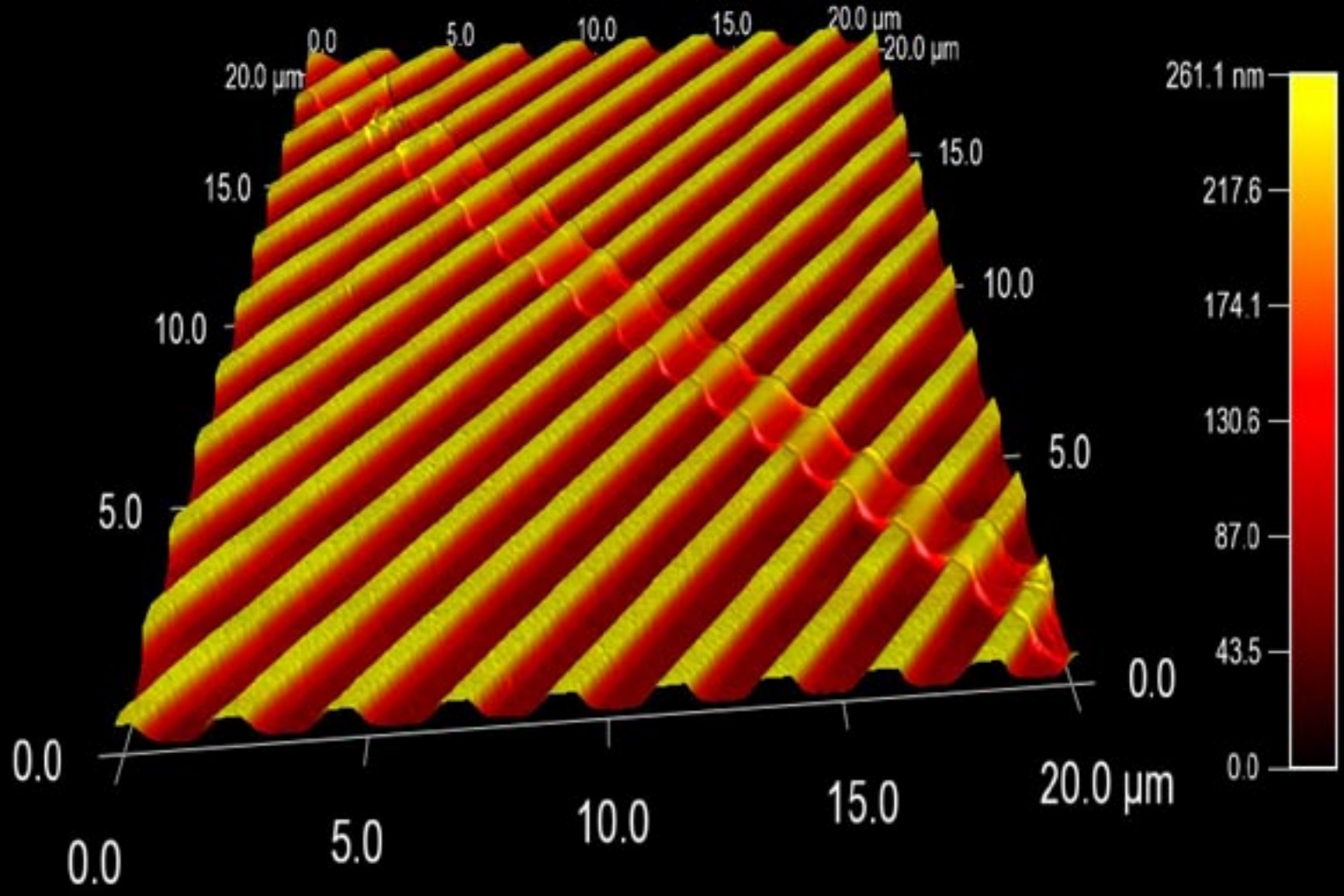
Electrical Parameters

Joyce Bradbury



First Edition, 2012

ISBN 978-81-323-3042-4

Published by:
Research World
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter 1

Bode Plot

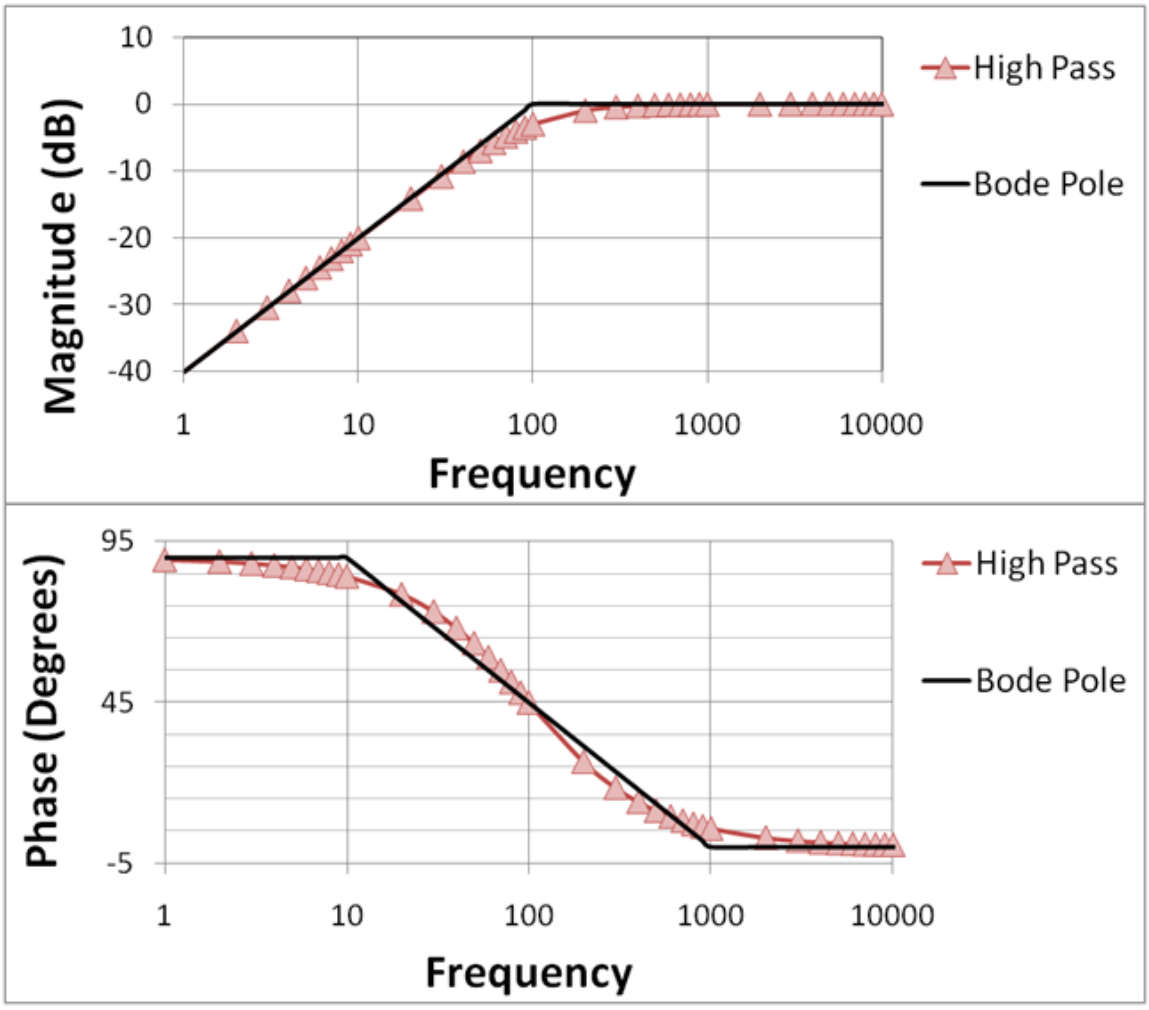


Figure 1(a): The Bode plot for a first-order (one-pole) highpass filter; the straight-line approximations are labeled "Bode pole"; phase varies from 90° at low frequencies (due to the contribution of the numerator, which is 90° at all frequencies) to 0° at high

frequencies (where the phase contribution of the denominator is −90° and cancels the contribution of the numerator).

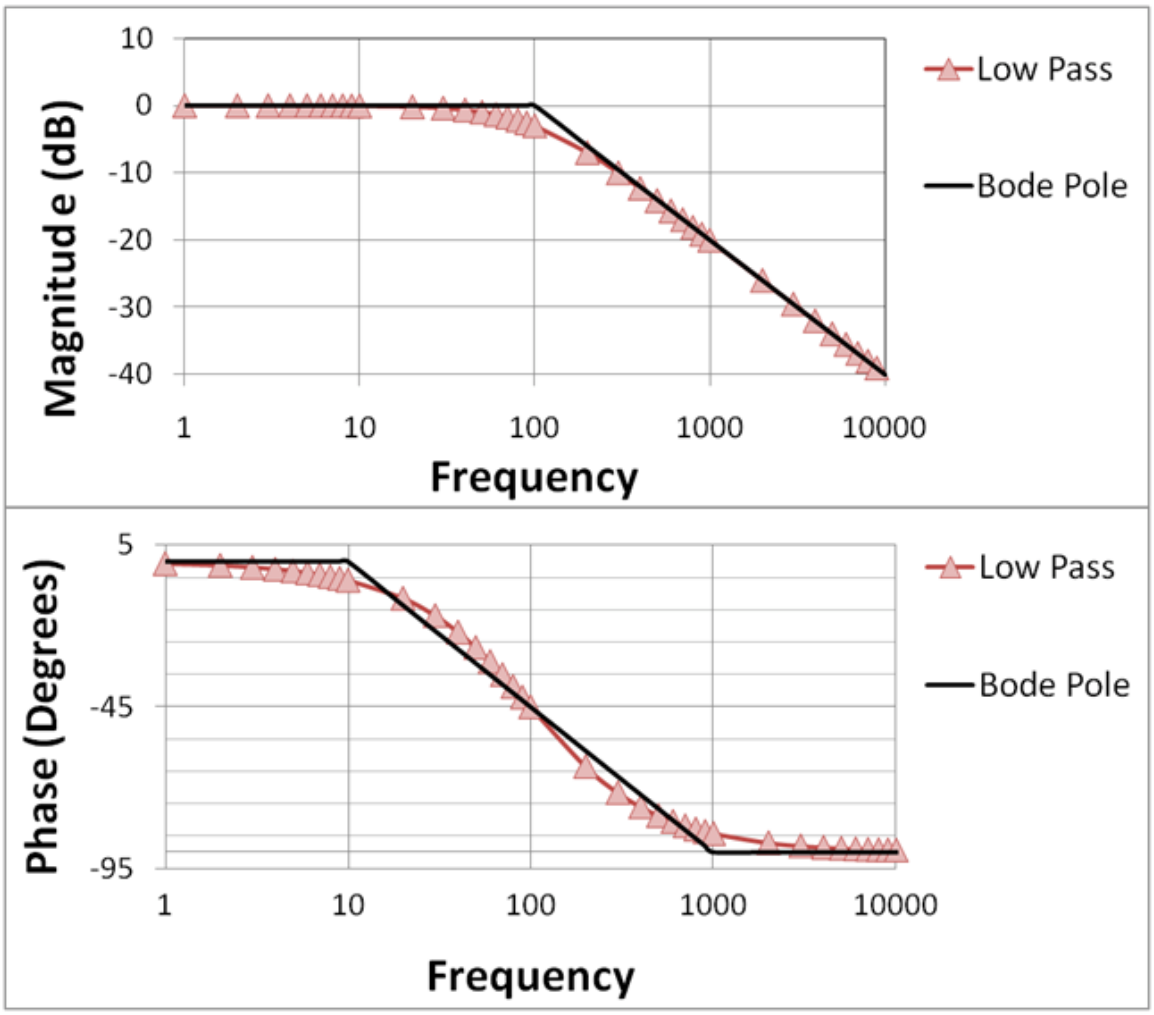


Figure 1(b): The Bode plot for a first-order (one-pole) lowpass filter; the straight-line approximations are labeled "Bode pole"; phase is 90° lower than for Figure 1(a) because the phase contribution of the numerator is 0° at all frequencies.

A **Bode plot** is a graph of the transfer function of a linear, time-invariant system versus frequency, plotted with a log-frequency axis, to show the system's frequency response. It is usually a combination of a **Bode magnitude plot,** expressing the magnitude of the frequency response gain, and a **Bode phase plot,** expressing the frequency response phase shift.

Overview

Among his several important contributions to circuit theory and control theory, engineer Hendrik Wade Bode (1905–1982), while working at Bell Labs in the United States in the 1930s, devised a simple but accurate method for graphing gain and phase-shift plots.

These bear his name, *Bode gain plot* and *Bode phase plot* (pronounced *Boh-dee* in English, *Bow-duh* in Dutch).

The magnitude axis of the Bode plot is usually expressed as decibels of power, that is by the 20 log rule: 20 times the common (base 10) logarithm of the amplitude gain. With the magnitude gain being logarithmic, Bode plots make multiplication of magnitudes a simple matter of adding distances on the graph (in decibels), since

$$\log(a \cdot b) = \log(a) + \log(b).$$

A **Bode phase plot** is a graph of phase versus frequency, also plotted on a log-frequency axis, usually used in conjunction with the magnitude plot, to evaluate how much a signal will be phase-shifted. For example a signal described by: $A\sin(\omega t)$ may be attenuated but also phase-shifted. If the system attenuates it by a factor x and phase shifts it by $-\Phi$ the signal out of the system will be $(A/x)\sin(\omega t - \Phi)$. The phase shift Φ is generally a function of frequency.

Phase can also be added directly from the graphical values, a fact that is mathematically clear when phase is seen as the imaginary part of the complex logarithm of a complex gain.

In Figure 1(a), the Bode plots are shown for the one-pole highpass filter function:

$$\mathrm{T_{High}}(f) = \frac{jf/f_1}{1 + jf/f_1}\ ,$$

where f is the frequency in Hz, and f_1 is the pole position in Hz, $f_1 = 100$ Hz in the figure. Using the rules for complex numbers, the magnitude of this function is

$$\mid \mathrm{T_{High}}(f) \mid = \frac{f/f_1}{\sqrt{1 + (f/f_1)^2}},$$

while the phase is:

$$\varphi_{T_{High}} = 90^\circ - \tan^{-1}(f/f_1).$$

Care must be taken that the inverse tangent is set up to return *degrees*, not radians. On the Bode magnitude plot, decibels are used, and the plotted magnitude is:

$$\begin{aligned} 20\log_{10} \mid \mathrm{T_{High}}(f) \mid &= 20\log_{10}(f/f_1) \\ &-20\log_{10}\left(\sqrt{1 + (f/f_1)^2}\right). \end{aligned}$$

In Figure 1(b), the Bode plots are shown for the one-pole lowpass filter function:

$$\mathrm{T_{Low}}(f) = \frac{1}{1 + jf/f_1}.$$

Also shown in Figure 1(a) and 1(b) are the straight-line approximations to the Bode plots that are used in hand analysis, and described later.

The magnitude and phase Bode plots can seldom be changed independently of each other — changing the amplitude response of the system will most likely change the phase characteristics and vice versa. For minimum-phase systems the phase and amplitude characteristics can be obtained from each other with the use of the Hilbert transform.

If the transfer function is a rational function with real poles and zeros, then the Bode plot can be approximated with straight lines. These asymptotic approximations are called **straight line Bode plots** or **uncorrected Bode plots** and are useful because they can be drawn by hand following a few simple rules. Simple plots can even be predicted without drawing them.

The approximation can be taken further by *correcting* the value at each cutoff frequency. The plot is then called a **corrected Bode plot**.

Rules for hand-made Bode plot

The premise of a Bode plot is that one can consider the log of a function in the form:

$$f(x) = A \prod (x + c_n)^{a_n}$$

as a sum of the logs of its poles and zeros:

$$\log(f(x)) = \log(A) + \sum a_n \log(x + c_n).$$

This idea is used explicitly in the method for drawing phase diagrams. The method for drawing amplitude plots implicitly uses this idea, but since the log of the amplitude of each pole or zero always starts at zero and only has one asymptote change (the straight lines), the method can be simplified.

Straight-line amplitude plot

Amplitude decibels is usually done using the $20\log_{10}(X)$ version. Given a transfer function in the form

$$H(s) = A \prod \frac{(s + x_n)^{a_n}}{(s + y_n)^{b_n}}$$

where x_n and y_n are constants, $s = j\omega$, $a_n, b_n > 0$, and H is the transfer function:

- at every value of s where ω = x_n (a zero), **increase** the slope of the line by $20 \cdot a_n \ dB$ per decade.
- at every value of s where ω = y_n (a pole), **decrease** the slope of the line by $20 \cdot b_n \ dB$ per decade.
- The initial value of the graph depends on the boundaries. The initial point is found by putting the initial angular frequency ω into the function and finding $|H(j\omega)|$.
- The initial slope of the function at the initial value depends on the number and order of zeros and poles that are at values below the initial value, and are found using the first two rules.

To handle irreducible 2nd order polynomials, $ax^2 + bx + c$ can, in many cases, be approximated as $(\sqrt{a}x + \sqrt{c})^2$.

Note that zeros and poles happen when ω is *equal to* a certain x_n or y_n. This is because the function in question is the magnitude of H(jω), and since it is a complex function, $|H(j\omega)| = \sqrt{H \cdot H^*}$. Thus at any place where there is a zero or pole involving the term $(s + x_n)$, the magnitude of that term is $\sqrt{(x_n + j\omega) \cdot (x_n - j\omega)} = \sqrt{x_n^2 + \omega^2}$.

Corrected amplitude plot

To correct a straight-line amplitude plot:

- at every zero, put a point $3 \cdot a_n \ \mathrm{dB}$ **above** the line,
- at every pole, put a point $3 \cdot b_n \ \mathrm{dB}$ **below** the line,
- draw a smooth curve through those points using the straight lines as asymptotes (lines which the curve approaches).

Note that this correction method does not incorporate how to handle complex values of x_n or y_n. In the case of an irreducible polynomial, the best way to correct the plot is to actually calculate the magnitude of the transfer function at the pole or zero corresponding to the irreducible polynomial, and put that dot over or under the line at that pole or zero.

Straight-line phase plot

Given a transfer function in the same form as above:

$$H(s) = A \prod \frac{(s + x_n)^{a_n}}{(s + y_n)^{b_n}}$$

the idea is to draw separate plots for each pole and zero, then add them up. The actual phase curve is given by − arctan(Im[$H(s)$] / Re[$H(s)$]).

To draw the phase plot, for **each** pole and zero:

- if A is positive, start line (with zero slope) at 0 degrees
- if A is negative, start line (with zero slope) at 180 degrees
- at every $\omega = x_n$ (for stable zeros – $Re(z) < 0$), **increase** the slope by $45 \cdot a_n$ degrees per decade, beginning one decade before $\omega = x_n$ (E.g.: $\frac{x_n}{10}$)
- at every $\omega = y_n$ (for stable poles – $Re(p) < 0$), **decrease** the slope by $45 \cdot b_n$ degrees per decade, beginning one decade before $\omega = y_n$ (E.g.: $\frac{y_n}{10}$)
- "unstable" (right half plane) poles and zeros ($Re(s) > 0$) have opposite behavior
- flatten the slope again when the phase has changed by $90 \cdot a_n$ degrees (for a zero) or $90 \cdot b_n$ degrees (for a pole),
- After plotting one line for each pole or zero, add the lines together to obtain the final phase plot; that is, the final phase plot is the superposition of each earlier phase plot.

Example

A passive (unity pass band gain) lowpass RC filter, for instance has the following transfer function expressed in the frequency domain:

$$H(jf) = \frac{1}{1 + j2\pi fRC}.$$

From the transfer function it can be determined that the cutoff frequency point f_c (in hertz) is at the frequency

$$f_c = \frac{1}{2\pi RC}$$

or (equivalently) at

$$\omega_c = \frac{1}{RC}$$ where $\omega_c = 2\pi f_c$ is the angular cutoff frequency in radians per second.

The transfer function in terms of the angular frequencies becomes:

$$H(j\omega) = \frac{1}{1 + j\frac{\omega}{\omega_c}}.$$

The above equation is the normalized form of the transfer function. The Bode plot is shown in Figure 1(b) above, and construction of the straight-line approximation is discussed next.

Magnitude plot

The magnitude (in decibels) of the transfer function above, (normalized and converted to angular frequency form), given by the decibel gain expression A_{vdB}:

$$A_{\text{vdB}} = 20\log|H(j\omega)| = 20\log\frac{1}{\left|1 + j\frac{\omega}{\omega_{\text{c}}}\right|}$$
$$= -20\log\left|1 + j\frac{\omega}{\omega_{\text{c}}}\right| = -10\log\left[1 + \frac{\omega^2}{\omega_{\text{c}}^2}\right]$$

when plotted versus input frequency ω on a logarithmic scale, can be approximated by two lines and it forms the asymptotic (approximate) magnitude Bode plot of the transfer function:

- for angular frequencies below ω_c it is a horizontal line at 0 dB since at low frequencies the $\frac{\omega}{\omega_{\text{c}}}$ term is small and can be neglected, making the decibel gain equation above equal to zero,
- for angular frequencies above ω_c it is a line with a slope of −20 dB per decade since at high frequencies the $\frac{\omega}{\omega_{\text{c}}}$ term dominates and the decibel gain expression above simplifies to $-20\log\frac{\omega}{\omega_{\text{c}}}$ which is a straight line with a slope of −20 dB per decade.

These two lines meet at the corner frequency. From the plot, it can be seen that for frequencies well below the corner frequency, the circuit has an attenuation of 0 dB, corresponding to a unity pass band gain, i.e. the amplitude of the filter output equals the amplitude of the input. Frequencies above the corner frequency are attenuated – the higher the frequency, the higher the attenuation.

Phase plot

The phase Bode plot is obtained by plotting the phase angle of the transfer function given by

$$\varphi = -\tan^{-1}\frac{\omega}{\omega_{\text{c}}}$$

versus ω, where ω and ω_c are the input and cutoff angular frequencies respectively. For input frequencies much lower than corner, the ratio $\frac{\omega}{\omega_{\text{c}}}$ is small and therefore the phase angle is close to zero. As the ratio increases the absolute value of the phase increases and

becomes –45 degrees when $\omega = \omega_c$. As the ratio increases for input frequencies much greater than the corner frequency, the phase angle asymptotically approaches −90 degrees. The frequency scale for the phase plot is logarithmic.

Normalized plot

The horizontal frequency axis, in both the magnitude and phase plots, can be replaced by the normalized (nondimensional) frequency ratio $\frac{\omega}{\omega_c}$. In such a case the plot is said to be normalized and units of the frequencies are no longer used since all input frequencies are now expressed as multiples of the cutoff frequency ω_c.

An example with pole and zero

Figures 2-5 further illustrate construction of Bode plots. This example with both a pole and a zero shows how to use superposition. To begin, the components are presented separately.

Figure 2 shows the Bode magnitude plot for a zero and a low-pass pole, and compares the two with the Bode straight line plots. The straight-line plots are horizontal up to the pole (zero) location and then drop (rise) at 20 dB/decade. The second Figure 3 does the same for the phase. The phase plots are horizontal up to a frequency factor of ten below the pole (zero) location and then drop (rise) at 45°/decade until the frequency is ten times higher than the pole (zero) location. The plots then are again horizontal at higher frequencies at a final, total phase change of 90°.

Figure 4 and Figure 5 show how superposition (simple addition) of a pole and zero plot is done. The Bode straight line plots again are compared with the exact plots. The zero has been moved to higher frequency than the pole to make a more interesting example. Notice in Figure 4 that the 20 dB/decade drop of the pole is arrested by the 20 dB/decade rise of the zero resulting in a horizontal magnitude plot for frequencies above the zero location. Notice in Figure 5 in the phase plot that the straight-line approximation is pretty approximate in the region where both pole and zero affect the phase. Notice also in Figure 5 that the range of frequencies where the phase changes in the straight line plot is limited to frequencies a factor of ten above and below the pole (zero) location. Where the phase of the pole and the zero both are present, the straight-line phase plot is horizontal because the 45°/decade drop of the pole is arrested by the overlapping 45°/decade rise of the zero in the limited range of frequencies where both are active contributors to the phase.

