# Electronic Instrumentation

Peter Stallinga

peter@stallinga.org

Electronic Instrumentation (paperback)
Peter Stallinga
v. 1.1 (Thursday 12th September, 2024)

ISBN: 978-1-71659-398-7

Typefaces: Times Roman, Helvetica, Free Sans
Typesetting: LaTeX2e with TexLive in Texmaker
Graphical: Inkscape and PjotrSoft. All pictures made by the author.

stallinga.org
Non-profit science organization

# Preface

This book is a written out version of the lecture notes used for the lectures of Electronic Instrumentation at The University of The Algarve, which is a third-year undergrad discipline. It also includes the other lectures of general physics, physics of electronics, and electronics for reference.

As much as possible an Open Source character of the lectures is maintained, with as many products used being of this type as possible. In fact, the practical lectures were often based on recycled electronics parts. For instance, the stepper motors used in a practical lectures in which the students have to make a motor were recovered from a broken hard disk. This approach makes everything presented here cheap, easy and readily available all over the world. Basically, only one exception was made, namely the ZigBee wireless units, which therefore immediately turned out to be very difficult to use and quite expensive.

This philosophy was also used in preparing this document. For that reason I would like to thank the makers of the software products used: Ubuntu and Mint (Linux distributions), Texmaker (LaTeXCompiler), Inkscape (vector graphics), M4 macros for electric circuit diagrams (draw electronic circuits, Dwight Aplevich) ImageMagick (graphics converter), Google (search engine), Wikipedia (on-line encyclopedia wiki).

My thanks also go to my colleague Igor Khmelinskii who had the patience to proofread everything. As well as all the students that passed through the lectures, the feedback over the years helped crystallize the ideas in the format you now see in front of you.

P. S.

Faro, Portugal
September, 2024

# 0 Contents

# 1 | General concepts

## 1.1 Introduction

Generally speaking, the art of Electronic Instrumentation can be described as all the steps of extracting information from the environment, either for just obtaining, processing and storing this information, or even in order to control it. One has to think here about the classical example of the temperature of a system.

We can, for instance, have the need to monitor this temperature over time and store this information on a computer and communicate it to the rest of the world by means of Internet. Especially relevant for the monitoring of the climate are such weather-monitoring measurement-huts, that apart from the temperature also measure the humidity, the dew point, the cloud cover, the amount of rainfall, the type of precipitation, etc. It is obvious that many parameters of the weather are important and this information has to be extracted somehow from the system.

In a more advanced system, for instance the conditions in a room, the parameters not only monitored, but also controlled. We can think here of a heating system that keeps the temperature in a room at a comfortable 20 degrees on a cold winter day. Other, more professional systems can include the stabilizing of temperature and humidity in a greenhouse where tomatoes are grown.

This book deals with all the aspects of the information gathering, processing and acquisition steps needed to make such systems. It consists basically of three steps

1. **Physics**. In the first step, the information that resides in the physical world has to be translated somehow to the electronics world. We can think here of temperature sensors, for example a resistor whose resistance value depends on the temperature. Then, knowing the resistance implies knowing the temperature.

2. **Electronics**. While in this level the signal is already in electronic format, its range may be inadequate, or of the wrong type. Think of a thermocouple whose output voltage is in the order of some milli-volts at best, or

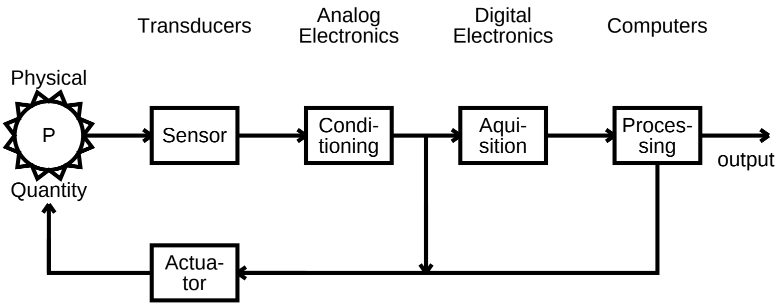


**Fig. 1.1**: Example of an electronic instrumentation system. A physical quantity (for instance temperature) is measured by a transducer sensor. Signal conditioning (amplification, filtering, off-set) takes places in the electronic world where also a decision can be taken about changing the physical quantity (or another) by a transducer actuator. Alternatively, the electronic signal can be converted to a digital signal and processing and decision making (partly) executed on a digital equipment such as a computer, or output generated for humans.

the temperature is translated to a resistance value in a thermistor while our ADC requires voltages.

3. **Informatics**. This level entails the translation of the analog electronic signal to a digital signal and the communication of the signal to or from a computer for further processing.

The book is organized in this way, with each aspect presented in a chapter, with the physics part divided into a scientific (pure physics) and engineering (applied physics) part: Electronics (Chapter 2), Physics (Chapter 3), Sensors & Actuators (Chapter 4) and Informatics (Chapter 5).

## 1.2 Signal generation; Transducers

The official meaning of the word ’transducer’ is: an element that converts one type of energy to another.In the framework of (electronic) instrumentation, the following definition is more adequate:

A transducer is an element that translates information from one physical domain to another

The magic keyword here is ’information’. Somehow, the information in one domain is translated into another domain. Two types of this are sensors and actuators which in the context of this book translate physical information to and from the electronic (analog or digital) domain, respectively. As a good example take the temperature sensor which translates temperature (unit: K) to electrical voltage (V) in a thermocouple or to current (A) in a semiconductor diode. The opposite direction also exists, a heating element is an actuator that does not directly translate an electric power into temperature, but into heat or heating power (W) that in turn will change the physical parameter temperature. Note that actuators that directly convert an electrical signal into temperature do not exist. In this way, in a closed system the temperature can be measured, monitored, and controlled.

In terms of signal processing we can see the sensors as transducer input elements 'reading' the state and actuators as output elements, 'writing' the state. A sensor translates from the physical domain to the electronic domain, while actuators do the opposite.

## 1.3 Parameters of sensors and systems

Sensors come in all sizes and prices. For some applications the cheapest sensors can be sufficient while in other systems we need high quality information and this implies better sensors with higher quality which are inherently also more expensive.

Imagine the sensor for measuring the temperature in a room to be used in an air-conditioning system. This is a simple on/off system; we want to switch on the air-conditioning when the temperature rises above a certain preset level. The absolute measurement value of the sensor ('accuracy') is not so relevant. The user will know "in *this* room I should put the temperature at 26 degrees to be comfortable", i.e., variations between sensors are not bothering us. At least not enough to justify an expensive sensor. The signal processing is simple, just on and off, this demands little in terms of linearity of the sensor. Moreover, the range at which the sensor should work is also limited, maybe 10 degrees at best. All in all, a simple thermistor probably suffices. Thermistors are little more expensive than resistors and for sure not a cost-determining element.

On the other hand, imagine a system for measuring the temperature in a scientific instrument for determining superconductivity. The temperature has to be accurately measured (for the lowest temperatures, close to the absolute zero, the accuracy should easily be in the milli-kelvin range). Moreover, it should have a high reproducibility. Every time we measure the temperature we would like to get the same temperature. Further limitations, such as the absence of effects of magnetic fields on the returned value, or the size can further increase the price. A ruthenium-oxide diode, used for scientific-grade low-temperature measurements can easily cost thousands of euros.

We can thus make a list of final system parameters that we should take into consideration when choosing the sensors for our system.

- **Price**. The most obvious is the price. We always aim to design the cheapest system. It does not make sense to use a ruthenium-oxide diode for an air-conditioning system. In general, for commercial systems, such as

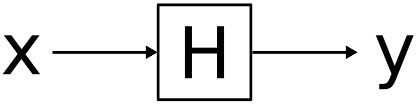


**Fig. 1.2**: The transfer function $H$ connects the input value to an output value of the transducer, $y = H(x)$

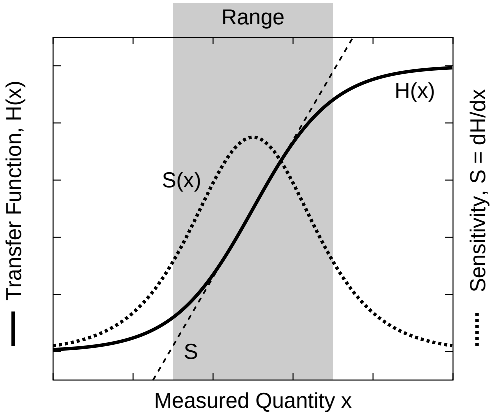


**Fig. 1.3**: Typical transfer function ($H$, solid line) of a sensor or system measuring quantity $X$. The sensitivity $S$ (dashed line) is the derivative of the transfer function. Ideally it is constant over a long range. In practice the transfer function is saturating at both ends and the sensitivity typically drops below an acceptable level outside the range, as indicated

those in the automotive industry, the price is the single-most-important parameter. Except maybe for some high-end luxury cars, if something cannot be measured cheaply then better not measure it at all.

The most important 'electronic' parameters are concerning the magnitude and quality of the signal generated:

- **Transfer function**. When the sensor connects one domain to the other, the response or transfer function $H$ is the function that describes that connection, see Figure 1.2. A Hall sensor that translates magnetic field to voltage has a transfer function $V(B)$ that describes how the output voltage depends on the input magnetic field. This can be a complex function. Obviously, for the sensor to work correctly, the function should have an inverse (the function should be a one-to-one relation), otherwise the obtained function value (for instance the voltage) is not enough to know the measured value (for instance the magnetic field). Other important parameters of the transfer function are sensitivity and linearity. See Figure 1.3.

- **Sensitivity and offset**. The sensitivity of the sensor is the derivative of the 'response' or 'transfer' function,

  $$S \equiv \frac{\mathrm{d}H(x)}{\mathrm{d}x}. \tag{1.1}$$

  For example, if a resistor has a resistance as a function of temperature given by $R(T)$ then the sensitivity is the value $S = \mathrm{d}R/\mathrm{d}T$. Since the transfer function in general is a non-linear function, the sensitivity depends on the point of operation, see Fig. 1.3.

  The sensitivity is probably the most important parameter of the sensor or system. This is the one that links the two worlds, something that is also visible in the units of this sensitivity. This unit consequently represents the final quantity divided by the original quantity. An electronic wind meter transducer, with an electronic output in current, will have a calibration factor of sensitivity with units '(amperes) per (meters per second)', A s/m. This is useful to bear in mind since it will give us a way to check our calculations; if the units are not correct, we did something wrong. See the section dedicated to units.

  The offset is given as the output value when the sensor is in its 'calibration point'. In many cases we want this offset to be zero. A temperature sensor giving 0 volt output when the temperature is equal to the nominal temperature. As we will see in the chapter on electronics (Chapter 2) this can easily be achieved by electronic techniques, such as the Wheatstone bridge, possibly in combination with (operational) amplifiers.

- **Linearity**. The linearity of a sensor or system is the degree in which the transfer function is linear. In other words, how constant the sensor sensitivity is over the input range. The advantages of high linearity is that the signal coming from the sensor is more easily processed by the rest of the circuit. Imagine a temperature sensor that is coupled to an ADC (maybe via amplifiers and filters). If the system up to the ADC is linear, with a linear transfer function from temperature to the input of the ADC, then the number found at the output is a direct measure for the temperature. Apart from a change in scale, the number directly represents the temperature. Any two consecutive numbers always represent the same temperature difference. This facilitates the processing and decision making later on and linear sensors have always preference. The trade-off is that linear sensors are invariably more expensive.

- **Gauge factor**. If the transfer function is known, also a so-called gauge factor or calibration factor can be determined. This is defined as the *relative* changes of the response, $\Delta y$, divided by the relative changes of the input value, $\Delta x$. For infinitely small variations

  $$k \equiv \frac{\mathrm{d}y/y}{\mathrm{d}x/x}. \tag{1.2}$$

It is obvious that for linear systems the gauge factor is unity. Take for instance a transfer function $y = H(x) = \alpha x$. Then

$$k \equiv \frac{\mathrm{d}y/y}{\mathrm{d}x/x} = \frac{\mathrm{d}y}{\mathrm{d}x} \cdot \frac{x}{y} = \alpha \cdot \frac{x}{\alpha x} = 1. \tag{1.3}$$

- **Range**. Considering the above, 'range' can be defined as the range of measured parameter values for which the sensor or system still behaves more-or-less linearly. A typical behavior of a sensor is an S-curve, with saturation for both low-end and high-end values and we can (arbitrarily) define the range as the range of values for which the response does not deviate more than 5% from linearity, there where the sensitivity does not deviate more than 5%, see Figure 1.4.

- **Accuracy**. If me make a million measurements of the same steady quantity with a certain instance of a sensor, we will find a distribution of measurements. The average value of this distribution ideally is equal to the real physical value, but individual instances of the sensor can vary and return values deviating from the real value. Even if we measure the physical quantity a million times, the average of these values will not be equal to the real value. Take for example the resistor whose value depends on the temperature (for example an increase of 0.1% per degree). Imagine at 300 kelvin, the resistor value is 1 kΩ. If the real temperature is 300 kelvin (room temperature) we should therefore measure 1 kΩ. Yet, we know from electronics that factories are not able to make all resistances equal. Every resistance value comes with a tolerance, typically 5% or 10%, indicating the statistical deviations from the nominal value. Our single resistance can therefore indicate a wrong 'temperature'; measuring the resistance, and therefore temperature, will be wrong even if we repeat it infinitely. The *expectation value* of the difference between the measured value and the real value is called 'accuracy'. This is the spread ('standard deviation') of distances from the real value to the average ('center of mass') of an infinite number of same-sensor-measurements repeated for an infinite number of sensors. The figure makes it clear. Note that the definition of 'accuracy' is slightly different from what we know from daily life. Robin Hood can shoot the arrow at the bulls eye quite accurately. The fact is that I can shoot equally accurately (the center of weight of all my attempts will probably be spot on the bulls eye). The difference is that he could do it reproducibly, hitting *all* arrows on the bulls eye.

- **Reproducibility**, analog resolution, analog noise, or precision. The reproducibility or precision is the statistical parameter telling us how close two different measurements probably will be. It is also called (analog) noise. It is the spread (standard deviation) in found values of an infinite number of measurements. If we measure every time nearly the same value, the reproducibility is high and the noise is low. That still does not necessarily mean that we know the measured parameter very accurately, as

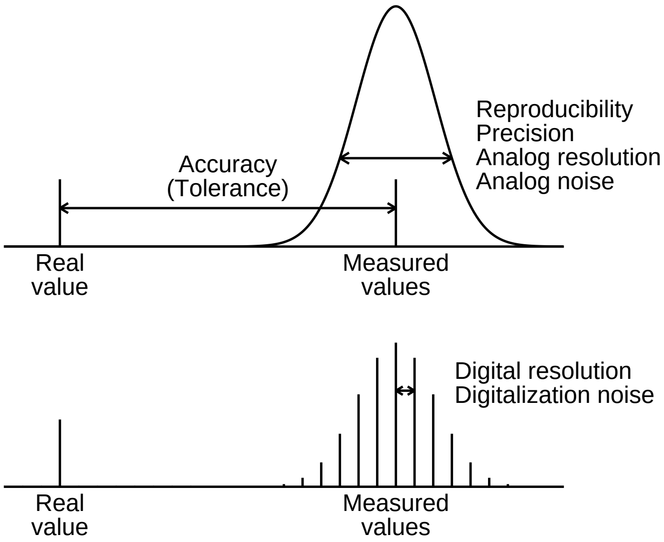


**Fig. 1.4**: Parameters of uncertainty in sensor values. (a) Accuracy (difference between real value and average of measured values). Reproducibility, precision or analog resolution (spread or standard deviation of measured values). (b) Digital resolution (difference between two adjacent possible values in digital systems)

seen before; another parameter is the accuracy. If the particular instance of the sensor itself is wrong, we will be measuring a million times the same erroneous value. We can call this reproducibility also the 'analog resolution', as it gives information about the resolving power of our sensor. Imagine two identical sensors, both with infinite accuracy (meaning that the average of an infinite number of measurements will always give the real physical value for both sensors). The 'reproducibility' or 'analog resolution' then tells us with how much certainty we can say that the temperature of the two sensors is different after a single measurement. This parameter, is not so much a parameter of the sensor as well as it is a parameter of the entire system. For instance, If measuring the temperature is done by a resistance, we can determine this resistance value very well, by low-pass filtering the signal, or by averaging many individual measured values upto achieving any desired analog resolution.

- **Systematic error**. Another name for accuracy is systematic error. This wording is especially popular among scientists. It is the error that is introduced into the measurements that cannot be eliminated by repeating and averaging the measurements, since every measurement (using the same sensor) will contain this exact same error. The only way to remove this error is by calibrating the system. Ideally, the systematic error is zero. If we cannot do so, we can deal with this error by analyzing it theoretically, on basis of the parameters of the sensors and the system and it turns into an uncertainty in the final presented values.

- **Tolerance**. Sellers of components often give a tolerance value for their product. This means the standard deviation of the distribution of systematic error. If we buy a thousand resistors, their systematic error will be a normal distribution (hopefully) around the nominal value written on the component. Each individual specimen will have a constant error (that can have high reproducibility; no noise), but the probability distribution function of the error has a standard deviation that is called the tolerance.

- **Digital resolution**. Not related to the analog resolution is the digital resolution. This parameter is only relevant for signals translated ('transduced') to the digital domain, for instance by analog-to-digital converters (ADCs). It tells us what is the distance between two adjacent *possible* measured values instead of the statistical spread of actual measured values. It is also often called 'digitalization noise' by many textbooks.

  Digital resolution is a form of systematic error, since it is an error that is predictably every time the same. Figure 1.5 exemplifies the systematic error associated with an analog-to-digital converter. The voltage as estimated by the ADC is a step function and the error voltage of difference of real voltage and estimated voltage is thus a saw-tooth function. Yet, it adds uncertainty to the final value obtained.

  This digital resolution of the overall system is given by the digital resolution at the entrance of the analog-to-digital converter divided by the sensitivity of the system up to that point.

  This resolution can be back-tracked into the final resolution of the measured quantity. As an example, for a system measuring X, converting to V at the entrance of a voltage ADC, the digital resolution of $X$ is given by

$$\Delta X = \frac{\Delta V}{S}, \tag{1.4}$$

  with $\Delta V$ the voltage resolution of the ADC and $S$ the sensitivity of the system. Section 1.7 explains how uncertainties in general propagate in the system, and how an uncertainty of the final measured value, such as voltage, can be back-tracked into an uncertainty at the source of the signal, for instance temperature.

**Question**: Imagine a sensor that converts temperature into voltage with a sensitivity of $S = 10$ mV/K (the sensor output voltage increases 10 mV for every degree temperature rise) and connected to an ADC of 8 bits with an input voltage range of 0 to 5 V. What is the digital resolution of our system?

**Answer**: On basis of this we can calculate the digital resolution

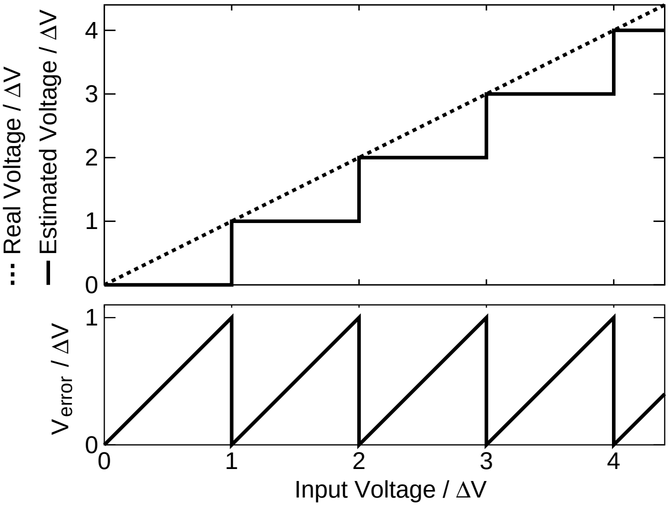


**Fig. 1.5**: Digitalization error $V_{\text{error}}$ in terms of digitalization step $\Delta V$ in a rounding-down ADC. The error voltage $V_{\text{error}}$, the difference between the real input voltage and the estimated value of the ADC, is a saw-tooth function shown in the bottom panel

of our temperature system: 8 bits means $2^8 = 256$ entrance-voltage levels, and thus a separation of $\Delta V = 5\text{V}/(256 - 1) = 19.6$ mV. The digital resolution in terms of temperature is now the voltage resolution at the entrance of the ADC divided by the sensitivity of the sensor: $\Delta T = \Delta V/S = 1.96$ K. A number at the output of the ADC 1 higher means the temperature of the sensor 1.96 K higher.

Later will be shown how adding noise can increase the resolution of an ADC. It is obvious that the better the digital resolution of the system, the better we will know the value of the measured quantity. Yet, we will still suffer from other (analog) problems. It makes no sense buying an expensive 24-bit ADC for our system when the analog resolution and accuracy are very low.

Other important parameters include

- **Selectivity**. If we have a sensor detecting a certain physical quantity, we want it to respond to only that quantity and nothing else. As a simple and obvious direct example, a carbon-monoxide gas sensor should not respond to ammonia. Taken this a little further, it can be stated that ideally the output of a sensor should only change when the measured quantity changes. As an example, a mass sensor should not respond to changes in temperature of the ambient. The signal of a temperature sensor should not depend on time (so-called drift). Etc.

- **Shelf life and operational life**. Some sensors, especially the biological sensors, can degrade, even when not used. They can consist of a biological material that degrades upon time. Other sensors can degrade upon use. Some sensors are one-time use, for instance because the surface is functionalized by a layer that reacts (or not) with the detected material.
- **Speed**. How fast does the sensor respond? Or how fast can a value change? When we measure the temperature with a massive sensor, obviously the measured temperature cannot change fast enough to accompany the measured objects temperature. Similar problems can be encountered when retarding elements such as capacitors are part of the electronic system. A pressure sensor can be made of a capacitor, as will be shown in Chapter 3. Combined with resistances this can cause relaxation times in the circuit, also called RC-times for the reason that they are equal to the product of resistance and capacitance, $\tau = RC$.
- **Durability**. If the sensor can stand the difficult operating conditions. Under water sensors. High pressure. Heat. Radiation. Mechanical stress.
- **Interference**. An ideal sensor, measuring quantity X should not change itself that quantity X, or any other property of the object. This is harder than it may seem at first. Ideally our measurement should not interfere with normal operation. However, we know from our daily life, that this is not easy. When the price of our products are 'measured' at a cashier, we know that our normal operation is interrupted. For the sensor to work we have to take our shopping out of the cart and place them in front of the sensor. We are used to this, but wouldn't it be good if the sensor did not interfere with our lives and we could just walk out of the shop (after paying, of course). Unfortunately this is not possible yet.

  Even worse it gets when the sensor itself is changing the quantity it is measuring. Quantum mechanics even teaches us that, ultimately, measuring a state changes that state. Yet, also in more macroscopic systems the measuring unit can easily change the measured quantity. In a furnace, where an object reaches thousands of degrees, placing a thermocouple can seriously change the object's temperature by the sheer thermal conductivity of the metal wires of the thermocouple. One side of the metal-wire pair is connected to the hot object, the other side connected to the cold environment, for example the multimeter in the operators room. Apart from electrical conductivity metals normally also have high thermal conductivity and a large amount of heat will be transported from measured object to the ambient, thus lowering the temperature of the object. We should always bear this in mind when we chose our sensor type. A couple of degrees interference probably is not harmful when measuring an object in a hot furnace in an industrial environment, yet it is disastrous when measuring temperatures close to absolute zero in a scientific experiment. For the latter, a thermocouple obviously is not adequate.

- **S/N (signal-to-noise) ratio**. A very important parameter and very similar to the reproducibility described above is the signal-to-noise ratio, often abbreviated with S/N. This tells us how good the signal is relative to the noise level. If we want to draw conclusions, it is always good to have a larger S/N. Methods for improving the S/N are filtering, either analog or digital. The simplest (digital) way of filtering is repeating the measurement and averaging. This can also be implemented in an analog way by a low-pass RC filter. The result is a better S/N ratio and a better precision (narrower distribution of reproducibility with smaller standard deviation). The price to pay is an increased time for taking a value, i.e., a reduced speed, also implying a reduced resolution in time.

- **Detection limit**. Often confused with sensitivity is the detection limit. Actually it is a combination of S/N ratio and sensitivity (and possibly digital resolution). This is used in phrases as ”The carbon-monoxide sensor system can detect concentrations as small as 3 ppm”. What it means is that for these concentrations the system has a S/N ration equal to 2 (or 3). The detection limit of $x$ is defined by

$$\frac{H(x_{\mathrm{DL}}) - H(0)}{H_{\mathrm{noise}}} = 2, \tag{1.5}$$

  with the noise being any of the types described above, for instance the digital resolution. Note also the difference between detection limit and quantification limit, a jargon especially used in chemistry environments. The detection limit specifies the lower limit of the value of the signal to detect its presence, whereas the quantification limit is the lower limit in concentration to allow for making quantitative statements.

- **Ease of operation / maintenance**. The best sensors are those that are contactless for two reasons. The first is that they are less likely to cause interference with the system. The second is that they are more likely to have an increased lifetime. A position sensor that consists of a variable resistor with a contact that is dragged over a carbon track will wear out faster than a distance sensor that is based on the ultrasonic principle.

  Thus, measuring at a distance, without mechanical contact to measured entity has preference. In some biological system invasive measurement techniques can be deadly for the studied object. The search is always for non-invasive measurement techniques.

  Other, more advanced, ways of increasing the S/N are taking differential values, for instance one current measurement with the light on and one with the light off. In an advanced version this idea is used with lock-in detectors (a.k.a. differential amplifiers) where only the signal is detected that is related to the modulation such as the switching on and off of the light. This is discussed in Chapter 2.

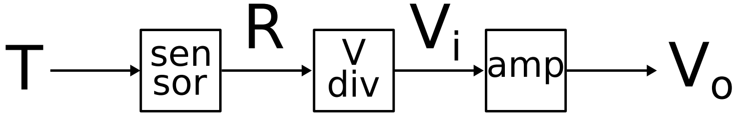


**Fig. 1.6**: System of a chain of signal-transducing elements

### 1.3.1 Sensitivity of a system

In some cases we have system that is a chain of transducers and signal changing elements. Take for example a temperature measurement system shown in Figure 1.6 consisting of a thermistor sensor that transduces the temperature value $T$ into a resistance value $R$. This resistance then used in a voltage divider to result in a voltage $V_{\mathrm{i}}$. The voltage amplified in an amplifier to yield $V_{\mathrm{o}}$. The question is now: what is the final sensitivity of the system, $S \equiv \mathrm{d}V_{\mathrm{o}}/\mathrm{d}T$? We have the situation that the voltage $V_{\mathrm{o}}$ is a function of voltage $V_{\mathrm{i}}$ is a function of resistance $R$ is a function of temperature $T$: $V_{\mathrm{o}} = V_{\mathrm{o}}(V_{\mathrm{i}}(R(T)))$.

From mathematics we know that to calculate the derivative of a composed function $y = f(g(x))$, we have to use the chain rule,

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}f(g(x))}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}g} \cdot \frac{\mathrm{d}g}{\mathrm{d}x}. \tag{1.6}$$

For example, in our thermometer-system case,

$$S \equiv \frac{\mathrm{d}V_{\mathrm{o}}}{\mathrm{d}T} = \frac{\mathrm{d}V_{\mathrm{o}}}{\mathrm{d}V_{\mathrm{i}}} \cdot \frac{\mathrm{d}V_{\mathrm{i}}}{\mathrm{d}R} \cdot \frac{\mathrm{d}R}{\mathrm{d}T}. \tag{1.7}$$

In this we can recognize the sensitivities of each of the components: amplifier, voltage divider, and sensor:

$$\begin{aligned} S_{\mathrm{Vi}\to\mathrm{Vo}} &\equiv \frac{\mathrm{d}V_{\mathrm{o}}}{\mathrm{d}V_{\mathrm{i}}} \\ S_{\mathrm{R}\to\mathrm{Vi}} &\equiv \frac{\mathrm{d}V_{\mathrm{i}}}{\mathrm{d}R} \\ S_{\mathrm{T}\to\mathrm{R}} &\equiv \frac{\mathrm{d}R}{\mathrm{d}T} \end{aligned} \tag{1.8}$$

which makes the overall sensitivity in terms of individual sensitivies equal to

$$S = S_{\mathrm{T}\to\mathrm{R}} \times S_{\mathrm{R}\to\mathrm{Vi}} \times S_{\mathrm{Vi}\to\mathrm{Vo}}. \tag{1.9}$$

In general terms, the sensitivity of a chain of elements is the product of sensitivities of the elements,

$$S = \prod_{\mathrm{i}} S_{\mathrm{i}}, \tag{1.10}$$

which is the chain rule of instrumentation.

### 1.3.2 Calibration of a sensor system

Calibration of a sensor or system consists of eliminating systematic error, thus increasing the accuracy. Statistical (unpredictable) errors cannot be eliminated, – by the sheer fact of their unpredictability – but calibration can remove the predictable errors. This causes that the average of a large number of measurements coincides with the real value. Or, in other words, after calibration of a sensor or system, the only uncertainty comes from statistical (random) fluctuations of the values. In Figure 1.4 this implies putting the center of the distribution curve on top of the real value.

Note that not every systematic error can be calibrated away. An example is the digitalization error described above (p. 8). In some cases it can be minimized. If we define the error as the root-mean-square of the error voltage, then, introducing an offset of half a digitalization level in a rounding-down ADC – effectively turning it into a rounding-to-closest value ADC – the error can be reduced by a factor 2.

Digitalization error, also sometimes confusingly called digitalization noise, can be reduced by using oversampling, a technique described in Chapter 5.

## 1.4 Signal conditioning; electronics

After the signal is translated from the physical domain into some kind of electronics domain (current, voltage, resistance, capacitance, etc.) the signal has to be prepared for acquisition or signal processing. As an example, a type-K thermocouple that translates temperature directly into a voltage has a sensitivity of $S = 41\mu\text{V}/°\text{C}$. If we were to connect it directly to a 5-volt ADC of 8 bit, the resolution in terms of temperature would be very low indeed: The 255 ($2^8 - 1$) different possibilities at the output of the ADC translate into a $\Delta V = (5\text{ V})/255 = 19.6$ mV voltage step (digital resolution) at the entrance of the ADC. This corresponds to a temperature resolution of $\Delta T = \Delta V/S = 19.6$ mV/(41 μV/°C) $= 480$ °C. Obviously an inadequate system, since even with my wet finger I can do better. We should either increase the voltage resolution ($\Delta V$) of the ADC (smaller voltage range, or larger number of bits) or increase the sensitivity of the system ($S$) up to the ADC. The latter can be achieved by amplification. Also, in some cases we will want to remove or introduce offset in our system, or remove noise by filtering. All this is called signal conditioning and will be discussed in Chapter 2.

**Question**: We have a type-K thermocouple ($S = 41$ μV/°C) that we will want to use between 20 °C and 100 °C. Our ADC has an input-voltage range of 0 - 5 V, see Figure 1.7. What is the gain (amplification factor) and offset of our signal-preparation circuit assuming that the reference point is at 0 °C?
**Answer**: At the lowest temperature the thermocouple will give a signal equal to 0.82 mV. This has to be removed. At the highest temperature, the sensor output signal is 4.1 mV. The output range is

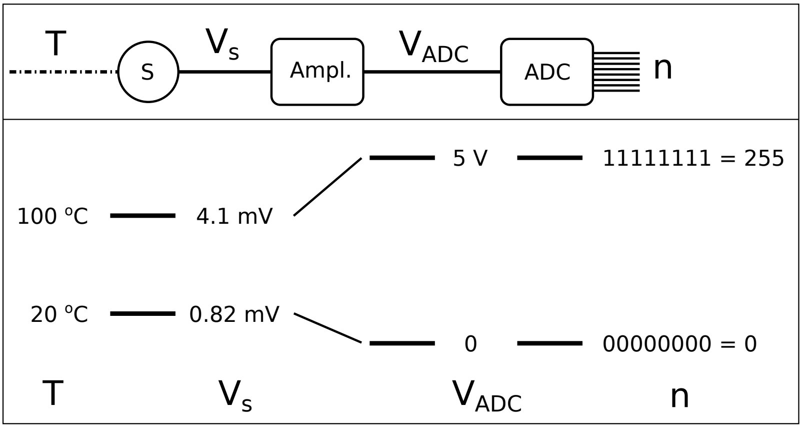


**Fig. 1.7**: Example of a mapping amplifier to prepare the signal coming from a temperature sensor (S) for a 0 - 5 V ADC. The output voltage $V_\mathrm{S}$ of a type-K thermocouple in the desired operating range 20 °C and 100 °C is between 0.82 mV and 4.1 mV. Therefore, a mapping amplifier $V_\mathrm{ADC} = 1500 \times (V_\mathrm{S} - 0.82$ mV) is required

thus 3.28 mV, which should be multiplied by 1524 to adjust it to the ADC input range. In total we should implement an amplifier between sensor and ADC that implements $V_\mathrm{ADC} = 1500 \times (V_\mathrm{S} - 0.82\ \mathrm{mV})$ or $V_\mathrm{ADC} = 1500 \times V_\mathrm{S} - 1.23$ V, whatever is more convenient, mapping the output of the sensor to the input of the ADC.

## 1.5 Data acquisition and signal processing

Signal acquisition in general terms is the total process of acquiring measurement data and storing it somewhere. In modern systems, invariably, the storage is done in digital format, there where in previous times, data could be stored in analog forms, such as the barographs storing pressure history on rolls of paper. Nowadays, analog storage is very rare. Even modern barographs store their data in digital format. Here starts the realm of informatics, beginning with converting the (by now) analog electronic signal to a digital (numeric) signal from what point a computer or any digital processing unit can take care of the processing.

The same that can be said about the acquisition can be said about the processing of data. In earlier days, the processing of data, like spectrum analysis, peak finding, signal integration, signal comparison, etc., was mostly done by human effort or by analog systems. Famous is the analog integration device, the differential machine that helped in solving differential equations. Modern processing is invariably done by digital computers or processors. While in principle a large part of the signal processing can be done in the (analog) electronic domain, such as filtering, integration, differentiating, thresholding, etc., any of these functions can also be done in the digital domain. It is up to the systems designer to decide where and how to implement what part of the functionality. For simple systems, like detecting the opening and closing of doors, and consequent switching on of warning lights, the processing at the digital level is an overkill; a simple threshold-like analog-trigger system is more than enough. For other systems, like the implementation of statistics of highway traffic, some digital data storage and (pre)processing may be adequate, and a tiny computer, such as the Atmel processor or Arduino controller board such as presented in Chapter 5 a good choice. For even more advanced signal processing, full-blown computer systems are probably needed. In the most advanced data-acquisition systems, a network of computers may be needed. We can think here of earthquake monitoring stations, where scientists can monitor the Earth's activity at a distance and compare this to other earthquake measurement huts, to accurately pinpoint seismic activity.

An advanced measurement system in a scientific laboratory can perform a lot of tasks like digital filter, curve fitting, statistics, decision making, graphical representation, etc. An example is the setup for measuring a luminescence spectrum as a function of temperature based on i) a monochromator, an actuator transducer that selects the wavelength of light to be measured ($\lambda$), ii) a photodetector, a sensor transducer that measures light intensity ($I$), and iii) a thermocouple, a sensor transducer that measures the temperature. iv) a heating element, an actuator to change the state of temperature. A computer receives the signals from the sensors, steers the actuator and combines all the information in a graph or saves it to file for further off-line processing. See Figure 1.8. Note the PID (proportional-integral-differential) stage in the computer. This is an algorithm for controlling the temperature, here shown inside the computer but can also easily be implemented by electronics outside the computer.

## 1.6 International System of Units (SI); *Système International d'Unités*

While not directly the scope of this book, a firm knowledge of quantities and units will help understanding and solving electronic instrumentation problems. While, as long as one is consistent, using any system of units is correct, the fact that the rest of the world has decided to use the International System of Units, or SI (short for the french Système International) makes that we best adopt that system as well. The international system of units consists of a set of units for basic quantities together with a set of prefixes. In many books the correct use of SI in communications is presented. See for instance the book of Almeida, *Sistema Internacional de Unidades (SI)*, or the publication of Cohen and Giacomo in Physica 146A (1987). Not only will our final presentation be unambiguous for our colleagues, but a careful use of units can even avoid

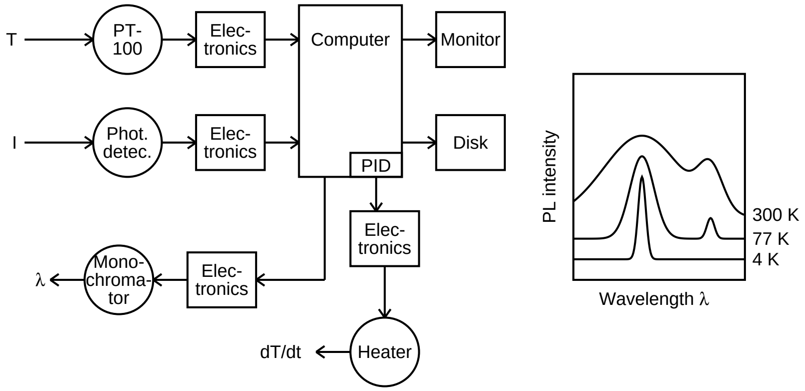


**Fig. 1.8**: Example of a scientific measurement setup. A photoluminescence (PL) spectrum analyzer measuring PL spectra as a function of temperature (T). It is based on two sensors - a temperature sensor (PT-100) and a photo-detector - and two actuators - a monochromator and a heater. A computer processes the signals and steers the equipment, shows the result on screen or saves it for off-line analysis. The right part of the figure shows an example of the output, PL spectra for various temperatures

major errors. Imagine that we present an electronic circuit with calculations and specify "the resistance at the emitter should be 3". For us this might be obvious; '3' means '3 kΩ'. For others this might be less obvious. On an electronic scheme some might not even know the component is a resistor and might misinterpret it for a capacitor; '3' is '3 nF'.

Another example. Imagine the specification for a temperature sensor specifying "The sensor is calibrated in the range 100 to 300 degrees". This is not unambiguous. Is here meant "100-300 kelvin" or "100-300 celsius"? In my village a road sign tells me the maximum velocity is there "50 km". My Internet service provider offers a connection speed of "5 Mb". All nonsense. Wrong units. To avoid confusion we can make a general rule:

Never, at any place - not even in pen-and-paper calculations or at 'intermediate' results - omit the (correct) units!

Trust me, this is going to save you from getting into trouble sooner or later. Rockets have crashed and people have died because of a wrong use of units (by the way closely followed by a wrong sign in the calculations). This is not a joke. According to CNN (September 30, 1999) "NASA lost a $125 million Mars orbiter because one engineering team used metric units while another

**Table 1.I**: SI base units

| *Quantity* | *Unit* | *Symbol* |
|---|---|---|
| Length | meter | m |
| Mass | kilogram | kg |
| Time | second | s |
| Electric current | ampere | A |
| Thermodynamic temperature | kelvin | K |
| Amount of substance | mole | mol |
| Luminous intensity | candela | cd |

used English units for a key spacecraft operation.” Therefore, always use units in calculations and communications and preferably use SI.

The units of SI can be divided into two subsets. There are the seven base units. Each of these base units is dimensionally independent, meaning that none of them can be expressed in terms of the others. From these seven base units several other units are derived by multiplication and/or division. For this, the normal rules of mathematics apply; if, for example, we divide two physical quantities with different units, the resulting unit is the division of the two units. Resistance with unit ohm (Ω) is the division of electrical potential with unit volt (V) and electric current with unit ampere (A). Thus, if resistance is potential divided by current, $R = V/I$, the resulting unit is volt divided by ampere, $[R] = [V]/[I]$. $\Omega = \mathrm{V/A}$. Note the writing convention: $[X]$ means ’the unit of quantity $X$’. Note also that for these calculations of divisions and multiplications of units, derivatives are considered divisions and integrals are considered multiplications. The dynamic resistance is the derivative of current vs. voltage, $R = \mathrm{d}V/\mathrm{d}I$, the unit of $\mathrm{d}V$ is volt and of $\mathrm{d}I$ is ampere resulting in the unit ohm for dynamic resistance, just as for the static resistance of Ohm’s law. In the equation of charge as the integral of current, $Q = \int I\mathrm{d}t$, the unit of charge (coulomb) is the product of the unit of $I$, current (ampere) and $\mathrm{d}t$, time (second): $\mathrm{C} = \mathrm{A\ s}$.

In this way, added to the basic SI units are a set of non-SI units accepted for use within SI. The base units and quantities are given in Table 1.I. Amazing that all humanly observable quantities in the universe can be expressed in these seven basic units. Note, for instance that the electric potential is not there, nor is the magnetic field. That is because they can be described in terms of standard SI units. Table 1.II shows some derivative units.

Some quantities are outside the scope of SI. One can think here of thinks as ’charm’, ’color’, or ’lepton number’ quantum numbers of elementary particles such as quarks. Anyway, this falls even further beyond the scope of this book.

**Question**: The magnetic field according to Ampere’s law is defined as the field at a distance of 1 meter from an infinite linear wire bearing a current of 1 ampere, but includes a conversion factor equal to $2\cdot10^{-7}$ N $\mathrm{A}^{-2}$. What is the unit of magnetic field?

**Table 1.II**: Derived SI units expressed in terms of basic SI units or other derived units

| *Quantity* | *Unit* | *Symbol* | *Basic SI* | *Other* |
|---|---|---|---|---|
| Plane angle | radian | rad | m/m | |
| Solid angle | steradian | sr | $m^2/m^2$ | |
| Frequency | hertz | Hz | $s^{-1}$ | |
| Force | newton | N | kg m $s^{-2}$ | J/m |
| Pressure | pascal | Pa | kg $m^{-1}$ $s^{-2}$ | $N/m^2$, $J/m^3$ |
| Energy, Work, Heat | joule | J | kg $m^2$ $s^{-2}$ | N m, C V |
| Power, Radiant flux | watt | W | kg $m^2$ $s^{-3}$ | J/s |
| Electric charge | coulomb | C | A s | |
| Electric potential | volt | V | kg $m^2$ $s^{-3}$ $A^{-1}$ | J/C, W/A |
| Electric capacitance | farad | F | $kg^{-1}$ $m^{-2}$ $s^4$ $A^2$ | C/V |
| Electric resistance | ohm | Ω | kg $m^2$ $s^{-3}$ $A^{-2}$ | V/A |
| Electric conductance | siemens | S | $kg^{-1}$ $m^{-2}$ $s^3$ $A^2$ | A/V, $\Omega^{-1}$ |
| Magnetic flux | weber | Wb | kg $m^2$ $s^{-2}$ $A^{-1}$ | V s |
| Magnetic flux density | tesla | T | kg $s^{-2}$ $A^{-1}$ | $Wb/m^2$ |
| Inductance | henry | H | kg $m^2$ $s^{-2}$ $A^{-2}$ | Wb/A |
| Luminous flux | lumen | lm | cd sr | |
| (Radio)activity | becquerel | Bq | $s^{-1}$ | |
| Absorbed dose | gray | Gy | $m^2$ $s^{-2}$ | J/kg |
| Dose equivalent | sievert | Sv | $m^2$ $s^{-2}$ | J/kg |

**Answer**: The unit according to Ampere's law is therefore N $A^{-2}$ × A / m = N $m^{-1}$ $A^{-1}$, which does not bring us much further, since it contains the non-basic-SI unit newton (N). Newton, however, can be converted to SI units when we remember the classic Newtonian equation, force equals mass times acceleration ($F = ma$), thus N = kg m $s^{-2}$. We arrive then at the unit for magnetic field: kg m $s^{-2}$ × $m^{-1}$ $A^{-1}$ = kg $s^{-2}$ $A^{-1}$. This, in fact, is equal to the unit tesla (T), see Table 1.II.

A physical quantity is expressed as the product of a numerical value (i.e. a pure number) and a unit. Some physical quantities do not have units. For example $n = 1.33$, the refractive index of glass, or $\varepsilon_r = 5$, the dielectric constant of polyethylene (making the permittivity of polyethylene five times that of vacuum, $\varepsilon = 5\varepsilon_0$). In some cases a unitless quantity can be 'forced' to have a unit, for instance "the gain of the amplifier is $A = 120$ V/V", informing us that the gain is defined in terms of voltage.

A prefix may be added to units to produce a multiple of the original unit. All multiples are integer powers of ten. For example, 'kilo' denotes a multiple of a thousand and 'milli' denotes a multiple of a thousandth. The prefixes are never combined: a millionth of a kilogram is a milligram not a microkilogram. Table 1.III gives the most common prefixes and in the last column gives and

**Table 1.III**: SI prefixes and their meaning

| *Prefix* | *Symbol* | *Factor* | *American* | *British* | *Used with meter* |
|---|---|---|---|---|---|
| yotta | Y | $10^{24}$ | septillion | quadrillion | clusters of galaxies |
| zetta | Z | $10^{21}$ | sextillion | trilliard | Milky Way |
| exa | E | $10^{18}$ | quintillion | trillion | 1000 stars in Milky Way |
| peta | P | $10^{15}$ | quadrillion | billiard | Oort Cloud diameter |
| tera | T | $10^{12}$ | trillion | billion | Jupiter orbit |
| giga | G | $10^{9}$ | billion | milliard | Earth-Moon distance |
| mega | M | $10^{6}$ | million | | country size |
| kilo | k | 1000 | thousand | | walking distance |
| hecto | h | 100 | hundred | | skyscraper height |
| deca | da | 10 | ten | | building height |
| - | - | 1 | one | | human size |
| deci | d | 0.1 | tenth | | hand |
| centi | c | 0.01 | hundredth | | finger width |
| milli | m | 0.001 | thousandth | | fingernail thickness |
| micro | µ | $10^{-6}$ | millionth | | hair thickness |
| nano | n | $10^{-9}$ | billionth | milliardth | molecule length |
| pico | p | $10^{-12}$ | trillionth | billionth | |
| femto | f | $10^{-15}$ | quadrillionth | billiardth | proton diameter |
| atto | a | $10^{-18}$ | quintillionth | trillionth | |
| zepto | z | $10^{-21}$ | sextillionth | trilliardth | |
| yocto | y | $10^{-24}$ | septillionth | quadrillionth | |

example for when the prefix is used in combination with the unit 'meter'. Note the use of lowercase for the 'k' of 'kilo'. Nearly always everywhere the kilo is written as 'K'. This is wrong. The uppercase 'K' should be used solely for 'kelvin'.

### 1.6.1 Writing conventions

While not strictly part of SI, which only tells which units are used for what, it also helps to format our text carefully, to make it more readable for others. Therefore it is useful to follow some convention when we write documents.

- Spelled-out units are written in lowercase even if they are named after people. Examples: 'watt', 'tesla', 'ampere', etc.

- Symbols for units are written in lower case, except for symbols derived from the name of a person. For example, the unit of pressure is named after Blaise Pascal, so its symbol is written 'Pa' whereas the unit itself is written 'pascal'. The one exception is the liter, whose original symbol 'l' is unsuitably similar to the numeral '1', 'L' can be used instead. The cursive '*l*' is occasionally seen, especially in Japan, but this is not recommended.

Another exception is the calorie, which has two abbreviations - to make it more confusing: The 'cal' (lowercase) is the energy it costs to heat 1 gram of water one degree celsius. The 'Cal' (uppercase) is the energy it costs to heat 1 liter of water 1 degree celsius, or in other words, 1 Cal = 1000 cal.

- Abbreviated units, unlike spelled-out full names of units, should not be pluralized. For example "25 kg" and not "25 kgs". For spelled-out unit names, all are made plural by adding an 's', except lux, hertz, and siemens, all of which are the same in singular and plural.

- Symbols do not have an appended period (.) unless at the end of a sentence.

- It is advised to write units in upright Roman type (m for meters, l for liters), so as to differentiate from the italic type used for symbols of variables and parameters ($m$ for mass, $l$ for length).

- A space should separate the number and the symbol, e.g. "2.21 kg", "$7.3 \times 10^2$ m$^2$", "22 °C". Exceptions are the symbols for plane angular degrees, minutes and seconds (°, ' and "), which are placed immediately after the number with no intervening space.

- Derived units formed from multiple units by multiplication are joined with a space or center dot (·), e.g. "N m" or "N·m". This also helps in things like distinguishing "ms" (millisecond) from "m s" (meter second).

- Units formed by division of two other units are joined with a solidus (/), or given as a negative exponent. For example, the "meter per second" can be written "m/s", "m s$^{-1}$", "m·s$^{-1}$". A solidus should not be used if the result is ambiguous, i.e., "kg·m$^{-1}$·s$^{-2}$" is preferable to "kg/m·s$^2$".

- When the value is zero or infinity, the unit does not have to be specified. We can say "the conductivity of the device $\sigma = 0$", without specifying 'siemens'. That is because zero is zero and infinity is infinity in any system of units. There is, of course, one exception, namely degrees centigrade 0 °C $\neq$ 0 K.

- In labels in plots of figures or table columns, the units are preferably shown after the solidus (/), showing that the quantity is 'divided' by this unit before plotting or listing. So "Drain-source current / A", or "Energy / J". This way, effectively unitless values result (that can be plotted and listed). Alternatively, but less correct, the units are placed in brackets: "Drain-source current (A)" and "Energy (J)". The difference is immediately obvious when we plot, for instance, the logarithm of current: "Log (Current / A)" is more unambiguous compared to "Log (Current) (A)", since for the latter it is not clear if the current is first made unitless by dividing by ampere before taking the logarithm (something that makes sense) or first the logarithm is taken and then the result is divided by A (something that does not make sense).

- Continuing the above: Note that functions cannot have units in their arguments! The logarithm of ampere does not make sense and neither does the exponent of joule, etc. If we find in our equation a function with an argument having a unit, we must have done something wrong. For instance we forgot to divide by the reverse-bias leakage current of the diode (to make it $\log[I/I_0]$) or the thermal energy (to make it $\exp[-E/kT]$) to give but some examples. Inside the functions no units! (Exceptions are simple 'polynomial' functions such as $x^2$ and $1/x$, as in "the frequency is given by the reciprocal time, $\omega = 1/t$").

## 1.6.2 Symbols for variables and parameters

Also for symbols of variables and parameters ('variables' for short) we need to be careful and follow some rules:

- Variables are single letters of the Latin or Greek alphabet.

- Chose names of symbols well. As a rule of thumb use the first letter of the word representing the meaning. 'P' for pressure, 'C' for capacitance. Obviously, sometimes we have to abandon this idea, namely if we already used a symbol for something else. If we have used 'C' for capacitance, we can no longer use it for current. In any case, it is best to use the convention of the area in which you are working. In my area of electronic measurements, for instance, current is normally indicated by $I$ and current density by $J$.

- To distinguish them from normal text, symbols for physical quantities and symbols for numerical variables are written in italics. Compare the sentences "We live in the house of a meter" (We bought a house from a person measuring things?) and "We live in the house of *a* meter". Yet, Greek symbols, descriptive subscripts and numerical subscripts are written in roman (since no distinction is necessary). Subscripts that themselves represent variables are again written in italics: $A_{\mathrm{v}}$ (voltage gain), $A_i$ (i-th element of vector A), $\mu$, "a house of $\alpha$ meter". Even better would be to write "a house with a length of $\alpha$ meter", of course, specifying the quantity, the symbol and the unit of the parameter.

- Always explain each and every used symbol at least once, namely at the first instance of use. Don't assume that people know that $R$ represents 'resistance'. Although it is immensely obvious that it is so for us, this is much less obvious for people that have never heard of electronics and that are working in for instance banking. $R$ probably (obviously!) means interest rate in their field. $R$ means 'diameter' for plumbers, 'regression coefficient' for statisticians.

- Multiplication of variables in an equation is done by placing them next to each other. A space or a multiplication symbol (×) can be used. Note that this now makes it obvious why symbols should be *single* letters. Otherwise in $ab$ we would not be able to distinguish between the product of $a$ and $b$ and the variable $ab$.

Table 1.IV gives a summary of some physical constants used in this book. Table 1.V shows some other useful units. Note the definition of lightyear in this table. While not very useful for the rest of the book, it shows how we should use symbols and units. The lightyear is a unit of length, defined as the distance that light travels in vacuum in one year, thus 1 ly $= c$ a. $c$ is the speed of light (unit: m/s) and a is the time unit equal to one (Julian) year (365.25 d = 31557600 s). This might be confusing at first sight; it looks as if the unit of $c$ is 'a'. Don't forget that the symbols $c$ and 'a' both represent a numerical value multiplied by a unit. Multiplying them will give a new value-unit combination, in this case the multiplication is 1 ly $= (2.99792458 \times 10^8$ m s$^{-1}) \times (31557600$ s$) =$ $(2.99792458 \times 10^8 \times 31557600)$ (m s$^{-1} \times$ s), resulting in a numerical value of $9.460716010 \times 10^{15}$ and a unit m.

Finally, it is worth to mention the unit bel. Named in honor of telecom pioneer Alexander Graham Bell, a bel represents a factor ten. Better known is the decibel, a tenth of a bel and thus representing a factor $\times 10^{1/10} = 1.258925412$. In engineering this is often used, especially in electronic engineering. We can for instance say that an amplifier has a voltage gain of 30 dB, meaning that the output voltage amplitude is a factor $\times 10^{30/10} = 1000$ times larger than the input voltage amplitude. It is more common to talk about the power gain of a circuit. If the voltage gain is a factor 1000 (30 dB), then the current gain, following Ohm's law $I = V/R$ (current is voltage divided by resistance) is also a factor 1000 bigger, making the total power $P = VI$ a factor 1,000,000 bigger, equaling 60 dB. Some confusion might remain when we read things such as "an amplifier has 40 dB gain" if it is not specified if a voltage or a power gain is meant. In such cases normally a power gain is meant, but exceptions always exist.

A factor two is approximately 3 dB. This is a recurring value. For instance, a filter, at its cut-off frequency, has an output voltage equal to $1/\sqrt{2}$ times the pass-band value. The power at this 3-dB point is thus halved or $-3$ dB. Far beyond this cut-off frequency, in a simple low-pass filter, the voltage drops proportional to the reciprocal frequency, $V_\mathrm{o} \propto 1/f$, or, in other words, a factor ten drop for every factor ten increase in frequency. We can call this $-10$ dB/dec – "ten decibels per decade" – in voltage terms, or $-20$ dB/dec in power terms.

A power can also be defined relative to a certain level, instead of relative to the input power. Often the milliwatt is used and we get a unit dBm. Expressed in dbm the power $P$ is $10 \times \log(P/1 \text{ mW})$. Also levels relative

**Table 1.IV**: Fundamental physical constants

| *Quantity* | *Symbol* | *Value* |
|---|---|---|
| Speed of light in vacuum* | $c$ | $2.99792458{\times}10^{8}$ m s$^{-1}$ |
| Permeability of vacuum* | $\mu_0$ | $4\pi \times 10^{-7}$ N A$^{-2}$ |
| Permittivity of vacuum* | $\varepsilon_0$ | $8.854187817 \times 10^{-12}$ F m$^{-1}$ |
| Gravitational constant | $G$ | $6.6725985 \times 10^{-11}$ m$^3$ kg$^{-1}$ s$^{-2}$ |
| Boltzmann constant | $k$ | $1.38065812{\times}10^{-23}$ J K$^{-1}$ |
| Planck constant | $h$ | $6.626075540{\times}10^{-34}$ J s |
| Elementary charge | $q$ | $1.6021773349{\times}10^{-19}$ C |
| Avogadro constant | $N_\mathrm{A}$ | $6.022136736{\times}10^{23}$ mol$^{-1}$ |
| Faraday constant | $F = qN_\mathrm{A}$ | $9.648533289{\times}10^{4}$ C mol$^{-1}$ |
| Stefan-Boltzmann constant | $\sigma$ | $5.6705119{\times}10^{-8}$ W m$^{-2}$ K$^{-4}$ |
| Bohr magneton ($qh/4\pi m_\mathrm{e}$) | $\mu_\mathrm{B}$ | $9.274015431{\times}10^{-24}$ J T$^{-1}$ |
| Nuclear magneton ($qh/4\pi m_\mathrm{p}$) | $\mu_\mathrm{N}$ | $5.050786617{\times}10^{-27}$ J T$^{-1}$ |
| Bohr radius | $a_0$ | $0.52917724924{\times}10^{-10}$ m |
| Electron mass | $m_\mathrm{e}$ | $9.109389754{\times}10^{-31}$ kg |
| Proton mass | $m_\mathrm{p}$ | $1.672623110{\times}10^{-27}$ kg |
| Neutron mass | $m_\mathrm{n}$ | $1.674928610{\times}10^{-27}$ kg |

*: $\mu_0\varepsilon_0 = 1/c^2$

to a microwatt are quite common, resulting in a unit dBμ, or 1 watt, resulting in dBW.

## 1.7 Propagation of uncertainty

An important issue for measurement systems – especially *scientific* measurement systems – is the propagation of uncertainty of our system. It is about making qualitative statements of the uncertainty of the final found value. It also specifies how to present the final result, i.e., how many decimal cases should be written out in the final value, how many decimal cases are significant.

First of all, 'uncertainty' is not the same as 'error'. Sometimes the two are mixed and we can find pages on 'Error propagation', when 'Uncertainty propagation' is meant. To make it clear, an error is something that is known and detected. I can for instance say that my thermometer has an error after I checked it with a calibrated high quality thermometer. I can then say "my thermometer has an error of 2 degrees" because when the thermometer indicates 21 degrees, I know it is in fact 23 degrees. The other thing is that, when I measure with my (uncalibrated) thermometer, I can say that "the temperature is 23 degrees in this room, but there is an uncertainty of 2 degrees in that value". Because my thermometer might have a systematic error, or because there is a random fluctuation in the measured values, this temperature has an uncertainty of 2 degrees; with 95% certainty the real temperature is between 21 and

**Table 1.V**: Other useful units

| *Unit* | *Symbol* | *Quantity* | *Value* |
|---|---|---|---|
| angstrom | Å | length | $10^{-10}$ m |
| electronvolt (= $q$ J $C^{-1}$) | eV | energy | 1.6021773349 $\times 10^{-19}$ J |
| standard gravity | $g$ | acceleration | 9.80665 m $s^{-2}$ |
| bar | bar | pressure | $10^5$ Pa |
| atmosphere | atm | pressure | $1.01325\times10^5$ Pa |
| torr | Torr | pressure | 133.322 Pa |
| calorie | cal | heat (energy) | 4.1868 J |
| calorie | Cal | heat (energy) | 1000 cal |
| curie | Ci | radioactivity | $3.7\times10^{10}$ $s^{-1}$ |
| rem | rem | equivalent dose | 0.01 Sv |
| rad | rd | absorbed dose | 0.01 Gy |
| minute | m | time | 60 s |
| hour | h | time | 3600 s |
| day | d | time | 86400 s |
| year | a | time | 365.25 d = 31557600 s |
| lightyear | ly | length | $c$ a = $9.460716010\times10^{15}$ m |
| gauss | G | magnetic field | $10^{-4}$ T |
| liter | L | volume | $10^{-3}$ $m^3$ |
| bel | B | ratio | 10× |
| (Unified) atomic mass unit | u | mass | $1.6605402 \times 10^{-27}$ kg |

**Table 1.VI**: Symbols used in this book

| *Symbol* | *Quantity* | *Unit* |
|---|---|---|
| $I$ | Electric current | A |
| $V$ | Electric potential | V |
| $R$ | Resistance | Ω |
| $G$ | Conductance | S |
| $C$ | Capacitance | F |
| $L$ | Inductance | H |
| $Z$ | Impedance | Ω |
| $Y$ | Admittance | S |
| $P$ | Pressure | Pa |
| $P$ | Power | W |
| $T$ | Temperature | K |
| $t$ | Time | s |
| $f$ | Frequency | Hz |
| $\omega$ | Radial frequency | rad/s |
| $x$, $y$, $z$, $L$, $h$, $W$ | Length | m |

25 degrees, because the exact value of temperature is not known but can only be estimated with some kind of probability-distribution function with an expectation (mean) value of 23, and a standard deviation of 2 degrees. Uncertainty is defined as the absence of certainty. A reduced quality of information.

We can now calculate the uncertainty of a found value if this value is the result of a function of input parameters, each with its own uncertainty. For instance, the temperature is a function of measured resistance(s) of a temperature-dependent resistor, placed in a voltage divider configuration (see Chapter 2). Each measured value comes with an error, as discussed above, in the form of precision or accuracy. Either causes an uncertainty of the measured value. For the calculation of uncertainty, the accuracy and precision are treated the same way. The uncertainty of the measurement is summarized in the parameter of standard deviation, $\sigma$, of the distribution if we were to repeat an infinite number of measurements with an infinite number of sensors. The standard deviation $\sigma$ is the width of the distribution if this probability of finding a specific value $x$ is normally distributed,

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \tag{1.11}$$

where $\mu$ is the median (average) value of the measurements. Often the parameter $\sigma^2$ is used, which is called variance.

The question now is: Given the probability distribution of each parameter used to calculate the final value, what is the uncertainty of this final value? While this is a complicated question (more precisely, it has a complex answer), we can simplify things. Here the following simplifications are made:

- Each variable is independent. Two measured quantities are always independent. This avoids talking about covariance, etc.

- Each variable has a probability-density function that is normally distributed as described above (Eq. 1.11): It is summarized by two parameters, $\mu$ and $\sigma$, representing the mean value and spread (standard deviation), respectively. This assumption is rather dubious, but the final analysis of uncertainty, actually is still quite good, even if the functions are not Gaussian.

- The relative standard deviation is small.

- Systematic and random errors – accuracy and precision – are treated equally. It does not matter if the uncertainty comes from a deviation from the real value because of a faulty sensor (systematic error), or from the deviation from the real value of a single measurement of an instance of the sensor (precision). Both are assumed to have a normal distribution, each with parameters $\mu$ and $\sigma$. The combination of the two results of a new set of parameters $\mu$ and $\sigma$ that can be found the same way as if the uncertainties came from two values, each with its own $\mu$ and $\sigma$.

- The uncertainties of the variables and parameters are known *a priori* (for instance from the datasheets of the sensors) and do not have to be experimentally found from the measurement data themselves.

With these assumptions we can analyze our systems. Assuming the final result is a function of variables $x_1$, $x_2$ ... $x_n$, i.e.,

$$y = f(x_1, x_2, \ldots, x_n), \tag{1.12}$$

each variable $x_i$ having an uncertainty $\sigma_i$, then the estimation value for $Y$ and its uncertainty $\sigma_\mathrm{y}$, based on measurements $X_i$ are given by respectively

$$Y = f(X_1, X_2, \ldots, X_n), \tag{1.13}$$

and

$$\sigma_\mathrm{y} = \sqrt{\sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_i^2}. \tag{1.14}$$

**Question**: Imagine a man walking with a speed of 6±2 km/h on a boat that travels with a speed of 10±3 km/h. What is the total speed, and its associated uncertainty, of the man?
**Answer**: The function is $y = v_1 + v_2$ and we easily find a total speed of 16.0 km/h by summing the two velocities. For the uncertainty we cannot simply sum the two individual uncertainties, $\sigma = \sigma_1 + \sigma_2 = 5$ km/h, but have to use the square-rooted-sum-of-squares instead. The derivatives are $\partial y/\partial v_1 = \partial y/\partial v_2 = 1$. Thus, $\sigma_\mathrm{y} = \sqrt{1^2 \times (2\text{ km/h})^2 + 1^2 \times (3\text{ km/h})^2} = 3.61$ km/h.

As a corollary, if a variable has both systematic error $\sigma_\mathrm{s}$ and random error $\sigma_\mathrm{r}$, then the total uncertainty in that variable is given by

$$\sigma = \sqrt{\sigma_\mathrm{s}^2 + \sigma_\mathrm{r}^2}, \tag{1.15}$$

from which moment on we can use this uncertainty for the variable, without remembering where the uncertainty came from exactly.

A special case is averaging. It is easy to show that by averaging, the relative uncertainty drops with a factor equal to the square root of the number of samples used in the averaging. The average of $N$ samples is defined as

$$y_\mathrm{avg} = \frac{1}{N}\sum_{i=1}^{N} y_\mathrm{i}. \tag{1.16}$$

Each sample comes with the same uncertainty $\sigma$. Since all derivatives $\partial y_\mathrm{avg}/\partial y_i$ in the function above are equal to $1/N$, the final uncertainty in the average is (Eq. 1.14)

$$\begin{aligned} \sigma_\mathrm{avg} &= \sqrt{N \times (1/N)^2 \times \sigma^2} \\ &= \frac{\sigma}{\sqrt{N}}. \end{aligned} \tag{1.17}$$

For example, taking 100 measurements of the same quantity reduces the uncertainty (noise) by a factor of 10.

An interesting case is a function that is a product or division of independent variables. For these functions the relative uncertainties can be quadratically summed:

$$\begin{aligned} f(x,y,z) &= \alpha xyz, && (1.18)\\ \sigma_{\mathrm{f}} &= \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 \sigma_{\mathrm{x}}^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_{\mathrm{y}}^2 + \left(\frac{\partial f}{\partial z}\right)^2 \sigma_{\mathrm{z}}^2} \\ &= \sqrt{(\alpha yz)^2\sigma_{\mathrm{x}}^2 + (\alpha xz)^2\sigma_{\mathrm{y}}^2 + (\alpha xy)^2\sigma_{\mathrm{z}}^2}, && (1.19)\\ \frac{\sigma_{\mathrm{f}}}{f} &= \sqrt{\left(\frac{\sigma_{\mathrm{x}}}{x}\right)^2 + \left(\frac{\sigma_{\mathrm{y}}}{y}\right)^2 + \left(\frac{\sigma_{\mathrm{z}}}{z}\right)^2}. && (1.20) \end{aligned}$$

In other words, as an example, if $x$, $y$ and $z$ have relative uncertainties of 5%, 1%, and 3%, respectively, the resulting uncertainty is 5.92%. A similarly interesting case is when the function is an algebraic sum:

$$\begin{aligned} f(x,y,z) &= \alpha x + \beta y + \gamma z, && (1.21)\\ \sigma_{\mathrm{f}} &= \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 \sigma_{\mathrm{x}}^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_{\mathrm{y}}^2 + \left(\frac{\partial f}{\partial z}\right)^2 \sigma_{\mathrm{z}}^2} \\ &= \sqrt{\alpha^2\sigma_{\mathrm{x}}^2 + \beta^2\sigma_{\mathrm{y}}^2 + \gamma^2\sigma_{\mathrm{z}}^2}. && (1.22) \end{aligned}$$

In this case, the absolute uncertainty is the weighed sum of individual uncertainties.

If we don't want to do the difficult mathematics, we can also find an estimation of the final uncertainty by first just substituting the nominal measured value and then by the nominal value plus or minus the uncertainty in this measured value. The final uncertainty is then equal to the spread divided by 2. Imagine we want to measure the volume of a cube, $V = a^3$, and measure its side, $a = 10 \pm 0.1$ cm. The estimated volume is thus 1000 cm$^3$. The above algorithm tells us to calculate the derivative, $\mathrm{d}V/\mathrm{d}a = 3a^2 = 300\ \mathrm{cm}^2$. The uncertainty in volume is then $\sigma_{\mathrm{v}} = \sqrt{(300\ \mathrm{cm}^2)^2 \times (0.1\ \mathrm{cm})^2} = 30\ \mathrm{cm}^3$. Alternatively, we might have substituted $a$, $a - \sigma_{\mathrm{s}}$ and $a + \sigma_{\mathrm{s}}$ in our formula for the volume and find an estimation of the uncertainty:

$$\begin{aligned} &V(10\ \mathrm{cm}) = 1000\ \mathrm{cm}^3 \\ &V(10 - 0.1\ \mathrm{cm}) = 970.299\ \mathrm{cm}^3 \\ &V(10 + 0.1\ \mathrm{cm}) = 1030.301\ \mathrm{cm}^3 \end{aligned}$$

and we find an uncertainty estimation of $\sigma_{\mathrm{v}} = [V(10 + 0.1\ \mathrm{cm}) - V(10 - 0.1\ \mathrm{cm})]/2 = 30.001\ \mathrm{cm}^3$, remarkably close to the uncertainty found the other way. This method is very good for single-variable functions, especially when the relative uncertainty is small, but no longer yields adequate results for multi-variable functions. However, these can be analyzed one variable at a time, and then be square-root-sum-of-squares as before.

Finally, if the analytical determination of uncertainty is impossible, or too cumbersome, we can always revert to Monte Carlo simulations. In this scheme, we take random samples $X_i$ from the distribution functions of each variable $x_i$ and calculate the final outcome $Y$ (Eq. (1.13)). Repeating this a large number of times will reconstruct the probability-density function of $y$, from which we can determine the mean $\mu_\mathrm{y}$ and standard deviation $\sigma_\mathrm{y}$. Generating random numbers from a uniform distribution (for instance floating point numbers between 0 and 1) is quite simple, most computer languages can do that directly. Other distributions are more complicated, though. To convert uniformly distributed random numbers to normally distributed numbers we can use the conversion

$$x = \sqrt{-2\ln(u_1)}\cos(2\pi u_2) \tag{1.23}$$

with $u_1$ and $u_2$ two independent drawn numbers from a uniform distribution between 0 and 1. This will result in a normal distribution of $x$ with average 0 and variance $\sigma^2 = 1$.

### 1.7.1 Number representation

After we have found the uncertainty in a value, we can determine how to write down this value. It does not make sense to write a value with 10 decimal cases, when the uncertainty is the first decimal case, for example, $y = 2.897865446553439 \pm 0.49329892435$. As a quite arbitrary rule we can use for rounding:

- The uncertainty (standard deviation) is rounded to (at most) two significant digits
- The number itself is rounded to the same decimal case as the uncertainty

Thus, $y = 245.897865446553439 \pm 0.49329892435$ becomes $y = 245.90 \pm 0.49$. If the uncertainty itself is not well known, it is better to present the uncertainty with only one digit, so the above becomes $y = 245.9 \pm 0.5$.

For rounding up or down a digit, the convention is to round to the nearest numeral: values between x0000... and x49999... are rounded down to x..., and between x5000...1 and x99999... are rounded up to (x+1).... A special case is the value exactly equal to x5000..., which is ambiguous. To avoid systematic error when using averaging, the number is rounded to the nearest even integer. So, 1.5 is rounded up to 2 and 4.5 is rounded down to 4. This makes sure that the average of a large number of rounded numbers is equal to the average of the numbers themselves.

For representing numbers we use the following conventions:

- When using scientific notation (number times exponent), use a number between 1 and 10, multiplied by the $10^{\text{th}}$ power exponent. For example: $6.022\times10^{23}$ and not $60.22\times10^{22}$ or $0.6022\times10^{24}$, although the latter is very popular in computing sciences, i.e. `.6022E24`.

- When using scaling factors in the units, use a number between 1 and 999.999.... For example: 378 nF, and not 0.378 μF.

If no uncertainty is given for a number, we must assume the uncertainty is half of the last digit supplied. For example, '8.34' is to be interpreted as '8.340±0.005.

Typically, as a rule of thumb one can say that in a multiplication, the number of significant digits in the final result is equal to the number of significant digits of the number with least significant digits. For example, $81.3 \times 0.0003 \times 654854 = 20000$ (although using the reasoning earlier above, we would arrive at $1.6 \pm 0.3 \times 10^4$). In this approach, a significant digit is any digit written, except when it is a leading or trailing zero before the floating point. For example: 0.3040 has four significant digits, 32000 has two. One more reason to use scientific notation, since $2.0 \times 10^3$ is the same as 2000, but the former has two significant digits, while the latter might be anything from one to four. To avoid confusion, the best is to supply an uncertainty. For additions and subtractions, the number of *decimal* places (not significant digits) of the result is equal to the number of decimal places of the number with least decimal places. For example: $1.3 + 10246.7889 = 10248.1$, because the first number has one decimal place. This is the result of the calculations presented in the previous section.

In principle, as a good advice, it can also be said that it is better to work with unrounded raw numbers until the last step of presentation of the results. It costs nearly no extra work (leave the numbers in the calculator) and it avoids introducing unnecessary errors.

### 1.7.2 Backtracking uncertainty

In most cases we wind up with a final value at the end of our sensing system, with its associated uncertainty, and we want to know not only what the value is of the measured physical parameter, but also its associated uncertainty. Imagine the temperature sensing system of Fig. 1.6. If we measure the final voltage with a multimeter that has an uncertainty of 1 mV, what is the uncertainty in terms of degrees centigrade?

This can easily be calculated if we know the sensitivity of the system. Figure 1.9 demonstrates this. If a sensor system has to start-to-end sensitivity of $S_{\mathrm{x}\rightarrow\mathrm{y}}$, and an uncertainty in the final measured value equal to $\sigma_{\mathrm{y}}$, then the uncertainty

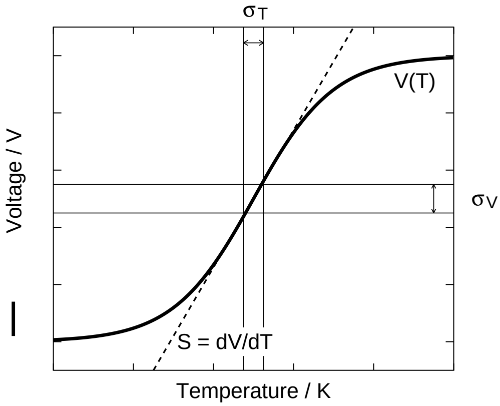


**Fig. 1.9**: Example of backtracking uncertainty in a temperature-to-voltage transducer system. An uncertainty $\sigma_{\mathrm{V}}$ in the final (measured) value $V$, translates into an uncertainty in the original measured value $T$ of $\sigma_{\mathrm{T}} = \sigma_{\mathrm{V}}/S$

in the initial parameter is the uncertainty in the measured parameter divided by the derivative of the transfer function, $\mathrm{d}H(x)/\mathrm{d}x$,

$$\sigma_{\mathrm{x}} = \frac{\sigma_{\mathrm{y}}}{\mathrm{d}H(x)/\mathrm{d}x} = \frac{\sigma_{\mathrm{y}}}{S}. \tag{1.24}$$

In our example, the case of a temperature system, the uncertainty in temperature is the uncertainty in measured voltage divided by the sensitivity of the system, $\sigma_{\mathrm{T}} = \sigma_{\mathrm{V}}/S$.

One element of uncertainty is the digital resolution. If the final measurement unit is an ADC, this has an associated uncertainty in the form of digital resolution $\Delta V$. In the same way as the uncertainty calculations shown above, the digital resolution of the measured parameter is given by

$$\Delta X = \frac{\Delta V}{S} \tag{1.25}$$

**Question**: What is the digital resolution of the system shown in Figure 1.7?
**Answer**: $\Delta V = (5 \text{ volt})/255 = 19.61$ mV, and $S = (41\ \mu\text{V}/^{\circ}\text{C}) \times (1500\ \text{V/V}) = 61.5$ mV/°C. Thus $\Delta T = 0.32$ °C.

## 1.8 Exercises

1. a) What is the resolution of a type-S thermocouple (5.88 µV/°C) connected to a 12-bit ADC with an input range of 0 - 5 V? b) What amplifier would be needed to make the sensor work between 0 °C and 40 °C (the reference temperature is 20 °C)? c) What would be the resolution then?

2. Show that the gauge factor of a power-law sensor, $H(x) \propto x^n$ is equal to $n$.

3. An alternative definition of magnetic field instead of the one given on page 17 is: "A particle having an electric charge, $q$, and moving in a magnetic field $B$ with a velocity $v$, experiences a force $F$, called the Lorentz force" (without conversion factor). On basis of this, find the SI unit for magnetic field.

4. Imagine you want to know what formula represents the energy of a charged capacitor, but you forgot. You remember it was something like the square of voltage, or the square of capacitance, or linear in both? On basis of units analysis determine which formula is correct.

5. a) The relaxation time of an electronic circuit is determined by the values of resistance and capacitance. Discuss, on basis of a units analysis, if the relaxation time is given by $\tau = RC$ or $\tau = 1/RC$.
   b) The thermal voltage of a bipolar transistor or diode depends on the temperature. Which one is correct? $V_\mathrm{T} = kT/q$, or $V_\mathrm{T} = q/kT$?
   c) Find expressions for the power of a circuit in terms of $V$, $I$, and $R$.

6. A temperature-dependent resistor is used in a voltage divider configuration (discussed in Chapter 2) with a 10 kΩ serial resistor R. The voltage divider gives an output voltage equal to

$$V = \frac{R}{R_\mathrm{T}(T) + R} V_\mathrm{CC}. \tag{1.26}$$

   If we use a 10-volt power-supply ($V_\mathrm{CC}$), and a thermoresistor given by

$$R_\mathrm{T}(T) = [1 + \alpha(T - 300\ \mathrm{K})]R_0, \tag{1.27}$$

   with $R_0$ equal to 10 kΩ and $\alpha = 1\%/\mathrm{K}$, and we measure $V = 6$ volt, what is the estimated temperature and its uncertainty? The values and relative uncertainties of the parameters are

| parameter | value | uncertainty |
|---|---|---|
| $R$ | 10 kΩ | 5% |
| $R_0$ | 10 kΩ | 3% |
| $\alpha$ | 1%/K | 0.1% |
| $V_\mathrm{CC}$ | 10 V | 1% |
| $V$ | 6 V | 2% |

## 1.9 Answers

1 a)

$$\left.\begin{array}{l} \mathrm{d}V/\mathrm{d}T = 5.88\ \mu\mathrm{V/°C} \\[6pt] \text{resolution: } \Delta T = \frac{\Delta V}{\mathrm{d}V/\mathrm{d}T} \\[6pt] \Delta V = \frac{5\ \mathrm{V}}{2^{12}-1} = 1.22\ \mathrm{mV} \end{array}\right\} \Delta T = 208\ \mathrm{°C}. \tag{1.28}$$

b) At 0 °C: $T_{\mathrm{dif}} = T - T_{\mathrm{ref}} = -20$ °C → $V = -20$ °C × 5.88 μV/°C = −117.6 μV. At 40 °C: $T_{\mathrm{dif}} = +20$ °C → $V = +20$ °C × 5.88 μV/°C = +117.6 μV. When we want an output between 0 and 5 V we have to do a voltage mapping +117.6 μV → +5 V, −117.6 μV → 0. This can be accomplished by an amplifier $V_{\mathrm{o}} = A \times (V_{\mathrm{i}} + V_{\mathrm{offset}})$, with $A = 21259$ and $V_{\mathrm{offset}} = 117.6$ μV.
c) The final resolution will then be $\Delta T = 40\ \mathrm{°C}/(2^{12} - 1) = 9.8 \times 10^{-3}$ °C (21 thousand times better).

2 If $H(x) = \alpha x^n$, the gauge factor is $k \equiv [\mathrm{d}H(x)/\mathrm{d}x] \times [x/H(x)] = n\alpha x^{n-1} \times x/\alpha x^n = n$.

3 $F = qBv$. $[F] = [q][B][v] \rightarrow [B] = [F]/[q][v] = \mathrm{N/(C\ m/s)}$. The newton is converted to SI units when we remember Newton's law, $F = ma$ (N = kg m/s$^2$). The coulomb is converted into SI when we realize that charge is the integral of current, $Q = \int I\mathrm{d}t$, or C = A s. Thus: $[B]$ = kg m/s$^2$ /(A s m/s) = kg A$^{-1}$ s$^{-2}$.

4 The unit of energy is joule (J).
Power: $P = IV$, W = (C/s) V = C V/s
Energy: $U = \int P\mathrm{d}t$, J = (C V/s) s = C V
Capacity: $C = Q/V$: F = C/V $[CV^2]$ = F V$^2$ = (C/V) V$^2$ = C V. Equal to energy!
In practice, the energy of a charged capacitor is given as $E = \frac{1}{2}CV^2$, see Chapter 3. Purely numerical constants, like 1/2, do not have units (or better to say, they have unit '1') and cannot be found by unit analysis.

5 a) Try $\tau = RC$. The unit of $R$ is Ω, which by virtue of Ohm's law ($R = V/I$) can be converted into Ω = V/A. The unit of capacitance is farad which is charge per volt, F = C/V. Thus, $[RC]$ = (V/A)×(C/V) = C/A. Coulomb is the integral of current, C = A s, thus $[RC]$ = s, which is the unit of time (what we were looking for).
b) Try: thermal voltage $V_{\mathrm{T}} = kT/q$. Substituting units left and right side of the equation: V = (J/K K)/C. The unit volt, as seen Exercise 4 is V = J/C. Joule in its turn can be converted to J = kg m$^2$ s$^{-2}$ (Einsteins famous equation, $E = mc^2$), and coulomb is ampere-second, C = A s. Thus we get

$$\frac{\mathrm{kg\ m^2}}{\mathrm{A\ s^3}} = \frac{\mathrm{kg\ m^2/s^2}}{\mathrm{A\ s}}, \tag{1.29}$$

which is correct.
c) The unit of power is energy per time, J/s.
J = kg m$^2$ s$^{-2}$ ($E = mc^2$)
V = J/C = kg m$^2$ s$^{-2}$ / A s = kg m$^2$ s$^{-3}$ A$^{-1}$.
R = V/A = kg m$^2$ s$^{-3}$ A$^{-2}$.
Then the following expressions have units of power:
$[VI]$ = kg m$^2$ s$^{-3}$ A$^{-1}$ × A = J/s
$[V^2/R]$ = (kg m$^2$ s$^{-3}$ A$^{-1}$)$^2$ /(kg m$^2$ s$^{-3}$ A$^{-2}$) = J/s
$[I^2R]$ = A$^2$ × kg m$^2$ s$^{-3}$ A$^{-2}$ = J/s

6 First we have to find the reverse function, temperature $T$ as a function of measured voltage $V$ and parameters $R$, $R_0$, $\alpha$:

$$T = \frac{1}{\alpha}\left[\frac{R}{R_0}\left(\frac{V_{\rm CC}}{V}-1\right)-1\right]+300\ \mathrm{K}, \tag{1.30}$$

and substituting the values we find $T = 266.66667$ K. We now use the following table:

$$\begin{aligned}
\frac{\partial T}{\partial \alpha} &= -\frac{1}{\alpha^2}\left[\frac{R}{R_0}\left(\frac{V_{\rm CC}}{V}-1\right)-1\right] = 3333.3\ \mathrm{K}^2, \\
\frac{\partial T}{\partial R} &= \frac{1}{\alpha}\frac{1}{R_0}\left(\frac{V_{\rm CC}}{V}-1\right) = 6.6667\ \mathrm{mK}/\Omega, \\
\frac{\partial T}{\partial R_0} &= -\frac{1}{\alpha}\frac{R}{R_0^2}\left(\frac{V_{\rm CC}}{V}-1\right) = 6.6667\ \mathrm{mK}/\Omega, \\
\frac{\partial T}{\partial V_{\rm CC}} &= \frac{1}{\alpha}\frac{R}{R_0}\frac{1}{V} = 16.667\ \mathrm{K/V}, \\
\frac{\partial T}{\partial V} &= -\frac{1}{\alpha}\frac{R}{R_0}\frac{V_{\rm CC}}{V^2} = 27.778\ \mathrm{K/V}.
\end{aligned}$$

The absolute uncertainties in the variables are

$$\begin{aligned}
\sigma_\alpha &= 10^{-5}/\mathrm{K}, \\
\sigma_{\rm R} &= 50\ \Omega, \\
\sigma_{\rm R0} &= 30\ \Omega, \\
\sigma_{\rm VCC} &= 100\ \mathrm{mV}, \\
\sigma_{\rm V} &= 120\ \mathrm{mV}.
\end{aligned}$$

The total uncertainty is thus

$$\begin{aligned}
\sigma_{\rm T} &= \sqrt{\left(\frac{\partial T}{\partial \alpha}\right)^2\sigma_\alpha^2+\left(\frac{\partial T}{\partial R}\right)^2\sigma_{\rm R}^2+\left(\frac{\partial T}{\partial R_0}\right)^2\sigma_{\rm R0}^2+} \\
&\quad \overline{+\left(\frac{\partial T}{\partial V_{\rm CC}}\right)^2\sigma_{\rm VCC}^2+\left(\frac{\partial T}{\partial V}\right)^2\sigma_{\rm V}^2} \\
&= 3.7471\ \mathrm{K}
\end{aligned} \tag{1.31}$$

The final presentation is thus $T = 266.7 \pm 3.7$ K.

# 2 Electronics

## 2.1 Introduction

This chapter discusses the electronic aspects of electronic instrumentation. It consists of altering the signal coming from sensors or going to actuators in magnitude, spectral distribution, or domain (e.g. current ↔ voltage) to make it suitable for the rest of the circuit. We will make use here of simple electronic circuits; no high level of electronics is needed, but some basic knowledge is required.

Electronics are, like everything in the real world, based on Physics. All behavior of electronic components and circuits can always be derived from physics laws. We will, for instance, see that the Kirchhoff's current law (KCL) is a charge-version of the law of physics that nothing is ever lost, and Kirchhoff's voltage law (KVL) a voltage-version of the law of symmetry of nature, that potentials are independent of the path taken.

Yet, in this book we first have a chapter on electronics and then the physics behind it is explained in the next chapter. At this stage we just need to know very few things of physics. In the opening chapter we have already seen the S.I. units. All electronics too are based on these units. More specifically, more relevant for electronics, we need to define the two concepts of current and voltage. The rest will then follow easily.

Current is the passage of charge. If through an area (for instance the cross-section of a wire) passes 1 coulomb per second, there is a current of 1 ampere. The driving force behind the current is electric potential, or voltage. We can visualize it as water flowing from a mountain. The difference in height between top an bottom of the mountain is the potential, the water flowing is the current.

This now immediately allows for two very essential concepts in electronics, two concepts that were formalized by Gustav Kirchhoff (1824-1887) in his two circuit laws, see Figure 2.1,

**KCL** Kirchhoff's current law: The sum of currents into/out of any point in space is zero.

$$\sum_{\mathrm{i}} I_{\mathrm{i}} = 0. \tag{2.1}$$

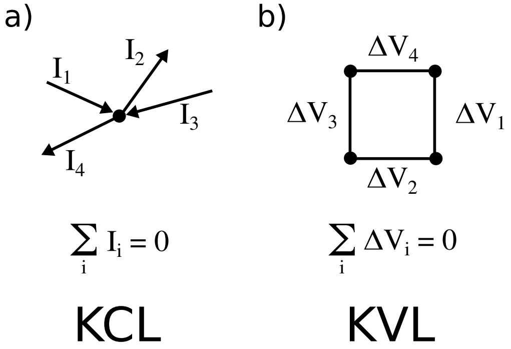


**Fig. 2.1**: Kirchhoff's circuit laws: Kirchhoff's current law (KCL, left) and Kirchhoff's voltage law (KVL, right)

This is due to the fact that charge cannot disappear, nor can it accumulate at any point.

**KVL** Kirchhoff's voltage law: In a closed path, the sum of voltage differences is zero.

$$\sum_{\mathrm{i}} \Delta V_{\mathrm{i}} = 0. \tag{2.2}$$

This is due to the fact that potential differences (integrals of force fields) are independent of the paths. So, any closed-path integral must be equal to the zero-path integral. The latter obviously being zero, any closed-path integral must be zero and thus the sum of voltage differences of a closed path must be zero.

Continuing with the analogy of current as water in a river flowing from a mountain, it is obvious, the larger the difference in height, the bigger the current. Also, the wider and deeper the channel in which the water flows, the bigger the current. This allows for the definition of resistance. For electric currents this is defined as the ratio between (electric) potential $V$ and resulting (electric) current $I$. This is so-called Ohm's law, named after Georg Simon Ohm (1789-1854),

$$R = \frac{V}{I}. \tag{2.3}$$

The units are Ω (ohm), V (volt) and A (ampere), respectively. Alternative representations of the same law are $V = RI$ and $I = V/R$. Note that, in the framework of instrumentation, a resistance can also be seen as a transducer translating information from one domain into the other, in this case from the voltage domain to the current domain, or vice versa. If any information is in the form of a voltage, we can use a resistance to convert this information into a current value. Equally, if we have information in the form of a current, we can force this current through a resistance and obtain the same information in the format of a voltage.

**Table 2.I**: Example of color coding resistances.

| Color | Value |  | Value |  | Value | Multiplier | Tolerance |
|---|---|---|---|---|---|---|---|
|  | Ring 1 |  | Ring 2 |  | Ring 3 | Ring | Ring |
| Silver |  |  |  |  |  | × 0.1 Ω | 10% |
| Gold |  |  |  |  |  | × 1 Ω | 5% |
| Black |  |  | 0.0 | + | 0.00 | × 10 Ω |  |
| Brown | 1 | + | 0.1 | + | 0.01 | × 100 Ω | 1% |
| Red | 2 | + | 0.2 | + | 0.02 | × 1 kΩ | 2% |
| Orange | 3 | + | 0.3 | + | 0.03 | × 10 kΩ |  |
| Yellow | 4 | + | 0.4 | + | 0.04 | × 100 kΩ |  |
| Green | 5 | + | 0.5 | + | 0.05 | × 1 MΩ | 0.5% |
| Blue | 6 | + | 0.6 | + | 0.06 | × 10 MΩ | 0.25% |
| Violet | 7 | + | 0.7 | + | 0.07 |  | 0.1% |
| Gray | 8 | + | 0.8 | + | 0.08 |  | 0.05% |
| White | 9 | + | 0.9 | + | 0.09 |  |  |

The reciprocal relation of the resistance also exists and is called conductance,

$$G = \frac{I}{V}, \tag{2.4}$$

the unit of which is siemens (S), named after the German inventor Werner von Siemens (1816-1892). It is obvious from the two above equations that conductance is the reciprocal of resistance.

When combining components, by placing them one after the other (in series), or next to each other (in parallel), they result in effective resistance and conductance. Since current cannot disappear in these elements - what comes in must come out - and the total voltage drop must be equal to the external supply voltage makes it easy for us to calculate the equivalent values. Placing two resistances in series:

$$\begin{aligned} V &= V_1 + V_2 \\ &= R_1 I + R_2 I, \\ R_\mathrm{s} &\equiv \frac{V}{I} = R_1 + R_2. \end{aligned} \tag{2.5}$$

Placing two resistances in parallel:

$$\begin{aligned} I_1 &= \frac{V}{R_1},\ \ I_2 = \frac{V}{R_2}, \\ I &= I_1 + I_2, \\ R_\mathrm{p} &\equiv \frac{V}{I} \\ &= \frac{1}{1/R_1 + 1/R_2} = \frac{R_1 R_2}{R_1 + R_2} = (R_1^{-1} + R_2^{-1})^{-1}. \end{aligned} \tag{2.6}$$

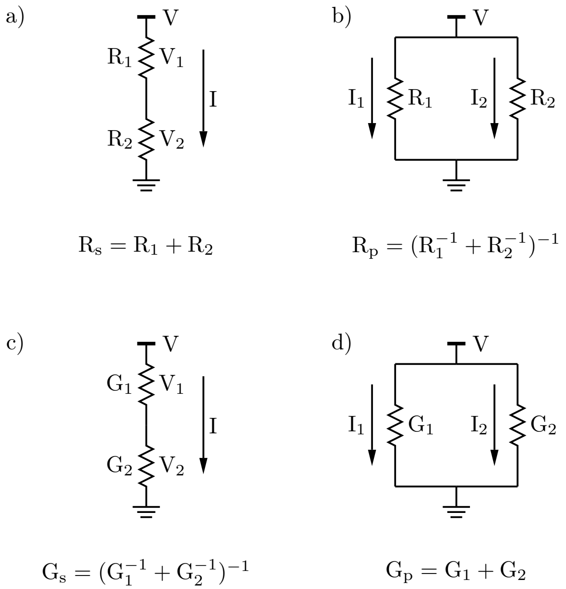


**Fig. 2.2**: Series and parallel circuits. a) The total resistance of resistances in series is the sum of resistances. b) The total resistance of resistances in parallel is the reciprocal of the sum of reciprocal resistances. c) The total conductance of conductances in series is the reciprocal of the sum of reciprocal conductances. d) The total conductance of conductances in parallel is the sum of conductances

**Question**: Show that placing two conductances in parallel and in series, results in an effective conductance equal to $G_\mathrm{p} = G_1 + G_2$ and $G_\mathrm{s} = (G_1^{-1} + G_2^{-1})^{-1}$, respectively.
**Answer**: We can either substitute the relation $G = 1/R$ everywhere above which will directly result in the correct expression, or start afresh. Placing two conductances in series:

$$\begin{aligned} V &= V_1 + V_2 \\ &= I/G_1 + I/G_2, \\ G_\mathrm{s} &\equiv \frac{I}{V} = (G_1^{-1} + G_2^{-1})^{-1}. \end{aligned} \tag{2.7}$$

Placing two conductances in parallel:

$$\begin{aligned} I_1 &= G_1 V, \;\; I_2 = R_2 V, \\ I &= I_1 + I_2, \\ G_\mathrm{p} &\equiv \frac{I}{V} = G_1 + G_2. \end{aligned} \tag{2.8}$$

Figure 2.2 summarizes these series and parallel calculations.

The power of a circuit or circuit element - the energy used by it and converted into heat - is given by the product of voltage drop and current, and this can also be rewritten with the use of Ohm's law,

$$P = \Delta V \times I = \Delta V^2/R = I^2 \times R. \tag{2.9}$$

### 2.1.1 Linear vs. non-linear circuits

Most of the electronic circuits presented here are linear circuits, which means that if we multiply the input signal by a factor $\alpha$, so does the output signal increase by the same factor $\alpha$. This implies that the output signal as a function of the input signal is a straight line *and* that the output signal is zero for a zero input signal. For these linear circuits, and exclusively for these circuits, we can use the superposition principle

Superposition Principle for a linear circuit:

The output signal of a sum of input signals is the sum of output signals for the individual input signals. Or to put it mathematically,

$$V_\mathrm{o}(V_\mathrm{i1} + V_\mathrm{i2}) = V_\mathrm{o}(V_\mathrm{i1}) + V_\mathrm{o}(V_\mathrm{i2}). \tag{2.10}$$

This will come in very handy later. All linear components, resistances, capacitors, operational amplifiers (when not in saturation), etc., can be analyzed this way. Note that non-linear components, such as diodes and transistors do not have this property.

If current is a non-linear function of potential, the definitions of resistance and conductance depend on the operating voltage, also known as 'bias'. We can then define small-signal equivalent, which are the derivative of the functions

$$r = \frac{\mathrm{d}V}{\mathrm{d}I}, \tag{2.11}$$

$$g = \frac{\mathrm{d}I}{\mathrm{d}V}. \tag{2.12}$$

Note the lower-case symbols which is standard convention in electronics. The interesting thing about the two definitions is that, if a small sinusoidal voltage signal is superimposed on a DC bias,

$$V(t) = V_\mathrm{DC} + v_\mathrm{ac}\sin(\omega t), \tag{2.13}$$

the resulting current is also a sinusoidal,

$$I(t) = I_\mathrm{DC} + i_\mathrm{ac}\sin(\omega t + \theta), \tag{2.14}$$

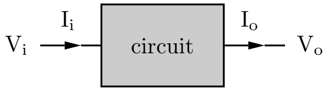


**Fig. 2.3**: Electronic circuit with input voltage and current ($V_\mathrm{i}$ and $I_\mathrm{i}$) and output voltage and current ($V_\mathrm{o}$ and $I_\mathrm{o}$) that allow to define the static and dynamic parameters presented in Table 2.II

**Table 2.II**: Definition of parameters of an electronic circuit.

| *Parameter* | *Large signal (static)* | *Small signal (dynamic)* |
|---|---|---|
| Voltage gain | $A_\mathrm{V} = V_\mathrm{o}/V_\mathrm{i}$ | $A_\mathrm{v} = v_\mathrm{o}/v_\mathrm{i} = \mathrm{d}V_\mathrm{o}/\mathrm{d}V_\mathrm{i}$ |
| Current gain | $A_\mathrm{I} = I_\mathrm{o}/I_\mathrm{i}$ | $A_\mathrm{i} = i_\mathrm{o}/i_\mathrm{i} = \mathrm{d}I_\mathrm{o}/\mathrm{d}I_\mathrm{i}$ |
| Transconductance | $G_\mathrm{m} = I_\mathrm{o}/V_\mathrm{i}$ | $g_\mathrm{m} = i_\mathrm{o}/v_\mathrm{i} = \mathrm{d}I_\mathrm{o}/\mathrm{d}V_\mathrm{i}$ |
| Transresistance | $R_\mathrm{m} = V_\mathrm{o}/I_\mathrm{i}$ | $r_\mathrm{m} = v_\mathrm{o}/i_\mathrm{i} = \mathrm{d}V_\mathrm{o}/\mathrm{d}I_\mathrm{i}$ |
| Input resistance | $R_\mathrm{i} = V_\mathrm{i}/I_\mathrm{i}$ | $r_\mathrm{i} = v_\mathrm{i}/i_\mathrm{i} = \mathrm{d}V_\mathrm{i}/\mathrm{d}I_\mathrm{i}$ |
| Output resistance | $R_\mathrm{o} = V_\mathrm{o}/I_\mathrm{o}$ | $r_\mathrm{o} = v_\mathrm{o}/i_\mathrm{o} = \mathrm{d}V_\mathrm{o}/\mathrm{d}I_\mathrm{o}$ |

with $V_\mathrm{DC}$ and $I_\mathrm{DC}$ following Ohm's law ($R = V_\mathrm{DC}/I_\mathrm{DC}$) and $v_\mathrm{ac}$ and $i_\mathrm{ac}$ following the small-signal Ohm's law ($r = v_\mathrm{ac}/i_\mathrm{ac}$). The phase shift $\theta$ can be caused by non-resistive elements such as capacitors and coils. See Figure 2.4.

Finally, for elements with input (i) and output signals (o), we can define the concepts, both large-signal and small-signal, as presented in Table 2.II.

