as *RFL.COEFF,E or I* gives us the absolute value of the reflection coefficient in linear scale. Note that since we have the modulus we can refer it both to voltage or current as we have omitted the sign. Obviously in the center the reflection coefficient is zero, at the boundary it is one.

The fourth is a Voltage transmission coefficient. Note that the modulus of the voltage (and current) transmission coefficient has a range from zero, i.e. short circuit, to +2 (open = 1+Γ with Γ = 1). This ruler is only valid for Z_{load} = real, i.e. the case of a step in characteristic impedance of the coaxial line.

Third ruler, right (see Fig. 65) marked as *TRANSM.COEFF.P* refers to the transmitted power as a function of mismatch and displays essentially the relation $P_t = 1 - |\Gamma|^2$. Thus, in the center of the SC full match, all the power is transmitted. At the boundary we have total reflection and e.g. for a Γ value of 0.5 we see that 75 % of the incident power is transmitted.

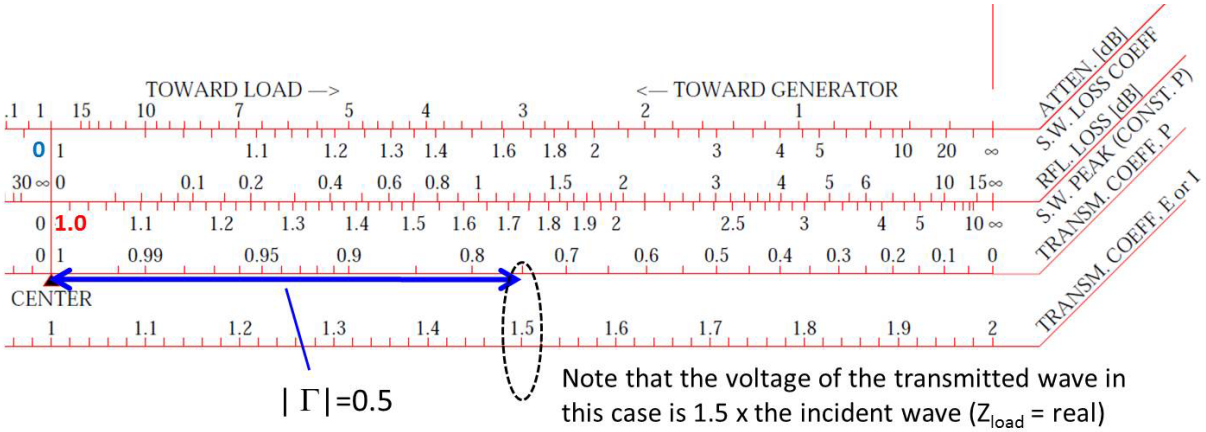


Fig. 65: Right part of the rulers usually plotted underneath the Smith Chart

Second ruler, right/upper part, denoted as *RFL.LOSS in dB* denotes reflection loss. This ruler refers to the loss in the transmitted wave, and should not be confounded with the return loss referring to the reflected wave. It displays the relation $P_t = 1 - |\Gamma|^2$ in dB. This ruler is nowadays rather not more

in use.

Let us analyse an example from Fig. 66: $|\Gamma| = 1/\sqrt{2} = 0.707$, transmitted power = 50 % thus loss = 50 % = 3 dB. Note that in the lowest ruler the voltage of the transmitted wave (Z_{load} = real) would be $V_t = 1.707 = 1 + 1/\sqrt{2}$ if referring to the voltage.

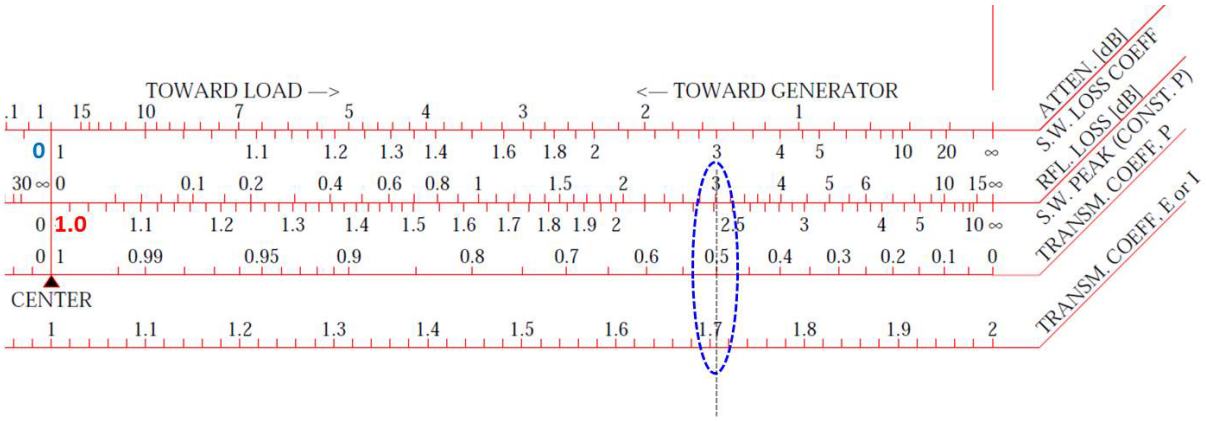


Fig. 66: Example for $|\Gamma| = 1/\sqrt{2} = 0.707$ and 50 % of transmitted power (i.e. 3 dB loss), see description in text

Finally, the First ruler, right/upper part, denoted as *ATTEN. in dB* assumes that one is measuring an attenuator or a lossy line which itself is terminated by an open or short circuit (full reflection). Thus the wave is traveling twice through the attenuator (forward and backward). The value of this attenuator can be between zero and some very high number corresponding to the matched case. The lower scale of first ruler displays the same situation just in terms of VSWR.