Example with pole and zero

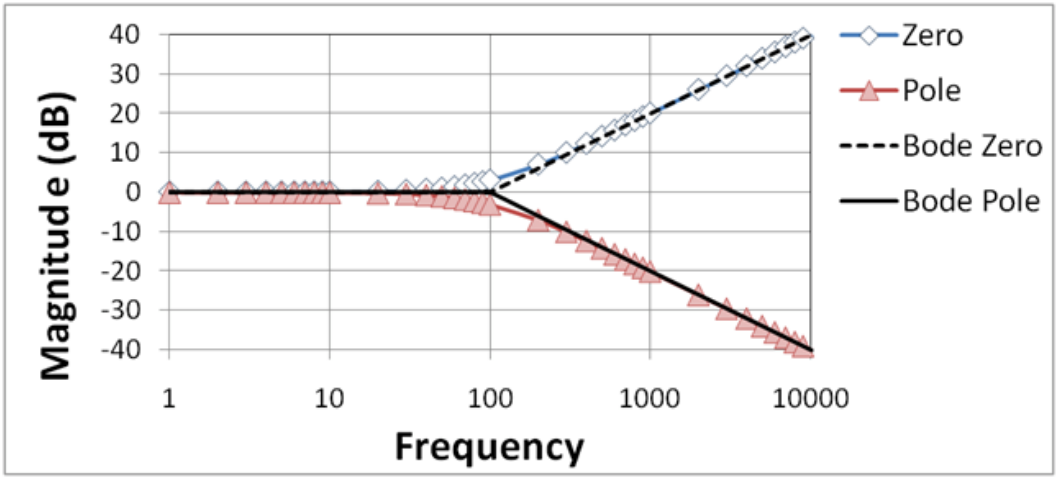


Figure 2: Bode magnitude plot for zero and low-pass pole; curves labeled "Bode" are the straight-line Bode plots

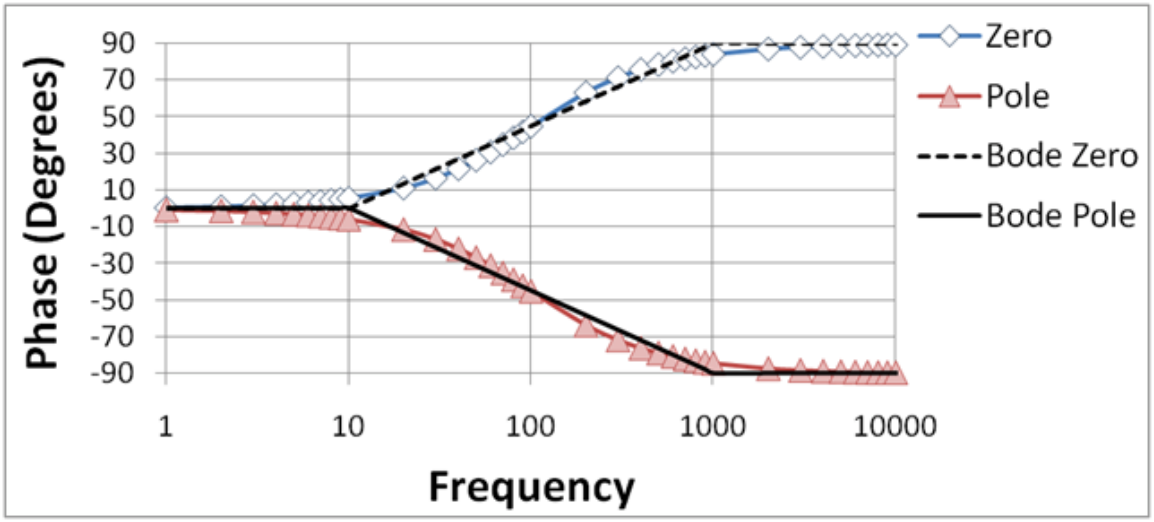


Figure 3: Bode phase plot for zero and low-pass pole; curves labeled "Bode" are the straight-line Bode plots

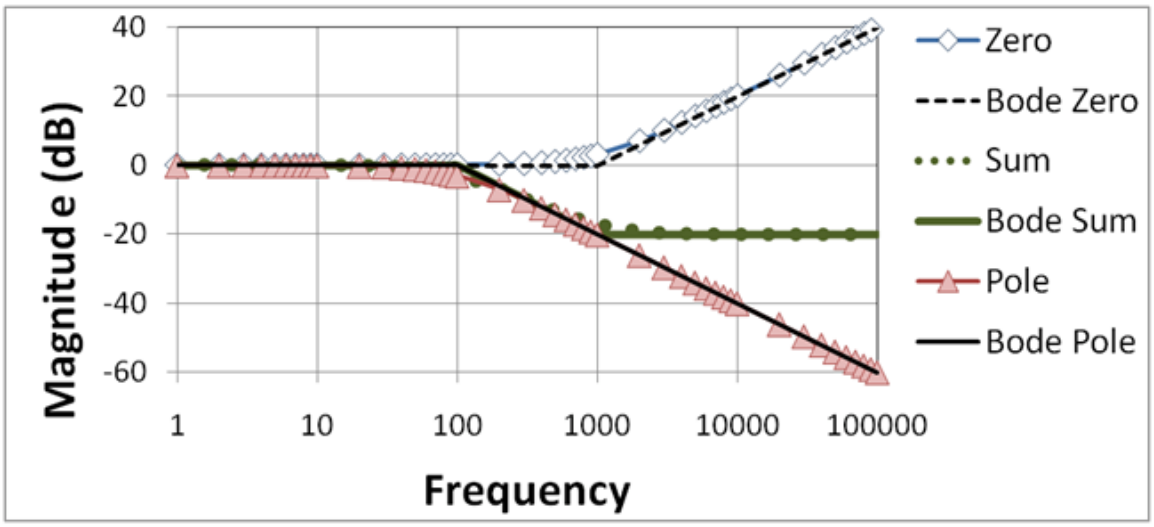


Figure 4: Bode magnitude plot for pole-zero combination; the location of the zero is ten times higher than in Figures 2&3; curves labeled "Bode" are the straight-line Bode plots

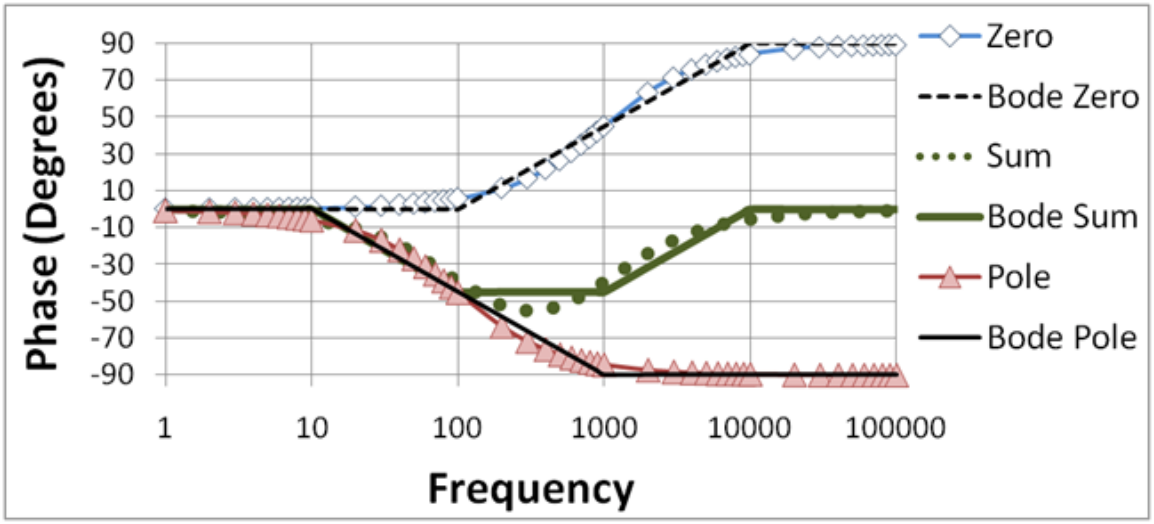


Figure 5: Bode phase plot for pole-zero combination; the location of the zero is ten times higher than in Figures 2&3; curves labeled "Bode" are the straight-line Bode plots

Gain margin and phase margin

Bode plots are used to assess the stability of negative feedback amplifiers by finding the gain and phase margins of an amplifier. The notion of gain and phase margin is based upon the gain expression for a negative feedback amplifier given by

$$A_{FB} = \frac{A_{OL}}{1 + \beta A_{OL}} ,$$

where A_{FB} is the gain of the amplifier with feedback (the **closed-loop gain**), β is the **feedback factor** and A_{OL} is the gain without feedback (the **open-loop gain**). The gain A_{OL} is a complex function of frequency, with both magnitude and phase. Examination of this relation shows the possibility of infinite gain (interpreted as instability) if the product $\beta A_{OL} = -1$. (That is, the magnitude of βA_{OL} is unity and its phase is −180°, the so-called **Barkhausen stability criterion**). Bode plots are used to determine just how close an amplifier comes to satisfying this condition.

Key to this determination are two frequencies. The first, labeled here as f_{180}, is the frequency where the open-loop gain flips sign. The second, labeled here f_{0dB}, is the frequency where the magnitude of the product $| \beta A_{OL} | = 1$ (in dB, magnitude 1 is 0 dB). That is, frequency f_{180} is determined by the condition:

$$\beta A_{OL}\left(f_{180}\right) = -\left|\beta A_{OL}\left(f_{180}\right)\right| = -\left|\beta A_{OL}\right|_{180},$$

where vertical bars denote the magnitude of a complex number (for example, $| a + j\, b | = [\, a^2 + b^2]^{1/2}$), and frequency f_{0dB} is determined by the condition:

$$\left|\beta A_{OL}\left(f_{0dB}\right)\right| = 1.$$

One measure of proximity to instability is the **gain margin**. The Bode phase plot locates the frequency where the phase of βA_{OL} reaches −180°, denoted here as frequency f_{180}. Using this frequency, the Bode magnitude plot finds the magnitude of βA_{OL}. If $|\beta A_{OL}|_{180} = 1$, the amplifier is unstable, as mentioned. If $|\beta A_{OL}|_{180} < 1$, instability does not occur, and the separation in dB of the magnitude of $|\beta A_{OL}|_{180}$ from $|\beta A_{OL}| = 1$ is called the *gain margin*. Because a magnitude of one is 0 dB, the gain margin is simply one of the equivalent forms: $20 \log_{10}(|\beta A_{OL}|_{180}) = 20 \log_{10}(|A_{OL}|_{180}) - 20 \log_{10}(1 / \beta)$.

Another equivalent measure of proximity to instability is the **phase margin**. The Bode magnitude plot locates the frequency where the magnitude of $|\beta A_{OL}|$ reaches unity, denoted here as frequency f_{0dB}. Using this frequency, the Bode phase plot finds the phase of βA_{OL}. If the phase of $\beta A_{OL}(f_{0dB}) > -180°$, the instability condition cannot be met at any frequency (because its magnitude is going to be < 1 when $f = f_{180}$), and the distance of the phase at f_{0dB} in degrees above −180° is called the *phase margin*.

If a simple *yes* or *no* on the stability issue is all that is needed, the amplifier is stable if $f_{0dB} < f_{180}$. This criterion is sufficient to predict stability only for amplifiers satisfying some restrictions on their pole and zero positions (minimum phase systems). Although these restrictions usually are met, if they are not another method must be used, such as the Nyquist plot.

Examples using Bode plots

Figures 6 and 7 illustrate the gain behavior and terminology. For a three-pole amplifier, Figure 6 compares the Bode plot for the gain without feedback (the *open-loop* gain) A_{OL} with the gain with feedback A_{FB} (the *closed-loop* gain).

In this example, $A_{OL} = 100$ dB at low frequencies, and $1 / \beta = 58$ dB. At low frequencies, $A_{FB} \approx 58$ dB as well.

Because the open-loop gain A_{OL} is plotted and not the product βA_{OL}, the condition $A_{OL} = 1 / \beta$ decides f_{0dB}. The feedback gain at low frequencies and for large A_{OL} is $A_{FB} \approx 1 / \beta$ (look at the formula for the feedback gain at the beginning of this section for the case of large gain A_{OL}), so an equivalent way to find f_{0dB} is to look where the feedback gain intersects the open-loop gain. (Frequency f_{0dB} is needed later to find the phase margin.)

Near this crossover of the two gains at f_{0dB}, the Barkhausen criteria are almost satisfied in this example, and the feedback amplifier exhibits a massive peak in gain (it would be infinity if $\beta A_{OL} = -1$). Beyond the unity gain frequency f_{0dB}, the open-loop gain is sufficiently small that $A_{FB} \approx A_{OL}$ (examine the formula at the beginning of this section for the case of small A_{OL}).

Figure 7 shows the corresponding phase comparison: the phase of the feedback amplifier is nearly zero out to the frequency f_{180} where the open-loop gain has a phase of −180°. In this vicinity, the phase of the feedback amplifier plunges abruptly downward to become almost the same as the phase of the open-loop amplifier. (Recall, $A_{FB} \approx A_{OL}$ for small A_{OL}.)

Comparing the labeled points in Figure 6 and Figure 7, it is seen that the unity gain frequency f_{0dB} and the phase-flip frequency f_{180} are very nearly equal in this amplifier, $f_{180} \approx f_{0dB} \approx 3.332$ kHz, which means the gain margin and phase margin are nearly zero. The amplifier is borderline stable.

Figures 8 and 9 illustrate the gain margin and phase margin for a different amount of feedback β. The feedback factor is chosen smaller than in Figure 6 or 7, moving the condition $|\beta A_{OL}| = 1$ to lower frequency. In this example, $1 / \beta = 77$ dB, and at low frequencies $A_{FB} \approx 77$ dB as well.

Figure 8 shows the gain plot. From Figure 8, the intersection of $1 / \beta$ and A_{OL} occurs at $f_{0dB} = 1$ kHz. Notice that the peak in the gain A_{FB} near f_{0dB} is almost gone.

Figure 9 is the phase plot. Using the value of $f_{0dB} = 1$ kHz found above from the magnitude plot of Figure 8, the open-loop phase at f_{0dB} is −135°, which is a phase margin of 45° above −180°.

Using Figure 9, for a phase of −180° the value of $f_{180} = 3.332$ kHz (the same result as found earlier, of course). The open-loop gain from Figure 8 at f_{180} is 58 dB, and $1 / \beta = 77$ dB, so the gain margin is 19 dB.

Stability is not the sole criterion for amplifier response, and in many applications a more stringent demand than stability is good step response. As a rule of thumb, good step response requires a phase margin of at least 45°, and often a margin of over 70° is advocated, particularly where component variation due to manufacturing tolerances is an issue.

Examples

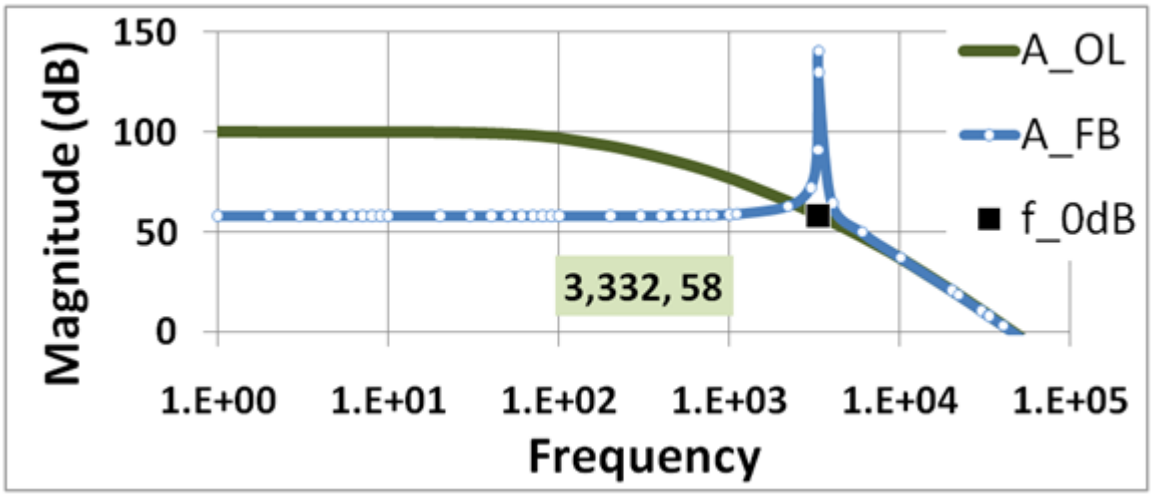


Figure 6: Gain of feedback amplifier A_{FB} in dB and corresponding open-loop amplifier A_{OL}. Parameter $1/\beta = 58$ dB, and at low frequencies $A_{FB} \approx 58$ dB as well. The gain margin in this amplifier is nearly zero because $|\beta A_{OL}| = 1$ occurs at almost $f = f_{180°}$.

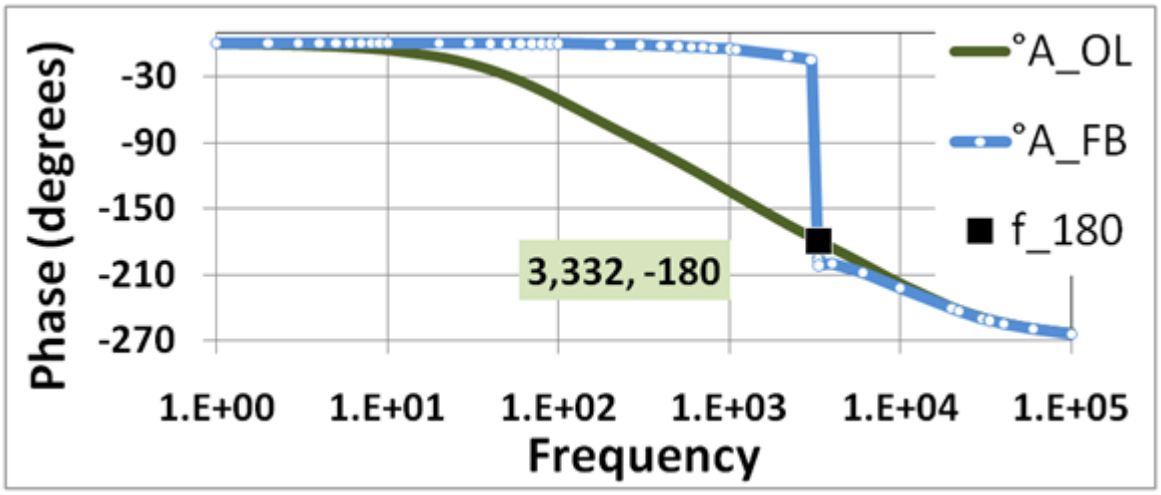


Figure 7: Phase of feedback amplifier $°A_{FB}$ in degrees and corresponding open-loop amplifier $°A_{OL}$. The phase margin in this amplifier is nearly zero because the phase-flip occurs at almost the unity gain frequency $f = f_{0dB}$ where $|\beta A_{OL}| = 1$.

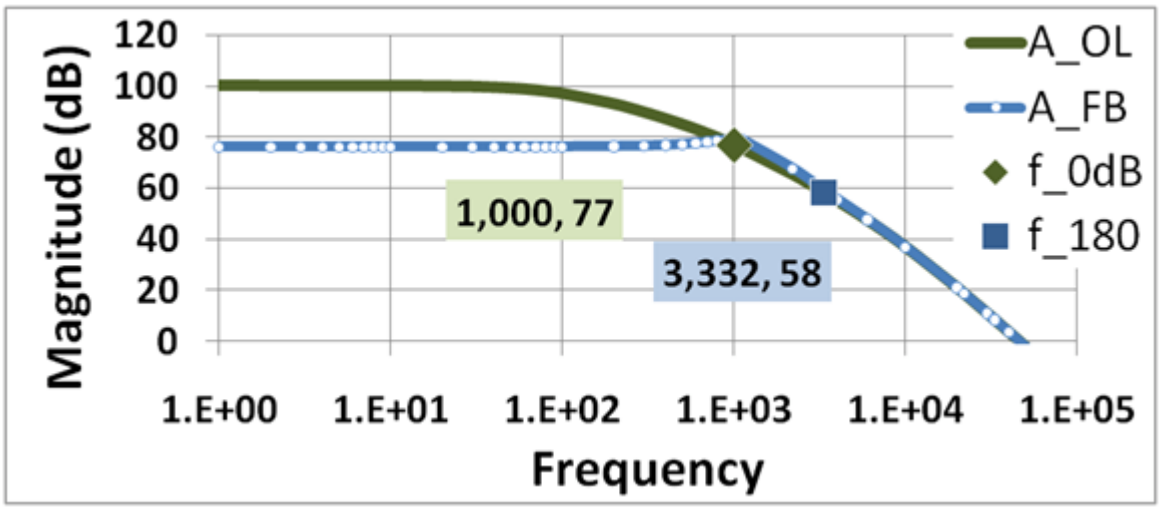


Figure 8: Gain of feedback amplifier A_{FB} in dB and corresponding open-loop amplifier A_{OL}. In this example, $1 / \beta = 77$ dB. The gain margin in this amplifier is 19 dB.

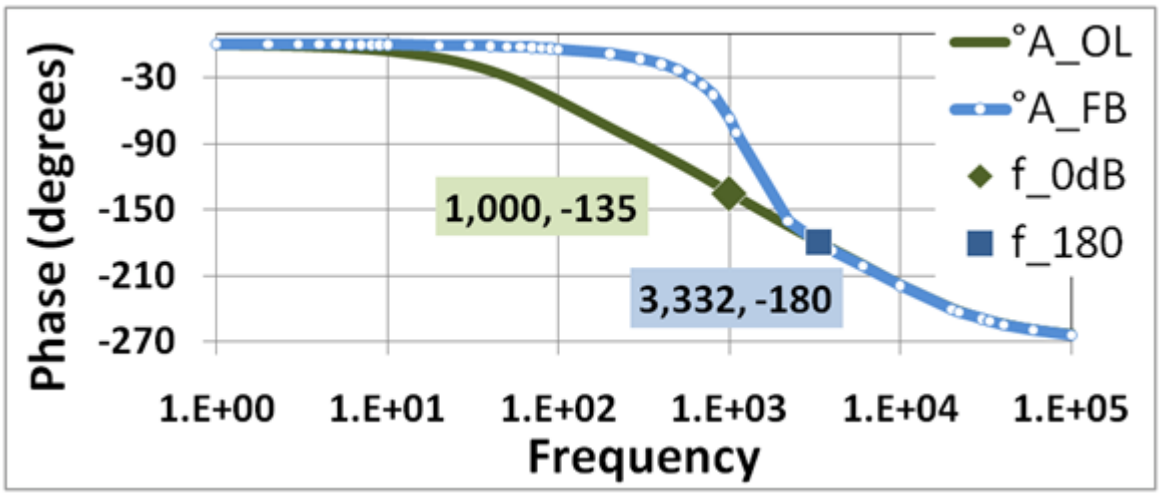


Figure 9: Phase of feedback amplifier A_{FB} in degrees and corresponding open-loop amplifier A_{OL}. The phase margin in this amplifier is 45°.

Bode plotter

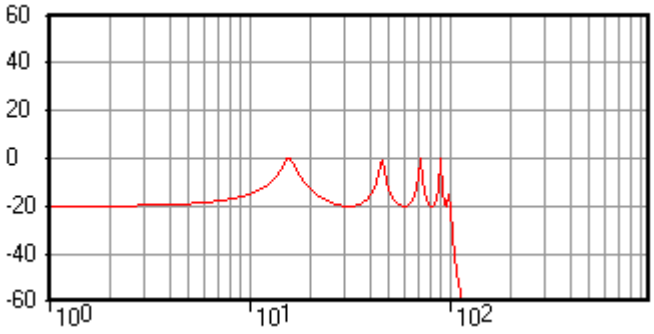


Figure 10: Amplitude diagram of a 10th order Chebyshev filter plotted using a Bode Plotter application. The chebyshev transfer function is defined by poles and zeros which are added by clicking on a graphical complex diagram.

The Bode plotter is an electronic instrument resembling an oscilloscope, which produces a Bode diagram, or a graph, of a circuit's voltage gain or phase shift plotted against frequency in a feedback control system or a filter. An example of this is shown in Figure 10. It is extremely useful for analyzing and testing filters and the stability of feedback control systems, through the measurement of corner (cutoff) frequencies and gain and phase margins.

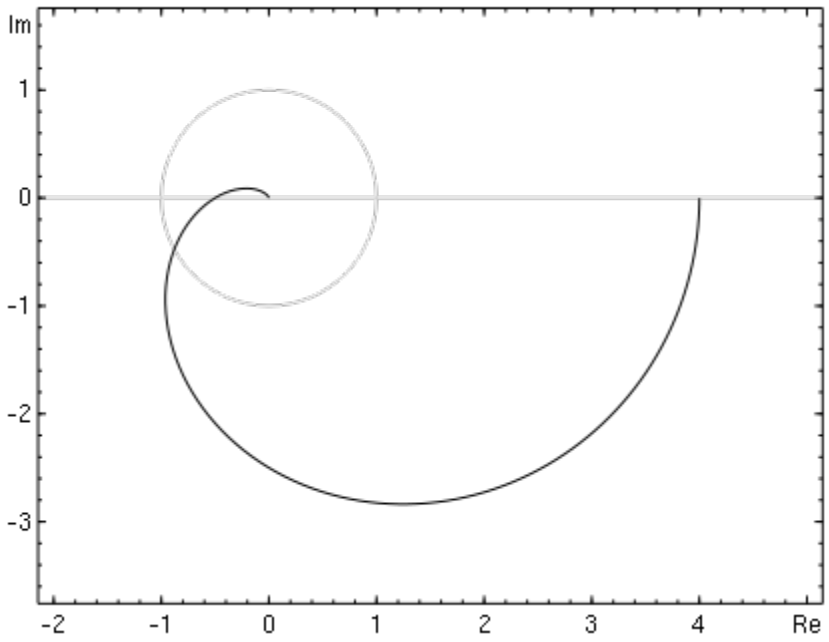
This is identical to the function performed by a vector network analyzer, but the network analyzer is typically used at much higher frequencies.

For education/research purposes usage of applications for plotting Bode diagrams for given transfer functions helps better understanding and faster getting of results.

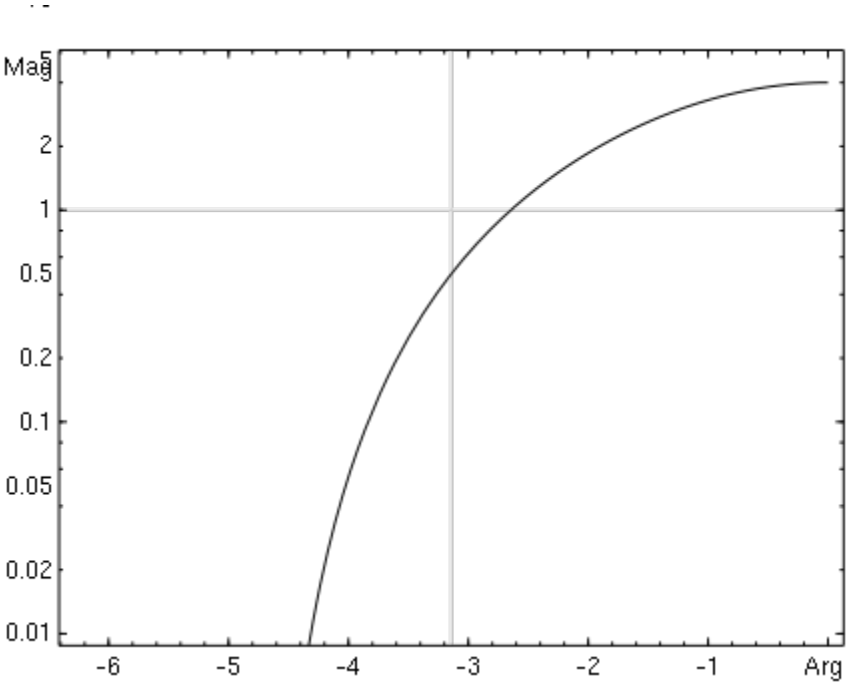
Related plots

Two related plots that display the same data in different coordinate systems are the Nyquist plot and the Nichols plot. These are parametric plots, with frequency as the input and magnitude and phase of the frequency response as the output. The Nyquist plot displays these in polar coordinates, with magnitude mapping to radius and phase to argument (angle). The Nichols plot displays these in rectangular coordinates, on the log scale.

Related Plots



A Nyquist plot.



A Nichols plot of the same response.

Chapter 2

Breakdown Voltage & Inrush current

Breakdown Voltage

High voltage dielectric breakdown within a block of plexiglas

The **breakdown voltage** of an insulator is the minimum voltage that causes a portion of an insulator to become electrically conductive.

The **breakdown voltage** of a diode is the minimum *reverse* voltage to make the diode conduct in reverse. Some devices (such as TRIACs) also have a *forward breakdown voltage*.

In Detail

Insulators

Breakdown voltage is a characteristic of an insulator that defines the maximum voltage difference that can be applied across the material before the insulator collapses and conducts. In solid insulating materials, this usually creates a weakened path within the material by creating permanent molecular or physical changes by the sudden current. Within rarefied gases found in certain types of lamps, **breakdown voltage** is also sometimes called the "striking voltage".

The breakdown voltage of a material is not a definite value because it is a form of failure and there is a statistical probability whether the material will fail at a given voltage. When a value is given it is usually the mean breakdown voltage of a large sample. Another term is also 'withstand voltage' where the probability of failure at a given voltage is so low it is considered, when designing insulation, that the material will not fail at this voltage.

Two different breakdown voltage measurements of a material are the AC and impulse breakdown voltages. The AC voltage is the line frequency of the mains (either 50 or 60 Hz depending on where you live). The impulse breakdown voltage is simulating lightning strikes, and usually uses a 1.2 microsecond rise for the wave to reach 90% amplitude then drops back down to 50% amplitude after 50 microseconds.

Two technical standards governing performing these tests are ASTM D1816 and ASTM D3300 published by ASTM.

Breakdown in vacuum

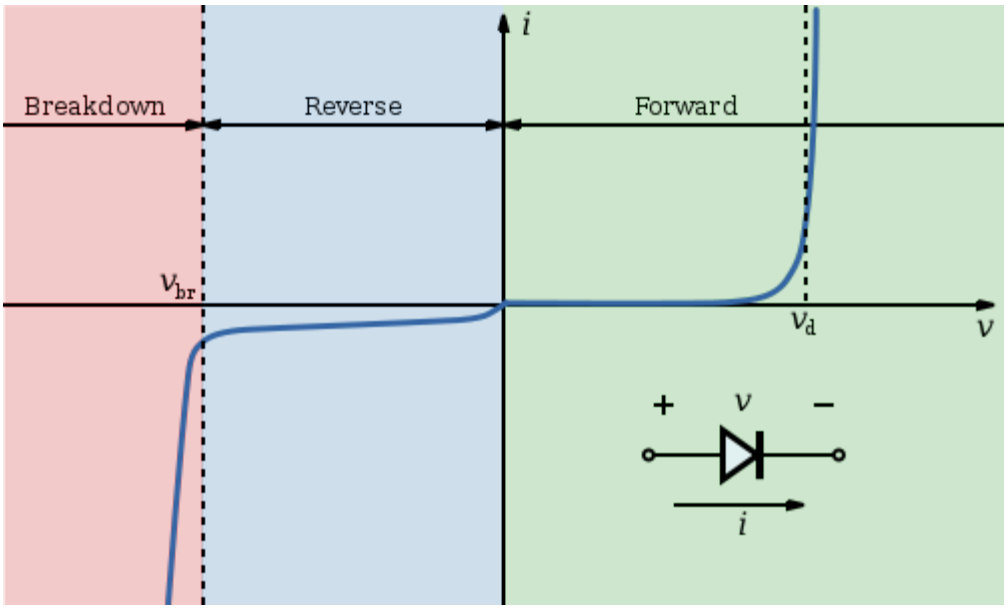
In standard conditions at atmospheric pressure, gas serves as an excellent insulator, requiring the application of a significant voltage before breaking down (e.g. lightning). In partial vacuum, this breakdown potential may decrease to an extent that two uninsulated surfaces with different potentials might induce the electrical breakdown of the surrounding gas. This has some useful applications in industry (e.g. the production of microprocessors) but in other situations may damage an apparatus, as breakdown is analogous to a short circuit.