## 2.2 Resistance circuits

The simplest of electronic components is the resistor and the simplest of all electronic circuits is then the voltage divider, composed of two resistances in series, as shown in Figure 2.5. The voltage between the resistance is somewhere between the supply voltages at the endpoints. Its exact value can be found easily if we realize that current cannot disappear, e.g., what passes the first resistance *has* to pass the second resistance. Combined with Ohm's law ($R = V/I$) in a rearranged version, $I = \Delta V/R$, with $\Delta V$ the voltage drop across the resistor, and the fact that the total voltage drop over the two resistance must necessarily be equal to the total supply voltage, we can find the voltage at midpoint, $V_\mathrm{x}$. For the current in the top resistance we have

$$I = (V - V_\mathrm{x})/R_1. \tag{2.15}$$

For the current in the second resistance

$$I = (V_\mathrm{x} - 0)/R_2. \tag{2.16}$$

These need to be equal, so

$$\frac{V - V_\mathrm{x}}{R_1} = \frac{V_\mathrm{x}}{R_2}. \tag{2.17}$$

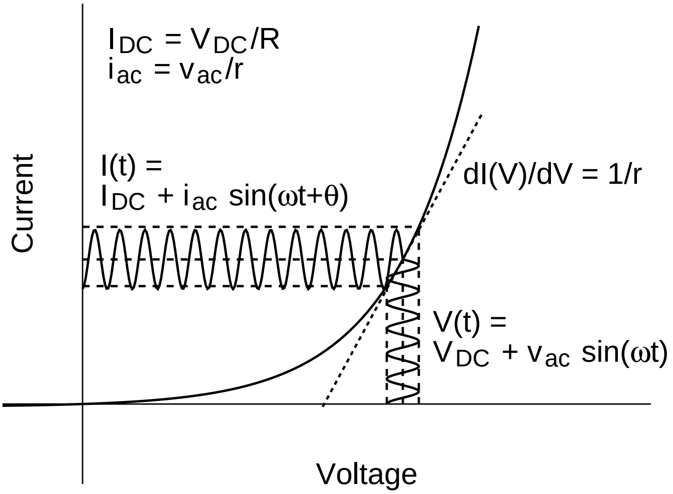


**Fig. 2.4**: Concept of dynamic resistance defined via the derivative of the I-V curve. A small sinusoidal AC signal ($v \sin(\omega t)$) superimposed on top of the bias ($V_{\mathrm{DC}}$) will cause a sinusoidal AC current - on top of the DC current - with the same frequency and with an amplitude $i$ proportional to the derivative of the I-V curve and the amplitude of the input signal, $i = v \times dI(V)/dV = v/r$. For linear electronics, following Ohm's law, the static resistance $R$ and dynamic resistance $r$ are identical. The phase shift θ can be caused by non-resistive elements such as the capacitor and the coil

Thus

$$V_{\mathrm{x}} = V\frac{R_2}{R_1 + R_2}. \tag{2.18}$$

**Question**: A voltage divider composed of $\mathrm{R}_1$ and $\mathrm{R}_2$, instead of having one side connected to ground is connected to $-V$. Use the superposition principle to find the output voltage.
**Answer**: Replace $-V$ with ground and find $V_{\mathrm{o}} = VR_2/(R_1 + R_2)$ as found above, then replace $+V$ with ground and find $V_{\mathrm{o}} = -VR_1/(R_1 + R_2)$. Summing the two gives

$$V_{\mathrm{o}} = V\frac{R_2 - R_1}{R_2 + R_1}. \tag{2.19}$$

An important parameter of the voltage divider is its output resistance. It is not difficult to show that it is equal to the two resistances in parallel. Remember that the output resistance is defined by the voltage drop that occurs when we start drawing current. While this is not so difficult to calculate, even easier is

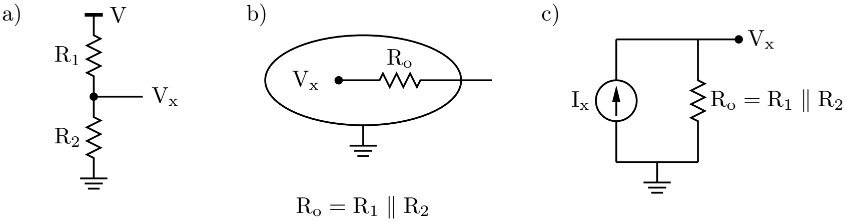


**Fig. 2.5**: a) Simple voltage divider composed of two resistances, $R_1$ and $R_2$ in series. The input voltage $V$ is divided into two parts, halfway the circuit the output voltage is $V_{\mathrm{x}} = V \times R_2/(R_1 + R_2)$. b) Thévenin equivalent, an (ideal) voltage source $V_{\mathrm{x}}$ with in series an output resistance in value equal to the two resistances in parallel. c) Norton equivalent, an (ideal) current source with in parallel a resistance in value equal to the two resistances in parallel

calculating the opposite, if we start drawing current that drops the voltage $V_{\mathrm{x}}$ by $\mathrm{d}V_{\mathrm{x}}$, what was the current $\mathrm{d}I$ that was drawn.

$$\mathrm{d}I_2 = \frac{\mathrm{d}V_{\mathrm{x}}}{R_2}, \tag{2.20}$$

$$\mathrm{d}I_1 = -\frac{\mathrm{d}V_{\mathrm{x}}}{R_1}, \tag{2.21}$$

$$\mathrm{d}I = \mathrm{d}I_1 - \mathrm{d}I_2 = -\mathrm{d}V_{\mathrm{x}}\left(\frac{1}{R_1} + \frac{1}{R_2}\right). \tag{2.22}$$

and finally

$$R_{\mathrm{o}} = -\frac{\mathrm{d}V_{\mathrm{x}}}{\mathrm{d}I} = \frac{1}{1/R_1 + 1/R_2} = R_1 \parallel R_2. \tag{2.23}$$

In other words, a (voltage) signal coming from a voltage divider has an effective output resistance equal to the two resistances in parallel. This is what is called in electronics the Thévenin-equivalent resistance (Theorem of Thévenin: Any black box containing only voltage sources, current sources, and other resistors can be converted to a Thévenin equivalent circuit, comprising exactly one voltage source and one resistor). The circuit behaves as if it is a non-ideal voltage source $V_{\mathrm{x}}$ with output resistance equal to $r_{\mathrm{o}} = R_1 \parallel R_2$. The advantage of this is that from that moment on we do not have to worry about this part of the circuit anymore. We can consider it a black box with parameters $V_{\mathrm{x}}$ and $R_{\mathrm{o}}$, as shown in Figure 2.5b. This is true for any circuit. We can always reduce it to a voltage source with an effective output resistance.

In the same way, any signal can be reduced to a current source outputting a current $I_{\mathrm{x}}$ with an equivalent resistance $R_{\mathrm{o}}$ in parallel. This is the so-called Norton equivalent, as shown in Figure 2.5c.

**Question**: Show that the resistance of the Norton-equivalent circuit is equal to $R_{\mathrm{o}} = R_1 \parallel R_2$; equal to the resistance of the Thévenin equivalent.

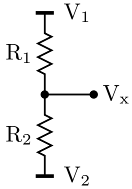


**Fig. 2.6**: Using the superposition principle we can find the voltage at $\mathrm{V_x}$ as the weighted average, $V_{\mathrm{x}} = \alpha V_1 + (1-\alpha)V_2$, with $\alpha = R_2/(R_1+R_2)$

**Answer**: Use the definition of the voltage $V_{\mathrm{x}}$, the open-circuit voltage and short-circuit current, given as, respectively:

$$V_{\mathrm{x}} = V\frac{R_2}{R_1+R_2} = R_{\mathrm{o}} I_{\mathrm{x}}, \tag{2.24}$$

$$I_{\mathrm{x}} = \frac{V}{R_1}. \tag{2.25}$$

Substitution gives

$$R_{\mathrm{o}} = \frac{R_1 R_2}{R_1+R_2} = R_1 \parallel R_2. \tag{2.26}$$

Alternatively, or in addition, we can verify that the voltage drops according to the earlier found value. If we start drawing current $\mathrm{d}I$ by connecting an external circuit, the voltage drops. Actually, this is easily verified, since this part of the current $I_{\mathrm{x}}$ is not passing the resistance $R_{\mathrm{o}}$, thus

$$\mathrm{d}V_{\mathrm{x}} = -\mathrm{d}I R_{\mathrm{o}}, \tag{2.27}$$

$$-\frac{\mathrm{d}V_{\mathrm{x}}}{\mathrm{d}I} = R_{\mathrm{o}}. \tag{2.28}$$

A slightly more advanced voltage divider is one in which both sides have a non-zero voltage, see Fig. 2.6. This results in a weighted average,

$$V_{\mathrm{x}} = \frac{R_2}{R_1+R_2}V_1 + \frac{R_1}{R_1+R_2}V_2 \tag{2.29}$$

$$= \alpha V_1 + (1-\alpha)V_2. \tag{2.30}$$

This can easily be found if we use the superposition principle. First we connect $\mathrm{V_2}$ to ground and calculate the output $V_{\mathrm{x}}$ and then connect $\mathrm{V_1}$ to ground and calculate the output. The total output is the sum of the contributions. The result is a voltage somewhere in between $V_1$ and $V_2$ and closer to the voltage connected to the lower of the two resistances.

In this book we will mostly be talking about voltage signals but sometimes also current signals. One can easily be converted to the other in the calculations.

Whatever comes in handy. In any case we *always* have to bear in mind that any signal at any place of the circuit or system comes from a source that has an effective output resistance, a series resistance in case of a voltage-signal source and a parallel resistance in case of a current source. The good thing is that from that moment on we can consider the signal as coming from a black box with those two parameters $V_\mathrm{x}$ and $R_\mathrm{o}$ or $I_\mathrm{x}$ and $R_\mathrm{o}$.

In view of the above, it is obvious that processing signals from a voltage source is best done by drawing little or no current by connecting an infinite load resistance $R_\mathrm{L} = \infty$, for instance the input resistance of the next amplifier stage. In that case the information (voltage $V_\mathrm{x}$) is not perturbed. In the case of a current signal it is best to connect a zero load resistance, which makes that the information in the form of the current signal $I_\mathrm{x}$ is not lost and totally passed to the next stage.

### 2.2.1 Wheatstone bridge

A common problem in signal processing is the offset of the signal. In case the sensor is a resistance, it is easy to remove this offset by using a Wheatstone bridge, a circuit invented by Samuel Christie and popularized by Charles Wheatstone. It consists of a double voltage divider, see Figure 2.7. One branch contains the unknown resistor to be measured, for instance a sensor resistance whose value depends on the measured quantity, and the other branch contains the reference resistances. Ideally, at the calibration point both voltage-divider branches produce the same voltage halfway and the voltage difference - measured by a multimeter, or further processed - is zero.

**Question**: A temperature-dependent resistance with nominal value of $R_\mathrm{S} = 1$ kΩ is measured with a standard 4-digit digital multimeter with scales ranging from 20 MΩ downto 2 Ω and 200 V downto 2 mV. What is the resolution when the resistance is measured directly and when the resistance is used in a Wheatstone bridge ($V_{++} = 10$ V) together with three resistances of $R = 1$ kΩ?
**Answer**: In the first case, the resolution is equal to the digital resolution of the multimeter. On the scale 2 kΩ the resolution is $\Delta R = 1$ Ω. In the second case, the resolution in resistance is given by the digital voltage resolution of the multimeter divided by the dependency of voltage on resistance, the sensitivity:

$$\Delta R = \frac{\Delta V}{\mathrm{d}V_\mathrm{o}/\mathrm{d}R_\mathrm{S}}. \tag{2.31}$$

In this case we can put the voltage scale to the minimum of 2 mV which has a resolution of $\Delta V = 1$ μV. The output voltage as a function of

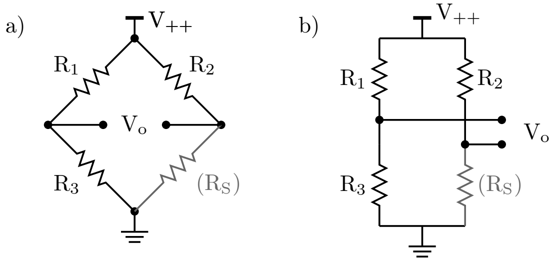


**Fig. 2.7**: Wheatstone bridge shown in two ways, classic and modern. The resistance $R_T$ is drawn gray and shown in brackets to indicate its variability and its function as a sensor. This sensor can be any of the four resistances. The other three resistances have values such that at the calibration point the output voltage difference $V_o$ is zero, for instance all equal to $R_S$

resistance and the derivative of this function are given by

$$V_o(R_T) = \left(\frac{R_S}{R_S+R} - \frac{R}{R+R}\right) \times V_{++} \tag{2.32}$$

$$\begin{aligned} \frac{dV_o}{dR_T} &= \frac{R}{(R_S+R)^2} \times V_{++} \\ &= \frac{1\ \mathrm{k\Omega}}{(1\ \mathrm{k\Omega}+1\ \mathrm{k\Omega})^2} \times 10\ \mathrm{V} \\ &= 2.5\ \mathrm{mV/\Omega} \end{aligned} \tag{2.33}$$

Substituting this value into Equation (2.31) gives $\Delta R = 0.4\ \mathrm{m\Omega}$, 2500 times better than by using direct measurement of resistance.

It can easily be shown that a Wheatstone bridge (and common voltage divider alike) are most sensitive when the resistance in series with the sensor is equal to the nominal sensor resistance, for instance in Figure 2.7 when $R_2 = R_S$. Remember that the sensitivity is given by the derivative of the output as a function of input quantity, in this case the derivative of $V_o(R_S)$,

$$S(R_2) \equiv \frac{dV_o(R_S)}{dR_S} = \frac{R_2}{(R_S+R_2)^2} \times V_{++}. \tag{2.34}$$

This function is plotted in Figure 2.8. To find the maximum, we have to calculate the derivative of $S$ with respect to the variable $R_2$ and set it to zero,

$$\begin{aligned} \frac{dS(R_2)}{dR_2} &= \frac{R_S - R}{(R_S+R)^3} V_{++} = 0 \Rightarrow \\ R_2 &= R_S. \end{aligned} \tag{2.35}$$

Thus we see that the sensitivity of the Wheatstone bridge is maximum if the shunt resistance is equal to the (nominal) resistance of the sensor.

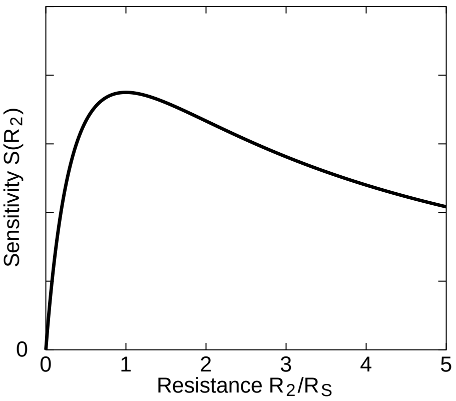


**Fig. 2.8**: Sensitivity as a function of shunt resistance in a Wheatstone bridge or voltage divider

## 2.3 Non-ohmic linear components

The resistors and resistive circuits of the previous section have the property that they all follow Ohm's law, with the current always proportional to the instantaneous voltage applied and nothing more, The current follows the voltage over time, $I(t) = V(t)/R$. Not all electronic elements and circuits behave in this way. As an example, a capacitor is an element that stores charge. When a voltage is applied charge flows into the capacitor — measurable as a current — and once the capacitor is full this current stops since no more charge flows into or out of the capacitor. The total charge $Q$ in a capacitor is given by its capacitance value $C$ and the applied voltage $V$,

$$C \equiv \frac{Q}{V}, \tag{2.36}$$

or

$$Q = CV. \tag{2.37}$$

The relation between current and bias in the time domain can then be found when we realize that current is the passage (derivative) of charge,

$$I(t) \equiv \frac{\mathrm{d}Q}{\mathrm{d}t} = C\frac{\mathrm{d}V}{\mathrm{d}t} \tag{2.38}$$

(assuming that the value of capacitance $C$ does not change with time). Note that the above temporal relation between current and voltage is still linear; doubling the voltage will also cause a doubling of current.

The effect this has on the behavior of the components and circuits can be analyzed in the time domain and in the frequency domain. (As we will see in a moment, the two are linked by the Laplace transform). A simple way to analyze the frequency domain is to see what happens when we apply a single frequency to the capacitor with amplitude $V_0$ and angular frequency $\omega = 2\pi f$. The voltage and current are then, respectively,

$$\begin{aligned} V(t) &= V_0\cos(\omega t), && (2.39)\\ I(t) &= C\frac{\mathrm{d}V(t)}{\mathrm{d}t} = -V_0\omega C\sin(\omega t) \\ &= V_0\omega C\cos(\omega t+\pi/2). && (2.40)\end{aligned}$$

## 2.3.1 Complex numbers; phasors

At this moment it is very useful to use the mathematical trick of complex numbers, with which we can easily calculate this behavior. The voltage of Eq. (2.39) can be written as a phasor (a sine wave with constant amplitude and phase),

$$\begin{aligned} V(t) &= V_0\mathrm{Re}[\cos(\omega t)+j\sin(\omega t)] \\ &= V_\mathrm{o}\mathrm{Re}[\mathrm{e}^{j\omega t}]. && (2.41)\end{aligned}$$

In other words, the voltage oscillation can be written in complex form as

$$V(t) = V_\mathrm{o}\mathrm{e}^{j\omega t}, \tag{2.42}$$

but we should not forget that the real measured current is only the real part of this expression, the imaginary part only helps our calculations. We also see that the current of Eq. (2.40) can be written as

$$\begin{aligned} I(t) &= V_\mathrm{o}\mathrm{Re}[\omega C\mathrm{e}^{j(\omega t+\pi/2)}] \\ &= V_\mathrm{o}\mathrm{Re}[j\omega C\mathrm{e}^{j\omega t}]. && (2.43)\end{aligned}$$

or in complex form,

$$I(t) = V_\mathrm{o}j\omega C\mathrm{e}^{j\omega t}, \tag{2.44}$$

where again we have to remember that the *observable* current is the real part of this expression.

Now, if we want to define a 'resistance' to model this behavior of the capacitor, in analogy to the resistance $R$ that is used in Ohm's law, we see that we can define a complex resistance ('impedance' as it is now called in general terms) equal to

$$Z_\mathrm{C} \equiv \frac{V(t)}{I(t)} = \frac{1}{j\omega C} \tag{2.45}$$

that will make us calculate the current through a capacitor as $I(t) = V(t)/Z_\mathrm{C}$ if the complex forms of voltage and capacitance are used (and the measured current being $\mathrm{Re}[I(t)]$). We see that complex numbers come in very handy when it comes to describing the behavior of non-resistive electronic components. A phase shift of 90° degrees just means a factor $j$. Moreover, taking the derivative adds a phase shift of 90°.

The same analysis we can perform on a coil or inductor. By definition, the inductance of a coil is defined through the electromotive force (EMF, or simply $V$) the coil generates in a changing magnetic field. Since this magnetic field depends linearly on the current through the coil we find the temporal relation between voltage and current as

$$V(t) = L\frac{\mathrm{d}I(t)}{\mathrm{d}t}, \tag{2.46}$$

or, reversing the equation and substituting our voltage of Eq. (2.39)

$$\begin{aligned} I(t) &= \frac{1}{L}\int V(t)\mathrm{d}t \\ &= \frac{1}{L}\int V_\mathrm{o}\cos(\omega t)\mathrm{d}t \\ &= \frac{1}{\omega L}V_\mathrm{o}\sin(\omega t) \\ &= \frac{1}{\omega L}V_\mathrm{o}\cos(\omega t - \pi/2). \end{aligned} \tag{2.47}$$

In complex numbers,

$$I(t) = \frac{V_\mathrm{o}}{j\omega L}\mathrm{e}^{j\omega t}, \tag{2.48}$$

where we used $j = \mathrm{e}^{j\pi/2}$ and the definition of $j$: $j^2 = -1$ and thus $1/j = -j$. The above equation allows for the definition of a complex resistance of a coil equal to (Eq. (2.39) and Eq. (2.48))

$$Z_\mathrm{L} \equiv \frac{V(t)}{I(t)} = j\omega L. \tag{2.49}$$

With these equations we can rapidly calculate the behavior of any circuit. We start with some simple examples of passive filters in the next section. See Table 2.III for the definitions of components in terms of impedance. As mentioned in the beginning of this chapter, the reciprocal of resistance $R$ is conductance $G = 1/R$. Likewise, the reciprocal for impedance $Z$ is called admittance $Y = 1/Z$. Table 2.IV gives a complete nomenclature used in what is called admittance spectroscopy. The last elements of this list ($\sigma$, $\rho$, $\varepsilon$, $K$ and $\chi$) are physical material parameters that will be discussed in Chapter 3.

The analysis of electronic circuits with the help of imaginary numbers is normally called phasor analysis, short for phase vector. It is important to note that this analysis is a purely mathematical aid. All measurable quantities are always real. The imaginary part of current, voltage, charge, or whatever, *never* exists in the circuit at any time. That is why it is called 'imaginary'.

### 2.3.2 Passive filters

In some cases we will want to remove noise or unwanted frequencies before continuing to process the signal. A simple way to do this is by filtering. Figure

**Table 2.III**: Impedance and admittance parameters of the resistor, capacitor and coil

| Component | Symbol | Impedance | Admittance |
|---|---|---|---|
| Resistance | R | $Z_{\mathrm{R}} = R$ | $G_{\mathrm{R}} = 1/R$ |
| Capacitance | C | $Z_{\mathrm{C}} = 1/j\omega C$ | $G_{\mathrm{C}} = j\omega C$ |
| Inductance | L | $Z_{\mathrm{L}} = j\omega L$ | $G_{\mathrm{L}} = 1/j\omega L$ |

**Table 2.IV**: Admittance spectroscopy device and material parameters and their relations. $A$ is cross section area of device, $d$ is device length. (From: Stallinga, *Electrical characterization of organic electronic materials and devices*).

| *Param.* | *Relation(s)* | *Name* | *Unit* |
|---|---|---|---|
| $Y$ | $Y = G + jB$<br>$Y = \mathrm{d}I/\mathrm{d}V$ | Admittance | S |
| $\lvert Y\rvert$ | $\lvert Y\rvert = \sqrt{G^2 + B^2}$ | Admittance magnitude | S |
| $Z$ | $Z = R + jX$<br>$Z = 1/Y$<br>$Z = \mathrm{d}V/\mathrm{d}I$ | Impedance | Ω |
| $G$ | $G = \mathrm{Re}(Y)$ | Conductance | S |
| $B$ | $B = \mathrm{Im}(Y)$ | Susceptance | S |
| $R$ | $R = \mathrm{Re}(Z)$ | Resistance | Ω |
| $X$ | $X = \mathrm{Im}(Z)$ | Reactance | Ω |
| $C$ | $Y_{\mathrm{C}} = j\omega C$<br>$Z_{\mathrm{C}} = -j/\omega C$<br>$C = Q/V$<br>$C = \mathrm{d}Q/\mathrm{d}V$ | Capacitance<br><br>(static; DC)<br>(dynamic; AC) | F |
| $L$ | $Y_{\mathrm{L}} = -j/\omega L$<br>$Z_{\mathrm{L}} = j\omega L$ | Inductance | H |
| $L$ | $L = G/\omega$<br>$L = 1/\omega R$ | Loss | F |
| $\tan\delta$ | $\tan\delta = L/C$<br>$\tan\delta = 1/\omega RC$ | Loss-tangent | - |
| $\sigma$ | $(d/A)G$ | Conductivity | S/m |
| $\rho$ | $(A/d)R$ | Resistivity | Ωm |
| $\varepsilon$ | $(d/A)C$ | Permittivity | F/m |
| $\varepsilon_{\mathrm{r}}$ | $\varepsilon = \varepsilon_{\mathrm{r}}\varepsilon_0$<br>$\varepsilon_{\mathrm{r}} = K$ | Dielectric constant | - |
| $\chi$ | $\varepsilon_{\mathrm{r}} = (1 + \chi)$ | Susceptibility | - |

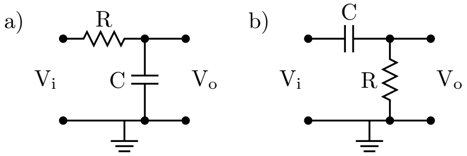


**Fig. 2.9**: Low-pass (a) and high-pass (b) passive RC filters

2.9 shows two examples of filters composed of a resistor-capacitor combination. These are passive filters since they do not contain amplifying elements and the output signal amplitude is for any frequency always lower than the input amplitude. In the previous section we have seen how capacitors and coils can be described with complex numbers. We will now use these definitions to calculate the frequency behavior of these filters.

As a first remark, analyzing the complex resistance (impedance) value we can make some very useful observations. It is clear from Table 2.III that the effective resistance of a capacitor is infinite at 0 Hz (DC) and vanishes for high frequencies. We can thus make a simple observation that will help us to rapidly determine the qualitative behavior of any circuit containing capacitors:

A capacitor can be considered open-circuit for low frequencies and short-circuit for high frequencies.

In other words, for low frequencies the capacitor is effectively pulled out of the circuit and for high frequencies replaced by a wire. Considering the fact that, upon closer examination, the circuits in Figure 2.9 are nothing more than simple voltage dividers, we can immediately say that for low frequencies, where the capacitor effectively are open circuits, the left circuit passes the entire input signal to the output and the right circuit has $V_\text{o}$ connected to (only) ground. On the other hand, for high frequencies, where the capacitors are effectively short-circuits, in the left circuit the output is shorted to ground and the right circuit is shorted to the input signal. We can thus label the left circuit a low-pass filter (LPF) since at low frequencies it copies the input and at high frequencies the signal is filtered off, and the right circuit a high-pass filter (HPF) since it passes the input signal at high frequencies and at low frequencies the signal is attenuated.

To mathematically prove this we use the voltage divider equation derived before (Eq. (2.18)), substituting one of the resistances for a capacitor. For the LPF this is

$$\begin{aligned} H_\text{LPF}(\omega) \equiv \frac{v_\text{o}(\omega)}{v_\text{i}(\omega)} &= \frac{Z_\text{C}}{Z_\text{C}+Z_\text{R}} = \frac{1/j\omega C}{1/j\omega C + R} \\ &= \frac{1}{1+j\omega/\omega_0}, \end{aligned} \tag{2.50}$$

$$H_\text{LPF}(f) = \frac{1}{1+jf/f_0}, \tag{2.51}$$

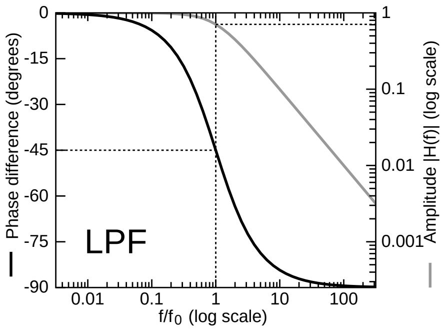


**Fig. 2.10**: Low-pass passive RC filter frequency response

with $v_\mathrm{i}$ and $v_\mathrm{o}$, the *amplitude* of the voltage signals, $\omega_0$ ($f_0$) the cut-off frequency, $\omega_0 = 1/\tau$ ($f_0 = \omega_0/2\pi = 1/2\pi RC$), with $\tau = RC$ the relaxation time - often called ’RC time’ - of the circuit. (Note that $\tau$ has units of time). This transfer function $H$ has two parts, a real part and an imaginary part. The real part represents the output signal that is in-phase with the input signal and the imaginary part is the 90°-out-of-phase signal. If the input signal is $V_\mathrm{i}(t) = \sin\omega t$, the output signal is given by $V_\mathrm{o}(t) = v_0 \sin\omega t + v_{90}\cos\omega t$. Together with the total amplitude $|v|$, and phase $\phi = \angle V_\mathrm{o}, V_\mathrm{i}$, these follow

$$v_0(f) = \mathrm{Re}(H(f)) = \frac{1}{1+(f/f_0)^2}, \tag{2.52}$$

$$v_{90}(f) = \mathrm{Im}(H(f)) = -\frac{f/f_0}{1+(f/f_0)^2}, \tag{2.53}$$

$$|v(f)| = \sqrt{v_0^2 + v_{90}^2} = |(H(f)| = \frac{1}{\sqrt{1+(f/f_0)^2}}, \tag{2.54}$$

$$\phi(f) = \tan^{-1}\left(\frac{v_{90}(f)}{v_0(f)}\right) = -\tan^{-1}\left(\frac{f}{f_0}\right). \tag{2.55}$$

At low frequencies ($f \ll f_0$) the real (in-phase) part is unity and the imaginary part is zero; at low frequencies the output copies the input. At high frequencies ($f \gg f_0$) the out-of-phase signal is much larger than the in-phase signal. The amplitude of the signal drops linearly with the frequency $|v_\mathrm{o}/v_\mathrm{i}| \propto 1/f$. At the cut-off frequency $f = f_0$ the in-phase and out-of-phase parts are equal and the phase is therefore $\phi(f_0) = -45°$. Moreover, substituting this frequency above shows that there the amplitude is $|v(f)| = 1/\sqrt{2}$, which is equal to $-3$ dB. Since the current delivered by the circuit, equal to $I_\mathrm{o} = V_\mathrm{o}/R_\mathrm{L}$ (with $R_\mathrm{L}$ the load resistance), also decreases a factor $\sqrt{2}$, and the power is voltage times current, the output power drops a factor 2 at the cut-off frequency (which equals $-6$ dB).

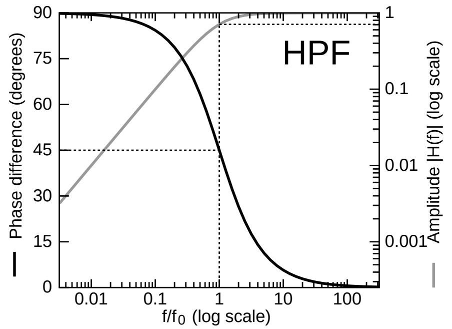


**Fig. 2.11**: High-pass passive RC filter frequency response

We can make the same analysis for the high-pass filter:

$$H_{\text{HPF}}(\omega) \equiv \frac{v_{\text{o}}(\omega)}{v_{\text{i}}(\omega)} = \frac{R}{R+1/j\omega C}$$

$$= \frac{1}{1-j\omega_0/\omega}, \tag{2.56}$$

$$H_{\text{HPF}}(f) = \frac{1}{1-jf_0/f}, \tag{2.57}$$

with $\omega_0$ and $f_0$ as defined before, $\omega_0 = 1/\tau = 1/RC$, etc. Note the different sign in the denominator term above, caused by the fact that $1/j = -j$. The in-phase, out-of-phase, amplitude and phase angle for a high-pass filter are given by, respectively,

$$v_0(f) = \text{Re}(H(f)) = \frac{1}{1+(f_0/f)^2}, \tag{2.58}$$

$$v_{90}(f) = \text{Im}(H(f)) = +\frac{f_0/f}{1+(f_0/f)^2}, \tag{2.59}$$

$$|v(f)| = \sqrt{v_0^2+v_{90}^2} = |(H(f)| = \frac{1}{\sqrt{1+(f_0/f)^2}}, \tag{2.60}$$

$$\phi(f) = \tan^{-1}\left(\frac{v_{90}(f)}{v_0(f)}\right) = +\tan^{-1}\left(\frac{f_0}{f}\right). \tag{2.61}$$

At high frequencies ($f \gg f_0$) the real (in-phase) part is unity and the imaginary part is zero; at high frequencies the output copies the input. At low frequencies ($f \ll f_0$) the out-of-phase signal is much larger than the in-phase signal. The amplitude of the signal drops linearly with the frequency $|v_{\text{o}}/v_{\text{i}}| \propto f$. At the cut-off frequency $f = f_0$ the in-phase and out-of-phase parts are equal and the phase is therefore $\phi(f_0) = 45°$. Again, substituting this frequency above shows that there the amplitude is $|v(f)| = 1/\sqrt{2}$ ($-3$ dB, or $-6$ dB in power).

### 2.3.3 Laplace transform

There is a price to pay for filtering signals. Or, to say it differently, there is a flip side of these filters. This becomes obvious when we analyze the filters in the time-domain instead of the frequency domain. This can be done with Laplace transforms.

$$F(s) = \int_0^\infty f(t)e^{-st}\mathrm{d}t, \tag{2.62}$$

with $s$ defined as $s = j\omega$. Without going into details about these transforms, suffices to say that if a filter has a response (transfer) function in frequency, it will have a temporal response to input signals, as given by the inverse Laplace transform. With the table of Laplace transformations in hand (Table 2.V) we can analyze what will happen to an input signal. For example a step in the input signal at $t = 0$,

$$u(t) = \left\{ \begin{array}{ll} 0 & \text{for } t < 0 \\ 1 & \text{for } t \geq 0 \end{array} \right. . \tag{2.63}$$

The Laplace transform of this function is given by $1/s$. If we multiply this by the transfer function $H_{\mathrm{LPF}}(s)$ we derived before, Eq. (2.50), we find the response in the frequency domain

$$F^u_{\mathrm{LPF}}(s) = \frac{\omega_0}{s(s+\omega_0)}. \tag{2.64}$$

Looking in the table, this corresponds to the exponential-approach function in time

$$f^u_{\mathrm{LPF}}(t) = (1 - \mathrm{e}^{-t/\tau})u(t), \tag{2.65}$$

with, as before, $\tau \equiv RC \equiv 1/\omega_0$. The response of a low-pass filter to a step in the signal is that it initially maintains the signal from before the step, but exponentially approaches the new value, see Figure 2.12. The characteristic time at which it follows the signals is the RC time.

We can make the same analysis for the high-pass filter. The step function in time $u(t)$, Laplace-transformed to the frequency domain (see Table 2.V), and multiplied by the transfer function $H_{\mathrm{HPF}}(s)$ (Eq. (2.56)), results in the response in the frequency domain equal to

$$H^u_{\mathrm{HPF}}(s) = \frac{1}{s+\omega_0}, \tag{2.66}$$

and this corresponds to the exponential-decay function in the time domain,

$$f^u_{\mathrm{HPF}}(t) = \mathrm{e}^{-t/\tau}u(t), \tag{2.67}$$

with, as before, $\tau \equiv RC \equiv 1/\omega_0$. The response of a high-pass filter to a step in the signal is that it instantaneously copies the new signal (or better to say signal *increase*), but then exponentially drops back to zero, see Figure 2.12. The characteristic time at which it drops is the RC time. This is useful if we

**Table 2.V**: Laplace transforms of some relevant basic functions and some examples. $F(s) = \mathcal{L}f(t)$, $G(s) = \mathcal{L}g(t)$, $s$ is defined as $s = j\omega$, $\delta(x)$ is zero everywhere except for $x = 0$ and its integral over all $x$ is unity

| Name | Time domain | Frequency domain |
|---|---|---|
| | $f(t)$ | $F(s)$ |
| Laplace | $f(t)$ | $\int f(t)\mathrm{e}^{-st}\mathrm{d}t$ |
| Inverse Laplace | $\int F(s)\mathrm{e}^{st}\mathrm{d}s$ | $F(s)$ |
| Sum | $af(t) + bg(t)$ | $aF(s) + bF(s)$ |
| Dirac-delta at $t = 0$ | $\delta(t)$ | $1$ |
| Unity | $1$ | $\delta(s)$ |
| Time derivative | $\mathrm{d}f(t)/\mathrm{d}t$ | $sF(s)$ |
| Time integral | $\int f(t)\mathrm{d}t$ | $F(s)/s$ |
| Delay | $f(t-b)$ | $\mathrm{e}^{-sb}F(s)$ |
| Frequency shift | $\mathrm{e}^{j\omega_0 t}f(t)$ | $F(s - j\omega_0)$ |
| Frequency differentiation | $tf(t)$ | $-\mathrm{d}F(s)/\mathrm{d}s$ |
| Unit step at $t = 0$ a.k.a. Heaviside | $u(t) = \int_{-\infty}^{t} \delta(t')\mathrm{d}t'$ | $1/s$ |
| Exponential decay | $\mathrm{e}^{-at}u(t)$ | $1/(s+a)$ |
| Exponential approach | $(1 - \mathrm{e}^{-at})u(t)$ | $a/s(s+a)$ |

want to only monitor changes; any offset, or slowly drifting levels are filtered off.

In conclusion: filters are good tools to remove unwanted frequencies in our signal, but we should be careful to not also filter out the information. Apart from that we have to bear in mind that we can introduce a delay in the response and in some cases this delay can be critical.

Remains only to prove the frequency-dependent impedance of a capacitor used above (Eq. (2.45)) in the framework of Laplace transforms. The capacitor is by definition an element that can store charge. The capacitance is defined as the amount of charge the device can store per unit voltage.

$$C \equiv \frac{Q}{V}. \tag{2.68}$$

Current is the movement of charge and is thus defined as the amount of charge passing per second. Applied to our capacitor this becomes the charge going in or coming out of the capacitor per second,

$$I(t) \equiv \frac{\mathrm{d}Q(t)}{\mathrm{d}t} = \frac{\mathrm{d}[C(t)V(t)]}{\mathrm{d}t} = C\frac{\mathrm{d}V(t)}{\mathrm{d}t}. \tag{2.69}$$

To this we can apply the Laplace transform on both sides of the equation. With the help of the knowledge that the Laplace transform of the derivative of a function equals $s$ times the Laplace transform of the function itself (see Table 2.V) we get

$$I(s) = sCV(s) \tag{2.70}$$

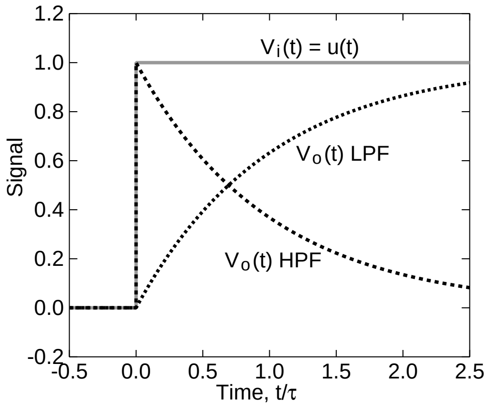


**Fig. 2.12**: Low-pass filter (LPF) and high-pass filter (HPF) repsonse to a step in the signal, $u(t)$

(The reusing of the function name may be confusing; if the argument is $s$, the function is the Laplace transform of the function with the argument $t$). The resistance is given by Ohm's law as the ratio of voltage and current, the Laplace transform of this resistance for a capacitor thus being

$$Z_{\mathrm{C}}(s) \equiv \frac{V(s)}{I(s)} = \frac{1}{sC}. \tag{2.71}$$

Substituting $s = j\omega$ directly yields Equation (2.45).

**Question**: Using Laplace transforms, turn the physical relation $V(t) = L \times \mathrm{d}I(t)/\mathrm{d}t$ for the inductance, Eq. (2.46), into the impedance expression $Z_{\mathrm{L}} = sL = j\omega L$ (Eq. 2.49).
**Answer**: Starting with the physical relation and applying Laplace transforms on both sides of the equal sign we get

$$\mathcal{L}[V(t)] = \mathcal{L}\left[L\frac{\mathrm{d}I(t)}{\mathrm{d}t}\right], \tag{2.72}$$

$$V(s) = sLI(s), \tag{2.73}$$

$$Z_{\mathrm{L}}(s) \equiv \frac{V(s)}{I(s)} = sL, \tag{2.74}$$

$$Z_{\mathrm{L}}(\omega) = j\omega L, \quad \text{q.e.d.} \tag{2.75}$$

Finally, alternatively — slightly more complex, and the reason we delayed it until here — the same circuit can be fully analyzed in the time domain. In fact, the time-domain analysis is the only one that has a physical basis! The others — frequency domain, complex numbers, Laplace transforms — are all mathematical tricks.It is therefore also useful to do the time-domain analysis, even if it is just once. As an example the LPF: The circuit is a voltage divider where the current through the resistor and the capacitor is given by Ohm's law and Eq. (2.69), respectively:

$$\begin{aligned} I_{\mathrm{R}} &= \frac{V_{\mathrm{i}}(t) - V_{\mathrm{o}}(t)}{R}, && (2.76) \\ I_{\mathrm{C}} &= C\frac{\mathrm{d}V_{\mathrm{o}}(t)}{\mathrm{d}t}. && (2.77) \end{aligned}$$

Since current cannot escape, these necessarily have to be equal. This results in a differential equation that can be solved. If the input voltage is a sinusoid, $V_{\mathrm{i}}(t) = \sin \omega t$, the trial solution is

$$V_{\mathrm{o}}(t) = A \sin \omega t + B \cos \omega t. \tag{2.78}$$

Substituting this in the equality $I_{\mathrm{R}} = I_{\mathrm{C}}$:

$$\sin \omega t - (A \sin \omega t + B \cos \omega t) = RC(\omega A \cos \omega t - \omega B \sin \omega t). \tag{2.79}$$

The solution for all $t$ is

$$\begin{aligned} 1 - A &= -\omega RCB, && (2.80) \\ -B &= \omega RCA. && (2.81) \end{aligned}$$

The amplitudes for the in-phase and out-of-phase signal are respectively given by

$$\begin{aligned} A &= \frac{1}{1 + \omega^2 (RC)^2}, && (2.82) \\ B &= -\frac{\omega RC}{1 + \omega^2 (RC)^2}, && (2.83) \end{aligned}$$

as found before (compare to Eqs. (2.52) and (2.53), $\omega \equiv 2\pi f$, $\omega_0 \equiv 1/RC$).

**Question**: Prove Eqs. (2.58) and (2.59) for the HPF.
**Answer**: The circuit is a voltage divider where the current through the resistor and the capacitor is given by Ohm's law and Eq. (2.69), respectively:

$$\begin{aligned} I_{\mathrm{R}} &= \frac{V_{\mathrm{o}}(t)}{R}, && (2.84) \\ I_{\mathrm{C}} &= C\frac{\mathrm{d}[V_{\mathrm{i}}(t) - V_{\mathrm{o}}(t)]}{\mathrm{d}t}. && (2.85) \end{aligned}$$

For the sinusoidal input voltage, $V_{\mathrm{i}}(t) = \sin \omega t$, we can substitute the same trial solution as used for the LPF, Eq. (2.78), in the equality $I_{\mathrm{R}} = I_{\mathrm{C}}$:

$$A \sin \omega t + B \cos \omega t = RC(\omega \cos \omega t - \omega A \cos \omega t + \omega B \sin \omega t). \tag{2.86}$$

The solution for all $t$ is

$$\begin{aligned} A &= \omega RCB, && (2.87) \\ B &= \omega RC(1-A). && (2.88) \end{aligned}$$

The amplitudes for the in-phase and out-of-phase signal are respectively given by

$$\begin{aligned} A &= \frac{\omega^2(RC)^2}{1+\omega^2(RC)^2} = \frac{1}{1+1/\omega^2(RC)^2}, && (2.89) \\ B &= \frac{\omega RC}{1+\omega^2(RC)^2} = \frac{1/\omega RC}{1+1/\omega^2(RC)^2}. && (2.90) \end{aligned}$$

### 2.3.4 Bode plots and Nyquist plots

A rapid and informative way of showing the frequency behavior of components and circuits is with Bode plots and Nyquist plots. A Bode plot is the logarithm of the absolute value of the transfer function (the ’gain’) plotted versus the logarithm of the frequency, often accompanied by a plot of the phase between input and output signals vs. the logarithm of frequency. A Nyquist plot is the imaginary part of the transfer function plotted versus its real part. The Bode plot we have already seen in the previous section. It will be discussed here in more detail using an impedance voltage divider (similar to the filters of the previous section X). It is composed of a resistance in series with a coil, with the output taken midway, see Figure 2.13.

Before we continue, we can already globally say how the circuit will behave, similarly to the way we analyzed the capacitor. Looking at the impedance of the inductor, Eq. (2.49), we can see that the impedance vanishes for low frequencies (DC), where it behaves like a simple wire, and is infinite for high frequencies, where it effectively becomes an open circuit. Applying this to our circuit we can expect that the transfer function is 1 for low frequencies and 0 for high frequencies and it is thus equivalent to the LPF based on a resistor-capacitor pair. This can also be formally calculated by looking at the explicit form of the transfer function.

An inductor can be considered open-circuit for high frequencies and short-circuit for low frequencies.

The transfer function between input and output is given by the voltage division,

$$\begin{aligned} H(\omega) &\equiv \frac{v_\mathrm{o}}{v_\mathrm{i}} = \frac{Z_\mathrm{R}}{Z_\mathrm{R}+Z_\mathrm{L}} \\ &= \frac{R}{R+j\omega L}. && (2.91) \end{aligned}$$

The magnitude and phase of this transfer function are respectively given by

$$\begin{aligned} |H(\omega)| &= \frac{1}{\sqrt{1+(\omega L/R)^2}}, & (2.92) \\ \phi &= -\tan^{-1}(\omega L/R). & (2.93) \end{aligned}$$

The Bode plot and phase plot are shown in Figure 2.13. We can make some interesting observations:

- At low frequencies ($\omega \ll R/L$, $\omega \ll RC$ for LPFs based on capacitors), the transfer function is close to 1 and independent of frequency.

- At high frequencies ($\omega \gg R/L$, $\omega \gg RC$ for LPFs based on capacitors) the transfer function is proportional to $1/\omega$. This means that the slope of the Bode plot is $-1$, since $\log_{10}(\omega^{-1}) = -\log_{10}(\omega)$. Looking at it another way, if the frequency increases by a factor 10 (a 'decade'), the transfer function drops a factor 10. A drop of a factor 10 is $-1$ bel, or $-10$ decibel (dB). The voltage amplitude at the output thus drops 10 db/decade. If the voltage drops a factor (10 dB), the power, the product of voltage and current ($P = VI = V^2/R$) then drops a factor 100 (20 dB). Classic electronics textbooks therefore call such behavior (after a cut-off frequency) $-20$ dB/decade, to describe the power drop with frequency.

- At the cut-off frequency, $\omega_0 = R/L$ (or $\omega_0 = RC$), the magnitude is exactly $1/\sqrt{2}$, as substitution of the cut-off frequency in the transfer function easily shows. This is equivalent to $-1.5$ dB or $-3$ dB power, the reason why this frequency is also called the 3-dB point.

- At the cut-off frequency the phase is $-\pi/4$ ($-45^\circ$), as substitution of $\omega_0$ in Equation (2.93) easily shows. In fact, a rapid way of finding the cut-off frequency is by looking for $45^\circ$-phase shifts. The phase starts changing more-or-less a decade before the cut-off frequency and reaches its final value ($-90^\circ$ in this case) one decade after the cut-off frequency.

**Question**: Some books, instead of 'decades' (factors 10) prefer to express the frequencies in terms of 'octaves' (factors 2). What is the value of the slope in decibels per octave?
**Answer**: 6 dB/octave, since $10 \times \log_{10}(2^{-2}) = 6$.

On basis of these observations we can rapidly construct the Bode plot and phase plot in a schematic way, see Fig. 2.13.

The same analysis can be done for a high-pass filter, see Fig. 2.14. The transfer function between input and output is given by the voltage division,

$$\begin{aligned} H(\omega) &\equiv \frac{v_\mathrm{o}}{v_\mathrm{i}} = \frac{Z_\mathrm{L}}{Z_\mathrm{L}+Z_\mathrm{R}} \\ &= \frac{j\omega L}{j\omega L + R}. & (2.94) \end{aligned}$$

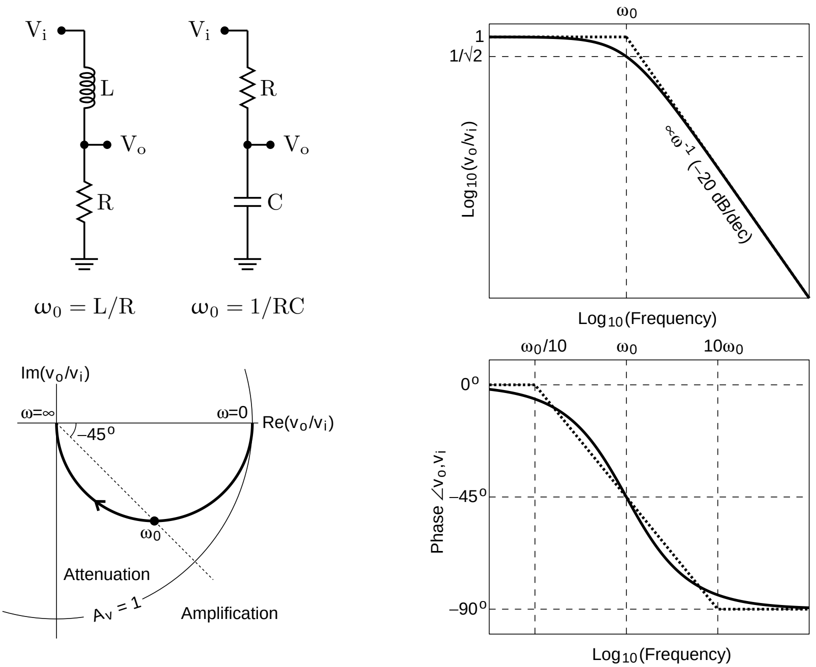


**Fig. 2.13**: Frequency behavior of a LPF made of an impedance voltage divider circuit. Top left: LPF circuits, resistor and coil or resistor and capacitor, with cut-off frequencies $\omega_0 = L/R$ and $\omega_0 = 1/RC$ respectively. Bottom left: Nyquist plot of real and imaginary part of transfer function $v_o/v_i$. The large circle shows the limit of voltage gain ($A_v = 1$). Passive circuits, including the ones shown here, always are completely within this circle. At low frequencies the gain is unity, at high frequencies it is zero. At the cut-off frequency $\omega_0$ the phase is $-45^o$. Top right: Bode plot (logarithm of gain vs. logarithm of frequency. Below the cut-off frequency $\omega_0$ the gain is nearly independent of frequency, far above it the gain drops of proportional to the frequency (voltage gain $-10$ dB/decade, power gain $-20$ dB/decade). At the cut-off frequency the voltage gain is $1/\sqrt{2}$. The behavior can be approximated by the dashed curve in cases we want to make rapid, schematic drawings of the behavior. Bottom right: Accompanying phase plot (phase vs. logarithm of frequency). At low frequencies the output is in-phase with the input. For high frequencies the phase difference is $-90^o$. At the cut-off frequency the phase is $-45^o$. The behavior can be approximated by the dashed curve; the phase starts dropping one decade before $\omega_0$ and reaches its final value one decade after $\omega_0$

The magnitude and phase of this transfer function are respectively given by

$$
\begin{aligned}
|H(\omega)| &= \frac{1}{\sqrt{1+(R/\omega L)^2}}, && (2.95)\\
\phi &= \tan^{-1}(R/\omega L). && (2.96)
\end{aligned}
$$

The Bode plot and phase plot are shown in Figure 2.14. Also here we can make some interesting observations:

- At high frequencies ($\omega \gg R/L$, $\omega \gg RC$ for HPFs based on capacitors), the transfer function is close to 1 and independent of frequency.

- At low frequencies ($\omega \ll R/L$, $\omega \ll RC$ for LPFs based on capacitors) the transfer function is proportional to $\omega$. This means that the slope of the Bode plot is $+1$. Looking at it another way, if the frequency increases by a factor 10 (a 'decade'), the transfer function also increases a factor 10. An increase of a factor 10 is 1 bel, or 10 decibel (dB). The voltage amplitude at the output thus increases 10 db/decade. If the voltage increases a factor (10 dB), the power, the product of voltage and current ($P = VI = V^2/R$) then increases a factor 100 (20 dB). Classic electronics textbooks therefore call such behavior (before a cut-off frequency) 20 dB/decade, to describe the behavior of power with frequency.

- At the cut-off frequency, $\omega_0 = R/L$ (or $\omega_0 = RC$), the magnitude is exactly $1/\sqrt{2}$, as substitution of the cut-off frequency in the transfer function easily shows. This is equivalent to $-1.5$ dB or $-3$ dB power, the reason why this frequency is also called the 3-dB point. (Equal to LPF).

- At the cut-off frequency the phase is $\pi/4$ ($+45^\circ$), as substitution of $\omega_0$ in Equation (2.96) easily shows. In fact, a rapid way of finding the cut-off frequency is by looking for $45^\circ$-phase shifts. The phase starts changing more-or-less a decade after the cut-off frequency and reaches its final value ($+90^\circ$ in this case) one decade below the cut-off frequency.

## 2.4 Non-linear components

Non-linear components are elements that don't follow the rule given before that increasing the input signal a factor will also cause the output signal to increase the same factor. The most famous are the diode and the transistor in all their various types. See Figure 2.15 for examples of non-linear electronic components. They will be discussed in this section.

### 2.4.1 Diodes

A diode is a non-linear element that has a current that highly depends on the voltage applied. Without, at this moment, going into detail about how that in

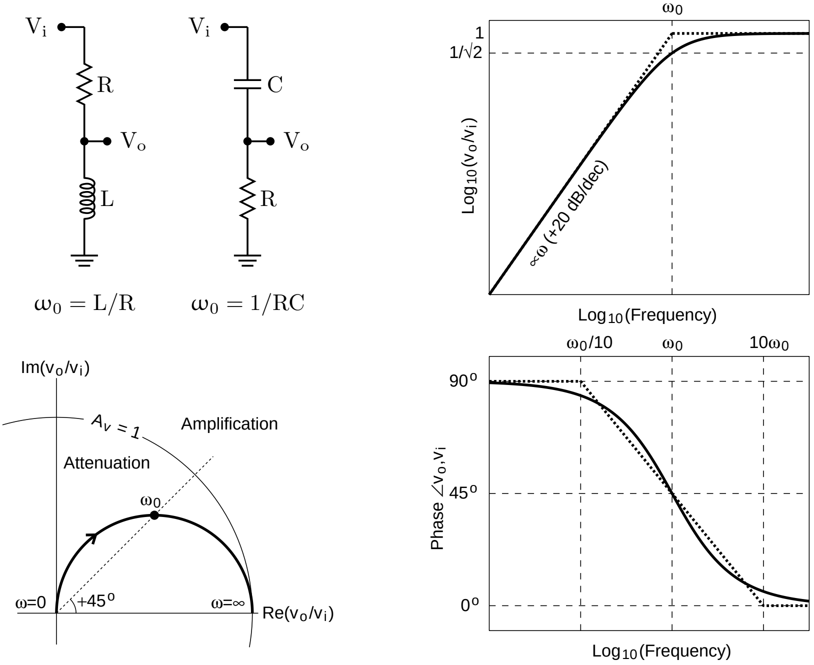


**Fig. 2.14**: Frequency behavior of an HPF made of an impedance voltage divider circuit. Top left: HPF circuits, resistor and coil or resistor and capacitor, with cut-off frequencies $\omega_0 = L/R$ and $\omega_0 = 1/RC$ respectively. Bottom left: Nyquist plot of real and imaginary part of transfer function $v_\mathrm{o}/v_\mathrm{i}$. The large circle shows the limit of voltage gain ($A_\mathrm{v} = 1$). All passive circuits, including the ones described here, always fall completely within this circle. At high frequencies the gain is unity, at low frequencies it is zero. At the cut-off frequency $\omega_0$ the phase is $+45^\mathrm{o}$. Top right: Bode plot (logarithm of gain vs. logarithm of frequency. Above the cut-off frequency $\omega_0$ the gain is unity and nearly independent of frequency, far below it the gain drops of proportional to the frequency (voltage gain $10$ dB/decade, power gain $20$ dB/decade). At the cut-off frequency the voltage gain is $1/\sqrt{2}$. The behavior can be approximated by the dashed curve in cases we want to make rapid, schematic drawings of the behavior. Bottom right: Accompanying phase plot (phase vs. logarithm of frequency). At high frequencies the output is in-phase with the input. For low frequencies the phase difference is $+90^\mathrm{o}$. At the cut-off frequency the phase is $+45^\mathrm{o}$. The behavior can be approximated by the dashed curve; the phase starts rising one decade after $\omega_0$ and reaches its final value one decade below $\omega_0$

**Table 2.VI**: Bias of a diode needed for various currents; A conducting diode under normal conditions has a voltage drop of about 0.7 V. ($I_\mathrm{S} = 10^{-14}$ A)

| Current | Voltage |
|---|---|
| 1 μA | 0.48 V |
| 1 mA | 0.66 V |
| 1 A | 0.83 V |

fact is accomplished, we give here the general function of current vs. voltage, which is called the Ebers-Moll equation

$$I(V) = I_\mathrm{S}\left(\mathrm{e}^{V/V_\mathrm{T}} - 1\right). \tag{2.97}$$

In this $V_\mathrm{T}$ is called the thermal voltage which depends on temperature, and is 26 mV for a diode at room temperature. $I_\mathrm{S}$ is what is called the reverse-bias saturation current, a name that makes sense, because if we substitute an infinite negative voltage, we see that the current saturates at this value. This current also depends on the temperature and we will see that a simple diode can be used as a temperature sensor that translates temperature to current. For normal temperatures, $I_\mathrm{S}$ is extremely small, in the order of $10^{-14}$ A. The smaller the better. Ideally we would like a diode to not conduct at all in reverse bias.

We see that the current grows rapidly - exponentially - with the voltage applied. Reasoning the other way around, we see that if we have any reasonable current, the voltage must be in the order of $V = 0.7$ V, see Table 2.VI. Ignoring the term $-1$ — which is anyway insignificant for measurable currents — we see that the current grows a factor $e = 2.72$ for every 26 mV increase in voltage. This rapidly becomes very large for voltages above 0.7 V, or immeasurable for voltages below 0.7 V. As an electronics engineer — engineering consists of knowing what approximations to make — we can safely say that "If there is current, the voltage drop of a diode is 0.7 V".

A (silicon) diode that is conducting current has a voltage drop of 0.7 V.

Another way at looking at a diode is that it passes a current only in one direction. If we ignore the voltage drop of 0.7 V we can say that the resistance of a diode is infinite for negative voltages and zero for positive voltages. It thus resembles a one-way street; current can only flow in one direction and is blocked in the other direction. The symbol for the diode incorporates this one-way idea; the current flows only in the direction of the arrow and is blocked in the other direction.

Special diodes are the light-emitting diode (LED), the Schottky diode and the Zehner diode. The LED follows the same I-V behavior, but because of it based on other materials (like GaAs or InP), it has a larger voltage drop, see

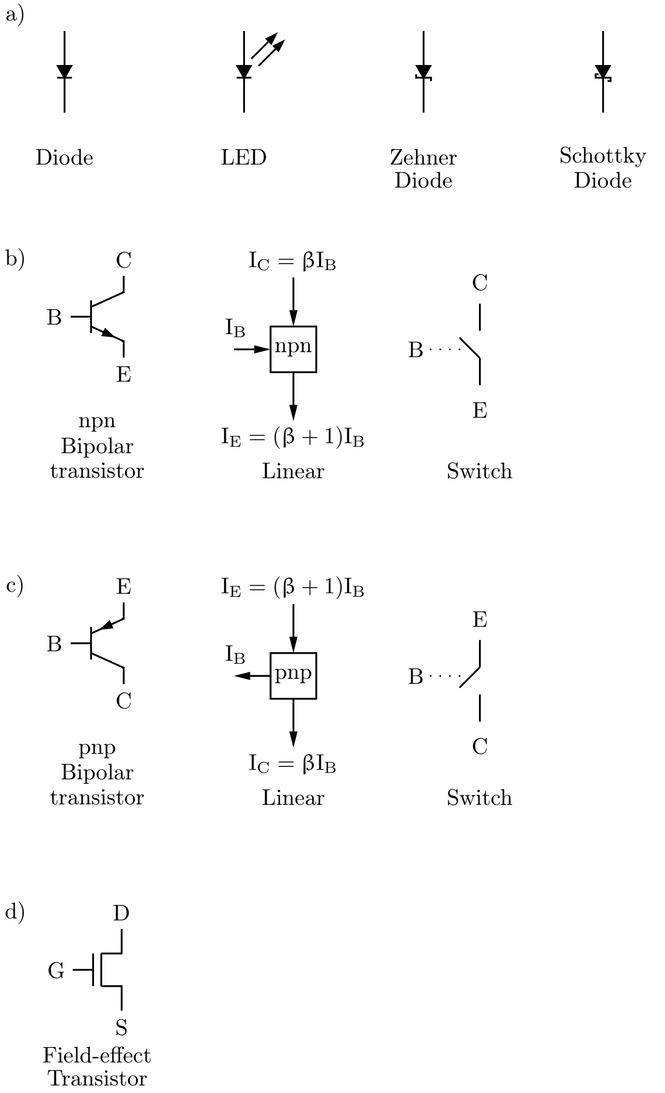


**Fig. 2.15**: Non-linear electronic elements: a) Diodes (normal, LED, Zehner and Schottky), b) npn bipolar transistor (symbol, linear operation, saturation operation) C=Collector, B=Base, E=Emitter, c) pnp bipolar transistor, d) Field-effect transistor, D=Drain, G=Gate, S=Source

**Table 2.VII**: Bias needed to turn light emitting diodes on

| Material | wavelength | Color | Voltage |
|---|---|---|---|
| GaAs | 850-940 nm | IR | 1.2 V |
| GaAsP | 650-660 nm | Red | 1.8 V |
| GaAsP | 605-620 nm | Orange | 2.0 V |
| GaAsP:N | 585-595 | Yellow | 2.2 V |
| AlGaP | 550-570 nm | Green | 3.5 V |
| SiC | 430-505 nm | Blue | 3.6 V |
| GaInN | 450 nm | White | 4.0 V |

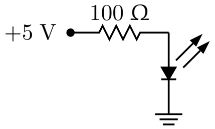


**Fig. 2.16**: LED with shunt resistor to limit the current to some tens of mA. An LED is a current-to-light actuator. With the resistor it becomes a voltage-to-light actuator; $I$ = (5 V - 1 V)/100 Ω = 40 mA

Table 2.VII). This being a small detail, what distinguishes it from a normal diode is that it emits light when enough current flows through it. An LED is thus an actuator that converts current into an optical signal. If we want to use a voltage signal to switch on and off an LED, this voltage first has to be converted into a current of the proper magnitude. The simplest way to do this is with a resistance; a resistance connected to an ideal voltage supply converts it into a (non-ideal) current supply. As an example, imagine we want to have light out of an LED with a 5 V signal. We know that an LED, to emit light, needs about 10 mA current. Combined with the fact that the voltage drop in the diode is about 1 V when conducting, we see that we need to place a resistance of about $R = (5 \text{ V} - 1 \text{ V})/10 \text{ mA} = 400\ \Omega$ in series with the LED. Normally we place resistances in the order of hundreds of ohms in series to 'protect' the LED, in order to limit the current. Without this shunt resistance the LED would blow up; substituting 5 V in the Ebers-Moll equation of the diode gives a current of about $3 \times 10^{69}$ A, i.e., there are not enough electrons in the universe to make up this current.

The Schottky diode behaves exactly like the normal diode. The difference is only in the materials used for the diode, semiconductors in the case of a normal diode and a semiconductor and a metal in case of a Schottky diode. This is completely irrelevant for us. More important is the Zehner diode. This is a diode placed in reverse bias. Whereas as normal diode saturates at the reverse-bias saturation current when placed under negative bias, as shown above, a Zehner diode 'breaks down' at a certain and precise negative voltage $V_Z$. This break-down voltage is determined by the material and design properties of the diode and is specified in the data sheet of the component and comes in a wide range. A Zehner diode is useful if we want have a well-defined reference voltage. Whatever current passes through the diode, the voltage drop across it is guaranteed $V_\mathrm{Z}$.

Because the diode is a non-linear element, its static and dynamic resistance are different. The static resistance is the ratio of voltage and current and can have any value. The dynamic resistance, defined as the derivative of the voltage-current function, is found easily if we ignore the "-1" term in the Ebers-Moll equation

$$r_\mathrm{diode} = \frac{1}{\mathrm{d}I/\mathrm{d}V} = \frac{1}{I_\mathrm{S}/V_\mathrm{T} \times \mathrm{e}^{V/V_\mathrm{T}}} = \frac{V_\mathrm{T}}{I}. \tag{2.98}$$

In other words, the dynamic resistance of a diode is given by the thermal voltage divided by the current of the diode.

The dynamic resistance of a diode with current $I$ is given by $r_\mathrm{diode} = V_\mathrm{T}/I$

To get an idea: a diode with 1 mA current, has 26 Ω dynamic resistance. In general, dynamic resistance is useful for when we want to calculate effects of small voltage increases on the current. For small changes, the function can be approximated by a line. A small voltage increase $\delta V$ will have an effect on the current equal to

$$\delta I = \delta V \times \frac{\mathrm{d}I(V)}{\mathrm{d}V} = \frac{\delta V}{r}, \tag{2.99}$$

with $r$ the dynamic resistance of the element. As an example, for a diode with 1 mA current (and thus 26 Ω dynamic resistance), increasing the voltage by 1 mV will increase the current by 0.04 mA. Superimposing a sinusoidal voltage of 1 mV on top of a DC voltage that caused 1 mA DC current will add a sinusoidal current with 40 μA amplitude to the current. In general if we add a sinusoidal voltage with amplitude $v$ superimposed on top of a DC bias $V_\mathrm{DC}$ we can expect a current of the form $I(t) = I_\mathrm{DC} + (v/r) \times \sin(2\pi ft + \theta)$. See Figure 2.4.

### 2.4.2 (Bipolar) transistors

A bipolar transistor is a three terminal device that can be considered a current amplifier. The connection from base to emitter is a diode with a current following the Ebers-Moll equation. The collector current is then $\beta$ times larger. In other words the base current $I_\mathrm{B}$ follows Ebers-Moll, and the collector current is

$$I_\mathrm{C} = \beta I_\mathrm{B}. \tag{2.100}$$

Applying the what-goes-in-must-come-out rule – also known as Kirchhoff's current law (KCL), the sum of currents going in and coming out must be zero – we see that the emitter current must be the sum of base and collector currents and is thus given by

$$\begin{aligned} I_\mathrm{E} &= (\beta + 1)I_\mathrm{B} \\ &= I_\mathrm{C}/\alpha, \end{aligned} \tag{2.101}$$

with $\alpha$ defined as $\alpha \equiv \beta/(\beta+1)$. This is the 'normal' mode of operation, where the transistor is effectively a current amplifier. We call this the 'linear' regime. The base-emitter junction is a diode that is forward biased. The resulting emitter current is following the Ebers-Moll equation. The base-collector junction is also a diode that should be reverse-biased in normal linear operation. The current of the collector is the current of the base multiplied by a factor, $\beta$. See Figure 2.15(e) and (f) for npn and pnp bipolar transistors respectively (named after the materials used for the collector, base and emitter respectively).

This gives us two possibilities, depending on the sign of the current and the relative voltages. If we want for our circuit the base to have a higher voltage compared to the emitter and we want base and collector currents to go into the transistor and the emitter current coming out of it, we use a so-called npn transistor. If, on the other hand, we need the base to have a lower voltage compared to the emitter we use a pnp transistor. In the symbol for the component we can recognize it by the arrow which represents the direction of current flowing at the emitter, current going coming out of it (npn) or going into it (pnp). The base-collector junction is also a diode, but for the linear operation this diode has to be reverse-biased. This way we can easily remember how to connect a bipolar transistor. For an n-p-n (collector-base-emitter) transistor we have to connect the base-emitter junction as p-n, 'positive-negative' because we want it in forward bias. Note also that current in a bipolar transistor can only flow in the direction of the arrow. Connecting the base-emitter with a bias less than about 0.7 V will close the transistor altogether. No current whatsoever will flow, $I_{\mathrm{B}} = I_{\mathrm{C}} = I_{\mathrm{E}} = 0$.

The current of the collector is as good as independent of the collector voltage. In other words, it does not matter what we connect to the collector, the current will stay the same. There are, however, limits. When the voltage at the collector becomes too low, the transistor no longer works in the linear regime because the diode at the collector-base junction becomes *forward* biased. This occurs when the collector has about the same voltage as the emitter. Then both diodes are similarly biased; the base-emitter and base-collector junction diodes are forward biased, both having the same 0.7 voltage drop. This is the 'switch' mode of operation - better known as 'saturation' - which we will use a lot in this book. For larger currents causing low collector voltages, the transistor works as a programmable switch; with the base current we can either short the collector to the emitter, or have the collector voltage 'floating', just like in a mechanical switch, which can either be a short or an open circuit.

Note that because the base-emitter junction is a diode, the base (or emitter) current has to be limited, for instance with a resistance, just like we have done for a (light-emitting) diode. A bipolar transistor is a current element and directly applying a voltage will probably destroy it.

Finally, for the connoisseurs, the Ebers-Moll equations for the base, collector and emitter currents are given by

$$
\begin{aligned}
I_\mathrm{B} &= \frac{I_\mathrm{S}}{\beta+1}\left(\mathrm{e}^{V_\mathrm{BE}/V_\mathrm{T}}-1\right), && (2.102)\\
I_\mathrm{C} &= \frac{\beta I_\mathrm{S}}{\beta+1}\left(\mathrm{e}^{V_\mathrm{BE}/V_\mathrm{T}}-1\right), && (2.103)\\
I_\mathrm{E} &= I_\mathrm{S}\left(\mathrm{e}^{V_\mathrm{BE}/V_\mathrm{T}}-1\right), && (2.104)
\end{aligned}
$$

(where the effect of the collector voltage on the currents has been completely neglected). This also allows for the definition of dynamic resistances, for which the $-1$ term in the above equations is neglected,

$$
\begin{aligned}
r_\mathrm{B} &\equiv \frac{\mathrm{d}I_\mathrm{B}}{\mathrm{d}V_\mathrm{BE}} = \frac{V_\mathrm{T}}{I_\mathrm{B}}, && (2.105)\\
r_\mathrm{E} &\equiv \frac{\mathrm{d}I_\mathrm{E}}{\mathrm{d}V_\mathrm{BE}} = \frac{V_\mathrm{T}}{I_\mathrm{E}} = \frac{r_\mathrm{B}}{\beta+1}. && (2.106)
\end{aligned}
$$

It was above assumed that the collector current is completely independent of the collector voltage. In practice is close follows the empirical relation $I_\mathrm{C} \propto (V_\mathrm{CB}+V_\mathrm{A})$, with $V_\mathrm{A}$ the Early voltage, in the order of some hundreds of volts. This allows for the definition of an 'output' (collector) resistance

$$r_\mathrm{C} \equiv \frac{\partial I_\mathrm{C}}{\partial V_\mathrm{CB}} = \frac{V_\mathrm{CB}+V_\mathrm{A}}{I_\mathrm{C}} \approx \frac{V_\mathrm{A}}{I_\mathrm{C}}. \qquad (2.107)$$

For a transistor with some milliamps collector current this collector resistance is in the order of a hundred kΩ; often negligible, meaning close-to-infinite, and the transistor collector is close to being an ideal current source, programmed by the base and there lies the 'power' of the transistor.

### 2.4.3 Examples of bipolar-transistor circuits

Figure 2.17 shows some examples of circuits with transistors to get some idea of how we should think when it comes down to transistors. Maybe the circuits are not useful, but they will give us some feeling of how transistors work. In these examples the supply voltage is $V_\mathrm{CC}$ = 10 V, and the resistor names are not shown, but follow the convention that the resistor connected to the base is called $\mathrm{R_B}$, etc.

a) We first calculate the base current. Assuming the transistor is working correctly, in the linear regime, the voltage drop base-emitter, $V_\mathrm{BE}$ is 0.7 V. Thus, we assume $V_\mathrm{B}$ = 0.7 V. We can then calculate the base current, $I_\mathrm{B}$ = (10 V - 0.7 V)/10 kΩ = 0.93 mA. Multiplying this with $\beta$ = 100, we find 93 mA for the collector current, and multiplying with $\beta + 1$ = 101 we find a 94 mA for the emitter current. Finally we verify that the collector-base junction is reverse biased. With $V_\mathrm{C}$ = 10 V and $V_\mathrm{B}$ = 0.7 V, this verifies.

b) The base voltage and base current are the same as above. Thus we might expect the same collector current. If we try a current of 93 mA we see that it

will induce a voltage drop at the collector resistor equal to $\Delta V = I \times R_\mathrm{C} = 93$ V. In other words, the current wants to force the collector voltage at $V_\mathrm{C} = V_\mathrm{CC} - 93$ V $= -83$ V. Clearly the transistor will not work in the linear regime. Instead it will work in the saturation regime and effectively $V_\mathrm{C}$ is connected to $V_\mathrm{E}$ in this case to ground, $V_\mathrm{C} = 0$. The collector current is thus $I_\mathrm{C} = (V_\mathrm{CC} - 0)/R_\mathrm{C}$ = 10 mA.
c) Through a resistor ($R_\mathrm{B}$ = 10 kΩ) the base is connected to ground, the same voltage as the emitter; there is no voltage drop $V_\mathrm{BE}$ and this diode junction thus has no current, according to Ebers-Moll: $I_\mathrm{B} = 0$. The collector current is thus also 0 and no voltage drop is induced in the collector resistor, $V_\mathrm{C} = C_\mathrm{CC} - 0 \times R_\mathrm{C} = V_\mathrm{CC} = 10$ V.
d) The 1 mA current from the current source passes partly through the base resistor of 1 kΩ and partly enters the base of the transistor. If we assume the transistor is working in the linear regime, the base is at $V_\mathrm{B} = V_\mathrm{EE} + V_\mathrm{BE} = 0$ + 0.7 V = 0.7 V. That means a current $I = V_\mathrm{B}/R_\mathrm{B}$ = (0.7 V)/(1 kΩ) = 0.7 mA passes through the base resistor. The rest coming from the current source, 0.3 mA, must enter the base of the transistor, $I_\mathrm{B}$ = 0.3 mA. The collector current is then $I_\mathrm{C} = \beta I_\mathrm{B} = 100\times$(0.3 mA) = 30 mA. The voltage drop at the collector resistor is then $\Delta V_\mathrm{C} = I_\mathrm{C} R_\mathrm{C}$ = (30 mA)×(100 Ω) = 3 V. That makes the collector voltage $V_\mathrm{C} = V_\mathrm{CC} - \Delta V_\mathrm{C}$ = 10 V − 3 V = 7 V. e) Whatever current we have passing through the resistor $\mathrm{R_B}$ (10 kΩ) and entering the base is multiplied by $(\beta + 1)$ and passes through resistor $\mathrm{R_E}$ at the emitter. We can find this current $I_\mathrm{B}$ if we know that the total voltage drop along this path must be 10 V:

$$I_\mathrm{B} \times (20\ \mathrm{k\Omega}) + 0.7\ \mathrm{V} + 101 I_\mathrm{B} \times (100\ \Omega) = 10\ \mathrm{V}. \tag{2.108}$$

This gives a base current equal to $I_\mathrm{B}$ = 0.309 mA, and a base voltage of $V_\mathrm{B} = V_\mathrm{CC}$ − (0.309 mA) × (20 kΩ) = 3.82 V. The emitter is 0.7 V lower, $V_\mathrm{E} = V_\mathrm{B} - V_\mathrm{BE}$ = 3.12 V, which we would also have found if we had calculated the emitter current, $I_\mathrm{C} = 101 I_\mathrm{B}$ = 31.2 mA and multiplied it by the emitter resistance, 100 Ω. The collector current is $\beta$ = 100 times the base current, $I_\mathrm{C}$ = 30.9 mA and the collector voltage $V_\mathrm{C} = V_\mathrm{CC}$ − (30.9 mA) × (100 Ω) = 6.91 V. We verify that the transistor is working properly in the linear mode: $V_\mathrm{C} > V_\mathrm{B} > V_\mathrm{E}$.
At this moment it is interesting to see where heat is being generated in the circuit. For the total heat we can multiply the voltages of the voltage supplies by their supplied current. There are two voltage supplies, one of 10 V ($V_\mathrm{CC}$ and one of 0 (ground). The first one supplies a current $I_\mathrm{B} + I_\mathrm{C}$ = 31.2 mA. The power is thus $P$ = (31.2 mA)×(10 V) = 0.312 W. The ground supplies the same current, but since the voltage is zero, it supplies no power. The total power of the circuit is 312 mW. We can look at this in more detail and specify where the heat is generated.

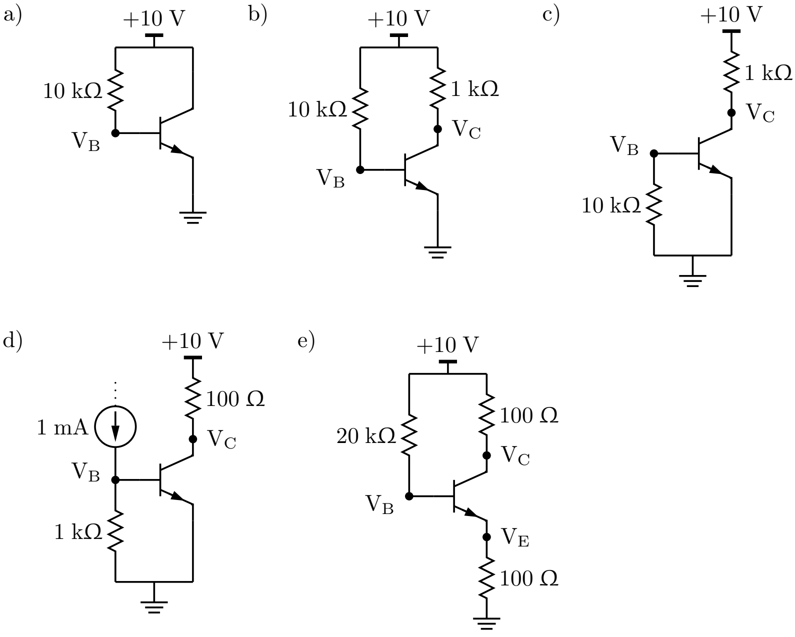


**Fig. 2.17**: Examples of transistor circuits. β = 100

| Location | Voltage drop | × | Current | = | Power |
|---|---|---|---|---|---|
| Base resistor | 6.18 V | × | 0.309 mA | = | 1.91 mW |
| Base-emitter junction | 0.7 V | × | 31.2 mA | = | 21.84 mW |
| Emitter resistor | 3.1 V | × | 31.2 mA | = | 97.38 mW |
| Collector-base junction | 3.09 V | × | 30.9 mA | = | 95.46 mW |
| Collector resistor | 3.09 V | × | 30.9 mA | = | 95.46 mW |
| Total | | | | | 312 mW |

### 2.4.4 The transistor as a small-signal amplifier

This circuit of Figure 2.18 has a novelty. It has a coupling capacitor at $V_\mathrm{B}$, where we connect a sinusoidal 'small' signal, $V_\mathrm{i}(t) = v_\mathrm{i}\sin(2\pi f t)$ (note the uppercase and lowercase symbols, lowercase is used for *amplitudes* of signals). At the collector we have a similar coupling capacitor where we monitor the output signal, $V_\mathrm{o}(t)$. The question is: what is the ratio between $v_\mathrm{o}$ and $v_\mathrm{i}$, the small signal voltage gain?

We first note that the capacitors are open circuits for low frequencies. For the moment it is not relevant what exactly means 'low frequencies'. Suffices to say there is a frequency below which the capacitors effectively are open circuits. DC, for sure, is below this frequency. The power supply is a DC element (0 Hz), and we thus find back the circuit of Fig. 2.17(b), but with different resistances. This we already know how to solve. The base current can

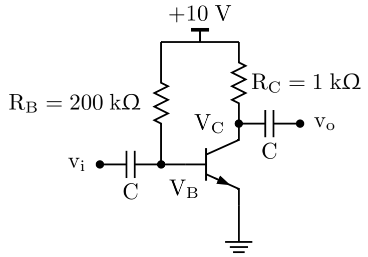


**Fig. 2.18**: Example of a small-signal amplifier

be found as $I_{\mathrm{B}} = (V_{\mathrm{CC}} - V_{\mathrm{BE}})/R_{\mathrm{B}} = (10\ \mathrm{V}\ \text{-}\ 0.7\ \mathrm{V})/(200\ \mathrm{k\Omega}) = 46.5\ \mu\mathrm{A}$. The collector current is $I_{\mathrm{C}} = \beta I_{\mathrm{B}} = 100{\times}(46.5\ \mu\mathrm{A}) = 4.65\ \mathrm{mA}$. The collector voltage is then $V_{\mathrm{C}} = V_{\mathrm{CC}} - I_{\mathrm{C}}R_{\mathrm{C}} = (10\ \mathrm{V}) - (4.65\ \mathrm{mA}){\times}(1\ \mathrm{k\Omega}) = 5.35\ \mathrm{V}$. This DC analysis we call the bias point of the circuit:

| Item | Symbol | Value |
|---|---|---|
| Base voltage | $V_{\mathrm{B}}$ | 0.7 V |
| Base current | $I_{\mathrm{B}}$ | 46.5 μA |
| Emitter voltage | $V_{\mathrm{E}}$ | 0 |
| Emitter current | $I_{\mathrm{E}}$ | 4.70 mA |
| Collector voltage | $V_{\mathrm{C}}$ | 5.35 V |
| Collector current | $I_{\mathrm{C}}$ | 4.65 mA |

For high frequencies (above a certain unspecified frequency) the capacitors are short-circuits and at the entrance we add the input signal $V_{\mathrm{i}}(t) = v_{\mathrm{i}}\sin(2\pi ft)$ to the base voltage. What will be the effect on the collector voltage? Imagine at a certain moment we are at 90° phase (sin(..) = +1) and have added $v_{\mathrm{i}}$. This will have increased the base-emitter diode-current by a value $i_{\mathrm{B}} = v_{\mathrm{i}}/r_{\mathrm{diode}}$. In the section on the diode we have seen that the dynamic resistance of the diode is $r_{\mathrm{diode}} = V_{\mathrm{T}}/I_{\mathrm{diode}}$, with $I_{\mathrm{diode}}$ the diode DC current and $V_{\mathrm{T}}$ the thermal voltage. In this case $r_{\mathrm{B}} = V_{\mathrm{T}}/I_{\mathrm{B}} = (26\ \mathrm{mV})/(46.5\ \mu\mathrm{A}) = 559\ \Omega$. (In electronics books, this resistance that models the small signal behavior of a bipolar transistor is often called $r_\pi$, named after the hybrid-π model used). Thus, the base current increases by $i_{\mathrm{B}} = v_{\mathrm{i}}/r_{\mathrm{B}}$. The collector current then increases β times, $i_{\mathrm{C}} = \beta i_{\mathrm{B}} = \beta v_{\mathrm{i}}/r_{\mathrm{B}}$. The collector voltage then *decreases* by an amount $v_{\mathrm{C}} = -i_{\mathrm{C}}R_{\mathrm{C}} = -\beta v_{\mathrm{i}} \times (R_{\mathrm{C}}/r_{\mathrm{B}})$. The negative sign stems from the fact that if current increases through $\mathrm{R_C}$, the voltage drop in this resistor increases. The capacitor at the collector then filters off the DC part and lets through this small-signal part, $v_{\mathrm{o}} = v_{\mathrm{C}}$. We thus find a small-signal voltage gain

$$A_{\mathrm{v}} \equiv \frac{v_{\mathrm{o}}}{v_{\mathrm{i}}} = -\beta\frac{R_{\mathrm{C}}}{r_{\mathrm{B}}} = -\alpha\frac{R_{\mathrm{C}}}{r_{\mathrm{E}}}. \tag{2.109}$$

In this case $A_{\mathrm{v}} = -179$.

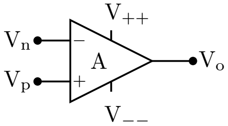


**Fig. 2.19**: Basic operational amplifier (opamp). The output is equal to the difference at the input terminals $V_\mathrm{p} - V_\mathrm{n}$, multiplied by $A$. The other terminal $V_{++}$ and $V_{--}$ represent the power supply voltages. For simplicity in our circuit diagrams we will omit these latter voltages, but don't forget that an opamp needs power, since it is an active element

## 2.5 Operational amplifiers (opamps)

The operational or differential amplifiers, normally shortened to 'opamps', are the most used elements in electronic instrumentation. In itself it is not so useful, unless for comparators, but the power of the opamp lies in its versatility. It can be used in simple comparators, amplifiers, (active) filters, etc. To understand this, it is good to define an opamp.

An operational amplifier is an active element (powered by an external source) that amplifies the difference between the two input voltage signals, see Figure 2.19,

$$V_\mathrm{o} = A \times (V_\mathrm{p} - V_\mathrm{n}), \tag{2.110}$$

with $V_\mathrm{o}$ the output voltage, and $V_\mathrm{p}$ and $V_\mathrm{n}$ signals at the non-inverting and inverting input terminals, respectively. The voltage amplification factor is normally very high. Commercial opamps typically have this open-loop voltage gain in the order $10^4$ to $10^6$, which implies that for even the tiniest difference of input signals the output saturates, since the output voltage cannot exceed the power-supply voltages $V_{++}$ and $V_{--}$. An opamp, powered by 10 volts, with an input-voltage difference larger than 10 µV, will saturate at 10 V. To make things 'worse', an ideal opamp is characterized by an infinite open-loop gain. Even so, factories try to make their opamps as ideal as possible.

### 2.5.1 Ideal opamp

In total, an ideal opamp is characterized by the following features, see also Fig. 2.20:

1. Infinite open-loop gain, $A = \infty$.
2. Infinite input resistance, $R_\mathrm{i} = \infty$ and thus also $r_\mathrm{i} = \infty$.
3. Zero (dynamic) output resistance $r_\mathrm{o} = 0$.
4. Infinite frequency bandwidth, $\Delta f = \infty$, and zero time delay, $\Delta t = 0$.

The **infinite open-loop gain** causes that the output voltage is infinite for *any* difference at the input terminals. Reasoning the other way around, if the

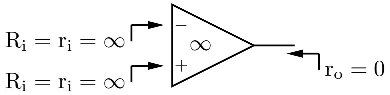


**Fig. 2.20**: Ideal operational amplifier (opamp). Infinite gain, infinite (static and dynamic) input resistance, zero (dynamic) output resistance

output voltage is *not* infinite (or saturates at one of the two power supply voltages) the difference at the input terminals must be zero. Rearranging Equation (2.110) we get

$$V_\mathrm{p} - V_\mathrm{n} = \frac{V_\mathrm{o}}{A}, \qquad (2.111)$$

and for $A = \infty$ this gives a zero difference. Thus, as a corollary of the first item above:

5. No saturation? Then $V_\mathrm{p} = V_\mathrm{n}$.

This will come in handy later. We can, for instance define one terminal the virtual ground when the other is factually grounded. It is called 'virtual' ground because for the purpose of the calculations we can consider the terminal at 0 volt.

The **infinite input resistance** of an opamp is useful since this implies that we will not draw current from the sensor or anything we connect to the input of the opamp. No current will enter the input terminals. Note, however, that the rest of the circuit and the things we will connect to the opamp can still draw current. Since no current enters the opamp for any voltage, both the static and dynamic input are zero, $R_\mathrm{i} = V_\mathrm{i}/I_\mathrm{i} = \infty$ and $r_\mathrm{i} = \mathrm{d}V_\mathrm{i}/\mathrm{d}I_\mathrm{i} = \infty$.

The **zero (dynamic) output resistance** implies that we can consider the opamp as an ideal voltage source; whatever we connect to the output will not change the output voltage. Since $r_\mathrm{o} = \mathrm{d}V_\mathrm{o}/\mathrm{d}I_\mathrm{o}$, if we start drawing output current and $I_\mathrm{o}$ changes, $V_\mathrm{o}$ does not change. That is useful, since, once we have the information coded in voltage at the output of an opamp, we can further process it at our will, without having to worry about destroying the information. Remember that one of the parameters of a sensor or system was the interference, that by measuring the signal we will destroy the signal. With the use of opamps we can avoid that. Opamps do not draw current and can supply whatever current at the output to maintain the programmed voltage.

The infinite bandwidth implies that the signal at the input can change infinitely fast, the amplifier will not hinder the passage of the signal in any way. There are no delays or filtering elements in the opamp. If, on the other hand, we do want to filter the signal, for instance to increase the signal-to-noise ratio, or to artificially introduce a time delay, we have to do this externally.

### 2.5.2 Non-ideal opamp

Real off-the-shelf commercial opamps are not ideal and have limitations.

- The open-loop gain is normally in the order of $10^5$ (then conventionally expressed as 100 dB, or 10 bell, the base-10-logarithm of the squared voltage gain indicating the power amplification since power $P = V^2/R$). However, this will not limit the operation very much, as we will see. For most purposes 100 dB ($A_\mathrm{p} = 10^{10}, A_\mathrm{v} = 10^5$) is as good as infinite.

- A commercial opamp normally does not only amplify the difference of the input signal, but also the sum, often called the 'common mode'. The quality of an opamp can then be expressed in a figure of merit that quantifies how well it amplifies the difference relative to the sum. This is the common-mode rejection ratio, or CMRR, a parameter that will normally be shown on the datasheet. Ideally, the CMRR is infinite, because the differential amplification is infinite and the common-mode amplification zero, giving an infinite ratio.

- The input resistance is in the order of some megaohms, some tiny current, in the order of microamperes will be drawn from the rest of the circuit.

- The output resistance in not zero. It lies normally in the order of some ohms. Also, we have to bear in mind that the output power of a commercial opamp lies in the order of tens of milliwatts. With 10 volts at the output, the opamp can deliver only some milliamperes at best. An opamp is not a power element! It is a signal element. If we want to use it to switch on a motor or anything else that consumes power we have to place some power circuit in between, for instance a relay or a power transistor.

- To avoid the possibility of oscillations, the bandwidth of opamps is normally on purpose limited by internal frequency-compensation. The result is a cut-off frequency in the order of tens of hertz. Since, as for any simple low-pass filter, after the cut-off frequency the output signal drops linearly with frequency, a unity gain is reached somewhere in the megahertz range. The interesting fact is that the product of gain and bandwidth is constant, independent of what we do with the opamp in our circuit. If we use it without feedback in an open-loop configuration, the gain is $10^5$ and the bandwidth 10 Hz, with a gain-bandwidth product of 1 MHz. If we use 100% feedback, in a voltage follower, the gain will drop to unity, but the bandwidth will increase to 1 MHz, maintaining the gain-bandwidth product of 1 MHz.

  In a datasheet of an opamp the frequency of unity gain $f_\mathrm{T}$ is normally specified. From this we can calculate the cut-off frequency of the opamp, $f_\mathrm{C} = f_\mathrm{T}/A_\mathrm{v}$

- Slew rate. This tells us how fast the output signal can change. While very similar to the frequency response, it is not exactly the same. The slew rate says how fast the output can change in terms of volts per second.

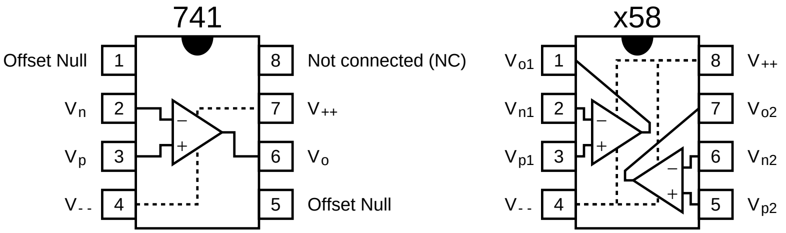


**Fig. 2.21**: Pin connection for a 741-family single opamp (left) and x58-family dual-opamp (right) in DIP (dual in-line package) version

**Table 2.VIII**: Parameters of a typical low-cost opamp, LM358 dual opamp of National Semiconductor

| *Parameter* | *Symbol* | *Value* | *Unit* |
|---|---|---|---|
| Voltage gain | $A_\mathrm{v}$ | 100 | dB |
| Bandwidth (unity gain) | $f_\mathrm{T}$ | 1 | MHz |
| Supply voltage | $V_{++} - V_{--}$ | 32 | V |
| Operating temperature range | | 0 - +70 | °C |
| CMRR | | 85 | dB |
| Max output current | $I_\mathrm{o,max}$ | $-20$ - $+40$ | mA |

### 2.5.3 Feedback

On basis of the ideal operational amplifier we can design some useful circuits. Before we continue, it is useful to design some feedback theory, since in most cases we will connect the output of the opamp to the input through a circuit. The input signal $V_\mathrm{i}$ is amplified by a factor $A$ yielding the output signal $V_\mathrm{o}$. A fraction $\beta$ of this output signal is added ('fed back') to the input, see Figure 2.22. $A$ is called the open-loop gain (the gain we would find if no feedback is used) and $\beta$ the feedback factor defined as

$$\beta \equiv \left.\frac{V_\mathrm{x}}{V_\mathrm{o}}\right|_{\mathrm{V_i=0}} \tag{2.112}$$

in passive circuits lying between $-1$ and $+1$. See Figure 2.22. Defining a voltage $V_\mathrm{x}$ just before the amplifier A we can easily find the relation between output and input signal:

$$V_\mathrm{x} = V_\mathrm{i} + \beta V_\mathrm{o}, \tag{2.113}$$

$$V_\mathrm{o} = A V_\mathrm{x}. \tag{2.114}$$

Thus

$$V_\mathrm{o} = V_\mathrm{i}\frac{A}{1 - A\beta}. \tag{2.115}$$

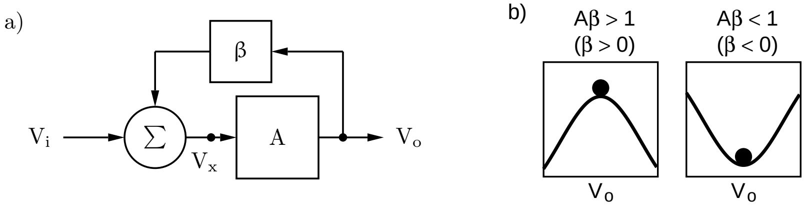


**Fig. 2.22**: a) Basic feedback diagram. The input signal $V_\mathrm{i}$ is amplified by a factor $A$ yielding the output signal $V_\mathrm{o}$. A fraction β of this output signal is added ('fed back') to the input. This results in the input-output relation as in Equation (2.115). Square boxes are multipliers, the circle represents summing. b) Schematic visualization of a system with positive feedback (left) and negative feedback (right). In brackets the condition for an ideal amplifier with infinite gain. The former situation is unstable, it will drift away from the metastable point in an accelerated way until it hits the limitations of the supply voltage. Negative feedback results in stable signals

For an ideal opamp ($A = \infty$) this reduces to

$$\frac{V_\mathrm{o}}{V_\mathrm{i}} = -\frac{1}{\beta}. \tag{2.116}$$

When analyzing an opamp circuit, sometimes it is easier to use feedback theory and sometimes direct calculations. (And sometimes it makes no difference). In any case, all roads lead to Rome and we can always find the correct answer using either technique. Although in case of saturation the feedback analysis can rapidly become more complicated.

This brings us to the subject of saturation and oscillation. Circuits with feedback can have this effect. Sometimes it is exactly the effect we want, while other times we will want to avoid it. In any case it is good to know where it comes from. As an example, substituting $\beta = +1$ in the equation above would result in a input-output relation equal to $V_\mathrm{o} = -V_\mathrm{i}$. In practice this will not happen and the circuit will saturate.

For that we have to analyze the feedback loop. This consists of the amplifier gain $A$ and the feedback fraction β. If the product of the two is larger than unity we have the situation of an increasing voltage at the output without anything at the input. Any (tiny) voltage (for example noise) that exists just before the amplifier, $V_\mathrm{x}$ in Figure 2.22, is multiplied by $A$ then multiplied by β and assigned to $V_\mathrm{x}$. We start with $V_\mathrm{x}$. After the first iteration the voltage is $A\beta V_\mathrm{x}$. The next iteration multiplies it again with $A\beta$ and the voltage is already $V_\mathrm{x} = (A\beta)^2 V_\mathrm{x}$. Then to $(A\beta)^3 V_\mathrm{x}$, etc. This rapidly grows and soon reaches saturation. We thus arrive at the conclusion that a circuit with a feedback loop $A\beta$ larger than unity will not be stable and will tend to saturate at either of the two supply voltages even without an input voltage.

Given the fact that the open-loop gain $A$ and the feedback factor β both can depend on the frequency, this condition can in some cases occur for only some frequencies. In these cases we have a sinusoid or a range of sinusoids at the output. In fact, we will make use of this when we describe oscillators that are based on this principle of positive feedback for some frequencies (Section 2.7).

$$A(f)\beta(f) \geq 1 : \text{Oscillations/saturation at frequency } f. \tag{2.117}$$

In more detail, if $A\beta$ is equal to unity, the oscillation is marginally maintained, we call this the Barkhausen criterion.

$$A(f)\beta(f) = 1 : \text{Barkhausen Criterion. Sustained oscillation at frequency } f. \tag{2.118}$$

The amplitude of the output voltage is neither rising, nor dropping. For $A\beta$ larger than unity the output is rising; with every passage through the feedback loop the output voltage is larger. The output increases with a speed determined by the propagation of the signal along the loop, in the order of sub-microseconds. That means that before too long - nearly instantaneously - the amplitude rises beyond the capabilities of the power supply and we see a highly-distorted square-wave-like signal at the output. For a feedback loop gain smaller than unity we see a damped oscillation at the output. Rapidly decaying to zero or settling at the value 'programmed' by the input $V_\mathrm{i}$. For a negative feedback loop gain every accidental increase of the signal (i.e. noise) is counteracted by a compensating signal that restores the equilibrium.

In this respect, DC signals can be seen as signals at 0 Hz. If the DC feedback loop $A\beta$ is larger than zero, then saturated 'oscillations' at 0 Hz occurs. In practice this means an output voltage equal to either of the supply voltages. Which of the two can depend on the input voltages and history of the device, as we will see in the comparator with hysteresis (Section 2.6.6).

For an ideal amplifier $A$ is infinite and we have two possibilities: If $\beta$ is larger than zero, then $A\beta$ is larger than unity and the circuit will oscillate. If $\beta$ is negative, then $A\beta$ is less than unity and the circuit will be stable. We conclude that, for an ideal amplifier, if positive feedback is used, the output will tend to saturate. Any tiny input signal will make make the output saturate at the output. Since any part of the circuit works like an antenna and any resistance generates noise (see section 2.8.1), tiny signals with a broad frequency spectrum can be found anywhere in the signals. We can visualize the system with positive feedback as shown in Figure 2.22(b). It is like the metastable system of a ball on a hill; any tiny vibration will make the ball roll off the hill in an accelerated way.

Negative feedback, on the other hand tends to stabilize the output, compare to the ball-in-a-valley situation of Fig. 2.22(b). The use of negative feedback is a powerful tool to get a stable and predictable output voltage. Moreover, it increases the input resistance, decreases the output resistance of a circuit, reduces the sensitivity of the overall gain. Something that is given here without proof.

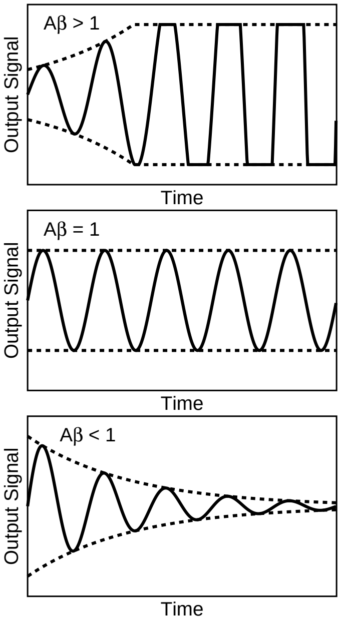


**Fig. 2.23**: Feedback and oscillations. If the feedback loop, the product of the amplifier gain $A(f)$ and the feedback factor $\beta(f)$, are unity for a given frequency $f$, an oscillation with that frequency is marginally sustained (middle panel). If the loop gain is smaller, a damped oscillation occurs that is extinguished after some time (bottom panel). For $A\beta$ larger than one, an oscillation with rising amplitude occurs that is clipped by the power supply and eventually results in a distorted oscillation. One has to bear in mind that the time scale to reach saturation or attenuation is determined by the feedback-loop delay and is of the order of microseconds, or faster

Negative feedback will result in an output value that stabilizes rapidly, positive feedback will result in signal run-away, instability and oscillations.