For the next example see Fig. 67: an 10 dB attenuator attenuates the reflected wave by 20 dB going forth and back and we get a reflection coefficient of Γ = 0.1. This correspond to the reflection of 10 % in voltage. Another example is 3 dB attenuator: for the forth and back transmission it gives 6 dB which correspond to half of the voltage. Table 5 is reprinted from an original paper of Phillip H. Smith [22] and summarizes reflection formulas discussed above.

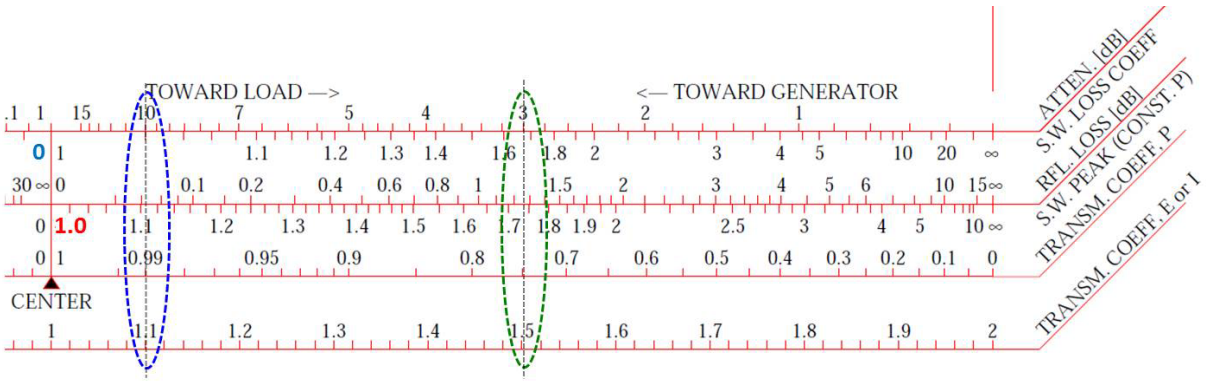


Fig. 67: Example for 10 dB and 3 dB attenuator, see description in text

Table 5: Reflection formulas

function	traveling waves	reflection coefficient	standing waves
VOLTAGE REFL. COEFF.	$\frac{r}{i}$	Γ	$\frac{S-1}{S+1}$
POWER REFL. COEF.	$(\frac{r}{i})^2$	Γ^2	$(\frac{S-1}{S+1})^2$
RETURN LOSS [dB]	$10 \cdot \log(\frac{i}{r})^2$	$-10 \cdot \log(\Gamma^2)$	$-10 \cdot \log(\frac{S-1}{S+1})^2$
REFLECTION LOSS [dB]	$10 \cdot \log(\frac{i^2}{i^2-r^2})$	$-10 \cdot \log(1-\Gamma^2)$	$-10 \cdot \log[1-(\frac{S-1}{S+1})^2]$
STDG. WAVE LOSS COEF.	$1-\frac{[(i+r)/(i-r)]^2}{2[(i+r)/(i-r)]}$	$\frac{1-\Gamma+\Gamma^2-\Gamma^3}{1-\Gamma-\Gamma^2+\Gamma^3}$	$\frac{1+S^2}{2S}$
STDG. WAVE RATIO [dB]	$20 \cdot \log(\frac{i+r}{i-r})$	$20 \cdot \log(\frac{1+\Gamma}{1-\Gamma})$	$20 \cdot \log(S)$
MAX. OF STDG. WAVE	$(\frac{i+r}{i-r})^{1/2}$	$(\frac{1+\Gamma}{1-\Gamma})^{1/2}$	$\sqrt{S}$
MIN. OF STDG. WAVE	$(\frac{i-r}{i+r})^{1/2}$	$(\frac{1-\Gamma}{1+\Gamma})^{1/2}$	$\frac{1}{\sqrt{S}}$
STANDING WAVE RATIO	$\frac{i+r}{i-r}$	$\frac{1+\Gamma}{1-\Gamma}$	S
ATTENUATION [dB]	$-10 \cdot \log(\frac{r}{i})$	$-10 \cdot \log(\Gamma)$	$-10 \cdot \log(\frac{S-1}{S+1})$

whereas: i = incident wave amplitude, r = reflected wave amplitude, Γ = reflection coefficient, $S \equiv$ SWR = voltage standing wave ratio.