The breakdown voltage in a partial vacuum is represented as :

$$V_\mathrm{b} = \frac{Bpd}{\ln Apd - \ln(1 + \frac{1}{\gamma_\mathrm{se}})}$$

where V_b is the breakdown potential in volts DC, A and B are constants that depend on the surrounding gas, p represents the pressure of the surrounding gas, d represents the distance in centimetres between the electrodes, and γ_{se} represents the Secondary Electron Emission Coefficient.

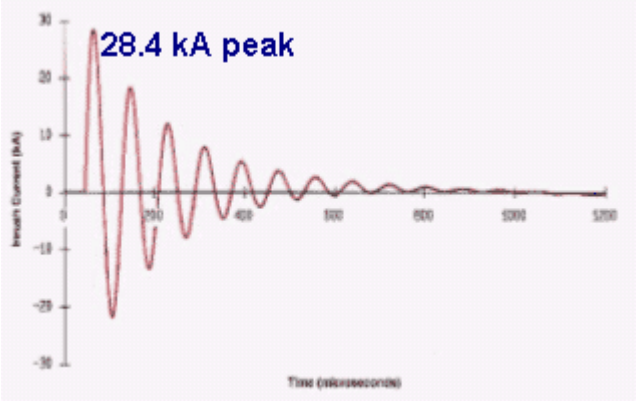
Diodes



Diode I-V diagram

Breakdown voltage is a parameter of a diode that defines the largest reverse voltage that can be applied without causing an exponential increase in the current in the diode. As long as the current is limited, exceeding the breakdown voltage of a diode does no harm to the diode. In fact, Zener diodes are essentially just heavily doped normal diodes that exploit the breakdown voltage of a diode to provide regulation of voltage levels.

Inrush Current

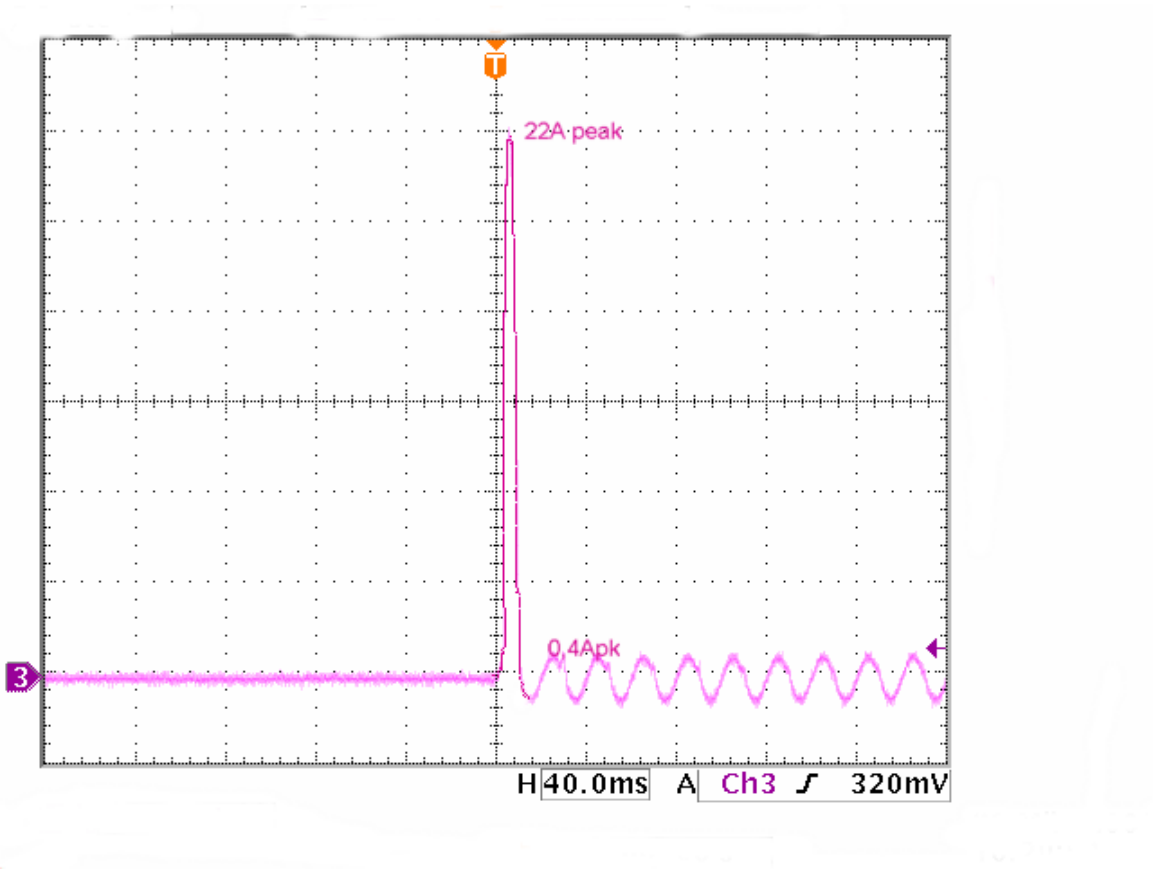


An example of inrush current transients during capacitor bank energization. The peak inrush current transient rises as high as 28.4 kA

Inrush current or **input surge current** refers to the maximum, instantaneous input current drawn by an electrical device when first turned on. For example, incandescent light bulbs have high inrush currents until their filaments warm up and their resistance increases. Alternating current electric motors and transformers may draw several times their normal full-load current when first energized, for a few cycles of the input waveform. Power converters also feature high inrush currents relative to their steady state currents. This is typically the charging current of the input capacitance. The selection of overcurrent protection devices such as fuses and circuit breakers is made more complicated when high inrush currents must be tolerated. The overcurrent protection must react quickly to overload or short circuit but must not interrupt the circuit when the (usually harmless) inrush current flows.

Transformers

When a transformer is first energized a transient current up to 10 to 50 times larger than the rated transformer current can flow for several cycles. This happens when the primary winding is connected around the zero-crossing of the primary voltage. For large transformers, inrush current can last for several seconds. Toroid transformers can have up to 80 times larger inrush, because the remnant magnetism is nearly as high as the saturation magnetism at the "knee" of the hysteresis loop. This is caused because the transformer will always have some residual flux density and when the transformer is re-energized the incoming flux will add to the already existing flux which will cause the transformer to move into saturation. Then only the resistance of the primary side windings and the power line are limiting the current.



An example of inrush current transients during a 100VA toroid transformer energization. Inrush peak raises to 50 times of nominal current.

Protection

A resistor in series with the line can be used to limit the current charging input capacitors. However, this approach is not very efficient, especially in high power devices, since the resistor will have a voltage drop and dissipate some power.

Inrush current can also be reduced by inrush current limiters. Negative temperature coefficient (NTC) thermistors are commonly used in switching power supplies, motor drives and audio equipment to prevent damage caused by inrush current. A thermistor is a thermally-sensitive resistor with a resistance that changes significantly and predictably as a result of temperature changes. The resistance of an NTC thermistor decreases as its temperature increases .

As the inrush current limiter self-heats, the current begins to flow through it and warm it. Its resistance begins to drop and a relatively small current flow charges the input capacitors. After the capacitors in the power supply become charged, the self heated inrush current limiter offers little resistance in the circuit, with a low voltage drop with respect to the total voltage drop of the circuit. A disadvantage is that immediately after the device is switched off, the NTC resistor is still hot and has a low resistance. It cannot

limit the inrush current unless it cools for more than 1 minute to get a higher resistance. Another disadvantage is that the NTC thermistor is not short circuit proof.

Another way to avoid the transformer inrush current is a "transformer switching relay". This does not need time for cool down. It can deal also with power line half-wave voltage-dips and is short circuit proof. This technique is Important for IEC 61000-4-11 tests.

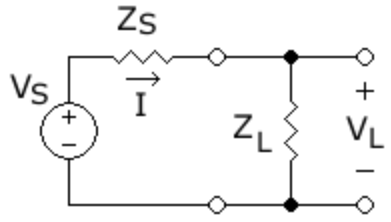
Another option, particularly for high voltage circuits, is to use a pre-charge circuit. The circuit would support a current limited precharge mode during the charging of capacitors, and then switch to an unlimited mode for normal operation when the voltage on the load is 90% of full charge.

Chapter 3

Input Impedance & Equivalent Series Resistance

Input Impedance

The **input impedance** of an electrical network is the equivalent impedance "seen" by a power source connected to that network. If the source provides known voltage and current, such impedance can be calculated using Ohm's Law. The input impedance is the Thévenin's equivalent circuit of the electrical network, modeled by an RL (resistor-inductor) or an RC (resistor-capacitor) combination, with equivalent values that would result in the same response as that of the network. It is also called Z11 in terms of Z-Parameters. Generally speaking, the exact definition depends on the particular field of study.



Simple source and load circuit (Z_S is the output impedance seen by the load, and Z_L is the input impedance seen by the source).

Termination requirements

Audio systems

Generally in audio and hi-fi systems, amplifiers have an input impedance several orders of magnitude higher than their output impedance. This concept is also called voltage bridging or impedance bridging. In this case,

Z_{load}, or the input of the driven stage >> Z_{source}, or the output of the driving stage.

In general, this configuration will be more resistant to noise (particularly power line hum). Also the loading effects on the driving amplifier stage are reduced. In certain circuits a voltage follower stage is used to match the source and load impedance, which results in maximum power transfer.

Video and high frequency (RF) systems

In RF systems, the input impedance of inputs, the characteristic impedance of the transmission line, and the load impedance all have to be equal (or "matched") to reduce signal reflections, which result in distortion and potential damage to the driving circuitry. This is known as a *matched connection*, and the process of correcting an impedance mismatch is called *impedance matching*. Typical values are 50ohm and 75ohm. In analog video circuits these reflections can cause "ghosting", where the time-delayed echo of the principle image appears as a weak and displaced image (typically to the right of the principal image). In high-speed digital systems, such as HD video, reflections result in interferences and potentially corrupt signal.

$$Z_{load} = Z_{line} = Z_{source}$$

Radio frequency power systems

In circuits carrying high power, matching the impedances is important for at least two reasons:

1. The maximum power at maximum efficiency will be transferred when the impedances are *complex conjugate matched* throughout the power chain, from the transmitter output, through the transmission line (a balanced pair, a coaxial cable, or a waveguide), to the antenna *system*, which consists of an impedance matching device and the radiating element(s). For maximum power, $Z_{load} = Z_{source}*$ (where * indicates the complex conjugate)
2. Failure to match impedances will create standing waves on the transmission line due to reflections. These will be periodic regions of higher than normal voltage. If this voltage exceeds the dielectric breakdown strength of the insulating material of the line then an arc will occur. This in turn can cause a reactive pulse of high voltage that can destroy the transmitter's final output stage. For reflectionless matching $Z_{load} = Z_{source}$ (no complex conjugate).

In the case of purely resistive impedances (no reactive components), the two types of impedance matching are identical.

Equivalent Series Resistance

Practical capacitors and inductors as used in electric circuits are not ideal components with only capacitance or inductance. However, they can be treated to a very good approximation as ideal capacitors and inductors in series with a resistance; this resistance is defined to be the **equivalent series resistance** (ESR).

Overview

Electric circuit theory deals with ideal resistors, capacitors and inductors, each assumed to contribute only resistance, capacitance or inductance to the circuit. However, all components have a non-zero value of each of these parameters. In particular, all physical devices are constructed of materials with finite electrical resistance, so that physical components have some resistance in addition to their other properties. The physical origins of ESR depend on the device in question.

One way to deal with these inherent resistances in circuit analysis is to use a lumped element model to express each physical component as a combination of an ideal component and a small resistor in series, the ESR. The ESR can be measured and included in a component's datasheet. To some extent it can be calculated from the device properties.

Q factor, which is related to ESR and is sometimes a more convenient parameter than ESR to use in calculations of high-frequency non-ideal performance of real inductors, is quoted in inductor data sheets.

Capacitors, inductors, and resistors are usually designed to minimise other parameters. In many cases this can be done to a sufficient extent that *parasitic* capacitance and inductance of a resistor, for example, are so small as not to affect circuit operation. However, under some circumstances parasitics become important and even dominant.

Component models

Actual passive two-terminal components can be represented by some network of lumped and distributed ideal inductors, capacitors, and resistors, in the sense that the real component behaves as the network does. Some of the components of the equivalent circuit can vary with conditions, e.g., frequency and temperature.

If driven by a periodic sinewave (alternating current) the component will be characterised by its (complex impedance $Z(\omega) = R + j\,X(\omega)$; the impedance can involve several minor resistances, inductances and capacitances in addition to the main property. These small deviations from the ideal behavior of the device can become significant under certain

conditions, typically high frequency, where the reactance of small capacitances and inductances can become a significant element of circuit operation. Models of lesser or greater complexity can be used, depending upon the accuracy required. For many purposes a simple model with an inductance or capacitance in series with an ESR is good enough.

These models, however simple or complex, can be inserted into a circuit to calculate performance. Computer tools are available for complex circuits; e.g., the SPICE program and its variants.

Pure capacitors and inductors do not dissipate energy; any process which dissipates energy must be treated as one or more resistors in the component model.

Inductors

Inductors have resistance inherent in the metal conductor, quoted as DCR in datasheets. This metallic resistance is small for small inductance values (typically below 1 Ω). The DC resistance is an important parameter in switch-mode power supply design. It can be modeled as a resistor in series with the inductor, therefore often leading to the DC resistance being referred to as the ESR. Though this is not precisely correct usage, the unimportant elements of ESR are often neglected in circuit discussion, since it is rare that all elements of ESR are significant to a particular application.

An inductor using a core to increase inductance will have losses such as hysteresis and eddy current losses in the core. At high frequencies there are also additional losses in the windings due to proximity and skin effect. These are in addition to wire resistance, and lead to a higher ESR.

Capacitors

In a non-electrolytic capacitor the metallic resistance of the leads and electrodes and losses in the dielectric comprise the ESR. Typically quoted values of ESR for ceramic capacitors are between 0.01 and 0.1 Ω. ESR of non-electrolytic capacitors tends to be fairly stable over time; for most purposes real non-electrolytic capacitors can be treated as ideal components.

Aluminium and tantalum electrolytic capacitors have much higher ESR values, and ESR tends to increase with frequency due to effects of the electrolyte. A very serious problem, particularly with aluminium electrolytics, is that ESR increases over time with use; ESR can increase enough to cause circuit malfunction and even component damage, although measured capacitance may remain within tolerance. While this happens with normal aging, high temperatures and large ripple current exacerbate the problem. In a circuit with significant ripple current, an increase in ESR will increase heat dissipation, much accelerating ageing.

Capacitors rated for high-temperature operation and of higher quality than basic consumer-grade parts are less susceptible to become prematurely unusable due to ESR increase. A cheap electrolytic capacitor may be rated for a life of less than 1000 hours at 85°C (a year is about 9000 hours). Higher-grade parts are typically rated at a few thousand hours at maximum rated temperature, as can be seen from manufacturers' datasheets. Electrolytics of higher capacitance have lower ESR; if ESR is critical, specification of a part of larger capacitance than is otherwise required may be advantageous.

Typical values of ESR for capacitors

Type	22 µF	100 µF
Standard aluminum	0.1 - 3.0 Ω	0.05 - 0.5 Ω
Ceramic	<0.015 Ω	

Chapter 4

Johnson–Nyquist Noise

Johnson–Nyquist noise (**thermal noise**, **Johnson noise**, or **Nyquist noise**) is the electronic noise generated by the thermal agitation of the charge carriers (usually the electrons) inside an electrical conductor at equilibrium, which happens regardless of any applied voltage.

Thermal noise is approximately white, meaning that the power spectral density is nearly equal throughout the frequency spectrum (however see the section below on extremely high frequencies). Additionally, the amplitude of the signal has very nearly a Gaussian probability density function.

History

This type of noise was first measured by John B. Johnson at Bell Labs in 1928. He described his findings to Harry Nyquist, also at Bell Labs, who was able to explain the results.

Noise voltage and power

Thermal noise is distinct from shot noise, which consists of additional current fluctuations that occur when a voltage is applied and a macroscopic current starts to flow. For the general case, the above definition applies to charge carriers in any type of conducting medium (e.g. ions in an electrolyte), not just resistors. It can be modeled by a voltage source representing the noise of the non-ideal resistor in series with an ideal noise free resistor.

The one-sided power spectral density, or voltage variance (mean square) per hertz of bandwidth, is given by

$$\bar{v_n^2} = 4k_BTR$$

where k_B is Boltzmann's constant in joules per kelvin, T is the resistor's absolute temperature in kelvins, and R is the resistor value in ohms (Ω). Use this equation for quick calculation:

$$\sqrt{\bar{v_n^2}} = 0.13\sqrt{R}\ \mathrm{nV}/\sqrt{\mathrm{Hz}}.$$

For example, a 1 kΩ resistor at a temperature of 300 K has

$$\sqrt{\bar{v_n^2}} = \sqrt{4 \cdot 1.38 \cdot 10^{-23}\ \mathrm{J/K} \cdot 300\ \mathrm{K} \cdot 1\ \mathrm{k\Omega}} = 4.07\ \mathrm{nV}/\sqrt{\mathrm{Hz}}.$$

For a given bandwidth, the root mean square (RMS) of the voltage, v_n, is given by

$$v_n = \sqrt{\bar{v_n^2}}\sqrt{\Delta f} = \sqrt{4k_B T R \Delta f}$$

where Δf is the bandwidth in hertz over which the noise is measured. For a 1 kΩ resistor at room temperature and a 10 kHz bandwidth, the RMS noise voltage is 400 nV. A useful rule of thumb to remember is that 50 Ω at 1 Hz bandwidth correspond to 1 nV noise at room temperature.

A resistor in a short circuit dissipates a noise power of

$$P = \bar{v_n^2}/R = 4k_B\, T \Delta f.$$

The noise generated at the resistor can transfer to the remaining circuit; the maximum noise power transfer happens with impedance matching when the Thévenin equivalent resistance of the remaining circuit is equal to the noise generating resistance. In this case each one of the two participating resistors dissipates noise in both itself and in the other resistor. Since only half of the source voltage drops across any one of these resistors, the resulting noise power is given by

$$P = k_B\, T \Delta f$$

where P is the thermal noise power in watts. Notice that this is independent of the noise generating resistance.

Noise current

The noise source can also be modeled by a current source in parallel with the resistor by taking the Norton equivalent that corresponds simply to divide by R. This gives the root mean square value of the current source as:

$$i_n = \sqrt{\frac{4k_B T \Delta f}{R}}.$$

Thermal noise is intrinsic to all resistors and is not a sign of poor design or manufacture, although resistors may also have excess noise.

Noise power in decibels

Signal power is often measured in dBm (decibels relative to 1 milliwatt). From the equation above, noise power in a resistor at room temperature, in dBm, is then:

$$P_{\mathrm{dBm}} = 10\ \log_{10}(k_B T \Delta f\ \times 1000)$$

where the factor of 1000 is present because the power is given in milliwatts, rather than watts. This equation can be simplified by separating the constant parts from the bandwidth:

$$P_{\mathrm{dBm}} = 10\ \log_{10}(k_B T \times 1000) + 10\ \log(\Delta f)$$

which is more commonly seen approximated as:

$$P_{\mathrm{dBm}} = -174 + 10\ \log_{10}(\Delta f)$$

Noise power at different bandwidths is then simple to calculate:

Bandwidth (Δ*f*)	**Thermal noise power**	**Notes**
1 Hz	−174 dBm	
10 Hz	−164 dBm	
100 Hz	−154 dBm	
1 kHz	−144 dBm	
10 kHz	−134 dBm	FM channel of 2-way radio
100 kHz	−124 dBm	
180 kHz	−121.45 dBm	One LTE resource block
200 kHz	−120.98 dBm	One GSM channel (ARFCN)
1 MHz	−114 dBm	
2 MHz	−111 dBm	Commercial GPS channel
6 MHz	−106 dBm	Analog television channel
20 MHz	−101 dBm	WLAN 802.11 channel

Thermal noise on capacitors

Thermal noise on capacitors is referred to as *kTC* noise. Thermal noise in an RC circuit has an unusually simple expression, as the value of the resistance (*R*) drops out of the equation. This is because higher *R* contributes to more filtering as well as to more noise. The noise bandwidth of the RC circuit is 1/(4*RC*), which can substituted into the above formula to eliminate *R*. The mean-square and RMS noise voltage generated in such a filter are:

$$\bar{v_n^2} = k_B T/C$$
$$v_n = \sqrt{k_B T/C}.$$

Thermal noise accounts for 100% of *kTC* noise, whether it is attributed to the resistance or to the capacitance.

In the extreme case of the *reset noise* left on a capacitor by opening an ideal switch, the resistance is infinite, yet the formula still applies; however, now the RMS must be interpreted not as a time average, but as an average over many such reset events, since the voltage is constant when the bandwidth is zero. In this sense, the Johnson noise of an RC circuit can be seen to be inherent, an effect of the thermodynamic distribution of the number of electrons on the capacitor, even without the involvement of a resistor.

The noise is not caused by the capacitor itself, but by the thermodynamic equilibrium of the amount of charge on the capacitor. Once the capacitor is disconnected from a conducting circuit, the thermodynamic fluctuation is *frozen* at a random value with standard deviation as given above.

The reset noise of capacitive sensors is often a limiting noise source, for example in image sensors. As an alternative to the voltage noise, the reset noise on the capacitor can also be quantified as the electrical charge standard deviation, as

$$Q_n = \sqrt{k_B T C}.$$

Since the charge variance is $k_B TC$, this noise is often called *kTC noise*.

Any system in thermal equilibrium has state variables with a mean energy of *kT*/2 per degree of freedom. Using the formula for energy on a capacitor ($E = \frac{1}{2}CV^2$), mean noise energy on a capacitor can be seen to also be $\frac{1}{2}C(kT/C)$, or also *kT*/2. Thermal noise on a capacitor can be derived from this relationship, without consideration of resistance.

The *kTC* noise is the dominant noise source at small capacitors.

Noise of capacitors at 300 K

Capacitance	$\sqrt{k_B T/C}$	Electrons
1 fF	2 mV	12.5 e^-
10 fF	640 μV	40 e^-
100 fF	200 μV	125 e^-
1 pF	64 μV	400 e^-
10 pF	20 μV	1250 e^-
100 pF	6.4 μV	4000 e^-
1 nF	2 μV	12500 e^-

Noise at very high frequencies

The above equations are good approximations at any practical radio frequency in use (i.e. frequencies below about 80 gigahertz). In the most general case, which includes up to optical frequencies, the power spectral density of the voltage across the resistor *R*, in V^2/Hz is given by:

$$\Phi(f) = \frac{2Rhf}{e^{\frac{hf}{k_B T}} - 1}$$

where *f* is the frequency, *h* Planck's constant, k_B Boltzmann constant and *T* the temperature in kelvins. If the frequency is low enough, that means:

$$f \ll \frac{k_B T}{h}$$

(this assumption is valid until few terahertz at room temperature) then the exponential can be expressed in terms of its Taylor series. The relationship then becomes:

$$\Phi(f) \approx 2Rk_B T.$$

In general, both *R* and *T* depend on frequency. In order to know the total noise it is enough to integrate over all the bandwidth. Since the signal is real, it is possible to integrate over only the positive frequencies, then multiply by 2. Assuming that *R* and *T* are constants over all the bandwidth Δ*f*, then the root mean square (RMS) value of the voltage across a resistor due to thermal noise is given by

$$v_n = \sqrt{4k_B TR\Delta f},$$

that is, the same formula as above.

Chapter 5

Nominal Impedance & Overdrive Voltage

Nominal Impedance

Nominal impedance in electrical engineering and audio engineering refers to the approximate designed impedance of an electrical circuit or device. The term is applied in a number of different fields, most often being encountered in respect of:

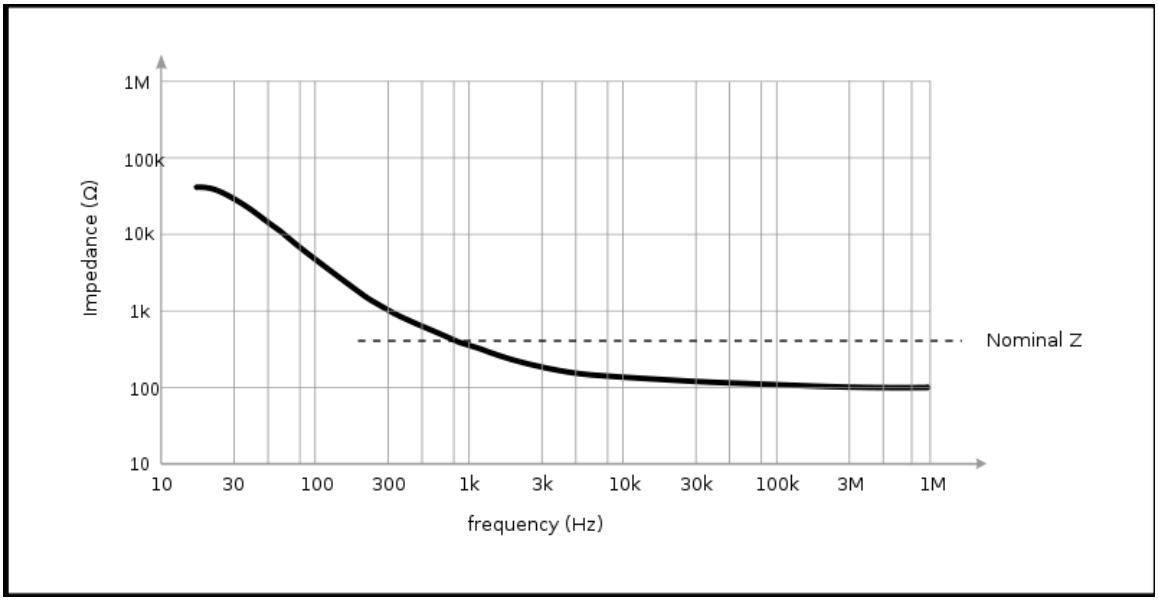
- The nominal value of the characteristic impedance of a cable of other form of transmission line.
- The nominal value of the input, output or image impedance of a port of a network, especially a network intended for use with a transmission line, such as filters, equalisers and amplifiers.
- The nominal value of the input impedance of a radio frequency antenna

The actual impedance may vary quite considerably from the nominal figure with changes in frequency. In the case of cables, there is also variation along the length of the cable. It is usual practice to speak of nominal impedance as if it were a constant resistance, that is, it is invariant with frequency and has a zero reactive component, despite this often being far from the case. Depending on the field of application, nominal impedance is implicitly referring to a specific point on the frequency response of the circuit under consideration. This may be at low-frequency, mid-band or some other point and specific applications are discussed in the sections below.