In practice, a circuit will oscillate at all frequencies at which the feedback loop gain is larger than one. We can even make use of it in the design of oscillators. A properly designed oscillator has the range of frequencies as small as possible. Ideally the Barkhausen Criterion is true for only one frequency and we get a nice monochromatic oscillation. In practice, oscillations can occur at other frequencies as well. We again take a look at the feedback loop $A\beta$. Both parameters can depend on frequency, but both can also contain phase shifts that also have to be taken into account. Note, for instance, that a phase shift of 180° is identical to and indistinguishable from a multiplication by $-1$. A phase shift of 180° in a negative feedback loop has the same effect as the same magnitude positive feedback, i.e., it can destabilize the signal. We must therefore analyze the full phase diagram of the feedback loop $A\beta$ to determine if the circuit will oscillate or not. A powerful tool to this end is a Nyquist plot, where the real part (phase 0°) is plotted versus the imaginary part (phase 90°). Figure 2.24 gives a schematic example. The Barkhausen Criterion $A\beta = 1$ can be found easily, since the real part is 1 and the imaginary part is equal to zero. Alternatively, we can say that the magnitude $|A\beta|$ is equal to 1 and the phase $\angle A\beta = 0$. Furthermore, we know from the discussion above that circuits with feedback loops with $A\beta > 1$ will also oscillate, albeit in a distorted way. Without phase shifts this is the line indicated in the figure, $\mathrm{Im}(A\beta) = 0$, $\mathrm{Re}(A\beta) > 1$. Without proof we will now state that a circuit will run the risk of oscillation when

- The magnitude of the feedback loop is equal or larger than one, $|A\beta| \geq 1$, and
- The phase is between $-45°$ and $+45°$.

The latter defines a phase margin of 45° from guaranteed oscillations. Inside this region the circuit "runs the risk of oscillating", which is an engineering way of saying that the circuit will oscillate when you don't want it to and it will not oscillate when you want it to. If we want it to oscillate we need to put the phase at zero. If we don't want it to oscillate we must stay out of the phase margin zone.

## 2.6 Basic opamp circuits

### 2.6.1 Voltage follower/Voltage buffer

The easiest circuit we can build with an opamp is the voltage follower. It consists of an opamp with the input signal connected to the positive terminal $V_\mathrm{p}$ and the output connected to the negative terminal, see Figure 2.25(a). We can analyze this circuit in two ways, either with the feedback ideas, or with the ideal-opamp rules. Starting with the latter, remember (Rule 5) $V_\mathrm{p} = V_\mathrm{n}$ in absence of saturation. Since the negative terminal is connected to the output as well, $V_\mathrm{o} = V_\mathrm{n}$ it follows that the output must be equal to the input, $V_\mathrm{o} = V_\mathrm{n} = V_\mathrm{p}$, in other words, the output 'copies' the input. And, indeed, the output does not saturate at the power supply voltages ($V_{++}$ or $V_{--}$) as long as the input does not reach these values.

By using the feedback ideas we arrive at the same conclusion. For connecting the output to the negative input terminal, the feedback factor equals $-1$ and the feedback equation, Eq. (2.115), becomes

$$V_\mathrm{o} = V_\mathrm{i}\frac{A}{1+A}. \tag{2.119}$$

Moreover, for an ideal amplifier the gain is infinite and the above equation reduces to $V_\mathrm{o} = V_\mathrm{i}$, the same result we found before.

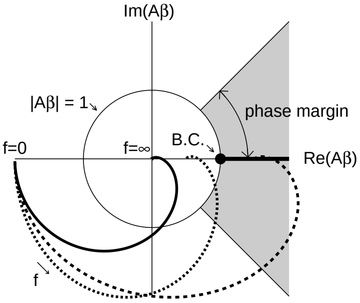


**Fig. 2.24**: Nyquist plot of the feedback loop $A(f)\beta(f)$; plotting the imaginary part vs. the real part. If this passes through the point $A\beta = 1$ - indicated by B.C., (Barkhausen Criterion) - for any frequency, the circuit will oscillate at that frequency. If $A\beta$ is real and larger than 1 (indicated by thick horizontal line), the circuit will also oscillate but in a saturated way. The gray zone of 45° phase margin indicates the situation where the circuit *might* oscillate. Three Nyquist plots are shown. All have negative feedback at low frequencies and are thus stable at DC. One (indicated by a solid line) is stable for all frequencies. The second (dotted line) might oscillate for a small range of frequencies, while the third (dashed line) oscillates for sure at two frequencies and runs the risk at oscillating in a larger frequency range

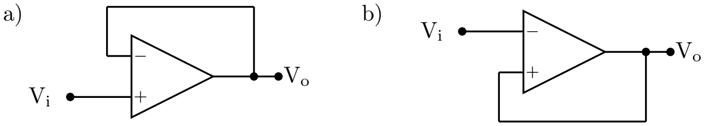


**Fig. 2.25**: a) Simple voltage follower using full negative feedback ($\beta = -1$). For an ideal opamp with infinite gain the output signal is equal to the input signal. b) Wrong voltage follower. While it mathematically also gives $V_o = V_i$ as a solution, positive feedback makes it unstable and it will rapidly wind up saturating at either supply voltage

Using similar reasoning we might have designed the voltage follower as in Figure 2.25(b), namely with full *positive* feedback, $\beta = +1$. After all, if we use the rule $V_\mathrm{p} = V_\mathrm{n}$ and $V_\mathrm{n}$ is connected to input and $V_\mathrm{p}$ to output, then $V_\mathrm{o}$ is necessarily equal to $V_\mathrm{i}$. No saturation, and this then retroactively justifies the rule of zero difference at the input terminals. However, as we have seen in the feedback section (Section 2.5.3), systems with positive feedback are unstable and they can drift rapidly into saturation. Any tiny difference at the input terminals is increased over and over again, until it saturates at either supply voltage. Note that the final situation with $V_\mathrm{p} = V_\mathrm{o}$ equal to $V_{++}$ or $V_{--}$ is also a solution. The conclusion is that, whereas both voltage followers of Figure 2.25 mathematically have the solution that the output is equal to the output, in practice only the one with negative feedback will produce this result because it is stable and settles within a couple of amplifier delays at the output value.

The question that now rises is What is the use of a circuit that does nothing to our signal, only copies it? The answer lies in the other points of the ideal opamp, namely the input and output resistance. Remember that an ideal opamp has infinite input resistance; no current enters the input terminals. This means that signals coming from something with finite, non-zero output resistance will not be distorted. This is why the circuit is also sometimes called Voltage Buffer.

Take for example the situation shown in Figure 2.26(a) of a sensor with output resistance $R_\mathrm{x}$ generating a signal $V_\mathrm{x}$ fed into a 10× amplifier with input resistance $R_\mathrm{i}$. Because of the voltage division at the input of the amplifier, the input voltage is only

$$V_\mathrm{i} = V_\mathrm{x}\frac{R_\mathrm{i}}{R_\mathrm{i} + R_\mathrm{x}}, \tag{2.120}$$

and the output voltage is

$$V_\mathrm{o} = (10 \times V_\mathrm{x})\frac{R_\mathrm{i}}{R_\mathrm{i} + R_\mathrm{x}}. \tag{2.121}$$

For a sensor with non-zero output resistance (non-ideal voltage source) connected to an amplifier with finite input resistance (drawing current) the signal amplification will be less than 10 times. Moreover, this factor lost, $R_\mathrm{i}/(R_\mathrm{i}+R_\mathrm{x})$ can depend on the temperature, the day, the specific sensor, etc. This makes the information in the signal unreliable. Placing a voltage follower/buffer in the circuit solves all problems, as can be seen in Figure 2.26(b). The voltage follower has infinite input resistance $R_\mathrm{i,oa}$ and draws no current, thus $V_\mathrm{p} = V_\mathrm{x}$,

$$\begin{aligned} V_\mathrm{p} &= V_\mathrm{x}\frac{R_\mathrm{i,oa}}{R_\mathrm{i,oa} + R_\mathrm{x}} && (2.122)\\ &= V_\mathrm{x}\frac{\infty}{\infty + R_\mathrm{x}} && (2.123)\\ &= V_\mathrm{x}. && (2.124) \end{aligned}$$

The voltage follower will copy - or *try* to copy - the input to the output $V_\mathrm{o,oa} = V_\mathrm{p}$. Because the opamp is an ideal voltage source, this output voltage is well defined, and the fact that the following amplifier stage has a finite

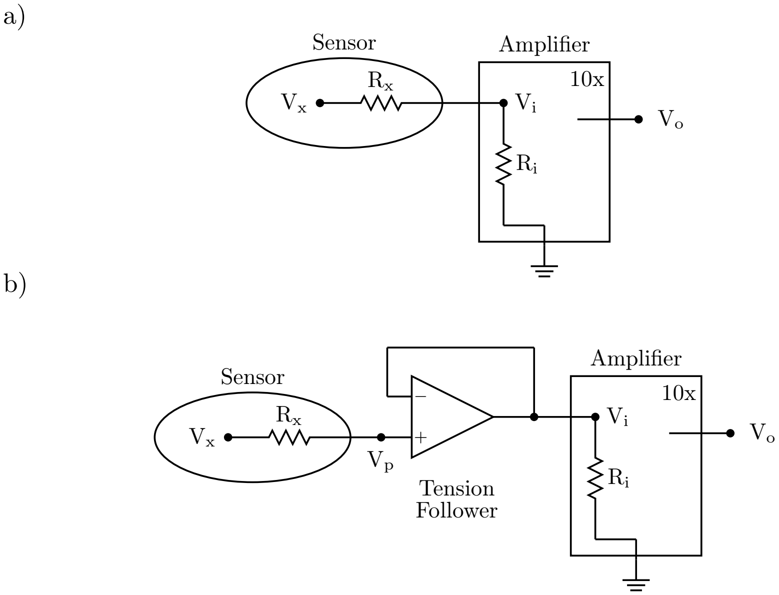


**Fig. 2.26**: a) Signal $V_\mathrm{x}$ from a sensor with an output resistance $R_\mathrm{x}$ amplified by a 10× amplifier with input resistance $R_\mathrm{i}$ (and without output resistance) is not amplified 10 times; $V_\mathrm{o} \neq 10 \times V_\mathrm{x}$. b) insertion of a voltage follower/buffer overcomes this problem; $V_\mathrm{o} = 10 \times V_\mathrm{x}$

input resistance does not matter. The output resistance of the voltage follower is zero, $R_\mathrm{o,oa} = 0$, thus

$$\begin{aligned} V_\mathrm{i} &= V_\mathrm{o,oa}\frac{R_\mathrm{i}}{R_\mathrm{i} + R_\mathrm{i,oa}} && (2.125)\\ &= V_\mathrm{o,oa}\frac{R_\mathrm{i}}{R_\mathrm{i} + 0} && (2.126)\\ &= V_\mathrm{o,oa} && (2.127)\\ &= V_\mathrm{p} && (2.128)\\ &= V_\mathrm{x}. && (2.129) \end{aligned}$$

This input voltage is multiplied by 10 and we get a neat and well defined $10 \times V_\mathrm{x}$ at the output, independent of temperature and other disturbing interferences.

**Question**: A temperature sensor has a resistance depending on the temperature with a nominal value of 100 Ω and is used with another (fixed) resistance of 100 Ω in a voltage divider with a ±10 volt supply, as shown in Figure 2.5. The output of this voltage divider is connected to an 100× amplifier with an input resistance of 1 kΩ. What is the signal at the output of this system when the temperature sensor has a resistance of 101 Ω? What is the error that is introduced by not using an amplifier with infinite input resistance?

**Answer**: We can follow two roads that will lead to the answer. First, the circuit can be redrawn if we realize that the input resistance of the amplifier is connected to the ground (0 V), see Fig. 2.27. Labeling the currents in the resistors by the subscripts of the resistors and realizing that current cannot disappear, we get the following set of equations

$$I_1 = \frac{10\ \mathrm{V} - V_\mathrm{i}}{R_1}, \tag{2.130}$$

$$I_2 = \frac{V_\mathrm{i} - (-10\ \mathrm{V})}{R_\mathrm{S}}, \tag{2.131}$$

$$I_\mathrm{i} = \frac{V_\mathrm{i}}{R_\mathrm{i}}, \tag{2.132}$$

$$I_1 = I_2 + I_\mathrm{i}. \tag{2.133}$$

This results in

$$V_\mathrm{i} = 10\ \mathrm{V} \times \frac{1/R_\mathrm{S} - 1/R_1}{1/R_\mathrm{S} + 1/R_1 + 1/R_\mathrm{i}}. \tag{2.134}$$

For the resistances $R_1 = 100\ \Omega$, $R_\mathrm{S} = 101\ \Omega$ and $R_\mathrm{i} = 1\ \mathrm{k}\Omega$ this gives $V_\mathrm{i} = 47.38$ mV. Multiplied by the (now) ideal 100× amplifier gives an output voltage equal to 4.738 V. Alternatively, we can see the sensor as a black box with signal voltage equal to $V_\mathrm{x} = 49.75$ mV, as found by a simple voltage divider, and an output resistance $R_\mathrm{x} = R_1 \parallel R_\mathrm{S} = 50\ \Omega$. Note that this output resistance changes with the signal, as the value of $R_\mathrm{S}$ changes. In any case, the signal is only linear for small deviations from the calibration point (100 Ω), a particularity of simple voltage dividers. This 49.75 mV is fed into a voltage divider composed of the 50-Ω output resistance of the sensor and the 1 kΩ amplifier input resistance, resulting in 47.38 mV signal that is then multiplied by 100 to give an output voltage of 4.74 V, as found before. If the amplifier had an infinite input resistance, the signal from the voltage divider would not have been lowered by the current drawn by the amplifier, $V_\mathrm{i} = V_\mathrm{x} = 49.75$ mV, and the output 4.975 V. This makes the error 0.237 V (5%). Finally, note that the output resistance value taken 'from the datasheet' of 50 Ω in practice varies and this introduces an error. All these problems can be avoided by using a voltage follower/buffer after the sensor to prevent current being drawn from it.

Thus we can conclude the following

Do not connect a voltage signal from an element with non-zero output resistance to a circuit with finite input resistance, or you will lose the information

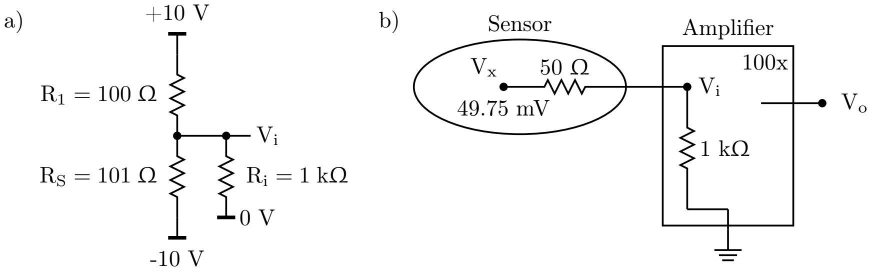


**Fig. 2.27**: a) Equivalent circuit of temperature sensor having 101 Ω resistance used in a voltage divider with a 100 Ω resistance connected to an amplifier with 1 kΩ input resistance. b) Equivalent circuit of the sensor by representing the signal coming from a signal source with 50 Ω output resistance connected to the amplifier with 1 kΩ input resistance. Both approaches lead to the same output voltage

contained within the signal.

### 2.6.2 Voltage amplifier

The next opamp circuit we will analyze is the voltage amplifier. This is achieved by using partial feedback of the output signal to the input signal. The best way to do it is with a voltage divider, see Figure 2.28. As shown above, halfway the voltage divider, the voltage is somewhere between the voltages connected at the extremes. In this case between the output voltage and ground. More precisely,

$$V_\mathrm{n} = V_\mathrm{o}\frac{R_1}{R_1 + R_\mathrm{f}}. \tag{2.135}$$

In other words, the feedback factor β, that is defined as the fraction of output voltage that is fed back into the input of the amplifying element is given by (subscript 'A' denotes the amplifier, compare to Figure 2.22).

$$\beta \equiv \frac{V_\mathrm{i,A}}{V_\mathrm{o,A}} = -\frac{R_1}{R_1 + R_\mathrm{f}}. \tag{2.136}$$

Note the negative sign, caused by the fact that the signal connected to the negative terminal that gets effectively multiplied by −1 before being fed into the amplifier. Substituting this into Eq. (2.116) directly gives the relation between input and output voltage:

$$\frac{V_\mathrm{o}}{V_\mathrm{i}} = -\frac{1}{\beta} = \frac{R_1 + R_\mathrm{f}}{R_1}. \tag{2.137}$$

It has a positive sign and that is why the circuit is sometimes called non-inverting amplifier.

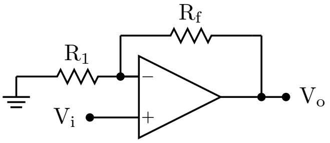


**Fig. 2.28**: Positive (non-inverting) amplifier composed of an opamp with negative feedback through a voltage divider resulting in a voltage gain $A_\mathrm{V} = V_\mathrm{o}/V_\mathrm{i} = (R_\mathrm{f} + R_1)/R_1$

The same result we can also find with our ideal-amplifier rules. First, the voltage divider causes the relation between $V_\mathrm{n}$ and $V_\mathrm{o}$ as given in Equation (2.135). Next, for an ideal amplifier (in the absence of saturation) the voltage at the negative terminal is equal to the one at the positive terminal, $V_\mathrm{p} = V_\mathrm{n}$. The positive terminal receives the input signal, thus

$$V_\mathrm{i} = V_\mathrm{o}\frac{R_1}{R_1 + R_\mathrm{f}}. \tag{2.138}$$

This directly leads to the voltage gain found in Equation (2.137).

**Question**: What would be the output if we switch the two inputs of the opamp, $V_\mathrm{p} \leftrightarrow V_\mathrm{n}$?
**Answer**: We might think that it doesn't matter, since $V_\mathrm{p} = V_\mathrm{n}$ and exchanging this still remains an equality; using our ideal opamp rules would result in the same input-output relation. Yet, it is obvious that in the feedback analysis something changes, since the feedback is now positive instead of negative, $\beta' = +R_1/(R_1 + R_\mathrm{f})$. We might think that the output is thus

$$\frac{V_\mathrm{o}}{V_\mathrm{i}} = -\frac{1}{\beta'} = -\frac{R_1 + R_\mathrm{f}}{R_1}. \tag{2.139}$$

However, this signal is metastable and in practice unstable, as discussed in the feedback section; any tiny disturbance (noise) will rapidly - and in an accelerated way - drift the output away from this metastable output voltage even if it starts there. In fact, it will probably never even arrive at this voltage, and the side at which the voltage will 'roll-off the voltage mountain' depends on the previous output voltage. See the section on hysteresis (Section 2.6.6). This shows we have to be careful with our analysis. As a rule of thumb we can say (repeat): Negative feedback will result in an output value that stabilizes, positive feedback will result in signal run-away and instability.

An inverting, negative-gain, amplifier can be made from the same circuit by applying the signal at the negative branch, as shown in Figure 2.29. By using the ideal opamp rules we easily find the output voltage. Starting with observing that $V_\mathrm{p}$ is connected to ground. Next, in the absence of saturation

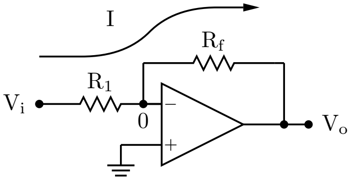


**Fig. 2.29**: Inverting, negative gain, amplifier composed of an opamp with negative feedback through a voltage divider resulting in a voltage gain $A_\mathrm{V} = V_\mathrm{o}/V_\mathrm{i} = -R_\mathrm{f}/R_1$

$V_\mathrm{p} = V_\mathrm{n}$, so the latter is also at 0 V. We call this *virtual* ground, denoted by '0' in the figure. For all purposes it is ground, but connecting it physically to ground would change the working of the circuit. Now our voltage divider $\mathrm{R_1}$-$\mathrm{R_f}$ is at one end connected to $V_\mathrm{i}$ and at the other end to $V_\mathrm{o}$. With the condition that halfway (at $V_\mathrm{n}$) the voltage is zero we can find that

$$\frac{V_\mathrm{o}}{V_\mathrm{i}} = -\frac{R_\mathrm{f}}{R_\mathrm{i}}. \tag{2.140}$$

There is an alternative way of finding this relation. Looking again at the circuit we see that resistor $\mathrm{R_1}$ feels a voltage drop from $V_\mathrm{i}$ to 0, virtual ground. The current through this resistor $\mathrm{R_1}$ is therefore given by

$$I = (V_\mathrm{i} - 0)/R_1. \tag{2.141}$$

This current has nowhere to go. Remember (Rule 2) that the input resistance of the opamp is infinite and no current enters it. There is no other way for this current but to pass through $\mathrm{R_f}$ and sink into the output of the opamp (Remember Rule 3, an opamp can sink or supply any current necessary to maintain the programmed voltage). By virtue of Ohm's law, this current induces a voltage drop in the resistance

$$\Delta V = IR_\mathrm{f}. \tag{2.142}$$

One end of the resistance is at virtual ground. The other end must therefore be at $-\Delta V$,

$$\begin{aligned} V_\mathrm{o} &= -\Delta V = -IR_\mathrm{f} \\ &= -V_\mathrm{i}\frac{R_\mathrm{f}}{R_1}, \end{aligned} \tag{2.143}$$

yielding the same input-output voltage relation as found before. We will use this 'current' analysis more often, since it is rather straightforward and the fastest way to an answer. We can use whatever strategy that comes in handy. In the case of the inverting amplifier the analysis using the current is easiest, and the feedback analysis more complicated. Still, it is very informative to perform this analysis.

The feedback diagram is shown in Figure 2.30. In this we define β as the voltage divider fraction

$$\beta \equiv \left.\frac{V_{\mathrm{n}}}{V_{\mathrm{o}}}\right|_{\mathrm{V_i=0}} = \frac{R_1}{R_1 + R_{\mathrm{f}}}, \tag{2.144}$$

which is now a value between 0 and 1. This is the part of the output that is fed back to the input. In the same manner we can define the part of the input signal that appears at the negative input terminal of the opamp. Not everything appears, because the input signal arrives at the negative input terminal through a voltage divider composed of $\mathrm{R_f}$ and $\mathrm{R_1}$

$$\left.\frac{V_{\mathrm{n}}}{V_{\mathrm{i}}}\right|_{\mathrm{V_o=0}} = \frac{R_{\mathrm{f}}}{R_{\mathrm{f}} + R_1} = (1 - \beta). \tag{2.145}$$

These two signals are summed at the input of the opamp

$$V_{\mathrm{n}} = (1 - \beta)V_{\mathrm{i}} + \beta V_{\mathrm{o}}. \tag{2.146}$$

Because we are dealing with the negative terminal of the opamp the gain is $-A$. We can also decompose this into a gain equal to $-1$ and a gain $A$, as shown in the diagram. This allows us to define an intermediate voltage $V_{\mathrm{x}}$.

$$V_{\mathrm{x}} = -V_{\mathrm{n}} = (\beta - 1)V_{\mathrm{i}} - \beta V_{\mathrm{o}}. \tag{2.147}$$

The output is this voltage multiplied by $A$ and we get the final equation

$$V_{\mathrm{o}} = AV_{\mathrm{x}} = [(\beta - 1)V_{\mathrm{i}} - \beta V_{\mathrm{o}}]A, \tag{2.148}$$

a simple linear equation with one unknown, easy to solve:

$$\frac{V_{\mathrm{o}}}{V_{\mathrm{i}}} = \frac{A(\beta - 1)}{1 + A\beta}. \tag{2.149}$$

Finally, for an ideal amplifier ($A = \infty$) this reduces to

$$\frac{V_{\mathrm{o}}}{V_{\mathrm{i}}} = \frac{\beta - 1}{\beta}. \tag{2.150}$$

Substituting the value of β from Eq. (2.144) results in Equation (2.143).

**Question**: Design the feedback diagram of the positive amplifier of Figure 2.28 that results in the input-output relation found in Equation (2.137). Note: use a positive β.
**Answer**: Figure 2.31 shows the diagram of the circuit with the relevant points.

$$\begin{aligned} V_{\mathrm{x}} &= V_{\mathrm{p}} - V_{\mathrm{n}} = V_{\mathrm{i}} - \beta V_{\mathrm{o}}, && (2.151)\\ V_{\mathrm{o}} &= AV_{\mathrm{x}}. && (2.152) \end{aligned}$$

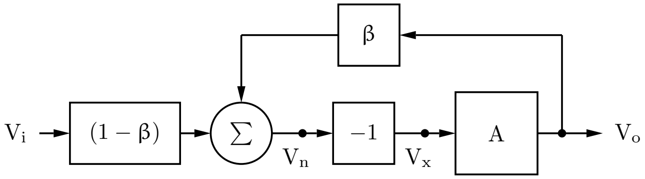


**Fig. 2.30**: Feedback diagram for a circuit using an inverting amplifier, or using the negative input terminal of an opamp (for example the circuit of Fig. 2.29). The result is the relation $V_\mathrm{o}/V_\mathrm{i} = (\beta - 1)/\beta$

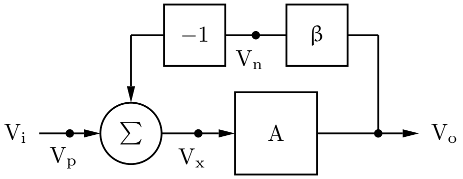


**Fig. 2.31**: Feedback diagram for a circuit using an opamp with input at the positive terminal and feedback over the negative terminal as for example in Fig. 2.28. The result is the relation $V_\mathrm{o}/V_\mathrm{i} = -1/\beta$

The solution being

$$\frac{V_\mathrm{o}}{V_\mathrm{i}} = \frac{A}{1 + A\beta}, \tag{2.153}$$

with $\beta \equiv R_1/(R_1 + R_\mathrm{f})$ positive. For an ideal opamp ($A = \infty$) this reduces to

$$\frac{V_\mathrm{o}}{V_\mathrm{i}} = \frac{1}{\beta} = \frac{R_1 + R_\mathrm{f}}{R_1}. \tag{2.154}$$

### 2.6.3 Differential amplifier

We can now calculate what happens if we connect signals to both entrances of the fed-back amplifier, as shown in Figure 2.32. The operational amplifier is a linear circuit, and this means that the output for any two signals connected is the sum of the output signals that would have occurred at the output had only one been connected to the input. This is the superposition principle that applies to all linear circuits. We can calculate the output for $\mathrm{V_{in}}$ connected to a signal and $\mathrm{V_{ip}}$ to ground, then calculate the output for $\mathrm{V_{in}}$ connected to ground $\mathrm{V_{ip}}$ to a signal, and then sum the results. We thus get

$$V_\mathrm{o} = V_\mathrm{ip}\frac{R_\mathrm{f} + R_1}{R_1} - V_\mathrm{in}\frac{R_\mathrm{f}}{R_1}. \tag{2.155}$$

As can be seen, the result is a differential amplifier - the amplifier amplifies the difference between the input signals, with a slightly unbalanced gain for the two

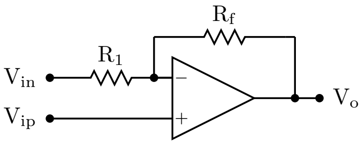


**Fig. 2.32**: Unbalanced differential amplifier composed of an opamp with negative feedback through a voltage divider resulting in an input-output relation $V_\mathrm{o} = V_\mathrm{ip} \times (R_\mathrm{f} + R_1)/R_1 - V_\mathrm{in} \times R_\mathrm{f}/R_1$

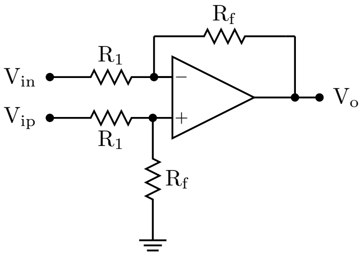


**Fig. 2.33**: Balanced differential amplifier, $V_\mathrm{o} = R_\mathrm{f}/R_1 \times (V_\mathrm{ip} - V_\mathrm{in})$

input signals. For large values of $R_\mathrm{f}$ compared to $R_1$ this unbalance disappears. To overcome this unbalance altogether, we can reduce the gain for the positive terminal by a voltage divider composed of $R_1$ and $R_\mathrm{f}$, as shown in Figure 2.33. This multiplies the signal $V_\mathrm{ip}$ with a factor $R_\mathrm{f}/(R_\mathrm{f} + R_1)$ before feeding it into the opamp. For this circuit we thus find the balanced differential amplification:

$$V_\mathrm{o} = \frac{R_\mathrm{f}}{R_1}(V_\mathrm{ip} - V_\mathrm{in}). \tag{2.156}$$

It is obvious that the balanced differential amplifier is only balanced if the resistors are balanced. The two $\mathrm{R_f}$s need to be exactly the same, as well as the two $\mathrm{R_1}$s.

**Question**: Calculate the output voltage for a balanced differential amplifier ($R_\mathrm{f} = 100\ \mathrm{k\Omega}$, $R_1 = 1\ \mathrm{k\Omega}$) for $V_\mathrm{ip} = V_\mathrm{in} = 1$ V. Repeat the same for when one of the $\mathrm{R_1}$ resistances is 1% higher.

**Answer**: For a balanced amplifier with balanced resistors the output is the input difference multiplied by $R_\mathrm{f}/R_1$, as in Eq. (2.156). Since the input difference is zero, so is the output voltage, $V_\mathrm{o} = 0$. Now we replace one resistor $\mathrm{R_1}$ with one with a slightly higher resistance, $1.01R_1$, for instance the one in the positive branch. The output voltage is now

$$V_\mathrm{o} = \left(\frac{R_\mathrm{f}}{R_\mathrm{f} + 1.01R_1} \times \frac{R_\mathrm{f} + R_1}{R_1} - \frac{R_\mathrm{f}}{R_1}\right) \times 1\ \mathrm{V}, \tag{2.157}$$

where we used the expressions for the individual contributions of the non-inverting and inverting amplifier, respectively, the former multi-

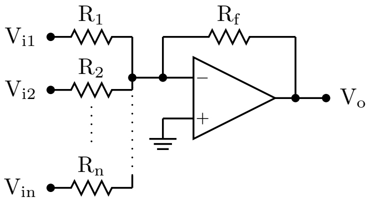


**Fig. 2.34**: Summing amplifier, $V_\mathrm{o} = -R_\mathrm{f} \sum(V_{\mathrm{i},i}/R_i)$

plied with an appropriate voltage-divider factor. With the knowledge that $R_\mathrm{f}/R_1 = 100$ we easily find that $V_\mathrm{o} \approx -10$ mV.

**Question**: What are the input and output resistance of the balanced differential amplifier?
**Answer**: The output resistance is zero; the opamp can supply whatever current necessary to maintain the programmed voltage. $r_\mathrm{o} \equiv \mathrm{d}V_\mathrm{o}/\mathrm{d}I_\mathrm{o} = 0$. The input resistance of the negative input is equal to $R_1$ because on the right side of this resistor is a fixed voltage, independent of $V_\mathrm{in}$: $V_\mathrm{n} = V_\mathrm{p} = V_\mathrm{ip} \times R_\mathrm{f}/(R_\mathrm{f} + R_1)$. The input current is $I_\mathrm{in} = (V_\mathrm{in} - V_\mathrm{p})/R_1$, and $r_\mathrm{in} \equiv \mathrm{d}V_\mathrm{in}/\mathrm{d}I_\mathrm{in} = R_1$. The input resistance of the positive input is equal to $R_1 + R_\mathrm{f}$. With the ideal opamp not drawing any current, the input feels only the voltage divider composed of R$_1$ and R$_\mathrm{f}$. The input current thus becomes $I_\mathrm{ip} = V_\mathrm{ip}/(R_1 + R_\mathrm{f})$ and the input resistance $r_\mathrm{ip} \equiv \mathrm{d}V_\mathrm{ip}/\mathrm{d}I_\mathrm{ip} = R_1 + R_\mathrm{f}$.

### 2.6.4 Summing amplifier

A summing amplifier can be made by adding more resistances to the negative terminal of the opamp to which a signal is fed, see Figure 2.34. Each signal comes with its own multiplication factor. Since $V_\mathrm{n}$ of the opamp is at virtual ground, each input signal brings a current equal to $I_{\mathrm{i},i} = V_{\mathrm{i},i}/R_i$. Through resistor R$_1$ passes a current $V_\mathrm{i1}/R_1$, through resistor R$_2$ passes a current $V_\mathrm{i2}/R_2$, etc. Once again, these currents have nowhere to go, and are forced through the feedback resistor R$_\mathrm{f}$ inducing a voltage drop in this resistor. With one foot of the resistor at (virtual) ground, the total output voltage thus becomes

$$V_\mathrm{o} = -R_\mathrm{f} \sum_i \left( \frac{V_{\mathrm{i},i}}{R_i} \right) . \tag{2.158}$$

**Question**: Inspired by the summing amplifier, and somewhat annoyed by the overall negative sign, somebody designed the circuit of Figure 2.35 for a *positive* summing amplifier. Why does this not work? What is the relation between output and input signals?
**Answer**: The difference between the negative summing circuit (Fig.

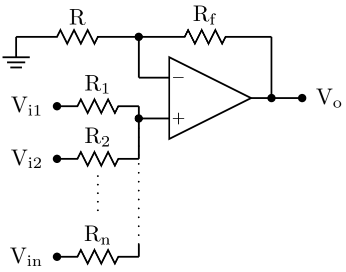


**Fig. 2.35**: Wrong design of a positive summing amplifier

2.34) and the positive circuit of Fig. 2.35 is that in the former the input signals come through resistances ($\mathrm{R}_1$ .. $\mathrm{R}_\mathrm{n}$) connected to (virtual) ground of $\mathrm{V}_\mathrm{n}$, while in the latter the resistances are connected to a 'floating' input, $\mathrm{V}_\mathrm{p}$. Given the fact that the input resistance of the opamp is infinite, the voltage at this terminal is the result of a multi-resistance voltage divider. Because this is a linear circuit we can use the superposition principle: calculate the contribution of each input voltage by setting the others to zero and then adding these to find $V_\mathrm{p}$. For instance, for $V_\mathrm{i1}$ we set all the other inputs to ground and we find

$$V_\mathrm{p} = V_\mathrm{i1} \times \frac{R_1}{R_1 + (R_2^{-1} + R_3^{-1} + ... + R_\mathrm{n}^{-1})^{-1}}. \tag{2.159}$$

The other contribution can be found in a similar way. It is obvious that this will not lead to a simple sum or (weighted) average of the input voltages, except for $n = 2$, a double-input summing amplifier. With the help of the non-inverting amplifier we find for the output voltage the complicated result

$$V_\mathrm{o} = \frac{R_\mathrm{f} + R}{R} \times \sum_{\mathrm{i}=1}^{n} \left[ \left( 1 + \left\{ R_\mathrm{i} \left[ \left( \sum_{\mathrm{j}=1}^{n} \frac{1}{R_\mathrm{j}} \right) - \frac{1}{R_\mathrm{i}} \right] \right\} \right)^{-1} \times V_{\mathrm{i},i} \right]. \tag{2.160}$$

A result you can immediately forget. (But, admit it, you are also impressed by the number of brackets!) This circuit is never used. A positive summing circuit can be made by the combination of a negative summing circuit and a negative amplifier. Interestingly, for $n = 2$ it does work,

$$V_\mathrm{o} = \frac{R_\mathrm{f} + R}{R(R_1 + R_2)} \times (V_1 R_1 + V_2 R_2). \tag{2.161}$$

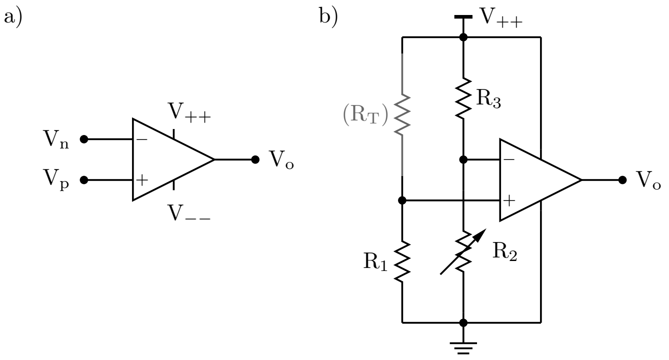


**Fig. 2.36**: a) Comparator, b) Example of a comparator used in a temperature-control system with a a Wheatstone bridge at the entrance. One of the resistances of the Wheatstone bridge is variable to be able to tune the circuit to have zero output for the quiescent 'bias' point. This allows for larger amplification in the amplifier and hence larger overall sensitivity

### 2.6.5 Comparator

A comparator is an opamp without any feedback, see Figure 2.36. Because the gain of an ideal opamp is infinite, any tiny difference at the input terminals will cause saturation at the output, either positive, when $V_\mathrm{p}$ is larger than $V_\mathrm{n}$, or negative when $V_\mathrm{n}$ is larger than $V_\mathrm{p}$, in other words,

$$V_\mathrm{o} = \begin{cases} V_{++} & \text{if } V_\mathrm{p} > V_\mathrm{n} \\ V_{--} & \text{if } V_\mathrm{p} < V_\mathrm{n}. \end{cases} \tag{2.162}$$

An ideal voltage comparator always saturates. Without feedback, or with $\beta = 0$, the gain becomes equal to $A$ and this is infinite for an ideal opamp. Note that because of saturation, the rule that $V_\mathrm{p} = V_\mathrm{n}$ is no longer valid.

The application of a comparator is that it converts the information to a binary value that can further be processed 'digitally'. This is especially useful if we have to make a yes/no decision, as in "Do we need to connect the heater, or not?". The information (temperature) from a sensor can be compared to a reference value and a decision can be made by electronics, instead of a computer. Such a system is shown in Figure 2.36b. The system is composed of a comparator and a Wheatstone bridge. The measurement branch of the Wheatstone bridge contains the temperature sensor (a temperature-dependent resistance) and the other branch - the reference branch - is tuned by the variable resistor $R_2$ to give the same voltage as the measurement branch at the commutation temperature, i.e. the temperature at which the output should change. Note the absence of feedback $V_\mathrm{o} \to V_\mathrm{i}$, a telltale sign that we are dealing with a comparator and that the output saturates at either of the supply voltages, in this case $V_{++}$ or 0.

While any opamp can be used in a comparator circuit, some are more adequate than others. Especially the "311" family is popular, due to its high commutation speed, switching.

### 2.6.6 Hysteresis. Schmitt trigger

Positive feedback can be used to introduce hysteresis in the comparator behavior, resulting in a so-called Schmitt trigger. While circuits that use negative feedback stabilize at a defined output voltage, allowing us to arrive at the equality $V_\mathrm{p} = V_\mathrm{n}$ for an ideal opamp found before (Rule 5), circuits that use no or positive feedback are unstable, as seen in Section 2.5.3. What this implies for our opamps is that their output voltage will saturate at the supply voltage. Moreover, or because of that, no longer the rule that both entrances are at the same voltage applies. An example of such a positive-feedback opamp-circuit is given in Figure 2.37. The circuit uses positive feedback through a voltage-divider composed of $\mathrm{R_f}$ and $\mathrm{R_1}$, making the feedback factor

$$\beta = \frac{R_1}{R_1 + R_\mathrm{f}}, \tag{2.163}$$

a number between 0 and +1. The output will saturate at either $V_{++}$ or $V_{--}$, the actual value depending not only on the input voltage, but also on the *history* of the input signal. Imagine $V_\mathrm{o}$ for some reason is at $V_{++}$. The positive input of the opamp is then at a steady $V_\mathrm{p} = \beta V_{++}$. As long as the negative terminal - the input signal - stays below this value, the comparator opamp output stays high. The moment $V_\mathrm{i}$ rises above $V_\mathrm{p} = \beta V_{++}$ the output of the comparator switches to $V_{--}$. From that moment on, the positive input terminal is at $V_\mathrm{p} = \beta V_{--}$, the low supply voltage. To switch it back to high it is not enough to lower the input voltage to below $\beta V_{++}$. To incur a switch of the output back to $V_{++}$ the input has to drop below $\beta V_{--}$. To summarize: For input voltages below $\beta V_{--}$ the output is high, $V_{++}$ and for input voltages above $\beta V_{++}$ the output is low. For voltages in between these two values, the output is undetermined; it depends on the history. See Figure 2.37(b). We can define the two threshold voltages for switching:

$$\begin{aligned} V_\mathrm{L} &= \beta V_{--}, & (2.164)\\ V_\mathrm{H} &= \beta V_{++}. & (2.165) \end{aligned}$$

When connecting the input at the positive terminal the behavior is slightly different, see Figure 2.38. The behavior can be calculated if we use the superposition principle. The voltage at the positive entrance of the opamp is given by the two voltages $V_\mathrm{i}$ and $V_\mathrm{o}$, the contribution of each is found by placing the other at zero. Then summing the individual voltage-divider contributions gives,

$$\begin{aligned} V_\mathrm{p} &= \left(\frac{R_\mathrm{f}}{R_\mathrm{f} + R_1}\right) V_\mathrm{i} + \left(\frac{R_1}{R_\mathrm{f} + R_1}\right) V_\mathrm{o} \\ &= (1-\beta)V_\mathrm{i} + \beta V_\mathrm{o}. & (2.166) \end{aligned}$$

This value is compared with the negative terminal which is in our circuit connected to ground, $V_\mathrm{n} = 0$. Once again, there are two possibilities. If the output

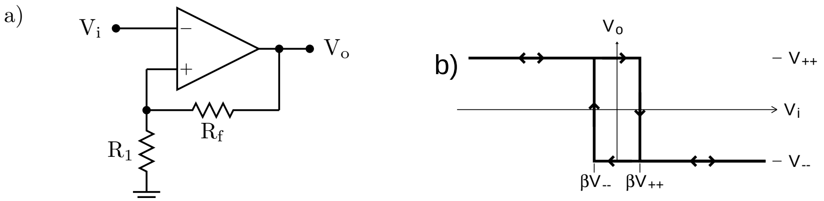


**Fig. 2.37**: a) Schmitt-trigger hysteresis circuit using an opamp with positive feedback. b) Behavior of the circuit. For some input voltages, the output voltage depends on the history. $\beta \equiv R_1/(R_1 + R_f)$

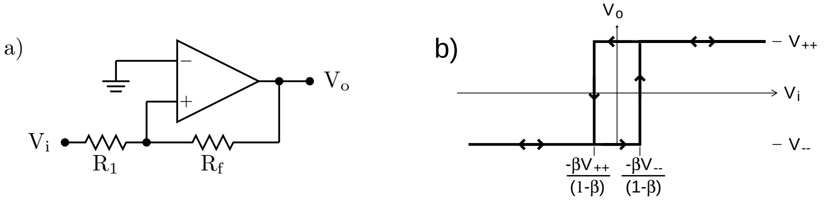


**Fig. 2.38**: a) Schmitt trigger: Hysteresis circuit using an opamp with positive feebdback with signal input at the positive terminal. b) Behavior of the circuit, $\beta \equiv R_1/(R_1 + R_f)$

is high ($V_o = V_{++}$), the output will not switch to low until $V_p < V_n = 0$, when $V_i$ drops below the lower threshold voltage ($V_L$). And if the output is low ($V_o = V_{--}$), it will not switch to high until $V_p > V_n = 0$, when the input voltage rises above the higher threshold voltage ($V_H$). These are give by

$$\begin{aligned} V_L &= -\frac{\beta}{1-\beta} V_{++}, & (2.167) \\ V_H &= -\frac{\beta}{1-\beta} V_{--}, & (2.168) \end{aligned}$$

respectively. An example is given in Figure 2.38(b).

The advantage of using hysteresis is that it avoids multiple commutations when the input signal is close to the switching value. Think of a system that switches on the illumination in a room if outside it is getting dark. When the sensor signal is close to the switching level, noise (passing of a cloud in the sky, or a car with lights on passing in the street) can temporarily lift the signal above or below the threshold. What happens is that we get multiple transitions in a short period of time. Or, in other words, our lights at home will become a stroboscope. To avoid this, we define two levels. To switch on the light, the sensor must signal above a certain level, while to switch them off again, the signal must reach below *another*, lower level. Figure 2.39 shows an example. When a single (central) level is defined, multiple commutations occur for noisy signals. When two levels are defined, one for switching on and

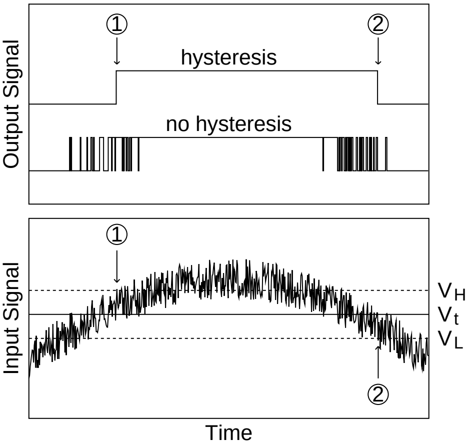


**Fig. 2.39**: Example of the use of hysteresis. Without hysteresis, the output signal is high when the input signal is above a certain threshold level $V_\mathrm{t}$ and low otherwise. The result is that, when the input signal is noisy, multiple transitions occur, as evident in the lower trace of the top panel. To avoid this we can define two levels. For the output signal to switch to high, the input signal has to be larger than the upper level $V_\mathrm{H}$, whereas to drop back to the low output state, the input signal has to be smaller than the lower threshold level, $V_\mathrm{L}$. The result is that, for the same input signal, only two transitions occur, Low → high (indicated by 1) and high → (indicated by 2), as shown in the top trace of the top panel

another for switching off, the multiple transitions are avoided. Note, however, the apparent 'delay' effect introduced by the hysteresis, an unavoidable side effect. An alternative way of avoiding multiple transitions is to define a 'dead time', a time in which the output is not allowed to change.

### 2.6.7 Current-to-voltage converter

With the same ideas we used for the voltage amplifier we can analyze the current-to-voltage converter. The basic circuit is shown in Figure 2.40(a). It consists of an opamp with negative feedback through a resistance $\mathrm{R_f}$. To this circuit, the signal, in the form of a current $I$, is applied at the negative terminal. Once again, due to the infinite input resistance of the opamp (ideal opamp Rule 2), this current has nowhere to go but pass through the resistance $\mathrm{R_f}$ and sink into the (zero resistance) output terminal of the opamp. This current $I$ thus forces a voltage drop in the resistance equal to $\Delta V = IR_\mathrm{f}$. One side of the resistance is connected to virtual ground (because $V_\mathrm{p}$ is at physical ground), so the other side is at $V_\mathrm{o} = -IR_\mathrm{f}$. The ideal opamp supplies through the output whatever current necessary to maintain this voltage. Depending on the sign it

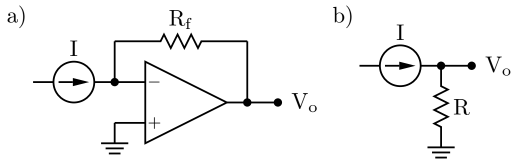


**Fig. 2.40**: a) Current-to-voltage converter. The output voltage is a function of the input current, $V_o = -IR_f$. b) An alternative current-to-voltage converter, $V_o = IR$ that has the disadvantage of having non-zero output resistance

will sink or supply this current.

The same functionality can be achieved by a simple resistance connected to ground, see Figure 2.40(b). In this case the output voltage depends on the input current as $V_o = I_i R$. Apart from the sign, the difference between the two circuits is (once again) the output resistance: For the current-to-voltage converter using an opamp this output resistance is zero - $r_o = \mathrm{d}V_o/\mathrm{d}I_o = 0$; the opamp can supply any current needed to maintain the voltage - while for the simple resistance it is equal to this resistance. Once we start drawing current by an external circuit, less current passes through $R_f$ and less voltage drop is induced, thus lowering $V_o$.

### 2.6.8 Integrator/differentiator

An integrator can be made by placing a capacitor in the feedback loop, as shown in Figure 2.41(a). With the negative terminal of the opamp at virtual ground, the input current is given by $I = V_i/R$. With nowhere to go, the current passes through the capacitor, charging it. In the section before on filters we have seen that the current is the time-derivative of charge, so the charge is the integral of current over time,

$$Q_C = \int I(t)\mathrm{d}t. \qquad (2.169)$$

With the voltage drop of a capacitor given as the charge divided by the capacitance, and because the left foot of the capacitor is connected at ground, the output voltage is given by

$$V_o(t) = -\Delta V_C(t) = -\frac{Q_C(t)}{C} = -\frac{1}{RC}\int V_i(t)\mathrm{d}t. \qquad (2.170)$$

In other words, the output voltage is the integral of the input voltage, divided by the integration time-constant $\tau = RC$.

To make a differentiator we can exchange the resistor and capacitor, see Figure 2.41(b). The input current is then given by

$$I = C\frac{\mathrm{d}V_i(t)}{\mathrm{d}t}. \qquad (2.171)$$

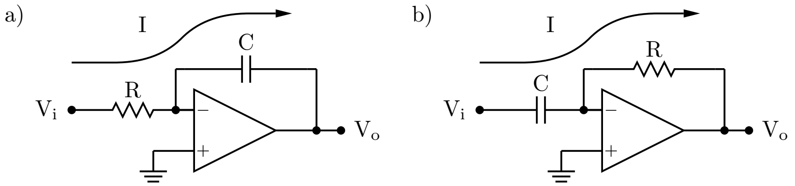


**Fig. 2.41**: a) Integrator. b) Differentiator

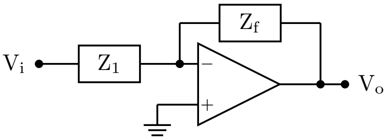


**Fig. 2.42**: Generalized circuit with impedance elements $Z_1$ and $Z_f$ as the feedback elements. These can depend on frequency and this results in a frequency dependent transfer function

The current is translated to a voltage drop by the resistance R. One side of the resistance being at virtual ground results in

$$V_o(t) = -RC\frac{dV_i(t)}{dt}. \tag{2.172}$$

In both cases, the differentiator and integrator, the signals at the output might be limited by the power supply. In the case of the integrator, it means that for instance DC signals et $V_i$ can only be integrated up to a certain time. For the differentiator is means that signals cannot change too fast.

### 2.6.9 Active filters

Active filters can be made with opamps by replacing the resistors by other elements, for instance capacitors or inductors. Figure 2.42 shows a general circuit of an opamp using feedback with generalized impedance elements, $Z_f$ and $Z_1$. The gain of a circuit can be given directly by replacing resistances by impedances, for example $R_1 \to Z_1$ and $R_f \to Z_f$. Alternatively we can use the feedback analysis for the circuits with feedback. The feedback factor β of the circuit in Figure 2.42 is given by

$$\beta = \frac{Z_1}{Z_1 + Z_f}, \tag{2.173}$$

and this can be a function of frequency, making the amplification of the circuit a function of the frequency.

As an example, take the circuit shown in Figure 2.43(a). Comparing to Figure 2.42 we see that $Z_1 = R_1$ and $Z_f = R_f \parallel 1/j\omega C_f = (1/R_f + j\omega C_f)^{-1}$ . Moreover, for this circuit, with negative feedback and input at the negative

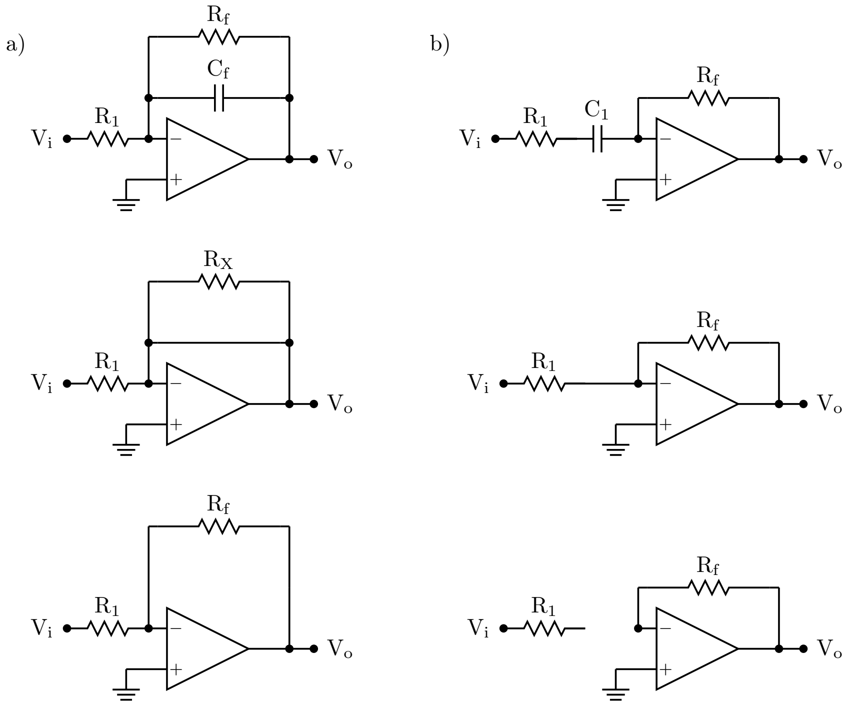


**Fig. 2.43**: Active low-pass (a) and high-pass (b) filters. Top: full circuit. Middle: equivalent circuit at high frequencies. Bottom: Equivalent circuit at low frequencies

side, we had found that the input-output relation is given by $V_\mathrm{o}/V_\mathrm{i} = (\beta - 1)/\beta$,

$$\begin{aligned} A_\mathrm{v}(\omega) \equiv \frac{V_\mathrm{o}}{V_\mathrm{i}} &= \frac{\beta - 1}{\beta} = -\frac{Z_\mathrm{f}}{Z_1} \\ &= -\frac{1}{R_1(1/R_\mathrm{f} + j\omega C_\mathrm{f})} \\ &= -\frac{R_\mathrm{f}}{R_1} \times \frac{1}{1 + j\omega R_\mathrm{f} C_\mathrm{f}}. \end{aligned} \tag{2.174}$$

Comparing this to the passive low filters we see that this active filter has a gain equal to $-R_\mathrm{f}/R_1$ at low frequencies, a gain we would have found if we had considered the capacitor as an open circuit for low frequencies. At high frequencies the gain falls off at a rate equal to the passive LPF. In fact, the circuit behaves like the combination of an amplifier with gain equal to $-R_\mathrm{f}/R_1$ followed (or preceded) by a passive low-pass filter with cut-off frequency equal to $f_0 = 1/2\pi R_\mathrm{f} C_\mathrm{f}$. Because the filter has an amplifier and the gain can be larger than unity, we call this an active filter.

For the high-pass filter of Figure 2.43(b) we can make the same sort of analysis. Given the fact that $Z_1 = R_1 + 1/j\omega C_1$ and $Z_\mathrm{f} = R_\mathrm{f}$ and that the

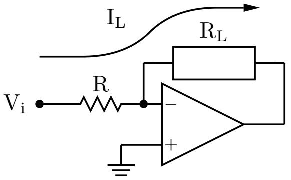


**Fig. 2.44**: Voltage-to-current converter. The current through the load resistance $\mathrm{R_L}$ is given by $I_\mathrm{L} = V_\mathrm{i}/R$

overall amplification is given by $v_\mathrm{o}/v_\mathrm{i} = -Z_\mathrm{f}/Z_1$ we find that

$$\begin{aligned} A_\mathrm{v}(\omega) \equiv \frac{v_\mathrm{o}}{v_\mathrm{i}} &= \frac{\beta - 1}{\beta} = -\frac{Z_\mathrm{f}}{Z_1} \\ &= -\frac{R_\mathrm{f}}{R_1} \times \frac{1}{1 + 1/j\omega R_1 C_1}. \end{aligned} \tag{2.175}$$

Comparing this to the passive high-pass filter of section 2.3 we see that the circuit behaves exactly like the combination of an amplifier with gain equal to $-R_\mathrm{f}/R_1$ followed by a passive high-pass filter with cut-off frequency equal to $f_0 = 1/2\pi R_1 C_1$. A qualitative behavior we would have found had we used the fact that a capacitor is effectively a short circuit at high frequencies and an open circuit at low frequencies; the gain at low frequencies would then be $-R_\mathrm{f}/(R_1 + \infty) = 0$ and at high frequencies $-R_\mathrm{f}/(R_1 + 0) = -R_\mathrm{f}/R_1$. These thoughts help us to verify our calculations; it is easy to make a sign error somewhere and wind up with erroneous results.

## 2.6.10 Voltage-to-current converter

With the current-to-voltage converter shown in Figure 2.40 sometimes we want to do the opposite, convert a voltage into a current. Figure 2.44 shows a possible implementation. Actually, this circuit is equal to the inverting-amplifier circuit (Fig. 2.29), with the feedback resistance $\mathrm{R_f}$ replaced by the load circuit $\mathrm{R_L}$. It is easy to calculate the current through this load. Assuming zero difference in the input terminals $\mathrm{V_p}$ and $\mathrm{V_n}$ and the latter thus being virtual ground, the input current can be calculated as $I_\mathrm{i} = V_\mathrm{i}/R$. This current is forced through the load since it cannot enter the input terminal of the opamp. Note that this makes the other side of the load to have a negative voltage, possibly a negative side effect. In any case, this circuit is just an example of what can be done with an opamp, without it being a useful circuit; a voltage-to-current converter can be much better made by transistors.

## 2.6.11 Instrumentation operational amplifiers

While for normal opamp applications it is enough that not too much signal is lost because of the non-infinite input resistance, for instrumentation applications, any loss of the signal due to current being drawn can seriously damage the

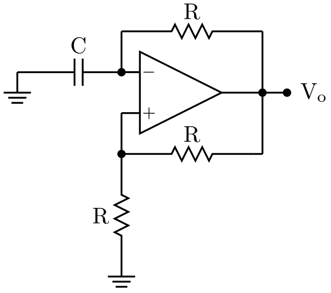


**Fig. 2.45**: Relaxation Oscillator

quality of the information in the signal. In some cases, like the pH sensor shown in the next chapter, the input resistance has to be as high as possible to avoid unwanted effects such as inducing a chemical reaction in the measured sample or warming up the sample by passing current through the sensor. For these situations manufacturers have special instrumentation amplifiers that have an extremely high input resistance.

## 2.7 Oscillators and timing circuits

The range of opamp circuits is sheer infinite. It is not possible to show them all here. Yet, there are some circuits that are useful and informative. In this section some advanced opamp circuits are described. We start with some oscillators.

### 2.7.1 Relaxation oscillator

Oscillators make use of the instability principle. Take for example the Relaxation Oscillator of Figure 2.45, named after the fact that after each commutation of the output voltage, the system tries to relax to a new stable state (but never reaches there; before it happens, the output already commutates again). The circuit has both positive and negative feedback. We can expect oscillations and/or saturation when - or for those frequencies where - the positive feedback wins. For low frequencies the capacitor is an open circuit and we have negative feedback equal to unity. The positive feedback is made up of resistors and is therefore independent of frequency and for the two equal resistances the voltage divider gives a feedback factor equal to 0.5. Thus, we conclude that the circuit is stable at DC and will not saturate. For high frequencies the capacitor is a short circuit and the negative feedback reduces to zero. Thus, for high frequencies the positive feedback wins and the circuit will oscillate.

Let us analyze the circuit in a different way, namely in the time domain. We can see that the circuit behaves like a comparator with time-dependent voltages at the two input terminals. Imagine the output at the positive supply voltage. For the sake of the calculation we will assume the supply voltages are ±1. (In fact, the functionality of the circuit does not depend on the exact values of the supply voltages). The output is thus at +1. The positive terminal is then at 0.5. Now imagine we start with the capacitor empty, like just after switching on the circuit. The negative terminal is thus at zero (remember $\Delta V_{\mathrm{C}} = Q/C$) and the comparator indeed results in a high level at the output. Seemingly we are in a stable situation. However, with one side of the resistor in the negative feedback loop at +1 and the other side at 0, a current will flow from $V_{\mathrm{o}}$ to $V_{\mathrm{n}}$. Since it cannot enter the (ideal) amplifier input, it is fully used to charge the capacitor C. Charging the capacitor will increase its voltage drop. But, this increase in voltage at $V_{\mathrm{n}}$ will decrease the voltage drop across the resistor and decrease the current of charging the capacitor. This is a classical relaxation system. One that we have seen in the RC filters. We know that such a system, composed of a resistor and a capacitor, will behave as an exponential decay or approach. If further nothing changes, the current to charge the capacitor will peter out. Eventually the current will be zero at which point the voltage drop across the resistor must be zero, thus $V_{\mathrm{n}}(t=\infty) = V_{\mathrm{o}}$.

However, it will never reach that value. Don't forget, the opamp works like a comparator. The moment $V_{\mathrm{n}}$ overtakes the positive terminal at +0.5, the output commutates, $V_{\mathrm{o}} \to -1$. Here we set the clock to zero and start calculating what happens. At this moment, $t = 0$, we have $V_{\mathrm{o}} = -1$, $V_{\mathrm{p}} = -0.5$ and $V_{\mathrm{p}} = +0.5$. The resistor of the negative-feedback loop feels a voltage drop of $V_{\mathrm{o}} - V_{\mathrm{n}} = -1.5$, i.e., a current is drawn out of the capacitor that is discharging it. The discharging continues and completely empties the capacitor. Still the negative current continues and the capacitor starts being charged negatively. Once again, this is a relaxation behavior. The current diminishes as the capacitor voltage changes. The tendency, for $t = \infty$ is to have zero current. This occurs for zero voltage drop across the resistor, $V_{\mathrm{n}} = V_{\mathrm{o}}$. This is a typical relaxation system. We can repeat here the calculation in frequency, or solve the differential equation as we have done for the RC filters, but instead of getting bogged down again in tedious calculations we will directly go to the trophy. We have a exponential decay/approach system that has the following conditions:

- The starting voltage at the negative terminal is $V_{\mathrm{n}}(0) = +0.5$.
- The final voltage (if it ever reaches there) is $V_{\mathrm{n}}(\infty) = -1$.
- The relaxation time is $\tau = RC$.

The solution to this is

$$V_{\mathrm{n}}(t) = -1 + 1.5\mathrm{e}^{-t/\tau}, \tag{2.176}$$

with $\tau = RC$. The voltage never reaches the final value, because the moment it drops below the voltage at the positive input terminal, $V_{\mathrm{p}} = -0.5$, the output commutates and we start the second part of a cycle. Analyzing the above equation we see that that happens at

$$t_1 = \tau \ln(3). \tag{2.177}$$

The output commutates, $V_{\mathrm{o}} \to +1$ and the second half of a cycle is started. Again we set the clock to zero and calculating what happens. With the positive terminal at $V_\mathrm{p} = 0.5$, initially the negative terminal is at $V_\mathrm{n} = -0.5$. A current flows from the output through the resistor in the negative-feedback loop. This cancels the (negative) charge in the capacitor and later charges it positively. The voltage drop at the capacitor therefore increases and, with one foot connected at ground, the other side, $V_\mathrm{n}$, increases with it. Everything in a typical relaxation manner resulting in a exponential decay/approach behavior with the following conditions:

- The starting voltage at the negative terminal is $V_\mathrm{n}(0) = -0.5$.
- The final voltage is the situation where there is zero current, $V_\mathrm{n}(\infty) = V_\mathrm{o} = +1$.
- The relaxation time is $\tau = RC$.

The solution to this is

$$V_\mathrm{n}(t) = 1 - 1.5\mathrm{e}^{-t/\tau}, \tag{2.178}$$

with $\tau = RC$. The moment this rises above the voltage at the positive input terminal, $V_\mathrm{p} = +0.5$, the output switches again and we start a new cycle with the first part as described before. Analyzing the above equation we see that that happens at

$$t_2 = \tau \ln(3). \tag{2.179}$$

If we replace the two resistances R of the positive feedback loop by unequal resistances $\mathrm{R}_1$ and $\mathrm{R}_2$, such that the feedback factor is $\alpha = R_1/(R_1 + R_2)$, it is not difficult to show that the times above are given by

$$t_1 = t_2 = \tau \ln\left(\frac{1+\alpha}{1-\alpha}\right). \tag{2.180}$$

The total period of a cycle is $T = t_1 + t_2$. With the frequency equal to the reciprocal of the period we reach the conclusion that the Relaxation Oscillator of Figure 2.45 generates a square-wave output with at a frequency

$$f_\mathrm{osc} = \left[2RC \ln\left(\frac{1+\alpha}{1-\alpha}\right)\right]^{-1}. \tag{2.181}$$

Figure 2.46 gives an example of the signals at the output and at the negative input terminal as a function of time. At the input terminal we can recognize the interrupted relaxations that look like shark fins. At the output (and at the positive input terminals scaled with a factor $\alpha$) we see the square wave.

### 2.7.2 Wien oscillator

The Wien bridge oscillator, named after Max Wien who developed it in the 19th century, is shown in Figure 2.47. It consists of an operational amplifier with negative feedback through resistances $\mathrm{R_f}$ and $\mathrm{R}_1$ (resulting in an amplifier with gain $1{+}R_\mathrm{f}/R_1$) and a positive feedback loop composed of two RC pairs, one

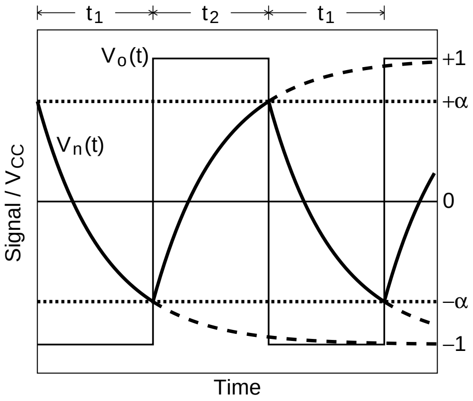


**Fig. 2.46**: Behavior of the relaxation oscillator shown in Fig. 2.45. Shark-fin waves are signs of RC-relaxation charge-discharge oscillations

connected in series and one in parallel. It is not difficult to see for what values of the components this circuit will oscillate. For that we use Barkhausen's criterion that tells us that for positive feedback the circuit will oscillate for that frequency where $A\beta$, the product of open-loop gain and feedback, is equal to unity. Our circuit has an open-loop gain given by

$$A = 1 + R_\mathrm{f}/R_1. \tag{2.182}$$

The positive feedback loop is made up of the voltage divider $Z_\mathrm{p}$ and $Z_\mathrm{s}$, which are the impedance of the parallel RC and the series RC circuit respectively. Given the fact that the impedance of a resistor is given by $R$ and of a capacitor by $1/sC$, with $s = j\omega$, we find

$$\begin{aligned} Z_\mathrm{p} &= (R^{-1} + sC)^{-1} = \frac{R}{1+sRC}, \\ Z_\mathrm{s} &= R + 1/sC, \\ \beta &\equiv \frac{V_\mathrm{p}}{V_\mathrm{o}} = \frac{Z_\mathrm{p}}{Z_\mathrm{p}+Z_\mathrm{s}} = \frac{1}{3+sRC+1/sRC}, \\ A\beta(\omega) &= \frac{1+R_\mathrm{f}/R_1}{3+j(\omega RC - 1/\omega RC)}. \end{aligned} \tag{2.183}$$

Barkhausen's Criterion, $A\beta = 1$, tells us that the imaginary part must be zero and the real part equal to 1. These conditions are met for

$$\begin{aligned} \omega &= 1/RC, \\ R_\mathrm{f} &= 2R_1. \end{aligned} \tag{2.184}$$

See the right side of Figure 2.47. It shows a Nyquist plot, Im($A\beta$) vs. Re($A\beta$), of the behavior of the circuit that is a displaced circle in this case. The effect

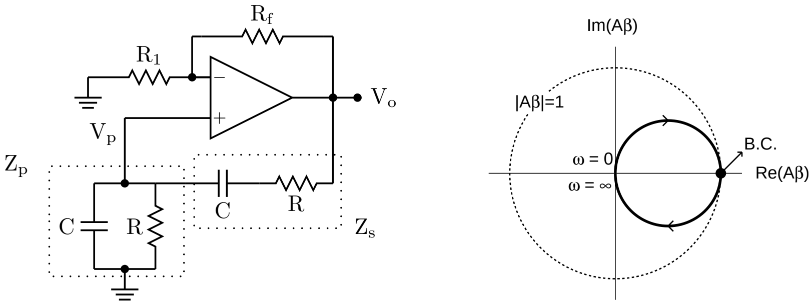


**Fig. 2.47**: Wien bridge oscillator circuit (left) and Nyquist plot (right). The circuit will oscillate at the frequency that meets the Barkhausen Criterion (B.C.), $A\beta(\omega) = 1$, thus $\mathrm{Re}[A\beta(\omega)] = 1$ and $\mathrm{Im}[A\beta(\omega)] = 0$, with $\beta = Z_\mathrm{p}/(Z_\mathrm{p}+Z_\mathrm{s})$ and $A = 1 + R_\mathrm{f}/R_1$

of $R_1$ and $R_\mathrm{f}$ is determining the size of the circle and this degree of freedom allows us to put the curve through the Barkhausen Criterion to make the circuit marginally sustain the oscillation.

### 2.7.3 Phase-shift oscillator

A phase shift oscillator is made by placing three RC pairs in the feedback loop of an opamp configured as an inverting amplifier (page 84) through the feedback loop composed of R and R$_\mathrm{f}$, resulting in a gain of $A = -R_\mathrm{f}/R$. See Figure 2.48. This amplifier is non-ideal in the sense that is has finite input resistance. Since the negative terminal $V_\mathrm{n}$ of the opamp is at virtual ground it is not difficult to show that this input resistance $\mathrm{d}V_\mathrm{i}/\mathrm{d}I_\mathrm{i}$ is equal to $R$. We can thus convert the non-ideal amplifier into an ideal amplifier by adding a resistance R to the feedback loop to result in the effective circuit as in Figure 2.48.

We now see that the feedback loop is composed of exactly three high-pass RC filters, as shown in the section on passive filters (Sec. 2.3). The cut-off frequency of each of these filters is not simply $1/RC$ because each filter RC pair loads the others. The calculation of the overall frequency behavior is quite complex. Yet we can make some global observations and qualitatively predict the behavior.

First of all, because a capacitor is open circuit at low frequencies and short circuit at high frequencies we can say that $\beta$ must be 0 at 0 Hz and 1 at infinite frequency. With $A$ equal to $-R_\mathrm{f}/R$, we find that $A\beta$ is then 0 at 0 Hz and $-R_\mathrm{f}/R$ at $\omega = \infty$. These two points can easily be located on the Nyquits plot. Moreover, each HPF has zero phase at high frequencies and $+90^\mathrm{o}$ at low frequencies as we have seen in Section 2.3. Thus, $A\beta$ has $-180^\mathrm{o}$ at high frequencies (due to the negative sign of A that is equivalent to $-180^\mathrm{o}$) and $-180^\mathrm{o} + 3 \times 90^\mathrm{o} = +90^\mathrm{o}$ at low frequencies. With this we can construct the Nyquist plot schematically: a counterclockwise spiral, spiraling in from $A\beta = -R_\mathrm{f}/R$ to 0. It crosses the line $\mathrm{Re}(A\beta) > 0$ and $\mathrm{Im}(A\beta) = 0$ somewhere

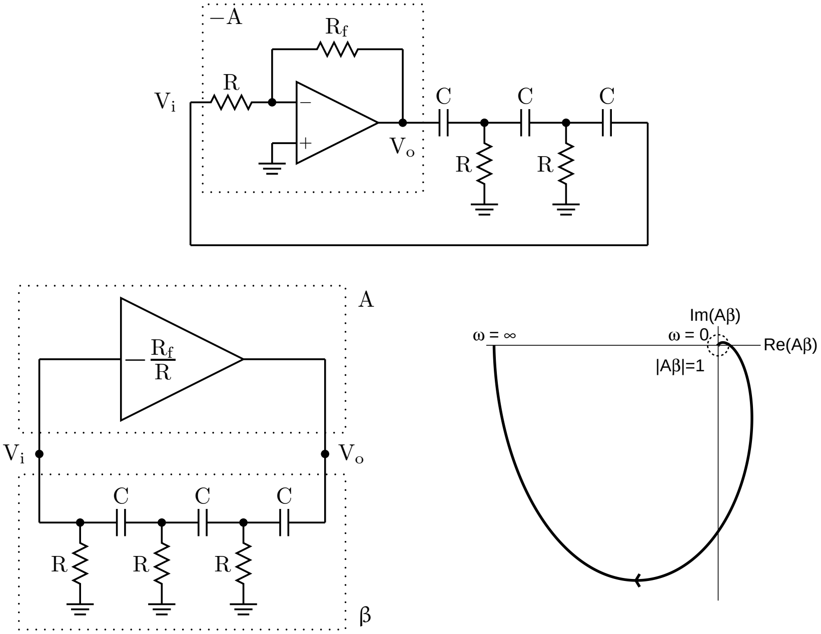


**Fig. 2.48**: Phase-shift oscillator, its equivalent circuit and a Nyquist plot for a well tuned oscillator that includes a frequency where $A\beta = +1$

at higher frequencies. The gain of our inverting amplifier, through changes of $R_\mathrm{f}$ can be tuned to make this crossing point at $\mathrm{Re}(A\beta) = +1$ to induce sustained oscillation. Calculation shows that this occurs for

$$\omega = \frac{1}{RC\sqrt{6}}, \tag{2.185}$$

$$R_\mathrm{f} = 29R. \tag{2.186}$$

### 2.7.4 Quartz-crystal oscillator

An often used oscillator is a quartz crystal. Because of the piezoelectric effect (described in the chapter on physics), a property of quartz, the mechanical oscillations of this crystal can be excited in an electrical way. Thus, the mechanical properties can be described by an electrical equivalent circuit, a so-called Butterworth Van Dyke circuit given in Figure 2.49. In fact, there is no way to distinguish the electrical behavior of the crystal from the behavior of the equivalent circuit and thus — for all electrical purposes — the crystal can be considered equal to the circuit.

The resistor R, capacitor C and inductor L represent the electrical behavior of the mechanical oscillations of the crystal. The parallel capacitance $\mathrm{C_p}$ is formed by the electrodes deposited on the crystal to excite the oscillations via the piezoelectric effect. The impedance of the Butterworth Van Dyke (BVD)

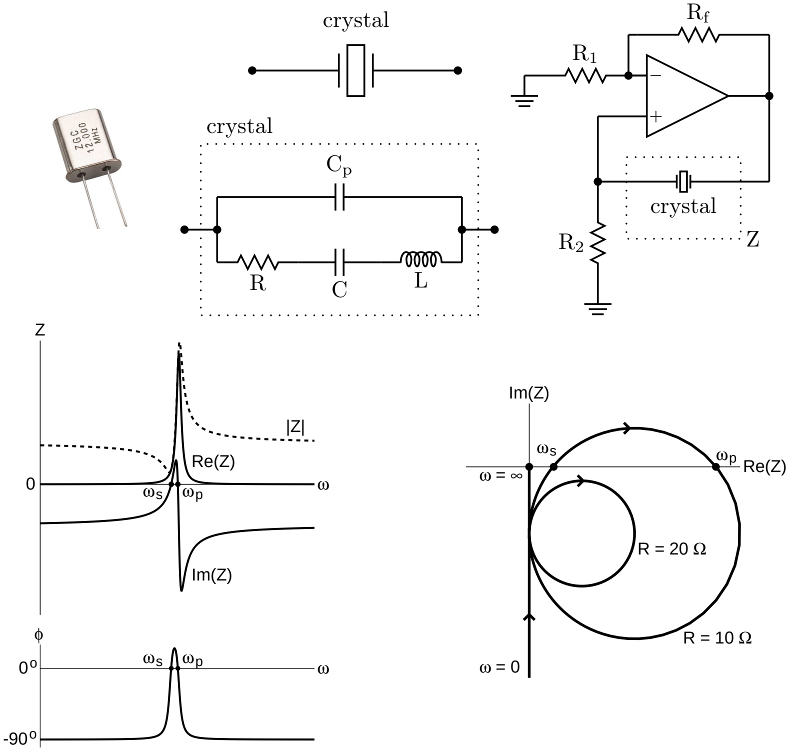


**Fig. 2.49**: Quartz-crystal oscillator. Top left: Quartz crystal component. Top middle: Electronic symbol (top) and Butterworth Van Dyke equivalent circuit (bottom). The resistor R, capacitor C and inductor L represent the electrical behavior of the mechanical oscillations of the crystal. The parallel capacitance $C_p$ is formed by the electrodes deposited on the crystal to excite the oscillations via the piezoelectric effect. Top right: Oscillator circuit used to make the crystal resonate at its fundamental frequency. Bottom Left: Spectrum of the admittance $Z$ for a crystal with $R = 10\ \Omega$, $C = 1$ pF, $L = 1$ mH, $C_p = 1$ nF. Bottom right: Nyquist plot of same crystal. In comparison also the crystal with $R = 20\ \Omega$ is shown that will never oscillate since the necessary condition $\mathrm{Im}(Z) = 0$ is never met

equivalent circuit is given by

$$Z(\omega) = \left( \frac{1}{R + 1/j\omega C + j\omega L} + j\omega C_{\mathrm{p}} \right)^{-1}. \tag{2.187}$$

When the crystal is used in a feedback circuit as shown on the right side of the figure, the crystal can be made to oscillate. To find the frequency of oscillation we have to look at the loop gain. The feedback loop gain is equal to $\beta = R_2/(R_2 + Z)$. Together with the gain of the non-inverting amplifier We get a total loop gain of

$$A\beta = \frac{R_{\mathrm{f}} + R_1}{R_1} \times \frac{R_2}{R_2 + Z}. \tag{2.188}$$

Barkhausen's Criterion tells us that this should be unity. This occurs for a frequency $\omega$ where $\mathrm{Im}[Z(\omega)] = 0$ and $R_{\mathrm{f}}R_2/R_1 = \mathrm{Re}[Z(\omega)]$. The latter means that we can use the amplifier resistances to tune the circuit to resonance. The first condition of vanishing reactance (imaginary part of impedance) occurs for two specific frequencies, $\omega_{\mathrm{s}}$ and $\omega_{\mathrm{p}}$. Vanishing reactance also means vanishing susceptance, which is the imaginary part of admittance, $B = \mathrm{Im}(Y)$, $Y = 1/Z = G + jB$. According to Equation (2.187) this susceptance is given by

$$B = \omega C_{\mathrm{p}} - \frac{(\omega L - 1/\omega C)}{R^2 + (\omega L - 1/\omega C)^2} = 0. \tag{2.189}$$

Standard electronics textbooks tell us that a good approximation is that the series resonance is the resonance frequency of the motional branch, found by assuming $C_{\mathrm{p}} = 0$,

$$\omega_{\mathrm{s}} = \frac{1}{\sqrt{LC}}, \tag{2.190}$$

and the parallel resonance is given by

$$\omega_{\mathrm{p}} = \sqrt{\frac{(C + C_{\mathrm{p}})}{LCC_{\mathrm{p}}}} = \omega_{\mathrm{s}} \times \sqrt{1 + C/C_{\mathrm{p}}}. \tag{2.191}$$

While for electronics purposes this is good enough, for our instrumentation applications this approximation is inadequate when we want to use the crystal as a measurement instrument, for instance an ultra-sensitive quartz-crystal microbalance (see the chapter on physics). In such cases, the electrostatic electrode capacitance $C_{\mathrm{p}}$ has to be compensated. Figure 2.49 shows simulations of the impedance spectrum ($Z$ vs. $\omega$) and Nyquist plot, $\mathrm{Im}(Z)$ vs. $\mathrm{Re}(Z)$, of a BVD circuit with $L = 1$ mH, $C = 1$ pF, $R = 10\ \Omega$ and $C_{\mathrm{p}} = 1$ nF. Note also that if $R$ or $C_{\mathrm{p}}$ gets too large, the condition $B = 0$, and with it $\mathrm{Im}(Z) = 0$, is no longer met for any frequency, and the crystal will not oscillate. The figure also shows a simulation of the Nyquist plot for the same circuit with $R = 20\ \Omega$ that never has a vanishing reactance. These effects can occur if we use the quartz crystal in liquid environments, where serious damping occurs (increasing $R$) or the dielectric changes of the surroundings make $C_{\mathrm{p}}$ larger. Submerging the crystal can cause a loss of resonance.

### 2.7.5 555 timer IC circuits

The 555 is a universal timer (integrated) circuit that can be used for a number of different time-applications, ranging from oscillators to single pulses. Inside the circuit (see Fig. 2.50) there are two operational amplifiers - the reason why it is presented in this section - that are connected, without direct feedback, as simple comparators (AR and AS). The comparing voltages of the two comparators are defined by the three internal equal resistances (of 5 kΩ) in a voltage-divider configuration to give one-third and two-thirds of the external supply voltage (VCC), respectively. The other signals of the comparators come from outside the IC and thus depend on the connections being made by the user. They are called 'threshold' (THR) and 'trigger' (TRI), respectively, the reason for these names we will see in a moment. The output of the comparators are fed to a flip-flop (FF) of the type set-reset. Note that the flip-flop responds to *changes* of the input signals, so-called flanks, and not to (steady) states. To be more precise, the flip-flop reacts to *positive* flanks, changes from low (GND) to high (VCC) voltage. The output of the flip-flop (Q) is also the output of the circuit. The inverted output ($\bar{\mathrm{Q}}$) is connected to the base of a bipolar transistor (Q1), current-protected with a 100 Ω resistance, that works as a switch; if $\bar{\mathrm{Q}}$ is high, the DIS line is effectively shorted to GND. If $\bar{\mathrm{Q}}$ is low, the DIS line is floating. The last connections are RST and CON, standing for 'reset' and 'control', respectively. The former gives us a way to manually reset the flip-flop and the latter gives a way to change the two comparator reference voltages from two-thirds and one-third VCC to any other desirable voltage. See Figure 2.51 for the pin and signal labeling of the 555 circuit in the DIL package. Not always are all pins used. It is advised to not let the unused pins floating where they can work as antenna and pick up high-frequency noise which may disturb the working of the IC. For instance, a bypass capacitor of 10 nF should be placed between the control-voltage pin (CON) and ground, and the reset pin (RST) should be directly connected to the supply voltage when these functionalities are not used.

With this circuit a variety of applications can be made. The three classic ones are the oscillator, the mono-stable (fixed pulse length) and fixed delay circuits.

#### 555 oscillator

The oscillator circuit, presented in the Philips NE555 Application Note (AN170), has two resistances and a capacitor placed in series from the supply voltage to ground. The DIS line is placed in between the resistances and both the TRI and THR lines connected to the capacitor, where we define a point x, with voltage $V_{\mathrm{x}}$, for our analysis. See Figure 2.52 for the circuit and its behavior.

Imagine the output is high, meaning Q is high and $V_{\mathrm{o}} = V_{\mathrm{CC}}$. This means that $\bar{\mathrm{Q}}$ is low and the transistor Q1, used as a switch, is not conducting; effectively the DIS pin is floating.

Effectively we have a simple RC circuit; a current starts charging the ca-

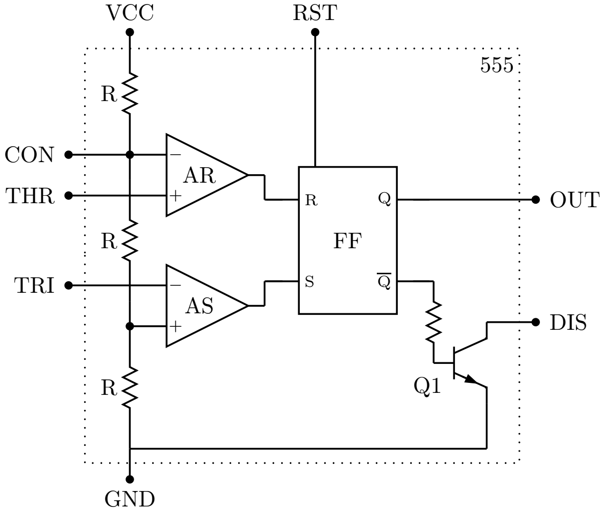
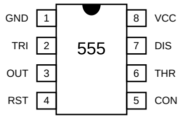


**Fig. 2.50**: 555 General purpose timer (integrated) circuit



| *Pin* | *Mnemonic* | *Function* |
|---|---|---|
| 1 | GND | Ground |
| 2 | TRI | Trigger |
| 3 | OUT | Signal out |
| 4 | RST | (Not) Reset/Enable |
| 5 | CON | Control |
| 6 | THR | Threshold |
| 7 | DIS | Discharge |
| 8 | VCC | Supply voltage |

**Fig. 2.51**: 555 pin connections

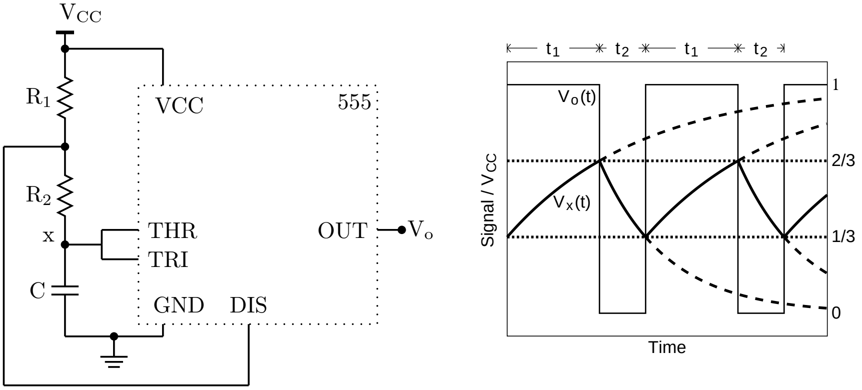


**Fig. 2.52**: Oscillator using a 555 Timer IC (left) and behavior of the circuit for $R_2 = R_1$ (right). The half-cycle times are $t_1 = (R_1 + R_2)C \times \ln(2)$ and $t_2 = R_2C \times \ln(2)$

pacitor C. We define a voltage $V_\mathrm{x}$ in between the resistor and the capacitor. Because the capacitor starts charging, this voltage $V_\mathrm{x}$ rises. Hence the current - given by $I(t) = [V_\mathrm{CC} - V_\mathrm{x}(t)]/(R_1 + R_2)$ - gradually drops. We recognize here a classical relaxation behavior, a voltage that exponentially approximates the supply voltage $V_\mathrm{CC}$ with a relaxation time given by $\tau_1 = (R_1 + R_2)C$. If $V_\mathrm{x}$ rises above one third the supply voltage $V_\mathrm{x} > V_\mathrm{CC}/3$ nothing happens; the flip-flop SET line is un-set, but the flip-flop does not respond to changes high-low. In any case, as we will see in a moment, the starting voltage at $t = 0$ is not zero, but $V_\mathrm{CC}/3$ instead. We thus arrive at an equation for the voltage $V_\mathrm{x}$ as a function of time in the first part of the oscillation period as

$$\begin{aligned} V_\mathrm{x}(t) &= V_\mathrm{x}(t=\infty) + [V_\mathrm{x}(t=0) - V_\mathrm{x}(t=\infty)]\mathrm{e}^{-t/\tau_1} \\ &= \left[1 - \frac{2}{3}\mathrm{e}^{-\frac{t}{(R_1+R_2)C}}\right] V_\mathrm{CC}. \end{aligned} \tag{2.192}$$

If left alone, the $V_\mathrm{x}$ would reach in this way a voltage equal to $V_\mathrm{CC}$. However, it never reaches this voltage. When it rises above two-thirds the supply voltage, $V_\mathrm{x} = (2/3)V_\mathrm{CC}$, the top comparator triggers. This thus happens at a time given by

$$t_1 = (R_1 + R_2)C \times \ln(2). \tag{2.193}$$

At this time the RESET line of the flip-flop is raised and the flip-flop output is set to low, $V_\mathrm{o} = 0$. Simultaneously, the negated output $\bar{\mathrm{Q}}$ is set to high. This opens the transistor Q1 and connects resistance $\mathrm{R}_2$ effectively to ground and the capacitor will start to discharge through this resistor with a time constant given by $\tau_2 = R_2C$. We thus arrive at an equation for the voltage $V_\mathrm{x}$ as a function of time in the second part of the oscillation period as

$$\begin{aligned} V_\mathrm{x}(t) &= V_\mathrm{x}(t=\infty) + [V_\mathrm{x}(t=0) - V_\mathrm{x}(t=\infty)]\mathrm{e}^{-t/\tau_2} \\ &= \frac{2}{3}\mathrm{e}^{-t/R_2C} V_\mathrm{CC}. \end{aligned} \tag{2.194}$$

The time for discharging from two-thirds to one-third the supply voltage is then given by

$$t_2 = R_2C \times \ln(2). \tag{2.195}$$

Combining the two we find a frequency of oscillation given by

$$f = \frac{1}{t_1 + t_2} = \frac{1.44}{(R_1 + 2R_2)C}. \tag{2.196}$$

(It beats me why the Philips Application Note has a factor 1.49 instead). Note that this results in a duty cycle, the percentage of the time the output is high, given by

$$D \equiv \frac{t_1}{t_1 + t_2} = \frac{R_1 + R_2}{R_1 + 2R_2}, \tag{2.197}$$

which is necessarily larger than 50%. This is because the resistance for charging is always larger than the resistance for discharging, while the voltage excursions

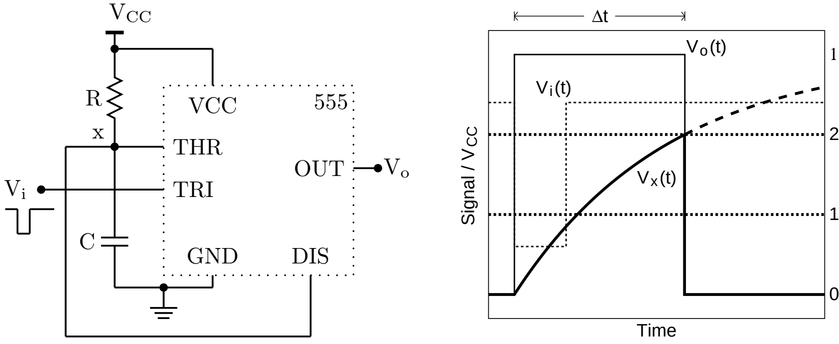


**Fig. 2.53**: Single-pulse circuit using a 555 IC (left) and its behavior (right) after receiving a signal $V_\mathrm{i} < V_\mathrm{CC}/3$. The pulse width is $\Delta t = RC \times \ln(3)$

are the same. To overcome this, a diode can be placed in parallel with resistor $\mathrm{R}_2$ to by-pass it in the charging cycle. This allows for more flexibility in the design of the duty cycle. However, as a trade off we will lose any grasp of the calculation, since ideal diodes do not exist; real didoes close with a voltage drop smaller than about 0.7 V.

Finally, the reset line (RST) can be used to 'gate' the oscillation. With this line high, the oscillation will be in free run, while putting a low voltage on the line will basically stop the working of the IC.

### 555 single-pulse Circuit

Figure 2.53(a) shows the 555 IC used in a mono-stable single-pulse circuit. Its output is low, except for a precisely determined time $\Delta t$ after the input is made low. This way any pulse, with any duration at the entrance is converted into a well determined pulse at the exit.

In steady state, long after receiving the last pulse at the entrance, the output is low. Consequently, the negated output $\bar{\mathrm{Q}}$ is high, the DIS line is effectively connected to ground, and the capacitor is fully discharged; $V_\mathrm{x} = 0$. The input is connected to the TRI input line and compared to a third of the supply voltage at the bottom comparator of the 555 IC. The moment this input voltage drops below $V_\mathrm{CC}/3$ in what is called a 'trigger' - hence the name of the input pin - the first comparator output changes state from low to high, i.e. a SET signal, and the flip-flop output state is changed from low to high. This open-circuits the transistor and the DIS line is effectively disconnected. The capacitor now starts charging with a time-constant given by $\tau = RC$, with a tendency to saturate at $V_\mathrm{CC}$,

$$V_\mathrm{x}(t) = \left[1 - \mathrm{e}^{-t/RC}\right] V_\mathrm{CC}. \tag{2.198}$$

When one-third of the supply voltage is reached, the bottom comparator output switches to low, but the flip-flop does not respond to changes low-high, so

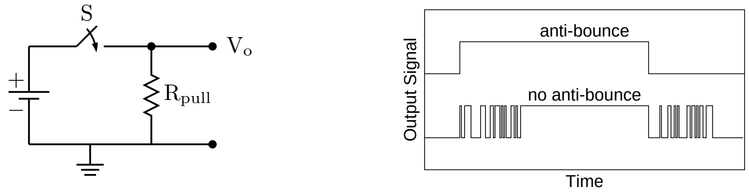


**Fig. 2.54**: Simple circuit of a switch (with pull-down resistor $R_{\text{pull}}$ to avoid 'floating' output if the switch is open). The result is a 'bouncing' of the output signal, as shown in the simulation on the right. Fake multiple transitions occur. To avoid this anti-bouncing circuits can be used which result in clean output signals with single transitions. Ideally the response is as on the top of the figure, an immediate, single switch after the first transition.

nothing happens. The moment two-thirds of the supply voltage are reached, the top comparator changes from low to high and the flip-flop is reset. The output is made low and the inverted output becomes high again. This opens the transistor, shorting the capacitor to ground and instantaneously discharging it. $V_{\text{x}} \to 0$. (The output of the top comparator switches from high to low but, once again, the flip-flop does not respond to this). This happens at a time $\Delta t$ after the start of the input pulse given by

$$\Delta t = RC \times \ln(3), \tag{2.199}$$

approximately 1.10 times the RC time. See Fig. 2.53(b) for the behavior of the circuit.

An application for this circuit is for instance the flushing of the water in a urinal for 5 seconds after receiving a signal. Note that new signals received within the period of the output pulse will *not* extend the pulse length. Somebody joining in for a piss will not change the moment the water will be flushed, nor the amount of flushing time. A new pulse will possibly again trigger a change of the output state of the bottom comparator from low to high, setting the flip-flop ... which was already set anyway. This circuit thus has a 'dead time' incorporated, a time in which input signals will do nothing to the operation of the circuit. A sequence, once set in motion, will take its course in exactly the programmed amount of time.

An important use of dead time is anti-bouncing of a switch. Ideally, a switch is either on or off and transitions occur cleanly, changing immediately and once from one state to the other. In practice, any switch will have multiple transitions before settling at the new state, see Figure 2.54 for an example. While the user has pressed the button only once, these fake transitions make it appear the button was pressed many times. While with informatics it is very easy to eliminate these fake transitions - we just wait a certain time before allowing any new input changes - with electronics it is more difficult.

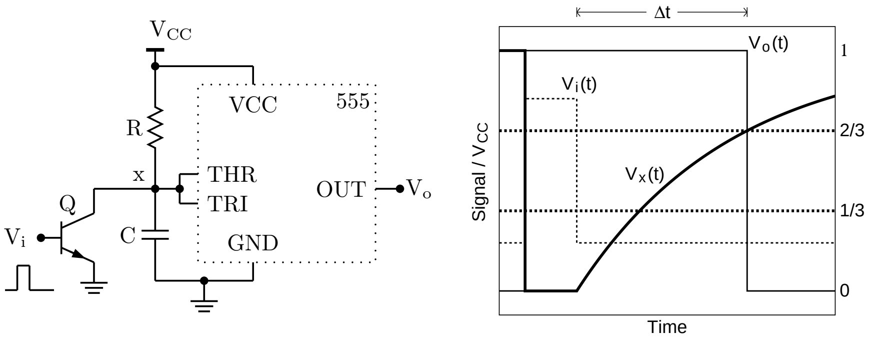


**Fig. 2.55**: Delay pulse circuit using a 555 IC (left) and its behavior (right) after receiving a signal large enough to open the transistor Q. The delay time is $\Delta t = RC \times \ln(3)$

### 555 delay circuit

The third 'classic' circuit with a 555 Timer IC is the programmed delay, an example of which is shown in Figure 2.55. An RC pair is used with the halfway point connected to both comparators, THR and TRI. An external transistor (Q) is configured as a switch; with low input signals the transistors is open-circuit, while with high input signals it is effectively a short circuit between collector and emitter. In steady state, the capacitor is fully charged, $V_x = V_{CC}$, the flip-flop is reset and the output low, $V_o = 0$. The input signal pulse switches a transistor from the low to high-conductive state. Effectively the capacitor is shorted to ground and fully discharged instantaneously, $V_x = 0$. This triggers the output state of the bottom comparator and a flip-flop SET signal is generated; the output of the 555 IC switches from low to high, $V_o = V_{CC}$. As long as the input signal is high, the transistor is a short circuit and the capacitor continues to be empty. When the external pulse stops, the transistor is open-circuit and the capacitor starts charging with a time constant $\tau = RC$ and a tendency towards $V_{CC}$. It rises above one-third $V_{CC}$, where, just like in the pulse circuit before, nothing happens because the bottom comparator output changes (from high to low), but the flip-flop does not respond to this type of change. Only when the voltage at x rises above two-thirds $V_{CC}$ does something happen. At that moment, given by

$$\Delta t = RC \times \ln(3) \tag{2.200}$$

after the beginning of the charging of the capacitor, does the top comparator output switch from low to high and is the flip-flop reset, $V_o \to 0$. See Figure 2.55 for the transient behavior of the circuit. Note that when a new high input signal is received before the end of the charging cycle is reached, the charging starts from scratch again. In other words, there is no 'dead-time'.