In most applications, there are a number of values of nominal impedance that are recognised as being standard. The nominal impedance of components and circuits are often assigned one of these standard values, regardless of whether the measured impedance exactly corresponds to it. The item is assigned the nearest standard value.

600 Ω

Nominal impedance first started to be specified in the early days of telecommunications. At first amplifiers were not available and when they did become available they were expensive. It was consequently necessary to squeeze every last drop of transmitted power from the cable at the receiving end in order to maximise the lengths of cables that could be installed. It also became apparent that reflections on the transmission line would severely limit the bandwidth that could be used or the distance that it was practicable to transmit. Matching equipment impedance to the characteristic impedance of the cable reduces reflections (and they are eliminated altogether if the match is perfect) and power transfer is maximised. To this end, all cables and equipment started to be specified to a standard nominal impedance. The earliest, and still the most widespread, standard is 600 Ω, originally used for telephony. It has to be said that the choice of this figure had more to do with the way telephones were interfaced into the local exchange than any characteristic of the local telephone cable. Telephones (old style analogue telephones) connect to the exchange through twisted pair cabling. Each leg of the pair is connected to a relay coil which detect the signalling on the line (dialling, handset off-hook etc). The other end of one coil is connected to volts and the second coil is connected to ground. A telephone exchange relay coil is around 300 Ω so the two of them together are terminating the line in 600 Ω.



Variation of characteristic impedance with frequency. At audio frequencies the impedance is far from constant and the nominal value is only correct at one frequency.

The wiring to the subscriber in telephone networks is generally done in twisted pair cable. This format at audio frequencies, and especially at the more restricted telephone band frequencies, is far from constant. It is possible to manufacture this kind of cable to have a 600 Ω characteristic impedance but it will only be this value at one specific

frequency. This might be quoted as a nominal 600 Ω impedance at 800 Hz or 1 kHz. Below this frequency the characteristic impedance rapidly rises and become more and more dominated by the ohmic resistance of the cable as the frequency falls. At the bottom of the audio band the impedance can be several tens of kilohms. On the other hand, at high frequency in the MHz region, the characteristic impedance flattens out to something almost constant. The reason for this response is explained at primary line constants.

Local area networks (LANs) commonly use a similar kind of twisted pair cable, but screened and manufactured to tighter tolerances than is necessary for telephony. Even though it has a very similar impedance to telephone cable, the nominal impedance is rated at 100 Ω. This is because the LAN data is in a higher frequency band where the characteristic impedance is substantially flat and mostly resistive.

Standardisation of line nominal impedance led to two-port networks such as filters being designed to a matching nominal impedance. The nominal impedance of low-pass symmetrical T- or Pi-filter sections (or more generally, image filter sections) is defined as the limit of the filter image impedance as the frequency approaches zero and is given by,

$$Z_{\mathrm{nom}} = \sqrt{\frac{L}{C}}$$

where *L* and *C* are as defined in constant k filter. As can be seen from the expression, this impedance is purley resistive. This filter transformed to a band-pass filter will have an impedance equal to the nominal impedance at resonance rather than low frequency. This nominal impedance of filters will generally be the same as the nominal impedance of the circuit or cable that the filter is working into.

While 600 Ω is an almost universal standard in telephony for local presentation at customer's premises from the exchange, for long distance transmission on trunk lines between exchanges other standard nominal impedances are used and are usually lower, such as 150 Ω.

50 Ω and 75 Ω

In the field of radio frequency (RF) and microwave engineering, by far and away the most common transmission line standard is 50 Ω coaxial cable (coax), which is an unbalanced line. 50 Ω first arose as a nominal impedance during world war two work on radar and is a compromise between two requirements. This standard was the work of the wartime US joint Army-Navy RF Cable Coordinating Committee. The first requirement is for minimum loss. The loss of coaxial cable is given by,

$$\alpha \approx \frac{R}{2Z_0} \text{nepers/metre}$$

where R is the loop resistance per metre and Z_0 is the characteristic impedance. Making the diameter of the inner conductor larger will decrease R and decreasing R decreases the loss. On the other hand, Z_0 depends on the ratio of the diameters of outer and inner conductors (D_r) and will decrease with increasing inner conductor diameter thus increasing the loss. There is a specific value of D_r for which the loss is a minimum and this turns out to be 3.6. For an air dielectric coax (wartime coax was rigid air insulated pipe and this remained the case for some time afterwards) this corresponds to a characteristic impedance of 77 Ω. The second requirement is for maximum power handling and was an important requirement for radar. This is not the same condition as minimum loss because power handling is usually limited by the breakdown voltage of the dielectric. However, there is a similar compromise in terms of the ratio of conductor diameters. Making the inner conductor too large results in a thin insulator which breaks down at a lower voltage. On the other hand, making the inner conductor too small results in higher electric field strength near the inner conductor (because the field lines are closer together on the smaller circumference) and again reduces the breakdown voltage. The ideal ratio, D_r, for maximum power handling turns out to be 1.65 and corresponds to a characteristic impedance of 30 Ω in air. 50 Ω was arrived at by taking the geometric mean of these two figures;

$$50 \approx \sqrt{30 \times 77}\ \Omega$$

and then rounding to a convenient whole number.

Wartime production of coax, and for a period afterwards, tended to use standard plumbing pipe sizes for the outer conductor and standard AWG sizes for the inner conductor. This resulted in coax that was nearly, but not quite, 50 Ω. Matching is a much more critical requirement at RF than it is at voice frequencies, so when cable started to become available that was truly 50 Ω a need arose for matching circuits to interface between the new cables and legacy equipment, such as the rather strange 51.5 Ω to 50 Ω matching network.

While 30 Ω cable is highly desirable for its power handling capabilities, it has never been in commercial production because the large size of inner conductor makes it difficult to manufacture. This is not the case with 77 Ω cable. 75 Ω nominal impedance cable has been in use from an early period in telecommunications for its low loss characteristic. According to Stephen Lampen of Belden Wire & Cable 75 Ω was chosen as the nominal impedance rather than 77 Ω because it corresponded to a standard AWG wire size for the inner conductor. 75 Ω is now the near universal standard nominal impedance for coaxial video interfaces and transmission lines.

Radio antennae

The widespread idea that 50 Ω and 75 Ω nominal impedances are connected with the input impedance of various antennae is, in fact, a myth. It is true, however, that several common antennae are easily matched to these cables. A quarter wavelength monopole has an impedance of 36.5 Ω, and a half wavelength dipole has an impedance of 72 Ω. A

half-wavelength folded dipole, commonly seen on television antennae, on the other hand, has an impedance four times that of a dipole, that is 288 Ω. The 0.5λ dipole and the 0.5λ folded dipole are commonly taken as having nominal impedances of 75 Ω and 300 Ω respectively.

Cable quality

One measure of cable manufacturing and installation quality is how closely the characteristic impedance adheres to the nominal impedance along its length. Impedance changes can be caused by variations in geometry along the cable length. In turn, these can be caused by a faulty manufacturing process or by faulty installation (such as not observing limits on bend radii). Unfortunately, there is no easy, non-destructive method of directly measuring impedance along a cables' length. It can, however, be indicated indirectly by measuring reflections, that is, return loss. Return loss by itself does not reveal much, since the cable design will have some intrinsic return loss anyway due to not having a purely resistive characteristic impedance. The technique used is to carefully adjust the cable termination to obtain as close a match as possible and then to measure the variation of return loss with frequency. The minimum return loss so measured is called the structural return loss (SRL). SRL is a measure of a cables' adherence to its nominal impedance but it is not a direct correspondence, errors further from the generator have less effect on SRL than those close to it. The measurement must also be carried out at all in-band frequencies to be significant. The reason for this is that equally spaced errors introduced by the manufacturing process will cancel and be invisible, or at least much reduced, at certain frequencies due to quarter wave impedance transformer action.

Audio systems

For the most part, audio systems both professional and domestic, have their components interconnected with low impedance outputs connected to high impedance inputs. These impedances are poorly defined and nominal impedances are not usually assigned for this kind of connection. The exact impedances make little difference to performance as long as the latter is many times larger than the former. This is a common interconnection scheme, not just for audio, but for electronic units in general which form part of a larger equipment or are only connected over a short distance. Where audio needs to be transmitted over large distances, which is often the case in broadcast engineering, considerations of matching and reflections dictate that a telecommunications standard is used, which would normally mean using 600 Ω nominal impedance (although other standards are sometimes encountered, such as sending at 75 Ω and receiving at 600 Ω which has bandwidth advantages). The nominal impedance of the transmission line and of the amplifiers and equalisers in the transmission chain will all be the same value.

Nominal impedance *is* used, however, to characterise the transducers of an audio system, such as its microphones and loudspeakers. It is important that these are connected to a circuit capable of dealing with impedances in the appropriate range and assigning a nominal impedance is a convenient way of quickly determining likely incompatibilities. Loudspeakers and microphones are dealt with in separate sections below.

Loudspeakers

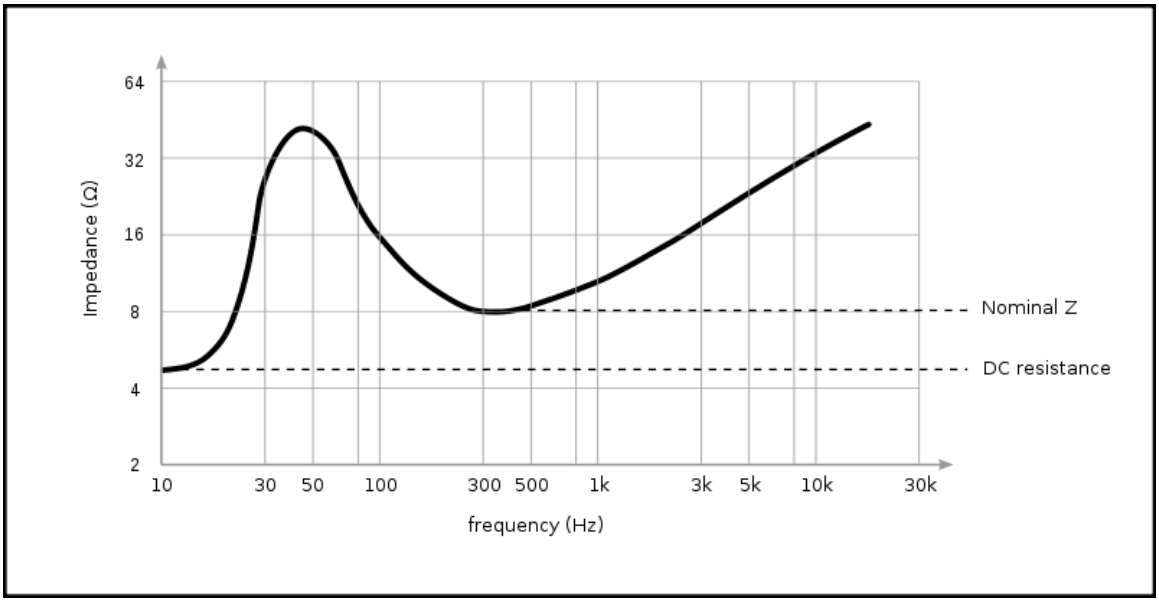


Diagram showing the variation in impedance of a typical mid-range loudspeaker. Nominal impedance is usually determined at the lowest point after resonance. However, it is possible for the low-frequency impedance to be still lower than this.

Loudspeaker impedances are kept relatively low compared with other audio components so that the required audio power can be transmitted without using inconveniently (and dangerously) high voltages. The most common nominal impedance for loudspeakers is 8 Ω. Also used are 4 Ω and 16 Ω. The once common 16 Ω is now mostly reserved for high frequency compression drivers since the high frequency end of the audio spectrum does not usually require so much power to reproduce.

The impedance of a loudspeaker is not constant across all frequencies. In a typical loudspeaker the impedance will rise with increasing frequency from its dc value (see diagram) until it reaches a point of mechanical resonance. Following resonance, the impedance falls to a minimum and then begins to rise again. Speakers are usually designed to operate at frequencies above their resonance, and for this reason it is the usual practice to define nominal impedance at this minimum and then round to the nearest standard value. The ratio of the peak resonant frequency to the nominal impedance can be as much as 4:1. It is, however, still perfectly possible for the low frequency impedance to actually be lower than the nominal impedance. A given audio amplifier may not be capable of driving this low frequency impedance even though it is capable of driving the nominal impedance, a problem that can be solved either with the use of crossover filters or underrating the amplifier supplied.

In the days of valves or vacuum tubes, most loudspeakers had a nominal impedance of 16 Ω. Valve outputs require an output transformer to match the very high output impedance and voltage of the output valves to this lower impedance. These transformers were

commonly tapped to allow matching of the output to a multiple loudspeaker setup. For example, two 16 Ω loudspeakers in parallel will give an impedance of 8 Ω. Since the advent of solid-state transformerless outputs, these multiple-impedance outputs have become rare, and lower impedance loudspeakers more common. The most common nominal impedance for a single loudspeaker is now 8 Ω. Most solid-state amplifiers are designed to work with loudspeaker combinations of anything from 4 Ω to 8 Ω.

Microphones

There are a large number of different types of microphone and there are correspondingly large differences in impedance between them. They range from the very low impedance of ribbon microphones (can be less than one ohm) to the very large impedance of piezoelectric microphones which are measured in megohms. The Electronic Industries Alliance (EIA) has defined a number of standard microphone nominal impedances to aid categorisation of microphones.

Range (Ω)	EIA nominal impedance (Ω)
20-80	38
80-300	150
300-1250	600
1250-4500	2400
4500-20,000	9600
20,000-70,000	40,000

The International Electrotechnical Commission defines a similar set of nominal impedances, but also has a coarser classification of low, medium and high impedance with medium being 600 Ω to 10 kΩ.

Oscilloscopes

Oscilloscope inputs are usually high impedance so that they only minimally affect the circuit being measured when connected. However, the input impedance is made a specific nominal value, rather than arbitrarily high, because of the common use of X10 probes. A common value for oscilloscope nominal impedance is 1 MΩ resistance and 20 pF capacitance. With a known input impedance to the oscilloscope, the probe designer can ensure that the probe input impedance is exactly ten times this figure (actually oscilloscope plus probe cable impedance). Since the impedance included the input capacitance and the probe is an impedance divider circuit, the result is that the waveform being measured is not distorted by the RC circuit formed by the probe resistance and the capacitance of the input (or the cable capacitance which is generally higher).

Overdrive Voltage

Overdrive voltage, usually abbreviated as V_{OV}, is typically referred to in the context of MOSFET transistors. The overdrive voltage is defined as the voltage between transistor gate and source (V_{GS}) in excess of the threshold voltage (V_t) where V_t is defined as the minimum voltage required between gate and source to turn the transistor on (allow it to conduct electricity). Due to this definition, overdrive voltage is also known as "excess gate voltage" or "effective voltage." Overdrive voltage can be found using the simple equation: $V_{OV} = V_{GS} - V_t$.

Technology

V_{OV} is important as it directly affects the output current (I_D) of the transistor, an important property of amplifier circuits. By increasing V_{OV}, I_D can be increased until saturation is reached.

Overdrive voltage is also important because of its relationship to V_{DS}, the drain voltage relative to the source, which can be used to determine the region of operation of the MOSFET. The table below shows how to use overdrive voltage to understand what region of operation the MOSFET is in:

Conditions	Region of Operation	Description
$V_{DS} > V_{OV}$; $V_{GS} > V_t$	Saturation (CCR)	The MOSFET is delivering a high amount of current, and changing V_{DS} won't do much.
$V_{DS} < V_{OV}$; $V_{GS} > V_t$	Triode (Linear)	The MOSFET is delivering current in a linear relationship to the voltage (V_{DS}).
$V_{GS} < V_t$	Cutoff	The MOSFET is turned off, and should not be delivering any current.

A more physics-related explanation follows:

In an NMOS transistor, the channel region under zero bias has an abundance of holes (ie, it is p-type silicon). By applying a negative gate bias ($V_{GS} < 0$) we attract MORE holes, and this is called accumulation. A positive gate voltage ($V_{GS} > 0$) will attract electrons and repel holes, and this is called depletion because we are depleting the number of holes. At a critical voltage called the THRESHOLD VOLTAGE (V_t or V_{th}) the channel will actually be so depleted of holes and rich in electrons that it will INVERT to being n-type silicon, and this is called the inversion region.

As we increase this voltage, V_{GS}, beyond V_{th}, we are said to be then OVERDRIVING the gate by creating a stronger channel, hence the OVERDRIVE VOLTAGE (Called often V_{ov}, V_{od}, or V_{on}) is defined as ($V_{GS} - V_{th}$)

Chapter 6

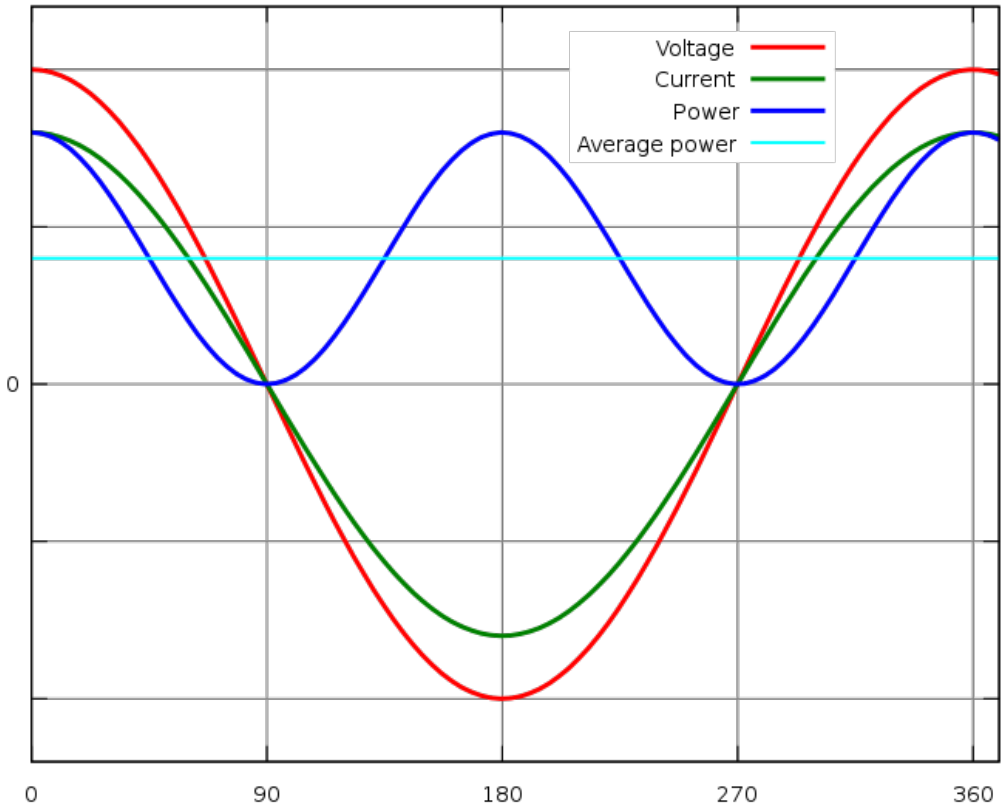
Power Factor

The **power factor** of an AC electric power system is defined as the ratio of the real power flowing to the load to the apparent power in the circuit, and is a dimensionless number between 0 and 1 (frequently expressed as a percentage, e.g. 0.5 pf = 50% pf). Real power is the capacity of the circuit for performing work in a particular time. Apparent power is the product of the current and voltage of the circuit. Due to energy stored in the load and returned to the source, or due to a non-linear load that distorts the wave shape of the current drawn from the source, the apparent power will be greater than the real power.

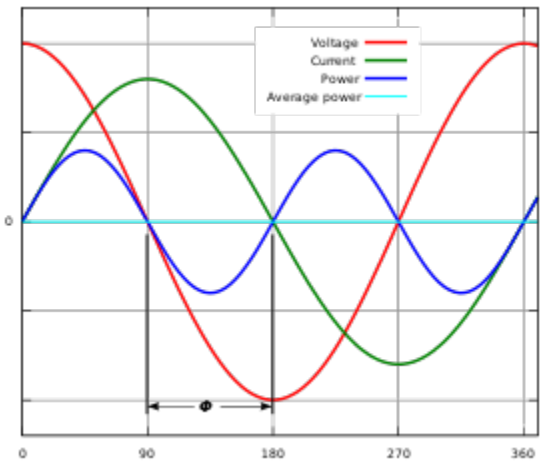
In an electric power system, a load with a low power factor draws more current than a load with a high power factor for the same amount of useful power transferred. The higher currents increase the energy lost in the distribution system, and require larger wires and other equipment. Because of the costs of larger equipment and wasted energy, electrical utilities will usually charge a higher cost to industrial or commercial customers where there is a low power factor.

Linear loads with low power factor (such as induction motors) can be corrected with a passive network of capacitors or inductors. Non-linear loads, such as rectifiers, distort the current drawn from the system. In such cases, active or passive power factor correction may be used to counteract the distortion and raise the power factor. The devices for correction of the power factor may be at a central substation, spread out over a distribution system, or built into power-consuming equipment.

Power factor in linear circuits

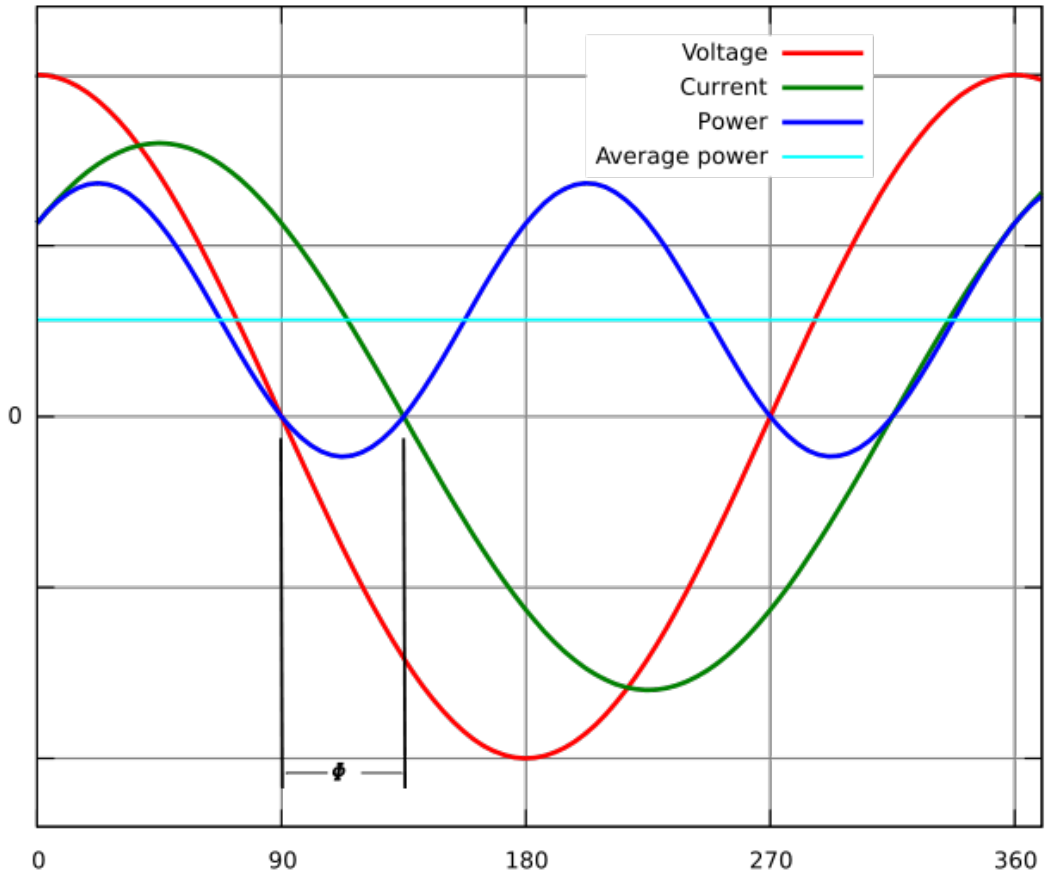


Instantaneous and average power calculated from AC voltage and current with a unity power factor (ϕ=0, cosϕ=1). Since the blue line is above the axis, all power is real power consumed by the load.



Instantaneous and average power calculated from AC voltage and current with a zero power factor (ϕ=90, cosϕ=0). The blue line shows all the power is stored temporarily in

the load during the first quarter cycle and returned to the grid during the second quarter cycle, so no real power is consumed.



Instantaneous and average power calculated from AC voltage and current with a lagging power factor (ϕ=45, cosϕ=0.71). The blue line shows some of the power is returned to the grid during the part of the cycle labelled φ

In a purely resistive AC circuit, voltage and current waveforms are in step (or in phase), changing polarity at the same instant in each cycle. All the power entering the loads is consumed. Where reactive loads are present, such as with capacitors or inductors, energy storage in the loads result in a time difference between the current and voltage waveforms. During each cycle of the AC voltage, extra energy, in addition to any energy consumed in the load, is temporarily stored in the load in electric or magnetic fields, and then returned to the power grid a fraction of a second later in the cycle. The "ebb and flow" of this nonproductive power increases the current in the line. Thus, a circuit with a low power factor will use higher currents to transfer a given quantity of real power than a circuit with a high power factor. A linear load does not change the shape of the waveform of the current, but may change the relative timing (phase) between voltage and current.

Circuits containing purely resistive heating elements (filament lamps, strip heaters, cooking stoves, etc.) have a power factor of 1.0. Circuits containing inductive or capacitive elements (electric motors, solenoid valves, lamp ballasts, and others) often have a power factor below 1.0.

Definition and calculation

AC power flow has the three components: real power (also known as active power) (P), measured in watts (W); apparent power (S), measured in volt-amperes (VA); and reactive power (Q), measured in reactive volt-amperes (var).

The power factor is defined as:

$$\frac{P}{S}.$$

In the case of a perfectly sinusoidal waveform, P, Q and S can be expressed as vectors that form a vector triangle such that:

$$S^2 = P^2 + Q^2.$$

If φ is the phase angle between the current and voltage, then the power factor is equal to the cosine of the angle, $|\cos \varphi|$, and:

$$|P| = |S| \, |\cos \varphi| \, .$$

Since the units are consistent, the power factor is by definition a dimensionless number between 0 and 1. When power factor is equal to 0, the energy flow is entirely reactive, and stored energy in the load returns to the source on each cycle. When the power factor is 1, all the energy supplied by the source is consumed by the load. Power factors are usually stated as "leading" or "lagging" to show the sign of the phase angle.

If a purely resistive load is connected to a power supply, current and voltage will change polarity in step, the power factor will be unity (1), and the electrical energy flows in a single direction across the network in each cycle. Inductive loads such as transformers and motors (any type of wound coil) consume reactive power with current waveform lagging the voltage. Capacitive loads such as capacitor banks or buried cable generate reactive power with current phase leading the voltage. Both types of loads will absorb energy during part of the AC cycle, which is stored in the device's magnetic or electric field, only to return this energy back to the source during the rest of the cycle.

For example, to get 1 kW of real power, if the power factor is unity, 1 kVA of apparent power needs to be transferred (1 kW ÷ 1 = 1 kVA). At low values of power factor, more apparent power needs to be transferred to get the same real power. To get 1 kW of real power at 0.2 power factor, 5 kVA of apparent power needs to be transferred (1 kW ÷ 0.2

= 5 kVA). This apparent power must be produced and transmitted to the load in the conventional fashion, and is subject to the usual distributed losses in the production and transmission processes.

Electrical loads consuming alternating current power consume both real power and reactive power. The vector sum of real and reactive power is the apparent power. The presence of reactive power causes the real power to be less than the apparent power, and so, the electric load has a power factor of less than 1.

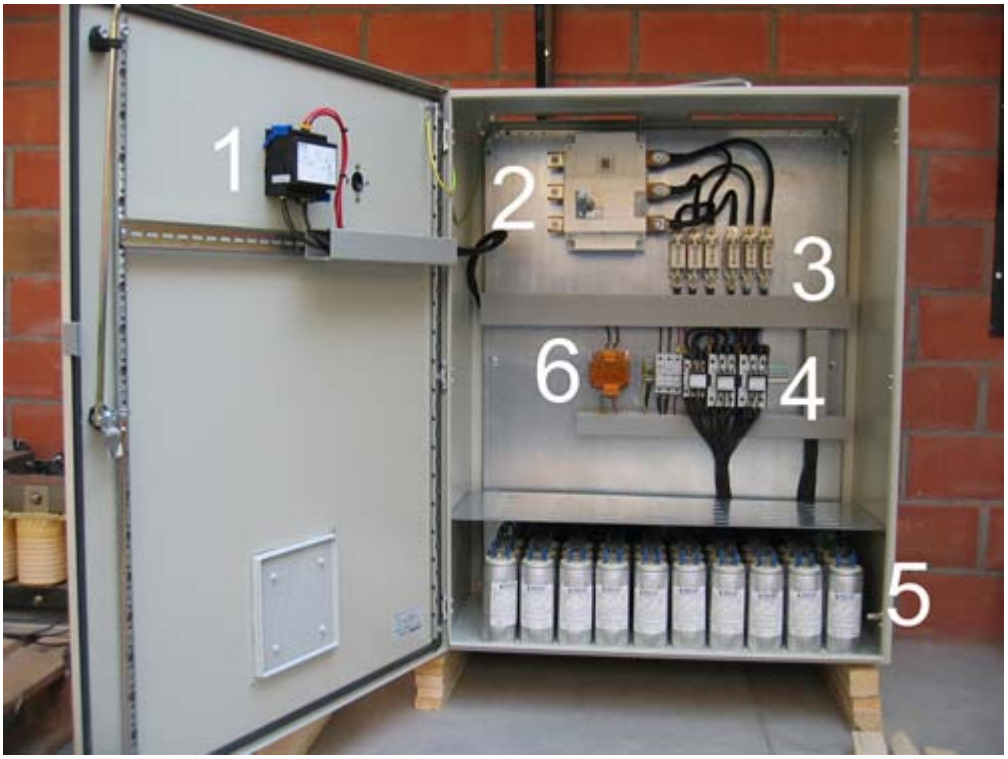
Power factor correction of linear loads

It is often desirable to adjust the power factor of a system to near 1.0. This *power factor correction* (PFC) is achieved by switching in or out banks of inductors or capacitors. For example the inductive effect of motor loads may be offset by locally connected capacitors. When reactive elements supply or absorb reactive power near the load, the apparent power is reduced.

Power factor correction may be applied by an electrical power transmission utility to improve the stability and efficiency of the transmission network. Correction equipment may be installed by individual electrical customers to reduce the costs charged to them by their electricity supplier. A high power factor is generally desirable in a transmission system to reduce transmission losses and improve voltage regulation at the load.

Power factor correction brings the power factor of an AC power circuit closer to 1 by supplying reactive power of opposite sign, adding capacitors or inductors which act to cancel the inductive or capacitive effects of the load, respectively. For example, the inductive effect of motor loads may be offset by locally connected capacitors. If a load had a capacitive value, inductors (also known as *reactors* in this context) are connected to correct the power factor. In the electricity industry, inductors are said to *consume* reactive power and capacitors are said to *supply* it, even though the reactive power is actually just moving back and forth on each AC cycle.

The reactive elements can create voltage fluctuations and harmonic noise when switched on or off. They will supply or sink reactive power regardless of whether there is a corresponding load operating nearby, increasing the system's no-load losses. In a worst case, reactive elements can interact with the system and with each other to create resonant conditions, resulting in system instability and severe overvoltage fluctuations. As such, reactive elements cannot simply be applied at will, and power factor correction is normally subject to engineering analysis.



1. Reactive Power Control Relay; 2. Network connection points; 3. Slow-blow Fuses; 4. Inrush Limiting Contactors; 5. Capacitors (single-phase or three-phase units, delta-connection); 6. Transformer Suitable voltage transformation to suit control power (contactors, ventilation,...)

An **automatic power factor correction unit** is used to improve power factor. A power factor correction unit usually consists of a number of capacitors that are switched by means of contactors. These contactors are controlled by a regulator that measures power factor in an electrical network. To be able to measure power factor, the regulator uses a current transformer to measure the current in one phase.

Depending on the load and power factor of the network, the power factor controller will switch the necessary blocks of capacitors in steps to make sure the power factor stays above a selected value (usually demanded by the energy supplier), say 0.9.

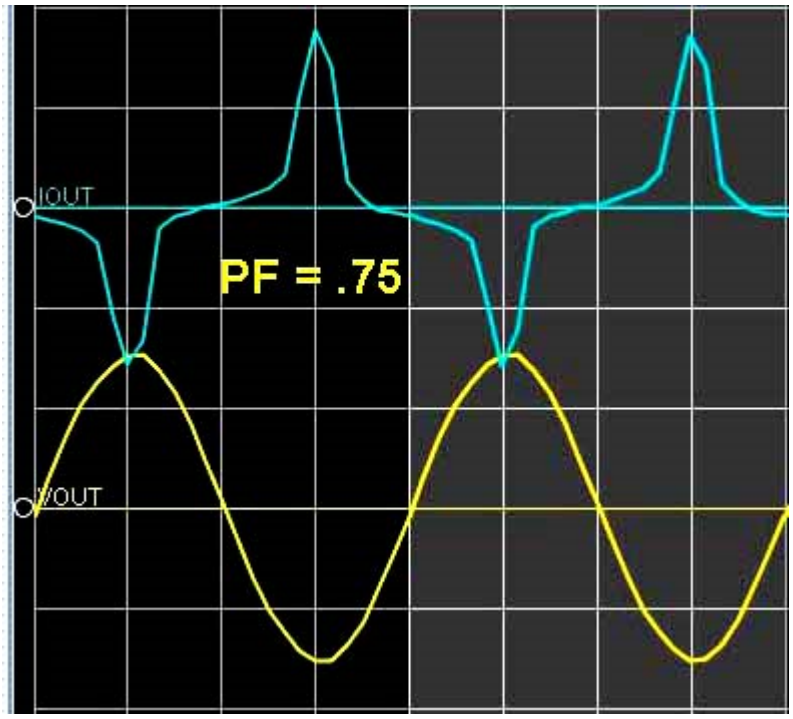
Instead of using a set of switched capacitors, an unloaded synchronous motor can supply reactive power. The reactive power drawn by the synchronous motor is a function of its field excitation. This is referred to as a *synchronous condenser*. It is started and connected to the electrical network. It operates at a leading power factor and puts vars onto the network as required to support a system's voltage or to maintain the system power factor at a specified level.

The condenser's installation and operation are identical to large electric motors. Its principal advantage is the ease with which the amount of correction can be adjusted; it behaves like an electrically variable capacitor. Unlike capacitors, the amount of reactive

power supplied is proportional to voltage, not the square of voltage; this improves voltage stability on large networks. Synchronous condensors are often used in connection with high voltage direct current transmission projects or in large industrial plants such as steel mills.

Non-linear loads

A non-linear load on a power system is typically a rectifier (such as used in a power supply), or some kind of arc discharge device such as a fluorescent lamp, electric welding machine, or arc furnace. Because current in these systems is interrupted by a switching action, the current contains frequency components that are multiples of the power system frequency. Distortion power factor is a measure of how much the harmonic distortion of a load current decreases the average power transferred to the load.



Sinusoidal voltage and non-sinusoidal current give a distortion power factor of 0.75 for this computer power supply load.

Non-sinusoidal components

Non-linear loads change the shape of the current waveform from a sine wave to some other form. Non-linear loads create harmonic currents in addition to the original (fundamental frequency) AC current. Filters consisting of linear capacitors and inductors can prevent harmonic currents from entering the supplying system.

In linear circuits having only sinusoidal currents and voltages of one frequency, the power factor arises only from the difference in phase between the current and voltage.

This is "displacement power factor". The concept can be generalized to a total, distortion, or true power factor where the apparent power includes all harmonic components. This is of importance in practical power systems which contain non-linear loads such as rectifiers, some forms of electric lighting, electric arc furnaces, welding equipment, switched-mode power supplies and other devices.

A typical multimeter will give incorrect results when attempting to measure the AC current drawn by a non-sinusoidal load; the instruments sense the average value of a rectified waveform. The average response is then calibrated to the effective, RMS value. An RMS sensing multimeter must be used to measure the actual RMS currents and voltages (and therefore apparent power). To measure the real power or reactive power, a wattmeter designed to work properly with non-sinusoidal currents must be used.

Distortion power factor

The *distortion power factor' describes how the harmonic distortion of a load current decreases the average power transferred to the load.*

$$\text{distortion power factor} = \frac{1}{\sqrt{1 + \text{THD}_i^2}} = \frac{I_{1,\text{rms}}}{I_{\text{rms}}}$$

THD_i is the total harmonic distortion of the load current. This definition assumes that the voltage stays undistorted (sinusoidal, without harmonics). This simplification is often a good approximation in practice. $I_{1,\text{rms}}$ is the fundamental component of the current and I_{rms} is the total current - both are root mean square-values.

The result when multiplied with the displacement power factor (DPF) is the overall, true power factor or just power factor (PF):

$$\text{PF} = \text{DPF}\frac{I_{1,\text{rms}}}{I_{\text{rms}}}$$

Switched-mode power supplies

A particularly important class of non-linear loads is the millions of personal computers that typically incorporate switched-mode power supplies (SMPS) with rated output power ranging from a few watts to more than 1 kW. Historically, these very-low-cost power supplies incorporated a simple full-wave rectifier that conducted only when the mains instantaneous voltage exceeded the voltage on the input capacitors. This leads to very high ratios of peak-to-average input current, which also lead to a low distortion power factor and potentially serious phase and neutral loading concerns.

A typical switched-mode power supply first makes a DC bus, using a bridge rectifier or similar circuit. The output voltage is then derived from this DC bus. The problem with

this is that the rectifier is a non-linear device, so the input current is highly non-linear. That means that the input current has energy at harmonics of the frequency of the voltage.

This presents a particular problem for the power companies, because they cannot compensate for the harmonic current by adding simple capacitors or inductors, as they could for the reactive power drawn by a linear load. Many jurisdictions are beginning to legally require power factor correction for all power supplies above a certain power level.

Regulatory agencies such as the EU have set harmonic limits as a method of improving power factor. Declining component cost has hastened implementation of two different methods. To comply with current EU standard EN61000-3-2, all switched-mode power supplies with output power more than 75 W must include passive PFC, at least. 80 PLUS power supply certification requires a power factor of 0.9 or more.

Power factor correction in non-linear loads

Passive PFC

The simplest way to control the harmonic current is to use a filter: it is possible to design a filter that passes current only at line frequency (e.g. 50 or 60 Hz). This filter reduces the harmonic current, which means that the non-linear device now looks like a linear load. At this point the power factor can be brought to near unity, using capacitors or inductors as required. This filter requires large-value high-current inductors, however, which are bulky and expensive.

A passive PFC requires an inductor larger than the inductor in an active PFC, but costs less.

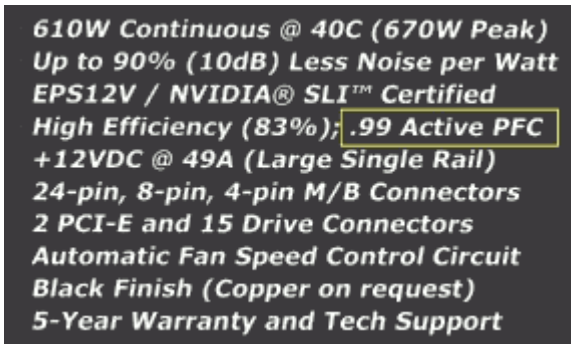
This is a simple way of correcting the nonlinearity of a load by using capacitor banks. It is not as effective as active PFC.

Passive PFCs are typically more power efficient than active PFCs. *Efficiency* is not to be confused with the PFC, though many computer hardware reviews conflate them. A passive PFC on a switching computer PSU has a typical power efficiency of around 96%, while an active PFC has a typical efficiency of about 94%.

Active PFC

An "active power factor corrector" (active PFC) is a power electronic system that controls the amount of power drawn by a load in order to obtain a power factor as close as possible to unity. In most applications, the active PFC controls the input current of the load so that the current waveform is proportional to the mains voltage waveform (a sine wave). The purpose of making the power factor as close to unity (1) as possible is to make the load circuitry that is power factor corrected appear purely resistive (apparent power equal to real power). In this case, the voltage and current are in phase and the

reactive power consumption is zero. This enables the most efficient delivery of electrical power from the power company to the consumer.



Specifications taken from the packaging of a 610W PC power supply showing Active PFC rating

Some types of active PFC are:

- Boost
- Buck
- Buck-boost

Active power factor correctors can be single-stage or multi-stage.

In the case of a switched-mode power supply, a boost converter is inserted between the bridge rectifier and the main input capacitors. The boost converter attempts to maintain a constant DC bus voltage on its output while drawing a current that is always in phase with and at the same frequency as the line voltage. Another switchmode converter inside the power supply produces the desired output voltage from the DC bus. This approach requires additional semiconductor switches and control electronics, but permits cheaper and smaller passive components. It is frequently used in practice. For example, SMPS with passive PFC can achieve power factor of about 0.7–0.75, SMPS with active PFC, up to 0.99 power factor, while a SMPS without any power factor correction has a power factor of only about 0.55–0.65.

Due to their very wide input voltage range, many power supplies with active PFC can automatically adjust to operate on AC power from about 100 V (Japan) to 230 V (Europe). That feature is particularly welcome in power supplies for laptops.

Importance of power factor in distribution systems

The significance of power factor lies in the fact that utility companies supply customers with volt-amperes, but bill them for watts. Power factors below 1.0 require a utility to generate more than the minimum volt-amperes necessary to supply the real power (watts). This increases generation and transmission costs. For example, if the load power factor were as low as 0.7, the apparent power would be 1.4 times the real power used by

the load. Line current in the circuit would also be 1.4 times the current required at 1.0 power factor, so the losses in the circuit would be doubled (since they are proportional to the square of the current). Alternatively all components of the system such as generators, conductors, transformers, and switchgear would be increased in size (and cost) to carry the extra current.

Utilities typically charge additional costs to customers who have a power factor below some limit, which is typically 0.9 to 0.95. Engineers are often interested in the power factor of a load as one of the factors that affect the efficiency of power transmission.

With the rising cost of energy and concerns over the efficient delivery of power, active PFC has become more common in consumer electronics. Current Energy Star guidelines for computers (ENERGY STAR® Program Requirements for Computers Version 5.0) call for a power factor of ≥ 0.9 at 100% of rated output in the PC's power supply. According to a white paper authored by Intel and the U.S. Environmental Protection Agency, PCs with internal power supplies will require the use of active power factor correction to meet the ENERGY STAR® 5.0 Program Requirements for Computers.

In Europe, IEC 555-2 requires power factor correction be incorporated into consumer products.

Measuring power factor

Power factor in a single-phase circuit (or balanced three-phase circuit) can be measured with the wattmeter-ammeter-voltmeter method, where the power in watts is divided by the product of measured voltage and current. The power factor of a balanced polyphase circuit is the same as that of any phase. The power factor of an unbalanced polyphase circuit is not uniquely defined.

A direct reading power factor meter can be made with a moving coil meter of the electrodynamic type, carrying two perpendicular coils on the moving part of the instrument. The field of the instrument is energized by the circuit current flow. The two moving coils, A and B, are connected in parallel with the circuit load. One coil, A, will be connected through a resistor and the second coil, B, through an inductor, so that the current in coil B is delayed with respect to current in A. At unity power factor, the current in A is in phase with the circuit current, and coil A provides maximum torque, driving the instrument pointer toward the 1.0 mark on the scale. At zero power factor, the current in coil B is in phase with circuit current, and coil B provides torque to drive the pointer towards 0. At intermediate values of power factor, the torques provided by the two coils add and the pointer takes up intermediate positions.

Another electromechanical instrument is the polarized-vane type. In this instrument a stationary field coil produces a rotating magnetic field, just like a polyphase motor. The field coils are connected either directly to polyphase voltage sources or to a phase-shifting reactor if a single-phase application. A second stationary field coil, perpendicular to the voltage coils, carries a current proportional to current in one phase of the circuit.

The moving system of the instrument consists of two vanes which are magnetized by the current coil. In operation the moving vanes take up a physical angle equivalent to the electrical angle between the voltage source and the current source. This type of instrument can be made to register for currents in both directions, giving a 4-quadrant display of power factor or phase angle.

Digital instruments can be made that either directly measure the time lag between voltage and current waveforms and so calculate the power factor, or by measuring both true and apparent power in the circuit and calculating the quotient. The first method is only accurate if voltage and current are sinusoidal; loads such as rectifiers distort the waveforms from the sinusoidal shape.

Mnemonics

English-language power engineering students are advised to remember: "ELI the ICE man" or "ELI on ICE" – the voltage E leads the current I in an inductor L, the current leads the voltage in a capacitor C.

Or CIVIL – in a capacitor(C) the current (I) leads voltage(V), voltage(V) leads current(I) in an inductor(L).

Chapter 7

Phase Margin, Power Gain & Power Rating

Phase Margin

In electronic amplifiers, **phase margin** (PM) is the difference between the phase, measured in degrees, of an amplifier's output signal (relative to its input) and 180°, as a function of frequency. The PM is taken as positive at frequencies below where the open-loop phase first crosses 180°, i.e. the signal becomes inverted, or antiphase; it is negative beyond. In negative feedback, a negative PM at a frequency where the loop gain exceeds unity guarantees instability, thus positive PM is a "safety margin" that ensures proper operation of an amplifier and, more generally, active filters, under various loads. In its simplest form, involving ideal negative feedback *voltage* amplifiers with non-reactive feedback, the phase margin is measured at the frequency where the open loop voltage gain of the amplifier equals the desired closed loop DC voltage gain.

More generally, PM is defined as that of the amplifier and its feedback network combined (the "loop", normally opened at the amplifier input), measured at a frequency where the loop gain is unity, and prior to the closing of the loop through tying the output of the open loop to the input source in such a way as to subtract from it.

In the above loop-gain definition, it is assumed that the amplifier input presents zero load. To make this work for non-zero-load input, the output of the feedback network needs to be loaded with an equivalent load for the purpose of determining the frequency response of the loop gain.

It is also assumed that the gain vs. frequency crosses unity gain with a negative slope and does so only once. This consideration matters only with reactive and active feedback networks, as may be the case with active filters.

Phase margin and its important companion concept, gain margin, are measures of stability in closed loop dynamic control systems. Phase margin indicates relative stability, the tendency to oscillate during its damped response to an input change such as a step function. Gain margin indicates absolute stability and the degree to which the system will oscillate without limit given any disturbance.

The output signals of all amplifiers exhibit a time delay when compared to their input signals. This delay causes a phase difference between the amplifier's input and output signals. If there are enough stages in the amplifier, at some frequency, the output signal will lag behind the input signal by one wavelength. In this situation, the amplifier's output signal will be in phase with its input signal though lagging behind it by 360°, i.e., the output will have a phase angle of –360°. This lag is of great consequence in amplifiers that use feedback. The reason: the amplifier will oscillate if the fed-back output signal is in phase with the input signal at the frequency at which its open loop voltage gain equals its closed loop voltage gain and the open loop voltage gain is one or greater. The oscillation will occur because the fed-back output signal will then reinforce the input signal at that frequency. In conventional operational amplifiers, the critical output phase angle is –180° because the output is fed back to the input through an inverting input which adds an additional –180°.

In practice, feedback amplifiers must be designed with phase margins substantially in excess of 0°, even though amplifiers with phase margins of, say, 1° are theoretically stable. The reason is that many practical factors can reduce the phase margin below the theoretical minimum. A prime example is when the amplifier's output is connected to a capacitive load. Therefore, operational amplifiers are usually compensated to achieve a minimum phase margin of 45° or so. This means that at the frequency at which the open and closed loop gains meet, the phase angle is -135°. The calculation is: -135° - (-180°) = 45°. Often amplifiers are designed to achieve a typical phase margin of 60 degrees. If the typical phase margin is around 60 degrees then the minimum phase margin will typically be greater than 45 degrees. A phase margin of 60 degrees is also a magic number because it allows for the fastest settling time when attempting to follow a voltage step input (a Butterworth design). An amplifier with lower phase margin will ring for longer and an amplifier with more phase margin will take a longer time to rise to the voltage step's final level.

A related measure is gain margin. While phase margin comes from the phase where the loop gain equals one, the gain margin is based upon the gain where the phase equals -180 degrees.

Power Gain

The **power gain** of an electrical network is the ratio of an output power to an input power. Unlike other signal gains, such as voltage and current gain, "power gain" may be ambiguous as the meaning of terms "input power" and "output power" is not always clear. Three important power gains are **operating power gain**, **transducer power gain** and **available power gain**. Note that all these definitions of power gains employ the use

of average (as oppose of instantaneous) power quantities and therefore the term "average" is often suppressed, which can be confusing at occasions.

Operating power gain

The operating power gain of a two-port network, G_P, is defined as:

$$G_P = \frac{P_{\text{load}}}{P_{\text{input}}}$$

where

- P_{load} is the maximum power delivered to the load
- P_{input} is the average power entering the network

Transducer power gain

The transducer power gain of a two-port network, G_T, is defined as:

$$G_T = \frac{P_{\text{load}}}{P_{\text{source,max}}}$$

where

- P_{load} is the average power delivered to the load
- $P_{source,max}$ is the maximum available average power at the source

In terms of y-parameters this definition can be used to derive:

$$G_T = \frac{4|y_{21}|^2 \Re(Y_L) \Re(Y_S)}{|(y_{11} + Y_S)(y_{22} + Y_L) - y_{12}y_{21}|^2}$$

where

- Y_L is the load admittance
- Y_S is the source admittance

This result can be generalized to z, h, g and y-parameters as:

$$G_T = \frac{4|k_{21}|^2 \Re(M_L) \Re(M_S)}{|(k_{11} + M_S)(k_{22} + M_L) - k_{12}k_{21}|^2}$$

where

- k_{xx} is a z, h, g or y-parameter
- M_L is the load value in the corresponding parameter set
- M_S is the source value in the corresponding parameter set

$P_{source,max}$ may only be obtained from the source when the load impedance connected to it (i.e. the equivalent input impedance of the two-port network) is the complex conjugate of the source impedance, a consequence of the maximum power theorem.

Available power gain

The available power gain of a two-port network, G_A, is defined as:

$$G_A = \frac{P_{\text{load,max}}}{P_{\text{source,max}}}$$

where

- $P_{load,max}$ is the maximum available average power at the load
- $P_{source,max}$ is the maximum power available from the source

Similarly $P_{load,max}$ may only obtained when the load impedance is the complex conjugate of the output impedance of the network.

Power Rating

In electrical and electronic engineering, the **power rating** of a device is a guideline set by the manufacturer as a maximum power to be used with that device. This limit is usually set somewhat lower than the level where the device will be damaged, to allow a margin of safety.

The power rating can actually mean a couple different things. In devices which primarily dissipate electric power or convert it into mechanical power, such as resistors, electric motors, and speakers, the power rating given is usually the maximum power that can be safely dissipated by the device. The usual reason for this limit is heat, although in certain electromechanical devices, particularly speakers, it is to prevent mechanical damage. When heat is the limiting factor, the power rating is easily calculated. First, the amount of heat that can be safely dissipated by the device, $P_{D,max}$, must be calculated. This is related to the maximum safe operating temperature, the ambient temperature or temperature range in which the device will be operated, and the method of cooling. If $T_{D,max}$ is the maximum safe operating temperature of the device, T_A is the ambient temperature, and θ_{DA} is the total thermal resistance between the device and ambient, then the maximum heat dissipation is given by

$$P_{D,max} = \frac{T_{D,max} - T_A}{\theta_{DA}}$$

If all power in a device is dissipated as heat, then this is also the power rating. On the other hand, if most of the power is converted into mechanical power, then we need to know the efficiency, η. Then, the power rating is given by

$$P_{max} = \frac{P_{D,max}}{1 - \eta}$$

Note that this is the real or effective power dissipated in the device.

In devices that primarily convert between different forms of electric power, such as transformers, or transport it from one location to another, such as transmission lines, the power rating almost always refers to the maximum power flow through the device. If an amount of power equal to the power rating were actually dissipated in the device, it would certainly be damaged. The usual reason for the limit is again heat, and the maximum heat dissipation is calculated as above. However, there may not be a direct relationship between power dissipated as heat and power converted by the device; simply put, the power converted depends on the power factor of the load whereas the heat dissipated does not. In this case, the maximum current is calculated and the power rating is given by

$$S_{max} = V_{nom} I_{max}$$

where V_{nom} is the nominal operating voltage. Note that the power rating in this case is an apparent power.

Power ratings are usually given in watts for real power and volt-amperes for apparent power, although for devices intended for use in large power systems, both may be given in a per-unit system. As the power rating depends on the method of cooling, different ratings may be specified for air cooling, water cooling, etc.

Exceeding the power rating of a device by more than the margin of safety set by the manufacturer usually does damage to the device by causing its operating temperature to exceed safe levels. In semiconductors, irreparable damage can occur very quickly. Exceeding the power rating of most devices for a very short period of time is not harmful, although doing so regularly can sometimes cause cumulative damage.

Audio amplifiers

Audio amplifier power ratings are typically established by driving the device under test to the onset of clipping, to a predetermined distortion level, variable per manufacturer or per product line. Driving an amplifier to 1% distortion levels will yield a higher rating than driving it to 0.01% distortion levels. Similarly, testing an amplifier at a single mid-range

frequency, or testing just one channel of a two-channel amplifier, will yield a higher rating than if it is tested throughout its intended frequency range with both channels working. Manufacturers can use these methods to market amplifiers whose published maximum power output includes some amount of clipping in order to show higher numbers. For instance, the Federal Trade Commission (FTC) established an amplifier rating system in which the device is tested with both channels driven throughout its advertised frequency range, at no more than its published distortion level. The Electronic Industries Association (EIA) rating system, however, determines amplifier power by measuring a single channel at 1,000 Hz, with a 1% distortion level—1% clipping. Using the EIA method rates an amplifier 10 to 20% higher than the FTC method.

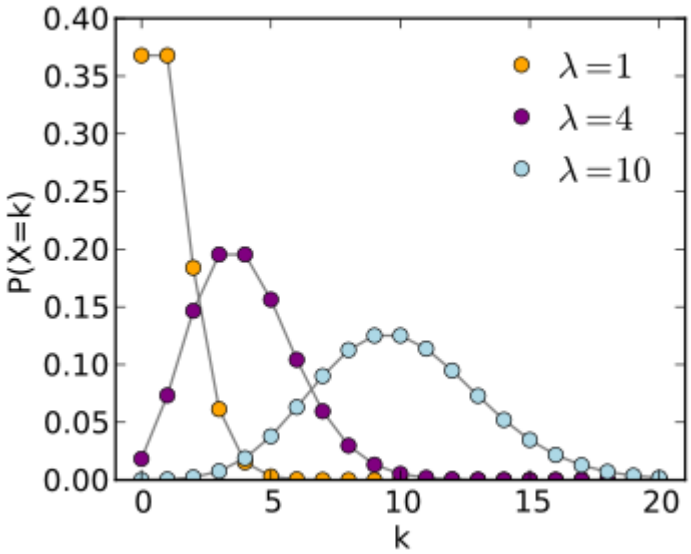
Chapter 8

Shot Noise

Photon noise simulation.

Shot noise is a type of electronic noise that occurs when the finite number of particles that carry energy (such as electrons in an electronic circuit or photons in an optical device) is small enough to give rise to detectable statistical fluctuations in a measurement. It is important in electronics, telecommunications, and fundamental physics.