An important difference between this delay circuit and the single-pulse circuit given before is that in this delay circuit a pulse is generated a certain time

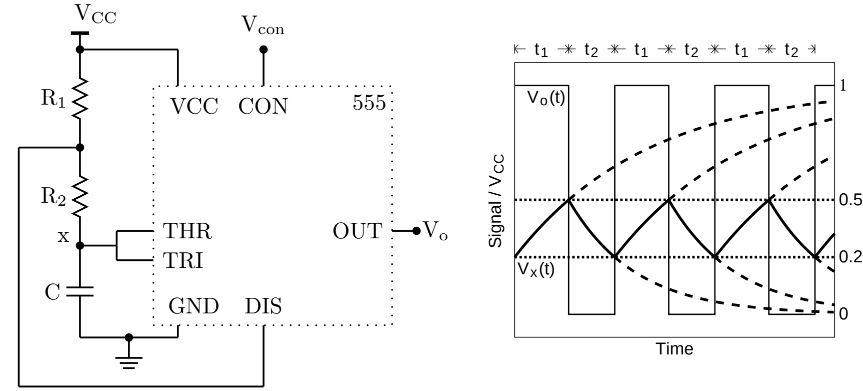


**Fig. 2.56**: Oscillator circuit as in Fig. 2.52, but with the reference voltages of the two 555-comparators externally defined by a control voltage $V_{\mathrm{con}}$ connected to the CON pin. The right side shows the behavior of the circuit for $R_1 = R_2$ and $V_{\mathrm{con}} = 0.5V_{\mathrm{CC}}$. For lower voltages the frequency increases because the first half of the cycle, $t_1$, shortens while the second half remains the same

after the *end* of the input pulse, where in the circuit before it is after the *beginning* of the input pulse. This makes the circuit even more adequate for urinal applications. Water will be flushed a certain time after we finish our business and not while we're in the middle of it. Note that the output of the circuit stays low (until a new trigger signal is received), an undesired side effect if we indeed want to use it for timing events as described here. However, we can connect the output of such a circuit as the input of a pulse-generating circuit as described before. For this purpose there exists the handy 556 IC, which consists of two 555 circuits integrated into one IC. In this way we can design a circuit that will have a high output for a certain amount of time and beginning a precise time after a certain event. Other applications of the 556 dual timer IC is the tone-burst generator, the first 555 generates a pulse of determined width, while the second 555 uses this as the gate (RST) of an oscillator. This finds applications in things like tone dialing for telephones.

### 555 use of the control voltage

The control signal can be used to change the timing of the circuits. A voltage applied to the CON pin 'overrides' the reference voltage of *both* comparators, instead of $2V_{\mathrm{CC}}/3$ and $V_{\mathrm{CC}}/3$ become $V_{\mathrm{con}}$ and $V_{\mathrm{con}}/2$ respectively. While the external circuit remains the same, and charging and discharging kinetics of the capacitor the same, the reference levels change and the relevant times with it. Take for example the oscillator of Figure 2.52. If we define the control voltage by an external source, as shown in Figure 2.56, the timing changes.

In the charging cycle, the voltage has to rise from $V_{\mathrm{con}}/2$ to $V_{\mathrm{con}}$. With a time constant, as before, $\tau_1 = (R_1 + R_2)C$ and a tendency to reach $V_{\mathrm{CC}}$. It is easy to show that this results in a first-half period of

$$t_1 = (R_1 + R_2)C \times \ln\left(\frac{V_{\mathrm{CC}} - V_{\mathrm{con}}/2}{V_{\mathrm{CC}} - V_{\mathrm{con}}}\right). \tag{2.201}$$

Likewise, the second-half period, the time it takes to discharge the capacitor from a voltage $V_{\mathrm{con}}$ to $V_{\mathrm{con}}/2$ through the resistor $R_2$ ($\tau_2 = R_2C$) and with a tendency towards 0, is given by

$$t_2 = R_2C \times \ln(2), \tag{2.202}$$

which is equal to the 'uncontrolled' version of the oscillator, which makes sense, since it still has to drop to half its starting value and a property of exponential decay is that the time it takes to drop to a certain fraction is independent of the starting value.

We thus get an oscillation frequency given by

$$f(V_{\mathrm{con}}) = \frac{1}{t_1 + t_2} = \left[(R_1 + R_2)C \times \ln\left(\frac{V_{\mathrm{CC}} - V_{\mathrm{con}}/2}{V_{\mathrm{CC}} - V_{\mathrm{con}}}\right) + R_2C \times \ln(2)\right]^{-1}. \tag{2.203}$$

In other words, the oscillation frequency is controlled by the external voltage. While technically speaking this is thus a voltage-controlled oscillator (VCO), in practice it is a very bad actuator $V \to f$, because of its non-linearity. As we have seen in the opening chapter, linearity is an important parameter of sensors and actuators. A much better VCO is the one given in Exercise 17, or the 231 and 331 integrated circuits. Or alternatively the VCO embedded in the 4046 PLL (phase-locked loop) IC.

In a similar way we can make a circuit that uses pulse-width modulation (PWM) techniques, by generating pulses with the mono stable circuit whose width depends on the control voltage. Once again, the pulse width will not be a nice linear function of the input bias $V_{\mathrm{con}}$ and the usefulness of the circuit is limited.

Finally, don't forget that the input resistance of the CON pin is $R_{\mathrm{in}} = R \parallel 2R$, and with the resistance in a standard 555 circuit equal to 5 kΩ, this is typically about 3.3 kΩ. This in case you want to apply a voltage source with non-zero output resistance (like a resistive network such as a voltage divider).

## 2.8 Signal processing; noise

### 2.8.1 Noise

Noise is in general all signal that contains no useful information. As such, noise is unwanted. In fact, the most important noise parameter is not so much the level of the noise, but the level of the signal relative to the noise, the so-called signal-to-noise ratio, S/N. The official definition is the power contained in the signal divided by the power contained in the noise.

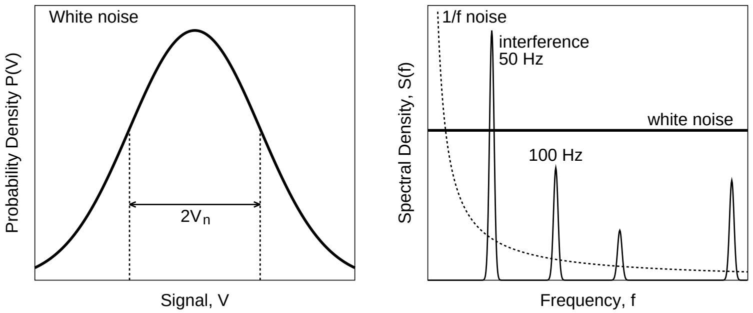


**Fig. 2.57**: Noise signal probability density and noise frequency density spectrum

In previous sections we found out how to increase the signal. In this section we talk about noise. There are several ways of eliminating noise and increasing the S/N. The simplest is by filtering it off, for instance with the filters described before (Section 2.3). More advanced filters are the Butterworth, Bessel en Chebyshev filters, among others. (See the book of Horowitz and Hill for an overview). Some types of noise can be eliminated at the source, others not. It is useful to classify the sources of noise, since once we know where the noise is coming from, it is easier to eliminate it.

### 2.8.2 Types of noise

**Johnson noise** is caused by thermal random effects in resistances and as such is also often called thermal noise. It can be compared to the random motion of dust particles in Brownian motion. Likewise, Johnson noise is caused by the random movement of charged particles in an electronic element. This random motion causes noise at the extremes of the element. It is thus an open-circuit, zero-current voltage source with a mean voltage amplitude (root-mean-square) given by

$$\text{Johnson noise: } V_{\text{rms}} = \sqrt{4kTR\Delta f}, \tag{2.204}$$

with $T$ the temperature, $k$ Boltzmann constant, $R$ the resistance, and $\Delta f$ the bandwidth. It is an unavoidable noise, meaning that a device with a certain resistance, working at a certain temperature and operating at a certain bandwidth will produce (at least) this noise. The exact value of the voltage is unpredictable (that is why it is noise), but has a Gaussian probability distribution, the probability $P$ to measure a voltage $V$ at the terminals is given by

$$P(V) = \frac{1}{V_{\text{rms}}\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{V}{V_{\text{rms}}}\right)^2}. \tag{2.205}$$

with $V_{\text{rms}}$ the standard-deviation ('root-mean-square') of the voltage. See Figure 2.57. It is a type of 'white noise', meaning that the spectral density is uniform, and that any two measurements are completely uncorrelated.

A second type of noise is **shot noise**. It occurs because of the quantization of current. As we know, current is the transport of charge and since charge is quantized (the elementary charge is $q = 1.6 \times 10^{-19}$ coulomb; all charges are integer multiples of this elementary charge), so is current. The random fluctuations of the number of charges causes a random fluctuation of current that is given by

$$\text{Shot noise: } I_{\text{rms}} = \sqrt{2qI\Delta f}, \tag{2.206}$$

with, as before, $\Delta f$ the bandwidth, $q$ the elementary charge and $I$ the magnitude of the current. It is obvious from the above equation that shot noise becomes more important for small currents; where the signal rises proportional to the current, the noise grows only with the square root of current, the important S/N parameters thus decreases rapidly for increasing currents, S/N $\propto \sqrt{I}$. This makes sense when we look at it from a mathematical point of view: For small currents, the number of charges involved is small and the random fluctuation of small numbers is relatively large; a stochastic process with small numbers has a large spread, there where the law of large numbers tells us that the average of a large number of processes will tend to its expectation value with a relatively small uncertainty.

Like Johnson's Noise, shot noise is a type of white noise, with a constant spectral density, and a probability function given by Eq. 2.205 (with $V$ replaced by $I$), as shown in Figure 2.57.

**1/f noise** is a frequently recurring type of noise that is named after its spectral distribution; it drops off as $1/f$ for increasing frequency (See Fig. 2.57). Other names for this types of noise are 'pink noise' (because of its higher density at lower frequencies like in the redder part of the optical spectrum) or 'flicker noise'. Not always is the spectral density following $1/f$. In general can be said that any noise spectrum having a spectral density proportional to $1/f^{\alpha}$ can be classified as 1/f noise, as long as the spectral density drops off with frequency. A special case is Brown noise, not named for its color, but named after random Brownian motion, which has a spectrum proportional to $1/f^2$.
Seen from the other point of view, for low frequencies this noise becomes prohibitive. Combined with the fact that 1/f noise is ubiquitous - annoying and unavoidable - this noise forces nearly every system to avoid measuring at DC (at 0 Hz the noise is infinite). We all know the side-effects of this 1/f noise. Even our human sensors (eyes and ears) are not sensitive for steady-state signals. We can very well distinguish differences in images - as in movement of objects - but have difficulty recognizing steady patterns. Police cars modulate their sirens for the sole purpose to avoid signals at low frequencies. In summary, 1/f noise makes measuring at low frequencies difficult. The answer normally is: modulation. Just like radio signals are modulated to avoid the impossible-to-overcome low S/N ratio of transmittance at DC through the atmosphere, any signal can be modulated to translate it away from the noise zones of the noise spectrum. Later we will see how this is done, when the lock-in amplifier (also known as phase-sensitive detection) is described.

The final type of noise described in this section is **interference**. Interference is in principle all sources of 'noise' that have their own fingerprint and their own origin. All unwanted parts of the spectrum caused by other equipment, etc. Notorious are the 50 Hz 'hum' of the electricity net (depending on where you live, it can be 60 Hz). Also its harmonics can be important. The second harmonic (first overtone) of the 50 Hz base frequency, i.e. at 100 Hz, can easily be found back in the blinking of luminescent light tubes. But not only the electricity net is a source o interference. Other sources may include your neighbor scientist modulating his magnetic field at 1.3 kHz, a vacuum pump rotating at 253 Hz in a nearby building, a car or truck with idling engine (and vibrating our experimental setup). The list of possibilities is sheer infinite.

### 2.8.3 Signal-to-noise ratio

Engineers (and scientists alike) when designing a measurement set up are always talking about "the signal-to-noise ratio". This, obviously, is the ratio of wanted information, the "signal" to the unwanted information, the "noise", defined in terms of power or voltage, it is (remember $P = V \times I = V^2/R$)

$$\text{SNR or S/N} = \frac{P_\text{s}}{P_\text{n}} = \frac{V_\text{s}^2}{V_\text{n}^2}. \tag{2.207}$$

Or in decibels,

$$\text{SNR}_\text{dB} = 10 \times \log\left(\frac{P_\text{s}}{P_\text{n}}\right) = 20 \times \log\left(\frac{V_\text{s}}{V_\text{n}}\right). \tag{2.208}$$

The SNR can also be defined in the mean value of the measured quantity relative to the variation of the quantity,

$$\text{SNR} = \left(\frac{V_\text{s}}{V_\text{rms}}\right)^2. \tag{2.209}$$

Generally speaking, as a rule of thumb, a signal is measurable if the S/N is greater than 2. Optimization of a measurement system consists of increasing the signal and reducing the noise. The latter can for instance consist of narrowing the bandwidth of a filter, but as a side effect this will then also limit the speed at which the signal can change (the information of *changes* of the signal appear in the sidebands of the spectrum and should not be filtered off!)

### 2.8.4 Eliminating noise

Knowing the source of the noise also gives us handles for reducing it. The resistance of the circuit is, in principle, a given parameter that cannot be altered. That leaves us with reducing the temperature and limiting the bandwidth when Johnson noise is giving us trouble. Lowering the temperature is not very easy, but the high-performance low-noise amplifiers are indeed cooled down, for instance to liquid-nitrogen temperatures. These can be found especially in scientific laboratories, where the most advanced instrumentation can be found.

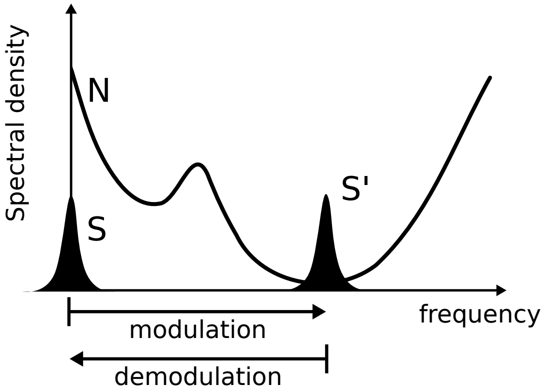


**Fig. 2.58**: Modulation and demodulation of a signal S to bring it to a noise-calm zone

Limiting the bandwidth is easy, see the sections on passive and active filters, Sections 2.3 and 2.6.9, respectively. However, a trade-off of filtering is that, whereas it blocks part of the noise, it also limits the response time of the circuit. As an example, a low-pass RC filter with a cut-off frequency given by $f_c = 1/2\pi RC$ will have a response time given by $\tau = RC$. Output signals cannot change faster than that. Any signals that change faster will be filtered off. Care has to be taken to not filter too much. Apart from that, a delay equal to $\Delta t = RC$ is introduced in the signal. When we want to make fast decisions and take fast action to signal changes, filtering becomes prohibitive. Think of airbags in cars. We want it to open before the head is smashed on the steering wheel.

### 2.8.5 Modulation; Lock-in detector

A powerful technique of reducing noise is modulation and demodulation of the signal. The modulation translates ('modulate') the signal to a zone in the spectrum with little noise, see Figure 2.58. It then amplifies the signal and demodulates it back to DC.

An example is the powerful lock-in amplification (LIA), also known as phase-sensitive detection (PSD). A lock-in amplifier translates ('modulate') the signal to a zone in the spectrum with little noise. It then amplifies the signal and demodulates it back to DC. It consists of applying a small AC voltage and determining the in-phase and 90°-out-of-phase current amplitudes, $I_{\mathrm{ac}}^{0}$ and $I_{\mathrm{ac}}^{90}$, respectively. Figure 2.59 shows this schematically. A sine-wave voltage, based on the internal reference of the lock-in detector is applied to the sample (for simplicity, the bias is not shown). The response current is converted to a voltage signal ("S") and multiplied with the square wave reference ("R"). In case the current is in-phase with the applied voltage, the resulting signal ("SxR") is a rectified signal with a DC component ('offset'), and components at $2\omega$, 4ω, etc. The low-pass filter, consisting of a simple RC circuit, removes the frequency components and passes only the DC part with its slow variations (ideally, everything below ω). With the help of Fourier analysis we can calculate the output signal of the lock-in detector. For an in-phase signal (conductance) with amplitude $A$:

$$\begin{aligned}
S(t) &= S\sin(\omega t), && (2.210)\\
R(t) &= \frac{4}{\pi}\sum_{\mathrm{n=1,3,5...}}^{\infty}\frac{1}{n}\sin(n\omega t), && (2.211)\\
S(t)\times R(t) &= \frac{4S}{\pi}\sum_{\mathrm{n=1,3,5...}}^{\infty}\frac{1}{n}\sin(n\omega t)\sin(\omega t) && (2.212)\\
&= \frac{4S}{\pi}\sum_{\mathrm{n=1,3,5...}}^{\infty}\frac{1}{2n}\left\{\cos\left([n-1]\,\omega t\right)+\cos\left([n+1]\,\omega t\right)\right\} && (2.213)\\
X(t) &= \frac{2S}{\pi}. && (2.214)
\end{aligned}$$

For an out-of-phase signal (susceptance):

$$\begin{aligned}
S(t) &= S\cos(\omega t), && (2.215)\\
R(t) &= \frac{4}{\pi}\sum_{\mathrm{n=1,3,5...}}^{\infty}\frac{1}{n}\sin(n\omega t), && (2.216)\\
S(t)\times R(t) &= \frac{4S}{\pi}\sum_{\mathrm{n=1,3,5...}}^{\infty}\frac{1}{n}\cos(n\omega t)\sin(\omega t) && (2.217)\\
&= \frac{4S}{\pi}\sum_{\mathrm{n=1,3,5...}}^{\infty}\frac{1}{2n}\left\{\sin\left([n-1]\,\omega t\right)+\sin\left([n+1]\,\omega t\right)\right\} && (2.218)\\
X(t) &= 0. && (2.219)
\end{aligned}$$

Thus, the output signal "X" of the lock-in detector is proportional to the amplitude of the in-phase component. The out-of-phase component is completely filtered off (by adequate selection of the cut-off frequency of the low-pass filter). By using a phase shift of 90° either for the sine wave before it is applied to the sample, or to the reference signal used in the multiplier, the system can be tuned for the 90° out-of-phase signal. Modern commercial (two-channel) lock-in detectors often have the possibility to measure both in-phase and out-of-phase signals simultaneously. As discussed earlier, the in-phase signal corresponds to conductance and the out-of-phase signal corresponds to susceptance.

A simplified version of this technique is by measuring at a certain frequency where there is little noise. Apply an AC signal of this frequency, band-pass filter the response signal, rectify it and low-pass filter the result, see Figure 2.61. Obviously it is then no longer phase sensitive, but the electronics are much cheaper.

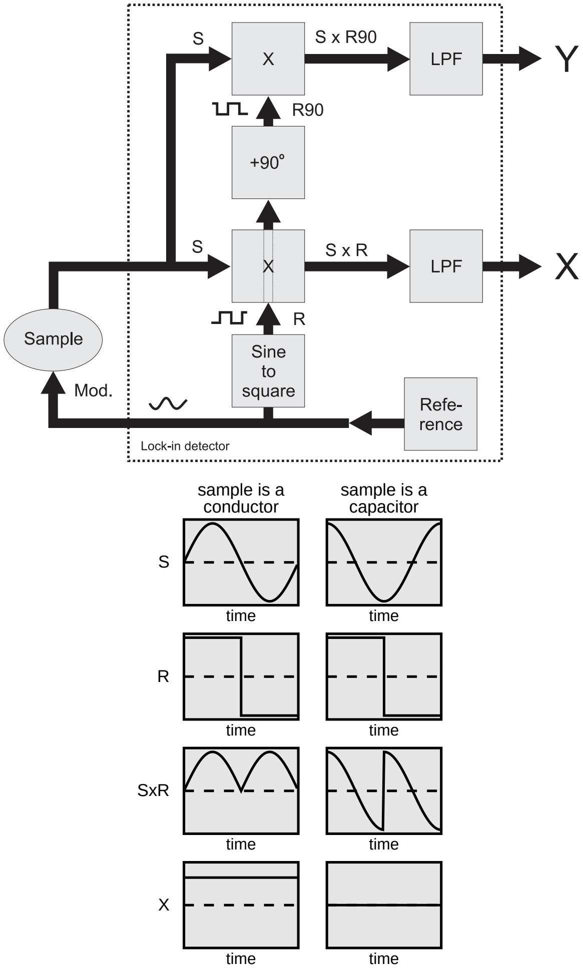


**Fig. 2.59**: Schematic showing how phase-sensitive detection can be used to build an RCL bridge. The reference is used to modulate the signal, which has in-phase and out-of-phase components. When the signal is multiplied by the square-wave reference ("R") and (low-pass) filtered, the output signal ("X") is proportional to the amplitude of the in-phase component only, as demonstrated in the bottom part of the figure. The in-phase part can be associated to conductance (and resistance), while the out-of-phase part (right) represents the susceptance (and capacitance). In a two-channel lock-in detector, as shown, the out-of-phase signal amplitude ("Y") can be found by using a 90$^{\text{o}}$ phase-shifted reference signal. Thus, $X$ is proportional to $G$ and $Y$ is proportional to $B$, the latter which, in turn, can be converted to capacitance (From: Stallinga, "Electrical characterization of organic electronic materials and devices")

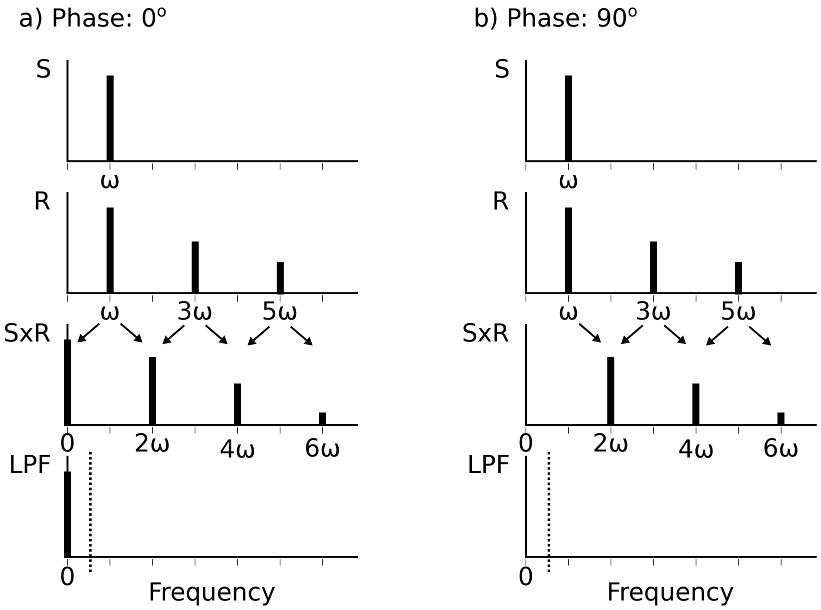


**Fig. 2.60**: Phase-sensitive detection (PSD). a) An in-phase signal S with frequency $\omega$, when multiplied with the square wave of the reference R, results in components at DC, $2\omega$, $4\omega$, etc. The low-pass filter (LPF) filters off everything but the DC component. An out-of-phase signal results in components only at $2\omega$, $4\omega$, etc. The low-pass filter (LPF) filters off everything

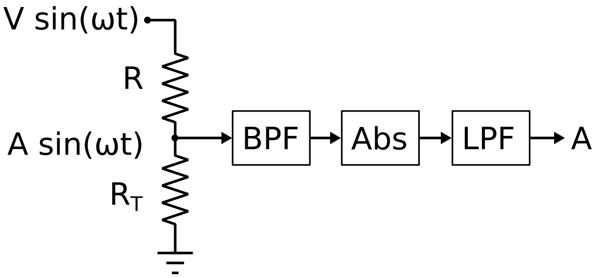


**Fig. 2.61**: Simple modulation technique. A sine wave is applied to the voltage divider containing the temperature sensor $R_T$. The signal is band-pass filtered, rectified and low-pass filtered to give the amplitude of the voltage-divider signal

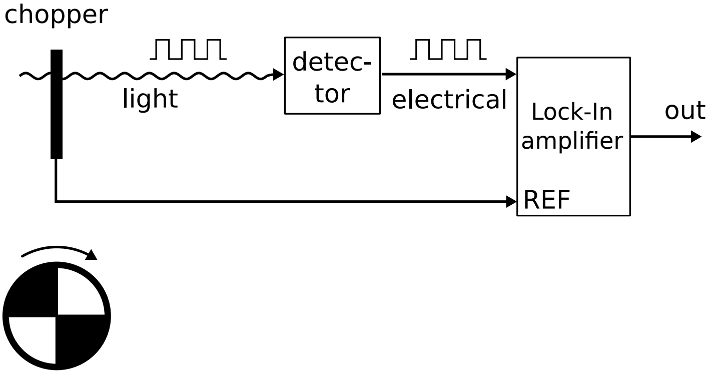


**Fig. 2.62**: A light-beam chopper used to modulate the optical signal

Modulation is not necessarily done electrically. A common technique is to use a 'chopper' in optical measurements. A chopper is a mechanical device and consists of a rapidly rotating disk with holes. This will modulate the light beam and the electrical signal of the detector.

### 2.8.6 Noise and cables

Noise often enters the signal via the cables. They can work as antennas for all the electromagnetic pollution in the air. This is especially important for radio-frequency noise (MHz-GHz), see Figure 2.63a. The easiest way to prevent this is to shield the noise by building a Faraday cage around the cable, as in coaxial cables (Fig. 2.63b), also sometimes called BNC for the type of connector normally used for the cable (Bayonet Neill-Concelman). This blocks (mostly) the electric part of the electromagnetic waves. On the inside, on basis of the magnetic part that is led through, the EM radiation is reconstructed, albeit attenuated. The Faraday cage does not block efficiently EM radiation, as can be tested by placing a mobile telephone inside a can. It often continues to ring.

Especially wires in a loop (but not only) are sensitive to the magnetic half of the electromagnetic waves (Fig. 2.63c). A loop has a large inductance, thus catching a lot of magnetic flux,

$$\Phi = AB, \tag{2.220}$$

with $A$ the area of the loop and $B$ the magnetic field perpendicular to this area. When the flux is changing, as in EM radiation, a current in the cable is induced that is equal to

$$I =\propto \frac{\mathrm{d}\Phi}{\mathrm{d}t}. \tag{2.221}$$

This can either be a change of magnetic field, for instance $B = B_0 \sin(\omega t)$, or a change of area, when the cables are moved,

$$I \propto A\frac{dB}{\mathrm{d}t} + B\frac{\mathrm{d}A}{\mathrm{d}t}. \tag{2.222}$$

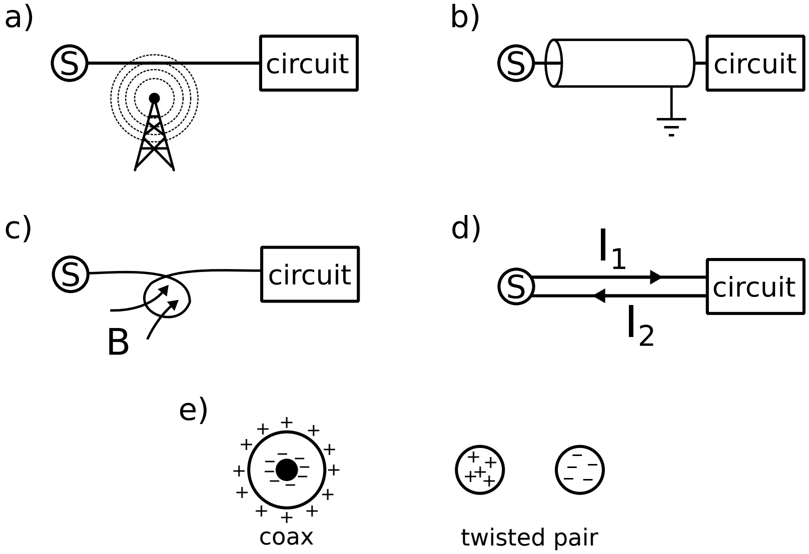


**Fig. 2.63**: a) Unshielded single-wire cables catch a lot of EM radiation. b) A coax cable shields (mostly) the electrical part. c) Wire loops catch magnetic part. d) Twisted pair differential signals to minimize noise (one wire bearing positive signal, the other negative signal). "S" represents a signal source

A simple solution is to use two wires (twisted pair), with the signal passed in positive current $I_\mathrm{s}$ in one wire and negative current $-I_\mathrm{s}$ in the other (Fig. 2.63d). The circuit processes the *difference* current. Any noise caught by the wire $\delta I$ is also picked up equally by the other cable and in the difference the noise is eliminated,

$$\left.\begin{array}{l} I_1 = I_\mathrm{s} + \delta I \\ I_2 = -I_\mathrm{s} + \delta I \end{array}\right\} I_1 - I_2 = 2I_\mathrm{s}, \tag{2.223}$$

no noise! Another advantage of differential currents is that the total magnetic field being built-up around the cable is zero, because the total current of the two wires is always zero. This reduces the losses of the cable.

Both types of cable (coax and twisted pair) have inherent capacitance (Fig. 2.63e). A 50 Ω coax cable has about 100 pF/m capacitance, while a twisted pair has about 60 pF/m. This is caused by the fact that any two metallic objects can store charge and work as capacitors. It can only be reduced by increasing the distance between the objects, or by making the wires very thin. That is also why twisted pair cables work slightly better, because the distance between the poles is somewhat larger. When this capacitance is a problem, single-pole cables should be used, at the cost of increased noise. The capacitance, in combination with the total resistance of the circuit causes a cut-off frequency given by $f_\mathrm{c} = 1/RC$.

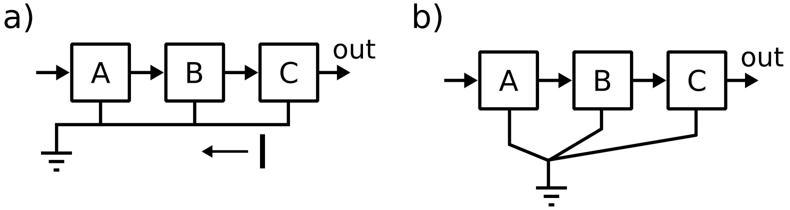


**Fig. 2.64**: a) Wrong grounding. Ground current from C combined with resistance of ground wire A-B causes a rise in the zero voltage of B. For this equipment this is indistinguishable from a signal coming from A. b) Correct star grounding. Equipment is independent

### 2.8.7 Noise and ground

Some forms of noise and offset are caused by (wrong) grounding of the equipment. The cable/wire used for grounding ideally has zero resistance, and all grounds have the same voltage. In reality, not all grounds are equal, and not all ground wires are of low enough resistance. This causes that when there is a current flowing to ground, this current causes a voltage drop in the wire and the equipment feels this as an offset. When the currents are coming from a noise source, these are then equivalent to a voltage source at the input of the equipment. Small as it may be, for tiny signals, this can become important.

A situation to avoid is one in which the equipment are ground 'in series', see Figure 2.64a. The current flowing to ground in one device will by the voltage drop induced in the ground wire create a voltage offset in the other equipment. To avoid this, star grounding has to be used, with each piece of equipment having its own path to a common ground.

### 2.8.8 Digital filtering

A simple way of eliminating noise - or better to say: increasing the signal to noise ratio - is with computers. With a disadvantage of necessarily needing expensive informatics equipment, it is a fast and reliable way of dealing with S/N issues. The easiest way is by averaging. Figure 2.65 shows an example of averaging entire measurement sets. In this example, the signal (a Gaussian peak of 1 V amplitude) is swamped in white noise, Eq. (2.205), with an amplitude of $V_\mathrm{n} = 1$ V. Somewhat arbitrarily this is assigned a S/N ratio of 1, which represents the S/N voltage ratio of a single point of the measurement, namely the maximum. The measured signal $Y$ is a superposition of signal $S$ and noise $N$:

$$Y(t) = S(t) + N(t). \tag{2.224}$$

Averaging implies repeating the measurement $Y(t)$, summing the individual measurements and dividing by $n$, the number of repetitions. By doing so, the signal strength adds up and is proportional to the number of measurements, while the noise amplitude increases only by a factor equal to the square-root of the number of measurements. This is evident when we look in detail at the noise. The sum of $n$ random numbers drawn from the normal distribution of white noise, Eq. (2.205), is again normally distributed, but with a standard deviation equal to the original standard deviation multiplied by $\sqrt{n}$. In other words, the amplitude of the sum *signal* is proportional to $n$, while the amplitude of the sum *noise* is proportional to $\sqrt{n}$. The ratio of the two, the S/N ratio, thus grows with the square root of $n$ (linear with $n$ if we define the S/N ratio in terms of power).

The example was constructed by generating white noise on basis of uniformly distributed random numbers in the interval 0 .. 1 converted to normally distributed numbers as described in Chapter 1, Eq. (1.23). This noise was added to a $m = 600$-point signal running from $t = 0$ s to $t = 10$ s containing a 1 volt amplitude Gaussian peak centered at $t = 5$ s and with a width (from peak to $e^{-0.5}$ V) of 1 s. In this case, the signal power and noise power are given by

$$S = \frac{1}{m}\sum_{\mathrm{i}=1}^{m} S_{\mathrm{i}}^2, \tag{2.225}$$

$$N = \frac{1}{m}\sum_{\mathrm{i}=1}^{m} N_{\mathrm{i}}^2 = \frac{1}{m}\sum_{\mathrm{i}=1}^{m}(Y_{\mathrm{i}} - S_{\mathrm{i}})^2, \tag{2.226}$$

and this gives a (power) S/N equal to 0.166, 0.379, 1.791 and 18.946 for averaging equal to 1, 2, 10 and 100 respectively, or an (amplitude) S/N equal to 0.408 ($n = 1$), 0.615 ($n = 2$), 1.338 ($n = 10$) and 4.353 ($n = 100$), respectively.

We see that in the original measurements $n = 1$ the signal is barely visible. Even averaging by a factor 2 will not help much and we can still not distinguish a peak in the signal. Only for substantial averaging does the signal get lifted out of the noise. Generally speaking, as a rule-of-thumb it can be said that a S/N of about 2 is needed before a signal can be recognized. This definition is rather arbitrary, though.

Another digital filtering technique is moving averaging. Instead of repeating the measurement $n$ times (single points or entire scans) and taking the average, which reduces the number of data points by a factor $n$, we can also take the average of the last $n$ data points and attribute each average to the instant measurement value. Every time a new data point comes in, the oldest data point is removed from the average, as if a window moves over the data points. This is called moving averaging. The amount of data points is not reduced by this technique.

The effect of averaging is the same as low-pass filtering.

As with electronic filtering, computational filtering has a side effect that an artificial delay is introduced in the response.

## 2.9 Lab projects

- Inside two identical boxes, labeled "A" and "B", are simple electronic circuits composed of a resistor and a capacitor. In one box these are

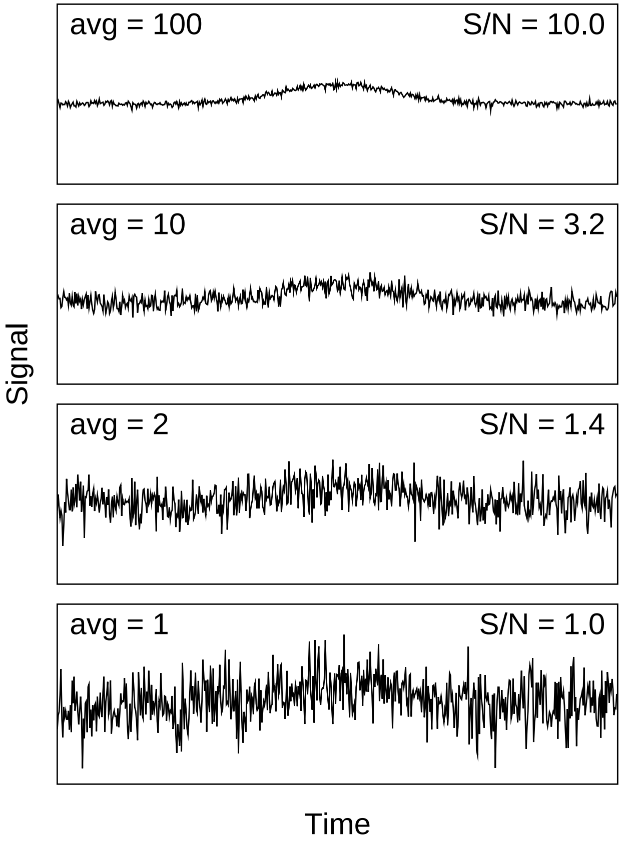


**Fig. 2.65**: Effect of repeating and averaging measurements to increase the S/N ratio. The bottom plot has a S/N of 1, with the amplitude of the signal $V_\mathrm{s}$ equal to the root-mean-square of the noise. Successive repetition of the measurement increases the S/N, proportional to $\sqrt{n}$

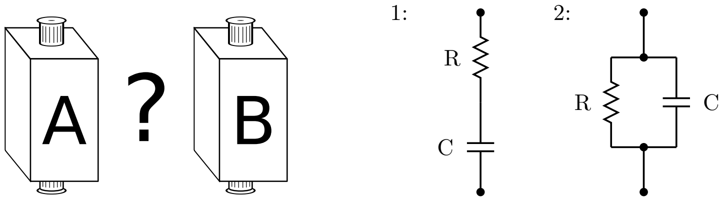


**Fig. 2.66**: Inside the sealed boxes A and B with BNC connectors are the circuits 1 and 2, but which circuit is in which box A is not known, nor are the values of $R$ and $C$

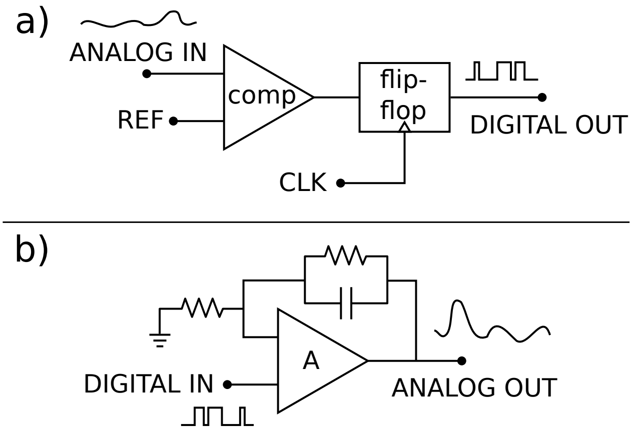
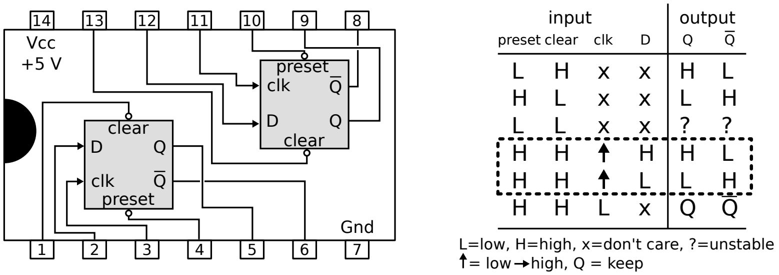


**Fig. 2.67**: a) 1-bit ADC and b) 1-bit DAC

placed in series, in the other in parallel, but it is not clear which circuit is in which box. See Figure 2.66. Find out experimentally which box contains which circuit, without opening them. The only thing allowed to do is connecting the boxes by their BNC connectors.

- Analog-digital converters (ADCs) are based on comparators, as described in Chapter 5. A simple 1-bit ADC can be made of a single comparator connected to a 'latch', the latter copies the input to the output when a clock pulse is received. . In this way a square pulse train is emitted at the output that represents the 1-bit sampling of the input. Design and implement a 1-bit ADC as in Figure 2.67. (A latch can be made from a D flipflop such as the 7474 integrated circuit, see Figure 2.68)
  A 1-bit digital-analog converter (DAC) can be made by low-pass filtering of the signal above. Design and implement a 1-bit DAC.

- Design and implement a hysteresis circuit with a switching window between 2 and 3 volt. An LED has to switch on when the input voltage rises above 3 volt and once switched on it should be switched off only when the voltage drops below 2 volt.



| input | | | | output | |
|---|---|---|---|---|---|
| preset | clear | clk | D | Q | $\overline{Q}$ |
| L | H | x | x | H | L |
| H | L | x | x | L | H |
| L | L | x | x | ? | ? |
| H | H | ↑ | H | H | L |
| H | H | ↑ | L | L | H |
| H | H | L | x | Q | $\overline{Q}$ |

L=low, H=high, x=don't care, ?=unstable
↑= low→high, Q = keep

**Fig. 2.68**: The 7474 integrated circuit containing two D flip-flops, and the D flip-flop truth table. To make a latch (copy input to output on clock flanks low-high), the highlighted inputs can be used

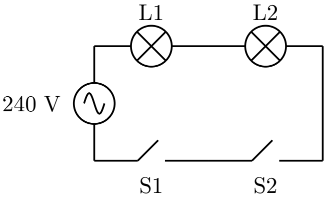


**Fig. 2.69**: Electronic enigma (Exercise 1)

## 2.10 Exercises

1. The circuit of Figure 2.69 contains an (AC) power source, two lightbulbs, L1 and L2, and two switches, S1 and S2. To switch on both lightbulbs we have to close both switches. By *adding* components we can make the circuit behave in such a way that closing S1 will switch on S1 only and S2 will switch on L2 only, independently. How?

2. Draw schematically the Bode plot of the circuit in Figure 2.70.

3. If an operational amplifier is not ideal, but has a finite gain $A$ instead, what is the relative 'loss', the error in a voltage follower? What is the error for a open-loop gain $A = 100$?

4. The circuit in Fig. 2.71 is strange. Using the rules of ideal opamp, what is the relation between $V_\mathrm{i}$ and $V_\mathrm{o}$?

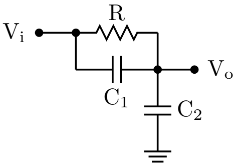


**Fig. 2.70**: Transfer function (Exercise 2)

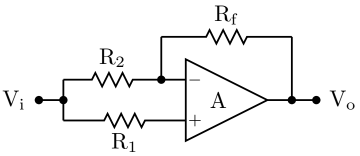


**Fig. 2.71**: Strange amplifier (Exercise 4)

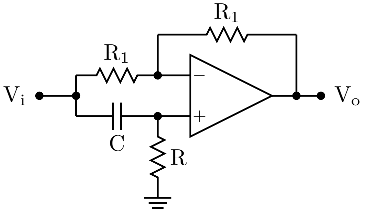


**Fig. 2.72**: All-pass filter (Exercise 5)

5. The filter in Figure 2.72 is called an all-pass filter. The gain at all frequencies is unity, and only the phase shifts along frequency. Explain why. (Hint: The magnitude of the ratio of two complex number is equal to the ratio of magnitudes of the individual numbers, $|a/b| = |a|/|b|$). Draw a Nyquist plot.

6. Using the ideal opamp rule $V_\mathrm{p} = V_\mathrm{n}$ for the unbalanced differential amplifier of Figure 2.32, prove Equations (2.155) and (2.156).

7. Calculate the transfer function $A_\mathrm{v}(\omega) = v_\mathrm{o}/v_\mathrm{i}$ of the integrator and differentiator.

8. Design an active LPF filter with the input at the positive terminal.

9. The circuit shown in Figure 2.73 is called a super-diode. It is an ideal rectifier in the sense that it copies the input for voltages larger than zero while the output is zero for negative voltages. Explain how the circuit works.

10. The circuit shown in Figure 2.74(a) is a logarithmic amplifier. It looks a lot like the super-diode of Exercise 9, but notice the subtle differences. Determine the relation between $V_\mathrm{o}$ and $V_\mathrm{i}$. Do the same for the exponential

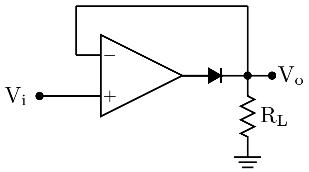


**Fig. 2.73**: Super diode (Exercise 9)

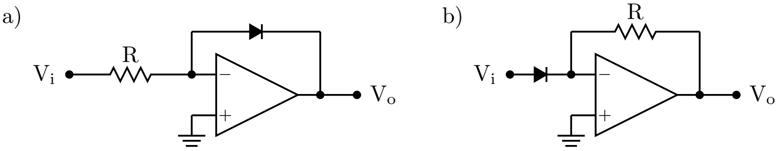


**Fig. 2.74**: (a) Logarithmic and (b) exponential amplifier (Exercise 10)

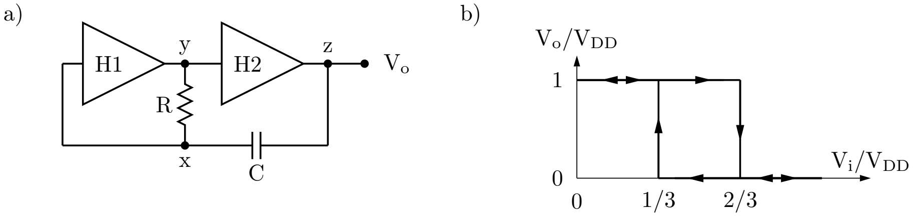


**Fig. 2.75**: a) Instable circuit (oscillator) made of two Schmitt triggers and an R-C pair. b) Behavior of each of the Schmitt triggers (Exercise 11)

amplifier in Figure 2.74(b) (NB: Ignore the "-1" term in the Ebers-Moll diode-current equation).

11. The circuit shown in Figure 2.75(a) is an oscillator based on two Schmitt triggers supplied with 0 and $V_{\rm DD}$. Imagine each having a behavior as shown in Figure 2.75(b), namely a hysteresis window from one third to two thirds of the supply voltage. The input of each Schmitt trigger is at a negative terminal of an opamp (see for instance Exercise 14) and thus has infinite input resistance; no current enters into the Schmitt triggers.

    a Explain how the circuit works by drawing the signals in time at points x, y and z.

    b Determine the oscillation period of the circuit

    c Give values for $R$ and $C$ to result in an oscillation of $f$ = 10 kHz.

12. The circuit in Figure 2.76 does not work as we want it to (rectifier). Explain why not.

13. The circuit in Figure 2.77 is a negative resistance converter, $R_{\rm i} \equiv V_{\rm i}/I_{\rm i} < 0$. Calculate the input resistance of the circuit.

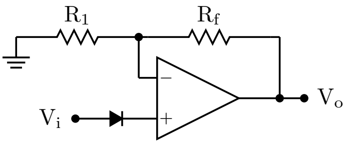


**Fig. 2.76**: Wrong rectifier. (Exercise 12)

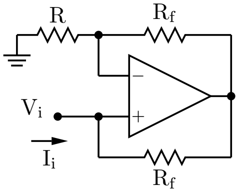


**Fig. 2.77**: Negative resistance circuit (Exercise 13)

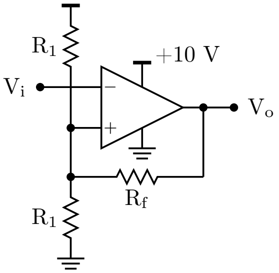


**Fig. 2.78**: Schmitt trigger (Exercise 14)

14. The circuit of Figure 2.78 is a comparator with hysteresis (Schmitt trigger). Determine the switching voltages $V_L$ and $V_H$ and draw the output voltage $V_o$ as a function of input voltage $V_i$. The resistances are $R_f = 100$ kΩ and $R_1 = 10$ kΩ. The supply voltages are +10 V and 0, respectively.

15. The circuit of Figure 2.79 is a Schmitt trigger (comparator with hysteresis). a) Give values for the components to result in a hysteresis window between 4 and 6 volt, as indicated in the right part of the figure. b) What is the input resistance of the circuit?

16. The circuit of Figure 2.80 is a Schmitt trigger (comparator with hysteresis). Give values for the components to result in a hysteresis window between $a = 1$ and $b = 2$ volt, as indicated in the right part of the figure.

17. The circuit of Figure 2.81 is a voltage-controlled oscillator (VCO). The

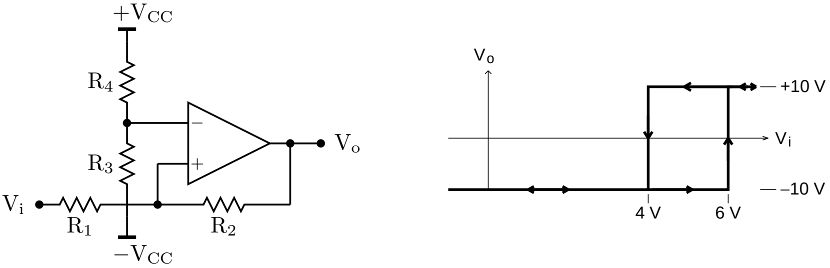


**Fig. 2.79**: Schmitt trigger with a hysteresis window between 4 V and 6 V (Exercise 15)

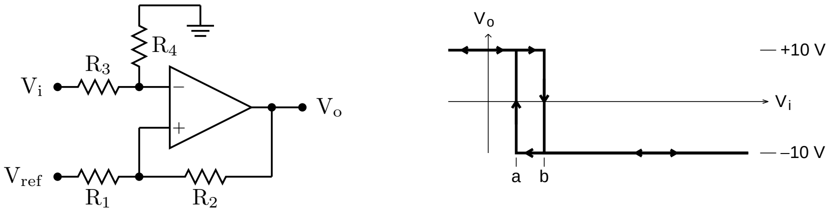


**Fig. 2.80**: Schmitt trigger with a hysteresis window between $a$ and $b$ (Exercise 16)

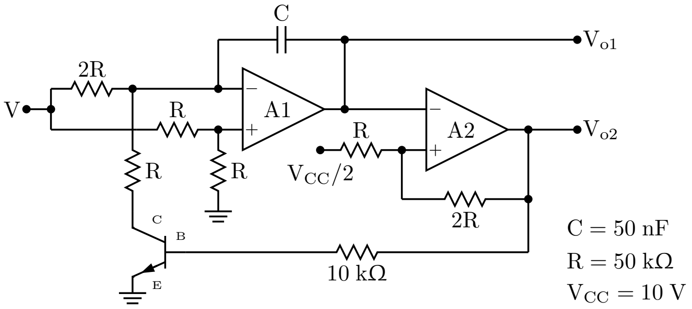


**Fig. 2.81**: Voltage-controlled oscillator (VCO) (Exercise 17)

output *frequency* is determined by the input *voltage*. The resistances are $R = 50$ kΩ and the capacitance is $C = 50$ nF. The single power supply is $V_{CC} = 10$ V (the other supply is ground). The transistor can be seen as an ideal switch: if a voltage at the base (B) is high, it connects the collector (C) to the emitter (E, in this case ground). The 10 kΩ is to protect the transistor by limiting the current.

a Explain how the circuit works

b Determine the relation between output frequency ($f$) and input voltage ($V$).

c What is the sensitivity $S$ of this actuator?

d What is the range of frequencies?

e Is this a good actuator in terms of linearity?

18. Figure 2.82 shows an oscillator based on the 555 timer circuit. How does this circuit work? Find an expression for the times $t_1$ and $t_2$ that the output is high, respectively low.

19. Based on a 555 timer IC design a circuit that lights an LED when an engine is running. A sensor is connected to the running engine that generates pulses with 1 ms interval. As long as these pulses keep coming in, the LED has to light up. Alternatively, we can have the LED lighting up in the absence of the pulses, to indicate a failure.

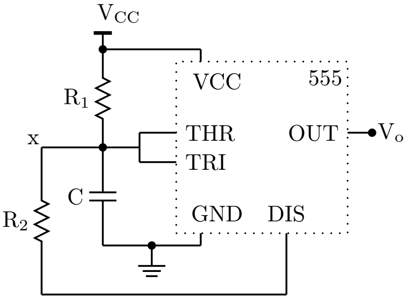


**Fig. 2.82**: Oscillator using a 555 Timer IC

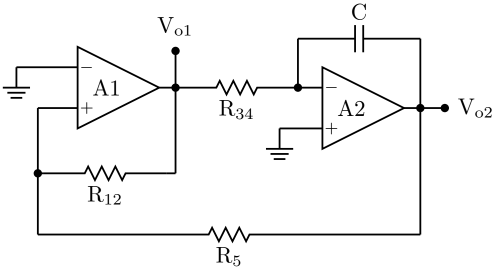


**Fig. 2.83**: Function generator using two opamps. (Exercise 20)

20. Figure 2.83 shows a function generator based on two op-amps. The left one is configured as a comparator with hysteresis, while the right one is configured as an integrator. Schematically draw the signals at output 1 and 2. Find values for the components to result in a wave of 1 V amplitude and 1 kHz frequency at output 2.

21. Figure 2.84 shows a Colpitts oscillator, similar in working as the Wien bridge oscillator (page 101). Find the condition for oscillation and make a Nyquist plot of $A\beta(\omega)$. Seemingly, the first capacitor, $\mathrm{C}_1$, does not do anything (because it is connected to ground and an ideal signal source $V_o$). The output impedance of the amplifier, however, cannot be neglected, although at the final calculation it does not play a part.

22. Figure 2.85 shows a Hartley oscillator, similar in working as the Colpitts bridge oscillator above (Ex. 21). Find the condition for oscillation and make a Nyquist plot of $A\beta(\omega)$.

## 2.11 Answers

1 Figure 2.86 shows the solution that adds 4 diodes (D1s, D2s, D1l, D2l) to the circuit. To understand how it works, imagine the AC power source is in the half-cycle in which the bottom voltage is higher than the top voltage. If we close switch S1, we now have a path for current to flow,

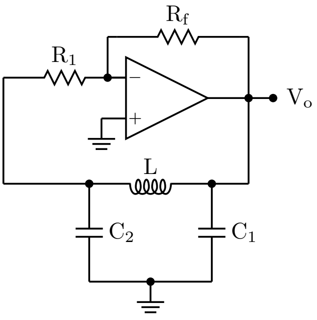


**Fig. 2.84**: Colpitts oscillator. (Exercise 21)

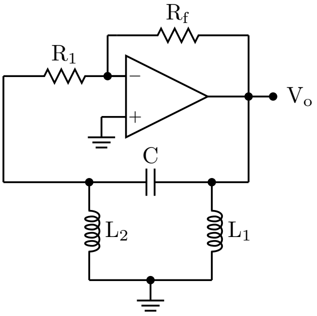


**Fig. 2.85**: Hartley oscillator. (Exercise 22)

namely S1-D2s-D2l-L1 and lightbulb L1 will switch on. Note that at the other half of the power cycle, the lightbulb will not be on. This circuit can easily be constructed and when hiding the diodes in the switches and the lightbulb fixtures the circuit will look like an enigma.

2 Given the fact that the impedance of a resistor is $R$ and a capacitor $1/sC$, we find the transfer of the voltage divider as

$$\frac{v_\mathrm{o}}{v_\mathrm{i}} = \frac{1/sC_2}{1/sC_2 + \frac{R/sC_1}{R+1/sC_1}} = \frac{1+sRC_1}{1+sR(C_1+C_2)}. \tag{2.227}$$

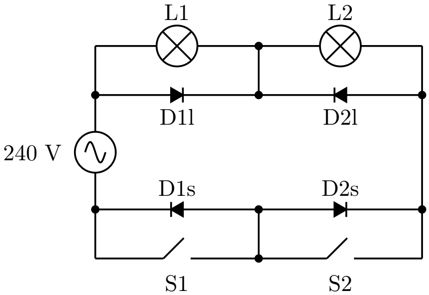


**Fig. 2.86**: Electronic enigma. (Exercise 1)

At low frequencies the capacitors are open circuit and the circuit lets through everything, $v_\mathrm{o}/v_1 = 1$.

$$\left.\frac{v_\mathrm{o}}{v_\mathrm{i}}\right|_{\mathrm{s}\to 0} = 1. \tag{2.228}$$

At high frequencies the capacitors become low impedance and the resistance R becomes relatively unimportant. In that case the voltage divider becomes

$$\left.\frac{v_\mathrm{o}}{v_\mathrm{i}}\right|_{\mathrm{s}\to\infty} = \frac{1/sC_2}{1/sC_2 + 1/sC_1} = \frac{C_1}{C_1 + C_2}, \tag{2.229}$$

which is a value between zero and one. Somewhere between the two extremes there are two critical frequencies, the first one where the output starts dropping as in a low-pass filter and the second where it starts stabilizing at the final value given above. Analyzing the above equation we can see that the first frequency is given by $\omega = 1/R(R_1 + C_2)$ and the second by $\omega = 1/RC_1$. Considering the fact that every every cut-off frequency changes the slope in a Bode plot by ± 20 dB/dec and changes the phase by $\pm 90^\mathrm{o}$ (depending on the type of cut-off frequency) we find the Bode plot as shown in Figure 2.87. As can be seen, the behavior is very similar to a low-pass filter, with the exception that the transfer function does not go to zero and the phase returns to zero degrees instead of $-90^\mathrm{o}$. Once again, the circuit consists of passive elements and the Nyquist falls entirely withing the unity circle (dashed line style) delimiting the area with gain from the area without gain

3 The relative gain can be defined as

$$A_\mathrm{r} \equiv \frac{V_\mathrm{o}(A)}{V_\mathrm{o}(A=\infty)} = \frac{A}{A+1} = \frac{1}{1+1/A}, \tag{2.230}$$

which we can Taylor expand ($x \equiv 1/A$) to give an approximation

$$A_\mathrm{r} \approx 1 - 1/A + \mathcal{O}([1/A]^2). \tag{2.231}$$

We can see that for $A = 100$ the error is about 1%.

4 The input resistance of an opamp is infinite, so no current enters its terminals. Specifically, $I_\mathrm{p} = 0$. This means there is no current passing through $\mathrm{R}_1$ and no voltage drop is induced in this resistance. This means that $V_\mathrm{p} = V_\mathrm{i}$. Another rule of an ideal opamp specifies that $V_\mathrm{p} = V_\mathrm{n}$. Thus, $V_\mathrm{n}$ is also equal to $V_\mathrm{i}$. With resistance $\mathrm{R}_2$ feeling no voltage drop, its current is zero. Using Kirchhoff's law for currents (KCL) at the point $\mathrm{V}_\mathrm{n}$, we find that the current through $\mathrm{R}_\mathrm{f}$ must also be zero (if nothing comes in, nothing can come out of point $\mathrm{V}_\mathrm{n}$). No current means no voltage drop and the right side of $\mathrm{R}_\mathrm{f}$ must have an equal voltage compared to its left side. Thus, $V_\mathrm{o} = V_\mathrm{i}$.

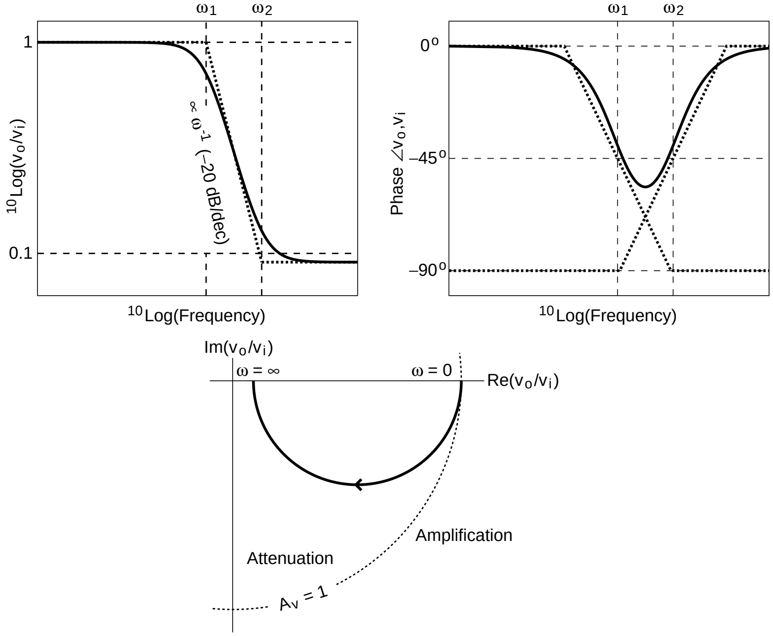


**Fig. 2.87**: Bode (dashed: approximation) and Nyquist plots of filter of Figure 2.70 with $R = 1$ kΩ, $C_1 = 1$ nF and $C_2 = 10$ nF. (Exercise 2)

5 The voltage at the positive input terminal is given by

$$V_\mathrm{p} = \frac{R}{R + 1/j\omega C} V_\mathrm{i}, \tag{2.232}$$

and for an ideal amplifier without saturation this must also be the voltage at the negative terminal, $V_\mathrm{n} = V_\mathrm{p}$. The input current through $R_1$ is then $I_\mathrm{i} = (V_\mathrm{i} - V_\mathrm{n})/R_1$. The output voltage is given by $V_\mathrm{o} = V_\mathrm{n} - I_\mathrm{i}R_1$ and we thus find a total gain $V_\mathrm{o}/V_\mathrm{i}$ equal to

$$A_\mathrm{V} \equiv \frac{V_\mathrm{o}}{V_\mathrm{i}} = \frac{j\omega RC - 1}{j\omega RC + 1}. \tag{2.233}$$

The amplitude of this gain is always unity, since generally $|(a - jb)/(a + jb)| = 1$,

$$\begin{aligned} |A_\mathrm{v}| &= 1, && (2.234) \\ \phi &\equiv \tan^{-1}\left[\frac{\mathrm{Im}(A_\mathrm{v})}{\mathrm{Re}(A_\mathrm{v})}\right] \\ &= \tan^{-1}\left[\frac{2\omega RC}{(\omega RC)^2 - 1}\right] = 180^\circ - 2\tan^{-1}(\omega RC). && (2.235) \end{aligned}$$

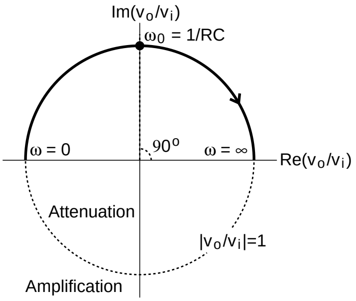


**Fig. 2.88**: Nyquist plot of an all-pass filter as shown in Figure 2.72. The gain is unity and only the phase changes from $180^\circ$ to $0^\circ$ with exactly $90^\circ$ at the frequency $\omega = 1/RC$

The gain is unity and the phase changes from $180^\circ$ for low frequencies to $0^\circ$ for high frequencies, with a phase of exactly $90^\circ$ at $\omega = \omega_0 = 1/RC$, as can be seen in the Nyquist plot of Fig. 2.88. Looking it at another way: At low frequencies the capacitor is open circuit and we have a 'classical' inverting amplifier (see Figure 2.29 of Section 2.6.2), with $R_\mathrm{f} = R_1$ this gives a gain $-1$ (Note that the resistance R has no effect, since no current passes through it since the input resistance of the opamp is infinite; effectively $\mathrm{V_p}$ is connected to ground). For high frequencies the capacitor is a short circuit and we have the circuit of Question 4 which, as we have seen, resulted in a gain of $+1$. These are the extreme points of the spectrum.

6 We use the equality $V_\mathrm{p} = V_\mathrm{n}$ and calculate the current $I_1$ through $\mathrm{R_1}$ which is also forced through $\mathrm{R_f}$, inducing a voltage drop relative to $\mathrm{V_n}$:

$$\begin{aligned}
V_\mathrm{n} &= V_\mathrm{p} = V_\mathrm{ip}. \\
I_1 &= (V_\mathrm{in} - V_\mathrm{n})/R_1 = (V_\mathrm{in} - V_\mathrm{ip})/R_1. \\
V_\mathrm{o} &= V_\mathrm{n} - I_1 R_\mathrm{f} \\
&= V_\mathrm{ip} - \frac{V_\mathrm{in} - V_\mathrm{ip}}{R_1} R_\mathrm{f} \\
&= V_\mathrm{ip} \frac{R_1 + R_\mathrm{f}}{R_1} - V_\mathrm{in} \frac{R_\mathrm{f}}{R_1}, \qquad (2.236)
\end{aligned}$$

which is equal to Eq. (2.155). For the balanced differential amplifier we get

$$\begin{aligned}
V_\mathrm{n} &= V_\mathrm{p} = V_\mathrm{ip}\frac{R_\mathrm{f}}{R_1 + R_\mathrm{f}}. \\
I_1 &= (V_\mathrm{in} - V_\mathrm{n})/R_1. \\
V_\mathrm{o} &= V_\mathrm{n} - \frac{V_\mathrm{in} - V_\mathrm{n}}{R_1}R_\mathrm{f} \\
&= \frac{R_1 + R_\mathrm{f}}{R_1}V_\mathrm{n} - \frac{R_\mathrm{f}}{R_1}V_\mathrm{in} \\
&= \frac{R_1 + R_\mathrm{f}}{R_1}V_\mathrm{ip}\frac{R_\mathrm{f}}{R_1 + R_\mathrm{f}} - \frac{R_\mathrm{f}}{R_1}V_\mathrm{in} \\
&= \frac{R_\mathrm{f}}{R_1}(V_\mathrm{ip} - V_\mathrm{in}), \qquad (2.237)
\end{aligned}$$

which is equal to Eq. (2.156).

7 Integrator: Replace in Equation (2.174) $Z_1 = R$, $Z_\mathrm{f} = 1/j\omega C$ to get

$$A_\mathrm{v} = -\frac{1/j\omega C}{R} = -\frac{1}{j\omega RC}. \qquad (2.238)$$

Differentiator: Replace in Equation (2.174) $Z_1 = 1/j\omega C$, $Z_\mathrm{f} = R$ to get

$$A_\mathrm{v} = -\frac{R}{1/j\omega C} = -j\omega RC. \qquad (2.239)$$

Note that the integrator has an infinite gain for zero Hz (DC). This makes sense; for constant input voltage, the integrator keeps on integrating and raising the output voltage forever. Equally, the differentiator has infinite gain for high frequencies. The response to an instant step function is a delta-Dirac function, a spike of infinite height at the moment of the step and zero otherwise. While similar to passive RC-filters, the behavior is not the same.

8 Figure 2.89 shows a possible solution. Since the opamp draws no current, specifically the positive terminal, we can decompose the circuit for our analysis: A passive low-pass RC filter followed by an amplifier. We arrive then at the final transfer function

$$\frac{V_\mathrm{o}}{V_\mathrm{i}} = \frac{1}{1 + jf/f_0} \times \frac{R_1 + R_\mathrm{f}}{R_1}, \qquad (2.240)$$

with $f_0 = 1/2\pi RC$.

9 The opamp will try to maintain the equality $V_\mathrm{p} = V_\mathrm{n}$. Since $V_\mathrm{p} = V_\mathrm{i}$ and $V_\mathrm{n}$ is also equal to the output voltage $V_\mathrm{o}$ connected to the load $\mathrm{R_L}$, in principle the output voltage is the input voltage, as long as the opamp manages to keep the equality. For positive voltages, the current through

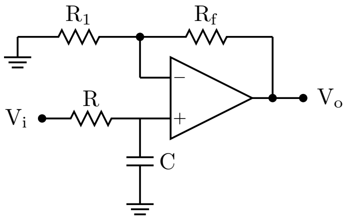


**Fig. 2.89**: Active filter with input at positive terminal. (Exercise 8)

the load resistance is positive and must come from the output of the opamp (input terminals of the opamp supply and sink no current). The opamp can supply this by putting at its output a voltage equal to $V_\mathrm{o}$ plus approximately 0.7 V, the nominal voltage drop of a diode operating in forward bias and supplying a reasonable current. So, as long as the input voltage is positive and lower than $V_{++}$ - 0.7 V, the opamp will not saturate and the output voltage is the input voltage.
When the input voltage is negative, it will try to make the output $V_\mathrm{o}$ also negative. However, this requires a current going in to the output of the opamp. This is blocked by the diode. No output voltage of the opamp will pull enough current through the diode. The current will be the reverse-bias saturation current (maybe as little as $10^{-14}$ A; effectively zero), and the output thus close to zero.
In conclusion we see that the output copies the input for positive voltages up to close to the supply voltage and zero for negative input voltages. Note that the circuit needs a load resistance, otherwise no current can flow anywhere and the diode remains closed and the output floating.

10 The negative terminal of the opamp is at virtual ground because of the ideal-opamp rule $V_\mathrm{p} = V_\mathrm{n}$ and $\mathrm{V_p}$ is connected to physical ground. This allows for the determination of the input current through the resistance, $I_\mathrm{i} = (V_\mathrm{i} - 0)/R = V_\mathrm{i}/R$. This current has nowhere to escape but through the diode. Using the Ebers-Moll current equation of a diode we can say that the current through the diode $I_\mathrm{D}$ and its voltage drop $\Delta V$ are related as

$$I_\mathrm{D} = I_\mathrm{S} \exp\left(\frac{\Delta V}{V_\mathrm{T}}\right) = \frac{V_\mathrm{i}}{R}. \tag{2.241}$$

Inverting this equation we get

$$V_\mathrm{o} = 0 - \Delta V = -V_\mathrm{T} \ln\left(\frac{V_\mathrm{i}}{RI_\mathrm{S}}\right). \tag{2.242}$$

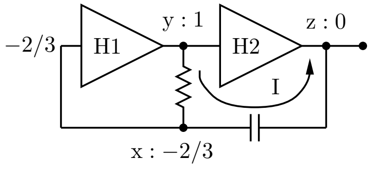
For the circuit in Fig. 2.74(a) we can make a similar analysis. The negative terminal is at virtual ground and we can calculate the diode current

$$I_\mathrm{D} = I_\mathrm{S} \exp\left(\frac{\Delta V_\mathrm{i}}{V_\mathrm{T}}\right). \tag{2.243}$$

This current is forced through the feedback resistor R and induces a voltage drop. With one side of the resistor at (virtual) ground, the other side, the output, is at

$$V_\mathrm{o} = -I_\mathrm{S} R \exp\left(\frac{\Delta V_\mathrm{i}}{V_\mathrm{T}}\right). \tag{2.244}$$

11 This calculation is done for $V_\mathrm{DD} = 1$ V. The functionality of the circuit does not depend on the exact value of $V_\mathrm{DD}$.
Start with $V_\mathrm{x} = 1$ V, $V_\mathrm{y} = 0$, $V_\mathrm{z} = 1$ V and the capacitor empty, $Q_\mathrm{C} = 0$. The resistor feels a voltage drop $V_\mathrm{x} - V_\mathrm{y} = 1$ V and a current will flow through it. This current can only come from the output of Schmitt trigger H2 (remember that they are made of opamps) and thus passes through the capacitor, charging it. Initially $I = \Delta V_\mathrm{R}/R = 1/R$, but the charging of the capacitor makes it having a voltage drop; the voltage at x is therefore decreasing and with y at a steady 0, the voltage drop across the resistor drops and the current with it. We recognize here a classic relaxation behavior. The current will continue to increase the charge in the capacitor in an ever-decreasing way. This, however, will not continue forever. When the voltage at x drops below 1/3 V the first Schmitt trigger (H1) will commutate; its output will switch $y : 0 \rightarrow 1$. In cascade the second Schmitt trigger will commutate, $z : 1 \rightarrow 0$.
At this moment we start the clock. Just before the commutation, the voltage was 1/3 V at x and 1 V at z; a voltage drop of 2/3 V across the capacitor. Capacitors have the property that voltage drops cannot change instantaneously (because $\Delta V_\mathrm{C} = Q/C$ and charge cannot disappear instantaneously; it takes current and time to remove charge). Thus, immediately after the switching of H2, the voltage at x must be $V_\mathrm{x} = V_\mathrm{z} - 2/3$ V $= -2/3$ V. At the other side of the resistor there is a voltage $V_\mathrm{y} = 1$ V and a current will come through the resistor equal to $I = (V_\mathrm{y} - V_\mathrm{x})/R$ (supplied by Schmitt trigger H1, passing through C and sinking into H2). The situation at $t = 0$ is as follows:



This current goes through the capacitor and is supplied by the output of H2. Seemingly the current goes against the voltage (from 1 V on the right side to +5/3 V on the left side). This is only seemingly. Don't forget that the current in a capacitor is not proportional to the voltage drop (as a resistor), but proportional to the *time-derivative* of this voltage. The current charges the capacitor in an ever-decreasing way. A classical relaxation behavior with for this case three boundary conditions for the voltage at x:

- Initially, $V_{\mathrm{x}}(0) = -2/3$ V.
- The final voltage, if nothing further happened is, $V_{\mathrm{x}}(\infty) = 1$ V, because at this value the current through the resistor would be zero $(\Delta V_{\mathrm{R}} = V_{\mathrm{x}} - V_{\mathrm{y}})$.
- The relaxation time is $\tau = RC$.

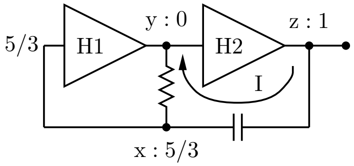
The solution to this exponential decay/approach is

$$V_{\mathrm{x}}(t) = 1 - \frac{5}{3}\exp(-t/RC). \tag{2.245}$$

When $V_{\mathrm{x}}$ reaches two 2/3 V, H1 will commutate. This occurs at a time

$$t_1 = RC\ln(5). \tag{2.246}$$

At this moment y switches $1 \to 0$ and in cascade H2 switches, z: $0 \to 1$. We reset the clock and start observing what happens. Again, the voltage drop in the capacitor is not affected by the switching. Before the switch, the voltage drop was $\Delta V_{\mathrm{C}} = V_{\mathrm{z}} - V_{\mathrm{x}} = 0 - 2/3$ V $= -2/3$ V. After the switch it must be the same, so the voltage at x is $V_{\mathrm{x}} = V_{\mathrm{z}} - \Delta V_{\mathrm{C}} = 1$ V $+2/3$ V $= 5/3$ V. We thus have the following situation at $t = 0$:



A current flows through the resistor and sinks into the output of H1. Thus, the capacitor is discharging. Once again in a relaxation way resulting in an exponential decay/approach. The conditions for the voltage at x are:

- Initially, $V_{\mathrm{x}}(0) = 5/3$ V.
- The final voltage, if nothing further happened is, $V_{\mathrm{x}}(\infty) = 0$ V, because at this value the current through the resistor would be zero $(\Delta V_{\mathrm{R}} = V_{\mathrm{x}} - V_{\mathrm{y}})$.
- The relaxation time is $\tau = RC$.

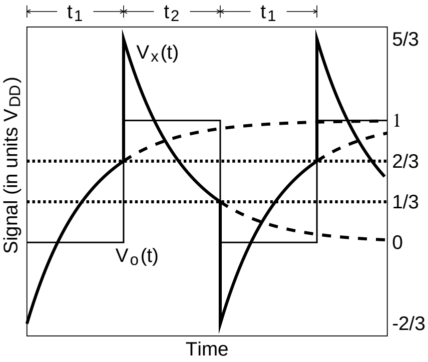
The solution to this exponential decay/approach is

$$V_{\mathrm{x}}(t) = \frac{5}{3}\exp(-t/RC). \tag{2.247}$$

It continues to drop until it reaches the commutation voltage of H1, namely 1/3 V. This happens at a time given by

$$t_2 = RC\ln(5). \tag{2.248}$$

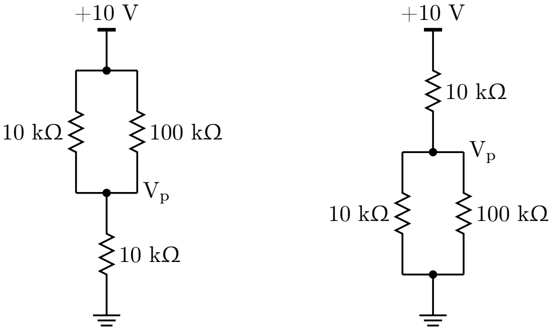
At this moment y switches 0 → 1 and in cascade H2 switches, z: 1 → 0 and we start a new cycle again.
The total period is given by $T = t_1 + t_2 = 2RC\ln(5)$. To get an oscillation frequency for the square-wave output at 10 kHz we can use values $C = 1$ nF and $R = 31$ kΩ.
Summarizing, the behavior of the circuit is as given below. The solid curve is the actual behavior, the dashed curves are guides-to-the-eye to show where the signal would have gone to. The horizontal dotted lines represent the switching levels, 1/3 V and 2/3 V.



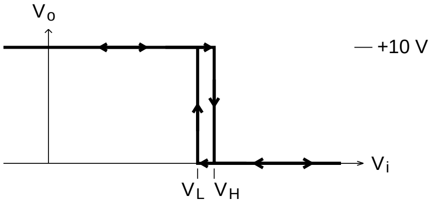
12 The rectifier does not work because the input resistance of the opamp is infinite. There is never going to be any current passing through the diode and it thus has no rectifying property.

13 Assume that $V_\mathrm{p} = V_\mathrm{n}$, that no current enters the input terminals and the output can sink or source whatever current necessary to maintain the condition. Now try again. Resistance R is with one foot connected to ground and the other to $V_\mathrm{n}$ which is equal to $V_\mathrm{i}$, it thus feels a voltage drop $V_\mathrm{i}$ and a current $I = V_\mathrm{i}/R$ is passing. This current must come from $\mathrm{V_o}$ via $\mathrm{R_f}$ (the one in the negative branch). The output voltage must thus be $V_\mathrm{o} = V_\mathrm{i} + IR_\mathrm{f}$. The bottom $\mathrm{R_f}$, part of the positive branch feels one side $V_\mathrm{i}$ and on the other side this $V_\mathrm{o}$. The current is thus equal to $(V_\mathrm{o} - V_\mathrm{i})/R_\mathrm{f} = I$. Note that the current goes from $\mathrm{V_o}$ to $\mathrm{V_i}$. Seen from the input terminal, for a positive voltage $V_\mathrm{i}$, a current equal to $I = V_\mathrm{i}/R_1$ comes out, or a current $I_\mathrm{i} = -I = -V_\mathrm{i}/R_1$ goes in. Applying Ohm's law gives the input resistance, $R_\mathrm{i} = V_\mathrm{i}/I_\mathrm{i} = -R$.

14 For $V_\mathrm{o} = 10$ V, the voltage at the positive terminal can be found easily if we realize that the situation is effectively a voltage divider as shown below (left).



This gives a voltage $V_\mathrm{p} = 5.24$ V. For $V_\mathrm{o} = 0$, the situation is effectively as in the right part of the figure, resulting in $V_\mathrm{p} = 4.76$ V. These are the $V_\mathrm{H}$ and $V_\mathrm{L}$ switching voltages, respectively, for the input voltage $V_\mathrm{i}$. We see from Figure 2.78 that the input signal is at the negative terminal of the opamp, and we can expect the output to be the positive supply voltage (10 V) for the smallest input voltages and the negative supply voltage (0) for the largest input voltages. We thus wind up with the behavior as shown below.



15 a) Define $\beta \equiv R_1/(R_1 + R_2)$, $(1 - \beta) = R_2/(R_1 + R_2)$. Then: $V_\mathrm{p} = \beta V_\mathrm{o} + (1 - \beta)V_\mathrm{i}$. In the right commutation point $V_\mathrm{i} = 6$ V, $V_\mathrm{o} = -10$ V, and $V_\mathrm{p} = V_\mathrm{n}$:

$$V_\mathrm{n} = \beta(-10\ \mathrm{V}) + (1 - \beta)(6\ \mathrm{V}) = (6 - 16\beta)\ \mathrm{V}. \tag{2.249}$$

In the other commutation point $V_\mathrm{i} = 4$ V, $V_\mathrm{o} = +10$ V, and $V_\mathrm{p} = V_\mathrm{n}$:

$$V_\mathrm{n} = \beta(10\ \mathrm{V}) + (1 - \beta)(4\ \mathrm{V}) = (4 + 6\beta)\ \mathrm{V}. \tag{2.250}$$

Two equations with two unknowns. The solution is $\beta = 1/11$ (for example $R_1 = 1$ kΩ and $R_2 = 10$ kΩ) and $V_\mathrm{n} = 50/11 = 4.55$ V (for example $R_3$ = 8 kΩ and $R_4 = 3$ kΩ).

b) The input resistance is defined as $r_\mathrm{i} = \mathrm{d}V_\mathrm{i}/\mathrm{d}I_\mathrm{i}$. The input current is given as $I_\mathrm{i} = (V_\mathrm{i} - V_\mathrm{o})/(R_1 + R_2)$. For a constant output voltage, the input resistance is thus equal to $R_1 + R_2$.

16 Define the two voltage dividers, $\beta_1 \equiv R_1/(R_1 + R_2)$ and $\beta_2 \equiv R_4/(R_3 + R_4)$. Then $V_\mathrm{p} = \beta_1 V_\mathrm{o} + (1 - \beta_1)V_\mathrm{ref}$ and $V_\mathrm{n} = \beta_2 V_\mathrm{i}$. Then, for the

commutation point $a$, we have $V_\mathrm{o} = -10$ V, $V_\mathrm{i} = a$, $V_\mathrm{n} = V_\mathrm{p}$, the latter rewritten as

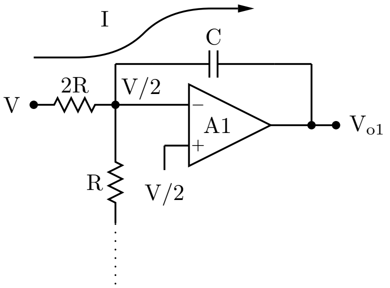
$$\beta_1(-10 \text{ V}) + (1 - \beta_1)V_\mathrm{ref} = a\beta_2. \qquad (2.251)$$

For the second commutation point ($b$), we have $V_\mathrm{o} = +10$ V, $V_\mathrm{i} = b$, and $V_\mathrm{n} = V_\mathrm{p}$, the latter resulting in

$$\beta_1(10 \text{ V}) + (1 - \beta_1)V_\mathrm{ref} = b\beta_2. \qquad (2.252)$$

Combining the two equations, subtracting (2.251) from (2.252), gives $\beta_1(20 \text{ V}) = (b - a)\beta_2$. For $a = 1$ V and $b = 2$ V this becomes $\beta_2 = 20\beta_1$. Substituting this into Eq. (2.252) gives $\beta_1 = V_\mathrm{ref}/(30 \text{ V} + V_\mathrm{ref})$. For example: $\beta_2 = 1$ ($R_3 = 0$), $\beta_1 = 1/20$, and $V_\mathrm{ref} = 30/19$ V.

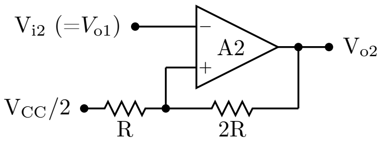
17 The opamps are ideal and have zero output resistance. For that reason we can analyze them separately. In the first situation the output of A2 is low. The transistor is effectively an open circuit. At the input of opamp A1 we have $V/2$ at the positive terminal because of the voltage divider made of two resistors R. Assuming no saturation, we also have $V/2$ at the negative terminal. We thus have the following situation:



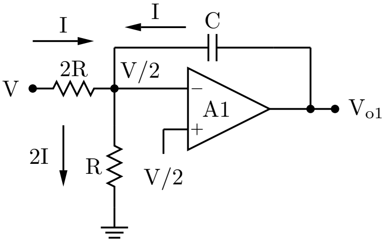
In this case the current is $I = (V - V/2)/2R = V/4R$. The output of this opamp is then the integral of the current, $V_\mathrm{o1} = -Q_\mathrm{C}/C = -\int(I/C)\mathrm{d}t + V_\mathrm{o1,0}$. The current is constant, since the negative terminal of A1 remains at $V/2$ (as long as $V_\mathrm{p}$ remains at this value). This is not a relaxation system. With a constant current, the output voltage as a function of time is given by

$$V_{01}(t) = -\frac{V}{4RC}t + V_\mathrm{o1,0}, \qquad (2.253)$$

i.e., a linearly dropping function with a slope equal to $-V/4RC$. The second opamp (A2) is configured as a Schmitt trigger. $V_\mathrm{p} = 2V_\mathrm{CC}/3$ if $V_\mathrm{o2} = V_\mathrm{CC}$ and $V_\mathrm{p} = +V_\mathrm{CC}/3$ when $V_\mathrm{o2} = 0$.



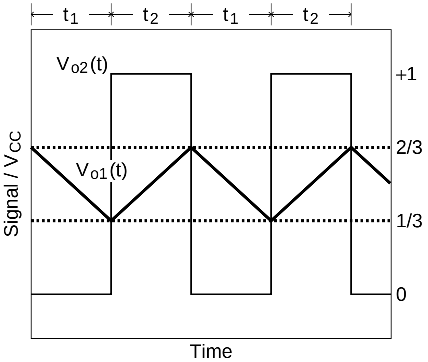
When the output of this opamp is high, the transistor is activated and effectively becomes a short circuit between collector and emitter. In other words, connecting the dangling resistor at opamp A1 to ground. We then get the following situation at A1:



$V_p$ and $V_n$ are still both at $V/2$. The current through the resistor 2R thus remains the same, as found above. However, this time an additional resistor R is connected to ground. This carries a current $2I$. Half of it is supplied by the branch with the 2R resistor connected to V. The rest must be coming from the branch with the capacitor C. The capacitor is thus discharged with a current equal to $I = V/4R$. The output is then

$$V_{01}(t) = +\frac{V}{4RC}t + V_{o1,0}, \tag{2.254}$$

i.e., a linearly rising function with a slope equal to $+V/4RC$. If we now connect all the parts together then we can find the behavior as shown below:



From there it is also easy to calculate the two half-periods. One time $t_1$ it takes to drop the output voltage $V_{o1}$ from $2V_{CC}/3$ to $V_{CC}/3$, and one time $t_2$ it takes to rise the output voltage $V_{o1}$ from $V_{CC}/3$ to $2V_{CC}/3$. Equations (2.253) and (2.254) respectively give $t_1 = t_2 = 4RC/3 \times V_{CC}/V$. The frequency of oscillation thus becomes

$$f = \frac{1}{t_1 + t_2} = \frac{3}{8RC} \times \frac{V}{V_{CC}}, \tag{2.255}$$

i.e., a frequency that linearly depends on the input voltage (which makes it an ideal actuator, answer [e]). The sensistivity of the actuator is given by

$$S \equiv \frac{\mathrm{d}f(V)}{\mathrm{d}V} = \frac{3}{8RCV_{\mathrm{CC}}}. \tag{2.256}$$

To find the range of frequencies we just substitute the range of input voltages. Given the fact that these voltages cannot go beyond the power supply voltages, we find in this case an input voltage range 0 - $V_{\mathrm{CC}}$ and a frequency range of 0 to $3/8RC$.

18 Imagine the output is high, meaning Q is high and $V_{\mathrm{o}} = V_{\mathrm{CC}}$. This means that $\bar{\mathrm{Q}}$ is low and the transistor Q1, used as a switch, is not conducting; effectively the DIS pin is floating. The resistance $\mathrm{R}_2$ is thus connected to nothing and effectively we have the circuit as the one of the left side of Figure 2.90. This is a simple RC circuit; a current starts charging the capacitor C. We define a voltage $V_{\mathrm{x}}$ in between the resistor and the capacitor. Because the capacitor starts charging, this voltage $V_{\mathrm{x}}$ rises. Hence the current - given by $I(t) = (V_{\mathrm{CC}} - V_{\mathrm{x}}(t))/R_1$ - gradually drops. We recognize here a classical relaxation behavior, a voltage that exponentially approximates the supply voltage $V_{\mathrm{CC}}$ with a relaxation time given by $\tau_1 = R_1C$. If $V_{\mathrm{x}}$ rises above one third the supply voltage $V_{\mathrm{x}} > V_{\mathrm{CC}}/3$ nothing happens; the flip-flop SET line is un-set, but the flip-flop does not respond to changes high-low. In any case, as we will see in a moment, the starting voltage at $t = 0$ is not zero, but $V_{\mathrm{CC}}/3$ instead. We thus arrive at an equation for the voltage $V_{\mathrm{x}}$ as a function of time in the first part of the oscillation period as

$$\begin{aligned} V_{\mathrm{x}}(t) &= V_{\mathrm{x}}(t=\infty) + [V_{\mathrm{x}}(t=0) - V_{\mathrm{x}}(t=\infty)]\exp(-t/\tau_2) \\ &= \left[1 - \frac{2}{3}\exp\left(-\frac{t}{R_1C}\right)\right] V_{\mathrm{CC}}. \end{aligned} \tag{2.257}$$

If left alone, the $V_{\mathrm{x}}$ would reach in this way a voltage equal to $V_{\mathrm{CC}}$. However, it never reaches this voltage. When it rises above two-thirds the supply voltage, $V_{\mathrm{x}} = (2/3)V_{\mathrm{CC}}$, the top comparator triggers. This thus happens at a time given by

$$t_1 = R_1C \times \ln(2). \tag{2.258}$$

At this time the RESET line of the flip-flop is raised and the flip-flop output is set to low, $V_{\mathrm{o}} = 0$. Simultaneously, the negated output $\bar{\mathrm{Q}}$ is set to high. This opens the transistor Q1 and connects resistance $\mathrm{R}_2$ effectively to ground. We have then the situation as in Fig. 2.90(b). Now the capacitor will start to discharge. Or not. If we have chosen wrong values for the resistances, the capacitor might continue to charge! Look at point x of the circuit. At this point we can apply Kirchhoff's law - 'what goes in, must come out' - and calculate the current to and from the

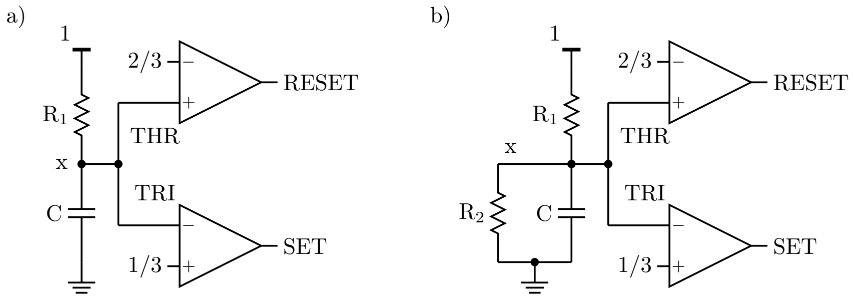


**Fig. 2.90**: Effective circuits when OUT is high (a) and low (b). Voltages in units of $V_{\mathrm{CC}}$

resistances; the difference is what enters the capacitor. With $V_{\mathrm{x}}$ at two-thirds $V_{\mathrm{CC}}$, through R$_1$ we have a current $I_1 = V_{\mathrm{CC}}/3R_1$, and through R$_2$ we have a current $I_2 = 2V_{\mathrm{CC}}/3R_2$. The latter should be larger, otherwise the capacitor will continue to charge. In other words, we have a condition for our circuit design, $R_2 < 2R_1$.

The capacitor will discharge and the voltage $V_{\mathrm{x}}$ decay exponentially. The quiescent point, the steady-state final value would be that value that no longer discharges the capacitor, thus where the current through R$_2$ and R$_1$ are equal,

$$\begin{aligned} \frac{V_{\mathrm{CC}} - V_{\mathrm{x}}(t=\infty)}{R_1} &= \frac{V_{\mathrm{x}}(t=\infty)}{R_2} \Rightarrow && (2.259) \\ V_{\mathrm{x}}(t=\infty) &= \frac{R_2}{R_1+R_2} V_{\mathrm{CC}}. && (2.260) \end{aligned}$$

If this final-voltage is not lower than one-third of the supply voltage, nothing more will happen and we do not have an oscillator. We thus get a condition for our oscillator circuit stronger than the one given above: $V_{\mathrm{x}}(t=\infty) < V_{\mathrm{CC}}/3$ implies $R_2 < R_1/2$.

With the starting voltage and the final voltage known, rests us only to find the relaxation time. In time and frequency analysis of circuits we know that voltage-supplies play the same role as ground. For calculating the time constants of circuits we can replace the $V_{\mathrm{CC}}$ with ground and we find the time constant equal to $\tau_2 = (R_1 \parallel R_2)C$. We thus arrive at the following equation for the voltage $V_{\mathrm{x}}$ as a function of time in the second part of an oscillation period

$$\begin{aligned} V_{\mathrm{x}}(t) &= V_{\mathrm{x}}(t=\infty) + [V_{\mathrm{x}}(t=0) - V_{\mathrm{x}}(t=\infty)] \exp(-t/\tau_2) \\ &= \left[\frac{R_2}{R_1+R_2} + \frac{2R_1 - R_2}{3(R_1+R_2)} \exp\left(-\frac{t}{(R_1 \parallel R_2)C}\right)\right] V_{\mathrm{CC}} && (2.261) \end{aligned}$$

The time it takes to reach one-third of the supply voltage is then given

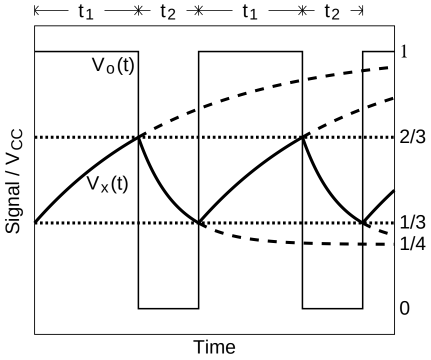


**Fig. 2.91**: Behavior of the oscillator circuit of Fig. 2.82 for $R_2 = R_1/3$

by

$$t_2 = (R_1 \parallel R_2)C \times \ln\left(\frac{2R_1 - R_2}{R_1 - 2R_2}\right), \qquad (2.262)$$

where we can also find back our condition; only for $R_2 < R_1/2$ is the argument of the logarithm positive (and larger than unity), to give a real solution. At this time the circuit starts the first part of an oscillation again, as described above. The frequency of oscillation is given by $f = 1/(t_1 + t_2)$ and an expression can easily be found. Figure 2.91 summarizes the behavior of this oscillator circuit based on the 555 timer IC.

19 The solution asks for a circuit without a dead time. The system is reset as long as the pulses come in. From the moment of receiving a pulse we start counting and if within let's say ten times the pulse rate the circuit does not receive a new pulse, the output LED will switch off. The 555 delay circuit of Section 2.7.5 seems very appropriate, which we project for a time delay $\Delta t$ equal to about 10 ms, for instance with a capacitor of 1 μF and a resistance of 10 kΩ. As can be seen in Figure 2.55, the output of the circuit is high as long as pulses keep coming in. If we want an LED switched on as long as the engine is running we connect an LED from the 555 output to ground ('protected' with a 100 Ω resistance). If on the other hand we want an LED to switch on when the engine stopped running, we connect the LED from output to supply voltage, see Figure 2.92. The resistor at the output is to limit the current through the diodes, which should be about 10 mA, since an LED is a *current*-to-light actuator and not a voltage-to-light actuator; first the output voltage has to be converted to a current by use of the actuator resistor (voltage-to-current converter).

20 The left opamp is configured as a comparator with hysteresis; positive feedback equal to $\beta = R_5/(R_5 + R_{12})$. The hysteresis window is given by $V_{\rm i1} = \pm\beta/(1-\beta) \times V_{\rm cc} = R_5/R_{12} \times V_{\rm cc}$. The right opamp is configured

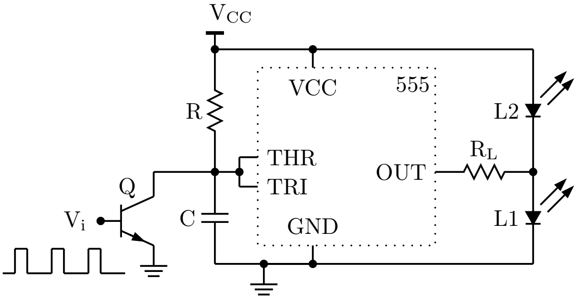


**Fig. 2.92**: Circuit for detection of presence or failure of a pulse train based on the 555 delay circuit of Section 2.7.5. An LED at the output either detects the presence of pulses (L1) or their absence (L2). (Exercise 19)

as an integrator with a time constant given by $\tau = R_{34}C$ and an output given by $V_{o2} = -(1/\tau)\int V_{i2}(t)dt + V_{o2,0}$, with $V_{i2} = V_{o1}$. Combining the two circuits we can see that the output of opamp 2 is rising linearly (with opamp 1 at $-V_{cc}$), until it hits the upper hysteresis level of opamp 1, which makes it commutate and starts lowering the output of opamp 2 linearly. Thus a triangular waveform results at output 2, with an amplitude of $\beta/(1-\beta) \times V_{cc} = R_5/R_{12} \times V_{cc}$, see Figure 2.93. The first half period of the oscillation can be found by writing the function of the transient and setting it to the hysteresis threshold, as in

$$V_{o2}(t) = \frac{\beta}{1-\beta}V_{cc} - \frac{V_{cc}}{R_{34}C}t = -\frac{\beta}{1-\beta}V_{cc}. \tag{2.263}$$

This gives a half-period time $\Delta T_1$. Because of the symmetry, the second half period is the same. Thus, the total period is

$$\Delta T = \Delta t_1 + \Delta t_2 = 4\frac{R_5}{R_12}R_{34}C. \tag{2.264}$$

For a triangular wave of 1 V amplitude and 1 kHz frequency we can use the following values: $V_{cc} = 10$ V, $R_5 = 10$ kΩ, $R_{12} = 100$ kΩ, $C = 1$ μF, $R_{34} = 2.5$ kΩ.