It also refers to an analogous noise in particle simulations, where due to the small number of particles, the simulation exhibits detectable statistical fluctuations not observed in the real-world system. Magnitude of this noise increases with the average magnitude of the current or intensity of the light. However, since the magnitude of the average signal increases more rapidly than that of the shot noise (its relative strength decreases with increasing signal), shot noise is often only a problem with small currents or light intensities.



The number of photons that are collected by a given detector varies, and follows a Poisson distribution, depicted here for averages of 1, 4, and 10.

The intensity of a source will yield the *average* number of photons collected, but knowing the average number of photons which will be collected will not give the actual number collected. The actual number collected will be more than, equal to, or less than the average, and their distribution about that average will be a Poisson distribution.

Since the Poisson distribution approaches a normal distribution for large numbers, the photon noise in a signal will approach a normal distribution for large numbers of photons collected. The standard deviation of the photon noise is equal to the square root of the average number of photons. The signal-to-noise ratio is then

$$\mathrm{SNR} = \frac{N}{\sqrt{N}} = \sqrt{N}$$

where N is the average number of photons collected. When N is very large, the signal-to-noise ratio is very large as well. It can be seen that photon noise becomes more important when the number of photons collected is small.

Explanation

Intuitive explanation

It is known to everyone that in a statistical experiment such as tossing a fair coin and counting the occurrences of heads and tails, the numbers of heads and tails after a great many throws will differ by only a tiny percentage, while after only a few throws outcomes with a significant excess of heads over tails or vice versa are common; if an experiment with a few throws is repeated over and over, the outcomes will fluctuate a lot. (It can be proved that the relative fluctuations reduce as the square root of the number of throws, a result valid for all statistical fluctuations, including shot noise.)

Shot noise exists because phenomena such as light and electric current consist of the movement of discrete, quantized 'packets'. Consider light—a stream of discrete photons—coming out of a laser pointer and hitting a wall to create a visible spot. The fundamental physical processes that govern light emission are such that these photons are emitted from the laser at random times; but the many billions of photons needed to create a spot are so many that the brightness, the number of photons per unit time, varies only infinitesimally with time. However, if the laser brightness is reduced until only a handful of photons hit the wall every second, the relative fluctuations in number of photons, i.e. brightness, will be significant, just as when tossing a coin a few times. These fluctuations are shot noise.

In electronic devices

Shot noise in electronic devices consists of random fluctuations of the electric current in many electrical conductors, due to the current being carried by discrete charges (electrons) whose number per unit time fluctuates. This is often an issue in p-n junctions. In metal wires this is not an issue, as correlations between individual electrons remove these random fluctuations.

Shot noise is distinct from current fluctuations in thermal equilibrium, which happen without any applied voltage and without any average current flowing. These thermal equilibrium current fluctuations are known as Johnson-Nyquist noise or thermal noise.

Shot noise is a Poisson process and the charge carriers which make up the current will follow a Poisson distribution. The current fluctuations have a standard deviation of

$$\sigma_i = \sqrt{2\,q\,I\,\Delta f}$$

where q is the elementary charge, Δf is the bandwidth in hertz over which the noise is measured, and I is the average current through the device. All quantities are assumed to be in SI units.

For a current of 100 mA this gives a value of

$$\sigma_i = 0.18\,\mathrm{nA}$$

if the noise current is filtered with a filter having a bandwidth of 1 Hz.

If this noise current is fed through a resistor the resulting noise power will be

$$P = 2\,q\,I\,\Delta f R.$$

If the charge is not fully localized in time but has a temporal distribution of $q\,F(t)$ where the integral of $F(t)$ over t is unity then the power spectral density of the noise current signal will be,

$$S_i(f) = 2\,q\,I\,|\Psi(f)|^2,$$

where $\Psi(f)$ is the Fourier transform of $F(t)$.

Note: Shot noise and Johnson–Nyquist noise are both quantum fluctuations. Some authors treat them as a single unified concept.

In quantum optics

In quantum optics, shot noise is caused by the fluctuations of detected photons, again therefore a consequence of discretization (of the energy in the electromagnetic field in this case). Shot noise is a main part of **quantum noise**.

Shot noise is measurable not only in measurements at the few-photons level using photomultipliers, but also at stronger light intensities measured by photodiodes when using high temporal resolution oscilloscopes. As the photocurrent is proportional to the light intensity (number of photons), the fluctuations of the electromagnetic field are usually contained in the electric current measured.

In the case of a *coherent light* source such as a laser, the shot noise scales as the square-root of the average intensity:

$$\Delta I^2 \ \stackrel{\mathrm{def}}{=}\ \langle (I - \langle I \rangle)^2 \rangle \propto I.$$

A similar lower bound of quantum noise occurs in linear quantum amplifiers. The only exception being if a squeezed coherent state can be formed through correlated photon generation. The reduction of uncertainty of the number of photons per mode (and therefore the photocurrent) may take place just due to the saturation of gain; this is intermediate case between a laser with locked phase and amplitude-stabilized laser.

Space charge

Low noise active electronic devices are designed such that shot noise is suppressed by the electrostatic repulsion of the charge carriers. Space charge limiting is not possible in photon devices.

Chapter 9

Stray Voltage

Stray voltage describes the occurrence of voltage between two objects that should not have any voltage difference between them. Small voltages are often measured between two grounded objects in distant locations, due to normal current flow in the power system. Large voltages can appear on the enclosures of electrical equipment due to a fault in the electrical power system, such as a failure of insulation.

Small stray voltages may never be noticed and may only be detected with a voltmeter. Larger voltages may have a range of effects, from barely perceptible to dangerous electric shocks. Normally, metal electrical equipment cases are bonded to ground to prevent a shock hazard if energized conductors accidentally contact the case. Where this bonding is not provided or has failed, a severe hazard of electric shock or electrocution is presented when circuit conductors contact the case.

In any place where equipment is in direct contact with a person or animal (such as pools, surgery, electric milking machines, and many others), particular attention must be paid to elimination of stray voltages. Dry intact skin has a higher resistance than wet skin or a wound, so voltages that would otherwise be unnoticed would be significant for a wet or surgical case.

Terminology

"Stray voltage" describes any case of elevated potential, but more precise terminology gives an indication of the source of the voltage. "Neutral to earth voltage" specifically refers to a difference in potential between a locally grounded object and the grounded return conductor, or neutral, of an electrical system. The neutral is theoretically at 0 V potential, as any grounded object, but current flows on the neutral back to the source, elevating the neutral voltage. This elevated potential between the neutral and earth is also called “neutral to earth voltage” or NEV. NEV is the product of current flowing on the neutral and the impedance of the neutral conductor between a given point and its source,

often a distant substation. NEV differs from accidentally energized objects because it is a result of normal system operation, not an accident or a fault in materials or design.

In 2005, the Institute of Electrical and Electronics Engineers (IEEE) convened working group 1695 to lay down definitions and guidelines for mitigating the various phenomena referred to as stray voltage. Two draft definitions emerged for stray voltage and contact voltage:

- *Stray voltage* defined as " A voltage resulting from the normal delivery and/or use of electricity (usually smaller than 10 volts) that may be present between two conductive surfaces that can be simultaneously contacted by members of the general public and/or their animals. Stray voltage is caused by primary and/or secondary return current, and power system induced currents, as these currents flow through the impedance of the intended return pathway, its parallel conductive pathways, and conductive loops in close proximity to the power system. Stray voltage is not related to power system faults, and is generally not considered hazardous. "
- *Contact voltage* defined as " A voltage resulting from abnormal power system conditions that may be present between two conductive surfaces that can be simultaneously contacted by members of the general public and/or their animals. Contact voltage is caused by power system fault current as it flows through the impedance of available fault current pathways. Contact voltage is not related to normal system operation and can exist at levels that may be hazardous."

In New York City, a woman named Jodie S. Lane was electrocuted by a manhole cover energized by an “improperly insulated wire” in January 2004. Her death made public electrical safety in the urban environment a major concern for utilities. The term “stray voltage” was used by the media and NY state regulatory agency. In 2005, the state acknowledged that “stray voltage” properly refers to neutral to earth voltage (NEV), but conceded that the notoriety of the Jodie S. Lane incident had caused stray voltage to be a term that is well understood by the public. At that point, the regulator used stray voltage to describe any “voltage conditions on electric facilities that should not ordinarily exist. These conditions may be due to one or more factors, including, but not limited to, damaged cables, deteriorated, frayed or missing insulation, improper maintenance, or improper installation.” In the same document, the commission accepted NEV to be a naturally occurring condition.

Since that time, the term “stray voltage” has two definitions. This decision is cause for confusion among utilities, regulators, and the public. Other more esoteric phenomenon also result in elevated voltages on normally non-energized surfaces, but are also referred to as “stray voltage.” Examples are voltage due to capacitive coupling, current induced by power lines, EMF, lightning, earth potential rise, and problems stemming from open neutrals.

Origins of stray voltages

Coupled and induced voltages

Ungrounded metal objects close to electric field sources such as neon signs or conductors carrying alternating currents can have measurable voltage levels caused by capacitive coupling. Capacitive coupling is the mechanism used by Electrical tester pen devices. Because the capacitance between an object and a current source is typically small, only very small currents can flow from the energized source to the coupled object. High impedance digital voltmeters may measure elevated voltages from non-energized objects due to this coupling, in effect providing a false reading. For this reason, voltage measurements of normally non-energized objects must be verified. Verification of a voltage reading is performed using a shunt resistor in parallel with voltmeter test leads. Another approach is to use a low impedance voltmeter. Since no current can flow from a coupled surface through the small shunt or meter resistance, capacitively coupled voltages will collapse to zero. If the object is in contact with a current source , the voltage will drop only slightly as dictated by Ohm's Law.

Classical induction does occur when long conductors form an open grounded loop under and parallel to transmission or distribution lines. In these cases, current is induced in the loop when a person makes contact with it and ground. Since this involves real current flow, it is potentially hazardous. This type of induced current occurs most often on long fences and distribution lines built under transmission.

The very small capacitance between overhead line and a fluorescent lamp provides enough current to cause the lamp to glow.

Since voltages detected by high-impedance instruments disappear or become greatly reduced when a low impedance is substituted, the effect is sometimes called **phantom voltage** (or *ghost voltage*). The term is often used by electricians, and might be seen, for example, when measuring the voltage at a lighting fixture after removing the bulb. It's not unusual to measure *phantom voltages* of 50-90 volts when testing the wiring of ordinary 120 V circuits with a high-impedance instrument. While the voltage produced may read almost to the full supply voltage, the capacitance or mutual inductance between the wires of building wiring systems is typically quite low and incapable of supplying significant amounts of current. However, in overhead transmission work on high-voltage lines, safety rules require connecting a conductor to earth ground during maintenance, since

induced voltages on a conductor may be sufficient to cause electrocution or serious injury.

Degraded insulation on power conductors

Another form of stray voltage is due to damaged or degraded insulation. Damaged insulation is essentially a high impedance fault which will allow current to flow through any available path to ground, a condition which can cause shocks or fires if left unmitigated. This leakage can occur when there is damage caused by physical, thermal, or chemical stresses to insulation on underground power lines. Examples of this damage is swollen or cracked insulation from overloading, abrasions caused by digging or ground seizing, and damage from salt or oil exposure. Leakage can also occur due to moisture, salt, dust, and dirt buildup on open air insulators in overhead distribution. If the leakage in these cases is severe enough, it can lead to a pole fire.

Electrolysis and corrosion

Dissimilar buried metals such as copper and steel can function as the poles of a galvanic cell, using moist soil as the electrolyte. Stray direct currents in soil may counteract the anti-corrosion effect of a cathodic protection system. Design of high voltage direct current transmission systems must take care so that current flowing in the earth does not cause objectionable corrosion to buried objects such as pipelines.

Typically an electric railway will have at least one of the rails used as a return conductor for the traction current. This rail is in contact with the earth at many places throughout its length. Since current will follow every parallel path between source and load, some part of the traction current will also flow through the earth. Where the railway uses direct current, this stray current can cause damage to other buried metallic objects by electrolysis and accelerate corrosion of metal objects in touch with the soil.

Capacitive leakage through insulation

Alternating current is different from direct current in that the current can flow through what would ordinarily seem to be a physical barrier. In a series circuit, a capacitor blocks direct current but passes alternating current.

In power transmission systems, one side of the circuit, known as the neutral, is grounded to dissipate static electricity. It is possible to get a shock by only touching the *hot* wire, due to the person's body being capacitively coupled to the ground upon which the person stands.

Leakage from single-wire earth return

The term "stray voltage" is used to describe the gradient (rate of change with respect to distance) of electrical potential in the surface of the soil, associated with single-wire earth return electricity distribution systems used in some rural locations. This gradient is low at

points far away from the earth return connections, but increases near the ground rods where the metallic circuit enters the earth.

Neutral return currents through the ground

In three phase four-wire ("wye") electrical power systems, when the load on the phases is not exactly equal, there is some current in the neutral conductor. Because both the primary and secondary of the distribution transformer are grounded, and the primary ground is grounded at more than one point, the earth forms a parallel return path for the neutral current, allowing part of the neutral current to continuously flow through the earth. This arrangement is partially responsible for stray voltage.

Stray voltage is a result of the design of a 4 wire distribution system and as such has existed as long as such systems have been used. Stray voltage became a problem for the dairy industry some time after electric milking machines were introduced, and large numbers of animals were simultaneously in contact with metal objects grounded to the electric distribution system and the earth. Numerous studies document the causes, physiological effects, and prevention, of stray voltage in the farm environment. Today, stray voltage on farms is regulated by state governments and controlled by the design of equipotential planes in areas where livestock eat, drink or give milk. Commercially available neutral isolators also prevent elevated potentials on the utility system neutral from raising the voltage of farm neutral or ground wires.

Public concerns about stray voltage

In metropolitan areas, stray voltage issues have become a major concern. These areas have large amounts of aging underground electrical distribution equipment in crowded public spaces. Even a low rate of insulation failures or current leakage can result in hazardous exposure of the general public.

Consolidated Edison has had frequent incidents of stray voltage. including the death of Jodie S. Lane in 2004, while walking her dog in Manhattan.

Toronto Hydro pulled all employees off regular duty on the weekend of January 30, 2009 to deal with ongoing stray voltage problems in the city. This came after as many as five children were shocked though none suffered serious injury. The stray voltage problem had claimed the lives of two dogs in the previous few months.

Effects on people

Potential differences between pool water and railings or shower facilities and grounded drain pipes are not uncommon as a result of neutral to earth voltages (NEV), and can be a major nuisance, but are not life threatening. Contact voltage results from damaged insulation on a current carrying conductor, however, can be very dangerous and can lead to shock or electrocution. Such a condition can arise spontaneously from mechanical, thermal, or chemical stress on insulation materials or from unintentional damage from

digging activity, freeze-frost seizing, corrosion and collapse of conduit, or even workmanship issues. Contact voltage energizes objects which are normally safe – fences, telephone booths, street signs, etc. Anywhere buried electric wiring exists, a failure can occur in that wiring and create conditions for allow electricity to flow into the immediate surroundings. Some systems have protective devices such as circuit breakers or GFCI, designed to isolate such a fault. However, in the absence of protective devices, if the devices fail, or if they are not installed correctly, a fault will go undetected until it either causes a failure of the circuit or until it is found by a person.

Effects on farm animals

Stray voltage can have harmful effects on animal health and productivity. Some dairy farmers have claimed damage to yields or stock caused by it.

Dr. Douglas J. Reinemann,Professor of Biological Systems Engineering at University of Wisconsin–Madison, reported on stray voltages on dairy farms in 2003. Investigation of stray voltage claims must also consider other animal health concerns.

Legal proceedings in Wisconsin

In 2003, the Wisconsin Supreme Court upheld a judgement of $1.2 million against the Wisconsin electrical utility WEPCO in *Hoffman* v. *Wisconsin Electric Power Company*. The Hoffman family, dairy farmers near New London, had sued WEPCO after several years of declining production. WEPCO had measured on the farm currents due to stray voltage below one milliampere, the "level of concern" set by the Wisconsin Public Service Commission, but the court ruled on procedural grounds that the utility could be found negligent under common law even though they met the state standard. The Hoffmans had presented, the court said, a viable alternative theory that stray voltage had caused them economic harm.

Stray/contact voltage detection

Stray voltage is generally found during routine electrical work or as a result of an customer complaint or shock incident. A growing number of utilities in urban areas conduct active tests for stray voltage, or more specifically, contact voltage for public safety reasons. Some electrical faults may also be discovered during routine work or inspection programs which are not specifically focused on stray voltage.

Equipment used to detect stray voltage vary, but common devices are electrical tester pens or electric field detectors, with follow up testing using a voltmeter. Electrical tester pens are hand-held devices which detect a potential difference between the user's hand and the object being tested. They generally indicate on contact with an energized object, if the potential difference is above the sensitivity threshold of the device. Accuracy can be affected if the user is at an elevated potential him/herself or if the user is not making firm contact with the device using a bare hand. Electric field detectors detect the electric field strength relative to the user or mounting platform. By sensing electric field gradients

at a distance they detect energized objects without making direct contact. A low electric field reading also provides a positive indication that no objects are energized in a tested area. Electric field detectors respond to all field sources, and positive indications must be verified with a voltmeter to eliminate false positives. Electric field proximity sensing has other industrial applications from manufacturing to building security.

Chapter 10

Voltage Drop & Threshold Voltage

Voltage Drop

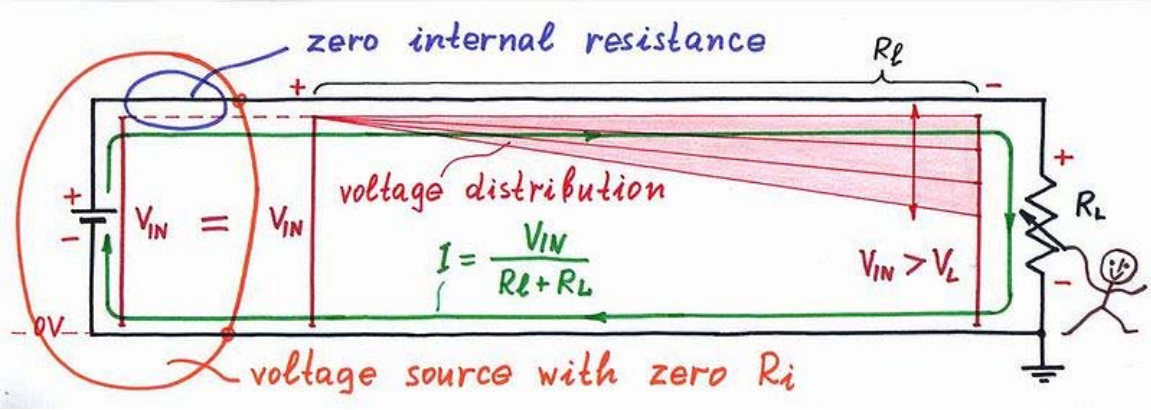
Voltage drop is the **reduction** in voltage in the passive elements (not containing sources) of an electrical circuit. Voltage drops across conductors, contacts, connectors and source internal resistances are undesired as they reduce the supplied voltage while voltage drops across loads and other electrical and electronic elements are useful and desired.

In electrical wiring, national and local electrical codes may set guidelines for maximum voltage drop allowed in a circuit conductors, to ensure reasonable efficiency of distribution and proper operation of electrical equipment (the maximum permitted voltage drop varies from one country to another). Voltage drop may be neglected when the impedance of the interconnecting conductors is small relative to the other components of the circuit. For example, an electric space heater may very well have a resistance of ten ohms, and the wires which supply it may have a resistance of 0.2 ohms, about 2% of the total circuit resistance. This means that 2% of the supplied voltage is actually being lost by the wire itself. Excessive voltage drop will result in unsatisfactory operation of electrical equipment, and represents energy wasted in the wiring system. Voltage drop can also cause damage to electrical motors.

In electronic design and power transmission, various techniques are used to compensate for the effect of voltage drop on long circuits or where voltage levels must be accurately maintained. The simplest way to reduce voltage drop is to increase the diameter of the conductor between the source and the load which lowers the overall resistance. The more sophisticated techniques use active elements to compensate the undesired voltage drop.

Voltage drop in direct current circuits

A current flowing through the non-zero resistance of a practical conductor necessarily produces a voltage across that conductor. The dc resistance of the conductor depends upon the conductor's length, cross-sectional area, type of material, and temperature.



The local voltages along a long line decrease gradually from the source to the load

If the voltage between the conductor and a fixed reference point is measured at many points along the conductor, the measured voltage will decrease gradually toward the load. As the current passes through a longer and longer conductor, more and more of the voltage is "lost" (unavailable to the load), due to the voltage drop developed across the resistance of the conductor. In this diagram the voltage drop along the conductor is represented by the shaded area. The local voltages along the line decrease gradually from the source to the load. If the load current increases, the voltage drop in the supply conductor also increases. Voltage drop exists in both supply and return wires of a circuit.

A principle known as Kirchhoff's circuit laws states that in any circuit, the sum of the voltage drops across each component of the circuit is equal to the supply voltage.

Voltage drop in alternating current circuits

In alternating current circuits, additional opposition to current flow occurs due to the interaction between electric and magnetic fields and the current within the conductor; this opposition is called "impedance". The impedance in an alternating current circuit depends on the spacing and dimensions of the conductors, the frequency of the current, and the magnetic permeability of the conductor and its surroundings. The voltage drop in an AC circuit is the product of the current and the impedance (Z) of the circuit. Electrical impedance, like resistance, is expressed in ohms. Electrical impedance is the vector sum of electrical resistance, capacitive reactance, and inductive reactance. The voltage drop occurring in an alternating current circuit is the product of the current and impedance of the circuit. It is expressed by the formula $E = IZ$, analogous to Ohm's law for direct current circuits.

Voltage drop in building wiring

Most circuits in a house do not have enough current or length to produce a high voltage drop. In the case of very long circuits, for example, connecting a home to a separate building on the same property, it may be necessary to increase the size of conductors over the minimum requirement for the circuit current rating. Heavily-loaded circuits may also require a cable size increase to meet voltage drop requirements in wiring regulations.

Wiring codes or regulations set an upper limit to the allowable voltage drop in a branch circuit. In the United States, the 2005 National Electrical Code (NEC) recommends no more than a 5% voltage drop at the outlet.. The Canadian electrical code requires no more than 5% drop between service entrance and point of use. UK regulations limit voltage drop to 4% of supply voltage. Following changes to the BS7671:2008 on consumers' installation, the following has become in force since 1 July 2008:

Type of Supply	**Voltage drop lighting**	**Voltage drop other**
DNO	3%	5%
Private	6%	8%

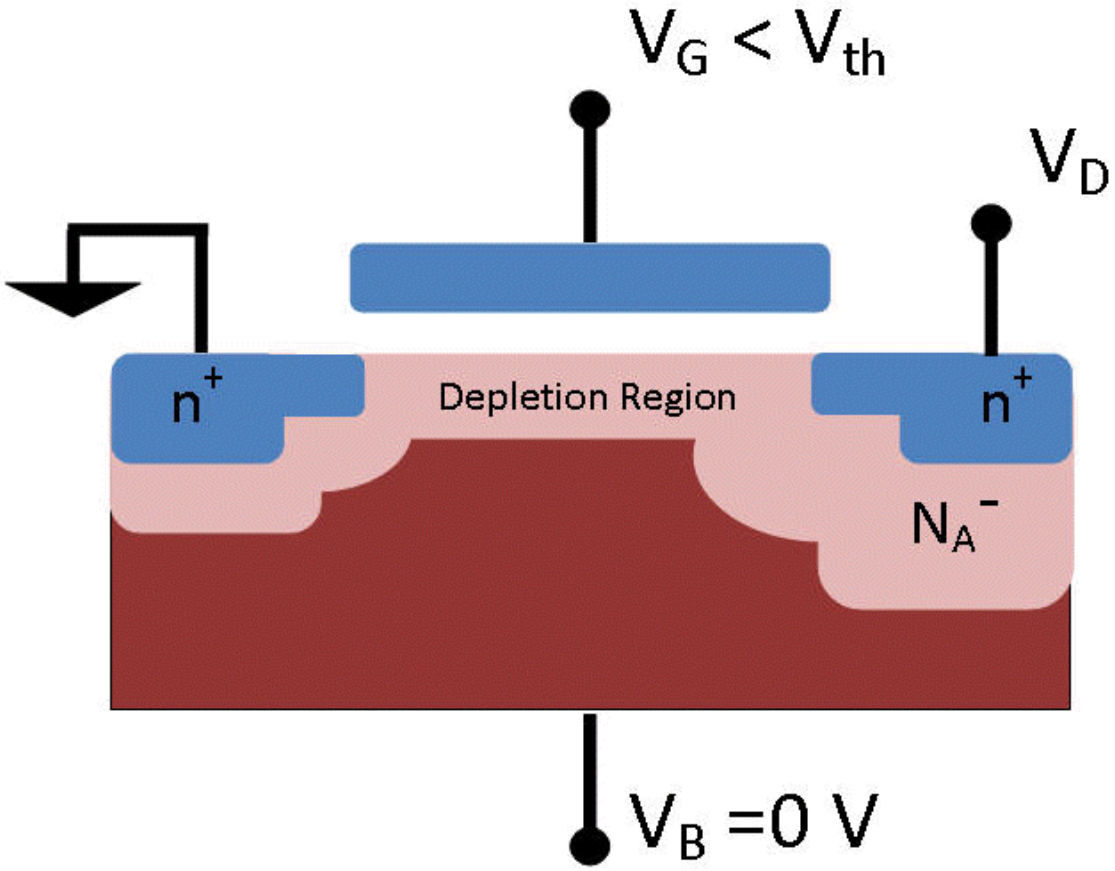
Voltage drop of a branch circuit is readily calculated, or less accurately it can be measured by observing the voltage before and after applying a load to the circuit. Excessive voltage drop on a residential branch circuit may be a sign of insufficiently sized wiring or of other faults within the wiring system, such as high resistance connections.

Using higher voltages

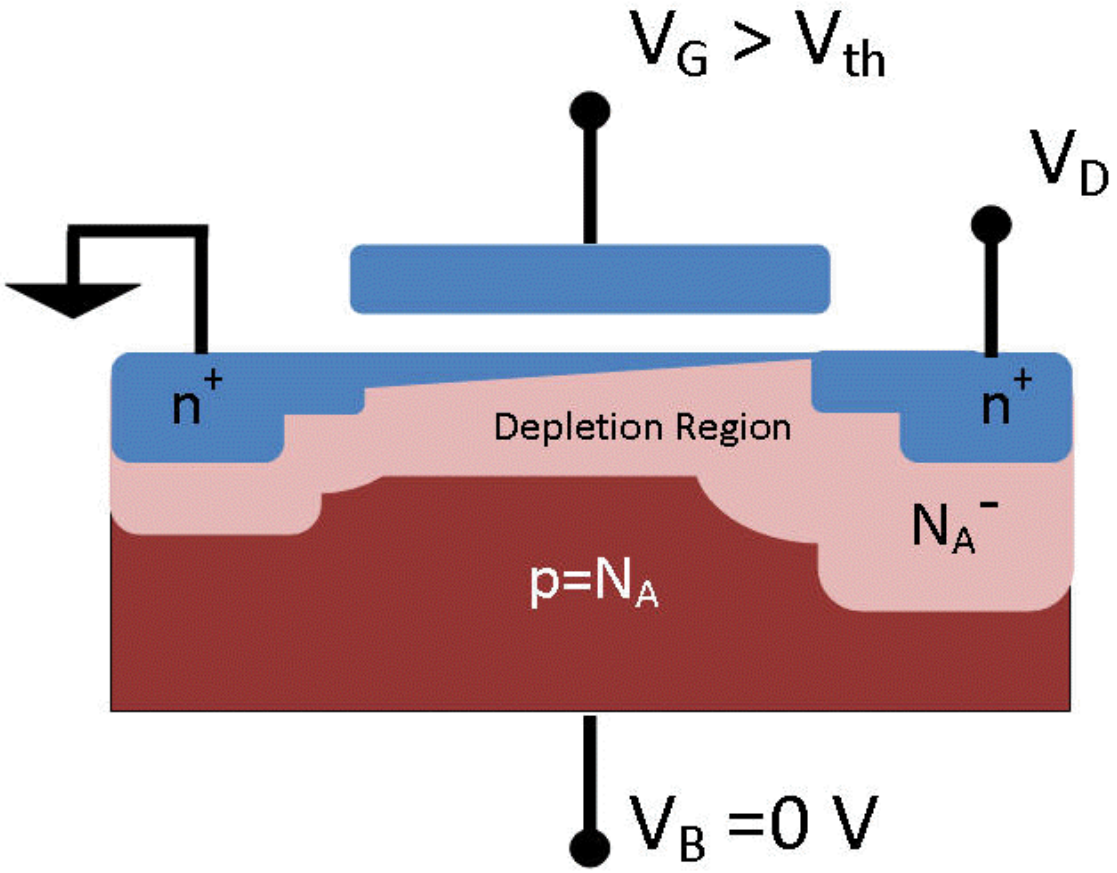
Over long distances, larger conductors become expensive, and it is preferable to redesign the circuit to operate at a higher voltage. Doubling the voltage halves the current required to deliver the same amount of power, halving the voltage drop, and an additional doubling in efficiency is realized because that drop is a smaller fraction of the total voltage.