21 Figure 2.94 shows the circuit of the feedback loop β for the Colpitts Oscillator. The resistance $R$ is the output resistance of the opamp. Ideally this is zero, but any real opamp has finite output resistance and this is essential, the feedback loop β is per definition the part of the output $V_o$ that appears at the input $V_i$ of our amplifier (that has a gain equal to $A = -R_f/R_1$), that can be found by considering the voltage dividers in the circuit:

$$\beta \equiv \frac{V_i}{V_o} = \frac{Z}{Z+R} \times \frac{Z_{C2}}{Z_{C2}+Z_L}. \tag{2.265}$$

First, the second voltage divider composed of C$_2$ ($Z_{C2} = 1/sC_2$) and L

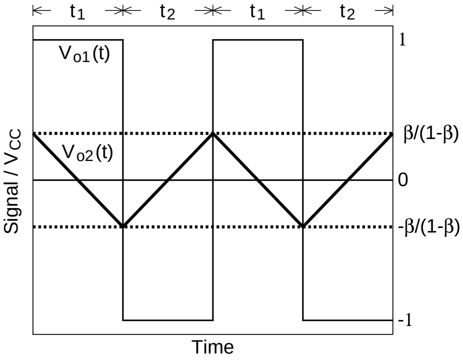


**Fig. 2.93**: Waveforms at output 1 and 2 of the function generator of Fig. 2.83

($Z_{\mathrm{L}} = sL$), with $s = j\omega$ is given by

$$\frac{Z_{\mathrm{C2}}}{Z_{\mathrm{C2}} + Z_{\mathrm{L}}} = \frac{1}{1 - \omega^2 L C_2}. \tag{2.266}$$

The impedance $Z$ is given by

$$Z = \left[(1/sC_1)^{-1} + (1/sC_2 + sL)^{-1}\right]^{-1}. \tag{2.267}$$

Substituting this in (2.265) gives

$$\beta = \frac{1}{(1 - \omega^2 L C_2) + j\omega R C_1 [C_2/C_1 + (1 - \omega^2 L C_2)]}. \tag{2.268}$$

Oscillation will occur when the loop gain $A\beta$ is unity. Since $A$ is purely real, this means that $\beta$ must be real too. This implies

$$R C_1 [C_2/C_1 + (1 - \omega^2 L C_2)] = 0. \tag{2.269}$$

Trivial solutions are $R = 0$ (ideal amplifier) or $C_1 = 0$. The non-trivial solution is

$$\omega = \sqrt{\frac{1}{L(C_1 \oplus C_2)}}, \tag{2.270}$$

with $C_1 \oplus C_2$ the series sum capacitance of $C_1$ and $C_2$, $C_1 \oplus C_2 = (1/C_1 + 1/C_2)^{-1}$. At this frequency the feedback loop is, according to Eq. (2.268), $\beta = -C_1/C_2$. Thus, if our amplifier has a gain $A$ such that the total loop-gain is unity the Barkhausen Criterion will be met; if

$$A\beta = \left(-\frac{R_{\mathrm{f}}}{R_1}\right) \times \left(-\frac{C_1}{C_2}\right) = 1, \tag{2.271}$$

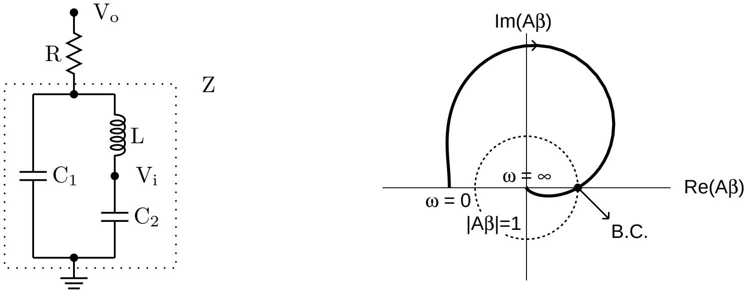


**Fig. 2.94**: Left: Circuit of the feedback loop of the Colpitts Oscillator. The resistance $R$ is the output resistance of the opamp. Right: Nyquist plot for an example that meets the condition (Barkhausen Criterion, B.C.) of oscillation, for $R_\mathrm{f}/R_1 = C_2/C_1$. (Exercise 21)

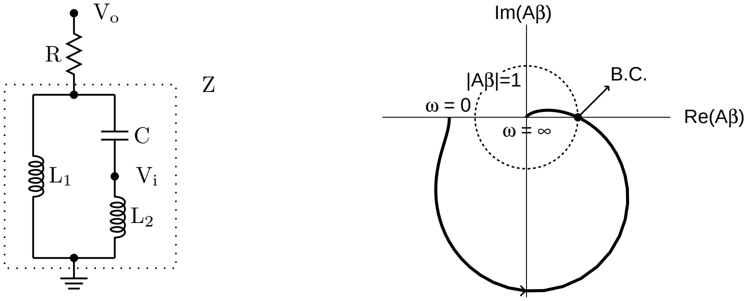


**Fig. 2.95**: Left: Circuit of the feedback loop of the Hartley Oscillator. The resistance $R$ is the output resistance of the opamp. Right: Nyquist plot for an example that meets the condition (Barkhausen Criterion, B.C.) of oscillation, for $R_\mathrm{f}/R_1 = L_1/L_2$. (Exercise 22)

the circuit will oscillate at the frequency given by

$$f = \frac{1}{2\pi}\sqrt{\frac{1}{L(C_1 \oplus C_2)}}, \tag{2.272}$$

Note that the output resistance of the opamp does not enter into the final results. It only has effect in the *quality* of oscillation; the lower $R$ the better ('sharper') the oscillation frequency.

22 Figure 2.95 shows the circuit of the feedback loop β for the Hartley Oscillator. The resistance $R$ is the output resistance of the opamp. Ideally this is zero, but any real opamp has finite output resistance and this is essential, the feedback loop β is per definition the part of the output $V_\mathrm{o}$ that appears at the input $V_\mathrm{i}$ of our amplifier (that has a gain equal to $A = -R_\mathrm{f}/R_1$), that can be found by considering the voltage dividers in

the circuit:

$$\beta \equiv \frac{V_\mathrm{i}}{V_\mathrm{o}} = \frac{Z}{Z+R} \times \frac{Z_\mathrm{L2}}{Z_\mathrm{L2}+Z_\mathrm{C}}. \tag{2.273}$$

First, the second voltage divider composed of $\mathrm{L}_2$ ($Z_\mathrm{L2} = sL_2$) and C ($Z_\mathrm{C} = 1/sC$), with $s = j\omega$ is given by

$$\frac{Z_\mathrm{L2}}{Z_\mathrm{L2}+Z_\mathrm{C}} = \frac{sL_2}{sL_2+1/sC}. \tag{2.274}$$

The impedance $Z$ is given by

$$Z = \frac{sL_1 \times (1/sC + sL_2)}{sL_1 + sL_2 + 1/sC}. \tag{2.275}$$

Substituting this in (2.273) gives

$$\beta = \frac{-\omega^2 L_1 L_2}{(L_1/C - \omega^2 L_1 L_2) + jR(\omega(L_1+L_2) - 1/\omega C)}. \tag{2.276}$$

Oscillation will occur when the loop gain $A\beta$ is unity. Since $A$ is purely real, this means that $\beta$ must be real too. This implies

$$\omega(L_1+L_2) - 1/\omega C = 0. \tag{2.277}$$

The solution is

$$\omega = \sqrt{\frac{1}{C(L_1+L_2)}}, \tag{2.278}$$

At this frequency the feedback loop is, according to Eq. (2.276), $\beta = -L_2/L_1$. Thus, if our amplifier has a gain $A$ such that the total loop-gain is unity the Barkhausen criterion will be met; if

$$A\beta = \left(-\frac{R_\mathrm{f}}{R_1}\right) \times \left(-\frac{L_2}{L_1}\right) = 1, \tag{2.279}$$

the circuit will oscillate at the frequency given by

$$f = \frac{1}{2\pi}\sqrt{\frac{1}{C(L_1+L_2)}}, \tag{2.280}$$

Note that the output resistance of the opamp does not enter into the final results. It only has effect in the *quality* of oscillation; the lower $R$ the better ('sharper') the oscillation frequency.

# 3 | Physics

## 3.1 Introduction

In this chapter the physical aspects of instrumentation are treated. In other words, *how* is it that the signal — the information — is translated to and from the electronics world. As we have seen in the first chapter, this is the realm of transducers. A transducer 'maps' the physical parameter onto an electronic parameter, such as voltage, current, resistance or capacitance. We learned how to use – process – the latter in electronic circuits, subjects ranging from amplification to filtering, in the previous chapter. In this chapter we will answer some fundamental questions about the underlying physics of transducers. For example how it is that the resistance value of a resistor depends on temperature and why this dependence is fundamentally different, even in sign, when going from metallic resistors to semiconductor resistors. For that we have to go back to the basics and first determine what it is exactly 'current', and 'voltage', etc.

## 3.2 Physical background of conduction

Current is the passage of charge. It is as simple as that. The more charge passes through a wire, the higher the current. Current is like moving water in a river — the name current is even derived from this analogy. What is different between the water in a river and electric current in a wire is that water in a river is continuous and electric charge is quantized. We could imagine a river consisting of droplets of water. The 'droplets' for electric current are so tiny though, that we might as well see charge as something continuous. Yet, in some cases the quantization is important. Especially when the currents get small we can see the quantization effects and some forms of noise are directly caused by this quantization (Shot noise).

The smallest unit of charge is the elementary charge that has a value of $q = 1.60217733 \times 10^{-19}$ C. (The strangeness of value is due to the fact that current was known before the underlying idea of current being the flow of charge was discovered). Everything, from the smallest to the biggest thing in this universe has an integral multiple of this charge (excluding quarks, sub-particles of atomic

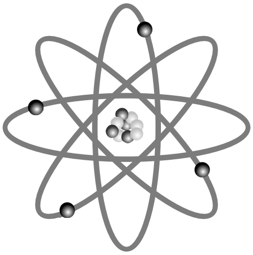


**Fig. 3.1**: Schematic representation of an atom of beryllium (Be). Inside the nucleus there reside 4 protons (dark) each with charge $+q$ and 5 neutrons (without charge). Around the nucleus is a cloud of electrons, each with charge $-q$. In a neutral atom the number of electrons and protons is equal and the total charge is zero

particles that can have multiples of one third of this elementary charge).

Upon closer inspection, charges can have both signs, positive and negative. When we zoom in to the atomic scale we see that the atom consists of a positively charged nucleus, consisting of protons with charge $+q$ and charge-neutral neutrons. This nucleus makes up most of the mass of atoms. Around it is a cloud of rapidly moving ultra-light electrons at a distance from the nucleus. Each electron carries a charge $-q$, and normally the number of electrons in an atom is equal to the amount of protons and the overall charge of an atom is zero. Obviously, moving atoms — objects in general, neutral objects that is — are not electrical current.

**Question**: How many electrons do we have approximately in our body? (Assume 70 kg, 90% water and 10% hydrocarbons of the [simplified] form $\mathrm{C_nH_{2n}}$).

**Answer**: Every proton is compensated by an electron. We just have to count the number of protons. Water is $H_2O$. One atom of hydrogen has one proton. One atom of oxygen has 8 protons and 8 neutrons. One molecule of water thus has 18 times the unified atomic mass unit (1 u = $1.6605402 \times 10^{-27}$ kg, see Chapter 1) and 8 + 2×1 = 10 protons (and electrons). With a carbon atom consisting of 6 protons and 6 neutrons, one moiety (group of atoms) of $CH_2$ a mass equal to (12 u) + 2×(1 u) = 14 u and 6 + 2×1 = 8 protons (and electrons). The total number of electrons thus becomes $2.5 \times 10^{28}$:

| Moiety | Total mass | Molecular mass | Electrons per moiety | Total electrons |
|---|---|---|---|---|
| $H_2O$ | 63 kg | 18 u | 10 | $2.34 \times 10^{28}$ |
| $CH_2$ | 7 kg | 14 u | 8 | $2.4 \times 10^{27}$ |
| | | | | Total: $2.6 \times 10^{28}$ |

It becomes interesting from a point of view of electronics when we separate the electrons from the nucleus. In that case, either the movement of free electrons or the remaining positively charged atom (ion) will represent current. In most electronics the moving charge consists of electrons, but this is not always the case. In a battery, for instance, the external current in the form of electrons from one pole to the other is matched by a current of ions inside the battery in the electrolyte such that neither pole accumulates charge. A battery always remains neutral overall. As we will see, this distinguishes it from a capacitor that becomes locally charged (one pole becomes positively charged while the other gets negatively charged, maintaining overall neutrality).

Now that we know that current is the movement of charge we can even make a numerical link between charge and current.

1 ampere of current is equal to the passage of 1 coulomb of charge per second

That puts things into a new perspective. If we have 1 ampere of current, really an astronomical amount of electrons pass every second through the wires, $N = (1\ \mathrm{C/s})/(q\ \mathrm{C/electron}) = 6.24 \times 10^{18}$ electrons/s. With these large numbers, it is meaningless to think of electrons as individual particles and our analogy of water running in a river is more adequate.

**Question**: How many electrons were flowing through my 50-watt television when I watched the football game yesterday?
**Answer**: If you are living in Europe, with 240 volt from the grid, the calculation is as follows: The grid in Europe has 240 volt voltage AC, but that is the effective voltage (that value that for a DC voltage would result in the same power). The amplitude of the applied voltage is a factor $\sqrt{2}$ larger, $V = 240\sqrt{2}\sin(2\pi f t)$, with $f$ the frequency (50 Hz). The current is the same sinusoid divided by the resistance of my television. With the power (product of $V$ and $I$) on averaging 50 watt, this resistance must be 1152 Ω. The current is thus of the form

$$I(t) = \frac{240\sqrt{2}}{1152}\sin(2\pi f t). \tag{3.1}$$

In half the cycle the current is positive, meaning that electrons flow through the wire into my TV from one side to the other, while in the other half of the cycle the direction is reversed and electrons flow back. The average current in either half of a cycle is ($T$ is the period of oscillation, $T \equiv 1/f$)

$$< I > \; = \; \frac{1}{T/2}\int_0^{T/2} \frac{240\sqrt{2}}{1152}\sin(2\pi f t)\mathrm{d}t \tag{3.2}$$

$$= \; \frac{2 \times 240\sqrt{2}}{1152\pi} = 0.188\ \mathrm{A}. \tag{3.3}$$

In 90 minutes that gives a total charge that passed equal to $Q = (90\ \mathrm{min}) \times (60\ \mathrm{s/min}) \times (0.188\ \mathrm{C/s}) = 1015$ coulomb. Divided by the elementary charge gives $N = Q/q = 7 \times 10^{21}$ electrons. Half of them in one direction and the other half in the other direction.

The next question is: what is it that makes current flow? Everything has a tendency to reach a state of minimal energy. Balls roll from the mountain. Raindrops fall from the sky. Etc. For electric current the charge wants to attain minimum electrical energy. The minimum energy for charges occurs when they are close together if they are of opposite sign and far apart when they are of same sign. (This is a difference with gravitational forces where all masses always attract each other; mass always has positive sign and forces are attractive).

The electric field at a distance $r$ of a charge $Q$ is given by

$$E(r) = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}. \tag{3.4}$$

The force acting upon a particle with charge $q$ is then this electric field times the charge $q$, $F = qE$ and points in the direction of the charge $Q$ if charges are different or away from it when charges are of same sign. We can thus define an energy of a particle, similar to gravitational forces: The energy difference of an object is the integral of force acting upon the object along a path. (Arbitrarily) defining the energy at $r = \infty$ as zero we get an electric energy of

$$U(r) = \int_{\infty}^{r} F(r')\mathrm{d}r' = \int_{\infty}^{r} qE(r')\mathrm{d}r' = -\frac{1}{2\pi\varepsilon_0}\frac{qQ}{r}. \tag{3.5}$$

The electric potential is defined as the energy per unit charge and is thus given by

$$V(r) = \int_{\infty}^{r} E(r')\mathrm{d}r' = -\frac{1}{2\pi\varepsilon_0}\frac{Q}{r}. \tag{3.6}$$

In conclusion: charge wants to 'annihilate', meaning that charges of opposite sign want to move closer together. If they have the possibility they will move closer together and this movement of charge constitutes current. The lowest energy is where the charges are as close together as possible (they cannot be in the same place, $r = 0$ for reasons of existence of forces other than electric). Associated to the energy is a 'potential', that is the energy 'per charge', The amount of electrical energy lost (and converted to another form of energy) when we move one coulomb (more precise, it is the energy of an infinitesimal charge divided by that charge $V = U/q$ for the limit of $q$ approaching zero).

So, in other words, electrical energy can be converted into other forms of energy, for instance light, or heat, or sound, or movement. The amount of energy available per charge is called the potential.

## 3.3 Classical physics of electronics

The properties and behavior of electronic devices can all be derived from the classical equations of Maxwell (see Table 3.I). While this is already a quite

**Table 3.I**: Maxwell equations. $\mathbf{E}$ is electric field, $\mathbf{D}$ is displacement, $\mathbf{P}$ is polarization, $\mathbf{H}$ is magnetic field, $\mathbf{B}$ is magnetic induction (or magnetic flux density), $\mathbf{M}$ is magnetization, $\mathbf{J}$ is current density, $\mu$ is permeability, $\varepsilon$ is permittivity (for both: subscript '0' means 'of vacuum'), $t$ is time

$$\begin{aligned} \text{Faraday's law}: \quad \nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t}, && (3.7) \\ \text{Ampere's law}: \quad \nabla \times \mathbf{H} &= \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J}, && (3.8) \\ \text{Gauss' law}: \quad \nabla \cdot \mathbf{D} &= \rho, && (3.9) \\ \nabla \cdot \mathbf{B} &= 0, && (3.10) \end{aligned}$$

$$\mathbf{B} \equiv \mu \mathbf{H} = \mu_0(\mathbf{H} + \mathbf{M}), \qquad (3.11)$$

$$\mathbf{D} \equiv \varepsilon \mathbf{E} = \varepsilon_0(\mathbf{E} + \mathbf{P}). \qquad (3.12)$$

difficult thing to do, imagine the scientists at the time of Maxwell. They reasoned the other way around: based on the measured behavior, what are the underlying mathematical equations? The resulting Maxwell equations cannot but be called brilliant. More so since it explains both static and dynamic electromagnetism and predicted also some till-then-unobserved phenomena. For years it was considered the final word on the subject, until Einstein showed that relativistic effects have to be taken into account in some cases. The Theory of Relativity, however, is not treated in this subject and we stick to classical physics of electronics. First we will treat the microscopic – physical – parameters and properties, like magnetic and electric fields, and then show how they are mapped to observable – macroscopic – properties, things like currents and voltages in electronic devices and components.

### 3.3.1 Microscopic electrical parameters

We are now ready to present some microscopic ('physical') and macroscopic ('electronic') electrical parameters. Starting with microscopic parameters. These are the quantities that are defined in 'any point in space', in contrast to the macroscopic quantities that belong to the 'overall device'. Of course, at the end they all boil down to physics, and are all based on the equations of Maxwell (Table 3.I). The first microscopic parameter is then

- Density of electrons, $n$ (unit: per cubic meter, $\mathrm{m}^{-3}$).

Then, the charge density is the electron density times the charge on each electron,

**Table 3.II**: Useful mathematical definitions and theorems. Divergence theorem: The volume integral of the divergence of a field is equal to the closed-surface integral of the normal component of that field. Kelvin-Stokes theorem: The surface integral of the curl of any vector field is equal to the closed contour-integral of the parallel component of that vector. This theorem helps us convert the derivative-forms of the Maxwell equations into the integral forms

$$\mathbf{F} = \begin{pmatrix} F_\mathrm{x}(x,y,z) \\ F_\mathrm{y}(x,y,z) \\ F_\mathrm{z}(x,y,z) \end{pmatrix}, \qquad \nabla \equiv \begin{pmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{pmatrix},$$

$$\text{Divergence}: \quad \nabla\cdot\mathbf{F} = \partial F_\mathrm{x}/\partial x + \partial F_\mathrm{y}/\partial y + \partial F_\mathrm{z}/\partial z, \tag{3.13}$$

$$\text{Curl}: \quad \nabla\times\mathbf{F} = \begin{pmatrix} \partial F_\mathrm{z}/\partial y - \partial F_\mathrm{y}/\partial z \\ \partial F_\mathrm{x}/\partial z - \partial F_\mathrm{z}/\partial x \\ \partial F_\mathrm{y}/\partial x - \partial F_\mathrm{x}/\partial y \end{pmatrix}. \tag{3.14}$$

$$\text{Divergence theorem}: \quad \iiint_\mathrm{V} (\nabla\cdot\mathbf{F})\,\mathrm{d}V = \oiint_\mathrm{S} \mathbf{F}\cdot\mathrm{d}\mathbf{S}. \tag{3.15}$$

$$\text{Kelvin Stokes theorem}: \quad \iint_\mathrm{S} (\nabla\times\mathbf{F})\cdot\mathrm{d}S = \oint_\mathrm{l} \mathbf{F}\cdot\mathrm{d}\mathbf{l}. \tag{3.16}$$

- Charge density, $\rho = -qn$ (unit: coulomb per cubic meter, $\mathrm{C/m^3}$).

Since the symbol $\rho$ is also used for resistivity, as we will see in a moment, it is better to forget this definition. Everywhere we need charge density, we will use the written-out form $-qn$.
Current is the passage of charge, thus if we imagine an area in space, see Figure 3.2, we can count how many electrons will be passing through it per second, $\mathrm{d}Q/\mathrm{d}t$. If the electrons are more abundant, or if they travel faster, or if we make the area larger, more will pass through the circle per second. So, the current through the circle is proportional to the speed of the charges, the density of the charges and the area. If we divide the current by the area, we can define

- Current density,

  $$\mathbf{J} = -qn\mathbf{v} \tag{3.17}$$

  (units: ampere per square meter, $\mathrm{A/m^2}$), where $\mathbf{v}$ is the average speed of the electrons. Both current and speed (velocity) are vectors and like all vectors in this book shown in bold font.

- Electric field, $\mathbf{E}$ (unit: volt per meter, V/m).

  From one of the Maxwell equations (Table 3.I), namely Equation 3.9, which states that the divergence of the displacement vector ($\mathbf{D} = \varepsilon\mathbf{E}$, with $\varepsilon$ the permittivity) is equal to charge density, we can derive an equation

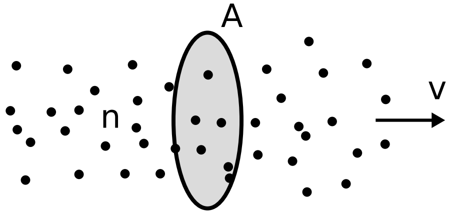


**Fig. 3.2**: Total current is the number of electrons passing through an area $A$ per second. This is equal to the product of area ($A$), charge density ($-qn$) and (average) speed $\mathbf{v}$. Current density is the current per area, thus $\mathbf{J} = -qn\mathbf{v}$ and has units A/m$^2$

that states that the integral of displacement over a surface S of a volume V is equal to the charge contained within that volume. For this we use the divergence theorem (Table 3.II), which states that the volume integral of a divergence of a field is equal to the surface integral of the normal of the field, applied to the vector field of electric field ($\mathbf{E} = \mathbf{D}/\varepsilon$) combined with the Maxwell equation of divergence of displacement (Eq. 3.9) it directly follows that

$$\oiint_{\mathrm{S}} \mathbf{E} \cdot \mathrm{d}\mathbf{S} = \frac{1}{\varepsilon} \iiint_{\mathrm{V}} \rho(x, y, z)\mathrm{d}x\mathrm{d}y\mathrm{d}z = Q/\varepsilon, \tag{3.18}$$

which is called Gauss' law. The closed-surface integral of (the normal component of the) displacement field is equal to the charge contained within the volume.

- An external electric field causes reorientation of internal electric dipoles that cause an internal electric field $\mathbf{P}$. This is called 'polarization'. The polarizability is parametrized in 'susceptibility', $\chi_{\mathrm{e}}$, a unitless quantity,

  $$\mathbf{P} = \varepsilon_0 \chi_{\mathrm{e}} \mathbf{E}. \tag{3.19}$$

  This combines with the electric field into the total displacement,

  $$\mathbf{D} = \varepsilon_0(\mathbf{E} + \mathbf{P}) = \varepsilon_0(1 + \chi_{\mathrm{e}})\mathbf{E} = \varepsilon\mathbf{E}, \tag{3.20}$$

  where $\varepsilon$ is called 'permittivity' (unit: farad per meter, F/m). It can be expressed relative to the permittivity of vacuum, defining a new unitless quantity $\varepsilon_{\mathrm{r}}$,

  $$\varepsilon = \varepsilon_{\mathrm{r}}\varepsilon_0. \tag{3.21}$$

  This relative permittivity $\varepsilon_{\mathrm{r}}$ is normally called 'dielectric constant' and is unity plus the electric susceptibility, $\varepsilon_{\mathrm{r}} = 1 + \chi_{\mathrm{e}}$.

- Electric force, $\mathbf{F} = Q\mathbf{E}$ (unit: newton, N). With $Q$ the charge (unit: coulomb, C). Remember that this is the only thing the old scientists in Maxwell's days had to work with, forces. The rest was deduced based on the observed forces.

- Electric potential, $V$ (unit: volt, V). In physics jargon, the potential is a mathematical aid to help us calculate things (note that it is not part of Maxwell's equations). More precisely, the electric potential is defined as the integral of the electric field along a path. The other way around, the electric field is a vector that is defined as the gradient of potential

  $$-\nabla V = \mathbf{E}. \tag{3.22}$$

  The minus sign comes from the fact that the electric field points 'downhill', i.e., to lower potentials. The inverse of the above equation gives us the voltage difference between two pints in space

  $$\Delta V = -\int \mathbf{E}(r) \cdot \mathrm{d}\mathbf{r}, \tag{3.23}$$

  integrating the electric field along a path, where the exact path is not important; only the end points matter. That is, the component of the electric field vector that is parallel to the path vector.

  A special form of Gauss' law occurs for the case where the system is symmetric, with all variables depending only on one coordinate. In this case Gauss' law (Eq. 3.9) becomes

  $$\frac{\mathrm{d}^2 V(x)}{\mathrm{d}x^2} = -\rho(x)/\varepsilon, \tag{3.24}$$

  which is Poisson's equation.

- Electric (potential) energy,

  $$U = QV \tag{3.25}$$

  (unit: joule, J).

- Conductivity,

  $$\sigma = \mathbf{J}/\mathbf{E} \tag{3.26}$$

  (unit: ampere per volt meter, A/Vm).

- Resistivity,

  $$\rho \equiv 1/\sigma \tag{3.27}$$

  which is thus equal to $\rho = \mathbf{E}/\mathbf{J}$ and has units Vm/A or $\Omega$m

- Mobility. As we will see, in contrast to electrons in vacuum, the electrons in materials in the influence of an electric field are not accelerated for ever. Because of interactions with the surroundings, the average speed of electrons rapidly thermalizes and is proportional to the electric field. This proportionality is what is called mobility of the charges, which can thus be defined as the ratio of (average) speed and electric field,

  $$\mu \equiv \frac{\mathbf{v}}{\mathbf{E}}. \tag{3.28}$$

While this symbol is equal to the one used for permeability of the material, and to the prefactor one-millionth, it will be clear enough when which one is meant.

We can now find an equation for the conduction in metals based on this classical physics picture. (Later we will also see the quantum mechanics picture). We use here the Drude model. If we follow an individual electron we see that it frequently and randomly bumps into the nuclei. At each bump, the speed of the electron is randomized to a value $v_0$, see Figure 3.3. After that, the electron is accelerated in an external electric field $E$, the force being $F = qE$, the acceleration $a = F/m_\mathrm{e} = qE/m_\mathrm{e}$ and the speed is thus found to be $v(t) = v_0 + qEt/m_\mathrm{e}$. Since the speed $v_0$ is random and averages out to zero, the average speed of the electrons is given by the average time $\tau$ between collisions,

$$<v> = \frac{qE\tau}{m_\mathrm{e}}, \tag{3.29}$$

with $\tau$ the mean free time between collisions (which is half the average time between collisions; any particular electron is statistically halfway between collisions). The current density $J$ is the charge density $qn$ ($n$ is the electron density, $q$ is the elementary charge) times the average speed, thus

$$J = qn<v>= \frac{nq^2\tau E}{m_\mathrm{e}}. \tag{3.30}$$

The current density is also per definition the conductivity $\sigma$ times the electric field, we can thus find an expression for the conductivity,

$$\sigma \equiv \frac{J}{E} = \frac{nq^2\tau}{m_\mathrm{e}}. \tag{3.31}$$

We see that the conductivity increases when the time between collisions increases. The resistivity, $\rho \equiv 1/\sigma$ increases when the collisions become more frequent. An important finding, since it quite adequately explains the behavior of metallic conduction. An important thing to note is that if the temperature increases, the atoms start vibrating more and the possibility of an electron hitting one of them increases. Thus, for increasing $T$ we can expect a decreasing $\tau$ and conductivity $\sigma$. Metal resistors have a behavior that in electronic instrumentation is called PTC, which stands for positive temperature coefficient, $\mathrm{d}R/\mathrm{d}T > 0$.

### 3.3.2 Microscopic magnetic parameters

The magnetic properties of space can also be summarized in a couple of parameters, all based on, or summarized in, Maxwell's equations:

- The main magnetic parameter is magnetic field, the vector field $\mathbf{H}$ (unit: A/m). A macroscopic current of 1 ampere in an infinite wire causes a

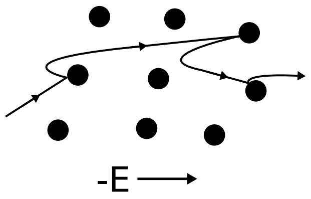


**Fig. 3.3**: Conduction in metals according to the Drude model. Electrons are bumping into atoms which randomizes their speed. After each bump, the electrons are accelerated in an electrical field until the next bump. An average speed in the direction of the electric field thus results

magnetic field of 1 A/m at 1 meter distance. This is the direct result of Ampere's law (Eq. (3.8)), as we will see later. The magnetic field is a microscopic parameter; every point in space has a magnetic field value and direction.

- Just like the electric field **E** can polarize space, expressed in **P**, the two combining into displacement **D**, so can magnetic field **H** magnetically polarize a material, or 'magnetize' it. The resulting internal magnetic field is called the magnetization vector field **M**, which combines with the magnetic field into magnetic flux density **B**,

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}) = \mu\mathbf{H}, \tag{3.32}$$

which has a unit of tesla (T, translated in SI into newtons per meter per ampere, N m$^{-1}$ A$^{-1}$) or weber per square meter (Wb/m$^2$). And note the definition of a new parameter $\mu$ which is called the 'permeability', with the subscript '0' denoting 'of vacuum'. (unit: henry per meter, H/m, which in SI units is N/A$^2$. Note this $\mu$ is not the same as charge mobility mentioned before). This field **B** is what causes forces on moving charges.

- The magnetizability of a material is parametrized in a unitless quantity called 'magnetic susceptibility',

$$\mathbf{M} = \mu_0\chi_\mathrm{m}\mathbf{H}. \tag{3.33}$$

Also here we can express the permeability relative to the permeability of vacuum,

$$\mu = \mu_\mathrm{r}\mu_0, \tag{3.34}$$

and thus

$$\mu_\mathrm{r} = 1 + \chi_\mathrm{m}. \tag{3.35}$$

- Like the electric potential was defined such that the electric field is the gradient (divergence) of this potential, $\mathbf{E} = -\nabla V$, a vector magnetic potential field **A** can be defined such that the magnetic flux density **B** is the curl of this vector potential,

$$\nabla \times \mathbf{A} = \mathbf{B}. \tag{3.36}$$

Note that this definition of **A** follows from, or is the direct result of, the condition of Maxwell's equations, namely the condition that the divergence of magnetic flux density is zero, $\nabla \cdot \mathbf{B} = 0$ (Eq. (3.10)). For the divergence of the curl of a vector field is zero, $\nabla \cdot (\nabla \times \mathbf{A}) = 0$. This because for any physical vector field $\partial^2 F_1/\partial x \partial y = \partial^2 F_1/\partial y \partial x$, etc.; the order of derivation does not matter.

- Maybe a more useful mathematical parameter, to help us calculate things, is magnetic flux, which is the area integral of magnetic flux density $B$,

$$\Phi = \iint_{\mathrm{S}} \mathbf{B} \cdot \mathrm{d}\mathbf{S}, \tag{3.37}$$

which has unit weber (Wb) or tesla square meter (T m$^2$). So we can define an area and imagine, 'count', the magnetic flux lines passing through the area to find the total flux $\Phi$. For many magnetic calculations this comes in handy, as we will see later.

For magnetic parameters we can also define relations. Let us first convert Ampere's law of the Maxwell equations, Eq. (3.8), to a version in integral form. We do that by taking a surface element on both sides and then integrating over a certain area S,

$$\iint_{\mathrm{S}} (\nabla \times \mathbf{H}) \cdot \mathrm{d}\mathbf{S} = \iint_{\mathrm{S}} \mathbf{J} \cdot \mathrm{d}\mathbf{S}. \tag{3.38}$$

Now using Kelvin-Stokes theorem, Eq. (3.16), on the left side and noting that on the right side the integral of the current density is the total current passing through the surface, or, in other words, is enclosed by the loop,

$$\oint_{\mathrm{l}} \mathbf{H} \cdot \mathrm{d}\mathbf{l} = I_{\mathrm{enclosed}}. \tag{3.39}$$

This is the integral form of Ampere's law. We can use it to calculate the behavior of macroscopic devices and electronic components. Note that, technically speaking, current $I$ is a macroscopic property – of a device such as a wire – while magnetic field **H** is a microscopic property – every point in space has a value. We have here a mixed equation.

### 3.3.3 Kinetics of charge; continuity

There are now in general three ways the density of charge somewhere in space can change:

- **Drift**. As discussed before, drift is the flow of charge as the result of electrical forces. We can define a drift current density by combining the current density found before (Eq. 3.17) and the definition of mobility (Eq. 3.28),

  $$\mathbf{J} = -qn\mu\mathbf{E}. \tag{3.40}$$

  If the current is constant, for instant in a wire carrying a current, then charge density will not change, since per second as much charge moves in as out any point in space. If on the other hand current does not balance, with more charge per second coming in or going out, with the incoming current differing from the outgoing current, charge accumulates or disappears. The total charge in the universe is conserved, the accumulation of charge in any point in space is thus necessarily equal to the divergence of current at that point,

  $$\frac{\partial\rho}{\partial t} = -\nabla\cdot\mathbf{J}. \tag{3.41}$$

- **Creation/annihilation**. Overall, electric charge cannot be created or annihilated, Yet, locally we can assume charge sources. For instance in a current source we can imagine charge being created on one pole and annihilated on the other pole. This while we know that internally, inside the source, the charge continuity will maintain. When analyzing the functioning of a device we can say things like "electrons are injected at the cathode and collected at the anode", as if they miraculously appear there out of nowhere, or disappear into nowhere.

  Another reason for still using the creation or annihilation formalism is the analysis of currents using the concept of holes. As we will see, in semiconductors charge exists as – or better to say, can be analyzed very elegantly using – electrons and holes having negative and positive charge respectively. While the total charge cannot change, as this is a conservation law of the universe, the individual contributions *can* change. Electrons can 'recombine' with holes and they both disappear (in the formalism; in reality no electron has disappeared). In this way, when we consider the partial currents of holes and electrons, indeed charges can disappear. We can for instance have the situation that electrons have a certain 'lifetime' $\tau$ defining a rate by which they recombine with holes, and we will use equations like $\mathrm{d}n/\mathrm{d}t = -n/\tau$.

- **Diffusion**. Nature has a tendency for chaos. This expresses itself in the tendency of densities to equalize in space. When we start with a certain disequilibrium, nature will restore the equilibrium in a process that is called diffusion and to this diffusion a current is associated. In fact, the current is proportional to the gradient of density, so for our electric charge we have an electric current equal to

  $$\mathbf{J} = -D\nabla\rho, \tag{3.42}$$

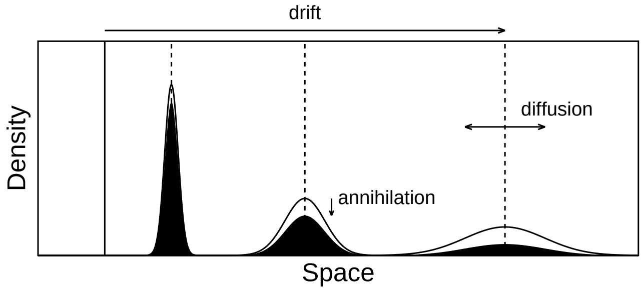


**Fig. 3.4**: Schematic drawing showing the evolution of a distribution when three processes actuate on it. Starting with a delta-Dirac distribution (shown as a vertical line), drift dislocates the distribution, diffusion spreads it, and annihilation lowers it

with $D$ the diffusion coefficient (unit: $\mathrm{m}^2/\mathrm{s}$). Note that the current is in the direction of diminishing density. For instance an electronic current is in the direction of increasing electron density

$$\mathbf{J}_{\mathrm{n}} = qD_{\mathrm{n}}\nabla n, \tag{3.43}$$

with $D_{\mathrm{n}}$ the electron diffusion coefficient.

In his Brownian-motion experiment of smoke particles, Einstein found that there is a correlation between diffusion coefficient and mobility, and this is now called Einstein's equation,

$$D = \frac{\mu kT}{q}. \tag{3.44}$$

Note that diffusion does not necessarily generate a macroscopic current. It can just spreads the charge in all directions, without any preferential direction. Electric field always causes a net current which we called 'drift current' above. An example is a delta distribution at $t = 0$. The diffusion spreads the distribution, with as much moving one direction as moving in the opposite direction, the drift dislocates the distribution in a certain direction, and annihilation reduces the integral of the distribution. See Figure 3.4.

## 3.3.4 Macroscopic parameters

### Current

Where current density is a microscopic parameter, the more famous current is a macroscopic parameter, a quantity that belongs to a device with its dimensions. It easily follows that current is current density $\mathbf{J}$ multiplied by the surface area

**S** through which this current density passes,

$$I = \iint_{\mathrm{S}} \mathbf{J} \cdot \mathrm{d}\mathbf{S}. \tag{3.45}$$

It can also be defined as the amount of charge that goes through the area per unit time. If we label this area $A$, we define the current as a time derivative,

$$I \equiv \frac{\mathrm{d}Q}{\mathrm{d}t} = -qn\mathbf{v} \cdot \mathbf{A} = \mathbf{J} \cdot \mathbf{A}. \tag{3.46}$$

As an example, a device with length, width and height given by $L$, $W$, and $h$, respectively, through which passes a current density along $L$ equal to $J$ will have a current equal to $I = JWh$ (Fig. 3.5).

### Resistance

Now we are ready to define some macroscopic (electronic) parameters. As we will see, these can be described in terms of microscopic (physical) parameters. To start, resistance, according to Ohm's law, is by definition the ratio of voltage and current,

$$R \equiv \frac{V}{I} \tag{3.47}$$

(unit: ohm, Ω. See Chapter 2). Conductance is per definition the inverse of resistance,

$$G \equiv \frac{1}{R} = \frac{I}{V} \tag{3.48}$$

(unit: siemens, S ≡ 1/Ω and sometimes called mho). Both device parameters, conductance and resistance, can easily be calculated in terms of the physical parameters, conductivity ($\sigma$) and resistivity ($\rho$). As an example, if we have a bar of material with length $L$, width $W$ and height $h$, see Figure 3.5, and we apply a bias voltage $V$, the electric field is given by $E = V/L$. The current density is then given by (Eq. 3.26) $J = \sigma E = \sigma V/L$. The cross section of the device is equal to $A = Wh$ and the current is thus $I = AJ = \sigma VWh/L$. Finally, the conductance, being the ratio of current and voltage is given by

$$G = \sigma\frac{Wh}{L}, \tag{3.49}$$

and resistance by

$$R \equiv \frac{1}{G} = \frac{L}{\sigma Wh} = \frac{\rho L}{Wh}. \tag{3.50}$$

We can now also understand physically the behavior of resistances when placed in series or parallel. For instance, when we place two equal resistances in series, effectively the length of the resistance is doubled and Equation 3.50 tells us that then the resistance doubles. On the other hand, when placing them in parallel, the cross-section $A = WL$ is doubled and the resistance is halved.

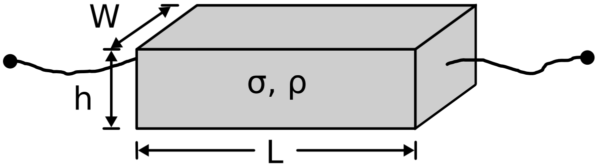


**Fig. 3.5**: Bar resistor. A bar of material, with dimensions $L$, $W$ and $h$ made of material with conductivity σ (and resistivity $\rho = 1/\sigma$) has a resistance equal to $R = \rho L/Wh$

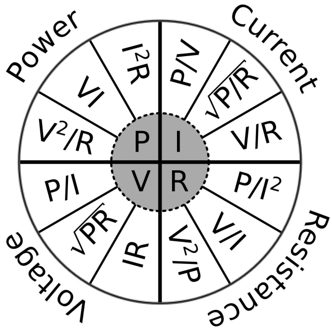


**Fig. 3.6**: Ohm's Disc: Various ways of expressing power, resistance, current and voltage in terms of each other, including Ohm's law ($R = V/I$) at the bottom-right

## Power

If every charge d$Q$ going from a potential energy loses an amount of energy given by d$U$ = d$QV$ (Eq. 3.25), then the change of energy per second – the power $P$ – is given by

$$P \equiv \frac{\mathrm{d}U}{\mathrm{d}t} = V\frac{\mathrm{d}Q}{\mathrm{d}t}. \tag{3.51}$$

But this contains the definition of current (Eq. 3.46), and power thus becomes current times voltage (drop):

$$P = VI. \tag{3.52}$$

With the help of this definition and Ohm's law, any member of the quartet $P$, $R$, $I$, $V$, can always be expressed in terms of two of the other three, see Figure 3.6.

## Capacitance

Capacitance is per definition the macroscopic capacity to store charge per voltage,

$$C \equiv \frac{Q}{V} \tag{3.53}$$

(unit: coulomb per volt, which is per definition farad, F). The above equation is the equivalent of Ohm's law for charge. Compare Equation 3.47 with Equation 3.53. Note, however, that in Ohm's law the voltage is in the *nominator*, while in the capacitance law it is in the *denominator*. This is an inconsistency in electronics; it would have made much more sense if Georg Simon Ohm had worked with conductance instead of resistance. Or if the concept of reciprocal capacitance was more commonly used. We are now left with some confusing nomenclature.

For a bar-capacitor, similar to the bar resistance above, the capacitance can easily be calculated. Assuming that it consists of metal plates separated by a bar of insulating material ($\sigma = 0$) with permittivity $\varepsilon$, see Figure 3.7. Imagine we put a charge of $Q$ on one of the metal plates And, because the total charge of any device must be neutral, the charge on the opposite metal plate is $-Q$. The charge will cause an electric field pointing from the positive charge to the negative charge. If we ignore boundary effects – if we imagine the device stretched very far in directions y and z – the electric field is perpendicular to the surface of the plates, along x. If we assume the material to be homogeneous, with constant $\varepsilon$, the electric field must be constant from one side to the other. We can now use Gauss' law, Eq. (3.18), that stated that the closed integral of electric field over a surface of a volume is equal to the charge contained within that volume, and analyze the situation of Figure 3.7. We choose a box, as indicated, with the bottom face fixed below the bottom plate of the capacitor. The top face of the box we will slowly raise. Because the electric field is always pointing in the direction x, the side panes of the box do not contribute to the integral (because $\mathbf{E}\cdot\mathrm{d}\mathbf{S}$ is always zero there, because $\mathbf{E}$ is in the plane of $S$) and we thus only have to integrate the bottom and top plate of our imaginary box, where $\mathbf{E}\cdot\mathbf{n} = |E|$. While we maintain the bottom face below the plate of our capacitor, we slowly raise the top face. When the face is still below the bottom plate of our capacitor, the charge contained within is zero. And thus the field must be equal on the bottom and top faces of our box, because the two halves of the integral have to cancel. To simplify, we set this field arbitrarily to zero, as if our capacitor is the only thing in the universe. We raise the top face of our box, and when it includes the bottom plate of the capacitor, at $x = 0$, the charge contained within is suddenly equal to $-Q$. The area of the integral is $L \times W$ (outside the capacitor the field is zero since we assumed no boundary effects). This area multiplied by the (constant) electric field $E$ is then equal to $-Q/\varepsilon$. In other words, the electric field is

$$E = -\frac{Q}{\varepsilon WL}. \tag{3.54}$$

Raising the top face of our box changes nothing – and the electric field remains the same at this value – until we reach the top plate of the capacitor, at $x = h$, after which the total charge is zero and the electric field drops back to 0. Setting (arbitrarily) the potential of the bottom plate of the capacitor zero, we find the

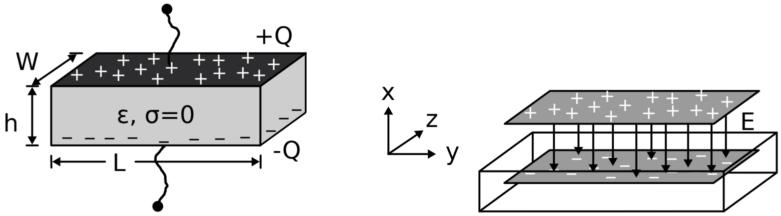


**Fig. 3.7**: Bar capacitor. Left: A bar of insulating material material ($\sigma = 0$), with dimensions $L$, $W$ and $h$ made of material with permittivity $\varepsilon$ placed between two metal plate electrodes has capacitance equal to $C = \varepsilon WL/h$. Right: Electric field and integration box used to find relation between voltage and charge and a capacitance value of $C = \varepsilon WL/h$

potential at any position $x$ as the integral of the electric field as

$$V(x) = -\int_0^x E(x')dx' = \frac{xQ}{\varepsilon WL}, \tag{3.55}$$

and the potential of the top plate thus, substituting $x = h$,

$$V(x = h) = \frac{hQ}{\varepsilon WL}. \tag{3.56}$$

The capacitance is per definition the ratio of charge per voltage (Eq. 3.53), and is then

$$C \equiv \frac{Q}{V} = \frac{\varepsilon A}{h}, \tag{3.57}$$

with $A = WL$ the area of a capacitor plate. Note that $Q$ is the charge on a single plate. The total charge in a capacitor is always zero!

**Question**: The famous Leyden jar (original 'Leidsche Fles'), see Figure 3.8, is a capacitor made of a glass bottle filled with conductive water in which an electrode is placed. The other electrode is something conductive placed on the outside of the bottle, for instance a hand. The insulator is the glass wall of the bottle. What is approximately the capacitance of a Leyden jar?
**Answer**: The area of the hand is about $A = 75$ cm$^2$. The dielectric constant of glass is $\varepsilon_\mathrm{r} = 4.7$ ($\varepsilon = \varepsilon_\mathrm{r}\varepsilon_0 = 4.7 \times 8.85418 \times 10^{-12}$ F/m $= 4.16 \times 10^{-11}$ F/m). The thickness of the glass wall is about $h = 3$ mm. Eq. 3.57 thus yields $C = 100$ pF. If the bottle is coated with a metal on the outside, the entire area of the bottle counts, instead of just the hand. We can expect up to one order of magnitude higher capacitance in such a bottle.
**Question**: What happens if we remove our hands?
**Answer**: Imagine we store with 1 V 100 pC in the bottle (Eq. (3.53)). If we first remove our top hand, charge cannot go out of the bottle.

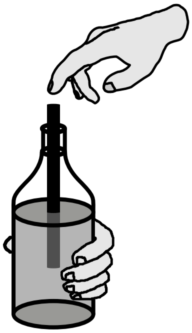


**Fig. 3.8**: Leidse Fles (Leyden jar), effectively a capacitor

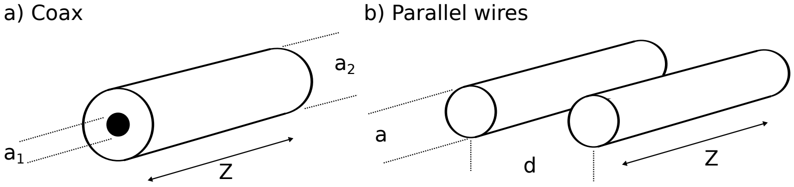


**Fig. 3.9**: a) Coax cable and b) parallel wires cable

Removing the other hand as well will increase the distance $h$, from 3 mm maybe to 3 m. The capacitance with thus go down a factor 1000. With the same amount of charge, the voltage will go up the same factor 1000, from 1 volt to 1000 volt.

Other configurations can also be easily calculated. Table 3.III shows capacitance expressions for some geometries.

We can calculate the energy of a charged capacitor. We start with an empty capacitor whose energy we set to zero and slowly add tiny charges, filling it up unto $Q$. The additional energy d$U$ when we add an infinitesimal incremental charge d$Q$ to the charge already inside the capacitor is (Eq. 3.25)

$$\mathrm{d}U = V\mathrm{d}Q. \tag{3.58}$$

But, with the definition of capacitance (Eq. 3.53), the voltage is equal to $V = Q/C$, thus the above equation reduces to

$$\frac{\mathrm{d}U}{\mathrm{d}Q} = \frac{Q}{C}. \tag{3.59}$$

Integrating and once again substituting the capacitance definition gives

$$U = \frac{Q^2}{2C} = \frac{1}{2}CV^2. \tag{3.60}$$

**Table 3.III**: Capacitances for some geometries. $Z$ is length, $A$ is area, $a$ is wire radius ($a_1$ is inner, $a_2$ is outer radius), $d$ is wire or plate (inter)distance

| **Geometry** | **Capacitance** | **Comment** |
|---|---|---|
| Parallel plates | $\frac{\varepsilon A}{d}$ | Equation 3.57<br>Figure 3.7 |
| Coax cable | $\frac{2\pi\varepsilon Z}{\ln(a_2/a_1)}$ | Exercise 2<br>Figure 3.9a |
| Parallel wires | $\frac{\pi\varepsilon Z}{\cosh^{-1}(d/a)}$ | Figure 3.9b |
| Concentric spheres | $\frac{4\pi\varepsilon}{1/a_1 - 1/a_2}$ | |

$\cosh^{-1}(x) = \ln(x + \sqrt{x^2 - 1})$

**Question**: What is the energy of a 1 nF capacitor charged to 10 V?
**Answer**: Substituting 1 nF and 10 V in Eq. 3.60 gives $U = 50$ nJ.

Now that we know what is the macroscopic energy of a charged capacitor, we can find the microscopic parameter of energy density, simply by dividing this energy by the volume. As we will see, the energy is related to the electric field. Assuming the electric field is constant between the plates and zero elsewhere, the energy density is the energy of the capacitor divided by its volume $WLh$. Using Equation 3.60, and $V = Eh$

$$u \equiv \frac{U}{\text{volume}} = \frac{CV^2/2}{WLh} = \frac{\varepsilon WL}{2h}\frac{(Eh)^2}{WLh} = \frac{1}{2}\varepsilon E^2. \tag{3.61}$$

The energy density is proportional to the square of electric field. This is actually a well known fact from electrodynamics. Our approximation actually resulted in an accurate expression. The energy of a device is the space integral of the above equation.

Also for the capacitance we can easily understand the behavior of capacitors placed in series or parallel. When placing two equal capacitors in parallel, the total area doubles, Equation (3.57), and the capacitance doubles. On the other hand, if we place them in series, effectively a new capacitor is created with double the height $h$, but with a metal plate somewhere in the middle. Since this plate is uncharged, it is effectively not there (since charge cannot reach there through the insulating medium, $\sigma = 0$), and we wind up with a simple capacitor of double the height and thus half the capacitance.

In most cases, instead of the static capacitance given above, it is more relevant for the working of the circuit to use a dynamic definition

$$C_{\text{AC}} \equiv \frac{\mathrm{d}Q(V)}{\mathrm{d}V}, \tag{3.62}$$

because it is the *current* that flows in and out a capacitor when we *change* the voltage $V(t)$ that counts for the working of a circuit. The total amount of charge inside the capacitor is irrelevant. It is like a black box; we do not see what is inside, and the only observable we have is the external current. It is easy to show that, in the case the total charge being a linear function of voltage ($Q \propto V$), the static and dynamic definitions are the same,

$$\begin{aligned} I(t) &= \frac{\mathrm{d}Q(t)}{\mathrm{d}t} = \frac{\mathrm{d}[C(V)V]}{\mathrm{d}t} = C(V)\frac{\mathrm{d}V}{\mathrm{d}t} + V\frac{\mathrm{d}C(V)}{\mathrm{d}t} \\ &= \left(C(V) + V\frac{\mathrm{d}C(V)}{\mathrm{d}V}\right)\frac{\mathrm{d}V}{\mathrm{d}t} \end{aligned} \tag{3.63}$$

which reduces to

$$I(t) = C\frac{\mathrm{d}V}{\mathrm{d}t}, \tag{3.64}$$

in case $C$ is constant. The implications of the above equation, in terms of signal phase and amplitude, and how it can be used to build circuits (for example filters) has been explained in the chapter on electronics (Chapter 2).

A special case of voltage-dependent capacitance is the varicap, a solid-state capacitor whose value is proportional to $1/\sqrt{(V - V_\mathrm{i})}$ and is based on the depletion zone of a reverse-biased junction diode, to be discussed later.

An additional effect for capacitors is that their capacitance value can (and often does) depend on the applied frequency. This is due to the nature of the permittivity $\varepsilon$ of the insulating material. Especially at higher frequencies the materials get 'lossy', meaning that the capacitor starts more and more behaving like a resistor, with the current in phase with the voltage, and its amplitude proportional to the voltage instead of its derivative. At very high frequencies, the resistance value is so low that it can be considered zero. This is a physics effect and should not be confused with the electronic effect that a capacitor, in series with a resistor, can be considered a short circuit for frequencies far above the cut-off frequency given by $\omega_\mathrm{c} = 1/\tau = 1/RC$, as explained in the chapter on electronics, an effect that takes place even if $C$ is constant. The physics effect tells us that the capacitor is a short circuit at high frequencies *irrespective* of the resistance connected to it.

There are various types of capacitors. Apart from the varicaps mentioned above, the general aim is to have a flat spectrum of the capacitance. Another requisite might be a high maximum allowed voltage (high breakdown voltage), or small size, or high operating temperatures. Apart from that, there always exists the need to make the device as cheap as possible. For these reasons, there exist a range of types of capacitors, each with their advantages and trade-offs. See Table 3.IV for a list. To increase the capacitance, while maintaining the size, the insulator can be replaced with an insulator of high dielectric constant. Apart from that, it allows for approximating the metal plates easily without the electrodes actually touching, thus even further increasing the capacitance. These insulating materials – dielectrics – can be anything ranging from plastics to glass, ceramics, mica and even paper, air or vacuum.

**Table 3.IV**: Types of capacitors

| Type | Application |
|---|---|
| Air | High frequency. Low loss |
| Ceramic | High frequency |
| Electrolytic | Low Frequency. Not heat resistant |
| Solid state | (Heat) robust |

The most often used capacitors are ceramic. They consist of foil that is either rolled-up, or interleaved (rectangular boxes) to have maximum area for minimal size. Once can think of a metal-coated foil.

In electrolytic capacitors, one of the electrodes is replaced by a conductive (ionic) solution and that makes them good for high-current low-frequency applications such as filters in power supplies. Care has to be taken polarizing these capacitors, since a persistent reverse bias will induce chemical reactions in the capacitor and will sooner or later blow it up. They are also very intolerant to heat; modern day electronics nearly always fail in the electrolytic capacitors of the power supplies. You can recognize such failed capacitors by their popped up belly. (Impress your friends and family by fixing their electronics in mere minutes!)

Solid state capacitors look very much like electrolytic capacitors, but they do not have the 'rupture cross' on top. That is because they do not rupture. They can stand heat much better. They are often fully encased in metal.

In some cases we want both high and low frequency behavior and it is not uncommon to use capacitors of two types in parallel, for instance in eliminating noise of the power supply. Most advanced electronic instrumentation circuits use this double-capacitor technique, with a electrolytic capacitor in parallel with a ceramic capacitor.

### Inductance

The electronic parameter inductance, which is caused by the magnetic behavior of space like capacitance was caused by its electric behavior, is per definition the ratio between voltage and changes of current,

$$\begin{aligned} L &\equiv \frac{V(t)}{\mathrm{d}I(t)/\mathrm{d}t}, \\ V(t) &= L\frac{\mathrm{d}I(t)}{\mathrm{d}t}, \end{aligned} \tag{3.65}$$

with the second form more conventionally found in the literature. We see that all three major electronic parameters, resistance, capacitance and inductance are defined as ratios of voltage and (derivatives of) charge, see Table 3.V.

The inductance can also be defined in terms of microscopic physical quantities. To understand this, we have to go back to the Maxwell equations. If we look at the definition of inductance, we see that it is defined in terms of voltage

**Table 3.V**: The three main electronic parameters are all based on the behavior of charge

| A | = | B | Ratio B:A |
|---|---|---|---|
| $Q$ | | $V$ | $1/C$ (Capacitance) |
| $\mathrm{d}Q/\mathrm{d}t$ | $I$ | $V$ | $R$ (Resistance) |
| $\mathrm{d}^2Q/\mathrm{d}t^2$ | $\mathrm{d}I/\mathrm{d}t$ | $V$ | $L$ (Inductance) |

and current changes. Voltage was the path-integral of electric field, electric field is connected to changes of magnetic field magnetic field by Faraday's law, while magnetic fields are caused by currents as stated in Ampere's law. So, we have the feeling that we have all the information needed to calculate the inductance for any configuration.

Let's start with Faraday's law of Eq. (3.7) and surface-integrate it on both sides:

$$\iint_{\mathrm{S}} (\nabla \times \mathbf{E}) \cdot \mathrm{d}\mathbf{S} = - \iint_{\mathrm{S}} \left( \frac{\partial \mathbf{B}}{\partial t} \right) \cdot \mathrm{d}\mathbf{S}. \tag{3.66}$$

If we apply the Kelvin-Stokes theorem to the left side and use the definition of magnetic flux of Eq. (3.37) to the right side, using the property that the integral of the derivative is the derivative of the integral, we get

$$\oint_{\mathrm{l}} \mathbf{E} \cdot \mathrm{d}\mathbf{l} = -\frac{\mathrm{d}\Phi}{\mathrm{d}t}. \tag{3.67}$$

This says that going around a closed path and integrating the induced electric field is equal to the speed of changes of the magnetic flux going through the area enclosed by the path. That is, if we cut open the path, this results at that point in a measurable voltage, which is often called the electromotive force (EMF). (Note that it is not a force, but a voltage and has unit volt, V). Faraday's law of electromagnetic induction,

$$V_{\mathrm{EMF}} = -\frac{\mathrm{d}\Phi}{\mathrm{d}t}. \tag{3.68}$$

This voltage is caused by changes of an external field that passes through the loop and is used in electricity generators, where for instance a turbine moves permanent magnets whose changing magnetic fields are then picked up by coils that generate this voltage and current.

The magnetic field can also be caused by the existing currents in the wire itself. Changing the currents in the loop will cause changes of magnetic fields which are then picked up by the loop and translated into an EMF that tries to annul changes in current, i.e., a coil has a certain inertia. This is modeled by the inductance parameter $L$ and can be calculated for each configuration. If we know how flux $\Phi$ depends on current we can substitute this in the definition of inductance,

$$L = -V_{\mathrm{EMF}}/(\frac{\mathrm{d}I}{\mathrm{d}\Phi}\frac{\mathrm{d}\Phi}{\mathrm{d}t}) = \frac{\mathrm{d}\Phi(I)}{\mathrm{d}I} = \frac{\Phi(I)}{I}. \tag{3.69}$$

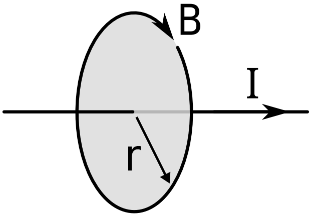


**Fig. 3.10**: Magnetic field at a distance $r$ away from a straight wire carrying current $I$. Eq.(3.73)

(In the last step we assume that the field and flux depend linearly on current). Lenz's law states that the voltage sign is such that the resulting current will oppose the changes in current that caused it. If the current decreased, it will try to increase the current.

The magnetic field $H$ and $B$ and thus magnetic flux $\Phi$ are determined by the current density through Ampere's law (Eq. (3.8)). Knowing the current in a structure will give us the magnetic field and thus magnetic flux. Substituting this into the equation above will give us the inductance $L$.

Let us first consider the magnetic field formed by a current through a straight wire (Figure 3.10). Ampere's law (Eq. 3.8) links the curl of magnetic field $H$ to the current density. If we integrate this equation in the static form (no time derivatives) over a surface S we get

$$\iint_S \left(\nabla \times \frac{\mathbf{B}}{\mu}\right) \cdot \mathrm{d}\mathbf{S} = \iint_S \mathbf{J} \cdot \mathrm{d}\mathbf{S}. \tag{3.70}$$

We can now use Kelvin-Stokes theorem that states that for any vector field $F$, the surface integral of its curl is equal to the closed contour integral of the vector, see Equation 3.16. For our magnetic inductance it means that

$$\oint_l \mathbf{B} \cdot \mathrm{d}\mathbf{l} = \mu \iint_S \mathbf{J} \cdot \mathrm{d}\mathbf{S} = \mu_0 I, \tag{3.71}$$

which is Ampere's law in integral form: The integral of magnetic field along a closed path, the edge of a surface, is equal to the total current passing through the surface. We can apply this now to a path along a circle with radius $r$ that encompasses the wire with a current $I$. For symmetry reasons, the magnetic field is constant on this circle, and always parallel to the line segment of the circle, $B \cdot \mathrm{d}l = B\ \mathrm{d}l$,

$$2\pi r B(r) = \mu I, \tag{3.72}$$

and we see that the magnetic field at a distance $r$ of a wire with current $I$ is given by

$$B(r) = \frac{\mu}{2\pi r} I. \tag{3.73}$$

A wire by itself does not have inductance, just like a single metal plate does not have capacitance. For a capacitance, the plate or any object always has to

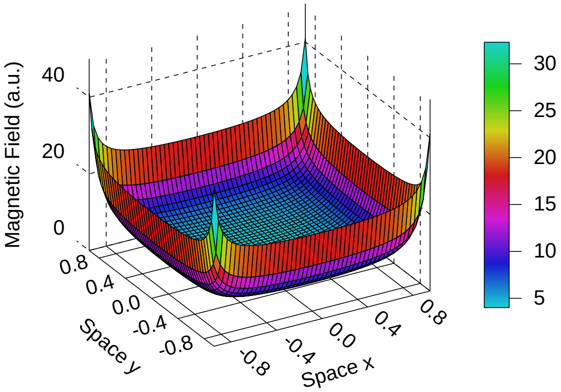


**Fig. 3.11**: Magnetic field of a square loop of 2 m × 2 m. At the wires the field goes to infinity

be considered with respect to something else, for instance a second plate. A lone plate has zero capacitance (any minute charge put on a two-dimensionally-infinite plate causes a constant field and infinite voltage at infinity, and thus $C = Q/V = 0$). Likewise, it does not make sense to talk about the inductance of a lone wire. The magnetic flux is the integral of $B$ and thus is infinite if integration of $1/r$ is done to infinity. It has to be part of a structure that encompasses a surface that is limited in space and contains a certain definable magnetic flux $\Phi$.

Moreover, since charge cannot disappear or accumulate in space, any physical configuration always must consist of such a closed loop anyway, a closed-loop current where charge returns to its origin.

What is more, the magnetic field of an infinitesimal thin wire is infinite at $r = 0$ as Eq. (3.73) shows. Even when combined into any closed-loop structure, when composed of thin wires, the flux $\Phi$ is infinite because the integral of the field function $\int(1/r)\mathrm{d}r$ diverges for $r = 0$. This seriously complicates our calculations since we cannot calculate ideal systems of infinitesimally thin wires. All calculations always depend on the thickness of the used wires.

Imagine four wires combined into a square loop passing a current $I$. The magnetic field of such a loop is plotted in Figure 3.11.

As an example of a calculation, imagine the situation of a square loop made of four wires of finite radius $a$ and a distance of $Z$ passing a total current $I$ as in Figure 3.12. The current inside the wire is the integral of the current density and the magnetic field inside the wire can thus easily be calculated as $B(r) = \mu I r/2\pi a^2$. Outside the wire it is the same as found for an infinitely thin wire, $B(r) = \mu I/2\pi r$. (Note that the permeability of copper is close to unity so $\mu = \mu_0$ inside and outside the wire). The contribution to the flux of a

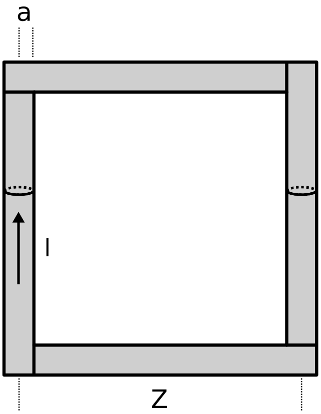


**Fig. 3.12**: Four wires in a square loop. Wires have radius $a$ and distance $Z$

single wire is

$$\Phi_{\text{wire}} = 2Z \int_0^a \frac{\mu I}{2\pi} \frac{r}{a^2} \mathrm{d}r + Z \int_{\mathrm{a}}^Z \frac{\mu I}{2\pi r} \mathrm{d}r.$$

The factor 2 in the first term is caused by the fact that the wire starts at $-a$ and not at 0. The total flux of a loop is four times the flux of a single wire and the total inductance can be found by dividing by $I$ (see Eq. (3.69)),

$$L_{\square} = \frac{4\Phi_{\text{wire}}}{I} = \frac{2Z\mu}{\pi} \left[1 + \ln\left(\frac{Z}{a}\right)\right]. \tag{3.74}$$

The next step is thus the calculation of the magnetic field and flux of a circular loop, see Figure 3.13. Because the symmetry is not so nice as a straight wire, we have to use the Biot-Savart version of Ampere's law, namely the incremental contribution d**B** of an infinitesimal current segment $I$ d$l$ to the magnetic field,

$$\mathrm{d}\mathbf{B} = \frac{\mu I \mathrm{d}\mathbf{l} \times \mathbf{1}_{\mathrm{r}}}{4\pi\rho^2}, \tag{3.75}$$

with $\mathbf{1}_{\mathrm{r}}$ a unity vector pointing from the current segment to the point of evaluation of the magnetic field, and ρ the distance to it. Applied to the current loop of Figure 3.13 the magnetic field at a general point with cylindrical coordinates ($r$, $z$, ϕ) in the z-direction, perpendicular to the loop – the component that interests us – is

$$B_{\mathrm{z}}(r, z) = \frac{\mu I}{2\sqrt{z^2 + (R+r)^2}} \left( \frac{R^2 - r^2}{z^2 + (r-R)^2} E_2(k) + E_1(k) \right), \tag{3.76}$$

with $E_1$ and $E_2$ elliptical functions, and $k$ given by

$$k = \frac{4ra}{\sqrt{(z^2 + R + r)^2}}. \tag{3.77}$$

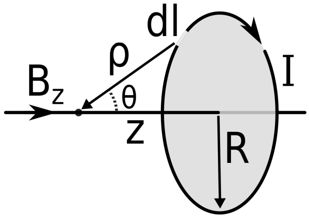


**Fig. 3.13**: Magnetic field caused by a current $I$ in a loop with radius $R$ at a distance $z$ on the axis. ρ is distance between ring element and the point

Evaluated on the axis at a distance $z$ results in a magnetic field in the direction $z$ perpendicular to the loop (all other directions, due to symmetry consideration, must result in zero). Substituting $r = 0$ in the above equation gives us the result below. This result can also be found by analyzing Figure 3.13 and rewriting the Biot-Savart equation,

$$\mathrm{d}B_\mathrm{z} = \frac{\mu I \mathrm{d}l \sin(\theta)}{4\pi\rho^2}. \tag{3.78}$$

The integration $\int \mathrm{d}l$ over the entire loop gives a factor $2\pi R$. Furthermore, $\rho^2 = R^2 + z^2$ and $\sin(\theta) = R/\rho$, and we arrive finally at a magnetic field

$$\begin{aligned} B_\mathrm{z}(z) &= \int_\mathrm{l} \mathrm{d}B_\mathrm{z} = \frac{\mu I 2\pi R}{4\pi(R^2+z^2)} \frac{R}{\sqrt{R^2+z^2}} \\ &= \frac{\mu I}{2} \frac{R^2}{(R^2+z^2)^{3/2}}, \end{aligned} \tag{3.79}$$

which is the same as by substituting $r = 0$ into the general equation. At the center of the loop ($z = 0$), the magnetic field is thus

$$B_\mathrm{z}(0) = \frac{\mu I}{2R}. \tag{3.80}$$

The next step is completely unjustified, but is the textbook way of treating the subject (exact solution given in Table 3.VI): If we simplify the equation and assume that the magnetic field is constant in the area of the loop, the total flux and inductance are, respectively,

$$\Phi_\bigcirc = \pi R^2 \times \frac{\mu I}{2R} = \frac{\pi\mu I R}{2}, \tag{3.81}$$

$$L_\bigcirc = \frac{\Phi}{I} = \frac{\pi\mu R}{2}. \tag{3.82}$$

If we place various loops close together, the inductance is proportional to the square of the number of loops. This is easy to understand, because the field is increasing linearly with the number of loops, but this field also passes through more loops, so the effect is squared.

$$L_\mathrm{N} = N^2 L. \tag{3.83}$$

Often the definition of flux linkage $\lambda$ is used, where

$$L_{\mathrm{N}} = \frac{\lambda}{I} = \frac{N\Phi}{I}. \tag{3.84}$$

A solenoid or coil is a spiral that resembles very much a set of $N$ loops spaced at a distance $\Delta Z$, with a total length of $Z = N\Delta Z$, see Figure 3.14. We can thus 'easily' find the total magnetic field from Equation 3.79. For instance, for the center of the solenoid

$$B_{\mathrm{z}} \sum_{\mathrm{n=-N/2}}^{N/2} B_{\mathrm{z}}\left(nZ/N\right). \tag{3.85}$$

However, there is an easier way to do this. Namely by going back to Ampere's law (Eq. 3.71) that states that the integral of magnetic field along a closed path is equal to the total current passing through the surface that is encompassed by the path. Since any path will do, we can chose a specific path that facilitates the calculation, see the path shown in Figure 3.15. The integration path is shown there dashed. The sides of this path do not contribute to the integral because the magnetic field is there perpendicular to the path. Also, the top segment of the path contributes relatively little (the further we place it away, the less it counts), so we remain with the bottom segment, where the field is along the path, and the integral is thus equal to $B_{\mathrm{z}}Z'$. The current enclosed in the path is equal to the number of loops passing through the path $N' = Z'/\Delta Z$, times the current in each loop, $I$, we thus find

$$B_{\mathrm{z}}Z' = \mu I \frac{Z'}{\Delta Z}, \tag{3.86}$$

thus

$$B_{\mathrm{z}} = \frac{\mu I}{\Delta Z}, \tag{3.87}$$

or alternative forms,

$$B_{\mathrm{z}} = \mu \frac{N}{Z} I = \mu n I, \tag{3.88}$$

with $N$ the total number of loops, $Z$ the total length of the solenoid, and $n$ the density of loops (per meter). The inductance of a solenoid is this filed multiplied by the area and divided by the current,

$$L_{\mathrm{solenoid}} = \frac{\pi R^2 \mu N^2}{Z}. \tag{3.89}$$

The energy stored in an inductor bearing current is easily calculated. We will slowly build up the current in the inductor and every increment in current costs a certain energy. The total energy stored in the inductor is then equal to the energy consumed in building up the current. Or, in other words, it is the integral of power over time:

$$U = \int_0^y P(t)\mathrm{d}t = \int_0^t V(t)I(t)\mathrm{d}t. \tag{3.90}$$

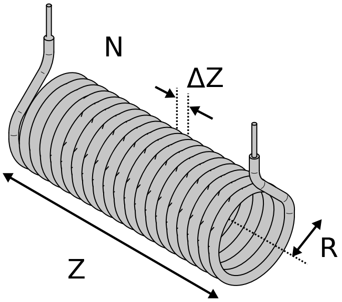


**Fig. 3.14**: Solenoid of $N$ windings, each with radius $R$, spaced $\Delta Z$ resulting in a total device length of $Z = N\Delta Z$. (Based on image by inductiveload found on Wikipedia)

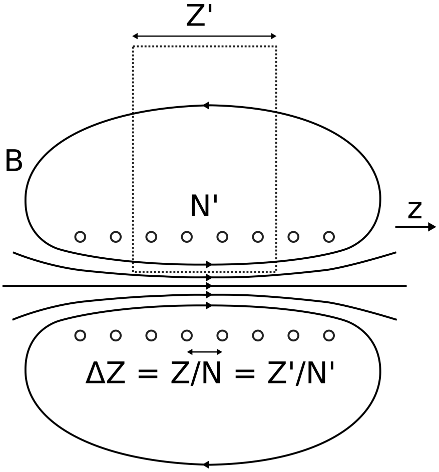


**Fig. 3.15**: Cross-section of a solenoid. Each circle is a coil with current $I$, spaced $\Delta z$. Using Ampere's law to find and approximation for the field inside the solenoid through integration along path shown here dashed. Only the bottom segment effectively contributes to the integral; the side segments are perpendicular to the field and the top segment is too far away and the field too small

**Table 3.VI**: Approximations of inductances for some geometries. $Z$ is length, $a$ is wire radius ($a_1$, $a_2$: inner and outer radius), $R$ is loop diameter, $d$ is wire (inter)distance, $N$ is number of coils

| **Geometry** | **Inductance** | Comment |
|---|---|---|
| Single square coil | $\frac{2\mu Z}{\pi}\left[1+\ln\left(\frac{Z}{a}\right)\right]$ | |
| Single circular coil | $\mu\pi R/2$<br>$\mu R\left[\ln\left(\frac{8R}{a}\right)-\frac{7}{4}\right]$ | approx.<br>exact |
| Solenoid | $\pi R^2\mu\frac{N^2}{Z}$ | Figure 3.15 |
| Parallel wires | $\frac{\mu Z}{\pi}\cosh^{-1}\left(\frac{d}{a}\right)$ | Figure 3.9 |
| Coax cable | $\frac{\mu Z}{2\pi}\ln\left(\frac{a_2}{a_1}\right)$ | Figure 3.9 |

$\cosh^{-1}(x)=\ln(x+\sqrt{x^2-1})$

We can now substitute the definition of inductance (Eq. 3.65) and work out the integral:

$$\begin{aligned} U &= \int_0^t V(t)I(t)\mathrm{d}t \\ &= \int_{\mathrm{I}(0)}^{I(t)} LI\frac{\mathrm{d}I}{\mathrm{d}t}\mathrm{d}t \\ &= \int_0^I LI'\mathrm{d}I' = \frac{1}{2}LI^2. \end{aligned} \tag{3.91}$$

This energy is stored in the form of a magnetic field. In an approximation, we can calculate this energy density, simply by dividing the total energy of the inductor by the volume effectively occupied by the inductor. As an example, the solenoid, with length $Z$ and coil area $A$ will have an energy density

$$u \equiv \frac{U}{V} = \frac{LI^2/2}{AZ}. \tag{3.92}$$

Substituting Eqs. 3.88 and 3.89 above

$$I^2 = \frac{B^2Z^2}{\mu^2N^2},\quad L=\pi R^2\mu\frac{N^2}{Z},\quad A=\pi R^2, \tag{3.93}$$

gives for the energy density

$$u=\frac{B^2}{2\mu}. \tag{3.94}$$

We see that both a charged capacitor (through its electric field), and a current-bearing inductor (through its magnetic field) store energy. This energy cannot disappear instantaneously! For a capacitor this is no big deal; if we disconnect a charge capacitor, the charge will remain stored in it, and the field and energy will remain until we reconnect it and draw the charge out of it; a capacitor can be used as a battery. For an inductor, however, the things are much more complicated. Since the energy is stored in the form of magnetic field that is associated to a current, the fact that energy cannot disappear instantaneously, implies that the current *must* continue, even if we disconnect the inductor! Imagine what will happen. Because the resistivity of air is quite high, opening the circuit at some place will effectively insert a high resistance into the circuit at that place. Where the current continues initially unaltered, this current induces a high voltage drop $\Delta V = RI$ across the gap. A voltage that can easily go far beyond the supply voltage. Moreover, the current through the resistance generates a power there equal to $P = I^2R$ (see Figure 3.6) and this can be a huge power. So much power deposited in air will heat it up to the point that it will start emitting light. In other words, by disconnecting a current-bearing inductor, we will create what is more-commonly known as a spark. The current and power will continue until the energy originally stored in the magnetic field is consumed and converted into heat and light. An example is a car battery. We know that we have to be very careful when disconnecting it. A 12-volt car battery can easily deliver tens of amperes and carelessly disconnecting it can be dangerous to our lives!

**Question**: A battery from a car with the two 40-watt headlights on is disconnected. What will be the voltage induced? (Resistivity of air: $\rho = 2 \times 10^{16}$ Ωm, cable diameter: $d = 1$ cm, gap: $L = 1$ cm).

**Answer**: Using $P = VI$, and $P = 80$ W, $V = 12$ V, we find $I = 6.67$ A. The resistance of the gap is $R = \rho L/A = 2 \times 10^{16}$ (Ωm) $\times$ 1 cm$/\pi(1 \text{ cm}/2)^2 = 2.5 \times 10^{18}$ Ω. The voltage is $V = RI = 1.7 \times 10^{19}$ V. If the leads have many loops and high inductance with a lot of energy stored, this astronomical voltage can continue for considerable time.

**Question**: If we open the circuit with our hands, what will be the voltage across our body? ($R \approx 2$ MΩ).

**Answer**: $V = RI = 12$ MV. Quite lethal, if there is enough energy stored in the inductor and the current and voltage will continue long enough.

This effect of creating sparks we all know from experience. Often when we disconnect power equipment (with high current) a spark is visible at the connector. Often also a spark is seen when we connect the device, but that is because it is not easy to make an instantaneous connection. In reality, when we connect something, in a split second thousands of times the connection will be established and broken. Every disconnection can give cause to a spark.

Finally, just as resistance is accompanied by its reciprocal conductance, so also is capacitance accompanied by its reciprocal quantity. This quantity is

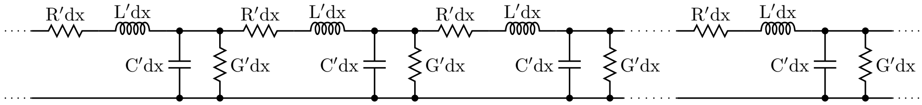


**Fig. 3.16**: Transmission line. A cable consists of an infinite number of tiny segments of length $\mathrm{d}x$, each having resistance $R'\mathrm{d}x$, inductance $L'\mathrm{d}x$, capacitance $C'\mathrm{d}x$ and conductance $G'\mathrm{d}x$

called 'elastance' with unit 'daraf' (1/F). The reciprocal quantity of inductance does not have a specific name, but the unit of reciprocal inductance is the 'yrneh'; we recognize the units written backwards. They are, however, seldom used. More common is to work with quantities called reactance ($X_\mathrm{C} = -1/\omega C$, $X_\mathrm{L} = \omega L$) and susceptance ($B_\mathrm{C} = \omega C$, $B_\mathrm{L} = -1/\omega L$) that have units ohm and siemens, respectively. See Chapter 2.

## 3.4 Cables as transmission lines

As seen in the previous section, cables unavoidably have both capacitance – because a cable always has a 'return line', and the couple forms a type of metal-plates capacitor – and inductance – because any current induced in the cable causes a magnetic field, the link between the two being self-inductance. Apart from that, a cable has its obvious resistance, limiting currents, and leakage, represented by a conductance short to the return wire. From an electronic point of view, a wire can then be divided into to an infinite number of infinitesimal tiny segments, $\mathrm{d}x$, each with its own resistance $R'\mathrm{d}x$, inductance $L'\mathrm{d}x$, capacitance $C'\mathrm{d}x$ and conductance $G'\mathrm{d}x$. Figure 3.16 shows an example of this. All the parameters have their usual units multiplied by 'per length' (F/m, etc.), hence the prime ($'$) added to the symbols to distinguish them from the conventional ones.

As an example may serve a coax cable. The capacitance per unit length and inductance per unit length of these cables were calculated in the previous sections. The cable resistance per meter $R'$ and leakage conductance per meter $G'$ can also easily be found. With a radius of $a_1$, a copper nucleus has a resistance per meter of

$$R' = \frac{\rho_\mathrm{c}}{\pi a_1^2}, \tag{3.95}$$

with $\rho_\mathrm{c}$ the resistivity of copper. The leakage conductance through the insulator can be found through calculation of a leakage resistance. This resistance can be imagined to be composed of concentric cylinders of insulator ($\rho_\mathrm{i}$ resistivity) with radius $r$ and thickness $\mathrm{d}r$, placed in series. Each cylinder, with length $Z$, contributing an incremental resistance equal to

$$\mathrm{d}R = \frac{\rho_\mathrm{i}\mathrm{d}r}{2\pi r Z}, \tag{3.96}$$

**Table 3.VII**: Parameters of coax cables and values for a typical copper-teflon cable ($\rho_c = 16.78$ n$\Omega$m, $\rho_i = 10^{16}$ $\Omega$m, $\varepsilon_i = 2.1\varepsilon_0$, $\mu_i = \mu_0$, $a_1 = 0.455$ mm, $a_2 = 1.49$ mm)

| Parameter | Expression | Typical value |
|---|---|---|
| $R'$ | $\frac{\rho_c}{\pi a_1^2}$ | $25.8$ m$\Omega$/m |
| $G'$ | $\frac{2\pi}{\rho_i \ln(a_2/a_1)}$ | $5.297 \times 10^{-16}$ S/m |
| $L'$ | $\frac{\mu}{2\pi} \ln(a_2/a_1)$ | $2.37 \times 10^{-7}$ H/m |
| $C'$ | $\frac{2\pi\varepsilon}{\ln(a_2/a_1)}$ | $98.5$ pF/m |
| $Z_0$ | $\sqrt{L'/C'}$ | $50\ \Omega$ |

and the total resistance can be found by the integral from the inner radius $a_1$ to the outer radius $a_2$,

$$R = \int_{a_1}^{a_2} \frac{\rho_i \mathrm{d}r}{2\pi r Z} = \frac{\rho_i}{2\pi Z} \ln\left(\frac{a_2}{a_1}\right). \tag{3.97}$$

Thus

$$G' = \frac{1}{ZR} = \frac{2\pi}{\rho_i \ln(a_2/a_1)}. \tag{3.98}$$

These parameters of a coax cable are summarized in Table 3.VII, together with typical values.

## 3.4.1 Impedance of a cable

If a voltage is applied to the cable, the capacitance will be charged, or in other words, a current will flow, even if nothing is connected to the other side of the cable. This current is not instantaneous, but the inductance limits the growth of the current. In general, when a voltage is applied, a complicated interplay between current, voltage, resistance, capacitance, inductance and conductance takes place. It is at this stage interesting to calculate what is the total impedance of a signal cable, defined as the ratio of voltage and current at the entrance of the cable.

For this, it is easiest to use an infinite cable. Figure 3.18 shows the first segment of this cable. Because the cable is of infinite length, the impedance looking into the cable $Z_0$ is the same as the impedance looking into the rest of the cable after the first segment. In other words,

$$Z_0 = A + (B \parallel Z_0), \tag{3.99}$$

with

$$A = R'\mathrm{d}x + j\omega L'\mathrm{d}x \tag{3.100}$$

$$B = (G'\mathrm{d}x + j\omega C'\mathrm{d}x)^{-1} \tag{3.101}$$

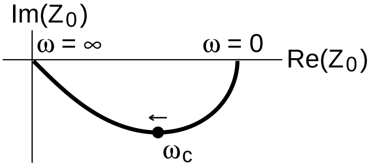


**Fig. 3.17**: Nyquist plot of the impedance of a cable based on parameters of Table 3.VII. The turning frequency $\omega_c$ is normally very small (typically in the order of μHz)

Solving this non-linear equation gives for the limit d$x \to 0$ ($A \to 0$, $B \to \infty$)

$$Z_0 = \sqrt{AB} = \sqrt{\frac{R' + j\omega L'}{G' + j\omega C'}}. \tag{3.102}$$

(see Figure 3.17 for a graphical representation of the real and imaginary part of this impedance). Note that the length of the segment d$x$ has disappeared from the equation. The parameters $R'$, $L'$, $G'$ and $C'$ are all 'per unit length', but the final value $Z_0$ does not have this unit 'per length' and it pertains to the total cable. Even more interesting, a cable without resistance ($R' = 0$) and without leakage ($G' = 0$), in other words, an ideal cable, or any cable at high frequency, has an impedance given by

$$Z_0 = \sqrt{\frac{L'}{C'}}, \tag{3.103}$$

which is a non-imaginary positive real number with a unit 'ohm'. In other words, it behaves like a normal resistor. Note that because the length cancels in the division, $L'$ and $C'$ can also be replaced by $L$ and $C$, the total inductance and capacitance of the cable. By designing the dimensions of the cable, like inner and outer diameter of the poles, or the distance between them, this characteristic resistance $Z_0$ can be made any value. Popular are the 50-Ω and 75-Ω coax cable families.

Note that the value $Z_0$ is also found for infinite frequencies. On the other hand, for low frequencies the impedance is given by

$$Z_0(\omega \to 0) = \sqrt{\frac{R'}{G'}}, \tag{3.104}$$

which once again is a real positive number with unit 'ohm'. The turning-point frequency that makes the cable go from low-frequency regime to the high-frequency regime is very low. Substituting typical values for cables, for example the coax cable of Table 3.VII, will result in a turning frequency in the micro-hertz range. For frequencies below this, the infinite cable will work as a network of resistances. For higher frequencies it works as an LC network. For all reasonable purposes, the impedance is the high-frequency variant.

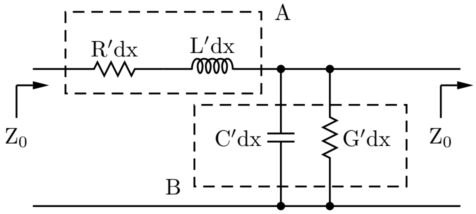


**Fig. 3.18**: The first segment of an infinite cable. The characteristic impedance $Z_0$ as looking into the cable is the same as the impedance after the first element

### 3.4.2 Wave propagation in a lossless cable

Going back to the ideal cable – without resistance ($R' = 0$) and leakage ($G' = 0$) – if a time-dependent signal is applied to it, for instance $V(t)$, the induced current variations cause voltage drops along the cable in the inductors, and part of the current is lost in charging the capacitors. Analyzing Figure 3.19,

$$\begin{aligned} \mathrm{d}V(x,t) &= -L'\mathrm{d}x\frac{\mathrm{d}I(x,t)}{\mathrm{d}t}, && (3.105)\\ \mathrm{d}I(x,t) &= -C'\mathrm{d}x\frac{\mathrm{d}V(x,t)}{\mathrm{d}t}. && (3.106)\end{aligned}$$

Moving the terms d$x$ to the left side and differentiating the first equation with respect to $t$ and the second to $x$ yields

$$\begin{aligned} \frac{\partial^2 V(x,t)}{\partial x \partial t} &= -L'\frac{\partial^2 I(x,t)}{\partial t^2}, && (3.107)\\ \frac{\partial^2 I(x,t)}{\partial x^2} &= -C'\frac{\partial^2 V(x,t)}{\partial x \partial t}. && (3.108)\end{aligned}$$

From which easily follows

$$\frac{\partial^2 I}{\partial x^2} = L'C'\frac{\partial^2 I}{\partial t^2}. \quad (3.109)$$

Likewise,

$$\frac{\partial^2 V}{\partial x^2} = L'C'\frac{\partial^2 V}{\partial t^2}. \quad (3.110)$$

These are classical wave equations which result in the propagation of the wave. Substituting a general wave function

$$V(x,t) = V_0 \exp\left[j\left(\omega t - 2\pi x/\lambda\right)\right], \quad (3.111)$$

and

$$I(x,t) = I_0 \exp\left[j\left(\omega t - 2\pi x/\lambda\right)\right], \quad (3.112)$$

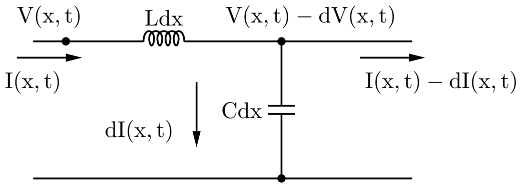


**Fig. 3.19**: One segment of a transmission line and the voltages and currents in it

with $V_0/I_0 = V(x=0,t)/I(x=0,t) \equiv Z_0$ in Eq. (3.109) or (3.110) above gives the constraint

$$\begin{aligned}\left(\frac{2\pi}{\lambda}\right)^2 &= \omega^2 L'C' \\ \omega &= \frac{2\pi/\lambda}{\sqrt{L'C'}}.\end{aligned} \tag{3.113}$$

The speed at which a wave 'travels', the so called phase velocity $v$, can be calculated by setting the phase constant and calculating the place where this phase occurs as a function of time:

$$\omega t - \frac{2\pi}{\lambda}x = \text{constant}, \tag{3.114}$$

from which follows the phase velocity, using the constraint of Eq. 3.113

$$v = \frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\lambda\omega}{2\pi} = \frac{1}{\sqrt{L'C'}}. \tag{3.115}$$

Assuming (without proof) that the signal travels in electromagnetic waves, this phase velocity has to be equal to the speed of light, and thus

$$\frac{1}{v^2} = \mu\varepsilon, \tag{3.116}$$

and we thus now find

$$L'C' = \mu\varepsilon. \tag{3.117}$$

Looking at the capacitances and inductances found (Tables 3.III and 3.VI, respectively), for instance the coax cable and parallel wires, we see that this is indeed the case.

If the cable is not infinite, sooner or later the wave will hit the end and is possibly reflected back with a factor $r$. A general wave, instead of Eq. 3.111 is

$$V(x,t) = V_0 \exp\left[j\left(\omega t - 2\pi x/\lambda\right)\right] + rV_0 \exp\left[j\left(\omega t + 2\pi x/\lambda\right)\right], \tag{3.118}$$

where the first term is the positive phase velocity and the second term negative phase velocity. The accompanying current is given by solving Equation 3.106:

$$I(x,t) = \frac{V_0}{Z_0} \exp\left[j\left(\omega t - 2\pi x/\lambda\right)\right] - r\frac{V_0}{Z_0} \exp\left[j\left(\omega t + 2\pi x/\lambda\right)\right], \tag{3.119}$$

where the definition of $Z_0 = V_0/I_0$ was used, as well as the constraint of Eq. 3.113. Note the minus sign in the second term of the current. If a load $Z_\mathrm{L}$ is placed at the end of the transmission line, and by, without loss of generality, defining $x = 0$ at this end,

$$Z_\mathrm{L} \equiv \frac{V(x=0)}{I(x=0)} = Z_0 \frac{1+r}{1-r}. \tag{3.120}$$

Inverting this equation:

$$r = \frac{Z_\mathrm{L} - Z_0}{Z_\mathrm{L} + Z_0}. \tag{3.121}$$

This is an important conclusion. If we do not want reflections, we have to terminate the line with a load with an impedance equal to the characteristic impedance of the cable, $Z_\mathrm{L} = Z_0$. If we have a 50-Ω cable, the equipment on the other end of the cable has to have a 50 Ω input resistance. If there is an impedance mismatch, if the load is more (or less) than the characteristic impedance of the cable, the signal is reflected back to the sender and absorbed there (if the sender has impedance equal to the cable). Specifically, if the end of a cable is left open ($Z_\mathrm{L} = \infty$), or shorted ($Z_\mathrm{L} = 0$), the wave is fully reflected. If we want to transfer the maximum amount of energy, the impedance of the receiver has to match the impedance of the cable.

### 3.4.3 Lossy cables

If the cable is lossy, the resistive elements in the transmission line segments cannot be ignored and the wave-propagation equations (Eqs. 3.105 and 3.106) are replaced by

$$\frac{\partial V}{\partial x} = -L'\frac{\partial I}{\partial t} - R'I, \tag{3.122}$$

$$\frac{\partial I}{\partial x} = -C'\frac{\partial V}{\partial t} - G'V. \tag{3.123}$$

Solutions of these equations give an impedance $Z_0 \equiv V_0/I_0$ equal to the one found before (Eq. 3.102). To calculate the exact wave form and propagation is quite complicated. Differentiating the first equation with respect to $x$ and the second with respect to $t$ and subsequent substitution gives

$$\frac{\partial^2 V}{\partial x^2} = (L'C')\frac{\partial^2 V}{\partial t^2} + (R'C' + G'L')\frac{\partial V}{\partial t} + (R'G')V, \tag{3.124}$$

to which we can apply a trial function of an exponentially-decaying sinusoidal wave,

$$V(x,t) = V_0 \exp(-ax) \exp[j(\omega t - 2\pi x/\lambda)]. \tag{3.125}$$

Substituting in the differential equation (Eq. (3.124)) results in the expression

$$(a + j2\pi/\lambda)^2 = -\omega^2 L'C' + j\omega(R'C' + G'L') + R'G'. \tag{3.126}$$

Because the real part and the imaginary part on both sides have to be equal, we find the attenuation factor

$$a = \frac{\omega(R'C' + G'L')\lambda}{4\pi}, \tag{3.127}$$

and the frequency-wavelength relation

$$\omega = \sqrt{\frac{(2\pi/\lambda)^2 + R'G'}{L'C' + (R'C' + G'L')^2(\lambda/2\pi)^2/4}}. \tag{3.128}$$

For a lossless cable, $R' = 0$ and $G' = 0$, the attenuation is zero ($a$=0), and the frequency-wavelength relation goes back to the one found earlier (Eq. (3.113)). The phase velocity, once again, is found by looking at the phase of the wave ($\omega t - 2\pi x/\lambda$) and setting it constant, the phase velocity then being the speed at which this constant phase travels:

$$v \equiv \frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\lambda\omega}{2\pi} = \sqrt{\frac{1 + R'G'/(2\pi/\lambda)^2}{L'C' + (R'C' + G'L')^2(\lambda/2\pi)^2/4}}. \tag{3.129}$$

At high frequencies (short wavelengths), or small resistance and leakage, this is equal to the speed of light in that insulator medium , $v = 1/\sqrt{\mu\varepsilon} = 1/\sqrt{L'C'}$. There is no dispersion to be expected. However, don't forget that all parameters of the cable depend on frequency and we can expect serious signal distortion when these effects are included.

## 3.5 (Quantum) physics of solid state materials and semiconductors

The previous section describes the 'classic' physics of electronics. With this we can explain everything ... *except* the behavior of semiconducting materials. To understand these peculiar materials, we have to take a look at quantum physics. We start with quantum mechanic atomic physics resulting in discrete energy levels and then we will join atoms to form solid state materials and describe them with solid-state physics theory. Here the individual atoms that constitute the crystal are of less importance and the behavior is of the overall crystal instead. As said, we start with atomic (quantum) physics.

As we all know, the electrons of isolated atoms have discrete energy levels. A schematic simplified picture of the Rutherford-Bohr model of these levels is given in Figure 3.20. For instance, the energy levels of hydrogen are given by

$$U_{\mathrm{H}} = -\frac{13.6\ \mathrm{eV}}{n^2}, \tag{3.130}$$

where the unit of energy, electronvolt (eV), is equal to the energy of an elementary charge passing one volt, thus 1 eV $= q \times$ (1 volt) $= 1.6 \times 10^{-19}$ J. We can thus recognize 'shells': K ($n = 1$), L ($n = 2$), M ($n = 3$), etc.

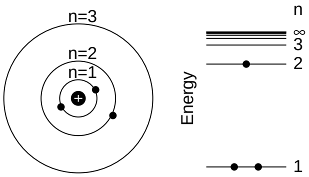


**Fig. 3.20**: Schematic representation of the electronic orbitals and energy levels of an isolated atom. The levels in this case are filled with three electrons (Lithium)

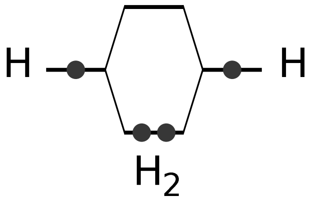


**Fig. 3.21**: Hydrogen molecule energy levels. Two separated hydrogen atoms (left and right), with the K-shell (1s electron, $n = 1$) levels shown, approach each other and they interact forming new molecular energy levels (middle). The two electrons (one from each hydrogen atom) can lower their energy this way and thus a stable 'bonding' molecule is formed

The electrons are being filled on the levels in a way that 1) The total energy is minimized, 2) The electrons do not occupy the same state, the so-called Pauli exclusion principle. The K shell consists of two states and can thus accommodate 2 electrons, the L shell can accommodate 8 electrons, etc.

When two atoms join ('interact'), their levels start mixing. In the first step, for instance combining two hydrogen atoms, we get molecular levels. Without going into details about how this exactly works and what the resulting levels will be we can get some idea. An example is given in Figure 3.21, where two atomic K-shell levels of hydrogen interact ('mix') to result in two molecular levels, one higher and one lower than the original atomic levels. The electrons both fall into the new lower level, showing that energy is gained by forming this bond and the hydrogen molecule is thus stable; it costs energy to separate them again.

When more atoms join, the mixing of the levels becomes more and more intricate. Every new atom brings with it its own levels for mixing and the levels of the total structure thus rapidly becomes abundant. In fact, so much so that individual levels are ever less distinguishable. More so, since they are no longer sharp, but rather diffuse, due to the Heisenberg principle that states that the accuracy of knowledge of energy of a level is diminishing if the electron spends less time on it. Since in a larger structure the electrons rapidly travel 'all over the place', the resulting energy level diagram is one big smear. The smearing is further augmented by the vibrational movement of the individual atoms.

However, we can get a general picture. In nearly all cases, in solid material crystals, the levels are 'lumped'; they fall in so-called 'bands' of energy, each one containing a gazillion levels (the same as the number of atoms in the solid). As we have seen, two interacting hydrogen atoms, each contributing one energy level results in a system with two levels and we could already have recognized some form of 'bands' in Figure 3.21. If we add more atoms, there will be many bands and the final number of levels in a single band is equal to the number of atoms of the system. We can now find a density of energy levels, commonly known as density-of-states DoS, that is equal to the density of atoms, and thus has a unit 'per cubic meter'. If we specify it as a function of energy (the vertical axis in the energy diagrams), we get a unit 'per cubic meter per joule', $\mathrm{m}^{-3}\mathrm{J}^{-1}$. Chemists and physicist prefer to use energy units calorie and electronvolt respectively, but the idea is the same. How many levels exactly are available in a band? For that we can use another analysis, that starts with free electrons

The above approach is atomic-physics and molecular-physics oriented. We started with atomic and molecular levels and slowly moved into larger clusters until eventually arriving at crystals. We can also reason from the other limit. We start with free electrons and let them interact with the positive matrix of the ions of the crystal. This is the approach mostly used for solid state materials, and in particular semiconductor physics. We also mix classical mechanics with quantum mechanics for this approach.

We start with a free electron. A free electron with mass $m_\mathrm{e}$ and speed $v$ has a kinetic energy given by

$$U_\mathrm{k} = \frac{1}{2} m_\mathrm{e} v^2. \tag{3.131}$$

From classical mechanics we also know that we can also talk in terms of momentum, $p = mv$. From quantum mechanics, the momentum, in turn, is expressed in terms of wave number $k$ ('quantum speed'), which is defined as

$$k \equiv \frac{2\pi}{h} p, \tag{3.132}$$

with $h$ Planck's constant. We thus get

$$U_\mathrm{k} = \frac{h^2 k^2}{8\pi^2 m_\mathrm{e}}. \tag{3.133}$$

So far, it is just a renaming of the symbols. Figure 3.22 shows the kinetic energy of a free electron as function of its speed $k$.

We will now place the electrons on top of the crystal lattice with three ingredients:

- They are still independent, meaning that they don't feel each other
- They feel the host lattice of periodically placed positively charged ions

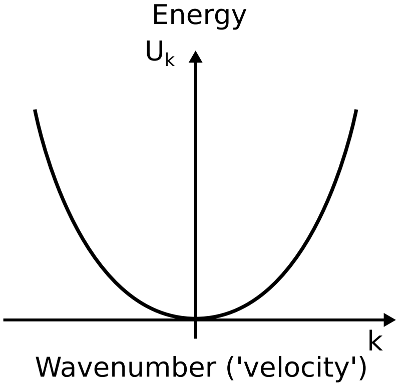


**Fig. 3.22**: Kinetic energy $U_\mathrm{k}$ of a free electron as a function of its speed (which is called wave number $k$ in quantum mechanics terms). The curvature of the plot is determined by the mass of the electron, see Eq. 3.133

- We use quantum mechanics coupling momentum $p$ to a wavelength $\lambda$ according to De Broglie Relation (which is given here without proof or further discussion)

$$\lambda = \frac{h}{p} = \frac{2\pi}{k} \tag{3.134}$$

**Question**: What is the wavelength of a person of 80 kg walking at 6 km/h?
**Answer**: $p = mv = 133$ kg m s$^{-1}$. $\lambda = h/p = 5 \times 10^{-36}$ m.