This is the motivation for commercial high voltage electrical power distribution, and for the use of the +12V power supply rail for high-power loads in modern personal computers.

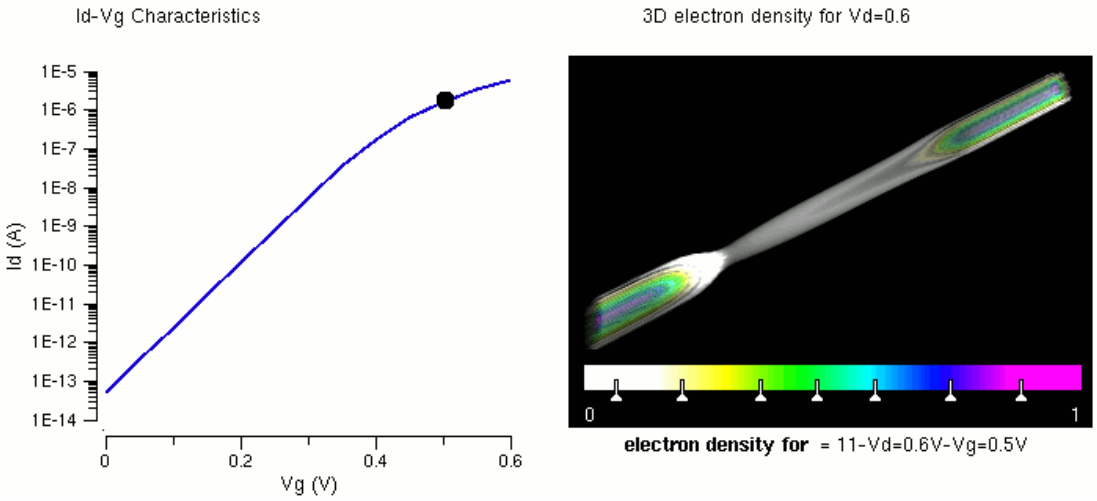
Threshold Voltage



Depletion region of an nMOSFET biased below threshold



Depletion region of an nMOSFET biased above threshold with channel formed



Simulation result for formation of inversion channel (electron density) and attainment of threshold voltage (IV) in a nanowire MOSFET. Note that the threshold voltage for this device lies around 0.45V.

The **threshold voltage** of a MOSFET is usually defined as the gate voltage where an inversion layer forms at the interface between the insulating layer (oxide) and the substrate (body) of the transistor. The purpose of the inversion layer's forming is to allow the flow of electrons through the gate-source junction. The creation of this layer is described next.

In an n-MOSFET the substrate of the transistor is composed of p-type silicon, which has positively charged mobile holes as carriers. When a positive voltage is applied on the gate, an electric field causes the holes to be repelled from the interface, creating a depletion region containing immobile negatively charged acceptor ions. A further increase in the gate voltage eventually causes electrons to appear at the interface, in what is called an inversion layer, or channel. Historically the gate voltage at which the electron density at the interface is the same as the hole density in the neutral bulk material is called the threshold voltage. Practically speaking the threshold voltage is the voltage at which there are sufficient electrons in the inversion layer to make a low resistance conducting path between the MOSFET source and drain.

In the figures, the source (left side) and drain (right side) are labeled $n+$ to indicate heavily doped (blue) n-regions. The depletion layer dopant is labeled N_A^- to indicate that the ions in the (pink) depletion layer are negatively charged and there are very few holes. In the (red) bulk the number of holes $p = N_A$ making the bulk charge neutral.

If the gate voltage is below the threshold voltage (top figure), the transistor is turned off and ideally there is no current from the drain to the source of the transistor. In fact, there is a current even for gate biases below threshold (subthreshold leakage) current, although it is small and varies exponentially with gate bias.

If the gate voltage is above the threshold voltage (lower figure), the transistor is turned on, due to there being many electrons in the channel at the oxide-silicon interface, creating a low-resistance channel where charge can flow from drain to source. For voltages significantly above threshold, this situation is called strong inversion. The channel is tapered when $V_D > 0$ because the voltage drop due to the current in the resistive channel reduces the oxide field supporting the channel as the drain is approached.

In modern devices the threshold voltage is a much less clear-cut parameter subject to variation with the biases applied to the device.

Body effect

The **body effect** describes the changes in the threshold voltage by the change in V_{SB}, the source-bulk voltage. Since the body influences the threshold voltage (when it is not tied to the source), it can be thought of as a second gate, and is sometimes referred to as the "back gate"; the body effect is sometimes called the "back-gate effect".

For an enhancement mode, n-mos MOSFET body effect upon threshold voltage is computed according to the Shichman-Hodges model (accurate for very old technology) using the following equation.

$$V_{TN} = V_{TO} + \gamma(\sqrt{V_{SB} + 2\phi_F} - \sqrt{2\phi_F})$$

where V_{TN} is the threshold voltage when substrate bias is present, V_{SB} is the source-to-body substrate bias, $2\varphi_F$ is the surface potential, and V_{TO} is threshold voltage for zero substrate bias, $\gamma = (t_{ox}/\epsilon_{ox})\sqrt{2q\epsilon_{si}N_A}$ is the body effect parameter, t_{ox} is oxide thickness, ε_{ox} is oxide permitivity, ε_{si} is the permitivity of silicon, N_A is a doping concentration, q is the charge of an electron.

Dependence on oxide thickness

In a given technology node, such as the 90 nanometer CMOS process, threshold voltage depends on the choice of oxide and on **oxide thickness**. Using the body formulas above, V_{TN} is directly proportional to γ, and t_{OX}, which is the parameter for oxide thickness.

Thus, the thinner the oxide thickness, the lower the threshold voltage. While this may seem to be an improvement, it is not without cost; for the thinner the oxide thickness, the higher the subthreshold leakage current flowing through the device will be. Consequently, the design specification for 90 nanometer gate oxide thickness was set at 1 nanometer to control the leakage current. This kind of tunneling, called Fowler-Nordheim Tunneling.

$$I_{fn} = C_1 W L (E_{ox})^2 e^{-E_0/E_{ox}}$$

where C_1 and E_0 are constants and E_{ox} is the electric field across the gate oxide.

Before scaling the design features down to 90 nanometers, a dual oxide approach for creating the oxide thickness was a common solution to this issue. With a 90 nanometer process technology, a triple oxide approach has been adopted in some cases. One standard thin oxide is used for most transistors, another for I/O driver cells, and a third for memory and pass transistor cells. These differences are based purely on the characteristics of oxide thickness on threshold voltage of CMOS technologies.

Dependence on temperature

As with the case of oxide thickness affecting threshold voltage, **temperature** has an effect on the threshold voltage of a CMOS device. Expanding on part of the equation in the body effect section

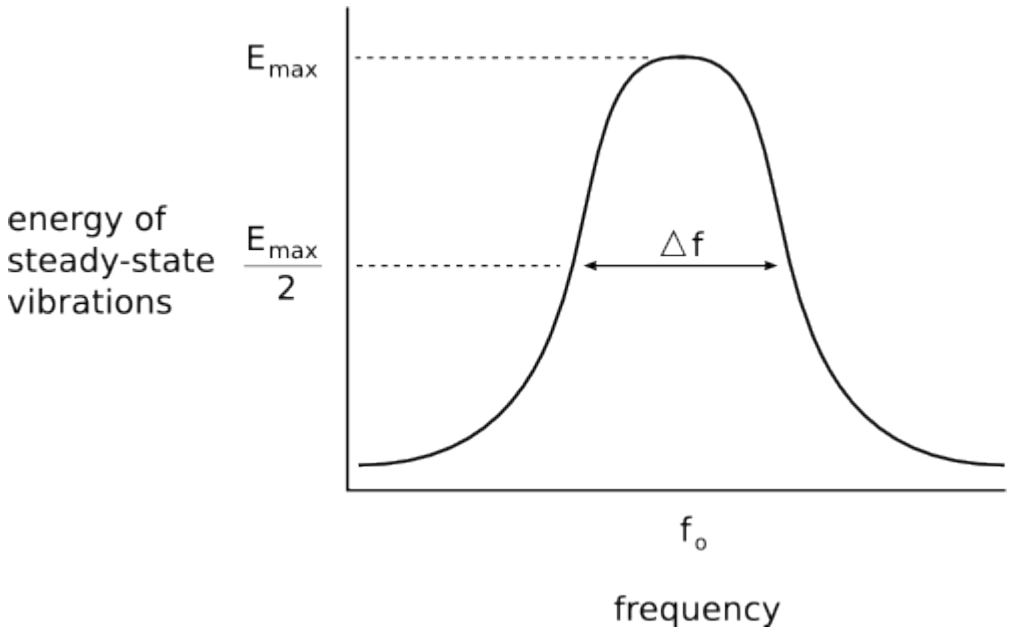
$$\varphi_f = (kT / q)\ln(N_A / N_i)$$

where k is Boltzmann's constant, T is Temperature, q is the charge of an electron, N_A is a doping parameter and N_i is the intrinsic doping parameter for the substrate.

We see that the surface potential has a direct relationship with the temperature. Looking above, that while the threshold voltage does not have a direct relationship but is not independent of the effects. On average this variation is between −4 mV/°C and −2 mV/°C depending on doping level. For a change of 30 °C this results in significant variation from the 500mV design parameter commonly used for the 90 nanometer technology node.

Chapter 11

Q Factor



The bandwidth, Δf, of a damped oscillator is shown on a graph of energy versus frequency. The Q factor of the damped oscillator, or filter, is $f_0 / \Delta f$. The higher the Q, the narrower and 'sharper' the peak is.

In physics and engineering the **quality factor** or **Q factor** is a dimensionless parameter that describes how under-damped an oscillator or resonator is, or equivalently, characterizes a resonator's bandwidth relative to its center frequency. Higher *Q* indicates a lower rate of energy loss relative to the stored energy of the oscillator; the oscillations die out more slowly. A pendulum suspended from a high-quality bearing, oscillating in air, has a high *Q*, while a pendulum immersed in oil has a low one. Oscillators with high quality factors have low damping so that they ring longer.

Sinusoidally driven resonators having higher Q factors resonate with greater amplitudes (at the resonant frequency) but have a smaller range of frequencies around that frequency for which they resonate; the range of frequencies for which the oscillator resonates is called the bandwidth. Thus, a high Q tuned circuit in a radio receiver would be more difficult to tune, but would have more selectivity; it would do a better job of filtering out signals from other stations that lie nearby on the spectrum. High Q oscillators oscillate with a smaller range of frequencies and are more stable.

The quality factor of oscillators varies substantially from system to system. Systems for which damping is important (such as dampers keeping a door from slamming shut) have $Q = ½$. Clocks, lasers, and other resonating systems that need either strong resonance or high frequency stability need high quality factors. Tuning forks have quality factors around $Q = 1000$. The quality factor of atomic clocks, superconducting RF cavities used in accelerators, and some high-Q lasers can reach as high as 10^{11} and higher.

There are many alternate quantities used by physicists and engineers to describe how damped an oscillator is and that are closely related to the quality factor. Important examples include: the damping ratio, relative bandwidth, linewidth and bandwidth measured in octaves.

The concept of Q factor originated in electronic engineering, as a measure of the 'quality' desired in a good tuned circuit or other resonator.

Definition of the quality factor

In the context of resonators, Q is defined in terms of the ratio of the energy stored in the resonator to the energy supplied by a generator, per cycle, to keep signal amplitude constant, at a frequency (the resonant frequency), f_r, where the stored energy is constant with time:

$$Q = 2\pi \times \frac{\text{Energy Stored}}{\text{Energy dissipated per cycle}} = 2\pi f_r \times \frac{\text{Energy Stored}}{\text{Power Loss}}.$$

The factor 2π makes Q expressible in simpler terms, involving only the coefficients of the second-order differential equation describing most resonant systems, electrical or mechanical. In electrical systems, the stored energy is the sum of energies stored in lossless inductors and capacitors; the lost energy is the sum of the energies dissipated in resistors per cycle. In mechanical systems, the stored energy is the sum of the potential and kinetic energies; the lost energy is the work done by an external conservative force, per cycle, to maintain amplitude.

For high values of Q, the following definition is also mathematically accurate:

$$Q = \frac{f_r}{\Delta f} = \frac{\omega_r}{\Delta \omega},$$

where f_r is the resonant frequency, Δf is the bandwidth, $\omega_r = 2\pi f_r$ is the angular resonant frequency, and $\Delta\omega$ is the angular bandwidth.

More generally and in the context of reactive component specification (especially inductors), the frequency-dependent definition of Q is used:

$$Q(\omega) = \omega \times \frac{\text{Maximum Energy Stored}}{\text{Power Loss}},$$

where ω is the angular frequency at which the stored energy and power loss are measured. This definition is consistent with its usage in describing circuits with a single reactive element (capacitor or inductor), where it can be shown to be equal to the ratio of reactive power to real power.

Q factor and damping

The Q factor determines the qualitative behavior of simple damped oscillators. (For mathematical details about these systems and their behavior see harmonic oscillator and linear time invariant (LTI) system.)

- A system with **low quality factor** ($Q < ½$) is said to be **overdamped.** Such a system doesn't oscillate at all, but when displaced from its equilibrium steady-state output it returns to it by exponential decay, approaching the steady state value asymptotically. It has an impulse response that is the sum of two decaying exponential functions with different rates of decay. As the quality factor decreases the slower decay mode becomes stronger relative to the faster mode and dominates the system's response resulting in a slower system. A second-order low-pass filter with a very low quality factor has a nearly first-order step response; the system's output responds to a step input by slowly rising toward an asymptote.

- A system with **high quality factor** ($Q > ½$) is said to be **underdamped.** Underdamped systems combine oscillation at a specific frequency with a decay of the amplitude of the signal. Underdamped systems with a low quality factor (a little above $Q = ½$) may oscillate only once or a few times before dying out. As the quality factor increases, the relative amount of damping decreases. A high-quality bell rings with a single pure tone for a very long time after being struck. A purely oscillatory system, such as a bell that rings forever, has an infinite quality factor. More generally, the output of a second-order low-pass filter with a very high quality factor responds to a step input by quickly rising above, oscillating around, and eventually converging to a steady-state value.

- A system with an **intermediate quality factor** ($Q = ½$) is said to be **critically damped.** Like an overdamped system, the output does not oscillate, and does not overshoot its steady-state output (i.e., it approaches a steady-state asymptote). Like an underdamped response, the output of such a system responds quickly to a

unit step input. Critical damping results in the fastest response (approach to the final value) possible without overshoot. Real system specifications usually allow some overshoot for a faster initial response or require a slower initial response to provide a safety margin against overshoot.

In negative feedback systems, the dominant closed-loop response is often well-modeled by a second-order system. The phase margin of the open-loop system sets the quality factor Q of the closed-loop system; as the phase margin decreases, the approximate second-order closed-loop system is made more oscillatory (i.e., has a higher quality factor).

Quality factors of common systems

- A unity gain Sallen–Key filter topology with equivalent capacitors and equivalent resistors is critically damped (i.e., $Q = 1 / 2$).
- A Butterworth filter (i.e., continuous-time filter with the flattest passband frequency response) has an underdamped $Q = 1/\sqrt{2}$.
- A Bessel filter (i.e., continuous-time filter with flattest group delay) has an underdamped $Q = 1/\sqrt{3}$.

Physical interpretation of Q

Physically speaking, Q is 2π times the ratio of the total energy stored divided by the energy lost in a single cycle or equivalently the ratio of the stored energy to the energy dissipated per one radian of the oscillation.

It is a dimensionless parameter that compares the time constant for decay of an oscillating physical system's amplitude to its oscillation period. Equivalently, it compares the frequency at which a system oscillates to the rate at which it dissipates its energy.

Equivalently (for large values of Q), the Q factor is approximately the number of oscillations required for a freely oscillating system's energy to fall off to $1 / e^{2\pi}$, or about 1/535, of its original energy.

The width (bandwidth) of the resonance is given by

$$\Delta f = \frac{f_0}{Q},$$

where f_0 is the resonant frequency, and Δf, the bandwidth, is the width of the range of frequencies for which the energy is at least half its peak value.

The factors Q, damping ratio ζ, and attenuation α are related such that

$$\zeta = \frac{1}{2Q} = \frac{\alpha}{\omega_0}.$$

So the quality factor can be expressed as

$$Q = \frac{1}{2\zeta} = \frac{\omega_0}{2\alpha},$$

and the exponential attenuation rate can be expressed as

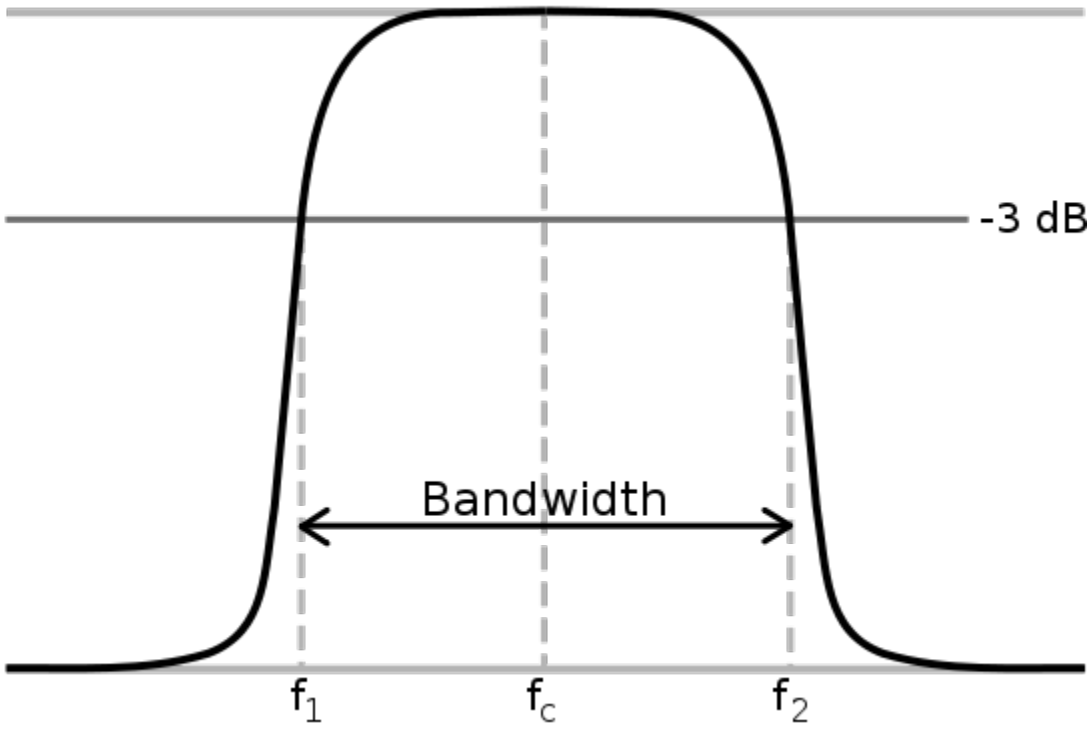
$$\alpha = \zeta\omega_0 = \frac{\omega_0}{2Q}.$$

For any 2nd order low-pass filter, the response function of the filter is

$$H(s) = \frac{\omega_0^2}{s^2 + \underbrace{\frac{\omega_0}{Q}}_{2\zeta\omega_0 = 2\alpha} s + \omega_0^2}$$

For this system, when $Q > 0.5$ (i.e., when the system is underdamped), it has two complex conjugate poles that each have a real part of α. That is, the attenuation parameter α represents the rate of exponential decay of the oscillations (e.g., after an impulse) of the system. A higher quality factor implies a lower attenuation, and so high Q systems oscillate for long times. For example, high quality bells have an approximately pure sinusoidal tone for a long time after being struck by a hammer.

Electrical systems



A graph of a filter's gain magnitude, illustrating the concept of -3 dB at a gain of 0.707 or half-power bandwidth. The frequency axis of this symbolic diagram can be linear or logarithmically scaled.

For an electrically resonant system, the *Q* factor represents the effect of electrical resistance and, for electromechanical resonators such as quartz crystals, mechanical friction.

RLC circuits

In an ideal series RLC circuit, and in a tuned radio frequency receiver (TRF) the *Q* factor is:

$$Q = \frac{1}{R}\sqrt{\frac{L}{C}},$$

where *R*, *L* and *C* are the resistance, inductance and capacitance of the tuned circuit, respectively. For a parallel RLC circuit, the Q factor is the inverse of the series case:

$$Q = R\sqrt{\frac{C}{L}},$$

Consider a circuit where R, L and C are all in parallel. The lower the parallel resistance, the more effect it will have in damping the circuit and thus the lower the Q. In this case the X and R are interchanged. This is useful in filter design to determine the bandwidth.

In a parallel LC circuit where the main loss is the R in series with the L, *Q* is as in the series circuit. This is a common circumstance for resonators, where limiting the resistance of the inductor to improve Q and narrow the bandwidth is the desired result.

Complex impedances

For a complex impedance

$$\tilde{Z} = R + j\mathrm{X}$$

the *Q* factor is the ratio of the reactance to the resistance (or equivalently, the absolute value of the ratio of reactive power to real power), that is:

$$Q = \left|\frac{\mathrm{X}}{R}\right|$$

Thus, one can also calculate the *Q* factor for a complex impedance by knowing just the power factor of the circuit

$$Q = \frac{|\sin\phi|}{|\cos\phi|} = \frac{\sqrt{1 - PF^2}}{PF} = \sqrt{\frac{1}{PF^2} - 1}$$

or just the tangent of the phase angle

$$Q = |\tan\phi|$$

where φ is the phase angle and *PF* is the power factor of the circuit.

Mechanical systems

For a single damped mass-spring system, the *Q* factor represents the effect of simplified viscous damping or drag, where the damping force or drag force is proportional to velocity. The formula for the Q factor is:

$$Q = \frac{\sqrt{Mk}}{D},$$

where M is the mass, k is the spring constant, and D is the damping coefficient, defined by the equation $F_{\text{damping}} = -Dv$, where v is the velocity.

Optical systems

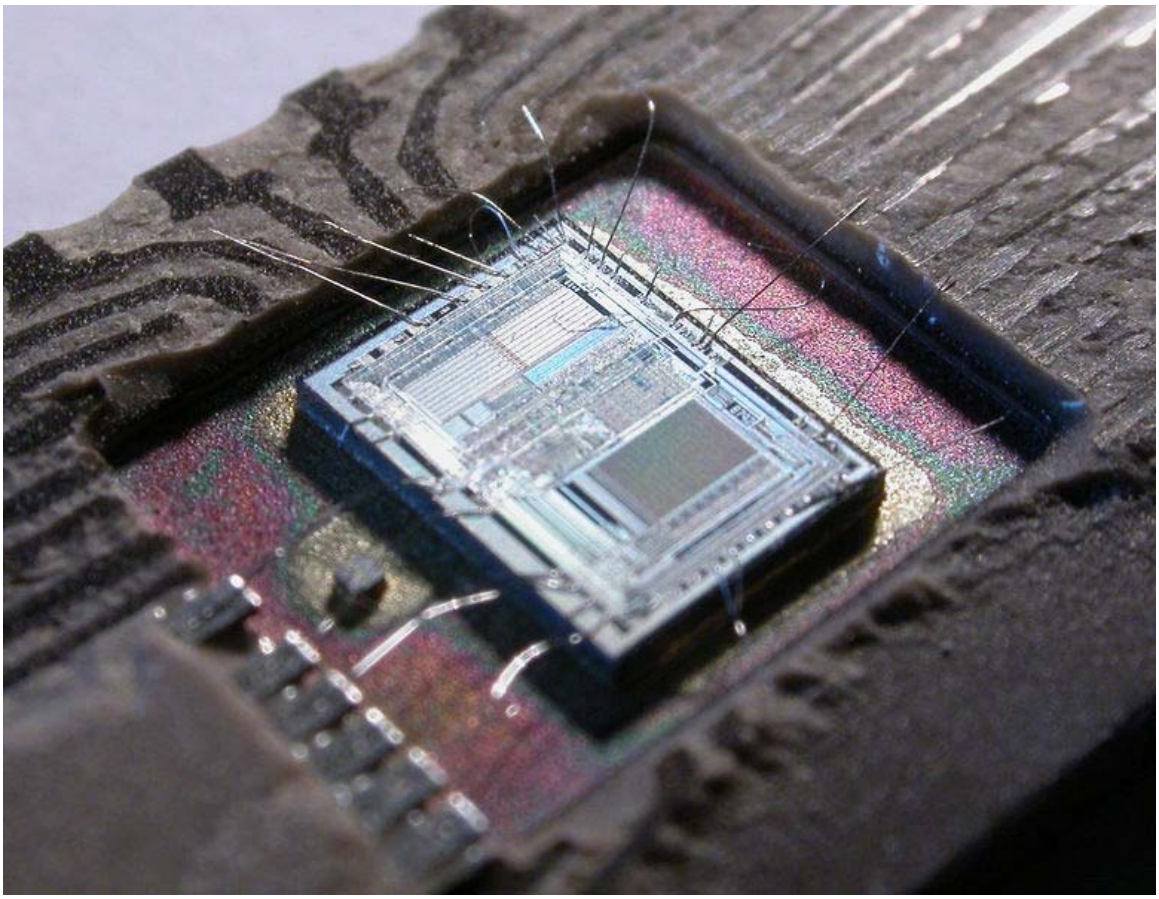
In optics, the *Q* factor of a resonant cavity is given by

$$Q = \frac{2\pi f_o \mathcal{E}}{P},$$

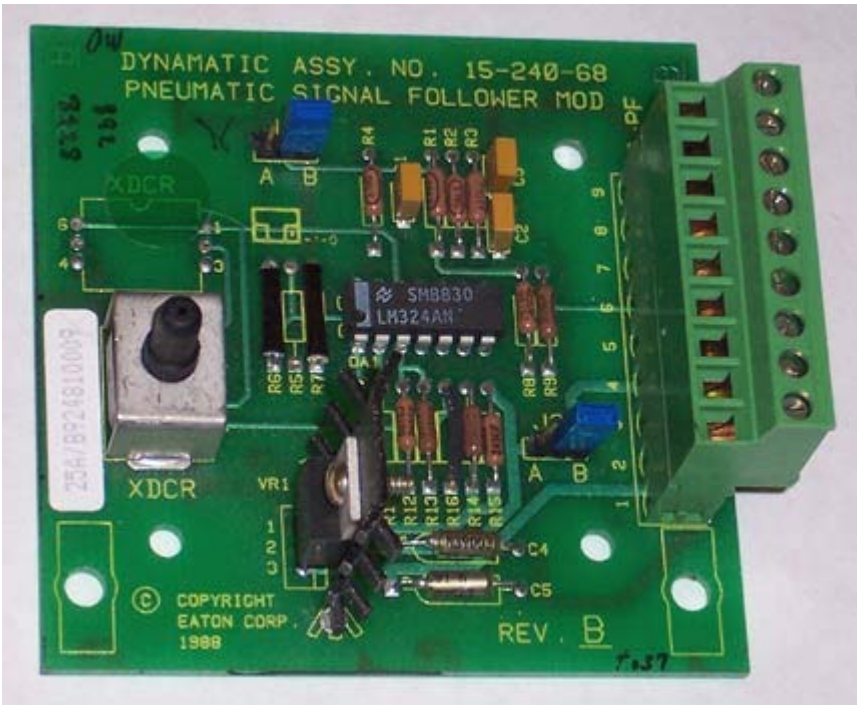
where f_o is the resonant frequency, $\mathcal{E}$ is the stored energy in the cavity, and $P = -\frac{dE}{dt}$ is the power dissipated. The optical *Q* is equal to the ratio of the resonant frequency to the bandwidth of the cavity resonance. The average lifetime of a resonant photon in the cavity is proportional to the cavity's *Q*. If the *Q* factor of a laser's cavity is abruptly changed from a low value to a high one, the laser will emit a pulse of light that is much more intense than the laser's normal continuous output. This technique is known as Q-switching.