**Question**: What is the wavelength of an electron with speed equal to 1% of speed of light?
**Answer**: $p = m_\mathrm{e}c/100 = 2.73 \times 10^{-24}$ kg m/s. $\lambda = h/p = 2.4\times10^{-10}$ m = 2.4 Å.

Now that we have established that speed comes with wavelength and (kinetic) energy, we can let the wavelength interact with the lattice and see what energy contribution that will give. To do that, first of all, we have to realize that the wave function of a particle represents the probability density function:

The probability $P$ to find the electron at a certain place $x$ is proportional to the square of the wave function $\Psi$ at that place,

$$P(x) = |\Psi(x)|^2. \tag{3.135}$$

Moreover, in a crystal, the electron is confined to the crystal and is, as it were, put in a box. From classical mechanics we know that confined wave functions are of the type standing wave functions. That is, the half wavelength fits exactly an integer number of times in the box.

$$\begin{aligned} n \times \frac{\lambda}{2} &= L = Na \\ \lambda &= \frac{2Na}{n} \end{aligned} \tag{3.136}$$

with $L$ the box (crystal) length, $N$ the number of atoms and $a$ the distance between atoms. If we substitute this in the De Broglie Relation (Eq. 3.134) we find that the wave number is quantized, namely

$$k = \frac{\pi}{Na} n. \tag{3.137}$$

The longest wavelength is thus equal to twice the dimension of the box, $\lambda = 2L = 2Na$. An interesting case occurs when the wavelength is equal to twice the interatomic distance, $\lambda = 2a$, thus $k = \pi/a$, see Figure 3.23 for a one-dimensional representation of a crystal and its waves for this special case.

$$k_{\text{edge}} = \frac{\pi}{a}. \tag{3.138}$$

For this wavelength and wave number there are two distinct possibilities, one with high probability density on the atomic nuclei and one with low probability density. Or, in other words, one situation in which the electron spends a lot of time close to the nuclei and one in which it spends more time further away. From classical mechanics we know that the potential energy of a charge $p$ at a distance $r$ of a second charge $q$ is given by Coulombs equation (see also Eq. 3.6),

$$U_{\text{coulomb}} = \frac{pq}{2\pi\epsilon r}. \tag{3.139}$$

If $p$ and $q$ are of opposite sign, the lowest energy is obtained when the charges are close together. In our case, when the electron is closer to the nuclei. We can thus expect two different energies for the two situations shown in Figure 3.23, namely lower energy for the top situation compared to the bottom situation. Adding this potential energy to the energy diagram of the free electron (kinetic) energy results in the diagram shown in Figure 3.24 (left). We can recognize here the formation of a band gap; there exist energies that are not possible to reach for electrons in a crystal. Note also that the link between wave number $k$ and momentum $p$ is lost, since momentum is no longer the only contribution to the energy of the electron.

Finally, because of the periodicity of the lattice, it is possible to represent the entire energy diagram in a zone between $-\pi/a$ and $+\pi/a$, the so-called first Brillouin zone, see 3.24 (right). Such pictures we call band diagrams. We can now clearly recognize energy bands and a 'forbidden gap' in the possible energies.

An important question now is: Since we know that the $k$ values are quantized (Eq. 3.137), how many levels are there in a band? This question is easily

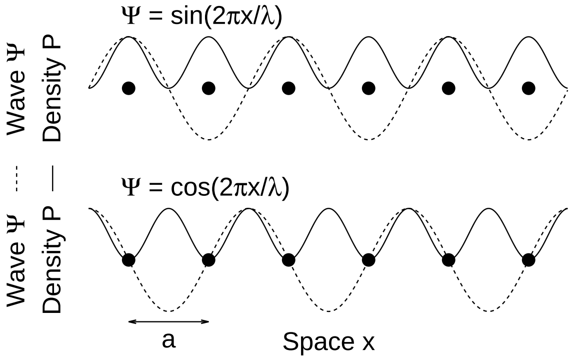


**Fig. 3.23**: Wave function and probability density of a special case where $\lambda = 2a$, thus $k = \pi/a$. There exist two possibilities, one with high density of wave function on the atomic nuclei (top) and one with no density on the nuclei (bottom) – all other possibilities can be expressed in combinations of these two. They have significantly different energy because of the large Coulomb interaction (Eq. 3.139)

answered. Equation 3.137 tells us that the distance between two values is

$$\Delta k = \frac{\pi}{Na}, \tag{3.140}$$

and the value of $k$ at the edge of a Brillouin zone is $k = \pi/a$, we have a total number of discrete k-values equal to

$$N_\mathrm{k} = \frac{k_\mathrm{edge}}{\Delta k} = \frac{\pi/a}{\pi/Na} = N. \tag{3.141}$$

If, instead of atoms we would have placed other repeating units – for instance pairs of Gallium and Arsenide in GaAs crystals – we would have had to talk about unit cells instead of atoms. The important conclusion is thus

The amount of levels in a band is equal to the number of nuclei/unit cells of the crystal.

$$N_\mathrm{k} = N. \tag{3.142}$$

The electrons contributed by the atoms or unit cells are now filling up the levels in the same way as we have seen for the hydrogen molecule, and they obeying the same basic quantum mechanic principles: 1) The total energy is minimized, 2) The electrons do not occupy the same state (Pauli's exclusion principle). Every level can contain two electrons, because an electron can have spin up or spin down. While this quantum-mechanic number of spin is irrelevant for our electronic analysis, it does mean that every level can accommodate two electrons, and thus every band can accommodate a total of $2N$ electrons.

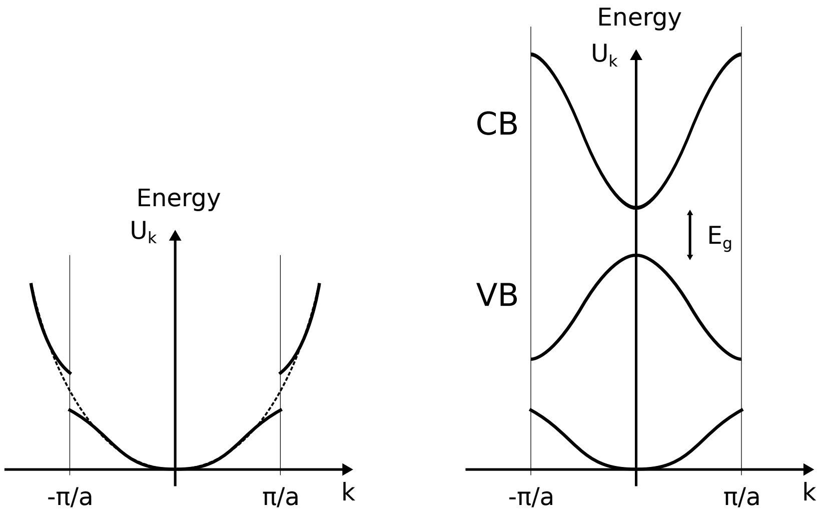


**Fig. 3.24**: Band diagram. Left: When the interaction with the lattice is taken into account, the free electron (kinetic) energy diagram (dashed line) is distorted, especially at wave numbers close to the values where the wavelength is equal to twice the inter-atomic distance, $\lambda = 2a$, thus $k = \pi/a$ (or odd multiples thereof, $k = (2n-1)\pi/a$), with the wave functions with high density at the nuclei having lower energy than the ones with low density at the nuclei. Right: Because of the periodicity of the lattice and the wave function, implying invariance over shifts of $2n\pi/a$, the entire energy diagram can be mapped to a zone between $-\pi/a$ and $+\pi/a$, the first Brillouin zone. Clear energy 'gaps' are visible, energy values that the electron cannot take. When filling up the energy levels from low-to-high (as nature wants), the last fully occupied band is called 'valence band' (VB), and the first unoccupied band is named 'conduction band' (CB), separated by the bandgap $E_\mathrm{g}$. Note also that the horizontal scale is quantized. The distance between two quantization levels is $\Delta k = \pi/Na$ and there thus exist exactly $N$ levels in each branch

There now exist two possibilities. Each atom (or repeating combination of atoms, like in Ga-As, etc) contributes an odd number of electrons, with a total of $N$, $3N$, etc, or an even number of electrons, $2N$, $4N$, etc. In the first case, that is commonly found for not nicely chemically bonded materials one band of levels is necessarily half full (or half empty, depending on your point of view). In the second case, where every atom is nicely covalently bond, every band is either filled up completely or completely empty. The first case, with half full bands is called a metal, the second a semiconductor, see Figure 3.25. A third case exists, where on basis of the number of electrons, the material would be a semiconductor, but because the bands overlap, two bands can be partially filled. These materials are called semi-metals and they behave like metals in terms of conduction.

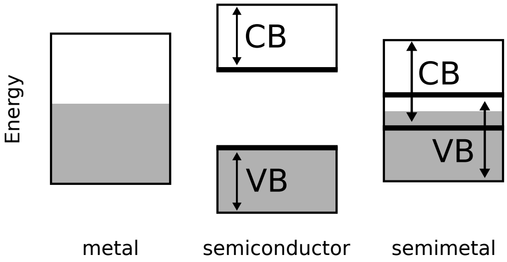


**Fig. 3.25**: Energy levels of a solid. Left: metal, one band of energy levels half filled with electrons (other bands not shown). Middle: semiconductor, all bands either completely full (valence band, VB), or completely empty (conduction band, CB). Right: Semi-metal, two bands overlap and they are both partly filled

The band that is full is called the valence band and the band that is empty is called the conduction band. This is mainly for historical and chemical reasons, since chemists call the electron(level)s that participate in chemical bonding 'valence electrons'. The reason for calling the conduction band by that name is less obvious. As we will see, conduction takes place through both the conduction *and* valence bands.

**Question**: If every atom in a silicon crystal contributes one electronic energy level to the energy band diagram, what is the density of states (DoS) of the bands?
**Answer**: The density of silicon is 2329 kg/m$^3$. The atomic mass is 28.0855 u. That is, 28.0855 g/mol. That is 28.0855 × $10^{-3}$ kg per (1 mol $\times N_\mathrm{A}$) atoms, with $N_\mathrm{A}$ Avogadro's constant (6.022 × $10^{23}$/mol). The DoS of states is thus (2329 kg/m$^3$) × (6.022 × $10^{23}$ atoms/mol) / (28.0855 × $10^{-3}$ kg/mol) × (2 levels / atom) = 9.99×$10^{22}$ m$^{-3}$.

### 3.5.1 Conduction in metals vs. conduction in semiconductors

There is a significant difference in the conduction mechanism of metals and semiconductors. It is true that both involve the movement of electrons – in the end it all boils down to the same thing – but the behavior is remarkably different.

In metals, there is a cloud of highly agile and mobile electrons on a background of positively charged static metal ions. The movement is so fast that we cannot discern individual electrons anymore and we see a negatively-charged cloud instead. In the absence of external forces (no voltage applied), there exist the same number of electrons moving in one direction as there exist electrons moving in the opposite direction; the net average speed is zero and no external current is observed. When a bias is applied, the electrons with positive speed

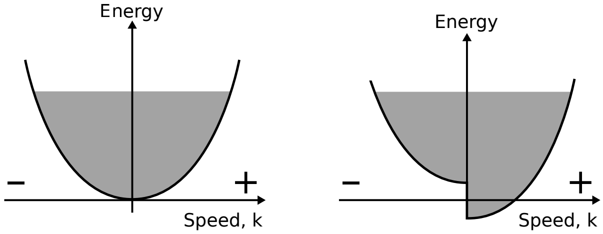


**Fig. 3.26**: Energy of electrons as a function of their speed in a metal. In absence of external force (voltage) the electrons with positive speed have the same energy as electrons of negative speed. The net average speed is zero (left). If a voltage is applied the electrons with positive speed become more abundant because they have lower energy. A net current results

are energetically favored, as if their energy is offset by a certain value. In Figure 3.26 this is schematically shown, where the energy of the electrons is plotted as a function of their speed. Because the electrons can easily occupy more levels on the positive-speed side – after all, the band was only have full and many levels are still available – there is no longer a balance between electrons with positive speed and negative speed. A net positive average velocity results which is effectively a current.

We have seen the conduction in a classical physics description of the Drude model, see Eq. (3.31). And we can now compare this to the situation in semiconductors. Again, like in metals, there is a cloud of highly mobile electrons on a background of a positively charged matrix. The difference, however, is that electrons cannot gain energy easily. In Figure 3.27 this is schematically shown. When the band with positive speed is offset, nothing happens, because the electrons do not have other states available; the band was filled to the rim and the electrons have nowhere to go but to remain in their original states. The result is that the balance in speed is maintained and the net average speed remains zero and no current is observed. This means that even when we apply a voltage, no current is observed. Semiconductors are insulators!

To make semiconductors conduct, electrons have to be somehow be injected into the conduction band. Once they are there, they can contribute to current, the same way electrons in metals can contribute to current. Alternatively, an electron can be taken away from the full (valence) band. This will leave behind a not completely full band that also can conduct. This is like a bubble of air in a otherwise full and closed box of water. If we incline the box (representing the application of a voltage), the bubble of air will move from one side to the other. What happens is that water (electrons) on average move in the opposite direction. The missing electron – the air bubble – is what is called a 'hole' in semiconductor physics jargon. Semiconductors can as easily conduct through electrons in a nearly empty conduction band as through 'holes' in a nearly full valence band. By the application of a voltage, we create a slope, a hill, from which the electron rolls down and the hole rolls up. Both movements represent

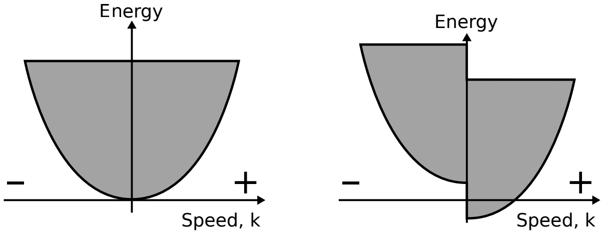


**Fig. 3.27**: Energy of electrons as a function of their speed in a semiconductor. In absence of external force (voltage) the electrons with positive speed have the same energy as electrons of negative speed. The net average speed is zero (left). If a voltage is applied the electrons with positive speed do not become more abundant, in spite of their lower energy, because no states are available; all states are filled to the rim. There remains an electron balance and no net current results

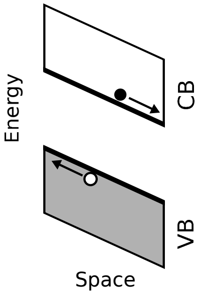


**Fig. 3.28**: Two ways of conduction. Electrons (•) roll down the slope created by the electric field by applying a bias. Missing electrons ('holes', ∘) bubble up the slope by the same bias

current.

There are basically four ways of creating electrons in the conduction band or holes in the valence band, see Fig. 3.29:

1. **Chemically**: The idea is to deliberately contaminate the material with chemically foreign dopants which either easily donate or accept electrons from the host material. This approach is what lays the basis for many semiconductor devices such as pn-junction diodes and bipolar transistors.

2. **Optically**: Electron-hole pairs can be created by the absorption of photons. The maximum of absorption of light occurs for photon energies close to the band gap.

3. **Electrically**: Electrical bias can modify the energetic band diagram in such a way as to allow for the presence of free carriers in the bands. As an example may serve the field-effect transistor. Even intrinsic materials can

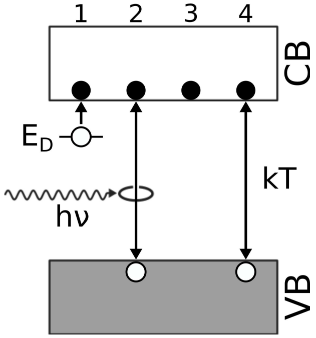


**Fig. 3.29**: Four ways of creating free charge carriers (holes ○ or electrons ●) contributing to external current: 1) Chemically, by adding donors to the material that can easily liberate electrons to the conduction band, thus creating free electrons ● and ionized donors, $N_{\mathrm{D}}^{+}$, 2) Optically, the energy of an absorbed photon can be used for the creation of electron-hole pairs ●−○, 3) Electrically, by changing the bias it can become energetically favorable to inject free electrons (or free holes), and 4) Thermally, very similar to optically, but the energy comes from the phonons (thermal vibrations) of the crystal.

be induced into conduction by the presence of an electric field. Likewise, with enough voltage, even the most insulating materials can be made to conduct current. For LEDs, this is the preferred method of carrier generation. Doping of the materials can result in non-radiative recombination paths that can kill the luminescence.

4. **Thermally**: This way of creating free charges in the bands is very similar to the optical method. With high enough temperature, the lattice vibrations can thermally excite electrons from the valence band to the conduction band.

Of these effects, the optical and thermal ones create an equal amount of free electrons and holes. The chemical and electrical ways can create a mismatch of carriers. The chemical way creates uniquely one type of carrier, either electrons or holes. In principle this also applies to electrical creation of carriers. The electrical way can create both carriers only at opposing electrodes in, for instance, LEDs. Moreover, the electrical creation of free carriers is the only way where charge neutrality is not maintained locally (throughout the entire device, charge neutrality is always conserved, of course).

Before we continue, it is good to make an approximation. While we can accurately calculate the number (and thus density) of levels in a band, see Equation 3.142, not all contribute to the following calculations. Some of the levels can be energetically too far away to have any significance. A more important quantity for semiconductor materials is the effective density of levels. This is a way of expressing effectively how many levels are available for conduction and not so much as how many there are in total. Moreover, these energy levels

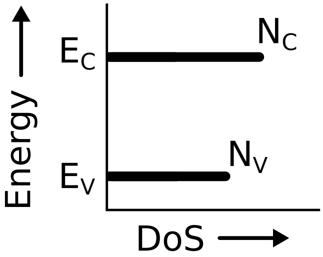


**Fig. 3.30**: Effective band diagram of a semiconductor, density-of-states (DoS) as a function of energy: $N_{\mathrm{C}}$ levels at $E = E_{\mathrm{C}}$ and $N_{\mathrm{V}}$ levels at $E = E_{\mathrm{V}}$

are by approximation all assumed to be all being at the same energy, see Figure 3.30:

The effective density-of-states (DoS) of a semiconductor can effectively be summarized by:

$$N_{\mathrm{C}} \text{ levels at } E = E_{\mathrm{C}} \text{ and } N_{\mathrm{V}} \text{ levels at } E = E_{\mathrm{V}},$$

see Figure 3.30. Note that these quantities have units $1/\mathrm{m}^3$.

### 3.5.2 The effect of temperature, the Fermi level

Electrons, like anything in nature, prefer to be in the lowest-energy state possible. As discussed above, this means that (at a temperature of 0 kelvin) all electrons are in the valence band, which is completely filled up, and no electrons are in the conduction band, which is completely empty. We can thus define an energy below which the states are occupied by electrons and states above which the states are empty. This energy is called the Fermi level, $E_{\mathrm{F}}$.

We can also define a function, called the Fermi-Dirac function $f(E)$, that is unity below the Fermi level and zero above it. The actual electrons, $n(E)$, with a certain energy $E$ is then given by the product of the density of levels available at that energy, $g(E)$, and this function:

$$n(E) = f(E)g(E). \tag{3.143}$$

We can visualize this as a box of marbles, see Figure 3.32. At low temperatures, the box is resting and the marbles are quietly lying at the bottom of the box. What happens when we increase the temperature is similar to shaking the box. The marbles start being agitated and no longer occupy the lowest energy possible. Some of them will jump up every now and then. The same occurs for our electrons in the materials. This is represented in the Fermi-Dirac distribution, which is no longer a Heaviside (step) function described above, instead it is

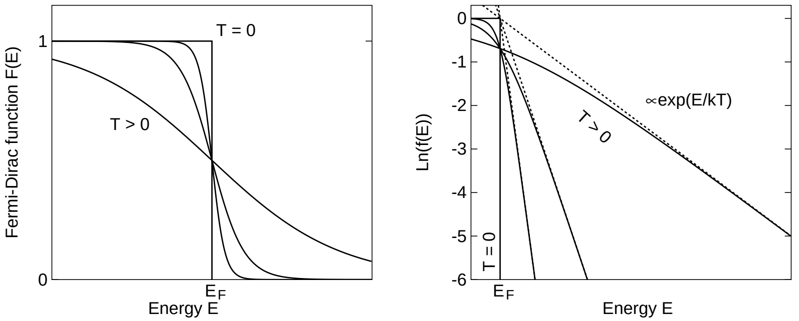


**Fig. 3.31**: Graphical representation of the Fermi-Dirac function of Equation (3.144) (left) and the Boltzmann approximation (dashed lines) for $E - E_\mathrm{F} \gg kT$ in the semilog-plot (right)

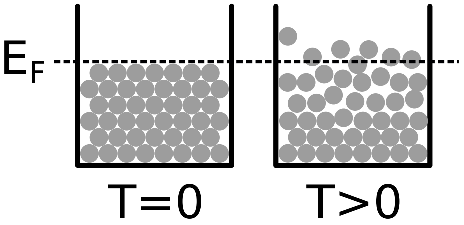


**Fig. 3.32**: Effect of temperature on electrons symbolized by shaking a box of marbles. At rest ($T = 0$) all the marbles (electrons) are at the lowest energy states, all are filling up the possibilities until $E_\mathrm{F}$. Raising the temperature agitates them and they are no longer all in the lowest possible energy state. Some of them are excited to higher states. The exact distribution is the Fermi-Dirac function of Equation (3.144)

defined by

$$f(E) = \frac{1}{1 + \exp\left(\frac{E - E_\mathrm{F}}{kT}\right)}. \tag{3.144}$$

We see this is a function of temperature. For $T$ approaching zero, the function goes back to the Heaviside function described before. Increasing the temperature causes a spreading of the step, see Figure 3.31. What this entails is a removing of some of the electrons at lower energies and placing them at higher energies. We call this 'thermal excitation'.

Note that at the Fermi level $f(E_\mathrm{F}) = 0.5$. Note also that the Fermi-Dirac functions is a purely mathematical tool to help us describe the behavior. No electron actually has to have this energy, nor do there have to be energy levels existing there. It is just an energy where the occupation of a level *theoretically* is half.

We can now calculate what the Fermi level would be in a typical semiconductor. We do this on basis of the following

- $N_\mathrm{C}$ levels at an energy $E_\mathrm{C}$
- $n$ electrons in these levels
- $N_\mathrm{V}$ levels at an energy $E_\mathrm{V}$
- $p$ electrons *missing* in these levels (in other words: $p$ holes *present* in these levels)
- The material is electrically neutral, $\rho = qp - qn = 0$, thus $n = p$

On basis of the above we first calculate the density of free electrons in the conduction band given a certain temperature and Fermi level:

$$n = \frac{N_\mathrm{C}}{1 + \exp\left(\frac{E_\mathrm{C} - E_\mathrm{F}}{kT}\right)}. \tag{3.145}$$

We then calculate the density of holes in the valence band. The amount of holes in the valence band is equal to the amount of valence-band levels minus the amount of these levels occupied with electrons:

$$p = N_\mathrm{V} - \frac{N_\mathrm{V}}{1 + \exp\left(\frac{E_\mathrm{V} - E_\mathrm{F}}{kT}\right)} = \frac{N_\mathrm{V}}{1 + \exp\left(\frac{E_\mathrm{F} - E_\mathrm{V}}{kT}\right)}. \tag{3.146}$$

These two have to be equal, and by combining the two equations above we can get an expression for the density of either one of them. It is easiest to use Boltzmann's approximation for the Fermi-Dirac function, namely removing the term '1' which is small compared to the exponential if $E_\mathrm{F}$ is far away from the bands, $E_\mathrm{C} - E_\mathrm{F} \gg kT$, and $E_\mathrm{F} - E_\mathrm{V} \gg kT$:

$$pn \approx N_\mathrm{C} N_\mathrm{V} \exp\left(\frac{-E_\mathrm{g}}{kT}\right), \tag{3.147}$$

with $E_\mathrm{g}$ the band gap, namely

$$E_\mathrm{g} = E_\mathrm{C} - E_\mathrm{V}. \tag{3.148}$$

The (intrinsic) Fermi level is close to the middle of the gap,

$$E_\mathrm{Fi} = \frac{E_\mathrm{C} + E_\mathrm{V}}{2} + \frac{kT}{2} \ln\left(\frac{N_\mathrm{V}}{N_\mathrm{C}}\right). \tag{3.149}$$

where the second term is negligible if the densities of states in the conduction band and valence band are comparable, which is normally the case. Finally, since $p = n$, we get

$$p = n = \sqrt{N_\mathrm{C} N_\mathrm{V}} \exp\left(\frac{-E_\mathrm{g}}{2kT}\right). \tag{3.150}$$

We see that, indeed, for intrinsic materials – that is, materials without doping – the material can be made conductive by heating it up:

$$\begin{aligned} \sigma &= qn\mu_\mathrm{n} + qp\mu_\mathrm{p} \\ &= q(\mu_\mathrm{n} + \mu_\mathrm{p})\sqrt{N_\mathrm{C} N_\mathrm{V}} \exp\left(\frac{-E_\mathrm{g}}{2kT}\right). \end{aligned} \tag{3.151}$$

**Table 3.VIII**: Properties of some important semiconductors

| Property | Symbol | Unit | Si | Ge | GaAs |
|---|---|---|---|---|---|
| Bandgap | $E_\mathrm{g}$ | eV | 1.12 | 0.66 | 1.42 |
| DoS (CB) | $N_\mathrm{C}$ | $\mathrm{cm}^{-3}$ | $2.8 \times 10^{19}$ | $1.04 \times 10^{19}$ | $4.7 \times 10^{17}$ |
| DoS (VB) | $N_\mathrm{V}$ | $\mathrm{cm}^{-3}$ | $1.04 \times 10^{19}$ | $6 \times 10^{18}$ | $7 \times 10^{18}$ |
| Mobility (electrons) | $\mu_\mathrm{n}$ | $\mathrm{cm}^2/\mathrm{Vs}$ | 1500 | 3900 | 8500 |
| Mobility (holes) | $\mu_\mathrm{p}$ | $\mathrm{cm}^2/\mathrm{Vs}$ | 450 | 1900 | 400 |
| Dielectric constant | $\varepsilon_\mathrm{r}$ | | 11.9 | 16.0 | 13.1 |
| Atom density | $N$ | $\mathrm{cm}^{-3}$ | $5.0 \times 10^{22}$ | $4.42 \times 10^{22}$ | $4.42 \times 10^{22}$ |
| Carrier concentration | $n = p$ | $\mathrm{cm}^{-3}$ | $1.45 \times 10^{10}$ | $2.4 \times 10^{13}$ | $1.79 \times 10^{6}$ |
| Resistivity | $\rho$ | $\Omega\mathrm{cm}$ | $2.3 \times 10^{5}$ | 47 | $2.25 \times 10^{3}$ |

We will later make use of this in temperature sensors. A very important conclusion we can draw at this stage is that semiconductors behave differently from metals in that they become *more* conductive when the temperature rises.

**Question**: Verify that the resistivity of silicon given in Table 3.VIII is consistent with the other parameters given there.
**Answer**: The conductivity as given by Eq. 3.151 is $\sigma = qn\mu_\mathrm{n} + qp\mu_\mathrm{p} = 4.53 \times 10^{-4}$ S/m and the resistivity thus $\rho = 1/\sigma = 2200\ \Omega$m, which is close to the value given in the table.

What happens if we add dopants to our material? For instance donors. Donors are impurities that can donate their electron to the host material. A donor can thus be either neutral, when this electron is still on the impurity, or positively charged, when the electron has been donated. We keep the restriction that the material must be neutral. That is, for every ionized donor, there must have been created a free electron in the conduction band. So, we add donors that are easily ionized, that is, their energy is quite close to the conduction band. We have the two new ingredients in our analysis:

- $N_\mathrm{D}$ donor levels at energy $E_\mathrm{D}$. Typically some tens of meV below the conduction band.
- The material remains electrically neutral, $p - n + N_\mathrm{D}^+ = 0$

Each of the quantities $p$, $n$, and $N_\mathrm{D}^+$ is given by the Fermi-Dirac function. Because the Fermi level can be close to the donor level, we can no longer use Boltzmann's approximation and an analytical solution is difficult to find. Yet, finding a graphical solution is commonly done and is in fact more informative. See Figure 3.33. The figure shows the densities of electrons, holes and ionized donors on a logarithmic scale as a function of Fermi level. Finding the Fermi then level consists of determining the charge-neutrality point, $p + N_\mathrm{D}^+ = n$.

Since the slopes of the curves of $n$ and $p$ and possibly also of $N_\mathrm{D}^+$ depend on temperature (slope $\propto 1/kT$), the crossing point depends on temperature.

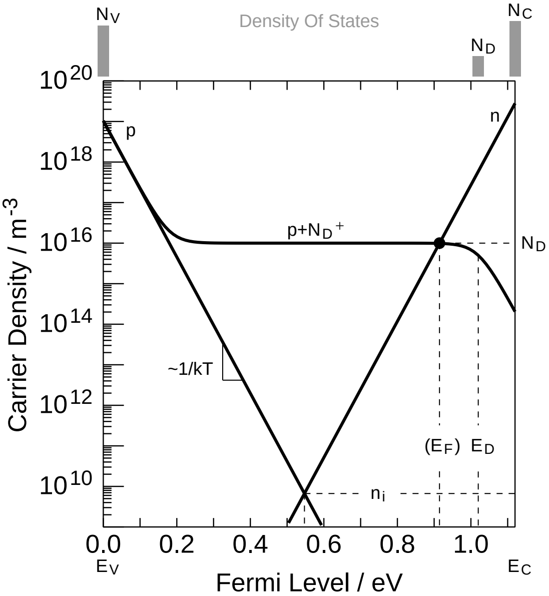


**Fig. 3.33**: Graphical method of finding the Fermi level and resulting carrier densities

Figure 3.33 shows an example. The Fermi level is given by the dot. It is clear that, in this case, $p \ll n$, i.e., holes play no role in the analysis at this point. Moreover, the Fermi-Dirac function for $N_\mathrm{D}^+$ is 1, meaning that all donors are ionized, $N_\mathrm{D}^+ = N_\mathrm{D}$. If we now increase the temperature, the slopes of the curves change. The charge-neutrality point moves to the left (the Fermi level drops), but the free-electron density remains the same. We are in a regime that is called 'saturation', since temperature has no effect on the densities. While the electron concentration is constant, the dependence of the position of the Fermi level on temperature in this regime is given by

$$E_\mathrm{F}(T) = E_\mathrm{C} - kT \ln\left(\frac{N_\mathrm{C}}{N_\mathrm{D}}\right). \tag{3.152}$$

This continues, until the temperature of the material has risen so much that the crossing point of the $n$ and $p$ curves rises above the donor level. In this case, the hole density starts playing a role again, and both $n$ and $p$ heavily depend on the temperature, as we have seen for the intrinsic material. The transition temperature at which this happens can easily be calculated by setting the intrinsic density of Equation (3.150) equal to the donor density, namely

$$T_\mathrm{is} = \frac{E_\mathrm{g}}{2k \ln\left(\frac{\sqrt{N_\mathrm{C} N_\mathrm{V}}}{N_\mathrm{D}}\right)}. \tag{3.153}$$

On the other hand, lowering the temperature starts having an effect when the Fermi level approaches the donor level. Lowering the temperature makes that the donors are no longer all ionized. Electrons return to their donors and we call this 'freeze-out'. The transition temperature between saturation and freeze-out is found as

$$T_{\text{fs}} = \frac{E_{\text{C}} - E_{\text{D}}}{2k \ln\left(\sqrt{\frac{N_{\text{C}}}{N_{\text{D}}}}\right)}, \tag{3.154}$$

below which the electron concentration is given by

$$n(T) = \sqrt{N_{\text{C}} N_{\text{D}}} \exp\left(-\frac{E_{\text{C}} - E_{\text{D}}}{2kT}\right). \tag{3.155}$$

Figure 3.34 summarizes this behavior on temperature, and Figure 3.35 the behavior of the Fermi level on temperature. We can recognize three zones in this figure: 1) For high temperatures, the material behaves intrinsic, that is, as if there were no donors. The densities of free carriers, and thus the conductivity of the material, heavily depend on temperature. 2) In the saturation regime, all donors are ionized, the hole density is negligible and the electron concentration is equal to the donor concentration. As a result, the material conductivity is independent of temperature. 3) In the freeze-out regime, the electrons return to their donors and the electron concentration and material conductivity depend weakly on temperature. We see that the resistivity of semiconductor materials indeed depends on the temperature in a negative way, NTC (negative temperature coefficient), if the material is intrinsic (undoped, $T_{\text{is}} = 0$, or generally if $T > T_{\text{is}}$) or if it is in the freeze-out regime, $T < T_{\text{fs}}$.

### 3.5.3 Optical properties

Luminescence in light-emitting diodes (LEDs) consists of electrons in the conduction band that fall back into the hole in the valence band, with the energy thus released (equal to the band gap $E_{\text{g}}$) converted into a photon of light with energy equal to $h\nu$ with $h$ Planck's constant and $\nu$ the frequency of the light, related to the wavelength by $\lambda = c/\nu$, with $c$ the speed of light. We can also call it electron-hole recombination and that is a simple concept, see Figure 3.36. A photo-detector is based on the opposite principle; a photon is absorbed and an electron is promoted from the valence band to the conduction band. It is called electron-hole-pair generation.

In practice, this idea is a little more complicated than that. When we look at real band structures of semiconductors, see Figure 3.37, we can somewhat recognize our simple calculations of band diagrams from before. The conduction band resembles very much the lower band of Figure 3.24, with the lowest energy at the smallest k values. In three dimensions, this minimum k-value is called Γ, while increasing values in a specific direction are called X or L. The valence band is an upside down conduction band, with the highest energy at the lowest k values.

We see that, upon closer inspection, the bands are a little more complicated than the simple parabolic functions. Moreover, they do not have extremes at

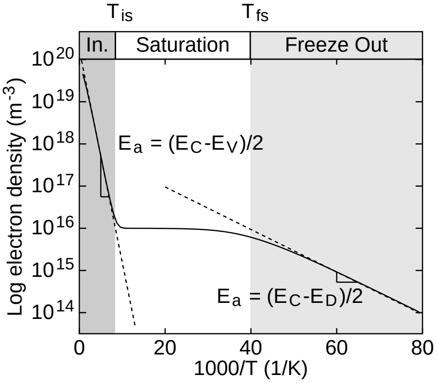


**Fig. 3.34**: An Arrhenius plot of carrier concentration ($n$ vs. $1/T$) in the presence of donors reveals three distinct regions. For high temperatures (left), the material behaves like undoped, intrinsic, material. Lowering the temperature levels the electron concentration at a value equal to the donor concentration, independent of temperature. All donors have donated their electron to the conduction band. This is called 'saturation'. Lowering the temperature further recaptures the electrons on the donors in what is called 'freeze-out'. The activation energy (slope in this plot) is given by $E_\mathrm{a} = E_\mathrm{g}/2 = (E_\mathrm{C} - E_\mathrm{V})/2$, 0 and $(E_\mathrm{C} - E_\mathrm{D})/2$ respectively

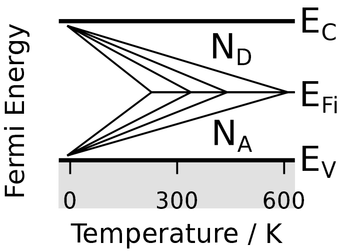


**Fig. 3.35**: Schematic diagram showing the dependence of Fermi level on temperature for four levels of doping levels (acceptors and donors, increasing from left to right). At high temperatures the Fermi level is fixed at the intrinsic value $E_\mathrm{Fi}$ of Eq. (3.149). At lower temperatures, in saturation, the Fermi level depends linearly on temperature with a slope proportional to the logarithm of $N_\mathrm{C}/N_\mathrm{D}$, or $N_\mathrm{V}/N_\mathrm{A}$ as in Eq. (3.152)

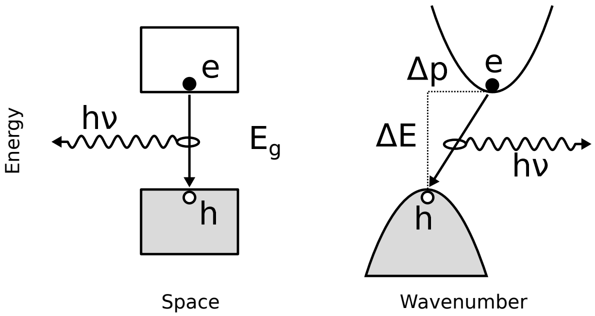


**Fig. 3.36**: Left: Luminescence (like in an LED) consists of electrons that fall back into the hole left behind in the valence band, where the gained energy $E_{\mathrm{g}}$ is converted into a photon with energy $h\nu$. Light absorption consists of the opposite process, namely the creation of electron-hole pairs, and this occurs if the energy of the incoming photon is large enough to promote an electron from the valence band to the conduction band. Right: Upon closer inspection, both energy and momentum need to be conserved in photon absorption and emission processes. If the bandgap is indirect, as shown here, both criteria cannot be met simultaneously, and the material is optically inactive

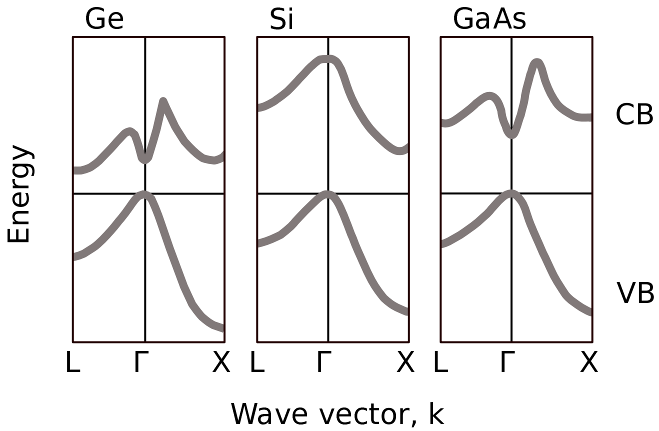


**Fig. 3.37**: Actual band structure of typical semiconductors, germanium, silicon and gallium arsenide in three dimensions. The labels L, Γ and X represent special values for the wave vector, with Γ: $k_{\mathrm{x}} = k_{\mathrm{y}} = k_{\mathrm{z}} = 0$, X: $k_{\mathrm{x}} = 2\pi/a$, $k_{\mathrm{y}} = k_{\mathrm{z}} = 0$, L: $k_{\mathrm{x}} = k_{\mathrm{y}} = k_{\mathrm{z}} = \pi/a$. It is clear that germanium and silicon are indirect bandgap materials, with the minimum of the conduction band not at the same place as the maximum of the valence band, while gallium arsenide is direct bandgap material. This has strong effect on the optical properties of the materials; of these shown, only a diode of GaAs will emit light

Γ. We see for instance that in silicon the minimum of the conduction band is not at the minimum k-values, Γ, but is close to the X-point instead. This has severe implications for the optical properties of the materials.

Imagine that we inject electrons into the device from one side and holes from the other side, as in a diode. The electrons go into the conduction band. And independently of where they started, they trickle to the bottom of the conduction band, close to the X point of the band structure. The holes, on the other hand, injected somewhere in the valence band, bubble up to the maximum of the valence band, at the Γ point, see the right side of Figure 3.36. The photon should carry away the energy, $\Delta E = E_{\mathrm{g}}$, but also the excess momentum, $\Delta p = h\Delta k$.

The first condition is easily met. For a silicon diode the bandgap is 1.15 eV ($1.84 \times 10^{-19}$ J) and the energy of a photon is $E = h\nu$, we thus find a frequency of the light equal to $\nu = 278$ THz and a wavelength of $\lambda = c/\nu = 1.08$ µm, i.e., infrared.

The second condition is as good as impossible to meet. The k-value of the minimum of the conduction band is at X, which means $k_{\mathrm{x}} = \pi/a$ and $k_{\mathrm{y}} = k_{\mathrm{z}} = 0$. With an inter-atomic distance of about 5 Å, the wave number is there $k = 6.3 \times 10^9/\mathrm{m}$. The maximum of the valence band is at $k = 0$ and we have $\Delta p = h \times 6.3 \times 10^9/\mathrm{m}$. The momentum of a photon is given by $p = h/\lambda$ and we thus need a wavelength of $\lambda = 1/\Delta k = 0.16$ nm to satisfy the conservation-of-momentum restriction. This is a value that lies in the X-ray part of the electromagnetic spectrum and is inconsistent with the earlier found restriction of energy conservation that gave values in the infrared range.

The conclusion is that for a semiconducting material to be able to emit light, to be 'optically active', the minimum of the conduction band has to be aligned with the top of the valence band. We call this direct-bandgap materials. A good example is gallium arsenide, see the band structure shown in the right of Figure 3.37. GaAs can thus be used in LEDs. Silicon cannot.

### 3.5.4 Semiconductor devices. pn-junction diodes

In the chapter on electronics we have seen how a diode is a device that has an exponential dependence of current on voltage, and how it could be used to rectify a voltage, since it effectively passes current in only one direction. Here we will see how a diode can be made by joining p-type semiconductor with acceptors to n-type material with donors. Figure 3.38 shows an example of a pn-junction before and after contact. The p-type material is the material doped with acceptors. If we assume saturation, all of the acceptors have accepted an electron from the valence band and are ionized $N_{\mathrm{A}}^{-} = N_{\mathrm{A}}$. The material is neutral because the number of holes in the valence band is equal to the number of ionized acceptors, $p = N_{\mathrm{A}}^{-} = N_{\mathrm{A}}$. Next to it we have n-type material with $N_{\mathrm{D}}$ donors that have all donated their electrons to the conduction band, they are all ionized and the material is neutral because the electron density is equal to the ionized donor density, $n = N_{\mathrm{D}}^{-} = N_{\mathrm{D}}$.

When the two materials are intimately joined, electrons will jump from the

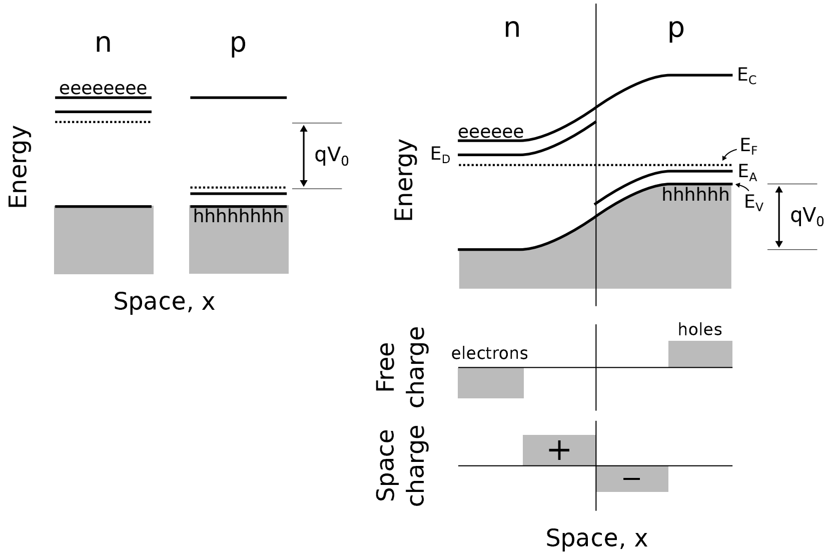


**Fig. 3.38**: Formation of a pn-junction diode. Left: Before contact the p-type material is full of acceptors that are all ionized. In the n-type material, all donors have donated their electrons to the conduction band. When contacting, energy can be gained by electrons jumping from the n-type to the p-type. The dopants are no longer compensated by free carriers and a space-charge region exists that is 'depleted of free carriers'. The space charge also causes a voltage drop and electrical field that forces the carriers back. In equilibrium the two phenomena are of equal intensity. In this situation the Fermi level is aligned and no more current flows

n-type side to the p-type side for two reasons. First, because they can gain energy like that. Note the difference in Fermi level on both sides of the barrier; electrons can gain energy by falling from the conduction band on one side into the holes on the other side. Second, diffusion is always trying to equalize densities. On the left side there is a high density of electrons, on the right side a low density. Diffusion will effectively involve a movement of electrons from left to right. Being helped by both diffusion and drift.

The result of this movement of electrons from n-type to p-type (and likewise holes from p-type side to n-type side) is that the p-type material becomes negatively charged and the n-type positively charged, i.e., we have a space-charge $\rho(x)$ in a region at the interface, called the 'depletion zone', since this space is depleted of free charge. According to Poisson's equation (Eq. 3.24), a space charge causes a voltage drop, and since the electron energy is given by $U = -qV$, an electron-energy rise. The transfer of electrons from one side to the other continues until the energy rise is equal to the difference in Fermi levels before contacting. In other words, the Fermi level has 'aligned' on both sides. We have the situation as in the right side of Figure 3.38.

To describe such a diode, normally the following reasoning is used. On the electrons actuate two 'forces'. The first force is the diffusion that tries to equalize the densities. In the figure, electrons travel from left-to-right, following the (negative) gradient of density, representing a negative diffusion current

$$J_{\mathrm{Dn}} = qD_{\mathrm{n}}\frac{\partial n}{\partial x} \tag{3.156}$$

On the other hand, electrons in the conduction band at the interface, want to roll down the hill, following the (positive) electric field, i.e., from right-to-left, representing a positive drift current

$$J_{\mathrm{En}} = q\mu_{\mathrm{n}} n E(x). \tag{3.157}$$

We can make use of the Einstein relation, $D = \mu kT/q$, and find a total electron current of

$$J_{\mathrm{n}} = J_{\mathrm{Dn}} + J_{\mathrm{En}} = q\mu_{\mathrm{n}}\left[n(x)E(x) + \frac{kT}{q}\frac{\partial n(x)}{\partial x}\right]. \tag{3.158}$$

Moreover, since in the Boltzmann approximation the density of free electrons is given by

$$n(x) = N_{\mathrm{C}}\exp\left[\frac{E_{\mathrm{F}}(x) - E_{\mathrm{C}}(x)}{kT}\right]. \tag{3.159}$$

The space-derivative can easily be found as

$$\frac{\partial n(x)}{\partial x} = \frac{n(x)}{kT}\frac{\partial[E_{\mathrm{F}}(x) - E_{\mathrm{C}}(x)]}{\partial x}. \tag{3.160}$$

But, the electric field is given as the derivative of potential, $E \equiv -\mathrm{d}V(x)/\mathrm{d}x$, and the derivative of conduction band energy, as discussed above ($U = -qV$), can be found as $\mathrm{d}E_{\mathrm{C}}(x)/\mathrm{d}x = -q\mathrm{d}V(x)/\mathrm{d}x$ and by substituting everything in Eq. (3.158) we arrive at an electron current given by

$$J_{\mathrm{n}} = \mu_{\mathrm{n}} n(x)\frac{\partial E_{\mathrm{F}}(x)}{\partial x}. \tag{3.161}$$

A similar analysis can be made for the hole current,

$$J_{\mathrm{p}} = \mu_{\mathrm{p}} p(x)\frac{\partial E_{\mathrm{F}}(x)}{\partial x}. \tag{3.162}$$

An important conclusion. The current is proportional to the density multiplied by the Fermi level gradient. It also directly means that if there is no net current, $J_{\mathrm{n}} = J_{\mathrm{p}} = 0$, the Fermi level is flat, exactly as shown in the figure. Either teh Fermi level is flat, or there are no free charges (as in an insulator).

In the equilibrium condition, as shown before, the product of electron and hole densities is constant everywhere (only depends on temperature). If, for some reason, the equilibrium is disturbed by for instance the injection of carriers

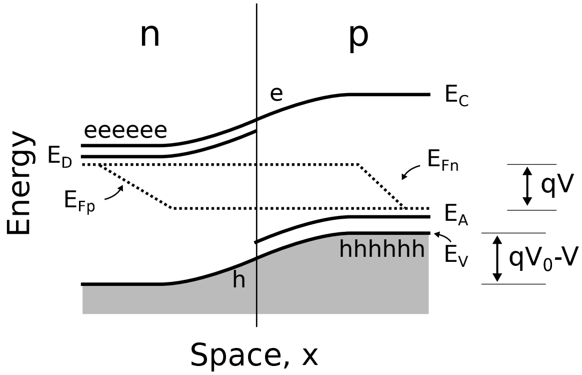


**Fig. 3.39**: Diode pn-junction in forward bias. The slope of the hill has reduced and carriers are injected to the other side. The product of electron density and hole density is no longer constant and this necessitates the definition of a quasi Fermi level $E_{\text{Fn}}$ and $E_{\text{Fp}}$ for either

from one side of the junction to the other by the application of a bias, there will be a reaction going on of the type

$$\mathrm{e} + \mathrm{h} \rightleftharpoons \text{nothing} + E_{\text{g}}. \tag{3.163}$$

I.e., electrons and holes will be generated or annihilated in an attempt to restore the equilibrium product $np$. There now exists another phenomenon that changes the density of charges, apart from drift (proportional to density and electric field) and diffusion (proportional to density gradient), namely generation-annihilation. Important to note is that this reaction is rather slow, of the order of microseconds. In that time, electrons and holes can travel a long distance, easily much longer than the depletion width. Electrons and holes can easily make it to the other side alive.

When a bias is applied, the voltage drop is changed. The balance of carrier movement no longer exists and a net current results, as we will now see. When this hill is reduced by a forward bias a situation as in Figure 3.39 exists. Electrons are injected from the electron-rich n-type material to the electron-lean p-type material representing a current $J_{\text{n}+}$. The distributions of electrons is therefore brought strongly off-equilibrium by $J_{\text{n}+}$. Equilibrium is restored by two processes, diffusion and recombination, respectively, both trying to reduce the electron density (see Figure 3.40),

$$\frac{\partial n(x,t)}{\partial t} = -D_{\text{n}} \frac{\partial^2 n(x,t)}{\partial x^2} \tag{3.164}$$

$$\frac{\partial n(x,t)}{\partial t} = -\frac{n(x,t)}{\tau_{\text{n}}}, \tag{3.165}$$

with the first equation Fick's law of diffusion. The steady state solution, inde-

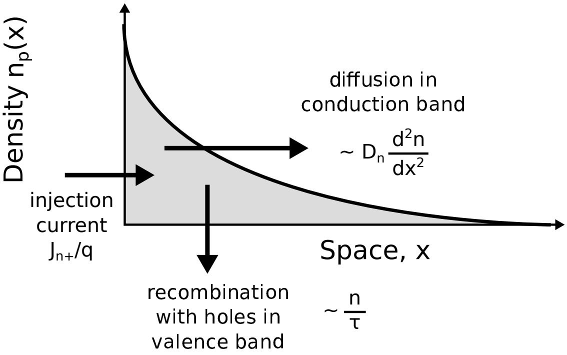

Density $n_p(x)$
diffusion in conduction band
$\sim D_n \frac{d^2n}{dx^2}$
injection current $J_{n+}/q$
Space, x
recombination with holes in valence band
$\sim \frac{n}{\tau}$

**Fig. 3.40**: Electrons are injected from the electron-abundant n-type into the other side of the junction equal to a current $J_{\mathrm{n}+}$. Electrons disappear from there by two processes, diffusion and recombination with holes. In steady state, what-comes-in-must-go-out, the injection is equal to the disappearance

pendent of time ($\partial n/\partial t = 0$) is given by

$$n(x) = n_{\mathrm{p}} \exp\left(-\frac{x}{\sqrt{D_{\mathrm{n}}\tau_{\mathrm{n}}}}\right), \tag{3.166}$$

with $n_{\mathrm{p}}$ the minority-carrier electron density in the beginning of the p-type material. We also require that in steady state the amount of electrons injected ('created') into the p-type material is equal to the ones that are annihilated by recombination (note that the diffusion does not make them disappear)

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \int_0^\infty -\frac{n(x)}{\tau_{\mathrm{n}}}\mathrm{d}x = -n_{\mathrm{p}}\sqrt{\frac{D_{\mathrm{n}}}{\tau_{\mathrm{n}}}}. \tag{3.167}$$

These disappearing electrons are continuously replenished by the injection current $J_{\mathrm{n}+}$ and we find that the current is

$$J_{\mathrm{n}+} = -q\frac{\mathrm{d}n}{\mathrm{d}t} = qn_{\mathrm{p}}\sqrt{\frac{D_{\mathrm{n}}}{\tau_{\mathrm{n}}}}. \tag{3.168}$$

Likewise, there is a (majority) electron current electrons from n-type to p-type side. But, although they are more numerous ($n_{\mathrm{n}}$), in this case, only those electrons with enough initial kinetic energy can make it up the hill of height $q(V_0 - V)$. This current is thus given by

$$J_{\mathrm{n}-} = -qn_{\mathrm{n}}\sqrt{\frac{D_{\mathrm{n}}}{\tau_{\mathrm{n}}}} \exp\left[-\frac{q(V_0 - V)}{kT}\right]. \tag{3.169}$$

From earlier calculations we know that the minority carrier electron density in the p-type material is equal to the majority carrier density on the other (n-type) side with a factor equal to the exponent of the Fermi level depth difference

$\Delta E_\mathrm{F} = qV_0$, namely $n_\mathrm{p}/n_\mathrm{n} = \exp(-qV_0)$. The total current then becomes

$$J_\mathrm{n} = J_\mathrm{n+} + J_\mathrm{n-} = qn_\mathrm{p}\sqrt{\frac{D_\mathrm{n}}{\tau_\mathrm{n}}}\left[\exp\left(\frac{qV}{kT}\right) - 1\right]. \tag{3.170}$$

Similarly we can find the hole current,

$$J_\mathrm{p} = qp_\mathrm{n}\sqrt{\frac{D_\mathrm{p}}{\tau_\mathrm{p}}}\left[\exp\left(\frac{qV}{kT}\right) - 1\right], \tag{3.171}$$

to find a total current given by the famous Schockley equation,

$$J = J_\mathrm{s}\left[\exp\left(\frac{V}{V_\mathrm{T}}\right) - 1\right], \tag{3.172}$$

with the reverse-bias saturation current given by

$$J_\mathrm{s} = qn_\mathrm{p}\sqrt{D_\mathrm{n}/\tau_\mathrm{n}} + qp_\mathrm{n}\sqrt{D_\mathrm{p}/\tau_\mathrm{p}}, \tag{3.173}$$

and the thermal voltage given by

$$V_\mathrm{T} = \frac{kT}{q}, \tag{3.174}$$

the latter being approximately 26 mV at room temperature and is also sometimes called diffusion voltage for reasons that are apparent here.

Because the product of electron density and hole density is no longer constant, holes and electrons no longer have the same Fermi level. Each has its own, often called quasi Fermi level, $E_\mathrm{Fn}$ and $E_\mathrm{Fp}$ for electrons and holes respectively. They are defined as *that* energy of the Fermi level that would result in the same amount of free carriers of that type,

$$n = N_\mathrm{C}\exp\left(\frac{E_\mathrm{Fn} - E_\mathrm{C}}{kT}\right), \tag{3.175}$$

$$p = N_\mathrm{V}\exp\left(\frac{E_\mathrm{V} - E_\mathrm{Fp}}{kT}\right). \tag{3.176}$$

In Figure 3.39 they are schematically plotted. The Fermi level is rather flat, until well on the other side of the interface where, by electron-hole recombination, the injected carriers resume their normal statistics and the two Fermi levels join again. Note that where the Fermi levels for holes and electrons are different, no equilibrium exists.

### Capacitance of a pn junction

Upon closer analysis, the diode has a capacitance. Every time the external bias is changed, the depletion width on either side changes and more or less space charge exists around the interface. This charge is supplied by the external circuit, just like in a normal capacitor. As an example, if the diode is placed in forward bias, with a positive voltage on the p-side relative to the n-side, the band bending reduces to $V_{\rm bi} - V$, the depletion width shrinks and less uncompensated acceptors (more holes) on the p-side and less uncompensated donors (more electrons) on the n-side. In other words, holes and electrons move into the device towards the interface. This moving of charge into the device is exactly the same as capacitance. As will be shown here, this capacitance depends on the bias and the diode works as a varicap.

The charge density $\rho(x) = -qN_{\rm A}$ or $qN_{\rm D}$, electric field $E(x) = \int \rho(x)/\varepsilon {\rm d}x$, potential $V(x) = \int E(x){\rm d}x$ and electron energy $U(x) = -qV(x)$ are shown in Figure 3.41. A non-symmetric diode is shown, with different densities of donors and acceptors, this results in different depletion widths on both sides, $W_{\rm n}$ and $W_{\rm p}$ respectively. Because the device has to be overall neutral, we have the condition that $W_{\rm n}N_{\rm D} = W_{\rm p}N_{\rm A}$. The electric field is the integral of the charge density and has a maximum at the interface equal to

$$\text{left side} : E(x) = \frac{qN_{\rm D}}{\varepsilon_{\rm s}}(x + W_{\rm n}) \tag{3.177}$$

$$\text{right side} : E(x) = \frac{qN_{\rm A}}{\varepsilon_{\rm s}}(W_{\rm p} - x). \tag{3.178}$$

The potential on both sides of the interface are given by

$$\text{left side} : V(x) = -\frac{qN_{\rm D}}{2\varepsilon_{\rm s}}(x + W_{\rm n})^2 \tag{3.179}$$

$$\text{right side} : V(x) = -V_{\rm bi} + \frac{qN_{\rm A}}{2\varepsilon_{\rm s}}(x - W_{\rm p})^2, \tag{3.180}$$

with $V_{\rm bi}$ the built-in voltage, the total voltage drop across the interface in the absence of bias. At the interface ($x = 0$) these two should be equal, and we find an expression for the built-in voltage:

$$V_{\rm bi} = \frac{qN_{\rm D}}{2\varepsilon_{\rm s}}(W_{\rm n}^2 + W_{\rm p}^2). \tag{3.181}$$

If bias is applied, the total band bending changes the above equation becomes

$$V_{\rm bi} - V = \frac{qN_{\rm D}}{2\varepsilon_{\rm s}}(W_{\rm n}^2 + W_{\rm p}^2). \tag{3.182}$$

Now substitute the charge neutrality condition, $W_{\rm p}N_{\rm A} = W_{\rm n}N_{\rm D}$, in the form $W_{\rm p}^2 = W_{\rm n}^2 N_{\rm D}^2/N_{\rm A}^2$ into the equation for the band bending $V_{\rm bi} - V$ (Eq. 3.182),

$$V_{\rm bi} - V = \frac{q}{2\varepsilon_{\rm s}} W_{\rm n}^2 \left( \frac{N_{\rm D}^2}{N_{\rm A}} + N_{\rm D} \right), \tag{3.183}$$

or

$$W_{\rm n} = \sqrt{\frac{2\varepsilon_{\rm s}(V_{\rm bi} - V)}{q} \frac{N_{\rm A}/N_{\rm D}}{N_{\rm D} + N_{\rm A}}}. \tag{3.184}$$

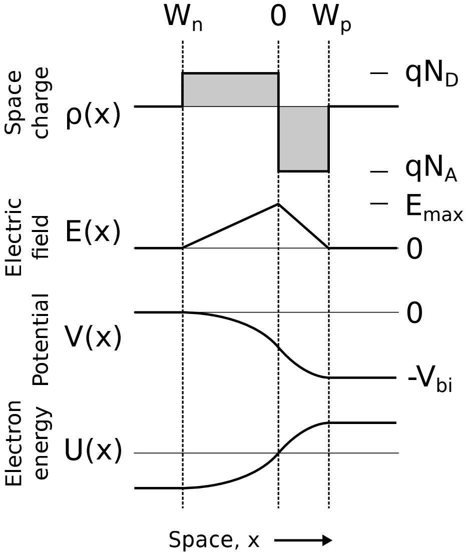


**Fig. 3.41**: pn-junction diode space charge ρ$(x)$, electric field $E(x)$, and voltage $V(x)$

Likewise, substituting $W_{\mathrm{n}}^2 = W_{\mathrm{p}}^2 N_{\mathrm{A}}^2 / N_{\mathrm{D}}^2$,

$$W_{\mathrm{p}} = \sqrt{\frac{2\varepsilon_{\mathrm{s}}(V_{\mathrm{bi}} - V)}{q} \frac{N_{\mathrm{D}}/N_{\mathrm{A}}}{N_{\mathrm{D}} + N_{\mathrm{A}}}}, \tag{3.185}$$

and the total width is then

$$\begin{aligned} W(V) = W_{\mathrm{p}} + W_{\mathrm{n}} &= \sqrt{\frac{2\varepsilon_{\mathrm{s}}(V_{\mathrm{bi}} - V)}{q(N_{\mathrm{D}} + N_{\mathrm{A}})}} \left( \sqrt{\frac{N_{\mathrm{D}}}{N_{\mathrm{A}}}} + \sqrt{\frac{N_{\mathrm{A}}}{N_{\mathrm{D}}}} \right) \\ &= \sqrt{\frac{2\varepsilon_{\mathrm{s}}}{q} \frac{(N_{\mathrm{A}} + N_{\mathrm{D}})}{N_{\mathrm{A}} N_{\mathrm{D}}} (V_{\mathrm{bi}} - V)}. \end{aligned} \tag{3.186}$$

The (dynamic) capacitance of a device is given as the derivative of the amount of charge as a function of voltage,

$$Q \equiv \frac{\mathrm{d}Q(V)}{\mathrm{d}V}. \tag{3.187}$$

We can exactly calculate how much charge is in the device. On the left side we have $qN_{\mathrm{D}}$ times the left-side depletion width $W_{\mathrm{n}}$,

$$\begin{aligned} Q_{+} &= qW_{\mathrm{n}}N_{\mathrm{D}} = qN_{\mathrm{D}} \sqrt{\frac{2\varepsilon_{\mathrm{s}}(V_{\mathrm{bi}} - V)}{q} \frac{N_{\mathrm{A}}/N_{\mathrm{D}}}{N_{\mathrm{D}} + N_{\mathrm{A}}}} \\ &= \sqrt{2q\varepsilon_{\mathrm{s}}(V_{\mathrm{bi}} - V) \frac{N_{\mathrm{A}} N_{\mathrm{D}}}{N_{\mathrm{D}} + N_{\mathrm{A}}}}. \end{aligned} \tag{3.188}$$

On the right side we have $-qN_\mathrm{A}$ times the right-side depletion width $W_\mathrm{p}$,

$$\begin{aligned} Q_- &= -qW_\mathrm{p}N_\mathrm{A} = -qN_\mathrm{A}\sqrt{\frac{2\varepsilon_\mathrm{s}(V_\mathrm{bi}-V)}{q}\frac{N_\mathrm{D}/N_\mathrm{A}}{N_\mathrm{D}+N_\mathrm{A}}} \\ &= -\sqrt{2q\varepsilon_\mathrm{s}(V_\mathrm{bi}-V)\frac{N_\mathrm{A}N_\mathrm{D}}{N_\mathrm{D}+N_\mathrm{A}}}. \end{aligned} \tag{3.189}$$

We see that indeed the amount of charge in the depletion width is balanced, with as much positive charge as negative charge (overall the device is neutral). When we substitute this in the definition of the capacitance above, we find

$$C = \frac{\mathrm{d}Q_-(V)}{\mathrm{d}V} = -\frac{\mathrm{d}Q_+(V)}{\mathrm{d}V} = \sqrt{\frac{2q\varepsilon_\mathrm{s}}{(V_\mathrm{bi}-V)}\frac{N_\mathrm{A}N_\mathrm{D}}{N_\mathrm{D}+N_\mathrm{A}}}. \tag{3.190}$$

This can be expressed in terms of the depletion width,

$$C(V) = \frac{\varepsilon_\mathrm{s}}{W(V)}. \tag{3.191}$$

In other words, the diode behaves like a normal metal-plates capacitor we have seen before, with $Q_+$ on one electrode and $Q_-$ on the other and with a distance between the plates equal to the total depletion width. Note also that the capacitance is not constant, but depends on the bias. With the bias we can effectively modulate the distance between the plates. This is what is called a varicap. The capacitance has a reciprocal square-root behavior and shoots to infinity for biases equal to the built-in voltage. Moreover, we can determine the doping densities by a so-called Mott-Schottky plot of $C^{-2}$ vs. $V$, especially when one of the densities is much smaller than the other, in a one-sided junction.

# 3.6 Semiconductor devices

## 3.6.1 Solar cells / optical detectors

An interesting device is a pn-junction that is illuminated. As seen before, illumination – absorption of photons – causes the creation of electron hole pairs if the energy of the photons is large enough. In the interface, the electrical field break them apart, and they drift towards the opposite electrodes, where they are collected and contribute to an external current.

## 3.6.2 Schottky (metal-semiconductor) diode

A Schottky diode is a diode like a pn-junction with a metal on one side instead of the semiconductor. The difference between a metal and a semiconductor is that a metal has plenty available carriers of both types, holes and electrons. Now we can repeat the calculation of the pn-junction, and for instance replace $N_\mathrm{D}$ with an infinite value and we get the correct behavior and correct parameters, such

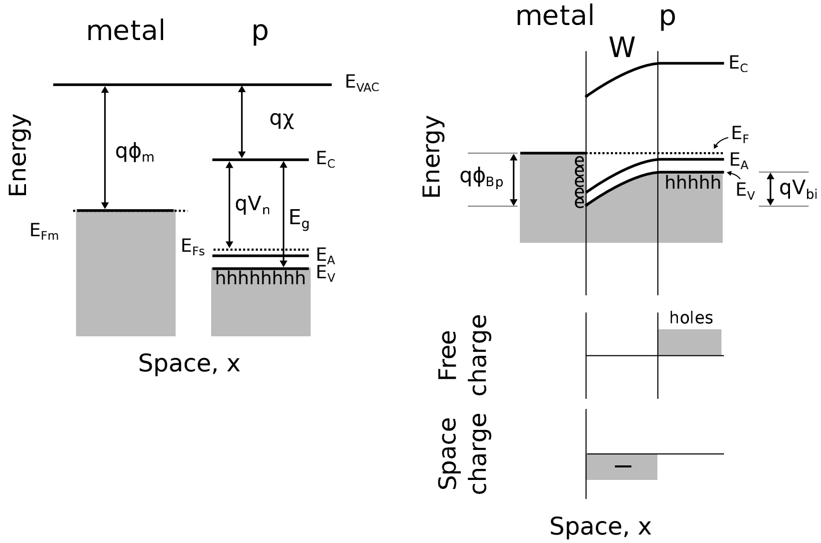


**Fig. 3.42**: Schottky barrier before and after contact

as depletion width (in the semiconductor; in the metal the depletion width is close to zero), etc. We can do a slightly modified analysis.

Before contacting (see Figure 3.42 that gives an example of a Schottky diode made of a metal a p-type semiconductor), the electrons in the metal at the top of the electron sea, at the Fermi level, have a certain energy relative to the vacuum level. This specifies the minimum energy it would cost to remove an electron from the metal. This is called the workfunction in jargon $q\phi_{\mathrm{m}}$. A similar energy exists for the energy to take an electron from the bottom of the conduction band, $q\chi$. Another parameter specifies the depth of the Fermi level in the energy gap, $qV_{\mathrm{n}}$.

Now we contact the two sides and, like in the pn-junction, charge will jump from one side to the other, in this case electrons will jump from the metal to the semiconductor, there recombining with the free holes of the valence band and leaving behind a space charge region of width $W$, see Figure 3.42. From this picture we can see that the amount of band bending – the built-in voltage – is given by

$$V_{\mathrm{bi}} = \chi + V_{\mathrm{n}} - \phi_{\mathrm{m}}. \tag{3.192}$$

This is the barrier as seen by (majority) carriers (holes in the case of p-type material) coming from the semiconductor traveling towards the metal. Moreover, we see a sharp drop at the interface equal to

$$q\phi_{\mathrm{Bp}} = (q\chi + E_{\mathrm{g}}) - q\phi_{\mathrm{m}}. \tag{3.193}$$

This is the barrier as seen by (majority) carriers coming from the metal going into the semiconductor.

We can make a similar calculation for the depletion width as the one made for the pn-junction (in fact, simpler), including space charge density, band bendings given by the Poisson equation and that should be equal to the Fermi-level difference before contact, etc. The result is a depletion width, total space charge density and capacitance (density) given by respectively

$$\begin{aligned} W(V) &= \sqrt{\frac{2\varepsilon_\mathrm{s}(V_\mathrm{bi}-V)}{qN_\mathrm{A}}}, && (3.194)\\ Q(V) &= qN_\mathrm{A}W = \sqrt{2q\varepsilon_\mathrm{s}N_\mathrm{A}(V_\mathrm{bi}-V)}, && (3.195)\\ C(V) &\equiv \frac{\mathrm{d}Q(V)}{\mathrm{d}V} = \sqrt{\frac{q\varepsilon_\mathrm{s}N_\mathrm{A}}{2(V_\mathrm{bi}-V)}} && (3.196)\\ &= \frac{\varepsilon_\mathrm{s}}{W}. && (3.197) \end{aligned}$$

The last equation shows that, once again, we have a device that behaves as a metal-plates capacitor, filled with dielectric material $\varepsilon_\mathrm{s}$ with a distance between the plates given by $W$.

For the current normally thermionic-emission-diffusion theory is used that takes into account that charges have to overcome the barrier by (thermal) kinetic energy, an effect that is balanced by a process of diffusion. The derivation of this current is well explained in the book of Sze (see the bibliography). The total current (density) is given by

$$J(V) = J_\mathrm{S}\left[\exp\left(\frac{qV}{kT}\right)-1\right], \quad (3.198)$$

with the reverse-bias saturation current given by

$$J_\mathrm{S} = A^{**}T^2\exp\left(-\frac{q\phi_\mathrm{Bp}}{kT}\right), \quad (3.199)$$

in which $A^{**}$ is called the effective Richardson constant. Note that the current of a Schottky diode depends strongly, exponentially, on temperature and it can have a positive or negative coefficient.

### 3.6.3 Bipolar junction transistor

A bipolar junction transistor consists of very thin layer of material embedded between two layers of the same material with opposing doping type. This thin layer, the 'base', controls the passage of current between the other two electrodes, called 'emitter' and collector', see Figure 3.43.

The working of a bipolar transistor is based on the fact that the base layer is extremely thin. Had the base layer been thick, the device would be indistinguishable from back-to-back connected pn-diodes. In a thin base, the charges can to make it to the other side before they recombine. This is the basic principle of the bipolar transistor.

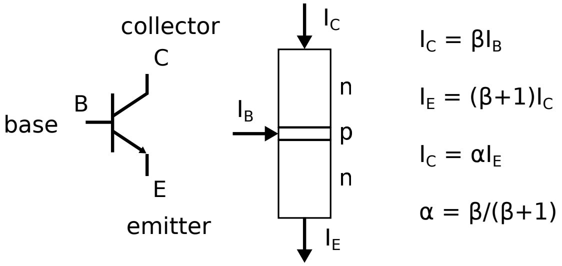


**Fig. 3.43**: npn-junction transistor. Small currents of the base, $I_\mathrm{B}$ modulate a large current at the collector, $I_\mathrm{C}$. The emitter current is the sum of the two, following Kirchhoff's law

By placing the base-emitter pn-junction diode in forward bias, approximately 0.7 volt, this diode starts conducting. Figure 3.44 shows a schematic diagram of a npn transistor without bias and under normal operation. When the transistor is polarized well, large quantities of electrons are injected from the emitter into the base and to a lesser extent holes from the base into the emitter. On the other hand, the collector-base junction is reverse biased ($V_\mathrm{C} > V_\mathrm{B}$) and the current is a tiny trickle, $J_\mathrm{s}$, as seen before for the pn-junction. However, electrons injected from the emitter into the base have a long lifetime, as we have seen before in the diode. They can easily diffuse to the other end of the base region. Their lifetime and diffusion length $L = \sqrt{D\tau}$ is further enhanced by minimizing the p-type doping in the base, thus limiting the probability of an electron finding a hole to recombine with. Once they reach the other side of the base and enter into the collector, they are rapidly swept away by the strong field in this reverse-biased region. In other words, nearly all of the electrons injected by the emitter make it to the other side of the base into the collector. That is, nearly all of the emitter current leaves the device though the collector, nearly nothing comes out at the base. Typically, the collector current is more than a hundred times bigger than the base current. The interesting thing about a bipolar transistor is that this ratio is as good as constant. The collector-base voltage is rather irrelevant, since the efficiency of the collector collecting electrons is not very much dependent on the field at this region.

In conclusion, by changing the base-emitter voltage drop, we control the amount of electrons being injected (emitted) by the emitter and a constant fraction $\alpha$ of these electrons (typically 99%) leave the transistor as a collector current. The rest (typically 1%) leaves the transistor as a base current. Seen from another point of view, the collector current is always a fixed multiple $\beta$ (typically 100) of the base current. In other words, even though the currents are controlled by the voltage drop at the base-emitter junction, the transistor can be seen as a *current* amplifier, with $I_\mathrm{C}$ a multiple of $I_\mathrm{B}$.

Figure 3.45 shows the electronic model of the bipolar npn transistor developed by Ebers and Moll. It consists of two diodes connected back-to-back and two current sources $\alpha_\mathrm{N} I_\mathrm{F}$ and $\alpha_\mathrm{I} I_\mathrm{R}$. The base-emitter diode is forward biased

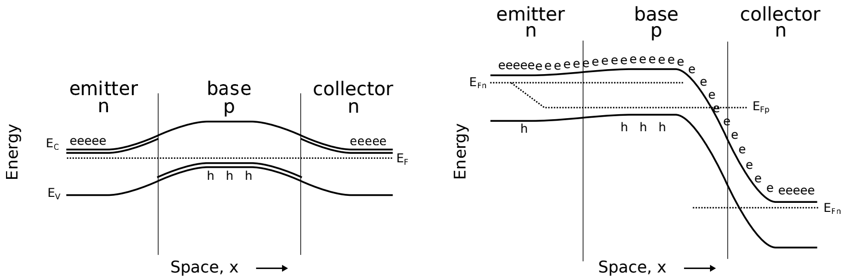


**Fig. 3.44**: npn bipolar junction transistor without bias (left) and in normal operation, with $V_{\mathrm{BE}} = 0.7$ V and $V_{\mathrm{CB}} > 0$

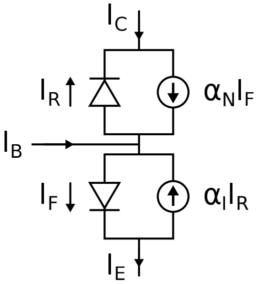


**Fig. 3.45**: Ebers-Moll electronic model of the bipolar npn-junction transistor

and has a current

$$I_{\mathrm{F}} = I_{\mathrm{F0}} \left[ \exp\left( \frac{qV_{\mathrm{BE}}}{kT} \right) - 1 \right]. \tag{3.200}$$

The collector-base diode is reverse biased and has a tiny (negative) current

$$I_{\mathrm{R}} = I_{\mathrm{R0}} \left[ \exp\left( -\frac{qV_{\mathrm{CB}}}{kT} \right) - 1 \right], \tag{3.201}$$

in which $V_{\mathrm{BE}} = V_{\mathrm{B}} - V_{\mathrm{E}}$ and $V_{\mathrm{CB}} = V_{\mathrm{C}} - V_{\mathrm{B}}$. As discussed in the physical model above, the transistor has current gain and this can be modeled electronically by the current sources. The total currents are thus

$$\begin{aligned} I_{\mathrm{E}} &= I_{\mathrm{F}} - \alpha_{\mathrm{I}} I_{\mathrm{R}}, && (3.202) \\ I_{\mathrm{C}} &= \alpha_{\mathrm{N}} I_{\mathrm{F}} - I_{\mathrm{R}}, && (3.203) \\ I_{\mathrm{B}} &= I_{\mathrm{E}} - I_{\mathrm{C}}. && (3.204) \end{aligned}$$

Functionally, the collector and emitter can be exchanged, and this implies that necessarily the condition $\alpha_{\mathrm{I}} I_{\mathrm{R0}} = \alpha_{\mathrm{N}} I_{\mathrm{F0}}$ should hold. Thus, we have the final

Ebers-Moll equations for current:

$$\begin{aligned} I_{\mathrm{E}} &= I_{\mathrm{F0}}\left[\exp\left(\frac{V_{\mathrm{BE}}}{V_{\mathrm{T}}}\right)-1\right]+\alpha_{\mathrm{N}}I_{\mathrm{F0}}\left[1-\exp\left(-\frac{V_{\mathrm{CB}}}{V_{\mathrm{T}}}\right)\right], && (3.205)\\ I_{\mathrm{C}} &= \alpha_{\mathrm{N}}I_{\mathrm{F0}}\left[\exp\left(\frac{V_{\mathrm{BE}}}{V_{\mathrm{T}}}\right)-1\right]+I_{\mathrm{R0}}\left[1-\exp\left(-\frac{V_{\mathrm{CB}}}{V_{\mathrm{T}}}\right)\right], && (3.206)\end{aligned}$$

with $V_{\mathrm{T}} = kT/q$. The first term in each equation is dominant and in ideal transistors the second term is negligible and the transistor works as an ideal current amplifier, with the collector current being determined solely by the base-emitter voltage. In other words, the transistor is nearly a perfect programmable ideal current source; whatever we connect to the collector does not influence the current. In practice, if we connect the high resistive load to the collector, a large voltage drop will be induced in that resistance and the collector voltage will drop, and this in turns reduces the collector current as is evident from the above equations.

Figure 3.46 shows the empirical curves of an npn junction transistor. The transfer curves, $I_{\mathrm{C}}$ vs. $V_{\mathrm{BE}}$, ideally are independent of the load ($V_{\mathrm{CE}}$), but in practice the diode currents depend on it. The same applies to the output curves. Ideally these are 'horizontal', with the current being independent on $V_{\mathrm{CE}}$. In practice they have a slope. For all biases $V_{\mathrm{BE}}$ these slopes point to a voltage on the horizontal axis, called the Early voltage $V_{\mathrm{A}}$, which is typically of the order of 100 volt. It defines the output resistance of the transistor, as discussed in the chapter on electronics,

$$r_{\mathrm{o}} \equiv \frac{1}{\mathrm{d}I_{\mathrm{C}}/\mathrm{d}V_{\mathrm{C}}} = \frac{V_{\mathrm{CE}}+V_{\mathrm{A}}}{I_{\mathrm{C}}} \approx \frac{V_{\mathrm{A}}}{I_{\mathrm{C}}}. \tag{3.207}$$

For a typical current in the order of milliamperes this resistance is about 100 kΩ.

Finally, note that the collector-base and base-emitter junctions are pn junction diodes and they thus have capacitance related to their depletion zone. They thus depend on the bias (a forward-biased diode has much larger capacitance than a reverse-biased diode), and also on the design. Typical values are some pF for the $C_{\mathrm{CB}}$ and some tens of pF for $C_{\mathrm{BE}}$. In electronic literature these are often called $C_{\mu}$ and $C_{\pi}$ respectively.

### 3.6.4 Field-effect transistors

In a metal-oxide-semiconductor field-effect transistor (MOS-FET), the resistance of a semiconductor to which source and drain electrodes are connected is modulated by the voltage of the metal gate that is electrically isolated from the semiconductor, see Figure 3.48.

Under normal operation, an inversion channel is created to make the device conduct. For instance, we start with p-type material for the semiconductor and connect a positive gate voltage. This voltage creates a field that drives out the holes from the semiconductor at the other side of the insulator. For strong

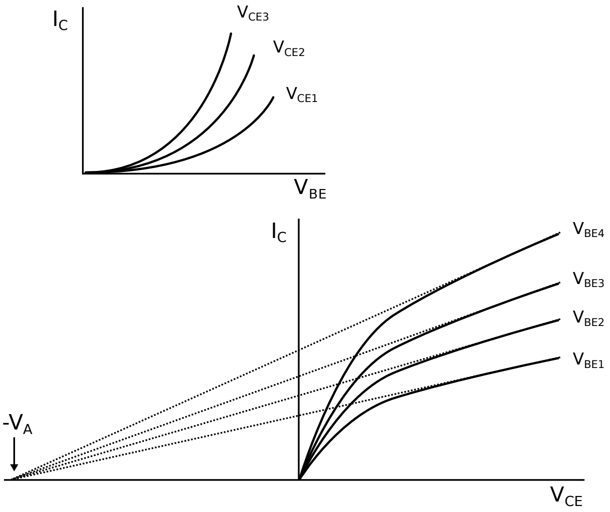


**Fig. 3.46**: Empirical transfer curves $I_{\mathrm{C}}$ vs. $V_{\mathrm{BE}}$ and output curves $I_{\mathrm{C}}$ vs. $V_{\mathrm{CE}}$ of an npn transistor. Ideally, the former are constant (independent of the parameter $V_{\mathrm{CE}}$) and the latter are horizontal (with the current independent on the load connected at the collector), but in practice the output curves point all to a voltage at the horizontal axis. This voltage is called the Early voltage $V_{\mathrm{A}}$ and defines the output resistance of the transistor

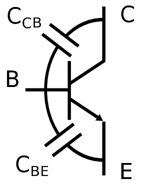


**Fig. 3.47**: Because the junctions of a transistor are np-diodes, they have internal capacitance. The forward biased junction BE has larger capacitance than the reverse biased CB junction

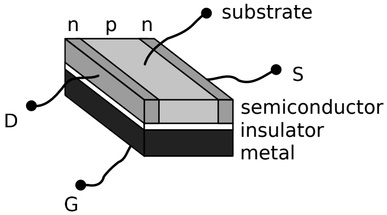


**Fig. 3.48**: Metal-oxide-semiconductor field-effect-transistor (MOS-FET). The gate (G), electrically isolated from it by an insulator, modulates the amount of charge in a semiconductor (here p-type). The conductance from drain (D) to source (S) (both n-type semiconductors here) is hereby changed. The substrate connection can also be used to change the conductive properties

voltages, even electrons are attracted to the semiconductor-insulator interface. When that happens, a so-called 'channel' develops, a high density sheet of electrons right at the interface. The conductance of the device suddenly shoots up and from that moment on increases linearly with the gate voltage.

In first order, for small voltages applied to the drain terminal (compared to the grounded source terminal), the conductance of the device is programmed by the gate voltage. Similar to the a pn-junction, the gate voltage induces a voltage drop and an electrical field. Part of it across the insulator and part of it in the semiconductor. It is not very difficult to calculate this if we make some basic assumptions. For instance for a n-channel inversion transistor:

- All acceptors are ionized everywhere in the semiconductor, $N_{\mathrm{A}}^{-} = N_{\mathrm{A}}$.
- Inside the depletion zone W, all holes have moved out and there is a space-charge density equal to $N_{\mathrm{A}}$. Outside it, their density is equal to the acceptor density and the material is neutral.

The example for a situation at the onset of creating an electron channel is shown in Figure 3.49. It has created a band bending (voltage drop) equal to $V_{\mathrm{s}}$ in the semiconductor just enough to force the Fermi level to be equal to the conduction band at the interface.

The curvature of the band bending can be calculated using Poisson's equation (Eq. 3.24), the voltage being the double integral of the space charge,

$$V(x) = \frac{qN_{\mathrm{A}}}{2\varepsilon_{\mathrm{s}}}(x - W)^2, \tag{3.208}$$

with the total voltage drop in the semiconductor thus given by

$$V_{\mathrm{s}} = V(x = 0) = \frac{qN_{\mathrm{A}}}{2\varepsilon_{\mathrm{s}}}W^2. \tag{3.209}$$

Apart from the voltage drop in the semiconductor, there is also a voltage drop $V_{\mathrm{ox}}$ across the insulator. The sum of the two is equal to the externally applied bias,

$$V_{\mathrm{G}} = V_{\mathrm{s}} + V_{\mathrm{ox}}. \tag{3.210}$$

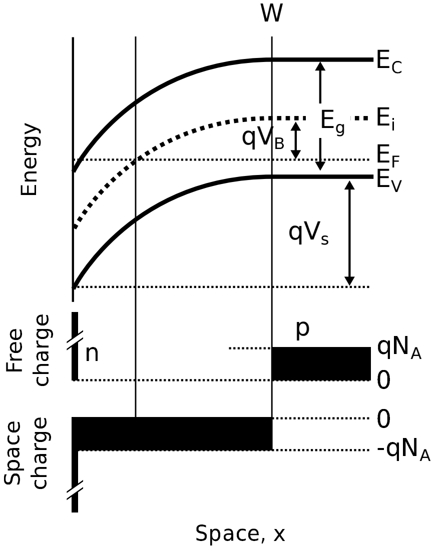


**Fig. 3.49**: Energy diagram, free charge and space charge distributions of an n-channel inversion MOS-FET slightly above the threshold voltage $V_{\rm G} = V_{\rm T}$. Parameters shown: $qV_{\rm s}$ band bending, $E_{\rm g}$ energy bandgap, $E_{\rm C}$ conduction band energy, $E_{\rm V}$ valence band energy, $E_{\rm F}$ Fermi level, $E_{\rm i}$ intrinsic level (if $E_{\rm F} = E_{\rm i}$, per definition $p = n$), $V_{\rm B}$ distance between $E_{\rm F}$ and $E_{\rm i}$ in bulk, $N_{\rm A}$ acceptor density

To find the voltage drop across the insulator, we can make use of Maxwell's continuity equations, more specifically the continuity of displacement $D = \epsilon E$. On the side of the semiconductor the displacement is given by

$$\begin{aligned} D_{\rm s}(0) &= -\varepsilon_{\rm s} \left.\frac{{\rm d}V(x)}{{\rm d}x}\right|_{\rm x=0} \\ &= qN_{\rm A}W. \end{aligned} \tag{3.211}$$

Since, by definition, there cannot be charge in the insulator, the field $E_{\rm ox} = D_{\rm ox}/\varepsilon_{\rm ox}$ is constant and the voltage drop in the oxide is thus given by this field multiplied by the oxide thickness $d_{\rm ox}$,

$$\begin{aligned} V_{\rm ox} &= D_{\rm ox}d_{\rm ox}/\varepsilon_{\rm ox} \\ &= qN_{\rm A}Wd_{\rm ox}/\varepsilon_{\rm ox} \\ &= \frac{\sqrt{2q\varepsilon_{\rm s}N_{\rm A}V_{\rm s}}}{C_{\rm ox}}, \end{aligned} \tag{3.212}$$

where the definition of capacitance (density), $C_{\rm ox} = \varepsilon_{\rm ox}/d_{\rm ox}$ (unit: F/m$^2$), was used. The gate voltage is then given by

$$V_{\rm G} = V_{\rm s} + \frac{\sqrt{2q\varepsilon_{\rm s}N_{\rm A}V_{\rm s}}}{C_{\rm ox}}. \tag{3.213}$$

**Table 3.IX**: Parameters of a typical n-type inversion-channel Si-SiO$_2$ MOS-FET at onset of inversion, $V_\mathrm{g} = V_\mathrm{T}$.

| *Parameter* | *Value* |
|---|---|
| $N_\mathrm{A}$ | $10^{15}$ cm$^{-3}$ |
| $d_\mathrm{ox}$ | 100 nm |
| $E_\mathrm{g}$ | 1.15 eV |
| $T$ | 300 K |
| $E_\mathrm{F}$ | $E_\mathrm{V}$ + 225 meV |
| $C_\mathrm{ox}$ | 345 μF m$^{-2}$ |
| $V_\mathrm{s}$ | 925 mV |
| $W$ | 630 nm |
| $V_\mathrm{ox}$ | 512 mV |
| $V_\mathrm{T}$ | 1.44 V |

In a typical MOS-FET, half of the voltage drop is across the interface and half in the semiconductor, see Table 3.IX.

The threshold voltage $V_\mathrm{T}$ , the onset of creating an inversion channel can be calculated by looking at the band bending figure (Fig. 3.49) and realizing that the Fermi level touches the conduction band at the interface when the band bending is equal to $qV_\mathrm{s} = (E_\mathrm{C} - E_\mathrm{F})$. Moreover, the Fermi level was found as a function of density of states $N_\mathrm{V}$ and $NA$, and we find a threshold voltage as a function of material and device parameters as

$$V_\mathrm{T} = \frac{1}{q}\left[E_\mathrm{g} - kT\ln\left(\frac{N_\mathrm{V}}{N_\mathrm{A}}\right)\right] + \frac{1}{C_\mathrm{ox}}\sqrt{2\varepsilon_\mathrm{s}N_\mathrm{A}\left[E_\mathrm{g} - kT\ln\left(\frac{N_\mathrm{V}}{N_\mathrm{A}}\right)\right]}. \qquad (3.214)$$

We make here now another approximation in assuming that the current is negligible for a gate bias bellow this threshold voltage, after which the current rises linearly. The latter makes sense. From the moment the threshold voltage has been overtaken, no further band bendings are possible inside the semiconductor, and all charge is created directly adjacent the insulator. In other words, from that moment on the device works as a capacitor with a linear charge-voltage relation (see the section on capacitors). The current is then proportional to the charge density, the electrical field created by the source-drain voltage, $E_\mathrm{DS} = V_\mathrm{DS}/L$, the electron mobility, and scaled with the device dimension (namely the electrode width $W$),

$$I_\mathrm{DS} = q\mu_\mathrm{n}V_\mathrm{DS}\frac{W}{L}\frac{V_\mathrm{G} - V_\mathrm{T}}{C_\mathrm{ox}}, \qquad (3.215)$$

which is the classical current equation for a MOS-FET in the linear regime. When we increase the drain voltage, the voltage drop across the interface diminishes at this place and it might be so that no longer the channel exists there. In this case the device enters the saturation regime and the current is

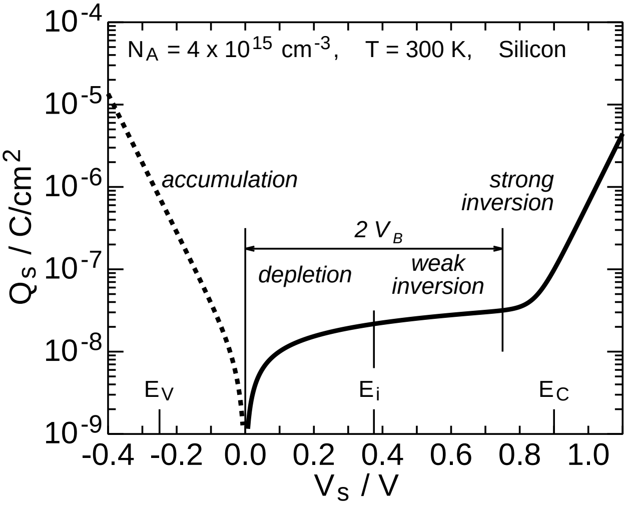


**Fig. 3.50**: Space charge as a function of total band bending in an inversion channel MOS-FET. The dashed line shows the charge in accumulation for comparison

quadratically depending on the gate bias. This is better explained in a thin-film transistor (TFT), which actually works much simpler.

Finally, it is interesting to see how much charge is in the device as a function of band bending. This is shown in Figure 3.50

### 3.6.5 Thin-film transistors (TFTs)

A TFT differs from a MOS-FET in that the active layer accommodating the channel can be as thin as one mono-layer. In other words, there is no room for band bendings. That is why TFTs work in accumulation instead of inversion, with carriers in the channel of the same type as the carriers in the bulk material. In accumulation no band bendings are needed and the device works right from the start as a parallel plates capacitor. Any charge induced by the gate is always directly at the interface and is always free charge. This makes sense, since no levels are available to store induced charge of any other type. Imagine p-type material doped with acceptors. To bring it into accumulation, a negative voltage is needed at the gate. Negative voltages attract positive charge on the other side of the insulator. This charge can only be free holes, since the acceptor can only be neutral or negatively charged. In other words, the dopants play no role in TFTs, or in general in devices working in accumulation. This simplifies the calculations. Let's remove them altogether from our model. A TFT is just a capacitor with on one side source and drain electrodes connected to the semiconductor material and on the other side an electrode, the gate, see Figure 3.51.

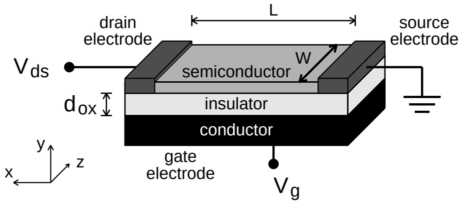


**Fig. 3.51**: Thin film transistor (TFT)

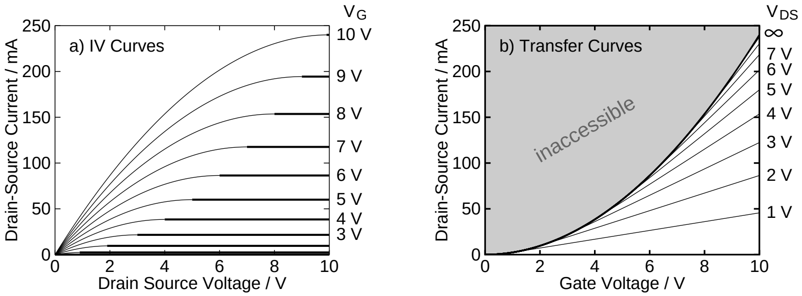


**Fig. 3.52**: Simulations of n-channel TFT and MOS-FET IV and transfer curves in the linear regime (thin lines) and the saturation regime (thick lines)

To describe the TFT, we make two assumptions: 1) The charge density at any point is proportional to the local voltage drop across the insulator, and 2) The local current is proportional to the local charge density. In equations:

$$p(x) = \rho(x)/q = C_{\text{ox}}\left[V(x) - V_{\text{G}}\right]/q, \tag{3.216}$$

$$I_{\text{x}}(x) = -qp(x)\mu_{\text{p}}W\frac{\mathrm{d}V(x)}{\mathrm{d}x}. \tag{3.217}$$

This differential equation can easily be solved. Using the boundary conditions that the current is constant everywhere ($I(x) = I_{\text{DS}}$ for all $x$), $V(x=0)=0$, and $V(x=L) = V_{\text{DS}}$, the solution is

$$I_{\text{DS}}^{\text{lin}}(V_{\text{G}}, V_{\text{DS}}) = \mu_{\text{p}}C_{\text{ox}}\frac{W}{L}\left(V_{\text{G}}V_{\text{DS}} - \frac{1}{2}V_{\text{DS}}^2\right). \tag{3.218}$$

This is equal to the behavior of a MOS-FET. When at the drain the potential is equal to the gate bias, no voltage drop exists there and no free charge. The channel closes, something that is called 'pinch-off'; the density of charges goes to zero. Further increase of the drain bias will not increase the current, any additional voltage beyond $V_{\text{DS}} = V_{\text{G}}$ is absorbed by an infinitesimal thin zone close to the drain ($p \to 0$ and $E \to \infty$, with the product of the two constant, implying constant $Ids$). The current in saturation is thus found as the current

**Table 3.X**: Parameters of TFTs used here

| *Parameter* | *Value* | *Unit* |
|---|---|---|
| $d_{\mathrm{ox}}$ | 216 | nm |
| $\varepsilon_{\mathrm{ox}}$ | 3.9 | $\varepsilon_0$ |
| $C_{\mathrm{ox}}$ | 160 | $\mu$F/m$^2$ * |
| $\varepsilon_{\mathrm{s}}$ | 11.9 | $\varepsilon_0$ |
| $L$ | 10 | $\mu$m |
| $W$ | 1 | cm |
| $\mu$ | 300 | cm$^2$/Vs |

*: In this chapter all capacitances are capacitance densities.

above with $V_{\mathrm{DS}}$ substituted by $V_{\mathrm{G}}$,

$$I_{\mathrm{DS}}^{\mathrm{sat}}(V_{\mathrm{G}}) = \mu_{\mathrm{p}} C_{\mathrm{ox}} \frac{W}{L} \frac{1}{2} V_{\mathrm{G}}^2. \tag{3.219}$$

The output curves ($I_{\mathrm{DS}}$ vs. $V_{\mathrm{DS}}$) and transfer curves ($I_{\mathrm{DS}}$ vs. $V_{\mathrm{G}}$) for TFTs and MOS-FETs alike are given in Figure 3.52, while the charge and voltage along the channel are given in Figure 3.53.

## 3.7 Exercises

1. What is the resistance of a cylindrical resistance of length $L$ and radius $r$?

2. What is the capacitance of a wire of radius $a_0$ in a tube of length $L$ and radius $a_1$? (see Figure 3.54).

3. The resistivity of air is about $10^{16}$ $\Omega$m. In a thunderstorm charge is building up like in a Van Der Graaff generator. It is stored for instance in the capacitor formed by cloud and ground, see Figure 3.55. When the associated voltage is large enough, a tiny current starts. This ionizes the air and the resistance drops, thus causing more current. A avalanche breakdown of the resistivity of air occurs that turns it into a plasma that has relatively low resistance. This hot plasma emits light in a short time, a so-called spark or flash. The breakdown field (also called dielectric strength) of air is about 1 MV/m.
a) What is the capacitance, charge, voltage and stored energy in a typical thunderstorm before a flash.
b) If the resistance suddenly drops to zero, what is the maximum current (if we assume a 1 cm radius bolt)?

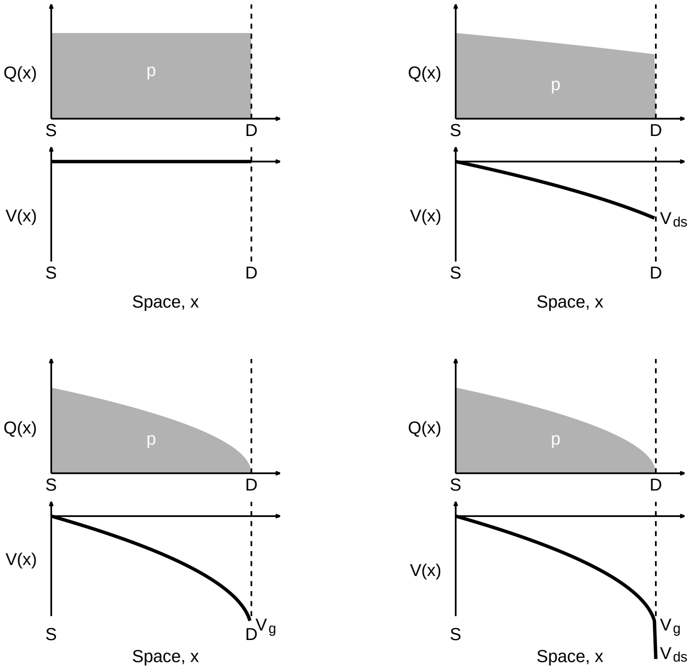


**Fig. 3.53**: Distribution of charge $Q(x) = qp(x)$ and potential $V(x)$ along the channel for various biases $V_{\mathrm{DS}}$ for a p-channel TFT (negative gate bias).