Chapter 12

Electronic Circuit



The die from an Intel 8742, an 8-bit microcontroller that includes a CPU, 128 bytes of RAM, 2048 bytes of EPROM, and I/O in the same chip.



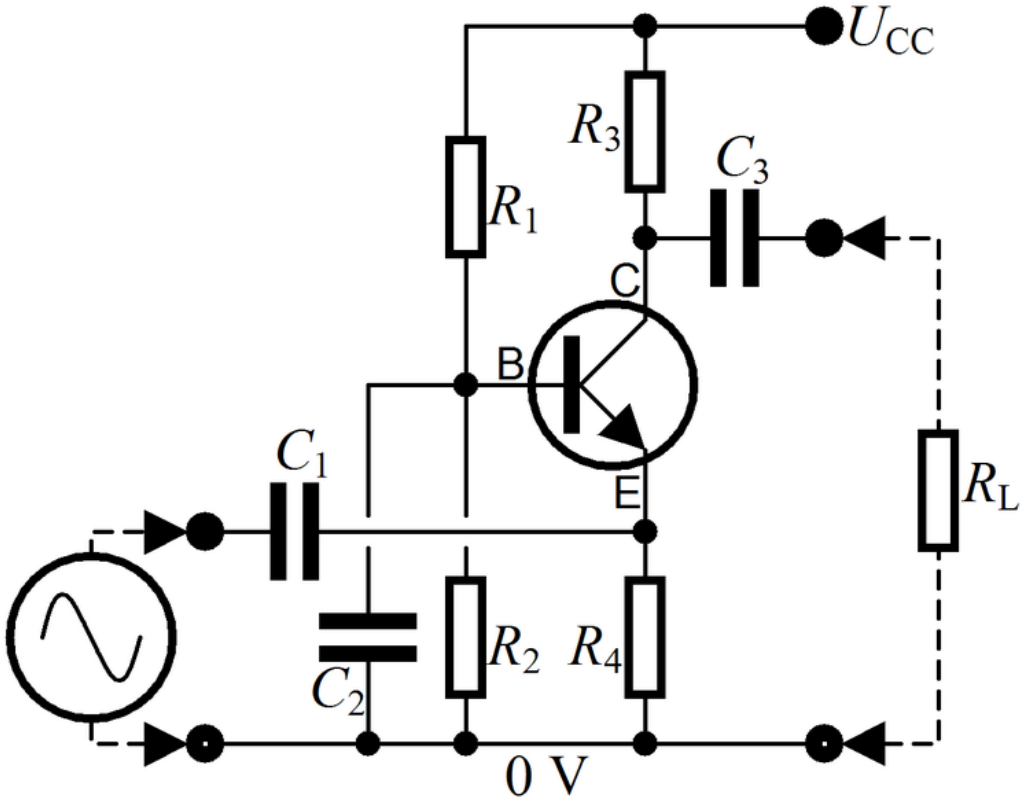
A circuit built on a printed circuit board (PCB).

An **electronic circuit** is composed of individual electronic components, such as resistors, transistors, capacitors, inductors and diodes, connected by conductive wires or traces through which electric current can flow. The combination of components and wires allows various simple and complex operations to be performed: signals can be amplified, computations can be performed, and data can be moved from one place to another. Circuits can be constructed of discrete components connected by individual pieces of wire, but today it is much more common to create interconnections by photolithographic techniques on a laminated substrate (a printed circuit board or PCB) and solder the components to these interconnections to create a finished circuit. In an Integrated Circuit or IC, the components and interconnections are formed on the same substrate, typically a semiconductor such as silicon or (less commonly) gallium arsenide.

Breadboards, perfboards or stripboards are common for testing new designs. They allow the designer to make quick changes to the circuit during development.

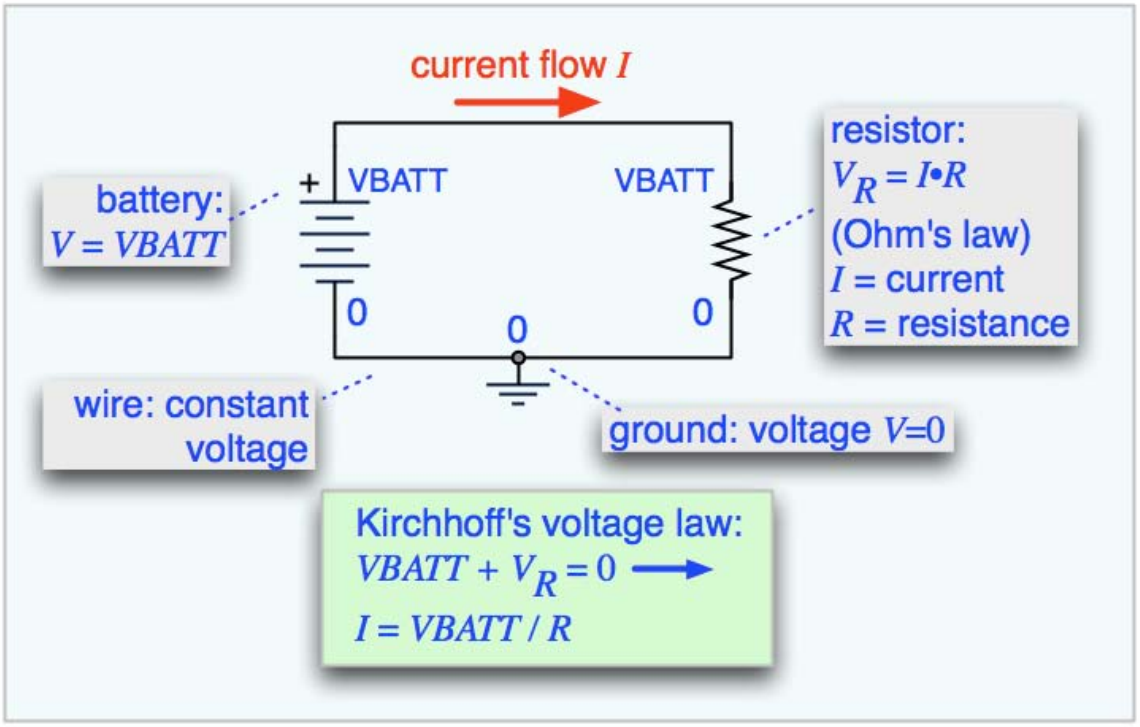
An electronic circuit can usually be categorized as an analog circuit, a digital circuit or a mixed-signal circuit (a combination of analog circuits and digital circuits).

Analog circuits



A circuit diagram representing an analog circuit, in this case a simple amplifier.

Analog electronic circuits are those in which current or voltage may vary continuously with time to correspond to the information being represented. Analog circuitry is constructed from two fundamental building blocks: series and parallel circuits. In a series circuit, the same current passes through a series of components. A string of Christmas lights is a good example of a series circuit: if one goes out, they all do. In a parallel circuit, all the components are connected to the same voltage, and the current divides between the various components according to their resistance.



A simple schematic showing wires, a resistor, and a battery.

The basic components of analog circuits are wires, resistors, capacitors, inductors, diodes, and transistors. (Recently, memristors have been added to the list of available components.) Analog circuits are very commonly represented in schematic diagrams, in which wires are shown as lines, and each component has a unique symbol. Analog circuit analysis employs Kirchhoff's circuit laws: all the currents at a node (a place where wires meet) must add to 0, and the voltage around a closed loop of wires is 0. Wires are usually treated as ideal zero-voltage interconnections; any resistance or reactance is captured by explicitly adding a parasitic element, such as a discrete resistor or inductor. Active components such as transistors are often treated as controlled current or voltage sources: for example, a field-effect transistor can be modeled as a current source from the source to the drain, with the current controlled by the gate-source voltage.

When the circuit size is comparable to a wavelength of the relevant signal frequency, a more sophisticated approach must be used. Wires are treated as transmission lines, with (hopefully) constant characteristic impedance, and the impedances at the start and end determine transmitted and reflected waves on the line. Such considerations typically become important for circuit boards at frequencies above a GHz; integrated circuits are smaller and can be treated as lumped elements for frequencies less than 10 GHz or so.

An alternative model is to take independent power sources and induction as basic electronic units; this allows modeling frequency dependent negative resistors, gyrators,

negative impedance converters, and dependent sources as secondary electronic components.

Digital circuits

In digital electronic circuits, electric signals take on discrete values, to represent logical and numeric values. These values represent the information that is being processed. In the vast majority of cases, binary encoding is used: one voltage (typically the more positive value) represents a binary '1' and another voltage (usually a value near the ground potential, 0 V) represents a binary '0'. Digital circuits make extensive use of transistors, interconnected to create logic gates that provide the functions of Boolean logic: AND, OR, NOT, and all possible combinations there of. Transistors interconnected so as to provide positive feedback are used as latches and flip flops, circuits that have two or more metastable states, and remain in one of these states until changed by an external input. Digital circuits therefore can provide both logic and memory, enabling them to perform arbitrary computational functions. (Memory based on flip-flops is known as SRAM (static random access memory). Memory based on the storage of charge in a capacitor, DRAM (dynamic random access memory) is also widely used.)

Digital circuits are fundamentally easier to design than analog circuits for the same level of complexity, because each logic gate regenerates the binary signal, so the designer need not account for distortion, gain control, offset voltages, and other concerns faced in an analog design. As a consequence, extremely complex digital circuits, with billions of logic elements integrated on a single silicon chip, can be fabricated at low cost. Such digital integrated circuits are ubiquitous in modern electronic devices, such as calculators, mobile phone handsets, and computers.

Digital circuitry is used to create general purpose computing chips, such as microprocessors, and custom-designed logic circuits, known as Application Specific Integrated Circuits (ASICs). Field Programmable Gate Arrays (FPGAs), chips with logic circuitry whose configuration can be modified after fabrication, are also widely used in prototyping and development.

Mixed-signal circuits

Mixed-signal or hybrid circuits contain elements of both analog and digital circuits. Examples include comparators, timers, PLLs, ADCs (analog-to-digital converters), and DACs (digital-to-analog converters). Most modern radio and communications circuitry uses mixed signal circuits. For example, in a receiver, analog circuitry is used to amplify and frequency-convert signals so that they reach a suitable state to be converted into digital values, after which further signal processing can be performed in the digital domain.

Chapter 13

Semiconductor Device

Semiconductor devices are electronic components that exploit the electronic properties of semiconductor materials, principally silicon, germanium, and gallium arsenide, as well as organic semiconductors. Semiconductor devices have replaced thermionic devices (vacuum tubes) in most applications. They use electronic conduction in the solid state as opposed to the gaseous state or thermionic emission in a high vacuum.

Semiconductor devices are manufactured both as single discrete devices and as *integrated circuits* (ICs), which consist of a number—from a few (as low as two) to billions—of devices manufactured and interconnected on a single semiconductor substrate.

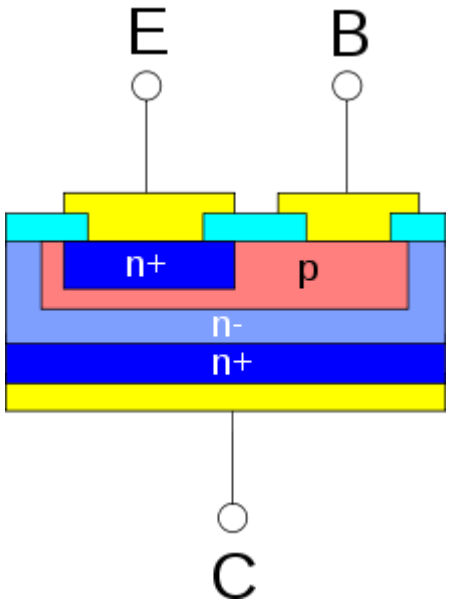
The main reason why semiconductor materials are so useful is that the behavior of a semiconductor can be easily manipulated by the addition of impurities, known as doping. Semiconductor conductivity can be controlled by introduction of an electric field, by exposure to light, and even pressure and heat; thus, semiconductors can make excellent sensors. Current conduction in a semiconductor occurs via mobile or "free" *electrons* and *holes*, collectively known as *charge carriers*. Doping a semiconductor such as silicon with a small amount of impurity atoms, such as phosphorus or boron, greatly increases the number of free electrons or holes within the semiconductor. When a doped semiconductor contains excess holes it is called "p-type", and when it contains excess free electrons it is known as "n-type", where *p* (positive for holes) or *n* (negative for electrons) is the sign of the charge of the majority mobile charge carriers. The semiconductor material used in devices is doped under highly controlled conditions in a fabrication facility, or *fab*, to precisely control the location and concentration of p- and n-type dopants. The junctions which form where n-type and p-type semiconductors join together are called p-n junctions.

Diode

The diode is a device made from a single p-n junction. At the junction of a p-type and an n-type semiconductor there forms a region called the depletion zone which blocks current conduction from the n-type region to the p-type region, but allows current to conduct from the p-type region to the n-type region. Thus when the device is *forward biased*, with the p-side at higher electric potential, the diode conducts current easily; but the current is very small when the diode is *reverse biased*.

Exposing a semiconductor to light can generate electron–hole pairs, which increases the number of free carriers and its conductivity. Diodes optimized to take advantage of this phenomenon are known as *photodiodes*. Compound semiconductor diodes can also be used to generate light, as in light-emitting diodes and laser diodes.

Transistor



An NPN bipolar junction transistor structure

Bipolar junction transistors are formed from two p-n junctions, in either n-p-n or p-n-p configuration. The middle, or *base*, region between the junctions is typically very narrow. The other regions, and their associated terminals, are known as the *emitter* and the *collector*. A small current injected through the junction between the base and the emitter changes the properties of the base-collector junction so that it can conduct current even though it is reverse biased. This creates a much larger current between the collector and emitter, controlled by the base-emitter current.

Another type of transistor, the field effect transistor operates on the principle that semiconductor conductivity can be increased or decreased by the presence of an electric field. An electric field can increase the number of free electrons and holes in a semiconductor, thereby changing its conductivity. The field may be applied by a reverse-

biased p-n junction, forming a *junction field effect transistor*, or JFET; or by an electrode isolated from the bulk material by an oxide layer, forming a *metal-oxide-semiconductor field effect transistor*, or MOSFET.

The MOSFET is the most used semiconductor device today. The *gate* electrode is charged to produce an electric field that controls the conductivity of a "channel" between two terminals, called the *source* and *drain*. Depending on the type of carrier in the channel, the device may be an *n-channel* (for electrons) or a *p-channel* (for holes) MOSFET. Although the MOSFET is named in part for its "metal" gate, in modern devices polysilicon is typically used instead. MOSFET is an IC which is semiconductor device.

Semiconductor device materials

By far, silicon (Si) is the most widely used material in semiconductor devices. Its combination of low raw material cost, relatively simple processing, and a useful temperature range make it currently the best compromise among the various competing materials. Silicon used in semiconductor device manufacturing is currently fabricated into boules that are large enough in diameter to allow the production of 300 mm (12 in.) wafers.

Germanium (Ge) was a widely used early semiconductor material but its thermal sensitivity makes it less useful than silicon. Today, germanium is often alloyed with silicon for use in very-high-speed SiGe devices; IBM is a major producer of such devices.

Gallium arsenide (GaAs) is also widely used in high-speed devices but so far, it has been difficult to form large-diameter boules of this material, limiting the wafer diameter to sizes significantly smaller than silicon wafers thus making mass production of GaAs devices significantly more expensive than silicon.

Other less common materials are also in use or under investigation.

Silicon carbide (SiC) has found some application as the raw material for blue light-emitting diodes (LEDs) and is being investigated for use in semiconductor devices that could withstand very high operating temperatures and environments with the presence of significant levels of ionizing radiation. IMPATT diodes have also been fabricated from SiC.

Various indium compounds (indium arsenide, indium antimonide, and indium phosphide) are also being used in LEDs and solid state laser diodes. Selenium sulfide is being studied in the manufacture of photovoltaic solar cells.

The most common use for organic semiconductors is Organic light-emitting diodes.

List of common semiconductor devices

Two-terminal devices:

- DIAC
- Diode (rectifier diode)
- Gunn diode
- IMPATT diode
- Laser diode
- Light-emitting diode (LED)
- Photocell
- PIN diode
- Schottky diode
- Solar cell
- Tunnel diode
- VCSEL
- VECSEL
- Zener diode

Three-terminal devices:

- Bipolar transistor
- Darlington transistor
- Field effect transistor
- GTO (Gate Turn-Off)
- IGBT (Insulated Gate Bipolar Transistor)
- SCR (Silicon Controlled Rectifier)
- SGCT (Switched Gate Commuted Thyristor)
- Thyristor
- TRIAC
- Unijunction transistor

Four-terminal devices:

- Hall effect sensor (magnetic field sensor)

Multi-terminal devices:

- Integrated Circuit (ICs)
- Charge-coupled device (CCD)
- Microprocessor
- Random Access Memory (RAM)
- Read-only memory (ROM)

Semiconductor device applications

All transistor types can be used as the building blocks of logic gates, which are fundamental in the design of digital circuits. In digital circuits like microprocessors, transistors act as on-off switches; in the MOSFET, for instance, the voltage applied to the gate determines whether the switch is on or off.

Transistors used for analog circuits do not act as on-off switches; rather, they respond to a continuous range of inputs with a continuous range of outputs. Common analog circuits include amplifiers and oscillators.

Circuits that interface or translate between digital circuits and analog circuits are known as mixed-signal circuits.

Power semiconductor devices are discrete devices or integrated circuits intended for high current or high voltage applications. Power integrated circuits combine IC technology with power semiconductor technology, these are sometimes referred to as "smart" power devices. Several companies specialize in manufacturing power semiconductors.

Component identifiers

The type designators of semiconductor devices are often manufacturer specific. Nevertheless, there have been attempts at creating standards for type codes, and a subset of devices follow those. For discrete devices, for example, there are three standards: JEDEC JESD370B in USA, Pro Electron in Europe and JIS in Japan.

History of semiconductor device development

Cat's-whisker detector

Semiconductors had been used in the electronics field for some time before the invention of the transistor. Around the turn of the 20th century they were quite common as detectors in radios, used in a device called a "cat's whisker". These detectors were somewhat troublesome, however, requiring the operator to move a small tungsten filament (the whisker) around the surface of a galena (lead sulfide) or carborundum (silicon carbide) crystal until it suddenly started working. Then, over a period of a few hours or days, the cat's whisker would slowly stop working and the process would have to be repeated. At the time their operation was completely mysterious. After the introduction of the more reliable and amplified vacuum tube based radios, the cat's whisker systems quickly disappeared. The "cat's whisker" is a primitive example of a special type of diode still popular today, called a Schottky diode.

Metal rectifier

Another early type of semiconductor device is the metal rectifier in which the semiconductor is copper oxide or selenium. Westinghouse Electric (1886) was a major manufacturer of these rectifiers.

World War II

During World War II, radar research quickly pushed radar receivers to operate at ever higher frequencies and the traditional tube based radio receivers no longer worked well. The introduction of the cavity magnetron from Britain to the United States in 1940 during the Tizard Mission resulted in a pressing need for a practical high-frequency amplifier.

On a whim, Russell Ohl of Bell Laboratories decided to try a cat's whisker. By this point they had not been in use for a number of years, and no one at the labs had one. After hunting one down at a used radio store in Manhattan, he found that it worked much better than tube-based systems.

Ohl investigated why the cat's whisker functioned so well. He spent most of 1939 trying to grow more pure versions of the crystals. He soon found that with higher quality crystals their finicky behaviour went away, but so did their ability to operate as a radio detector. One day he found one of his purest crystals nevertheless worked well, and interestingly, it had a clearly visible crack near the middle. However as he moved about the room trying to test it, the detector would mysteriously work, and then stop again. After some study he found that the behaviour was controlled by the light in the room– more light caused more conductance in the crystal. He invited several other people to see this crystal, and Walter Brattain immediately realized there was some sort of junction at the crack.

Further research cleared up the remaining mystery. The crystal had cracked because either side contained very slightly different amounts of the impurities Ohl could not remove–about 0.2%. One side of the crystal had impurities that added extra electrons (the carriers of electrical current) and made it a "conductor". The other had impurities that wanted to bind to these electrons, making it (what he called) an "insulator". Because the two parts of the crystal were in contact with each other, the electrons could be pushed out of the conductive side which had extra electrons (soon to be known as the *emitter*) and replaced by new ones being provided (from a battery, for instance) where they would flow into the insulating portion and be collected by the whisker filament (named the *collector*). However, when the voltage was reversed the electrons being pushed into the collector would quickly fill up the "holes" (the electron-needy impurities), and conduction would stop almost instantly. This junction of the two crystals (or parts of one crystal) created a solid-state diode, and the concept soon became known as semiconduction. The mechanism of action when the diode is off has to do with the separation of charge carriers around the junction. This is called a "depletion region".

Development of the diode

Armed with the knowledge of how these new diodes worked, a vigorous effort began to learn how to build them on demand. Teams at Purdue University, Bell Labs, MIT, and the University of Chicago all joined forces to build better crystals. Within a year germanium production had been perfected to the point where military-grade diodes were being used in most radar sets.

Development of the transistor

After the war, William Shockley decided to attempt the building of a triode-like semiconductor device. He secured funding and lab space, and went to work on the problem with Brattain and John Bardeen.

The key to the development of the transistor was the further understanding of the process of the electron mobility in a semiconductor. It was realized that if there was some way to control the flow of the electrons from the emitter to the collector of this newly discovered diode, one could build an amplifier. For instance, if you placed contacts on either side of a single type of crystal the current would not flow through it. However if a third contact could then "inject" electrons or holes into the material, the current would flow.

Actually doing this appeared to be very difficult. If the crystal were of any reasonable size, the number of electrons (or holes) required to be injected would have to be very large -– making it less than useful as an amplifier because it would require a large injection current to start with. That said, the whole idea of the crystal diode was that the crystal itself could provide the electrons over a very small distance, the depletion region. The key appeared to be to place the input and output contacts very close together on the surface of the crystal on either side of this region.

Brattain started working on building such a device, and tantalizing hints of amplification continued to appear as the team worked on the problem. Sometimes the system would work but then stop working unexpectedly. In one instance a non-working system started working when placed in water. Ohl and Brattain eventually developed a new branch of quantum mechanics known as surface physics to account for the behaviour. The electrons in any one piece of the crystal would migrate about due to nearby charges. Electrons in the emitters, or the "holes" in the collectors, would cluster at the surface of the crystal where they could find their opposite charge "floating around" in the air (or water). Yet they could be pushed away from the surface with the application of a small amount of charge from any other location on the crystal. Instead of needing a large supply of injected electrons, a very small number in the right place on the crystal would accomplish the same thing.

Their understanding solved the problem of needing a very small control area to some degree. Instead of needing two separate semiconductors connected by a common, but tiny, region, a single larger surface would serve. The emitter and collector leads would both be placed very close together on the top, with the control lead placed on the base of

the crystal. When current was applied to the "base" lead, the electrons or holes would be pushed out, across the block of semiconductor, and collect on the far surface. As long as the emitter and collector were very close together, this should allow enough electrons or holes between them to allow conduction to start.

The first transistor

A stylized replica of the first transistor

The Bell team made many attempts to build such a system with various tools, but generally failed. Setups where the contacts were close enough were invariably as fragile as the original cat's whisker detectors had been, and would work briefly, if at all. Eventually they had a practical breakthrough. A piece of gold foil was glued to the edge of a plastic wedge, and then the foil was sliced with a razor at the tip of the triangle. The result was two very closely spaced contacts of gold. When the plastic was pushed down onto the surface of a crystal and voltage applied to the other side (on the base of the crystal), current started to flow from one contact to the other as the base voltage pushed the electrons away from the base towards the other side near the contacts. The point-contact transistor had been invented.

While the device was constructed a week earlier, Brattain's notes describe the first demonstration to higher-ups at Bell Labs on the afternoon of 23 December 1947, often

given as the birthdate of the transistor. The "PNP point-contact germanium transistor" operated as a speech amplifier with a power gain of 18 in that trial. Known generally as a point-contact transistor today, John Bardeen, Walter Houser Brattain, and William Bradford Shockley were awarded the Nobel Prize in physics for their work in 1956.

Origin of the term "transistor"

Bell Telephone Laboratories needed a generic name for their new invention: "Semiconductor Triode", "Solid Triode", "Surface States Triode" [sic], "Crystal Triode" and "Iotatron" were all considered, but "transistor", coined by John R. Pierce, won an internal ballot. The rationale for the name is described in the following extract from the company's Technical Memoranda (May 28, 1948) calling for votes:

Transistor. This is an abbreviated combination of the words "transconductance" or "transfer", and "varistor". The device logically belongs in the varistor family, and has the transconductance or transfer impedance of a device having gain, so that this combination is descriptive.

Improvements in transistor design

Shockley was upset about the device being credited to Brattain and Bardeen, who he felt had built it "behind his back" to take the glory. Matters became worse when Bell Labs lawyers found that some of Shockley's own writings on the transistor were close enough to those of an earlier 1925 patent by Julius Edgar Lilienfeld that they thought it best that his name be left off the patent application.

Shockley was incensed, and decided to demonstrate who was the real brains of the operation. Only a few months later he invented an entirely new type of transistor with a layer or 'sandwich' structure. This new form was considerably more robust than the fragile point-contact system, and would go on to be used for the vast majority of all transistors into the 1960s. It would evolve into the bipolar junction transistor.

With the fragility problems solved, a remaining problem was purity. Making germanium of the required purity was proving to be a serious problem, and limited the number of transistors that actually worked from a given batch of material. Germanium's sensitivity to temperature also limited its usefulness. Scientists theorized that silicon would be easier to fabricate, but few bothered to investigate this possibility. Gordon K. Teal was the first to develop a working silicon transistor, and his company, the nascent Texas Instruments, profited from its technological edge. Germanium disappeared from most transistors by the late 1960s.

Within a few years, transistor-based products, most notably radios, were appearing on the market. A major improvement in manufacturing yield came when a chemist advised the companies fabricating semiconductors to use distilled water rather than tap water: calcium ions were the cause of the poor yields. "Zone melting", a technique using a

moving band of molten material through the crystal, further increased the purity of the available crystals.