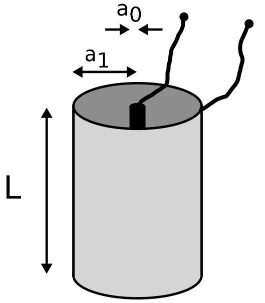


**Fig. 3.54**: Wire-in-tube capacitor. Exercise 2

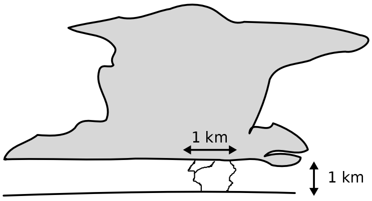


**Fig. 3.55**: Cross section of a thunderstorm. Exercise 3

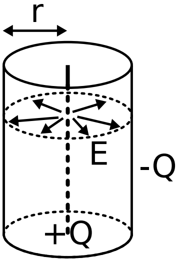


**Fig. 3.56**: Wire-in-tube capacitor. Exercise 2

## 3.8 Answers

2 We place a charge $+Q$ on the surface of the wire and $-Q$ on the wall of the tube. The electric field is radially pointing outward from the wire. We surface-integrate the electric field and use Gauss' law (Eq. 3.9) of a closed surface perpendicular to the electric field, namely a cylinder of radius $r$, as shown in Figure 3.56. The area multiplied by the electric field is then equal to charge contained withing the volume divided by $\varepsilon$. In this case, since we can ignore the end caps,

$$E \cdot 2\pi r L = Q/\varepsilon, \tag{3.220}$$

for $a_0 < r < a_1$ and 0 elsewhere. The potential is the integral of the electric field. Setting it (arbitrarily) to zero at $r = a_1$ we find

$$V(r) = -\int_{\mathrm{a}_1}^{r} \frac{Q}{2\pi\varepsilon r' L}\mathrm{d}r' = \frac{Q}{2\pi\varepsilon L}\ln\left(\frac{a_1}{r}\right). \tag{3.221}$$

Substituting $r = a_0$ and using the definition of capacitance as ratio of charge and voltage we get

$$C = \frac{2\pi\varepsilon L}{\ln(a_1/a_0)}. \tag{3.222}$$

Typical values of coax cables are about 100 pF/m.

3 a) A cloud is about 1 km from the ground and has an area of 1 km by 1 km. The relative permittivity of air is close to unity, so we can use the dielectric constant of vacuum $\varepsilon = \varepsilon_0$ and we get a capacitance, Eq. (3.57), equal to $C = \varepsilon A/h$ = (8.85 $10^{-12}$ F/m) × ($10^6$ m$^2$)/($10^3$ m) = 8.85 nF. The voltage at breakdown is ($10^3$ m) × ($10^6$ V/m) = 1 GV. The charge is thus $Q = CV$ = (8.85 $10^{-9}$ F) × ($10^9$ V) = 8.9 C. And the energy, Eq. (3.60), is $U = CV^2/2$ = (8.9 $10^{-9}$ F) × ($10^{18}$ V$^2$)/2 = 4.4 GJ (enough to light a 100-watt light bulb for three months!)
b) The inductance of a 1 km bolt with radius of 1 cm is (Table 3.VI) $L = (2Z \times 10^{-7}\ \mathrm{N/A^2}) \times [\ln(2Z/a) - 3/4]$ = 2.3 mH. If all the electric field

energy of the charged capacitor is converted into magnetic field energy of the current-bearing inductor, Eq. (3.91), then we get $I = \sqrt{2U/L} = 2$ MA.

# 4 | Sensors & Actuators

## 4.1 Introduction

In this chapter the sensors and actuator are studied. The amount of sensors and actuators in the world is enormous. It is not possible to describe them all. That is why here only some examples are given, to give you an idea how you should think when dealing with sensors. It is always important to find out about a sensor *how* they work from a scientific point of view. How is it achieved that the physical parameter is translated into an electronic parameter? This is important, because with this knowledge, the limitations of the device will be understood. Or the interference, or the accuracy, shelf life, etc.

The next step is to find out how to *make* them work, from an engineering point of view. What is the sensitivity? What is the accuracy? Etc. An engineer does not care (much) about where it comes from. An engineer just wants to know the basic facts. That is why an engineer just needs a datasheet, and that's it! From an engineering point of view the datasheet is the only thing that counts. From that moment on, the sensors and actuators can be considered black boxes. ”There can be Martians in there pulling strings, for all I care. I just want to know how to use the thing!”

In this chapter the workings of the sensors and actuators will be discussed, together with how to use them. Things as determining the sensitivity, etc.

## 4.2 Optical sensors & actuators

### 4.2.1 Light-emitting diode (LED)

A light emitting diode is the simplest actuator imaginable where the information in the form of an electrical signal is translated into an optical signal, for instance a warning light when the petrol tank is (nearly) empty.

An LED is a normal diode from which light is emitted as a side effect. Electrons are injected into the barrier from the n-type side and holes from the p-type side. What makes an LED different from a normal diode is that in an LED these electrons and holes recombine ’radiatively’, i.e., where the energy of

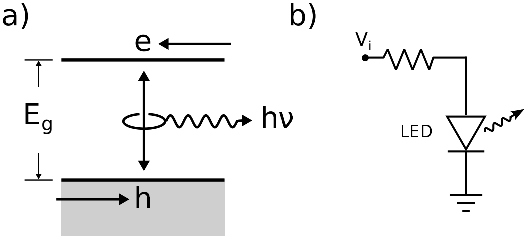


**Fig. 4.1**: a) Schematic drawing of an LED. Electrons and holes are injected into the conduction and valence band respectively. They recombine and the energy is converted into photons. The wavelength of light depends on the bandgap $E_\mathrm{g}$ of the material. b) An LED is a current-to-illumination actuator. A resistor should be used in series to define the current of a voltage source

**Table 4.I**: Typical semiconductor materials and their possible use for LEDs

| Material | Bandgap | | Wavelength | Color |
|---|---|---|---|---|
| InN | 0.65 eV | | | Infrared |
| Si | 1.15 | (indirect) | - | - |
| GaAs | 1.42 eV | | 800 nm | Infrared |
| AlAs | 2.12 eV | (indirect) | - | - |
| AlGaAs | (variable) | | (700 nm) | Red |
| GaP | 2.26 eV | | 600 nm | Yellow |
| AlP | 2.5 | (indirect) | - | - |
| AlGaP | (variable) | | (565 nm) | Green |
| ZnSe | 2.82 eV | | 470 nm | Blue |
| GaN | 3.4 eV | | 500 nm | Blue |
| InGaN | (variable) | | (405 nm) | Violet |
| C | 5.47 eV | | | Ultraviolet |
| AlN | 6.2 eV | | 210 nm | Ultraviolet |

recombination is converted into photons. Before recombination, the electrons relax to the bottom of the conduction band and the holes float upwards to the top of the valence band. The energy lost by recombination – the energy of the photon – is thus equal to the bandgap of the material, see Figure 4.1.

The various colors of LEDs stem from the different bandgaps of the materials used, see Table 4.I. By mixing materials, for instance gallium-arsenide and gallium-phosphide, $\mathrm{GaAs_xP_{1-x}}$, the color can be tuned. The table gives typical values. A white LED can be made by a UV or blue LED in combination with phosphorescent powder to down-convert it to visible light in a broad emission spectrum.

Note that not every material can be used for LEDs. Many materials, including silicon, do not have a direct bandgap, meaning that the bottom of the conduction band is not aligned with the top of the valence band and elec-

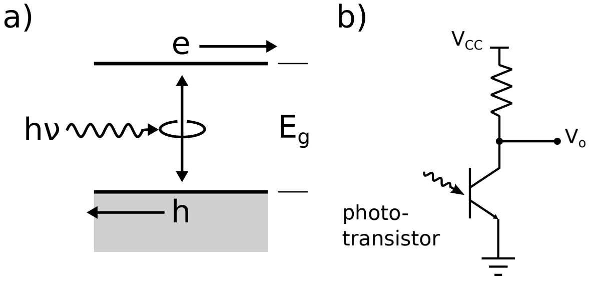


**Fig. 4.2**: a) Schematic drawing of a photo-transistor. Photons are absorbed and the energy converted into electron-hole pairs at the base of a junction transistor if the photon energy is large enough $h\nu > E_\mathrm{g}$. b) A photo-transistor in series with a resistor to convert the light into a voltage signal

trons and holes cannot recombine radiatively. In these cases, electrons fall back into the holes in the valence band via jumping to intermediate impurity states. Moreover, it can be stated that the making of short-wavelength LEDs is technologically more difficult compared to the long wavelength ones. This finds its origin in the stability of the material and the high density of impurities in such materials, impurities that will kill the radiation efficiency by supplying non-radiative recombination paths.

The first critical observation is that an LED translates *current* into light, and not voltage. If we directly apply a voltage source to an LED, it will burn. While it needs a minimal voltage to turn on, as determined by (but not only by) the bandgap of the material, the current grows exponentially with voltage – about a factor 10 for every 60 mV (more precisely, a factor $e$ for every $kT/q$ = 26 mV). There is a small range in which an LED will operate. Between switching on and burning is a narrow margin. That is why we have to treat the LED as a current-to-illumination actuator. To achieve this, we either use a current source, or place a resistance in series with the voltage source to convert it effectively in a sort of current source.

### 4.2.2 Photo-transistor

The opposite of an LED is a photo-diode, or photo-transistor. It converts light into current. Or, better to say, it converts light into conductance. The resistance of a photo-transistor is reduced by illumination. It works by converting the photon energy in electron-hole pairs at the base of a transistor. This is equivalent to injecting a current in the base of a normal junction transistor; the collector-emitter is strongly increased by such tiny injections of free carriers. To translate the current into a voltage signal, a resistor can be placed in series with the photo-transistor. If the resistance is large, the transistor works as a switch, for enough light, the output is effectively connected to ground. With little light, the output is 'pulled-up' to $V_\mathrm{CC}$, see Figure 4.2.

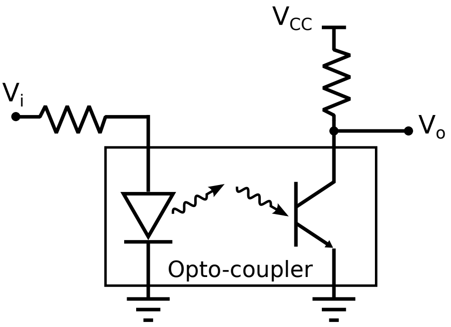


**Fig. 4.3**: Opto-coupler

### 4.2.3 Opto-coupler

An opto-coupler combines an LED with a photo-transistor, see Figure 4.3. With the diode, the voltage signal is translated into light and the photo-transistor converts it back into voltage. There are basically two advantages to using opto-couplers.

First, note that the input signal and output signal are electrically isolated. This means that, for instance, dangerous supply voltages can be kept away from the low-voltage signal part. Moreover, if a spark on the line occurs, it will not burn the circuit. Another reason for using these techniques is to avoid ground loops, with current being transported over the ground line. For these applications, integrated sealed 4-pin packages exists, with an infrared LED and a photo-transistor hidden inside.

Note that it is not important to pass *power* through an opto-coupler. It is not as if we are going to supply our computer with power to run it independently from other power sources. An opto-coupler is only to pass *information* from one side to the other.

A second important application of opto-couplers, and one probably even more often used, is the light-interruption detection. An example is the mouse of your computer, detecting the middle-wheel movement (and on older ball mouses, also the X and Y movements). The LED and the photo-transistor are physically separated to allow for objects to pass inbetween them and block the light. The distance between the emitter and detector can be even large, for instance in the entrance of a shop, to detect clients coming in. Or the man-leaving-urinal detectors found in many public toilets.

### 4.2.4 Light-dependent resistor (LDR)

Another way of detecting light is with a light-dependent resistor (LDR). This consists of a piece of semiconductor, normally cadmium sulfide, without any junctions, just a piece of material with ohmic (non rectifying) contacts, see Figure 4.4. They can be recognized by their 'wriggly' pattern, which is actually the active material. The rest is contacting electrodes.

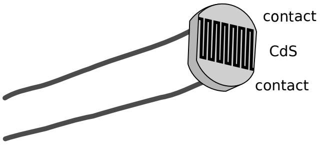


**Fig. 4.4**: Light-dependent resistor (LDR) consisting of a semiconductor (the dark pattern), usually cadmium sulfide, embedded between two ohmic-contacts

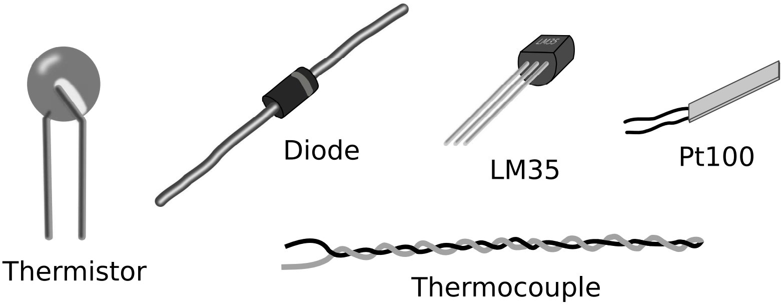


**Fig. 4.5**: Some temperature sensors

Like a photo-diode and photo-transistor, in an LDR the photons are converted into electron-hole pairs and this increases the conductivity (lowers the resistivity) by creating a high density of free carriers, both holes and electrons. The advantage of LDRs is their relatively low cost.

## 4.3 Temperature sensors

One of the most common applications of electronic instrumentation is the measurement of temperature. Temperature sensors come in all sizes, qualities, prices. The ones described here are summarized in Figure 4.5.

### 4.3.1 Thermocouple

A thermocouple is a transducer that translates temperature *difference* directly into voltage. It consists of a wire of two types of metal (alloy). These alloys have different electron affinity meaning that they to different extent attract electrons. If they are joined, electrons will thus have a tendency to go from one alloy to the other, which is equivalent to an electric field. As a result, a voltage drop is induced at the junction. Because in a closed loop a junction A/B is always accompanied by a junction B/A, the total voltage drop is always zero.

However, since the electron-affinity and thus electric field at the junction depend on the temperature, the voltage drop at a junction depends on the temperature, the so-called Seebeck effect. If the junctions are at a different temperature, the voltage drops do not sum up to zero anymore and finite voltage

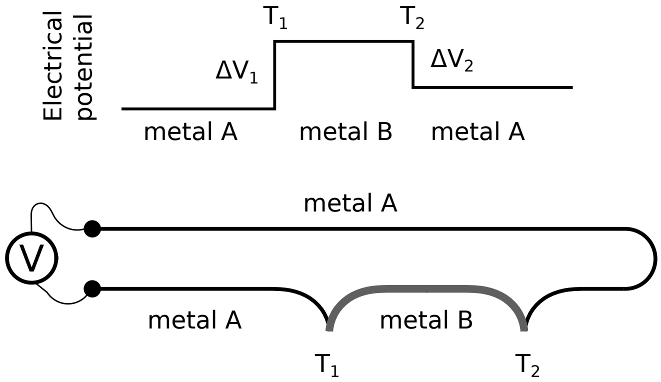


**Fig. 4.6**: Schematic drawing of a thermocouple temperature difference sensor consisting of wires of different alloys. At the junctions a voltage drop is induced. If the temperatures $T_1$ and $T_2$ are not equal, a resulting voltage can be measured at the end of the wires pair

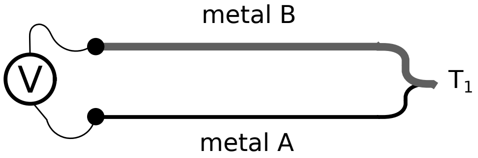


**Fig. 4.7**: Cold junction compensation technique

remains. Since the Seebeck effect is quite linear with temperature, a voltage is induced at the end of a A/B/A thermocouple that is linear with the temperature difference of the two junctions A/B and B/A, see Figure 4.6. Table 4.II gives a list of the most used thermocouples.

Note that the voltmeter also introduces metal-metal junctions that include voltage drops (if the wires in the voltmeter are not made of metal A). These two junctions – A/C and C/A – however, are at the same temperature and they cancel each other.

The operational procedure is to maintain one of the two junctions, A/B or B/A, at a calibrated temperature, for instance in a bucket of melting ice (per definition at the melting point of water, 273.15 K at 1 atm), the so-called cold junction. The voltmeter then measures a voltage that is linearly proportional to the *difference* temperature. This procedure is quite cumbersome in practice. Modern systems, for instance advanced multimeters, use a technique called cold-junction compensation, in which only one thermocouple is used to measure the temperature. The meter now also has asymmetric junctions, the temperature of which are measured using another temperature sensor, for instance a diode or a transistor. The system then compensates for this junction temperature. See for instance the book of Horowitz and Hill, *The Art of Electronics*.

The advantage of thermocouples is that they are relatively cheap and ro-

**Table 4.II**: Most popular thermocouples (Horowitz and Hill). In round brackets: the color code of wires in Europe (IEC). In square brackets the color code of wires in the USA (ANSI)

| Type | Alloy A<br>+ | Alloy B<br>− | Sensitivity<br>(μV/°C) | Max. Temp.<br>(°C) |
|---|---|---|---|---|
| J | Iron<br>(black)<br>[white] | Constantan<br>(white)<br>[red] | 51.45 | 760 |
| K | Chromel<br>(green)<br>[yellow] | Alumel<br>(white)<br>[red] | 40.28 | 1370 |
| T | Copper<br>(brown)<br>[blue] | Constantan<br>(white)<br>[red] | 40.28 | 400 |
| E | Chromel<br>(purple)<br>[purple] | Constantan<br>(white)<br>[red] | 60.48 | 1000 |
| S | Platinum<br>(orange)<br>[black] | Alloy 11<br>(white)<br>[red] | 5.88 | 1750 |

bust. Their main disadvantage is that they are interfering with the system. Don't forget that metals are not only good electrical conductors, but also good thermal conductors (the reason is that heat and electrical conduction both use the electronic energy properties of materials); it is difficult to find materials that are good electrical conductors and bad heat conductors, and vice versa. If we place a metal wire touching a hot (or cold) object, with the other end at our meter at room temperature, the wire will transport huge amounts of heat and the measured object can substantially change temperature because of the measurement. This is interference and can be quite damaging.

### 4.3.2 Diode

The diode as described in the chapter on physics has a current that not only depends on applied bias, but also on temperature. In general,

$$I = I_{\mathrm{S}}(T)\left[\exp\left(\frac{qV}{kT}\right) - 1\right], \tag{4.1}$$

with also the reverse-bias saturation current possibly depending on temperature, apart from the obvious dependencies of material choice and interface area.

Ideally, a sensor is linear, meaning that it has a constant sensitivity, i.e., the derivative of the transfer function, and we can achieve this for a diode sensor by using a current source to drive the diode. An example is given in Figure 4.8.

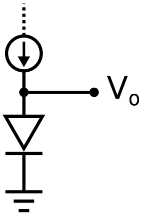


**Fig. 4.8**: Diode sensor, a diode used in combination with a current source resulting in a linear dependence of voltage $V_o$ on temperature

If the reverse-bias saturation current $I_S$ is a constant, the temperature is given by the voltage as

$$V(T) = \frac{k \ln(I/I_S)}{q} T. \tag{4.2}$$

The sensitivity of the sensor is given by

$$S \equiv \frac{dV}{dT} = \frac{k}{q} \ln\left(\frac{I}{I_{S0}}\right). \tag{4.3}$$

A typical reverse bias current is $10^{-14}$ A and if we drive the diode at 1 μA, we find a sensitivity of 1.6 mV/K. In practice often negative values are found. That is because the reverse-bias saturation current is not constant but depends on the mobility and diffusion coefficient (see Chapter 3) that normally also depend on the temperature. Also for a Schottky diode, the reverse-bias saturation current is temperature dependent. We find a voltage

$$V(T) = \frac{kT}{q} \ln\left(\frac{I}{T^2 A^{**}}\right) + \phi_{Bp}, \tag{4.4}$$

with $\phi_{Bp}$ the barrier height of the Schottky diode. In this case the sensor is no longer linear. The sensitivity can even have a negative value. Typical values for diodes are $-2$ mV/K.

A good approach is to use two diodes with two different currents, the voltage difference of the two very nicely follows

$$\Delta V = \frac{kT}{q} \ln\left(\frac{I_1}{I_2}\right). \tag{4.5}$$

For instance, if we use two currents, an order of magnitude apart, we get a sensitivity of $S = 0.2$ mV/K.

To drive the diodes we should be careful not to use too much current. If we use a value of 1 mA, combined with the fact that a conducting diode has a voltage drop of about 0.7 volt, the diode consumes 0.7 mW. Apart from the fact that that uses the battery too rapidly in a battery-operated application, this energy is transformed into the form of heat and that might interfere with the measurements, especially at low temperatures, when the diode can be substantially hotter than the environment being measured. On the other hand, too low a current will reduce the signal-to-noise ratio.

While the lone diode is not a very high quality temperature sensor, it has one major advantage and that is that it can easily be incorporated into integrated circuits at nearly no cost. That makes it an ubiquitous element of modern electronics.

### 4.3.3 LM35

An LM35 temperature sensor has, at its heart, diodes and current sources as described above. It also uses intelligent circuits and calibrated components on a tiny IC. The signal is further amplified to result in a well-calibrated output voltage given by

$$V_{\mathrm{LM35}}(T) = (10\ \mathrm{mV}) \times \frac{T}{{}^{\circ}\mathrm{C}}. \tag{4.6}$$

It is one of easiest to operate temperature sensors. It just needs a power of something like 5 volt, although the exact value is not very important. It is robust (as long as we do not invert the polarity of the power supply). Its main disadvantage is its cost. While that is no problem for our one-piece electronic trials, it might be prohibitive for large volume electronic applications. Like always, we have to decide if we really need to have too high quality signals and if the increased costs are not too much for our budget.

### 4.3.4 PT100

A very good and accurate temperature sensor is a PT100 resistor, named after the fact that it is made of platinum deposited on a ceramic substrate and has a calibrated resistance of exactly 100.000 Ω at 0 °C. The high accuracy is achieved by laser ablation of the platinum in the factory until the desired resistance is reached. This complicated method also immediately explains why the sensors are expensive. Yet, because of their high accuracy and their linearity down to low temperatures, the sensor is widely used.

At temperatures other than zero celsius the resistance is quite linearly depending on temperature, and is up to third order given by

$$R(t) = (100\ \Omega) \times \left[1 + \alpha \left(\frac{T}{{}^{\circ}\mathrm{C}}\right) + \beta \left(\frac{T}{{}^{\circ}\mathrm{C}}\right)^2 + \gamma \left(\frac{T}{{}^{\circ}\mathrm{C}}\right)^3\right]. \tag{4.7}$$

Typical values are $\alpha = 3.9 \times 10^{-3}$, $\beta = -5.8 \times 10^{-7}$, $\gamma = -4.4 \times 10^{-12}$. Thus, in first order, the sensitivity of the sensor is

$$S = \frac{\mathrm{d}R}{\mathrm{d}T} = 0.39\ \Omega/{}^{\circ}\mathrm{C}. \tag{4.8}$$

Measuring such low resistance as the one of the PT100 sensor in the hundred ohms range is not very easy. Normally the cables of a multimeter have a resistance in the order of ohms and that introduces an unacceptable error of some

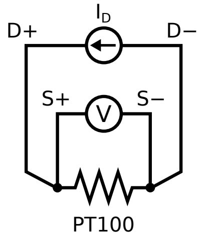


**Fig. 4.9**: 4-wire measurements. A drive current is supplied to the PT100 through two wires, D+ and D−. Two other wires, S+ and S−, measure the voltage drop across the sensor. These wires do not carry current and therefore do not include voltage drops. Also, because they are connected as close as possible to the sensor, they do not measure voltage drops of the drive cables

degrees. After all that trouble of laser-calibration in the factory, that is rather unsatisfactory. That is why normally 4-wire measurements technique is used to measure PT100 sensors. This works as follows: To the sensor two wires, called 'drive' (D+ and D−), are connected that supply a steady current. Any voltage drop in the wires is irrelevant. The voltage across the sample is then measured by two other wires, called 'sense' (S+ and S−), see Figure 4.9. These wires do not carry any current (the input resistance of a voltmeter is infinite) and because they are connected as close as possible to the sensor, do not measure any voltage drop in the cables of the driving current. The resistance can then be found as

$$R = \frac{V_{\mathrm{S+}} - V_{\mathrm{S-}}}{I_{\mathrm{D}}}. \tag{4.9}$$

Note that also for the PT100 sensor we have to be careful with the choice of the driving current. Too much and the sensor will heat up and interfere with the system measured. Too low and the signal-to-noise ratio will become too low.

### 4.3.5 Thermistor

A thermistor is characterized by its negative temperature coefficient, NTC, which means that the resistance drops when the temperature is increased. This distinguishes them from metal resistors such as the PT100 that have a positive temperature coefficient. Thermistors are normally made of semiconductors and as we have seen in the chapter on physics, they indeed become more conductive with higher temperatures, because electrons are thermally promoted from the valence band to the conduction band, thus increasing the number of electrons and holes available for conduction.

Typically, a thermistor has a temperature-resistance relation of the form

$$R(T) = R_0 \exp\left[\beta\left(\frac{1}{T} - \frac{1}{T_0}\right)\right], \tag{4.10}$$

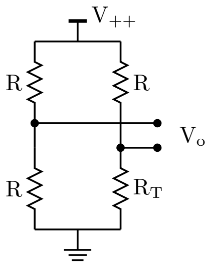


**Fig. 4.10**: A thermistor $R_{\mathrm{T}}$ used in a Wheatstone bridge to convert the information form the resistance domain to the voltage domain and to remove the offset

and this gives a sensitivity $S$, also sometimes called $\alpha$ for thermistors,

$$S \equiv \frac{\mathrm{d}R}{\mathrm{d}T} = -\frac{\beta R}{T^2}, \tag{4.11}$$

with typical values $R_0 = 1$ kΩ, $S = -45$ mΩ/K and a useful temperature range between −60 °C and +150 °C. (below this range the resistance becomes too large and above it too small).

NTC thermistors are good for measuring the temperature, but not good for controlling it. That is because positive feedback exists; if the temperature increases, the resistance decreases and the resistor draws more current and consumes more energy, thus heating up the resistor even more. This form of positive feedback is unstable and results in a runaway situation, as we have seen in the chapter on electronics. Thermal runaway in this case. If we want to protect an element from overheating we should place a PTC in series with it.

A thermistors main advantage lies in the fact that it is very cheap, easily an order of magnitude cheaper than a LM35 or PT100. If we want to use the sensor to convert temperature to voltage, we use a Wheatstone bridge (see Section 2.2.1 of Chapter 2), a differential voltage divider with in one leg two normal resistances not connected to the measured object or not susceptible to resistance changes, and in the other leg one normal resistor and the thermistor. These Wheatstone bridges introduce non-linearities, but for a thermistor this is not a serious problem, considering their own highly non-linear nature.

### 4.3.6 Remote sensing

The best sensors are those that do not make contact with the measured object since that minimizes the interference (though from a meta-physical point of view, it is *impossible* to measure something without interference!). A good example of contact-less measurement is remote sensing. An example of this is the determination of temperature of an object through its radiation emission spectrum.

It is based on the principle that hot bodies emit light and the color of the radiation is determined by its temperature. If we determine the spectrum of the

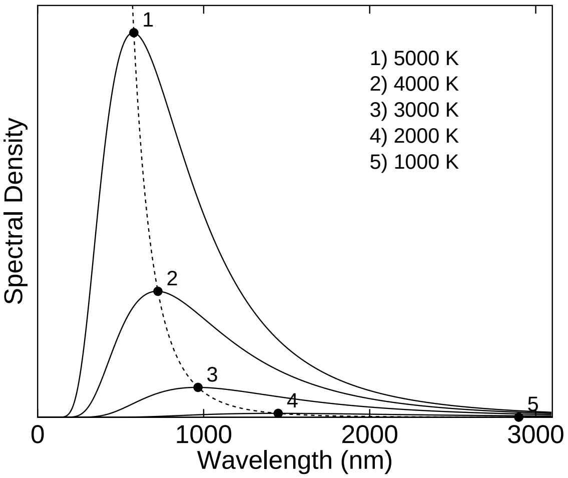


**Fig. 4.11**: Planck's black-body radiation spectrum. For higher temperatures, the total amount of radiation increases and the maximum shifts to shorter wavelengths given by Wien's law (dashed line). Determination of intensity at three fixed wavelength allows for reconstruction of curve and determination of the temperature

radiation, we know the temperature of the object, without even getting near to it. Hot objects emit infrared radiation, that what we normally call 'heat'. When we increase the temperature, the object starts glowing red. Further increasing the temperature makes it turn yellow, green and even blue. An example is lava coming out of a volcano, which is so hot – approximately 1000 °C to 1200 °C – that it is 'red-hot'. Our sun has a temperature of about 6000 °C and it is yellow. Such objects have a continuous emission spectrum and are called black bodies. That means that they follow Planck's law of black-body radiation, with a spectrum density as a function wavelength given by

$$I(\lambda) = \frac{hc^2}{\lambda^5}\frac{1}{\exp(h/\lambda kT) - 1}, \tag{4.12}$$

with $h$ Planck's constant, $c$ the speed of light, $k$ Boltzmann's constant, and $\lambda$ the wavelength, see Figure 4.11. The maximum of radiation occurs at a wavelength

$$\lambda_{\mathrm{max}} = \frac{b}{T}, \tag{4.13}$$

which is called Wien's law, with $b = 2.898\times10^{-3}$ Km (kelvin meter).

The total emitted power is proportional to the area of the surface of the object and the fourth power of its temperature in what is called Stefan-Boltzmann law. However, we do not have easy access to the measurement of the area, nor is it easy to determine the total power emitted (we would have to measure

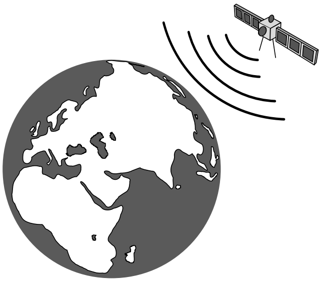


**Fig. 4.12**: Example of remote sensing, satellite measuring the temperature of the Earth at a distance from space

in all directions, placing an integrating sphere around the object). Nor is it needed. To measure the temperature, the intensity is measured at three fixed wavelengths from which the black-body radiation curve can be reconstructed and the temperature estimated.

This technique is used in remote sensing our planet by satellites and is one of the most reliable methods for determining the average temperature. Other famous remote-sensing techniques is the sensor for determining car speeds by radar, see Section 4.5.6.

### 4.3.7 Non-electronic sensor

Not always is exact knowledge of the temperature needed. Not always do we need to design a complicated and expensive electronic circuit. For security reasons, not always can we rely on the availability of electricity. A good example is a temperature sensor that detects if the flame burning gas is still on. If not, the gas flow has to be cut as fast as possible, to not create an explosive mixture in the room. We could now go about designing an expensive electronic gas sensor or an advanced temperature measurement system. Yet, that would be overkill *and* unreliable.

For these circumstances we can make use of a simple mechanical switch that is based on the principle of thermal expansion. This is like the iron rails of rail tracks that on a hot summer's day expand faster than the ground and start creating havoc by starting to touch and dislocate each other. Such forces can be enormous, easily enough to open or close a mechanical valve. Likewise, a mechanical 'bi-metal' sensor consists of two layers of different metals welded together. When this stick heats, the metals expand, but because they are different metals, they expand at a different rate. The effect is that the stick curls up in a arc, with the longer metal on the outside of the arc and the shorter metal on the inside, see Figure 4.13.

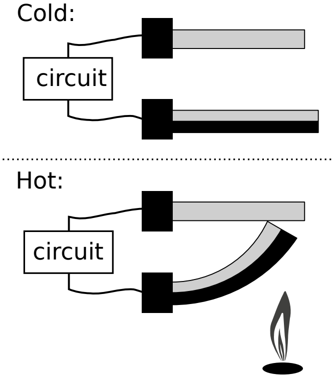


**Fig. 4.13**: A bi-metal mechanical sensor. When the bi-metal heats up, the metal strip composed of two metals with different expansion coefficients heats up and the electronic circuit is closed.

## 4.4 Frequency counters

To convert a frequency into a voltage or to a digital value there are two options. In the first option, the signal is amplified unto distortion, putting the amplifier in saturation, resulting in a conversion of the sinusoidal frequency into a square wave signal. This can then be fed to digital electronics circuits such as counter. The counter is periodically read ('latched') and reset, see Figure 4.14.

An alternative, if we want an analog output signal, an output voltage proportional to the input frequency, a circuit can be used that detects zero crossings of the signal. Every time this happens, a short constant-width pulse $\Delta t_1$ of $V_\mathrm{CC}$ is generated at the output, for instance with the 555 IC of Chapter 2. Since the distance $\Delta t_0$ between these two pulse is depending on the frequency, $\Delta t_0 = 1/f$, the average voltage is given by the weighing of time of the pulses are on, $\Delta t_1$, and the time they are off, $\Delta t_0 - \Delta t_1$,

$$V_\mathrm{o} = \frac{\Delta t_1}{\Delta t_0} \times V_\mathrm{CC} = (\Delta t_\mathrm{t} V_\mathrm{CC}) \times f. \tag{4.14}$$

In other words, a linear sensor is made this way, with the output voltage linearly depending on the input frequency.

## 4.5 Spatial sensors; position, speed

### 4.5.1 Displacement sensor: potentiometer

These are sensors to measure position in space. The simplest one one can imagine is a linear variable resistor, or linear potentiometer that directly translates position into resistance and voltage, respectively. A linear potentiometer consists of a sliding contact B at a resistor connected at its extremes A and C.

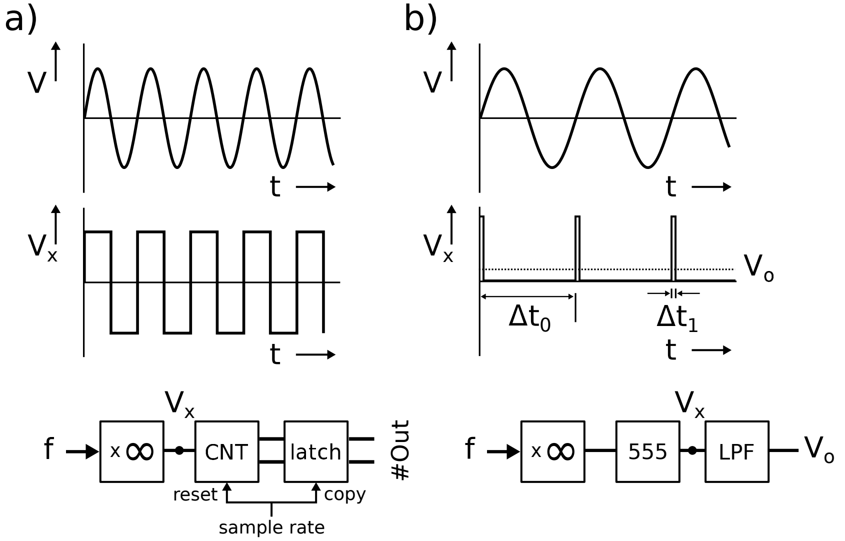


**Fig. 4.14**: a) Converting frequency to a digital frequency implies amplifying the signal into saturation and using digital counters. b) Converting to an analog voltage consists of a zero-crossing-detection circuit (basically an infinite amplifier as well) fed to a single-pulse circuit (based on a 555 IC) and a low-pass filter

The resistance between A and B and between B and C is variable, depending (linearly) on the position $x$ of the contact point B, for instance

$$R_{\mathrm{AB}} = \frac{x}{x_{\max}} R, \tag{4.15}$$

with $R$ the total, nominal resistance of the potentiometer. When used as a voltage divider between voltages connected at A and C, the voltage at B is a weighed average of the voltages at A and C, for $x = 0$ the weight is totally on A and for $x = x_{\max}$ the weight is totally on C. For instance, when A is connected to ground,

$$V_{\mathrm{B}} = \frac{R_{\mathrm{AB}}}{R_{\mathrm{AB}} + R_{\mathrm{BC}}} V_{\mathrm{C}} = \frac{R_{\mathrm{AB}}}{R} = \frac{x}{x_{\max}} V_{\mathrm{C}}, \tag{4.16}$$

and the potentiometer is a linear displacement sensor with a sensitivity equal to $S \equiv \mathrm{d}V/\mathrm{d}x = V_{\mathrm{C}}/x_{\max}$.

Also existing, and actually more often used, are rotary (angular) potentiometers. In this case a knob rotates the core of the potentiometer, to which the sliding contact B is connected. In this case the angle $\alpha$ is linearly translated to a resistance and voltage,

$$R_{\mathrm{AB}} = \frac{\alpha}{\alpha_{\max}} R. \tag{4.17}$$

It is also possible to convert an rotary potentiometer to a linear displacement sensor by attaching a rope to the axis; unrolling the rope will rotate the axis

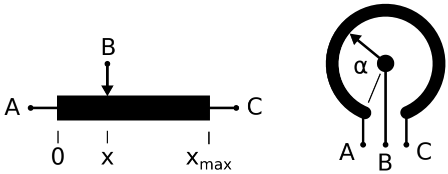


**Fig. 4.15**: Potentiometers. Linear (left) and rotary (right). The total resistance between A and C is constant, and equal to the nominal value written on the element. B is a sliding contact on top of the resistor. The resistance between A and B and between B and C is variable, depending (linearly) on the position $x$ in a linear potentiometer and on the angle $\alpha$ in a rotary potentiometer. When used as a voltage divider between voltages connected at A and C, the voltage at B is a weighed average of the voltages at A and C, for $x = 0$ or $\alpha = 0$ the weight is totally on A and for $x = x_{\max}$ or $\alpha = \alpha_{\max}$ the weight is totally on C

and hence change the resistance value,

$$x = \frac{\alpha}{2\pi} r, \tag{4.18}$$

with $r$ the radius of the axis of the potentiometer around which the rope is wound.

### 4.5.2 Displacement sensor: LVDT (linear differential variable transformer)

A more advanced, contactless, distance sensor is the LDT (linear displacement transducer) working on the principle of conductance compared to the principle of conductance of the potentiometers. It is based on the effect ferromagnetic material has on the inductance of a coil when placed inside it. The inductance is linearly dependent on the value of the permeability $\mu$ of the material. The permeability of a typical ferromagnetic material used, for instance mu-metal, or permalloy, easily lies in the tens of thousands. That makes the inductance of an empty coil as good as negligible compared to a coil filled with the material.

This effect is then used in a voltage divider consisting of two inductors $\mathrm{L}_1$ and $\mathrm{L}_2$ placed in series and driven with a sinusoidal voltage, $V_{\mathrm{ac}} = v_{\mathrm{ac}} \sin(\omega t)$. Alternatively, it can be seen as one inductor coil, tapped at the middle, see Figure 4.16. When the ferromagnetic bar is totally inside the top inductor, $\mathrm{L}_1$, this inductance value is much larger, $L_1 \gg L_2$, and the output voltage amplitude is nearly zero. On the other hand, when the bar is inside inductor $\mathrm{L}_2$, we have the situation $L_1 \ll L_2$ and the output voltage amplitude is equal to the driving voltage amplitude $v_{\mathrm{p}} = v_{\mathrm{ac}} * L_2/(L_1 + L_2) \approx v_{\mathrm{ac}}$. When the bar is halfway, both inductances are equal, $L_1 = L_2$ and $v_{\mathrm{p}} = v_{\mathrm{ac}}/2$. In a differential setup, using a Wheatstone bridge at the entrance of a differential amplifier, comparing the signal $v_{\mathrm{p}}$ with a perfect voltage divider consisting of two equal

**Table 4.III**: Wind direction codes and Gray codes. The latter to avoid multiple transitions between two adjacent states

| MSB - LSB | Digital | Direction | Gray code |
|---|---|---|---|
| 0 0 0 0 | 0 | 0°-22.5° | 0 0 0 0 |
| 0 0 0 1 | 1 | 22.5°-45° | 0 0 0 1 |
| 0 0 1 0 | 2 | 45°-67.5° | 0 0 1 1 |
| 0 0 1 1 | 3 | 67.5°-90° | 0 0 1 0 |
| 0 1 0 0 | 4 | 90°-112.5° | 0 1 1 0 |
| 0 1 0 1 | 5 | 112.5°-135° | 0 1 1 1 |
| 0 1 1 0 | 6 | 135°-157.5° | 0 1 0 1 |
| 0 1 1 1 | 7 | 157.5°-180° | 0 1 0 0 |
| 1 0 0 0 | 8 | 180°-202.5° | 1 1 0 0 |
| 1 0 0 1 | 9 | 202.5°-225° | 1 1 0 1 |
| 1 0 1 0 | 10 | 225°-247.5° | 1 1 1 1 |
| 1 0 1 1 | 11 | 247.5°-270° | 1 1 1 0 |
| 1 1 0 0 | 12 | 270°-292.5° | 1 0 1 0 |
| 1 1 0 1 | 13 | 292.5°-315° | 1 0 1 1 |
| 1 1 1 0 | 14 | 315°-337.5° | 1 0 0 1 |
| 1 1 1 1 | 15 | 337.5°-0° | 1 0 0 0 |

resistances, $v_{\mathrm{n}} = v_{\mathrm{ac}}/2$, an output voltage amplitude response is as shown in Figure 4.16.

### 4.5.3 Angular sensor

To measure an angle we can make use of a rotary potentiometer as described above. However, these have contact (and friction) with the object measured. Moreover, they have a limited angle range; even if we use multi-turn potentiometers, there comes a point when it reaches its limit. This is especially troublesome when we want to measure an angle that can vary infinitely. Take for example a wind direction meter. This has to be able to keep on rotating in the same direction forever, something that a simple potentiometer cannot do.

The solution is a set of LED-optical sensor pairs (optocouplers) described before. Between them is sandwiched a disk with holes, see Figure 4.17. Since the holes in the disk are binary coded to represent the angle of the disk, the binary combination of the LEDs is an indication of the wind direction. With 4 LEDs there are $2^4 = 16$ states, see Table 4.III.

The problem with this system is that at some transitions more than one bit changes state. Take for example the state 15 changing into state 0 when the wind is veering (changing direction clockwise). All four bits change at this transition. The problem is that they do not change simultaneously. However hard we try to make the system as aligned as possible. The sensors always respond a little bit differently. Imagine first the outside bit (LSB) changes,

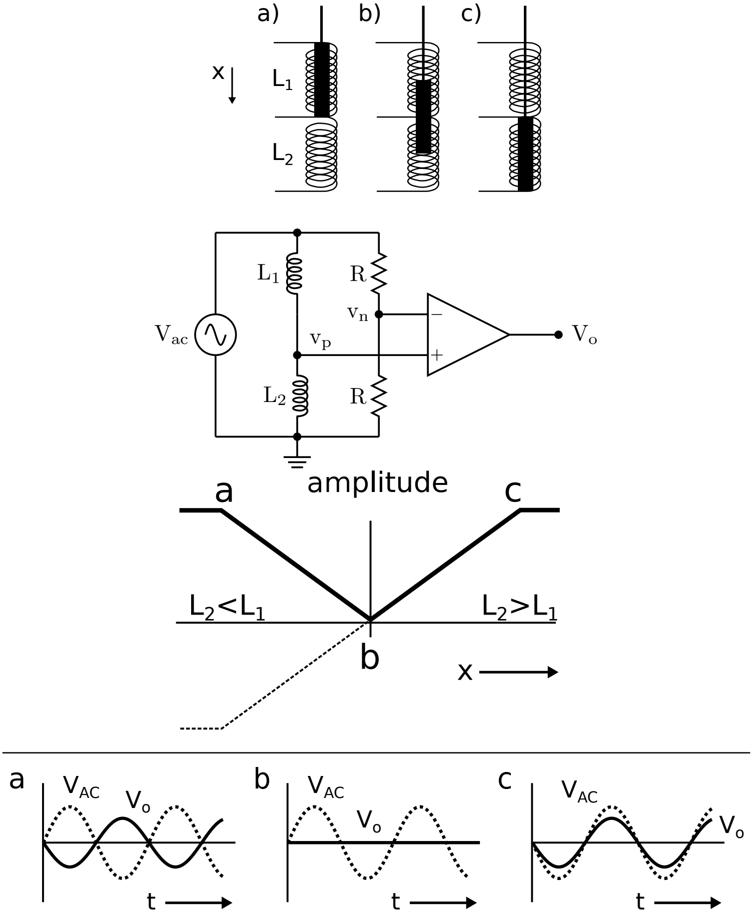


**Fig. 4.16**: LVDT sensor consists of two coils through which a ferromagnetic bar moves. This changes the inductance of the coils. Three situations are show, a) with bar inside coil L1, b) with bar halfway and c) with bar in coil inside coil 2. With a Wheatstone bridge at the entrance of a differential amplifier the inductor-voltage divider is compared with the resistor-voltage divider, and gives signal amplitudes as shown below

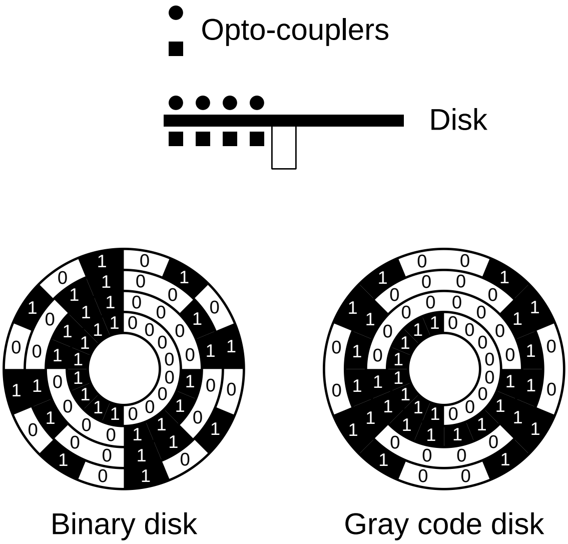


**Fig. 4.17**: A continuous (infinite range) angular sensor, for instance for measuring wind direction can be made of a disk with holes coding the angle. To avoid multiple transitions, Gray code can be used where only one bit between two adjacent states changes

then the inside bit (MSB), then bit 2, then bit 3. What we get in a short interval of time is the changes 1111 (15) → 1110 (14) → 0110 (6) → 0010 (2) → 0000 (0). The wind seems to move all over the place.

To avoid this, Gray code has been invented. By careful choice of bit patterns, only one bit can change at a transition between adjacent states, see Figure 4.17 and Table 4.III.

### 4.5.4 Angular speed sensor, RPM (rotations per minute)

The Maxwell equations (Table 3.I of Chapter 3) tell us that current causes a magnetic field and changing magnetic field causes electric field and thus current. We can make use of this second effect for a movement sensor. More precisely, for a sensor that detects passage of an object. We connect a magnet to the object and place a coil where we want the passage of the object to be detected, see Figure 4.18. When the magnet is moving past the coil, the time derivative of the magnetic field in the coil is large and the current large as well. Exactly when the magnet is aligned with the coil, the current inverts.

A rotational speed (RPM, rotations per minute) sensor can be made by counting the frequency of spikes. Disk drives normally use this technique, either with a coil, or with a Hall sensor, to be discussed later, where it is also used to synchronize the disk, to know the beginning of sectors on the disk, apart from

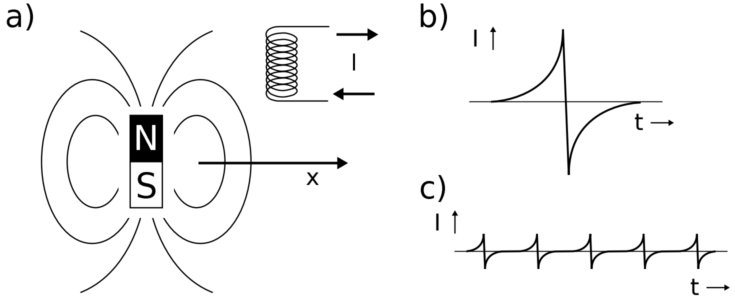


**Fig. 4.18**: a) A moving magnet passing under a coil induces a current as shown in b). In c) the repetition rate of the pulses are the rotation rate of the disk to which the magnet is connected

knowing the rotational speed.

Similar to this is the optical binary detector, that consists of an LED combined with an optical sensor, a so-called opto-coupler, as shown also before in the angular sensor. These opto-couplers are also used in the write-protect detector of legacy floppy disks.

### 4.5.5 Speed sensor; RPM

A simple system for measuring speed, and one used in most cars, is measure the RPM of one of its wheels and translate this into a speed. For this the only thing to be done is multiply the angular velocity by the radius of the wheel,

$$v = \omega R, \tag{4.19}$$

with $v$ the speed, $\omega$ the angular velocity ($2\pi$ times the angular frequency $f$), and $R$ the radius of the wheel. As an example,

**Question**: A car with 15-inch wheels has them rotating at 700 RPM. What is its speed?
**Answer**: 15 inch (37.8 cm) wheels have a circumference of 2.375 m. If they rotate 700 times per minute they rotate 42 thousand times per hour. That is a then distance of $4.2 \times 10^4 \times 2.375 = 99.75$ km per hour.

Because every pulse of the RPM sensor corresponds to a certain distance, the same sensor can also be used for the distance measurement, a.k.a. odometer. It is obvious that the sensor has to be recalibrated if the wheels are changed to other sizes. Also, tires are slightly smaller when the profile wears out. Then it seems the car is going faster than it actually is. This explains the error most cars have on the speedometer, because the factory is calibrating the sensor on the safe side for the largest tires imaginable, and anyway this gives the client the notion of having bought a fast and economical car, while in fact it is probably nothing special.

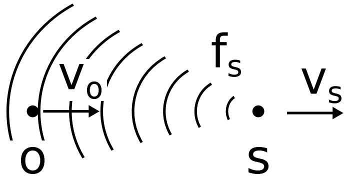


**Fig. 4.19**: Doppler effect. A source s going with speed $v_{\mathrm{s}}$ is emitting wave with frequency $f_{\mathrm{s}}$. The object o moving with a speed $v_{\mathrm{o}}$ is receiving it with a frequency $f_{\mathrm{o}}$ depending on the speed of the source and object and the wave velocity of the medium.

### 4.5.6 Speed sensor; Doppler

A nice contactless way of measuring speed is by using the Doppler effect of waves, named after scientist Christian Doppler, who described it early in the 19th century. We all know this effect, an approaching ambulance with the siren on emits a higher frequency sound. When it passes us, the frequency suddenly drops. The same effect occurs when we are ourselves moving past a stationary sound source; approaching makes the frequency higher, distancing makes it lower. This can be put in a general form, the Doppler equation,

$$f_{\mathrm{o}} = \left( \frac{v_{\mathrm{m}} + v_{\mathrm{o}}}{v_{\mathrm{m}} + v_{\mathrm{s}}} \right) f_{\mathrm{s}}, \tag{4.20}$$

in which $v_{\mathrm{s}}$ and $v_{\mathrm{o}}$ are the speed of the source and the object receiving the sound respectively, $f_{\mathrm{s}}$ the frequency as emitted by the source, and $v_{\mathrm{m}}$ the wave velocity of the medium, for instance the speed of sound in air, or the speed of light when electromagnetic waves are used. From the above equation it is clear that if the object is approaching the source, either $v_{\mathrm{o}}$ positive, or $v_{\mathrm{s}}$ negative, the frequency increases.

An interesting singularity exists when the speed of the source is equal to the speed of propagation of the waves in the medium, $v_{\mathrm{s}} = v_{\mathrm{m}}$. This occurs, for instance, when an airplane is traveling with the speed of sound. At that moment, the airplane travels with the same speed as that of the waves, it keeps on adding acoustic energy to the same wavefront, which to the observer standing on the ground, by the time this shockwave is received there is an instantaneous arrival of a lot of acoustic energy, comparable to the sound of an explosion. This is called a sonic boom.

The Doppler effect can be used in a speed detector, for instance as used by the police to determine the speed of a car with a radar gun. In this equipment, waves are emitted by the gun, reflected by the moving car and received again by the gun. The difference frequency by emitted and received waves is proportional to the speed of the car. For a static emitting source, the Doppler equation (Eq. 4.20) tells that the moving (o) car receives the waves at a frequency

$$f_{\mathrm{o}} = \frac{v_{\mathrm{m}} + v_{\mathrm{o}}}{v_{\mathrm{m}}} f_{\mathrm{s}}. \tag{4.21}$$

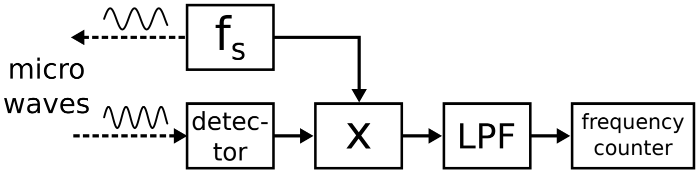


**Fig. 4.20**: A Doppler gun as used by the police. A source is emitting microwaves (GHz) waves. They are reflected by the moving car and the returning frequency is multiplied with the source frequency resulting in sum and difference frequencies. The sum frequency is filtered off with a low-pass filter. The difference frequency is counted and this is directly proportional to the speed of the car, as long as it is not moving at relativistic speeds

They are reflected back and received at a frequency

$$f_{\mathrm{s}}' = \frac{v_{\mathrm{m}}}{v_{\mathrm{m}} - v_{\mathrm{o}}} f_{\mathrm{o}} = \frac{v_{\mathrm{m}} + v_{\mathrm{o}}}{v_{\mathrm{m}} - v_{\mathrm{o}}} f_{\mathrm{s}}. \tag{4.22}$$

A Taylor expansion of this equation is

$$f_{\mathrm{s}}' = \left[1 + 2\frac{v_{\mathrm{o}}}{v_{\mathrm{m}}} + \mathcal{O}\left(\frac{v_{\mathrm{o}}^2}{v_{\mathrm{m}}^2}\right)\right] f_{\mathrm{s}}. \tag{4.23}$$

Thus, the difference frequency is in first order directly proportional to the speed of the car,

$$\Delta f = f_{\mathrm{s}}' - f_{\mathrm{s}} \approx 2\frac{v_{\mathrm{o}}}{v_{\mathrm{m}}} f_{\mathrm{s}}, \tag{4.24}$$

as long as the car is going much slower than the speed of light. The sensitivity of this transducer is

$$S_{\mathrm{v}\to\mathrm{f}} = \frac{\mathrm{d}f}{\mathrm{d}v} = \frac{2}{v_{\mathrm{m}}} \ (\mathrm{m}^{-1}). \tag{4.25}$$

The idea is now to multiply the returning signal with the original source frequency. Mathematics tells that multiplying two frequencies results in the difference and sum frequencies,

$$f_{\mathrm{s}} \otimes f_{\mathrm{s}}' = \Delta f \oplus (f_{\mathrm{s}}' + f_{\mathrm{s}}). \tag{4.26}$$

The sum frequency can easily be filtered off with a low-pass filter and the remaining difference frequency can be measured with a frequency counter as described before. See Figure 4.20.

The same microwaves can also be used for detecting the position of airplanes in the sky, or (less bellicose) precipitation. A source is emitting microwave pulses and they are reflected by the object and received by the source. However, instead of detecting the frequency shift, which would be a measure of the speed of the object as discussed above, instead the time of arrival of the reflected microwave pulse is measured. Since the wave velocity of the radiation is well known (namely the speed of light), the time difference between sending the

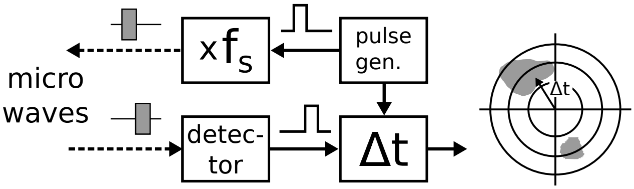


**Fig. 4.21**: A weather radar detecting precipitation. It consists of a source of microwave pulses that are reflected on the rain drops and a time-delay detection circuit. The time delay is a measure for the distance to the rain

pulse and receiving the reflected signal is directly proportional to the distance to the object. By careful choice of microwave frequency, the type of object can be selected. For weather radar, the wavelength is chosen such that they are reflected only by droplets of water of a certain size and not by clouds.

### 4.5.7 Extensometer (strain gauge)

'Strain' is the deformation of an object as caused by external forces, the latter being called 'stress'. We are familiar with that; if we press on an object, the object normally gets smaller. When we stress an object, it gets strained. This section is about quantifying such strain. But first we have to make some definitions of the strain, stress and ratio of the two.

Strain $\epsilon$ is the relative change in size in a dimension of an object, for instance the change in length,

$$\epsilon = \frac{\delta L}{L}. \tag{4.27}$$

They result from applying a pressure $P$. The ratio between applied pressure and strain is called Young's modulus,

$$E \equiv \frac{P}{\epsilon}. \tag{4.28}$$

Table 4.IV gives some values of this parameter of some materials. The higher $E$, the more pressure we need to deform the material by one percent in length. Concrete is 'hard' and needs a lot of force, $E$ is large. Rubber on the other hand is easy to deform and has a small Young's modulus.

To measure the strain a strain gauge can be used, see Figure 4.22. It consists of a strip resistor of length $L$, width $W$ and height $h$ made of a material with resistivity $\rho$. From the physics chapter we know that such a strip has a resistance given by $R = \rho L/Wh$.

If we stretch or compress the resistor, the dimensions change and the resistance value will change with it. A gauge factor can be defined that is the ratio of the relative change in resistance to the relative change in length,

$$k \equiv \frac{\delta R/R}{\delta L/L} = \frac{\delta R/R}{\epsilon}. \tag{4.29}$$

**Table 4.IV**: Material parameters

| Material | Young's modulus | Poisson ratio | Resistivity |
|---|---|---|---|
| Concrete | 27.3 GPa | 0.2 | 100 $\Omega$m |
| Rubber | 7.9 MPa | 0.5 | $10^{13}$ $\Omega$m |
| Cork | 32 MPa | 0 | $8.3 \times 10^{9}$ $\Omega$m |
| Copper | 130 GPa | 0.34 | $16.8 \times 10^{-9}$ $\Omega$m |
| Aluminum | 70 GPa | 0.35 | $28.2 \times 10^{-9}$ $\Omega$m |

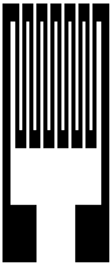


**Fig. 4.22**: Strain sensor (gauge). Changes in length are translated into changes in resistance

In first order, if stress is applied (along L), only that dimension changes size. The gauge factor is then unity:

$$k = \frac{\mathrm{d}R/R}{\mathrm{d}L/L} = \frac{\mathrm{d}R}{\mathrm{d}L} \cdot \frac{L}{R} = \frac{\rho}{Wh} \cdot \frac{L}{\rho L/Wh} = 1. \tag{4.30}$$

However, most materials also change size in the two dimensions perpendicular to the one of the applied pressure. If we press on top of a rubber ball, it gets squeezed and the height is reduced. But, simultaneously, the width is increased; the ball becomes an oblate spheroid (disk-shaped). In fact, the volume of a rubber ball stays the same. That is a property of rubber. For most materials, $W$ and $h$ increase as $L$ is decreased. In other words, $W$ and $h$ are functions of $L$. The gauge factor can then be expressed in terms of the derivative using the chain rule,

$$\begin{aligned} \frac{\mathrm{d}R}{\mathrm{d}L} &= \frac{\partial R}{\partial L} + \frac{\partial R}{\partial W} \cdot \frac{\mathrm{d}W}{\mathrm{d}L} + \frac{\partial R}{\partial h} \cdot \frac{\mathrm{d}h}{\mathrm{d}L} + \frac{\partial R}{\partial \rho} \cdot \frac{\mathrm{d}\rho}{\mathrm{d}L} \\ &= \frac{\rho}{Wh} - \frac{\rho L}{W^2 h} \cdot \frac{\mathrm{d}W}{\mathrm{d}L} - \frac{\rho L}{W h^2} \cdot \frac{\mathrm{d}h}{\mathrm{d}L} + \frac{L}{Wh} \cdot \frac{\mathrm{d}\rho}{\mathrm{d}L}. \end{aligned} \tag{4.31}$$

The gauge factor of Equation 4.29 then becomes

$$k = \frac{\mathrm{d}R/R}{\mathrm{d}L/L} = \frac{\mathrm{d}R}{\mathrm{d}L} \cdot \frac{L}{R} = 1 + 2\nu + \frac{d\rho/\rho}{\epsilon}, \tag{4.32}$$

with Poisson's ratio $\nu$ defined as the relative changes in a perpendicular direction divided by the relative changes in size in the direction parallel to the force,

$$\nu \equiv -\frac{\mathrm{d}W/W}{\mathrm{d}L/L} = -\frac{\mathrm{d}h/h}{\mathrm{d}L/L}. \tag{4.33}$$

Here also the piezoelectric effect is included that states that when the material changes size, the resistivity is changed. When we take a closer look at the gauge factor, we see that it is not needed to know the sizes in the three dimensions to calculate the effect of percentual changes on the resistance.

Poisson's ratios for some materials are given in Table 4.IV. A material that keeps its volume under stress has a Poisson's ratio equal to 0.5 (see Exercise 1). We see that rubber is such a material. Interestingly, cork has a Poisson's ratio equal to zero. That implies that if we press cork, it will not expand in the two perpendicular dimensions, and that is the reason why you can press a cork into a bottle. This unique property makes cork a fascinating material and valuable for everybody who likes a good wine.

**Question**: A copper resistance of nominal 1 kΩ is stretched 1% in length. What is the new resistance value?
**Answer**: The increase in resistance is given by

$$\begin{aligned}\Delta R &= \frac{\mathrm{d}R}{\mathrm{d}L}\Delta L = \left(\frac{\mathrm{d}R}{\mathrm{d}L}\frac{L}{R}\right)\cdot\frac{R}{L}\Delta L = kR\frac{\Delta L}{L} = (1+2\nu)R\frac{\Delta L}{L} \\ &= 1.68 \times (1\ \mathrm{k}\Omega) \times 1\% = 16.8\ \Omega. \end{aligned} \tag{4.34}$$

The new resistance value is thus 1017 Ω. Note that it was not needed to know the dimensions of the resistance.

Strain sensors, or extensometers as they are often called, are quite sensitive and find their use in places where tiny changes in length need to be measured. Another example is an earthquake meter, where a heavy ball is placed on the end of a metal bendable strip, see Figure 4.23a. The extensometer is glued on both sides of the strip. The ball has high inertia, and if the earth moves, the metal strip will bend, expand on one side and contract on the other. With a differential Wheatstone configuration the movement can be translated into a a voltage, see Figure 4.23b. The advantage is that it is sensitive and linear, and the voltage swing can easily and reliably be calculated.

Finally, we can also make a gauge sensor of a capacitor, following similar reasoning. This is done in Exercise 2.

## 4.6 Chemical sensors

Chemical sensors are based on chemical reactions. A reaction has a tendency to go in a certain direction and this tendency one can imagine as a force. A force, like any other force, then also has a potential, which is called the chemical

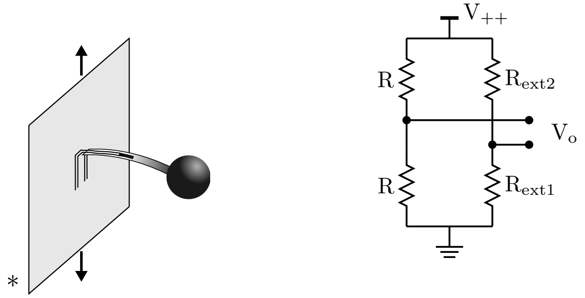


**Fig. 4.23**: Earthquake sensor. A heavy ball on a flexible rod. On two sides of the rod extensometers are placed. If the Earth shakes, the inertia of the ball makes the rod extend on one side and contract on the other. The extensometers translate these changes into resistance changes. In a Wheatstone bridge (right) these changes are translated into a voltage $V_\mathrm{o}$. Even better would be using two sets put in a cross-link Wheatstone configuration

potential, $\mu$ (unit: J/mol; how much energy is liberated when one mole of the substance is actually reacted). Because chemical reactions involve the transfer of charge (either electrons or protons), it is also often very easy to convert this chemical potential into an electrical potential if the charge is supplied through an external (electrical) circuit, like in a voltaic cell (a.k.a. battery). Here two examples are given, namely protonic acid-base reactions in a pH sensor and electronic reduction-oxidation reactions in lead-acid battery.

### 4.6.1 Acid-base reactions. A pH sensor

The pH is a measure of acidity of a solution. It is defined as the logarithm (base 10) of the molar concentration or activity of hydrogen ions (protons) in solution. In water, the hydrogen is bonded to a water molecule, in a so-called hydronium ion, $H_3O^+$, thus

$$\mathrm{pH} \equiv -\log_{10}\left(\frac{[\mathrm{H_3O^+}]}{\mathrm{mol/L}}\right). \tag{4.35}$$

Because of a minus sign, the higher the pH, the lower the concentration of ions. The pH of neutral is equal to 7, something that can easily be understood when we realize that, in water, self ionization occurs,

$$2\mathrm{H_2O} \rightleftharpoons \mathrm{H_3O^+} + \mathrm{OH^-}. \tag{4.36}$$

The unitless reaction constant of this equation is given by

$$K = \frac{[\mathrm{H_3O^+}][\mathrm{OH^-}]}{[\mathrm{H_2O}]^2}. \tag{4.37}$$

(The water concentration enters squared in the denominator because there are two water molecules on the left side of the reaction equation). The concentration of water can be considered constant, and thus a new constant can be defined,

$$K_\mathrm{w} = [\mathrm{H_3O^+}][\mathrm{OH^-}]. \tag{4.38}$$

At 298 kelvin, this constant has a value

$$K_\mathrm{w} = 1.00 \times 10^{-14}\ \ \mathrm{mol^2/L^2}. \tag{4.39}$$

Chemists normally like to omit the unit, and simply say that $K_\mathrm{w} = 10^{-14}$. Likewise, they omit the unit in the definition of pH given above, since it is assumed obvious that concentrations are in mol/L. While the implicit unit is obvious for them, its omission can cause serious confusion among non-chemists and we'd better always keep it. The exercise of the pH sensor (5) at the end of this chapter is an example of this effect of obtaining ridiculous results if we forget the units.

Neutral water has (necessarily) an equal number of $\mathrm{H_3O^+}$ ions as $\mathrm{OH^-}$ ions and the pH is thus 7. Note also that the similarly defined pOH as the logarithm of concentration of $\mathrm{OH^-}$ ions is also equal to 7. In fact, pOH is always equal to 14 − pH, as easily follows from the above equations.

A pH sensor measures the concentration of hydrogen ions by measuring the electrical potential of a certain chemical reaction whose chemical potential depends on the concentration of hydrogen ions. Typically the reaction is the one of hydrogen ions with glass in which protons are passed from $\mathrm{H_3O^+}$ to $\mathrm{SiO_2}$,

$$\mathrm{SiO} + \mathrm{H_3O^+} \rightleftharpoons \mathrm{SiOH^+} + \mathrm{H_2O}. \tag{4.40}$$

The above reaction takes place on both sides of a glass membrane. On the inside of a glass bulb, the ion concentration is constant, normally set to a pH of 7, while the outside of the glass bulb is exposed to the solution being measured. See Figure 4.24. The external voltmeter measures the electrical potential caused by the chemical potential difference on both sides of the membrane.

The chemical potential $\mu_\mathrm{A}$ (unit: J/mol) of a reagent A in a chemical reaction is proportional to the logarithm of the chemical activity of the reagent. This activity is normally proportional to the concentration of the reagent, and the chemical potential is thus of the form

$$\mu_\mathrm{A} = \mu_\mathrm{A,0} + RT \ln\left(\frac{[A]}{\mathrm{mol/L}}\right), \tag{4.41}$$

with $R$ the molar gas constant equal to 8.314 J mol$^{-1}$ K$^{-1}$, and $T$ the absolute temperature. The total chemical potential of a reaction measures how much energy is gained or lost when the reaction takes place and is given by the difference in sum of chemical potentials of the reagents on the left side and right side of the reaction equation. Moreover, since we are interested only in the differences in both sides of the membrane, the chemical potential difference is only due to the difference in proton concentration, since the rest is the same on both sides of the membrane.

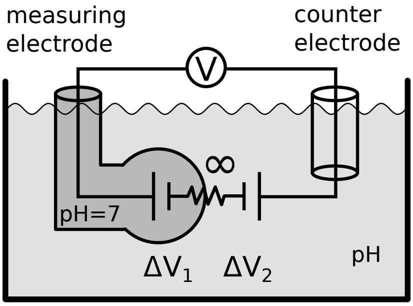


**Fig. 4.24**: Schematic diagram of a pH sensor in solution. It measures the difference in potential of a reaction of protons with glass on two sides of a glass membrane. On one side the pH is constant at 7, while the other side is exposed to the solution. Ideally, the membrane is electrically insulating, so that no reaction takes place and we measure only the potential of the reaction. But for the sensor to work, tiny currents have to pass in the form of protons passing through the membrane

The *electrical* potential is now given by the chemical potential divided by the amount of charge in one mole of reaction, namely one mole of $H^+$. This is Avogadro's number $N_A = 6.02214129 \times 10^{23}$ mol$^{-1}$ times the elementary charge $q = 1.602176565 \times 10^{-19}$ C, the product of which is called the Faraday constant ($F$). The measured external voltage is then the difference of the two potential drops on both sides of the glass membrane (Nernst Equation):

$$\begin{aligned} V &= \Delta V_1 - \Delta V_2 = \frac{RT}{qN_A}\left(\ln\left(\frac{[H_3O^+]}{\text{mol/L}}\right) - \ln(10^{-7})\right) \\ &= \frac{RT}{qN_A}\ln(10) \times (\text{pH} - 7), \end{aligned} \tag{4.42}$$

where the original definition of pH was used, Eq. (4.35), together with the mathematical property that $\log_{10}(x) = \ln(x)/\ln(10)$. The sensitivity of the sensor at room temperature (298 K) is given by

$$S \equiv \frac{\mathrm{d}V}{\mathrm{dpH}} = \frac{RT}{qN_A}\ln(10) = 59.13 \text{ mV/pH}. \tag{4.43}$$

It is obvious from the above equation that the temperature has to be known to make an accurate estimation of the pH on basis of the measured potential. Most commercial pH sensors have temperature sensors incorporated to facilitate this correction.

Ideally, the resistance of the membrane is infinite, since we do not want to let protons pass from the reference liquid into the liquid under test or back. Remember that an ideal sensor cannot cause itself changes of the measured object, something that is called interference and should be avoided; if we want to measure pH, we do not want to *change* the pH by letting protons migrate

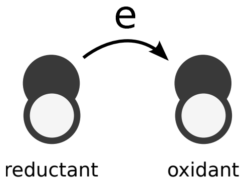


**Fig. 4.25**: The transfer of an electron from a reductant to an oxidant in a redox reaction. The oxidant oxidizes the reductant, or seen from the other side, the reductant reduces the oxidant

to the measured liquid. Moreover, we also do not want to change our reference solution inside the glass, which ideally remains at a constant pH. However, for the sensor to work in a system, tiny currents are needed, and the effective internal resistance is large, but cannot be infinite. As we have seen in the chapter on electronics, dealing with signals coming from high-resistive sources is difficult. We need amplifiers with input resistances much higher than the output resistance of the sensor. The requirements for the electronics are severe. This becomes even more evident if we realize that any current passing through the high-resistance membrane will not only cause interference, but also causes a voltage drop there, remember Ohm's $\Delta V = IR$, that directly distorts our measurements.

### 4.6.2 Redox reactions. A lead-acid battery

In a battery electronic 'redox' reactions take place with transfer of electrons from the reductant to the oxidant. Maybe somewhat confusingly, the process of adding an electron to the oxidant is called reduction and the process of taking an electron from a reductant is called oxidation. Thus, an oxidant oxidizes the reductant and a reductant reduces the oxidant, see Figure 4.25.

This is best explained by an example. In fact something that we all use daily, namely a lead-acid battery normally found in a car. It starts with lead-sulfate ($PbSO_4$) electrodes immersed in water ($H_2O$). When we connect a charger and supply a voltage (larger than 2.05 volt, as we will see), it works as an electrolytic cell, with chemical reactions induced by the external (electric) force. The electrons supplied at one electrode (the so-called cathode) are used in a reaction

$$\mathrm{PbSO_4(s) + 2e^- \longrightarrow Pb(s) + SO_4^{2-}(aq).} \tag{4.44}$$

On the other electrode (the so-called anode), electrons are extracted in a chemical reaction of the type

$$\mathrm{PbSO_4(s) + 2H_2O(l) \longrightarrow PbO_2(s) + SO_4^{2-}(aq) + 4H^+ + 2e^-,} \tag{4.45}$$

with in brackets indicated the state of the material, (s): solid, (l): liquid, (aq): aqueous solution. Thus, the current converts one electrode into lead, the other into lead oxide and the water turns into sulfuric acid, $H_2SO_4$, see Figure 4.26.

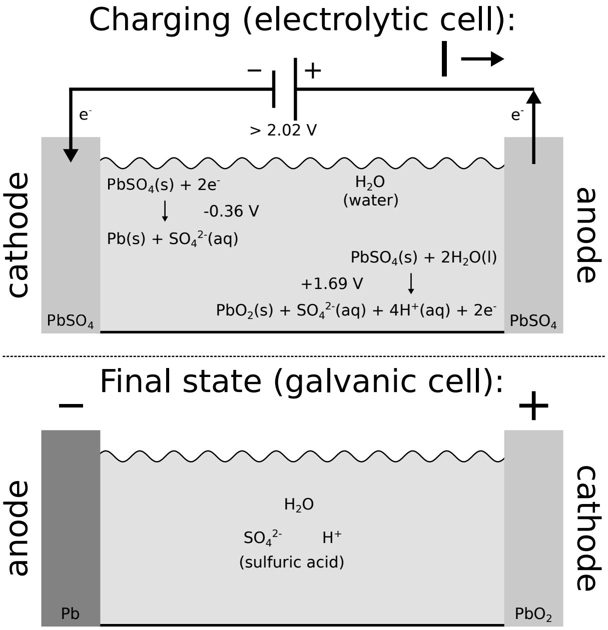


**Fig. 4.26**: Schematic of a lead-acid voltaic cell. Discharged it consists of lead-sulfate electrodes immersed in water. When a current is passed charging the battery it works as an electrolytic cell. The lead sulfate is converted into lead on the negative electrode (cathode) and lead oxide on the positive electrode (anode). The sulfate winds up in the water and together with the protons, this forms sulfuric acid. When discharging, the battery is a galvanic cell, with the negative electrode (now called anode) supplying electrons passing through an external circuit and entering into what is now called the cathode. The state of the material is indicated in brackets: (s): solid, (l): liquid, (aq): aqueous solution. A car battery is a pile of 6 such cells in series

We can now calculate what voltage should be applied to keep the reaction running. We can look at two half-reactions above and look up in a table their standard electrode potentials relative to the reference reaction of hydrogen molecule dissociation ($H_2(g) \rightleftharpoons 2H^+ + 2e^-$, which is defined as 0). Once again, this electrical potential is the chemical potential divided by the charge. We find in the table that at a concentration of 1 mol/L the first reaction has an electrical potential of -0.36 V and the second +1.69 V. A voltage of at least 2.05 volt is thus needed to keep the reaction running. A typical car battery consists of 6 such cells connected in series – hence the word 'battery' – and a total voltage of about 12 volt results.

By charging the battery we have converted electrical energy into chemical energy. To extract the energy and use it again in the form of electric energy, we let the reverse reactions happen by allowing charge to pass through an external electric circuit. This is called a galvanic cell, see the bottom drawing of Figure 4.26. Electrons come out of the Pb-electrode (which is now called the anode, $-$) as the result of the reaction of changing lead back into lead sulfate and on the other electrode (cathode, $+$) the electrons are used to convert the lead oxide back into lead sulfate.

In a battery, it has to be prevented that the reacting species meet each other. In the example above, this was easy, since both sides – lead, and lead oxide – are solid materials that remain fixed at the anode and cathode respectively. In other batteries, membranes have to be used to prevent materials to mix. These make sure that the reactions can only takes place through the passage of electrons via an external circuit. If not, the reactions would take place inside the liquid, which would just heat up.

The difference between the working of the acid-base sensor and the redox battery is not so much as one may think. Basically both boil down to the idea of a chemical potential that translates into an electrical potential. The big difference between the pH sensor and the battery is that in the pH sensor the concentration varies significantly; it can be orders of magnitude. In fact, it is that concentration that is measured through its (logarithmic) impact on the chemical (and thus electrical) potential. On the other hand, in the redox reaction of the battery, the concentrations are rather constant. They are assumed here to be 1 mol/L for the calculation and large variations of this value are not expected; in cases large changes do occur the voltages will change 59 mV for every order of magnitude concentration change, just like for the pH sensor.

Moreover, a pH sensor measures the potential (difference) of the *same* reaction on two sides of a membrane, while the battery voltage is the result of the chemical potential difference of two *different* reactions.

**Question**: What are the reductants and oxidants in the chemical reactions?
**Answer**: The oxidant is the one that receives electrons and the reductant is the one that gives them. Thus, at the anode we find the reductants and at the cathode the oxidants.

Finally, some confusion might exist about the nomenclature of anode and cathode. It is always the anode that emits electrons to an external circuit and where current thus enters. In an electrolytic cell, where chemical reactions are induced by an external force, this means the electrode where the positive pole of an external power source is connected. On the other hand, in a galvanic cell, where internal chemical energy is converted to an external electrical power, this means the negative pole, the electrode with the lower voltage.

### 4.6.3 Examples of chemical sensors

Chemical sensors then can consist of measuring the potential or current of the reaction. As an example may serve a sensor to detect carbon monoxide in the air. It consists of an electrochemical cell where carbon monoxide is oxidized to

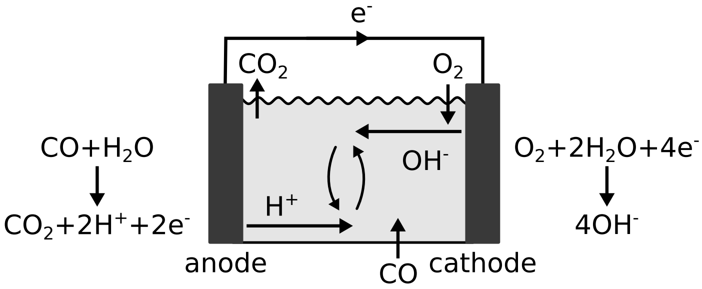


**Fig. 4.27**: Electrochemical carbon oxide sensor detecting the chemical reactions of Eqs. (4.46) and (4.47). The byproducts $H^+$ and $OH^-$ diffuse towards the opposing electrode and somewhere in between they recombine in order to try to maintain equilibrium with the product of the two concentrations given Eq. (4.38)

carbon dioxide in a reaction

$$CO(g) + H_2O(l) \rightarrow CO_2(g) + 2H^+(aq) + 2e^-, \tag{4.46}$$

and at the other electrode, oxygen is reduced,

$$O_2(g) + 2H_2O(l) + 4e^- \rightarrow 4OH^-(aq). \tag{4.47}$$

The byproducts, protons and hydroxide ions are drifting to the opposing electrodes and recombine, in an attempt to restore equilibrium defined in Eq. (4.38). This causes effectively a net current in the solution and closes the electronic circuit. This current can be externally measured and directly depends on the concentration of carbon monoxide.

The same principle is used in oxygen sensors (also known as lambda sensors in cars), for instance one using a Clark electrode, see Figure 4.28. Oxygen enters the potassium-chloride-filled chamber through a membrane (usually polytetrafluoroethylene [PTFE, or 'Teflon'] or fluorinated ethylene propylene, FEP) and is reduced by the reaction

$$O_2(g) + 4H^+(aq) + 4e^- \rightarrow 2H_2O(l), \tag{4.48}$$

at the (inert, non-reacting) platinum cathode. The electrons for this reduction of oxygen are supplied by the electrons generated at the silver anode in the reaction

$$Ag(s) + Cl^-(aq) \rightarrow AgCl(s) + e^-. \tag{4.49}$$

The electrochemical cell is polarized at a certain voltage (typically 0.6 V) and the measured current of this reaction is a direct measure of the concentration of oxygen.

A principal problem with chemical sensors is their specificity. As an example, we want our carbon monoxide sensor to detect *only* carbon monoxide and

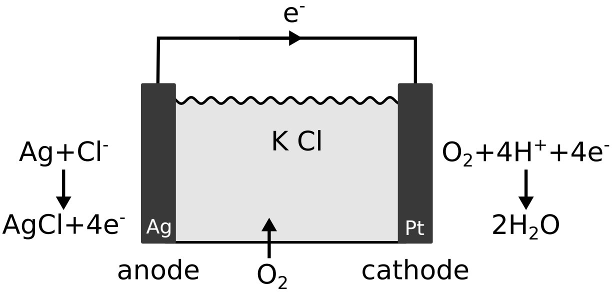


**Fig. 4.28**: Oxygen sensor using a Clark electrode for the cathode

nothing else and our oxygen sensor to be responsive to oxygen only. A way to ensure this is by a careful choice of a membrane that admits only the interesting species to the reaction chamber. Often optical properties of materials are used, since they can be more specific. As an example, carbon dioxide concentrations are nowadays measured through the absorption spectrum.

Also, often indirect methods of detection are used. Again, carbon monoxide is a good example. As a principle of working, a CO sensor contains (synthetic) hemoglobin – an ingredient that is also present in blood – and the reaction of it to CO changes its color from white to dark brown. An optical sensor is then used to convert the luminosity into an electrical signal.

In some cases electrical properties are used, the presence of the material to be detected changes directly the electrical properties of another, electronic, material used in the sensor. As we have seen, semiconductor properties depend a lot on the presence of dopants. Some gas sensors rely on the fact that the simple presence of a substance changes the density of dopants and thus the electrical properties of the device.

## 4.7 Power switching actuators

A common problem in the design of systems with actuators is that the electronics of signals is normally low power, while the final element to be switched (lightbulb, engine, air conditioning, etc.) is high power. An example is a 40-watt warning light that must be switched on when the temperature is too high. It is not possible to directly connect the lightbulb to a opamp circuit, for instance a comparator. The maximum output power supplied by an opamp is some tens of milliwatts at best. For sure it will not switch on a lightbulb.

A solution is to use a relay, see Figure 4.29. This is an element that can switch a large current with the use of a small current. The small current goes through a coil with a large number of windings and creates a strong magnetic field. This field attracts a piece of ferromagnetic material that is connected at the loose end of a cantilever. The cantilever rotates around the pivot and touches the other side of the metal, thus closing the circuit and large currents

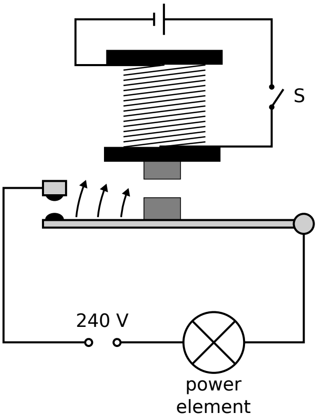


**Fig. 4.29**: Relay. A tiny current in the primary loop creates a strong magnetic field, attracting a cantilever and closing the secondary loop supplying the current to the power element

can flow as supplied by a secondary supply source, for instance the 240 volt net power. The small current of the primary loop can easily be supplied by an opamp.

The relay can be recognized by its characteristic 'clicking' sound when it is switching, for instance the indicators in older cars. One disadvantage of the relay is that it is a mechanical element and as such subject to wear and resulting in a limited lifetime. An additional problem is that every time the current is interrupted, a spark can be generated (see the chapter on physics). These sparks evaporate the contacts or can weld them together. In either case, the relay will eventually fail.

A low-power variant of the relay is the reed relay, before widely used in telephone switching centrals. It consists of two strips of ferromagnetic metal inside an inert-gas-filled glass tube. When a magnetic field is present, independent of the direction of the magnetic field, the metal gets magnetized and the two strips attract each other. Figure 4.30 shows two reed relays. One is switched by an external magnetic field (a), and the other by a magnetic field created by a current in the windings around the relay (b).

A modern solution to a power relay is the solid-state switch, or TRIAC. It has no mechanical elements, as the current is controlled by transistors. The TRIAC starts conducting when the control signal is active. It has the additional advantage that currents are switched off only when they are zero. To say it in another way: even if the control voltage is interrupted, the currents continue until they reach a zero level. This way, it waits for the next zero-crossing level of the AC cycle to switch off the power. This greatly improves the lifetime of the device. Moreover, if a large current is abruptly interrupted, electromagnetic waves are emitted all over the spectrum because the Laplace transform of a step (Heaviside) function is proportional to $1/f$. The switching of devices can

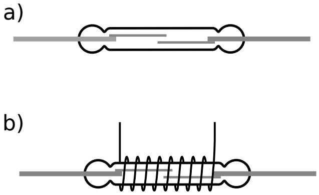


**Fig. 4.30**: Reed relay. A magnetic field magnetizes the blades inside the glass tube and they attract each other, thus closing the circuit. a) Switched by external field. b) Switched by current in the windings around the relay

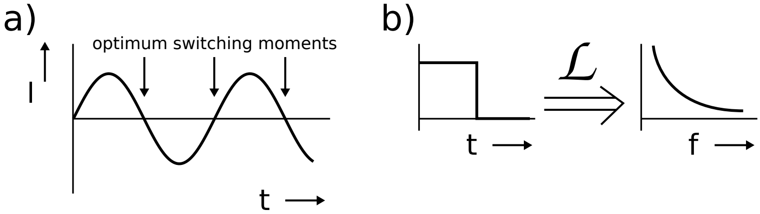


**Fig. 4.31**: Solid-state switch (TRIAC). a) When current is switched at zero-crossing points of a sine-wave current, problems of sparks caused by current inertia because of inductive effects are prevented. Also, the emission of electromagnetic waves (the Laplace transform of a step function is proportional to $1/f$, see b) is avoided

often be heard and seen on radios and TVs. A solid state switch avoids these problems, see Figure 4.31. In the chapter on informatics, an example is given how to use a TRIAC in a light dimmer.

## 4.8 Air & pressure sensors and actuators

### 4.8.1 Piezo sensors and actuators

The piezoelectric effect was discovered by the brothers (Pierre and Jacques) Curie. A sample that is squeezed (the Greek word for 'to squeeze' is 'πιεστε', pieste) by pressure, creates a voltage at it's ends. Necessary for this to happen, two criteria have to be met

- The sample should be a single crystal
- The crystal should not have inversion symmetry, in other words, the coordinate transformation $(x, y, z) \rightarrow (-x, -y, -z)$ should not map atoms onto similar atoms.

This is because the basic ingredient of the piezoelectric effect is a molecule that, when deformed, becomes polarized, see Figure 4.32b. A good example is silica. In its crystalline form (quartz) it is piezoelectric. In its amorphous form (glass,

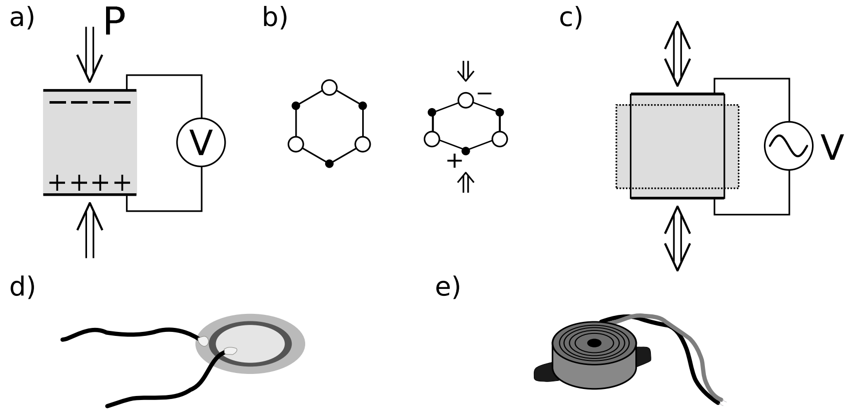


**Fig. 4.32**: Piezo: a) in a piezo sensor pressing a crystal without inversion symmetry, causes a voltage at its extremes. b) example of a non-inversion symmetry lattice unit of silica. c) Inverse piezoelectric effect; in a piezo actuator applying a voltage deforms the crystal. d) Typical piezo sensor. e) Typical piezo actuator

or sand) it is not. The combined effect of all molecules or unit cells of a crystal, when nicely aligned, contribute to a macroscopic electric effect, an externally measurable voltage.

The opposite effect also exists. When a voltage is applied to a crystal with non-inversion-symmetric structure such as quartz it deforms. The piezoelectric effect can thus be used for a pressure sensor, but also for a pressure actuator (loudspeakers). They are famous for being used in watches, and other places where low-grade (LoFi) sound is needed. For instance the beep when switching on a computer.

The piezo crystal has one more interesting property and that is that it has a very well defined mechanical oscillating resonance frequency. Together with the fact that these oscillations can be electrically excited through the piezoelectric effect makes them good elements for frequency standards. This has been described in the chapter on electronics (section of quartz-crystal oscillator, Section 2.7.4) and consists of placing the crystal in a feedback loop that maintains the phase at zero degrees. All modern equipment that needs a clock signal to function uses such crystal oscillators.

### 4.8.2 (Dynamic) pressure actuator: loudspeaker

The opposite of an air pressure sensor is a pressure actuator, $V \to P$. When the pressure is not static, but dynamic, the waves of pressure are called acoustic wave, or 'sound'. In other words, a dynamic air pressure actuator is a speaker. It normally works on the principle of magnetic forces, see Figure 4.33 for a schematic drawing of a speaker. A coil is placed inside a circular permanent magnet. When a current passes through the coil, it generates a magnetic field of

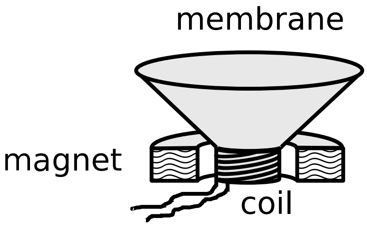


**Fig. 4.33**: Schematic drawing of a loudspeaker. A coil is connected to a lightweight conical membrane and is placed inside a permanent magnet (half of which is shown). Currents in the coil cause magnetic fields that interact with the permanent magnet field. A movement of the coil results and through the cone, this is translated into pressure waves

its own. These two magnetic fields interact and cause forces on each other; the permanent magnetic field causes forces on the electrons in movement (current in the coil), which are confined within the wire of the coil. The result is that the coil will move (since the magnet is fixed to the rest of the world), i.e., the coil. Connected to the coil is a lightweight membrane, normally in the form of a cone. This conveys the up-down movement to air-pressure waves.

Other, less used methods of producing sound are based on electrostatic attraction and repulsion, similar to a microphone, but inverse direction. In fact, any speaker can function as a microphone and any microphone can act as a speaker.

### 4.8.3 Pressure (difference) sensor; Pitot tube

Bernoulli's equation states that the sum of all forms of energy in a fluid in a closed system is constant. In a simplified form it is

$$\frac{v^2}{2} + gh + \frac{P}{\rho} = \text{constant}, \tag{4.50}$$

with $v$ the speed of the medium, $\rho$ its density, $h$ the height and $g$ gravitational constant. The energies are respectively kinetic energy, (gravitational) potential energy and (thermodynamic) internal energy.

How we should understand it is that, where one parameter changes, at least one of the others also has to change. We can make use of this in many ways. The most famous is the use in aviation. In fact, it is the reason why airplanes manage to stay in the air. Figure 4.34 shows a cross-section of a wing and the air-flow pattern around it. The path to travel from the front of the wing to the back of the wing is substantially longer above the wing compared to the path below the wing. This is one of the reasons why the air flows faster above the wing (though not the only reason; air has no physical obligation to 'meet' at the other end, nor does it actually do). And, because the changes in height $h$ are insignificant, Bernoulli's equation then tells us that the pressure above

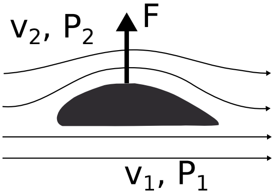


**Fig. 4.34**: Cross section of a wing and the flow pattern around it. Because of the shape of the wing, the air flowing above it has to take a longer path and thus has to travel faster $v_2 > v_1$. Bernoulli's equation (4.50) then tells us that the pressure above the wing must be lower than below the wing $P_2 < P_1$ and the wing is pushed upwards, $F$

the wing *must* be lower than below the wing. That means that the wing is pushed/sucked upwards.

For our subject of instrumentation it means that we can make use of the Bernoulli equation to measure one parameter by measuring the other. We can make a speed sensor by measuring the pressure (while staying on the same hight). Or we can make a altitude sensor by measuring pressure. It is obvious that this only works in 'closed' systems. For instance, if we take our altimeter for a field-trip into the mountains, pressure changes might as well be due to weather changes, and not necessarily only to altitude changes.

An example is the Pitot tube, see Figure 4.35. It consists of two chambers separated by a pressure (difference) sensor. The first chamber has an entrance at the front of the tube. Since the velocity there is zero, the pressure is the static outside pressure $P_0$. The other chamber has an opening on the side and is exposed to air with the velocity $v$ to determine. Assuming constant density $\rho$, the pressure difference between $\Delta P = P_0 - P_\mathrm{v}$ is given by the Bernoulli's equation (4.50),

$$v = \sqrt{\frac{2\Delta P}{\rho}}. \tag{4.51}$$

This technique is used a lot in aviation industry. Note however that knowledge of the density $\rho$ is needed to accurately convert the pressure to speed. In fact, the Pitot tube assumes

- Constant density $\rho$. In practice we know from bicycling that our speed causes a pressure gradient; we compress the air in front of us and create vacuum behind us, while on both sides the speed is zero.

- No viscosity and steady flow.

- Both entrances on same height (no gravitational effect).

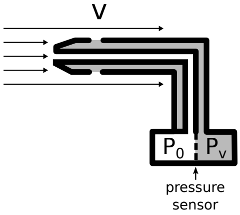


**Fig. 4.35**: Pitot tube. It has two chambers. The first chamber has an entrance at the front of the tube. The velocity there is zero and the pressure is the static outside pressure $P_0$. The other chamber (gray) has an opening on the side and is exposed to air with the velocity to determine. Assuming constant density ρ, the pressure difference between $P_0 - P_\mathrm{v}$ is given by the Bernoulli's equation (4.50)

### 4.8.4 Dynamic pressure sensor (microphone)

A very common pressure sensor is a microphone. More precisely, it is a differential (dynamic) pressure meter, in the sense that it measures pressure *waves.* It converts this into a voltage or a current. The working of a microphone is very simple.

A microphone consists of a simple capacitor made of two metal sheets. As explained in the chapter on physics, the capacitance of such parallel plates depends on the area and the distance between the plates,

$$C = \frac{\varepsilon_0 A}{d}. \tag{4.52}$$

In a microphone, one of the plates is a thin, flexible membrane that can easily move (bend) upon changes of pressure. In an electret microphone, one of the plates is permanently (electrostatically) charged. The other (membrane) is metallic and has the compensating mobile charge, see Figure 4.36a. The effect is that changes of the distance $d$ between the sheets causes a change in output voltage $V_\mathrm{o}$. Remember that charge, voltage and capacitance are directly related in the way

$$C = \frac{Q}{\Delta V}. \tag{4.53}$$

With the amount of charge being fixed, changes of capacitance are directly translated into changes of voltage:

$$V_\mathrm{o}(d(t)) = \frac{Q}{C(d(t))} = \frac{Qd(t)}{\varepsilon_0 A}. \tag{4.54}$$

An alternative scheme is to use a constant bias in combination with normal metal electrodes, one being thin and flexible and the other fixed. Any change in the membrane position then causes a change in stored charge,

$$Q(d(t)) = VC(d(t)) = \frac{\varepsilon_0 AV}{d}. \tag{4.55}$$

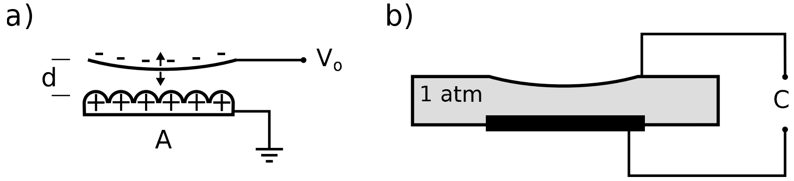


**Fig. 4.36**: a) Electret microphone consisting of capacitor with a fixed (electrostatic) charge on one electrode and a compensating mobile charge on the other. Changes in distance $d$ between electrodes translate into changes of output voltage $V_\text{o}$. b) A capacitor used in measuring static pressure difference between the ambient and a 1-atmosphere closed cell. In this case an electret capacitor will not do

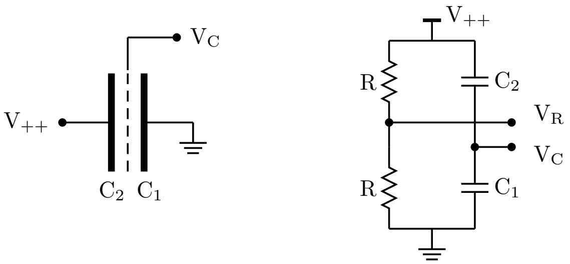


**Fig. 4.37**: Microphone part of a Wheatstone bridge. The thin membrane is placed between two fixed electrodes thus forming two capacitors, $C_1$ and $C_2$. The Wheatstone bridge gives an output voltage equal to $V_\text{o} = V_\text{R} - V_\text{C} = 0$ in rest and a differential voltage when the membrane moves

With charge implicitly depending on time, the current can be calculated

$$I(t) = \frac{\mathrm{d}Q}{\mathrm{d}t} = \frac{\mathrm{d}Q}{\mathrm{d}d} \cdot \frac{\mathrm{d}d}{\mathrm{d}t} \tag{4.56}$$

$$= -\frac{\varepsilon_0 AV}{d^2} \cdot \frac{\mathrm{d}d}{\mathrm{d}t}. \tag{4.57}$$

The signal coming from microphones is normally very weak and has to be amplified.

This second scheme can only measure changes in pressure as in acoustic waves (sound). With a static pressure the signal is zero. Technically speaking, it does not even measure pressure, but air displacement, with the membrane attempting to follow the movement of air. It can be placed in a Wheatstone bridge, with one leg made of resistances and the other leg of capacitances, see Figure 4.37.

The first scheme can in principle give directly a signal even for static pressures. However, since the output resistance of such a capacitor is infinite, no static voltage can be maintained (otherwise we would have a perpetual-motion machine, a source of voltage and energy without limit). Constant capacitance

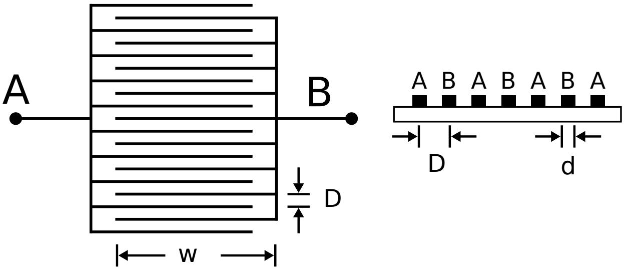


**Fig. 4.38**: Humidity sensor consisting of interdigitated electrodes on a plate. a) Top view, b) North-south cross-section

has to be measured dynamically, for instance converting it into a frequency as done by an oscillator of Chapter 2. Then the capacitor can be used to measure static pressure.

### 4.8.5 Humidity sensor

Humidity can be determined by measuring the conductivity of air. Because this conductivity is rather low (about $\sigma = 5 \times 10^{-15}$ S/m; the resistivity is $\rho = 2 \times 10^{16}$ $\Omega$m) large cross-section of air is needed to result in reasonable resistance values that are processable, in the order of kilo-ohms to mega-ohms. Normally a plate with interdigitated electrodes is used, a geometry that increases the effective total electrode width, see Figure 4.38. It is quite difficult to calculate the total resistance of such a system. However, we can make some observations. The resistance between two cylindrical wires of (large) length $w$ is given by (A.E. Kennelly, Proc. Am. Phys. Soc. vol. 48, p. 142, 1909)

$$R = w\frac{\rho}{2\pi}\cosh^{-1}\left(\frac{D}{d}\right), \tag{4.58}$$

with $D$ the distance between the centers of the wires, $d$ their diameter, $\rho$ the resistivity of the medium. The overall factor 2 comes from the fact that the wires are on an insulating plate and only half of space around it conducts.

To reduce the resistance, the total length $w$ can be increased by using more fingers. It also helps to reduce the distance $D$ between them. In any case, the resistance scales linearly with the resistivity $\rho$, as long as the conductance of the plate itself is negligible. Note also that the resistivity of air not only depends on humidity, but also on temperature, pressure and purity.

To drive the sensor, it is best to use AC signals, as DC signals can turn the sensor into an electro-chemical cell, as described before, especially when the humidity and the currents go up.

All-in-all, it is not easy to make a reliable air-humidity sensor. It is more common to use a mechanical relative humidity sensor, normally consisting of a string of horse hair twisted tightly and connected to a cantilever. Changes in

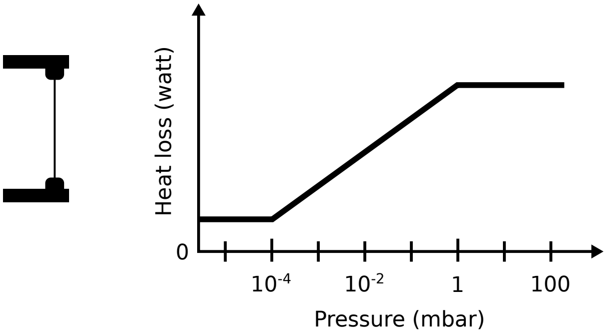


**Fig. 4.39**: A Pirani vacuum sensor. A wire is heated to a constant temperature. The power needed to keep the balance with dissipation of heat to the gas is proportional to the logarithm of gas density for pressures between approximately $10^{-4}$ mbar and 1 mbar

humidity will cause absorption and desorption of water, shrinking and expanding the string and moving the needle of the cantilever.

### 4.8.6 Vacuum sensor; Pirani

A capacitive pressure sensor works well for pressures close to ambient pressure. In some cases pressures close to vacuum are needed to be known, especially in scientific laboratory settings. In these cases, the linear difference pressure sensors are not adequate, since they are not capable to distinguish between, say 1 mbar and 0.1 mbar; both would measure a pressure difference with the ambient pressure of close to 1 atm and the error margin of the pressure is then easily much larger than the measured value itself. For vacuum measurements other types of phenomena have to be used. Two popular techniques are Pirani and Penning, that measure thermal conductivity and ionizability, respectively.

In a Pirani sensor, a thin wire is heated up to 100-150 °C and the heat loss is measured. This conductive heat loss is proportional to the logarithm of air pressure. The wire can be placed in a constant current configuration, with the hot wire as one resistance in a Wheatstone bridge. If the wire is in good vacuum, it has difficulty loosing heat and the wire heats up. This increases the resistance value and brings the Wheatstone bridge off balance. In modern equipment, no Wheatstone bridge is used, but the power supplied is regulated to keep the resistance at constant temperature (resistance). The power supplied is then equal to the power dissipated in the form of heat.

The Pirani sensor range is approximately from $10^{-4}$ mbar to 1 mbar. It is obvious that the heat loss also depends on the thermal properties of the gas measured. The sensor thus has to be calibrated for the type of gas measured. The gas in the sensor has to be static. Actually, the same device can also be used to measure air flow, as will be discussed later on.

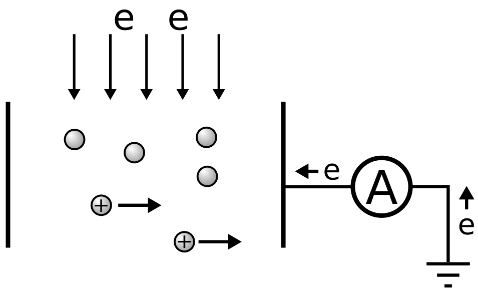


**Fig. 4.40**: A Penning vacuum sensor. High-kinetic-energy electrons ionize the gas molecules which drift towards the measuring electrode where they are neutralized by electrons. This current is directly proportional to the gas density.

### 4.8.7 Vacuum sensor; Penning

The Penning sensor, also known as 'cold cathode sensor', is based on the ionizability of a gas. The gas is partially ionized by electrons that are accelerated in a strong electric field. The positively charged ions of the gas are collected and neutralized by the measuring electrode, see Figure 4.40

The Penning sensor works with high voltages of about 2 kV. To increase the interaction time of the electrons with the gas molecules and to increase the efficiency of ionization, the chamber is placed in a strong (permanent) magnetic field. The electrons then have a spiral path because of the Lorentz forces. This increases their probability to find a molecule on its path. That makes that the sensor works until very low pressures, down to the $10^{-9}$ mbar range. The combination of high voltage and magnetic field makes the device rather expensive.

As for the Piranni sensor, the Penning sensor has to be calibrated for the type of gas being measured. That is because not all gas molecules are ionizable in the same way.

### 4.8.8 Gas-flow meters

The Pirani sensor concept can also be used to measure the *speed* of air. That is because the dissipation of heat not only depends on the pressure of the gas, but also on the amount of gas that passes by the wire. This is a phenomenon we all know, since we feel much colder if the wind is blowing compared to a quiet day; in meteorology this effect is called chill factor. The chill factor is a combination of temperature, wind force, humidity and pressure and determines how cold we *feel*, instead of just the simplistic value of temperature. The chill factor determines how much heat we lose when exposed to the air. One of the factors is pressure (as studied with the Pirani sensor). Another factor is air speed.

The amount (power) of heat transferred from a hot object to a cold medium

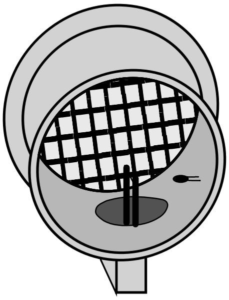


**Fig. 4.41**: A hot-wire air flow sensor as used in the air inflow of a turbo compressor of a car. The hot wire is connected from the bottom. On the side a temperature sensor (thermistor) is connected

is given by

$$P = \alpha A \Delta T, \tag{4.59}$$

with $\Delta T$ the temperature difference between the object and the medium, $A$ the area of contact, and $\alpha$ the heat-transfer coefficient that depends on the speed $v$ of the medium,

$$\alpha = \alpha_0 + \alpha_1 \sqrt{v}, \tag{4.60}$$

with $\alpha_0$ and $\alpha_1$ constants (still depending on pressure). A flow meter is a hot-wire contraption, with $A$ the surface of the wire, placed in a airflow of known temperature.

The power consumed by the object is $V^2/R$ or $I^2R$ (as by Ohms law). The resistance value depends on the temperature. The trick is thus to keep $R$ constant and measure the current or voltage needed to maintain this situation. This will then give the speed of the air via the above two equations if $\alpha_0$, $\alpha_1$, $A$ and $\Delta T$ are known. Especially the temperature is a variable and needs to be known. An example of a hot-wire air-flow sensor is shown in Figure 4.41, namely as the sensor of air inflow of a turbo compressor of a car engine. Note also the temperature sensor (thermistor) connected on the side. The board computer makes the calculation of the air flow based on these measurements.

Less advanced, but more common is the anemometer of cups connected on a vertical rotating shaft, as used in meteorological stations. They normally can be seen in combination with wind-direction meters, see Figure 4.42. Both use opto-couplers, or reed relays to transduce the signal into electronic format. The anemometer then converts the frequency of the pulses into a wind speed signal.

Another interesting flow meter from the point of view of physics is the mechanical floating-ball device. It consists of a light ball floating in a vertical airstream inside a conical tube, see Figure 4.43. Because the ball is floating stationary in the airflow, the forces of friction are equal to the force of gravity. The latter is

$$F_{\downarrow} = mg, \tag{4.61}$$

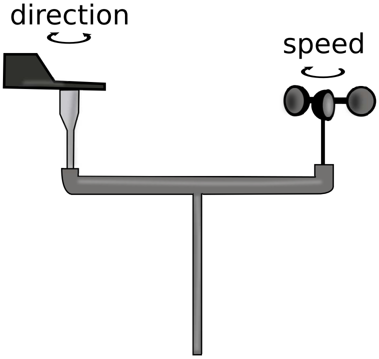


**Fig. 4.42**: Combination of anemometer and wind-direction meter.

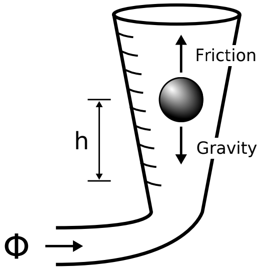


**Fig. 4.43**: A flow meter based on the principle of friction being balanced by gravity

with $m$ the mass of the ball, and $g$ the gravitational constant 9.81 m/s$^2$. The friction is given by Stokes equation that is good for small spheres:

$$F_{\uparrow} = 6\pi r \eta v, \tag{4.62}$$

with $r$ the radius of the sphere, $\eta$ the viscosity of the gas, and $v$ the velocity of the sphere relative to the medium. At equilibrium they are equal,

$$v = \frac{mg}{6\pi r \eta}. \tag{4.63}$$

The speed of the ball at the place of the ball is given by the gas debit $\Phi$ (unit: L/s) divided by the area $A$, the cross-section of the tube not covered by the ball, $A = \pi(R^2 - r^2)$, with $R$ the inner radius of the tube. Finally, this radius depends linearly on the height $h$, $R(h) = r + \alpha h$, where at $h = 0$ the ball is marginally blocking the opening, $R(0) = r$. Ignoring the term $\alpha^2 h^2$, the flux of gas is given by

$$\Phi = \frac{mg\alpha}{3\eta} h. \tag{4.64}$$

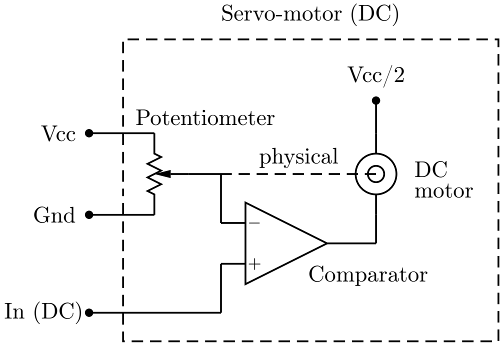


**Fig. 4.44**: Simplified effective (internal) circuit of a servo motor. It constantly compares the position of the motor via a potentiometer with the desired value (In) and supplies the necessary voltage to the DC motor. There exists a physical connection between the axis of the motor and the axis of the potentiometer

## 4.9 Motors as actuators

### 4.9.1 Positional servo motor

A servo motor is very simple. It combines an actuator and sensor in one, using feedback to make the actual angle of the axis equal to the desired angle. The basic ingredient of a (positional) servo motor is a normal DC motor that actuates changes of the angle. Connected to the axis of the motor is a potentiometer that supplies feedback to the angle. A controller reads the actual angle and operates the DC motor. A simplified effective internal circuit of a servo motor is given in Fig. 4.44. In this scheme, the only thing we have to do is feed a voltage between the 0 and $V_{cc}$ to the input pin of the servo motor.

More advanced servo motors – nearly all modern ones – use pulse-width modulation techniques to pass the information of the desired angle to the servo-motor controller. The *length* of the pulse supplied to the controller contains the information of the desired angle. Typically, a pulse width of 544 µs implies 0° and a pulse of 2400 µs length codes an angle of 180°. If we supply a pulse with 1 microsecond resolution, this translates into about 0.1 degree angular (digital) resolution. The accuracy and reproducibility can be much lower than that, though. We have to read the datasheet of the motor to find out about these other parameters and things as torque (the moment of the engine) and speed (how fast it can move from one angle to the other). The effective equivalent circuit of such a servo motor is given in Fig. 4.45

### 4.9.2 Velocity servo motor

A second type of servo motor is one in which there is a feedback loop of the *speed* of the motor. This we can call velocity servo motor. It means measuring the velocity of the motor accurately, probably by measuring pulses in an angular

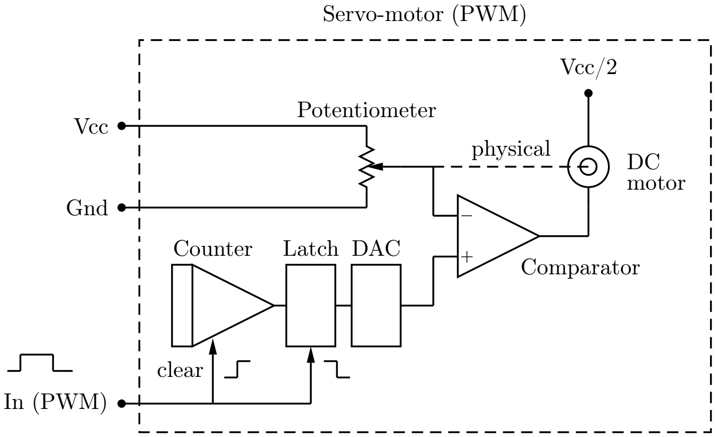


**Fig. 4.45**: Simplified effective (internal) circuit of a servo motor that uses a pulse-width modulation technique. A free-running counter is reset to zero by the start of an incoming pulse. When the pulse ends, the value of the counter is copied to a latch (memory) and fed into a digital-to-analog converter. This represents the desired angle as an analog voltage. This voltage is compared to the actual angle and the motor is rotated clockwise or anticlockwise depending on the sign of the comparator signal

speed sensor (see Section 4.5.4), and feeding it into a control circuit. An example of such a servo motor is given in Figure 4.45. This particular type receives the information of the desired speed by a pulse-width modulation (PWM) signal.

### 4.9.3 Stepper motor

A stepper motor is a motor that can make small *exact* steps in a rotation. For each pulse supplied to the motor, the axis advances one step, typically a few degrees. This then has two applications, first of all, by supplying a certain *number* of pulses relative to a known, calibrated, position the angle of the motor axis is known. Another uses of the stepper motor is that by supplying pulses with a certain determined *frequency*, the motor will rotate at a well defined speed (RPM). A stepper motor is thus an actuator that transduces either a number into an angle, or a frequency into an angular speed.

It is not difficult to understand how this is achieved. The central core of the motor, the 'axis' or 'rotor', consists of a permanent magnet. Around it, the fixed part, 'stator' are placed electro-magnets, basically coils. When current is passing through the coils, a magnetic field is created and the permanent magnet connected to the rotor will align itself to this field, see Figure 4.46. For a positive current in coil A, the north pole of the rotor aligns itself closest to this coil and then stays there, even if the current stops. If we now pulse a current in coil B, the north pole is attracted to that coil. A pulse sequence ABCD will cause a complete revolution. The motor and object connected to it through gears,

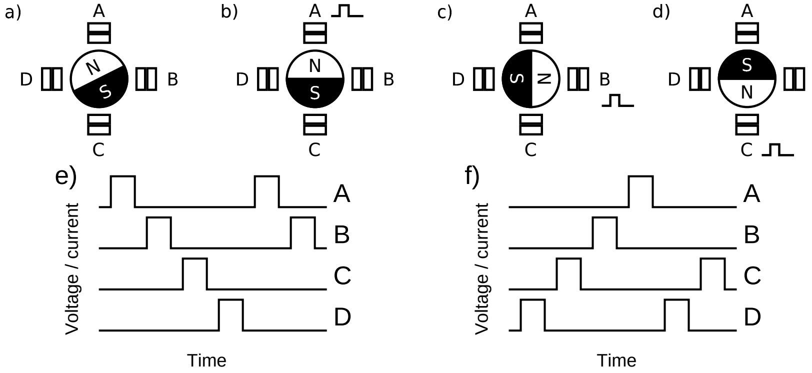


**Fig. 4.46**: Working of a stepper motor explained for a two-pole-four-coil stepper motor. The axis starts in a random position (a). By applying a pulse to coil (electromagnet) A, the permanent magnet on the axis aligns itself to that coil (b). Subsequent pulses to coils B (c) and C (d) further rotate the axis. A pulse sequence ABCD will cause a full clockwise rotation. The repetition rate of the pulse sequence the translates in an angular velocity (RPM). The width of the pulses is irrelevant. The pulse sequence in (e) causes clockwise rotation and the one in (f) causes a anti-clockwise rotation

etc., can be placed in an exact relative position by an exact number of pulses. Note however that we do not have knowledge about the absolute position of the motor and object, we only know the relative position of the object. This is a difference – the main difference – with a servo motor.

The repetition rate of the pulse sequence is translated in an angular velocity of the motor. Moreover, if we want the motor to rotate in the other direction, we can supply an inverse pulse sequence, DCBADCBA, etc.

Commercial stepper motors have more coils and/or magnet poles and can achieve a much higher resolution than the 90 degrees shown above. Typically, a stepper motor has 96 steps per turn (which represents a resolution of $3.75^o$). We can also make half-steps of the motor if we supply pulses to two coils simultaneously. In the example of Fig. 4.46, supplying a pulse to coil A and coil B at the same time will position the magnet halfway between A and B, at an angle of $45^o$.

Note that the lengths of the pulses are not relevant. As long as they are long enough to make the motor rotate to the next position. On the other hand, if the pulse is too long, the current will be high in the idle state and we run the risk of wasting energy and overheat the motor.

To drive a stepper motor we cannot connect it directly to transistors or other signal elements such as microcontrollers (see for instance the Arduino in Chapter 5). It is not even a good idea to connect it to simple power transistors, because they seriously run the risk of burning out. The problem is that a (stepper) is not only a device that converts electric energy into kinetic energy (rotation),

**Table 4.V**: Comparison of typical stepper and servo motors

| | Stepper motor | Servo motor |
|---|---|---|
| Angle | Relative | Absolute |
| Speed | Very well determined | Unpredictable (positional)<br>Well determined (velocity) |
| Resolution | Small | High |
| Accuracy | High | Medium |
| Angular range | Infinite | Limited (positional)<br>Infinite (velocity) |
| Pulse width | Irrelevant | Coding |

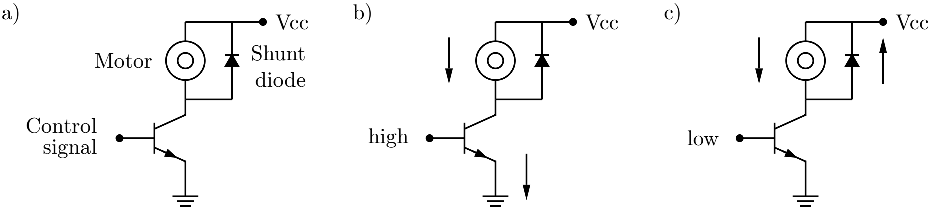


**Fig. 4.47**: a) Shunt diodes to protect stepper motors, engines, circular filamentary lightbulbs and other inductive elements. b) In normal operation (transistor is short circuit) when element is actuated. c) When switching off the element (transistor open circuit), the continuing current is dissipated through the shunt diode

but also the reverse. When the motor rotates, for instance by inertia after an induced rotation, a large current is induced in the coils. This current has nowhere to go; the transistor of the motor-driver is closed. Charge accumulates in the junction of the transistor and a large build-up of electric field occurs, a field that potentially may burn the transistor. Thus, we have to allow for a way for the induced current to dissipate. In fact, this effect occurs for all inductive elements (stepper motors, engines, and even lightbulbs with coiled filaments). The current, when in operation causes a build-up of magnetic field. When the driving force stops, the current continues, because the energy associated to the magnetic field (proportional to the space-integral of the magnetic field squared) cannot dissipate instantaneously and the current therefore *has* to continue until this energy is dissipated in the form of heat. We have to make sure this heat is dissipated far away from fragile elements. This is conventionally done by shunt diodes, see Figure 4.47.

Stepper motors exist in many types and sizes. The most important distinctions are

Unipolar ↔ Bipolar
Permanent magnet ↔ Reluctance magnet

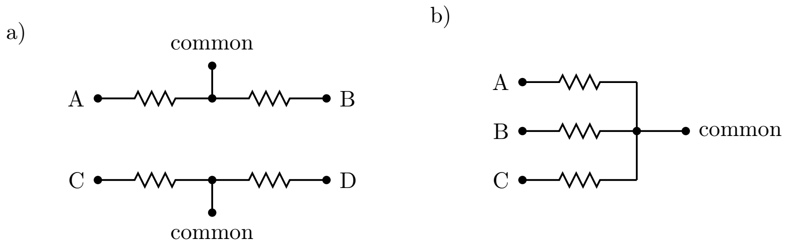


**Fig. 4.48**: Examples of stepper-motor coil configurations: a) unipolar 4-phase (6 wires),and b) bipolar 3-phase (4 wires)

The ones with permanent magnet can be recognized by the friction – the steps – that are felt when rotating it manually with the power off. This allows you even a way to determine the angular resolution without connecting it. Reluctance-magnet stepper motors, on the other hand, run smoothly with the power off and there is no way in finding how many steps for a full turn manually. The working principle is basically the same, although in a reluctance magnet, both negative and positive pulses will attract the closest 'pole' (protrusion), rather than the closest magnetic pole of correct sign. This is similar to the effect that a magnet is always attracting ferromagnetic material, and never repelling it. The south pole of a magnet will attract the north pole of another magnet (and repel another south pole) or any ferromagnetic material.

Unipolar vs. bipolar refers to the way the coils are connected. If they operate with both polarities of current or not. Remember, a positive current will cause a north-south magnetic field and a negative current a south-north magnetic field. Examples are given in Figure 4.48. We can easily find out what the internal configuration of the coils is with a multimeter. Measuring all combinations of wires will place them in two groups, one groups has $R$ resistance ($R$ typically some ohms) and the other group has $2R$ resistance. After that, it is a matter of trial-and-error to find the correct pulse sequence in order to make it work.

Some final observations. If you are a hobbyist and have taken a stepper motor from destroyed equipment, such as a printer or a harddisk, you now have an unknown motor in your hands. It will be very difficult to find the datasheet for it. With the above method you should be able to find how how to operate it. Here are some more hints to help you:

- A stepper motor with 5 wires is almost certainly 4-phase unipolar.

- A stepper motor with 6 wires is probably also 4 phase unipolar, but with to common power wires. They may both be the same color

- A stepper motor with only 4 wires is most likely bipolar.

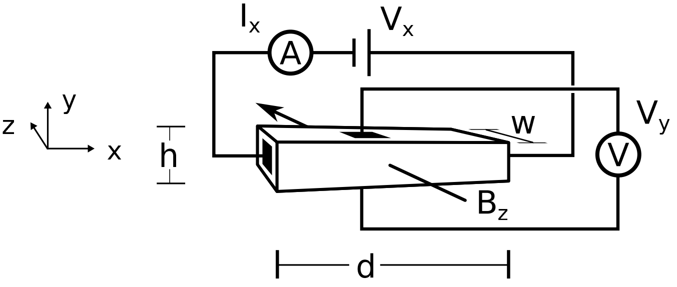


**Fig. 4.49**: Hall sensor for measuring magnetic field

# 4.10 Magnetic field sensors

Magnetic field can be measured in various ways. If just the presence of a (sufficient) field has to be detected, a simple coil can be used (Fig. 4.18) if the field is changing, or a reed relay (Fig. 4.30) if the field is static. For *quantifying* the field strength, basically two sensors exists, the Hall probe and NMR probe, discussed here.

## 4.10.1 Hall probe

The Hall probe to measure magnetic fields is based on the Hall effect discovered by Edwin Hall. The Hall effect is based on the Lorentz force, discovered by the Dutch scientist Hendrik Antoon Lorentz. Charges moving in a magnetic field feel a force that is perpendicular and proportional to the field and the magnetic field and has a magnitude equal to

$$F_{\mathrm{B}} = qB_{\mathrm{z}}v_{\mathrm{x}}. \tag{4.65}$$

In Figure 4.49, a positive charge moving along the x direction in a field along z, will be bend towards y. Charges build up there and create a space charge that causes a compensating electric field in this direction (y),

$$F_{\mathrm{E}} = qE_{\mathrm{y}} = q\frac{V_{\mathrm{y}}}{h}. \tag{4.66}$$

At steady state, there is no net current in this direction and the two forces $F_{\mathrm{B}}$ and $F_{\mathrm{E}}$ cancel. Further, the current $I_{\mathrm{x}}$ is charge density $n$ times (average) speed and multiplied by device cross section $A = wh$,

$$I_{\mathrm{x}} = qnvwh, \tag{4.67}$$

and thus

$$B_{\mathrm{z}} = \frac{nqV_{\mathrm{y}}w}{I_{\mathrm{x}}}. \tag{4.68}$$

Alternatively, if we don't know the density of charges, we can use the relation between velocity and electric field, called mobility (see Chapter 3), $v_\mathrm{x} = \mu_\mathrm{n} E_\mathrm{x}$, to get

$$B_\mathrm{z} = \frac{V_\mathrm{y} d}{\mu_\mathrm{n} V_\mathrm{x} h}. \tag{4.69}$$

### 4.10.2 NMR (nuclear magnetic resonance) probe

Another way of measuring the magnetic field is with a nuclear magnetic resonance (NMR) probe which is based on the quantization of magnetic energy levels of particles. In the chapter of physics this was already briefly mentioned. Apart from charge, which is quantized, always being a multiple of $q$, an electron also has a magnetic moment – called 'spin' S – that, when placed in a magnetic field, interacts with it and can have only two orientations resulting in two distinct energy levels. Likewise, every nucleon – proton or neutron – has a magnetic moment, nuclear spin, I. In an NMR probe, hydrogen is used. Because it has a single nucleon, it has a nuclear spin equal to $I = 1/2$. The two distinct energy levels corresponding to the two possibilities for the quantum number $m_\mathrm{n} = \pm I$, are given by

$$U_\mathrm{n} = m_\mathrm{n} g_\mathrm{n} \mu_\mathrm{N} B, \tag{4.70}$$

with $g_\mathrm{n}\mu_\mathrm{N}$ the natural constant quantifying the interaction of magnetic field with nuclear spin (the nuclear g-factor $g_\mathrm{n} = 5.585$ and the nuclear magneton $\mu_\mathrm{N} = 5.05079 \times 10^{-27}$ J/T). The energy of the two possibilities depend linearly on the magnetic field, see Figure 4.50. We can bring these hydrogen protons into resonance by radiating them with photons. Resonance will occur when the photon energy equals the difference between the energy levels, thus when

$$h\nu = g_\mathrm{n} \mu_\mathrm{N} B, \tag{4.71}$$

with $h$ Planck's constant, and $\nu$ the frequency of the photon. To give an example, for hydrogen atoms placed in a 1-tesla field, the resonance frequency is 42.576 MHz. In an NMR probe the system is kept in resonance by feedback and the frequency is then directly a measure for the magnetic field,

$$B = \frac{h}{g_\mathrm{n}\mu_\mathrm{N}} \nu. \tag{4.72}$$

An NMR is highly sensitive and accurate. The factors in the equation above are natural constants that do not depend on temperature, or pressure, or any other possible interfering effect. They can only be influenced by interactions with other magnetic fields, for instance magnetic moments of neighboring atoms, called 'chemical shift'. That is why the hydrogen in an NMR probe is normally diluted. On the other hand, in NMR imaging this chemical shift is used to determine and map the environment of the hydrogen atoms.

The same technique can also be used with the electron spin, but because the mass of electrons is about three orders of magnitude smaller than that of protons, and since the magnetic moment is reciprocal with the mass, the resonance

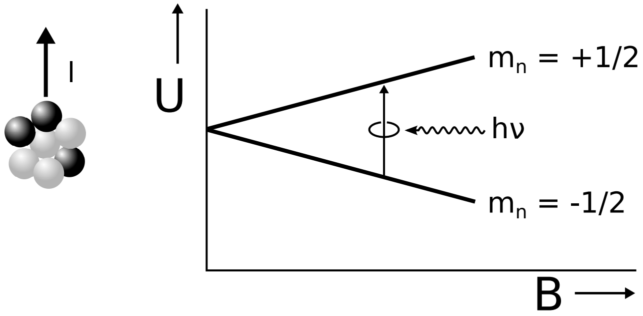


**Fig. 4.50**: Nuclear magnetic resonance (NMR)

frequency is about three orders of magnitude higher than that of protons, typically 28 GHz at a 1-tesla field. EPR is much less used for determining magnetic field strengths because the electrons interact much more with the environment and because electronics at the GHz frequency range is much more difficult to make than electronics at the MHz range, the latter being quite trivial.

## 4.11 Scientific instrumentation

Scientific instrumentation differs from industrial instrumentation in that it is normally state-of the art and highly sophisticated. Money normally is not a problem. The limitation of industrial equipment as needing to be robust is also not a big issue in scientific instruments. A good example is a quartz-crystal microbalance (QCM), so named because it can measure minute quantities of mass with a quartz crystal. We will now see just how sensitive the QCM is.

The basic component of a QCM is a quartz crystal as used in a crystal oscillator. The difference between a quartz crystal for frequency applications and a normal piezo sensor is that the former is of higher quality with respect to the (uniform) thickness which gives it a better defined resonance frequency (called Q-factor, $Q = f/\Delta f$). The mechanical resonance frequency of a crystal with thickness $d$ when oscillating in the sheer modus (waves in the plane of the surface) is given by

$$f = \frac{\sqrt{\mu\rho}}{2M}, \tag{4.73}$$

with $\mu$ the shear modulus of quartz ($2.947 \times 10^{11}$ g cm$^{-1}$ s$^{-2}$), $\rho$ the density of quartz (2.648 g cm$^{-3}$) and $M$ the mass per unit area of the crystal (g/cm$^2$), $M = m/A$, with $m$ the mass and $A$ the area. The mass per area is also given as $M = \rho d$ with $d$ the crystal thickness, and a crystal made of quartz thus has a frequency $f =$ 3.336 kHz m. A 5 MHz crystal has a thickness of about $d =$ 0.67 mm.

The mass that is added to the surface changes the value of $M$, and thus its resonance frequency. Crystals can be bought with calibrated nominal resonance

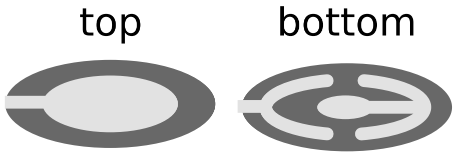


**Fig. 4.51**: A QCM crystal with electrodes on both sides (the disks on each side; the active area is the overlap between the disks). The contacting with the electrodes is done on one side, somewhere on the two semi-circles. A connection to the electrode on the other side is made on the edge

frequencies at 5 MHz, 10 MHz, etc. From the above equation it is easily seen that the quartz crystal has a gauge factor equal to (minus) unity,

$$\frac{\mathrm{d}f/f}{\mathrm{d}M/M} = -1. \tag{4.74}$$

Another way of expressing this is defining the mass sensitivity, in first order,

$$S \equiv \frac{\mathrm{d}f}{\mathrm{d}m} = -\frac{2f^2}{A\sqrt{\mu\rho}}. \tag{4.75}$$

As an example, a 5 MHz quartz crystal of active area with 1 cm diameter has a sensitivity of 72 GHz/kg. A change of resonance frequency of 1 Hz is induced by depositing a mass of 13.8 nanogram. Breathing on the sensor easily moves it some hundreds of hertz. Calculated in another way: if 5 MHz is equal to 0.67 mm, 1 Hz is equal to 1.3 Å, about the thickness of one monolayer.

The electronics of a QCM are the same as the electronics of a normal quartz oscillator crystal, namely the placement of the crystal in a oscillator circuit with feedback to keep it at the resonance frequency. More exact, the circuit keeps the crystal at the frequency where the phase-shift of the loop is zero (see Section 2.7.4 of Chapter 2). That might not be exactly the same thing in practice. The contacting of the crystal is slightly different, since one surface of the crystal is reserved for the measured specimen. The contacting thus has to be done on one side only, with the connection to the other side done through a via on the edge of the crystal, see Figure 4.51.

## 4.12 Lab projects

- Take a standard diode. Design a way to use it as a temperature sensor (what would be more linear? Constant voltage - measuring current, or constant current - measuring voltage?). Determine the theoretical curve of output vs. temperature. Place it in a water boiler (in a protective environment like a plastic bag) together with a normal thermometer (one that can stand 100 °C) and determine the experimental curve, see Fig. 4.52. Now discuss the aspects about the sensor system as described in

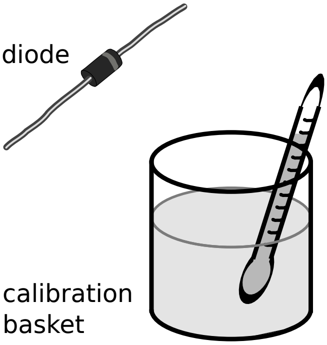


**Fig. 4.52**: A diode as a temperature sensor, calibrated with a boiler and a conventional thermometer

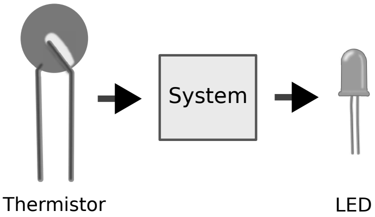


**Fig. 4.53**: A thermistor as temperature sensor used with a warning LED

Chapter 1, such as: linearity, range, sensitivity, etc. Overall, is this a good sensor?

- Based on a thermistor temperature sensor, design and implement a system that switches on a warning LED when the temperature rises above 40 °C, see Fig. 4.53. (See the comparator circuits in Chapter 2).

- Based on an LM35 temperature sensor, design and implement a system that switches on a power element (such as a computer cooling fan) when the temperature rises above 40 °C. Once switched on, the fan is only to be switched off when the temperature drops below 30 °C. (See the hysteresis circuits in Chapter 2). For the fan, a power transistor has to be used configured not as a signal amplifier, but as a switch.

- Glue two extensometers (strain gauges) (See Section 4.5.7) on two sides of a flexible bar. Use them as one leg in a Wheatstone bridge, with the other leg composed of normal resistances, and embed this in an electrical circuit as shown in Figure 4.54. Design the circuit in such a way that the total differential gain is approximately 100. What would be the expected output for an excursion of 10 cm?
  Make the bar oscillate and watch the output of the circuit on an oscilloscope. Is the amplitude comparable to what you expected? Figure 4.55

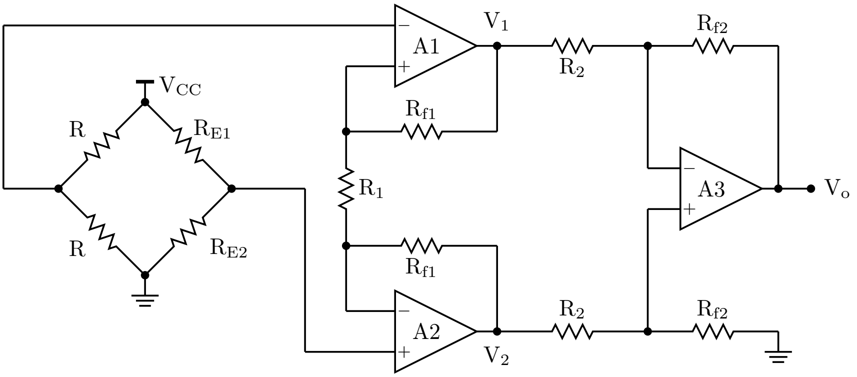


**Fig. 4.54**: Circuit for two strain gauge sensors placed in a Wheatstone bridge and differential amplifier

might help in analyzing the situation.

## 4.13 Exercises

1. A strip resistor of length $L$, width $W$, height $h$, made of a material with resistivity $\rho$ (see Figure 4.56) has a resistance value of

$$R = \frac{\rho L}{Wh}. \tag{4.76}$$

The gauge factor is defined as

$$k \equiv \frac{dR/R}{dL/L}. \tag{4.77}$$

a) Show that the gauge factor $k$ for a strip resistor with all parameters $W$, $h$ and $\rho$ functions of $L$, is given by

$$k = 1 + 2\nu + \frac{\mathrm{d}\rho/\rho}{\epsilon_\mathrm{L}}, \tag{4.78}$$

with $\epsilon_\mathrm{L}$ the strain along L ($\epsilon_\mathrm{L} = dL/L$), and $\nu$ Poisson's ratio defined as

$$\nu \equiv -\frac{\mathrm{d}W/W}{\mathrm{d}L/L} = -\frac{\mathrm{d}h/h}{\mathrm{d}L/L}. \tag{4.79}$$

b) What is the gauge factor $k$ for a material that has the property that the volume does not change when it's length is altered (assume constant $\rho$),

$$\frac{\mathrm{d}V}{\mathrm{d}L} = 0. \tag{4.80}$$

c) A strip resistor of 1 k$\Omega$ is extended 1% in length. What is the new resistance value? (Assume constant value and no piezoelectric effect.

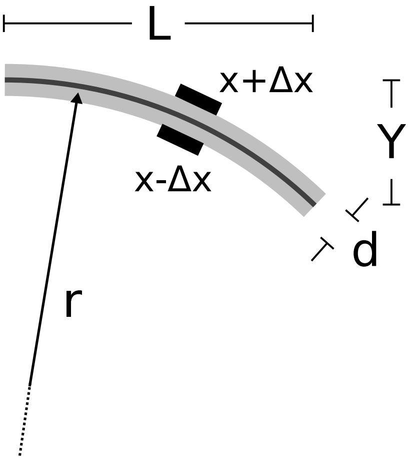


**Fig. 4.55**: Two strain gauges (extensometers) placed on sides of a metal plate. When the plate bends down, the top sensor extends and the bottom one contracts by an amount $\Delta x$ which is a function of $Y$ and the other parameters ($x$, $L$ and $d$). This can be calculated if we make use of the bending radius $r$

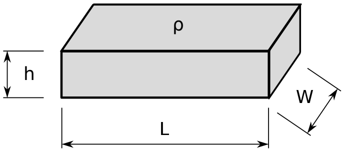


**Fig. 4.56**: Bar resistance. Width $W$, height $h$, length $L$ and resistivity ρ. The resistance is $R = \rho L/Wh$. (Exercise 1)

2. To weigh objects, we take a capacitor and let it be deformed when the weight is placed on top or hanging from it. Our capacitor consists of metal (thin) foil glued on two sides of a bar of material with dimensions $L$, $W$ and $h$. For this calculation we will extend the direction $h$. The capacitance of such a parallel-plates capacitor is given by $C = \varepsilon A/h$, with ε the permittivity of the material, $A$ the area of the plates and $h$ the distance between them.

   a) Derive an expression for the gauge factor $k_\mathrm{C}$ of the capacitor

$$k_\mathrm{C} \equiv \frac{\mathrm{d}C/C}{\mathrm{d}h/h}, \tag{4.81}$$

   using the definition of Poisson's ratio, ν

$$\nu \equiv -\frac{\mathrm{d}L/L}{\mathrm{d}h/h} = -\frac{\mathrm{d}W/W}{\mathrm{d}h/h} \tag{4.82}$$

   b) What is the gauge factor for a material that has constant volume?

c) A constant-volume capacitor of 1 μF is extended 1% in length. What is the new value of the capacitance?

Young's Modulus $E$ describes the deformation of a material when a force is acting upon it. It is the ratio between stress, pressure $P$ (Pa), and relative deformation, strain $\epsilon_{\mathrm{L}}$ (unitless),

$$E = P/\epsilon_{\mathrm{L}}, \tag{4.83}$$

$$\epsilon_{\mathrm{L}} \equiv \mathrm{d}h/h. \tag{4.84}$$

Note that the pressure is the force per area, $P = F/WL$ in our case. The material we use for the capacitor is rubber which has the following properties:

| | Rubber |
|---|---|
| Young's Modulus: | $E = 0.05$ GPa |
| Permittivity: | $\varepsilon = 7\varepsilon_0$ ($\varepsilon_0 = 8.85418 \times 10^{-12}$ F/m) |
| Poisson's ratio: | $\nu = 0.50$ |

The dimensions of our capacitor are:

$$\begin{aligned} L &= 10\ \mathrm{cm} \\ W &= 1\ \mathrm{cm} \\ h &= 10\ \mu\mathrm{m} \end{aligned}$$

d) What is the nominal capacitance? (unit: F)
e) What is the sensitivity of the sensor? (unit: F/N)
The weight $F$ and mass $m$ of an object are related to $F = mg$, with $g = 9.81\ \mathrm{m/s^2}$.
f) What is the sensitivity of the sensor? (unit: F/kg)
g) When measured with a multimeter with 4 decimal places at any scale, measuring directly the capacitance, what would be the resolution of the system? (unit: kg)

3. Take a diode and measure its sensitivity when used as a temperature sensor. To emulate a current source, place a 1 MΩ resistance in series with a voltage source. As long as the resistance of the diode is smaller than this shunt resistance, but larger than the input resistance of the voltmeter, the measurements can be made. You will probably find a negative temperature coefficient. What happens if you use double-current technique as explained in the text?

4. A Doppler gun is used by the police has a source frequency of 3 GHz. In a system with a counter that periodically measures every 1 second, what is the resolution of the system in terms of speed (km/h)?

5. A pH sensor is based on the chemical potential (unit: volt) of an acid reaction which, in turn, is based on the concentration (unit: mol per liter) of hydrogen ions, $[H^+]$. More precisely, $\mathrm{pH} = -\log_{10}([H^+])$, Eq. (4.35), and the resulting voltage is $V_{\mathrm{sensor}} = 59.13\ \mathrm{mV} \times (7 - \mathrm{pH})$, Eq. (4.43). The sensor is placed in a 3 liter solution with a pH of 5. To measure voltages we have a multimeter with 4 digits resolution (for example, at a scale of 20 V the resolution is 0.01 V) and a minimum scale of 20 mV.
   a) Design an electronic circuit to prepare the signal to work with the highest possible resolution in pH.
   b) What is this resolution in terms of number of hydrogen ions ($\Delta N_{\mathrm{H}+}$) of the system of a)?
   c) A pH sensor needs a high-impedance amplifier to work properly. Explain why.

6. A quartz crystal is a disk with a diameter of 2.5 cm and a calibrated resonance frequency of $f_0 = 5.000000$ MHz when nothing is deposited on top. The sensor is connected to an oscillator circuit that maintains the oscillations at its resonance value, as shown in the figure above. The figure shows sensor disks, a sample holder (the 'pipe') and the resonance circuit with a frequency counter.
   a) If the frequency counter gives a value once per second, what is the resolution of the system in terms of mass?
   To mechanically excite the crystal, the piezoelectric effect is used.
   b) Explain what is the piezoelectric effect.

7. In this exercise we will study luminance, $L$. Luminance is the intensity of light per area (unit cd/m$^2$, sometimes called 'nits') and represents the common idea of 'intensity'. Lower luminance means it is darker in the room. A certain family of sensors, with various sizes, has as sensitivity of 10 mA per cd/m$^2$ per m$^2$. For example, a sensor of 1 cm$^2$ exposed to ambient light of 200 cd/m$^2$ generates a current of 200 μA.
   a) Design a circuit (powered by 10 V) that translates the luminance to a voltage to be measured by a multimeter, using a sensor of 2 cm$^2$ area. The circuit has to be prepared for luminances between 20 and 800 cd/m2.
   b) Connected to a multimeter of 4 decimal cases and a minimal scale of 2 mV, what would be the final uncertainty, $\Delta L$, of the system?

8. The difference between a full battery and an empty one is that, while the open-circuit voltage is more or less equal, an empty battery does not manage to maintain that voltage supplying a current. A full battery supplies about 100 mW and and empty one barely 1 mW. Design a circuit that tests batteries. The circuit has to light an LED when the battery is empty. (The power for the circuit, obviously, does not come from the battery being tested, but from a separate power supply).

## 4.14 Answers

1 Given the fact that the resistance is $R = \rho L/Wh$, the full derivative, when $W$, $h$ and $\rho$ are functions of $L$, is given by

$$\frac{\mathrm{d}R}{\mathrm{d}L} = \frac{\rho}{Wh} \cdot \frac{\mathrm{d}L}{\mathrm{d}L} - \frac{\rho L}{W^2 h} \cdot \frac{\mathrm{d}W}{\mathrm{d}L} - \frac{\rho L}{Wh^2} \cdot \frac{\mathrm{d}h}{\mathrm{d}L} + \frac{L}{Wh} \cdot \frac{\mathrm{d}\rho}{\mathrm{d}L}. \tag{4.85}$$

In this case

$$\begin{aligned} k &\equiv \frac{\mathrm{d}R}{\mathrm{d}L} \cdot \frac{L}{R} = \frac{\mathrm{d}R}{\mathrm{d}L} \cdot \frac{Wh}{\rho} \\ &= 1 - \frac{L}{W} \cdot \frac{\mathrm{d}W}{\mathrm{d}L} - \frac{L}{h} \cdot \frac{\mathrm{d}h}{\mathrm{d}L} + \frac{L}{\rho} \cdot \frac{\mathrm{d}\rho}{\mathrm{d}L} \\ &= 1 - \frac{\mathrm{d}W/W}{\mathrm{d}L/L} - \frac{\mathrm{d}h/h}{\mathrm{d}L/L} + \frac{\mathrm{d}\rho/\rho}{\mathrm{d}L/L} \\ &= 1 + 2\nu + \frac{\mathrm{d}\rho/\rho}{\mathrm{d}L/L}. \end{aligned} \tag{4.86}$$

For constant volume,

$$\begin{aligned} V &= LWh = \text{constant}, \\ \frac{\mathrm{d}V}{\mathrm{d}L} &= Wh + Lh \cdot \frac{\mathrm{d}W}{\mathrm{d}L} + LW \cdot \frac{\mathrm{d}h}{\mathrm{d}L} = 0 \\ & \quad 1 + \frac{L}{W} \cdot \frac{\mathrm{d}W}{\mathrm{d}L} + \frac{L}{h} \cdot \frac{\mathrm{d}h}{\mathrm{d}L} = 0, \\ & \quad 1 - \nu - \nu = 0, \\ \nu &= 0.5, \\ k &= 1 + 2\nu = 2. \end{aligned} \tag{4.87}$$

For a resistance of 1 kΩ:

$$\begin{aligned} \Delta R &= \Delta L \cdot \frac{\mathrm{d}R}{\mathrm{d}L} \\ &= \frac{R}{L} \cdot \Delta L \cdot \left( \frac{\mathrm{d}R}{\mathrm{d}L} \cdot \frac{L}{R} \right) \\ &= R \cdot \frac{\Delta L}{L} \cdot k \\ &= 1\ \mathrm{k\Omega} \cdot 1\% \cdot 2 \\ &= 20\ \Omega. \end{aligned} \tag{4.88}$$

2 a) Given the fact that the capacitance is $C = \varepsilon LW/h$, the full derivative, when $W$, $h$ and $\varepsilon$ are functions of $L$, is given by

$$\frac{\mathrm{d}C}{\mathrm{d}h} = -\frac{\varepsilon WL}{h^2} \cdot \frac{\mathrm{d}h}{\mathrm{d}h} + \frac{\varepsilon L}{h} \cdot \frac{\mathrm{d}W}{\mathrm{d}h} + \frac{\varepsilon W}{h} \cdot \frac{\mathrm{d}L}{\mathrm{d}h} + \frac{WL}{h} \cdot \frac{\mathrm{d}\varepsilon}{\mathrm{d}h}. \tag{4.89}$$

In this case

$$\begin{aligned} k &\equiv \frac{\mathrm{d}C}{\mathrm{d}h}\cdot\frac{h}{C} = \frac{\mathrm{d}C}{\mathrm{d}L}\cdot\frac{h^2}{\varepsilon WL} \\ &= -1 + \frac{h}{W}\cdot\frac{\mathrm{d}W}{\mathrm{d}h} + \frac{h}{L}\cdot\frac{\mathrm{d}L}{\mathrm{d}h} + \frac{h}{\varepsilon}\cdot\frac{\mathrm{d}\varepsilon}{\mathrm{d}h} \\ &= -1 - \nu - \nu + \frac{\mathrm{d}\varepsilon/\varepsilon}{\mathrm{d}L/L} \\ &= -1 - 2\nu. \end{aligned} \tag{4.90}$$

The last step is for constant ε.
b) For constant volume,

$$\begin{aligned} V &= LWh = \text{constant}, \\ \frac{\mathrm{d}V}{\mathrm{d}h} &= WL + Lh\cdot\frac{\mathrm{d}W}{\mathrm{d}h} + Wh\cdot\frac{\mathrm{d}L}{\mathrm{d}h} = 0 \\ & \quad 1 + \frac{h}{W}\cdot\frac{\mathrm{d}W}{\mathrm{d}h} + \frac{h}{L}\cdot\frac{\mathrm{d}L}{\mathrm{d}h} = 0 \\ & \quad 1 - \nu - \nu = 0, \\ \nu &= 0.5, \\ k &= -1 - 2\nu = -2. \end{aligned} \tag{4.91}$$

c) For a capacitor of 1 μF:

$$\begin{aligned} \Delta C &= \Delta h\cdot\frac{\mathrm{d}C}{\mathrm{d}h} \\ &= \frac{C}{h}\cdot\Delta h\cdot\left(\frac{h}{C}\cdot\frac{\mathrm{d}C}{\mathrm{d}L}\right) \\ &= C\cdot\frac{\Delta h}{h}\cdot k \\ &= 1\ \mu\mathrm{F}\cdot 1\%\cdot(-2) = -20\ \mathrm{nF}, && (4.92) \\ C' &= C + \Delta C = 1\ \mu\mathrm{F} - 20\ \mathrm{nF} = 980\ \mathrm{nF} && (4.93) \end{aligned}$$

d) $\varepsilon = 7\varepsilon_0 = 6.198$ F/m. $C = \varepsilon WL/h = 6.2$ nF.
e) The sensitivity to force is given by

$$\begin{aligned} S_\mathrm{F} &\equiv \frac{\mathrm{d}C}{\mathrm{d}F} = \frac{\mathrm{d}C}{\mathrm{d}h}\cdot\frac{\mathrm{d}h}{\mathrm{d}F} = \left(\frac{\mathrm{d}C}{\mathrm{d}h}\cdot\frac{h}{C}\right)\cdot\frac{\mathrm{d}h}{\mathrm{d}F}\cdot\frac{C}{h} \\ &= k\cdot\frac{\mathrm{d}h/h}{\mathrm{d}F}\cdot C. \end{aligned} \tag{4.94}$$

We can now calculate the effect of an infinitesimal force d$F$. Because $E = P/\epsilon = \mathrm{d}F/WL\epsilon$, with $\epsilon = \mathrm{d}h/h$, the small force is equivalent to $\mathrm{d}F = EWL \times (dh/h)$, and thus (substituting above)

$$S_\mathrm{F} = \frac{kC}{EWL} = \frac{2\cdot 1\ \mu\mathrm{F}}{5\times 10^7\ \mathrm{Pa}\cdot 0.01\ \mathrm{m}\cdot 0.1\ \mathrm{m}} = 4\times 10^{-11}\ \mathrm{F/n} = 40\ \mathrm{pF/N}. \tag{4.95}$$

f) The sensitivity to mass is given by

$$S_{\mathrm{m}} \equiv \frac{\mathrm{d}C}{\mathrm{d}m} = \frac{\mathrm{d}C}{\mathrm{d}F}\cdot\frac{\mathrm{d}F}{\mathrm{d}m} = S_{\mathrm{F}}\cdot\frac{\mathrm{d}F}{\mathrm{d}m} = S_{\mathrm{F}}\cdot g = 40\ \mathrm{pF/N}\cdot 9.81\ \mathrm{N/kg} = 392.4\ \mathrm{pF/kg}. \tag{4.96}$$

g) On a scale of 20 nF, the digital resilution of the multimeter is $\Delta C = 0.01$ nF, and thus

$$\Delta m = \frac{\Delta C}{S_{\mathrm{m}}} = \frac{10\ \mathrm{pF}}{392.4\ \mathrm{pF/kg}} = 22.5 \times 10^{-3}\ \mathrm{kg}. \tag{4.97}$$

3 Find it out experimentally.

4 The received frequency and sensitivity are given by, resp.

$$\begin{aligned} f &= \left(\frac{v_{\mathrm{m}}+v}{v_{\mathrm{m}}-v}\right)\cdot f_{\mathrm{s}} \\ &\approx \left(1+\frac{2v}{v_{\mathrm{m}}}\right)\cdot f_{\mathrm{s}}, && (4.98) \\ S_{\mathrm{v}\to\mathrm{f}} &= \frac{\mathrm{d}f(v)}{\mathrm{d}v} = \frac{2f_{\mathrm{s}}}{v_{\mathrm{m}}}. && (4.99) \end{aligned}$$

With $\Delta f = 1/(1\ \mathrm{s}) = 1$ Hz, $v_{\mathrm{m}} = 3 \times 10^8$ m/s and $f_{\mathrm{s}} = 3$ GHz, with the sensitivity $S_{\mathrm{v}\to\mathrm{f}} = (6\ \mathrm{GHz})/(3 \times 10^8\ \mathrm{m/s}) = 20/\mathrm{m}$, this gives

$$\Delta v = \frac{\Delta f}{S_{\mathrm{v}\to\mathrm{f}}} = \frac{1\ \mathrm{Hz}}{20/\mathrm{m}} = 0.05\ \mathrm{m/s} = 0.18\ \mathrm{km/h}. \tag{4.100}$$

5 a) The signal at a pH of 5 is equal to 118.26 mV. To get the highest resolution we have to subtract this voltage. Assuming a 10 V power supply we can design the following circuit that includes two voltage followers (to not draw current from the sensor or voltage reference) and an adder (−1× amplifier). See Figure 4.57. b) The resolution $\Delta N_{\mathrm{H^+}}$ is given by the resolution of the multimeter $\Delta V_{\mathrm{m}}$ and the sensitivity of the system, $S = \mathrm{d}V_{\mathrm{m}}/\mathrm{d}N_{\mathrm{H^+}}$,

$$\Delta N_{\mathrm{H^+}} = \frac{\Delta V_{\mathrm{m}}}{\mathrm{d}V_{\mathrm{m}}/\mathrm{d}N_{\mathrm{H^+}}}, \tag{4.101}$$

$$\frac{\mathrm{d}V_{\mathrm{m}}}{\mathrm{d}N_{\mathrm{H^+}}} = \frac{\mathrm{d}V_{\mathrm{m}}}{\mathrm{dpH}}\cdot\frac{\mathrm{dpH}}{\mathrm{d}N_{\mathrm{H^+}}} = \frac{\mathrm{d}V_{\mathrm{m}}}{\mathrm{d}V_{\mathrm{pH}}}\cdot\frac{\mathrm{d}V_{\mathrm{pH}}}{\mathrm{dpH}}\cdot\frac{\mathrm{dpH}}{\mathrm{d}[H^+]}\cdot\frac{\mathrm{d}[H^+]}{\mathrm{d}N_{\mathrm{H^+}}} \tag{4.102}$$

We can calculate the individual terms and substitute in the equations

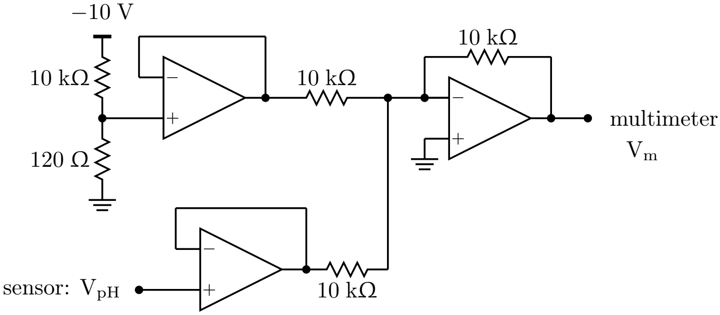


**Fig. 4.57**: Circuit for a pH sensor amplifier. The voltage follower ensures high input impedance. The other branch is the reference to subtract the offset and the third opamp is an adder. (Exercise 5)

above:

$$\begin{aligned}
&\Delta V_{\mathrm{m}} = 0.01\ \mathrm{mV},\\
&\mathrm{pH} = 5 \Rightarrow [\mathrm{H}^+] = 10^{-5}\ \mathrm{mol/L},\\
&V_{\mathrm{m}} = V_{\mathrm{pH}} - 118\ \mathrm{mV} \Rightarrow \frac{\mathrm{d}V_{\mathrm{m}}}{\mathrm{d}V_{\mathrm{pH}}} = 1 \text{ (see circuit in Figure 4.57)},\\
&V_{\mathrm{pH}} = (59.13\ \mathrm{mV}) \times (\mathrm{pH} - 7) \Rightarrow \frac{\mathrm{d}V_{\mathrm{pH}}}{\mathrm{dpH}} = 59.13\ \mathrm{mV},\\
&\mathrm{pH} = -\log_{10}\left(\frac{[\mathrm{H}^+]}{\mathrm{mol/L}}\right) \Rightarrow \frac{\mathrm{dpH}}{\mathrm{d}[\mathrm{H}^+]} = \frac{1}{\ln(10) \times [\mathrm{H}^+]} = 4.34 \times 10^4\ \mathrm{L/mol},\\
&[\mathrm{H}^+] = \frac{N_{\mathrm{H}^+}}{N_{\mathrm{A}} \times 3\ \mathrm{L}} \Rightarrow \frac{\mathrm{d}[H^+]}{\mathrm{d}N_{\mathrm{H}^+}} = 5.54 \times 10^{-25}\ \mathrm{mol/L}. \qquad (4.103)
\end{aligned}$$

Substituting in Eqs. (4.102) and (4.101) gives:

$$\frac{\mathrm{d}V_{\mathrm{m}}}{\mathrm{d}N_{\mathrm{H}^+}} = 1.42 \times 10^{-24}\ \mathrm{V/atom\ H}^+ \qquad (4.104)$$

$$\Delta N_{\mathrm{H}^+} = \frac{10\ \mu\mathrm{V}}{1.42 \times 10^{-24}\ \mathrm{V/atom\ H}^+} = 7.0 \times 10^{18}\ \text{atoms H}^+ \qquad (4.105)$$

# 5 Informatics

## 5.1 Introduction

The next step is bringing the information from the sensors to digital processing units, for instance computers. While this step is not essential – any processing can always also be done in the electronic level – some advanced processing is more readily done on digital processing equipment. The transfer to digital processors consists of three steps

1. Conversion from the analog to the digital electronic world.
2. Communication to the digital processor.
3. Signal processing

At the end the opposite path can be followed, i.e., communication from the digital equipment and conversion to the analog electronic world.

## 5.2 Analog-digital and digital-analog conversion

After the signal is prepared by electronic circuit to be of the correct amplitude, offset, and frequency spectrum, it is time to convert it to a digital format to be further processed numerically. While this is not necessary – simple instrumentation systems can do with only electronics – advanced systems use some form of on-line or off-line processing and for that the information is needed in a digital format. This is done by analog-digital converters (ADCs) and digital-analog converters (DACs).

### 5.2.1 Parameters of ADCs and DACs

For linear ADCs, the output bit-pattern (the 'number') is a linear function of the input voltage (ignoring offset). These ADCs have three important parameters:

- Input voltage range, $V_{\text{min}} - V_{\text{max}}$.

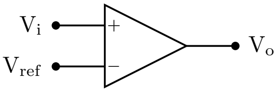


**Fig. 5.1**: A 1-bit ADC is a simple comparator. It outputs $+V_{\mathrm{CC}}$ if $V_{\mathrm{i}} > V_{\mathrm{ref}}$ and $-V_{\mathrm{CC}}$ if $V_{\mathrm{i}} < V_{\mathrm{ref}}$

- (Digital) resolution in voltage $\Delta V$, determined by the above and the number of bits $n$ and is defined as the distance between quantization levels.
- Resolution in time $\Delta t$ determined by the sampling rate $f$

Analyzing the first two, the voltage resolution is given by

$$\Delta V = \frac{V_{\mathrm{max}} - V_{\mathrm{min}}}{2^n - 1} \tag{5.1}$$

As an example, an ADC with a 0 to 5 volt range with 8 bits has a digital resolution of $\Delta V = 5\mathrm{V}/255 = 19.6$ mV.

The simplest ADC is a 1-bit ADC that only determines if the signal is positive or negative, or in general below or above a certain reference level, see Figure 5.1. This we called a comparator in the chapter on electronics.

The most accurate ADCs have 24 bits and thus have digital voltage resolutions of about $\Delta V = 10\mathrm{V}/(2^24-1) = 0.60$ µV. Increasing the resolution further has no use, since the digital resolution will drop below the analog noise level. Moreover, if we use the ADCs (and DACs) for human-destined signals, there is no sense in increasing the resolution, since the human mind cannot distinguish higher resolutions. The maximum resolution is some 14 bits for audio signals and 24 bits for video signals.

The other important parameter is the sampling rate. It is no use sampling faster than (two times) the maximum frequency of information, as it will only increase the cost of the equipment.

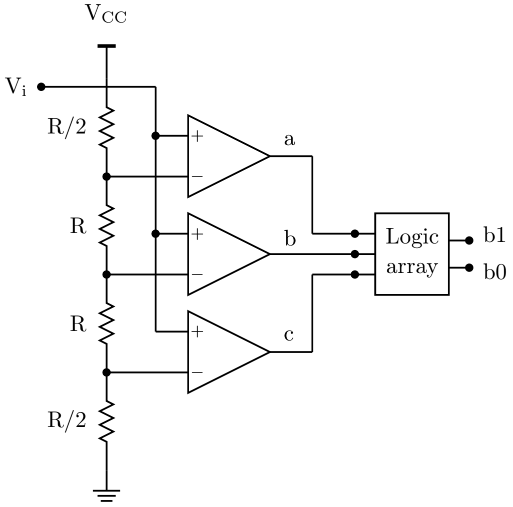
$$f_{\mathrm{sample}} = 2f_{\mathrm{max}} \tag{5.2}$$

which is called the Shannon-Nyquist sampling theorem.

### 5.2.2 ADCs

There exist several types of ADCs, which differ much in cost (complexity) and speed. Overall it can be said that if the number of bits increases, the complexity increases and the cost increases or the speed goes down. On the other hand, if the speed requirements go up, the number of bits must be reduced, or a high price has to be paid. Some examples of ADC are given

A flash ADC has $2^n - 1$ comparators that compare the signal to $2^n - 1$ reference signals. Each comparator has a binary output signal telling if the signal is larger ('1', $V_{\mathrm{o}} = +V_{\mathrm{CC}}$) or smaller ('0', $V_{\mathrm{o}} = -V_{\mathrm{CC}}$) than the reference signal of that comparator. An example of a 2-bit flash-ADC might look something



Logic array

| a | b | c | b1 | b0 |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | $V_{CC}$ | 0 | 1 |
| 0 | $V_{CC}$ | $V_{CC}$ | 1 | 0 |
| $V_{CC}$ | $V_{CC}$ | $V_{CC}$ | 1 | 1 |

**Fig. 5.2**: 2-bit flash ADC consisting of three comparators and a logic array

like the circuit shown in Figure 5.2. It consists of three comparators comparing to three equidistant voltages given by the voltage divider given by four equal resistances $R$, with the extreme two half that value. Each comparator gives an out of 0 or $+V_{CC}$, and there are thus theoretically $2^3 = 8$ possibilities. However, a state where the a lower comparator gives 0 and one above it $V_{CC}$ is not possible (it is not possible that the voltage is larger than the reference voltage at one comparator, but not larger than the reference voltage at a comparator below it). Physically, there are only four possible states. A logic array converts these four states into a 2-bit output, as shown by the conversion table of the logic array in the figure.

The advantage of the flash ADC, as its name implies, is that the conversion to digital is nearly instantaneous and is only limited by the electronics rise times of the comparators and the logic array. The price to pay is a very complicated circuit that grows rapidly with the number of bits: $\propto 2^n - 1$ comparators are needed for an n-bit flash ADC. Another disadvantage is that the resistances R of the reference voltages have to be equal to a large accuracy, in order to guarantee linearity of the circuit; if the input voltage is twice as large, we want twice as large digital number at the output. For that reason, flash converters are normally limited to 8-bits, but on the other hand easily go up to GHz sampling rates. A good example is the digital oscilloscopes.

The second type of ADC is the delta-encoded ADC. In this ADC, a digital counter is fed into a DAC (to be discussed later) and compared with the input voltage in a single comparator. The counter stops when the DAC signal overtakes the input signal. The value of the counter is then copied to the ADC output using a latch. See for example the 2-bit delta-encoded ADC of Figure 5.3

The electronics of this type of ADC are simpler, but it is also much slower.

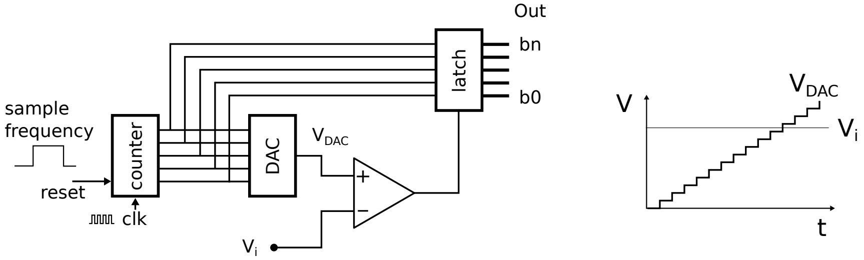


**Fig. 5.3**: n-bit delta-encoded ADC

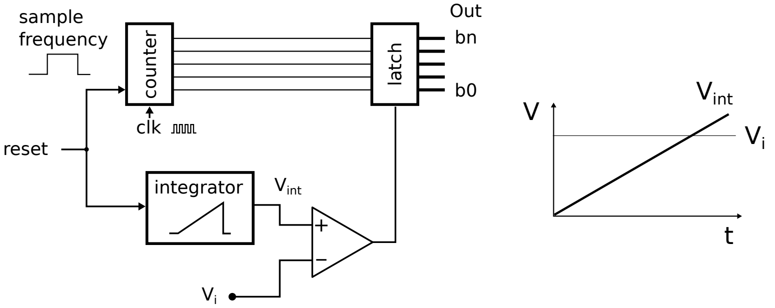


**Fig. 5.4**: Analog-ramp ADC

In the worst case, the sample is ready only after $2^n$ (counter) clock cycles. For example, a 24 bit ADC with a clock frequency of 1 GHz can only sample at a frequency of 60 Hz. There exist improved versions of this ADC, in which the last value is kept and used as a starting point for the next sampling, hence the name delta ADC. The counter will count up or down from that point. In principle the same limitation then exists – theoretically the first sample can be 0 and the next $2^n - 1$ – but the dynamic range can be limited by limiting the difference between two consecutive samples.

Another improvement is to first determine the MSBs (most significant bits) and then subtract this coarse voltage with a DAC and a differential amplifier from the input signal. In the second step the remaining bits (LSBs) are found.

The last ADC type discussed here is the analog-ramp ADC. This is similar to the delta ADC, but instead of a DAC, an analog counter (integrator) is used to compare the signal. Once again, the value of the digital counter is copied to the output of the latch once the voltage at the comparator overtakes the input voltage. Figure 5.4 shows an example.

### 5.2.3 DACs

A DAC (digital-analog converter) is having the opposite function compared to an ADC, it is converting from the digital domain to the electrical domain. Its workings are quite simple, it is based on the summing amplifier opamp

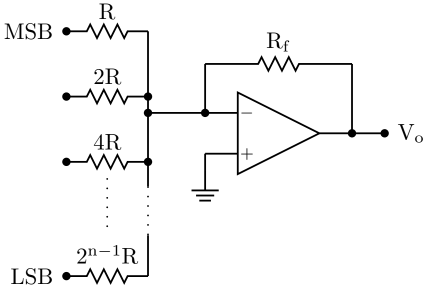


**Fig. 5.5**: An implementation of a DAC (digital-analog converter) based on a summing amplifier opamp configuration

configuration of Chapter 2, and might be something like in Figure 5.5. Each bit $b$ input voltage is weighted differently by an appropriate choice of input resistance; two consecutive bits differ in input resistance by a factor two. Together with the feedback resistance $R_\text{f}$, each bit $b$ gets amplified a factor $R_\text{f}/(2^{n-b}R)$, with the MSB ($b = n - 1$) thus getting a weight $R_\text{f}/R$ and the least-significant bit ($b = 0$) a weight $R_\text{f}/(2^n R)$ in the output voltage.

It is obvious that for good linearity, the resistances have to be of extremely high precision. For instance, the conductance variation of the MSB resistance cannot be larger than half the conductance value of the LSB resistance. This puts severe demands on the precision of the resistances. Imagine a 10-bit ADC with the MSB resistance being 1 kΩ (conductance 1 mS). The LSB resistance is then $2^9$ times larger, 512 kΩ (1.95 µS). The MSB resistance should have a precision of (1.95 µS)/(2 ms) = 0.10%, as good as impossible to achieve.

### 5.2.4 ADC/DAC oversampling

An elegant technique to increase the (digital) resolution of an ADC or DAC is by use of oversampling. This can actually be achieved by *adding* noise to the input signal and using oversampling, i.e., sampling more than once for every desired sample.

Imagine an ADC with 1 volt resolution that receives a noiseless signal of 2.1 V. This would result in an ADC value of $n$, representing 2 V (we assume we have an ADC that rounds down), see Figure 5.6a. For one sample, or one hundred, or one thousand, the result will always be 2 volt. Adding noise (and offset) at the input (much) larger than the digital quantization distance (1 V) will cause the ADC to every now and then return $n + 1$ (3 V), or $n + 2$ (4 V), or $n - 1$ (1 V). For a large factor of oversampling, the average of this will give 2.1 V, see Figure 5.6b. In fact, the resolution is increased by a factor $N$ equal to the oversampling rate,

$$\Delta V = \frac{V_\text{max} - V_\text{min}}{(2^n - 1)N} \tag{5.3}$$

Also for DACs oversampling can be used. Imagine we have an n-bit DAC

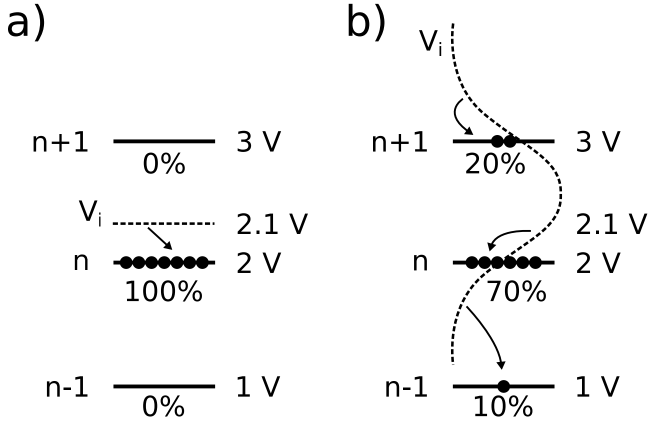


**Fig. 5.6**: ADC oversampling. a) Without adding noise, the input voltage $V_\text{i}$ = 2.1 V will always result in the same number $n$ corresponding to 2 V. b) By adding noise (here Gaussian) and offset, the input voltage can be reconstructed if oversampling is used in combination with averaging

from 0 to 10 V. The resolution is thus $10/(2^n - 1)$ and we have a value $x$ between 0 and $2^n - 1$. The output voltage, $V_\text{o} = x \times (10 \text{ volt})/(2^n - 1)$ can equally well be generated by placing $x$ times a logical 1 (10 volt) at the output and $2^n - 1 - x$times a logical 0 (0 volt) and then taking the average by a low-pass filter (LPF). In other words, we can use a 1-bit DAC plus LPF instead of a n-bit DAC, see Figure 5.7a.

The advantage of this scheme is that the DAC is extremely linear and the electronics are simple. To increase the frequency characteristics, the bit pattern of $x$ 1s and $2^n - 1 - x$ 0s is randomized before submitting it to the 1-bit DAC, see Figure 5.7b. This randomization of the bits places the non-interesting transitions higher up in the spectrum (because they are more frequent) from where they can be more easily filtered off by the LPF.

## 5.3 Communication

The communication between equipment in general in modern times occurs mostly through USB cables (universal serial bus). To understand them, it is better to take one step back and analyze the serial communications. We set the clock back some 20 years and talk about how computers were communicating between them and with peripheral equipment such as printers.

Communication in general can be divided into two types, serial and parallel, which describes the amount of bits that are being sent simultaneously; one bit at a time for serial communications and several bits at once (for instance a complete byte of 8 bits) for parallel communications. Since each bit being sent needs at least one wire in the cable used for communication (actually, most

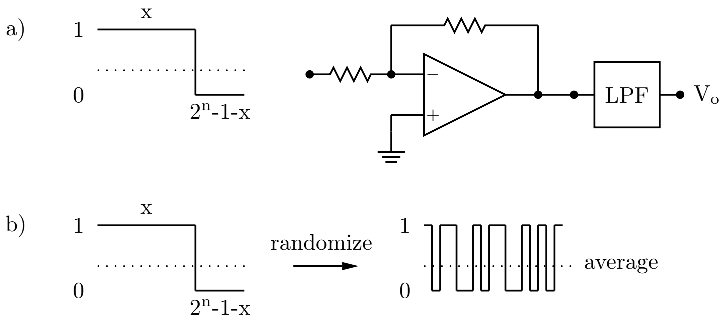


**Fig. 5.7**: DAC oversampling. a) The technique of oversampling used to turn a 1-bit DAC effectively into an n-bit DAC; $x$ 1s and $2^n - 1 - x$ os are supplied to the 1-bit DAC and the output is low-pass filtered. b) To have better frequency characteristics, the bit pattern is randomized before submitting to the DAC

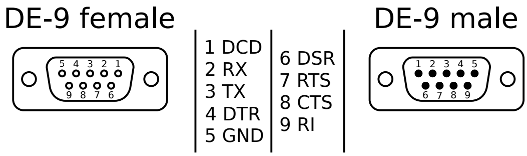


**Fig. 5.8**: DE-9 connector pin-out as seen looking at the cable or computer for female connectors with 9 holes (left) and male connectors with 9 pins (right). Note that the pin numbers have inverted. The pin labeling applies to its function as an RS-232 serial communications connector

use two for every bit, for instance each has its own ground reference wire), apart from the wires used for 'overseeing' the transfer of data, as we will see, the serial cables have less wires compared to the parallel cables and this is the reason why initially serial communications were used for long distances and parallel communications for short distances; at every application a trade-off is made between cost and speed. Speed is sacrificed to reduce the cost for long-distance communication.

### 5.3.1 Serial communication

At long distances the serial protocol is used. The most famous is the RS232 (Recommended Standard) protocol. Most early-days computers used this protocol to communicate between each other and with modems (modulator-demodulator). It can be recognized by the characteristic DB-25 connectors which were later replaced by the even more famous DE-9 connectors, for instance in the first computer mouses connected to Intel 8086 computers. See Figure 5.8. Note that female connectors have inverted numbering compared to the male connectors. A mouse cable normally had female connectors and the computer male connectors.

**Table 5.I**: RS232 signals (on DE-9 connector)

| Name | Pin | Direction | Meaning |
|---|---|---|---|
| GND | 5 | - | Ground |
| TX | 3 | DTE → DCE | Transmitted data |
| RX | 2 | DTE ← DCE | Received data |
| RTS | 7 | DTE → DCE | Request to send |
| CTS | 8 | DTE ← DCE | Clear to send |
| DTR | 4 | DTE → DCE | Data terminal ready |
| DSR | 6 | DTE ← DCE | Data set ready |
| DCD | 1 | DTE ← DCE | Data carrier detect |
| RI | 9 | DTE ← DCE | Ring indicator |

The standard of RS-232 serial communication is so widespread that even nowadays it is very often used. Nearly all low-transfer rate communication between equipment is done by RS-232, also because dedicated communications ICs – so-called UART (universal asynchronous receiver/transmitter) – are readily available and are cheap. The first version of RS-232 was designed for linking a typewriter to a modem (modulator-demodulator, to be described later on), or in general a piece of DCE (data communication equipment) to a DTE (data terminal equipment) like a terminal in a setting of a mainframe computer with many terminals. The original connector pin numbering is given in Figure 5.8. For such situations, each pin on the terminal is connected to the same pin on the modem. For instance, TX (transmitted data) on the terminal is connected to TX on the modem, see Table 5.I. This might seem confusing in modern terms, since the modem *receives* data on this pin. Yet, from the overall point of view of modem communication it makes sense, since this data is then modulated and *transmitted* by the modem via a telephone line; overall the data is transmitted to the listener on the other side of the telephone via the TX pins.

Apart from the data information signals, there also exist so-called handshake signals. The terminal signals the modem it has data ready to send by asserting the RTS (request to send) line. The terminal, however, only actually starts transmitting the data on the TX line when it receives a signal that the modem is ready to receive the data (when the modem asserts DSR, data set ready). Likewise, in the opposite direction, when the modem has some data for the terminal, it asserts the CTS (clear to send) line and waits for the terminal to acknowledge it is ready for this data (when the terminal asserts DTR, data terminal ready). This ensures that data is never lost. Original UART chips had no or few data buffers and without some sort of flow control, there is a very high risk of data being lost.

Other signals that were used only for modem communication is are DCD (carrier detect, if the phone line is actually present) and RI (if a call is coming in).

In modern times, much more often serial communication is used between

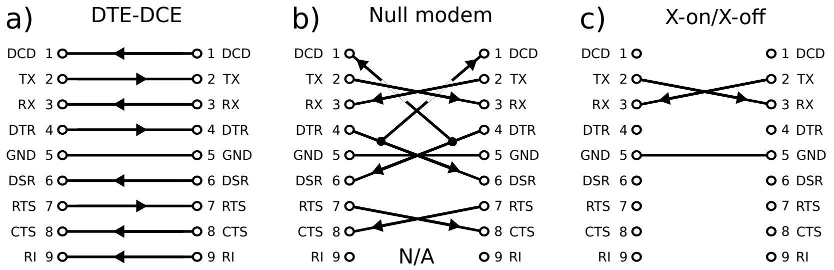


**Fig. 5.9**: RS232 connections types a) Original (terminal-modem) DTE-DCE, b) Null modem (cross-linked), c) Modern cable, software handshaking (X-on/X-off)

equipment of the same type, for instance two computers. It is immediately obvious that the above connections will not work. Both computers would see the TX line as output, etc. For these kind of communications null modem cables are used, they cross-connect the various signals. The pin 3 (TX) on one side is connected to pin 2 (RX) on the other side and vice versa. Also the handshake signals are cross-linked (RTS-CTS, DTR-DSR). Note that for such connections the ring indicator does not make sense anymore (computers normally do not ring each other). Figure 5.9 shows the pin connections in terminal-modem cables and computer-computer (null) modem cables. The cables normally (but not always) have female connectors on both sides (and male connectors on the computers). The only way of being sure is reading the manuals of the equipment and measure the cable with a multimeter.

In modern systems buffers can be very large and it is nearly impossible to have an overflow and a loss of data. It is quite safe to use no handshaking whatsoever. Still, many systems use some form of flow control, mostly by software in the X-on/X-off protocol. That is, the equipment sends a byte to pause transmission (X-off, ASCII character 19) on the TX data line when its buffer gets too full and another byte to resume transmission (X-on, ASCII character 17) when it is ready to receive again. This requires some intelligence in the systems and is only found in modern systems. See Figure 5.9c.

More complicated it gets when one side requires handshaking and the other side does not. We can then use loopback techniques, to fool the device that needs handshaking into thinking there is actually handshaking going on. We achieve this by soldering the RTS and CTS signals on one side to each other. Whenever the device is ready-to-send by asserting the RTS signal, it also thinks the other side is clear-to-send because it sees the CTS line asserted, and sends the data. The same is done with the DSR, DCD and DTR lines.

Serial communication consists of sending of small packages, 7 or 8 bits, one bit at a time, over the TX/RX line. Such signals have the following properties:

- Consist of logical 0s (called 'space') and 1s (called 'mark') where a logical 0 is a voltage equal to +12 V (normally between +3 and +10 V) and a logical 1 is a voltage equal to −12 V (normally between −3 and −10 V). Modern communications can use different levels, in the 0-5 volt range, in so called pseudo-RS232, with a 1 being +5 V and a 0 equal to −5 V, i.e., inverted sign compared to true RS232.

- Start with a start bit, a logical 0.
- Are followed by a 'character' that contains the information (the 'data' being transmitted), and consists of 7 or 8 data bits with the least significant bit (LSB) first and most significant bit (MSB) last.
- In case the character has 7 bits, the sequence is followed by a parity bit that verifies the integrity of the databits. It is either 'E' (even) or 'O' odd, depending on the number of 1s in the total 8-bit sequence, including the parity bit. Example, if the data bits are 0110100 and we chose even parity, the parity bit is 1 because that makes the total number of 1s even. In case the character is 8 bit, there is no parity bit, which is called parity 'N' (none).
- Followed by a stop bit, a logical 1, of arbitrary length (minimum 1 bit).
- Can be sent at various speeds, which is called the bitrate, or 'baud', describing the number of bits per second transmitted (including parity bit). Standards are 75, 150, 300, 1200, 2400, 4800, 9600 and 19200 baud with the first one barely adequate for typewriters and the last one barely adequate for transmitting files from one computer to another. Faster rates have been implemented but they are not standard. Default normally is 9600.

The communication is then often summarized in a single designation, for instance "9600N81", signifying: speed (9600), parity (none), data bits (8) and stop bit length (1). A typical RS232 bit pattern (9600N71-codification of the number 105) is given in Figure 5.10. In many cases, serial communication is also used between TTL (transistor-transistor logic) components. While the timing of these serial signals is the same as true RS232, the voltage levels are different. In TTL, a logical "1" is equal to +5 volt and a logical "0" is 0 volt. Note the absence of 'inversion'. Figure 5.10 shows an example. To make a TTL component communicate with other equipment through true RS232, so-called level-shifter circuits exist, the most famous integrated circuits being those of MAXIM. An even more remarkable feature of this chip is that it can translate the 0-5 volt signals to −12-12 volt signals using only 5 volt power supply. It makes use of charge-pumped DC-DC converters such as the Dickson charge pump.

The next step is the codification of information. After determining how to send the bit pattern as done above, it has to be decided to give meaning (information) to the patterns. From the very start, a convention was designed to attribute a meaning to every binary bit pattern (probably the other way around, a bit pattern was assigned to a meaning). Simultaneously, a number

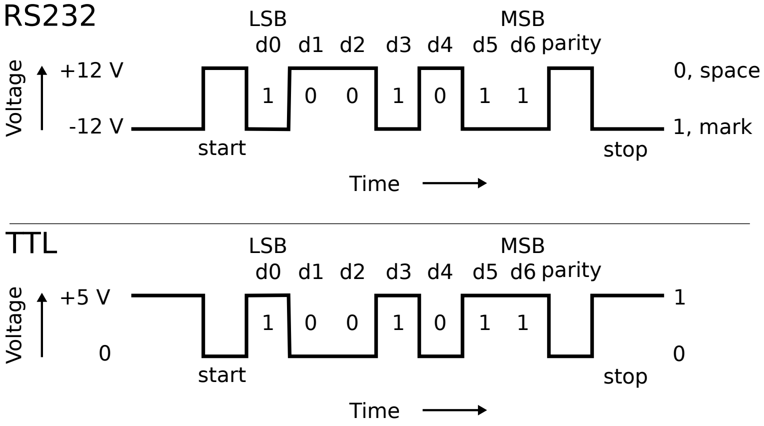


**Fig. 5.10**: Example of 9600N71 RS-232 pattern sending of the character 1101001 (ASCII 'i', 105). The pattern starts with a start bit (logical 0) followed by 7 data bits in reverse order (LSB first), followed by a parity bit (0 in this case, because the number of 1s is already even), followed by a stop bit (logical 1). A logical 1 is called a 'mark' and is a low voltage, while a logical 0 is called a 'space' and is high voltage. The bottom part shows the same pattern in TTL CMOS 'pseudo RS232' signal, as for instance used by the Arduino

was attributed to a bit pattern by given bit 1 to 7 an ascending weight from 1, 2, 4, 8, 16, 32, and 64 (and 128 for the 8th bit that was later added). I.e., the least significant, lowest weight, bit was the first one. It was also convention to express the numbers in hexadecimal, with the number $mn$ representing $16 \times m + n$. In hexadecimal digits with values between 10 and 15 were represented by letters 'A' until 'F'. In this way we can reconstruct the official table of ASCII (American standard code for instruction interchange), see Tables 5.II and 5.III. The first table shows the control codes (like the X-off and X-on codes discussed above). Interesting are the two characters 'line feed' (LF) and 'carriage return' (CR), inherited from old (line) printers. A line feed would put the printing head one line down (actually, it would put the paper one line up, of course), whereas a carriage return would put the printing head at the beginning of the current line. Humans normally understand a 'new line' as a combination of both, *and* move one line down *and* place the head (pen) at the beginning of the line. One instruction would suffice. UNIX (and modern day Linux) used the character LF for that purpose from the start. Microsoft invented for their MS-DOS the idea to use both CR+LF for a newline. In the 21st century the world still suffers from this difference between the two major operating systems.

The second table is the 'human' information table, with standard characters of the standard English alphabet in both uppercase ('A'-'Z', 65-90) and lowercase ('a'-'z', 97-122) plus most-used punctuation characters translated into codes. This latter table also shows one more control code, namely 'delete', a control code to indicate the elimination of the character below the cursor.

**Table 5.II**: ASCII control codes

| Dec | Hex | Bin | Value | Meaning |
|---|---|---|---|---|
| 0 | 00 | 0000000 | NUL | Null character |
| 1 | 01 | 0000001 | SOH | Start of header |
| 2 | 02 | 0000010 | STX | Start of text |
| 3 | 03 | 0000011 | ETX | End of text (Ctrl-C) |
| 4 | 04 | 0000100 | EOT | End of transmission |
| 5 | 05 | 0000101 | ENQ | Enquiry |
| 6 | 06 | 0000110 | ACK | Acknowledge |
| 7 | 07 | 0000111 | BEL | Bell |
| 8 | 08 | 0001000 | BS | Back space |
| 9 | 09 | 0001001 | HT | Horizontal tab |
| 10 | 0A | 0001010 | LF | Line feed * |
| 11 | 0B | 0001011 | VT | Vertical tab |
| 12 | 0C | 0001100 | FF | Form feed |
| 13 | 0D | 0001101 | CR | Carriage return * |
| 14 | 0E | 0001110 | SO | Shift out |
| 15 | 0F | 0001111 | SI | Shift in |
| 16 | 10 | 0010000 | DLE | Data link escape |
| 17 | 11 | 0010001 | XON | Device control 1 |
| 18 | 12 | 0010010 | DC2 | Device control 2 |
| 19 | 13 | 0010011 | XOFF | Device control 3 |
| 20 | 14 | 0010100 | DC4 | Device control 4 |
| 21 | 15 | 0010101 | NAK | Not acknowledge |
| 22 | 16 | 0010110 | SYN | Synchronous idle |
| 23 | 17 | 0010111 | ETB | End of transfer block |
| 24 | 18 | 0011000 | CAN | Cancel |
| 25 | 19 | 0011001 | EM | End of medium |
| 26 | 1A | 0011010 | SUB | Substitute (Ctrl-Z) |
| 27 | 1B | 0011011 | ESC | Escape |
| 28 | 1C | 0011100 | FS | File separator |
| 29 | 1D | 0011101 | GS | Group separator |
| 30 | 1E | 0011110 | RS | Record separator |
| 31 | 1F | 0011111 | US | Unit separator |

$*$: UNIX (Linux): newline is LF, MS-DOS (Windows): newline is CR+LF

**Table 5.III**: ASCII character codes

| Dec | Hex | Bin | Value | Dec | Hex | Bin | Value | Dec | Hex | Bin | Value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 32 | 20 | 0100000 | space | 64 | 40 | 1000000 | @ | 96 | 60 | 1100000 | ’ |
| 33 | 21 | 0100001 | ! | 65 | 41 | 1000001 | A | 97 | 61 | 1100001 | a |
| 34 | 22 | 0100010 | " | 66 | 42 | 1000010 | B | 98 | 62 | 1100010 | b |
| 35 | 23 | 0100011 | # | 67 | 43 | 1000011 | C | 99 | 63 | 1100011 | c |
| 36 | 24 | 0100100 | $ | 68 | 44 | 1000100 | D | 100 | 64 | 1100100 | d |
| 37 | 25 | 0100101 | % | 69 | 45 | 1000101 | E | 101 | 65 | 1100101 | e |
| 38 | 26 | 0100110 | & | 70 | 46 | 1000110 | F | 102 | 66 | 1100110 | f |
| 39 | 27 | 0100111 | ’ | 71 | 47 | 1000111 | G | 103 | 67 | 1100111 | g |
| 40 | 28 | 0101000 | ( | 72 | 48 | 1001000 | H | 104 | 68 | 1101000 | h |
| 41 | 29 | 0101001 | ) | 73 | 49 | 1001001 | I | 105 | 69 | 1101001 | i |
| 42 | 2A | 0101010 | * | 74 | 4A | 1001010 | J | 106 | 6A | 1101010 | j |
| 43 | 2B | 0101011 | + | 75 | 4B | 1001011 | K | 107 | 6B | 1101011 | k |
| 44 | 2C | 0101100 | , | 76 | 4C | 1001100 | L | 108 | 6C | 1101100 | l |
| 45 | 2D | 0101101 | - | 77 | 4D | 1001101 | M | 109 | 6D | 1101101 | m |
| 46 | 2E | 0101110 | . | 78 | 4E | 1001110 | N | 110 | 6E | 1101110 | n |
| 47 | 2F | 0101111 | / | 79 | 4F | 1001111 | O | 111 | 6F | 1101111 | o |
| 48 | 30 | 0110000 | 0 | 80 | 50 | 1010000 | P | 112 | 70 | 1110000 | p |
| 49 | 31 | 0110001 | 1 | 81 | 51 | 1010001 | Q | 113 | 71 | 1110001 | q |
| 50 | 32 | 0110010 | 2 | 82 | 52 | 1010010 | R | 114 | 72 | 1110010 | r |
| 51 | 33 | 0110011 | 3 | 83 | 53 | 1010011 | S | 115 | 73 | 1110011 | s |
| 52 | 34 | 0110100 | 4 | 84 | 54 | 1010100 | T | 116 | 74 | 1110100 | t |
| 53 | 35 | 0110101 | 5 | 85 | 55 | 1010101 | U | 117 | 75 | 1110101 | u |
| 54 | 36 | 0110110 | 6 | 86 | 56 | 1010110 | V | 118 | 76 | 1110110 | v |
| 55 | 37 | 0110111 | 7 | 87 | 57 | 1010111 | W | 119 | 77 | 1110111 | w |
| 56 | 38 | 0110100 | 8 | 88 | 58 | 1011000 | X | 120 | 78 | 1110100 | x |
| 57 | 39 | 0111001 | 9 | 89 | 59 | 1011001 | Y | 121 | 79 | 1111001 | y |
| 58 | 3A | 0111010 | : | 90 | 5A | 1011010 | Z | 122 | 7A | 1111010 | z |
| 59 | 3B | 0111011 | ; | 91 | 5B | 1011011 | [ | 123 | 7B | 1111011 | { |
| 60 | 3C | 0111100 | < | 92 | 5C | 1011100 | \ | 124 | 7C | 1111100 | \| |
| 61 | 3D | 0111101 | = | 93 | 5D | 1011101 | ] | 125 | 7D | 1111101 | } |
| 62 | 3E | 0111110 | > | 94 | 5E | 1011110 | ˆ | 126 | 7E | 1111110 | ˜ |
| 63 | 3F | 0111111 | ? | 95 | 5F | 1011111 | _ | 127 | 7F | 1111111 | delete |

Later extensions were made to the ASCII table. The first extension was adding one bit. The newly created 128 patterns could thus include special characters, like regional accents. The problem with this 8-bit table is that it is no longer standard and two people or two computers communicating with 8-bit code possibly do not understand each other and the text might arrive garbled, with funny symbols in the text. Later extensions try to resolve this problem by redesigning a planet-wide standard, called UTF-8 and UTF-16, which now include all character codes in the world, including all Asiatic character sets. However, ASCII is still the most widely used communication standard.

### 5.3.2 Using RS-232 in modern computers

Modern computers do no longer have the RS-232 port and mainly come with USB ports (discussed later on). If we want to use the RS-232 communication,

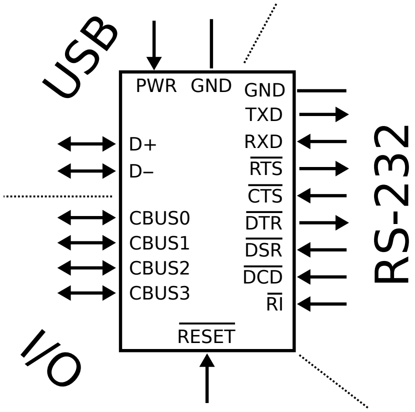


**Fig. 5.11**: FTDI chip functionality translating USB signals into (pseudo) RS-232 signals (TTL). Mnemonics with a dash indicate inverted signals (low is active), which is standard in RS-232. Apart from USB signals and RS-232, the chip also has general I/O (input/output) ports and a reset

we can buy a USB-serial adapter. These use a dedicated IC from FTDI (Future Technology Devices International) to do the signal translation from USB to RS-232, see Figure 5.11. Since the chip is powered by the USB power line (5 V) it does not work with standard RS-232 signals of ±12 volt, but 0 V and 5 volt logic levels instead.

Once we plug in the adapter into a USB port, a virtual port is created. In Linux these can be found in the directory `/dev/`, and should be something like `ttyACM0`, or `ttyUSB0`. We can configure the software to use this device. For instance in the `dosbox.conf` file of the DOSBox MS-DOS emulator we add a line

```
serial1=directserial realport:ttyACM0
```

Or for any Windows application in the WINE (Windows-like) environment, simply create a symbolic link in the directory `/.wine/dosdevices/` with the name of the device, for instance

```
ln -s /dev/ttyACM0 ./COM1
```

After that, effectively a serial port exists that is indistinguishable from a real one. The FTDI IC can even handle all kinds of handshake signals.

### 5.3.3 Modem communication

The modem was already mentioned in the previous section. It is a piece of equipment that modulates the digital bit stream onto a carrier and sends it to a receiver via a telephone line. The idea is to code a 0 into one frequency and a 1 into another frequency. This is a technique called frequency shift key (FSK), see Figure 5.12. Both frequencies should be below the cut-off frequencies as imposed by the telephone operator, which is normally around 3 kHz. Since for

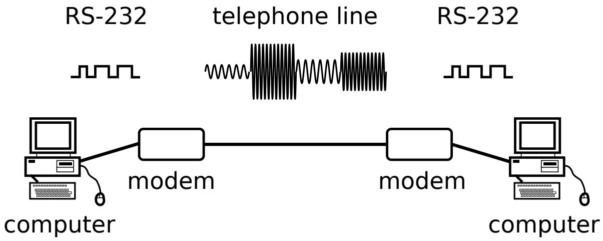


**Fig. 5.12**: Use of a modem (modulator-demodulator) to transmit information over a long distance using a telephone line. The modem modulates the individual bits into frequencies and amplitudes

a frequency to be determined, at least one full cycle should be transmitted, this limits the bitrate to about 3 kb/s. Modern modems also use amplitude modulation techniques to increase the bitrate. Before the advent of ADSL and optical connections, 56 kb/s modems were very popular.

To configure the modem the same RS-232 line is used. The Commands are sent to the modem to set it up. The most used is the standard invented by Hayes. Commands start with 'AT', and that is why they are also called AT commands. Now the problem is: how does the modem know the information is to be used to configure the modem or to be sent through telephone to the other side? Maybe the user wanted to send the text "AT HOME" to the other computer. The answer is that a special dedicated escape sequence is used to put the modem into setup mode. In the Hayes protocol, this is

a one-second pause, +++, a one-second pause

It is very unlikely our data contains this sequence, especially considering the pause. From that moment on, the modem interprets all text as set-up instructions. For instance `ATZ` to reset the modem. Table 5.IV shows the most used ones. The modem automatically goes back to data mode, with the received data forwarded to the computer on the other side of the telephone line, after some idle time of silence without receiving any characters from the local computer. Typically, a command `ATDT2075285716` will be issued to let the modem connect to another modem listening at telephone number 20756285716. That modem has a computer connected that waits for a text "RING" coming from the modem. That computer issues a command `ATA` and from that moment on the two computers can talk to each other as if they were connected by a null modem RS-232 cable.

### 5.3.4 USB

USB (universal standard bus) is a natural evolution of RS-232. Constantly the speed of transmission was increased. Partly due to increased quality of electronics and partly because of limiting the maximum distance of communication. A technological advance was the use of twisted pair combined with differential signals. On two wires the same signal was passed, positive in one wire and

**Table 5.IV**: Some Hayes modem setup commands

| Command | Meaning |
|---|---|
| `ATZ` | Reset modem |
| `ATH0` | Hang up phone (connection will be closed) |
| `ATH1` | Pick up phone (dial tone will be heard) |
| `ATA` | Attend incoming call |
| `ATDxxxxx` | Dial number xxxxx |
| `ATDT` | same as above; dial-by-tones |
| `ATDP` | same as above; dial by-pulses |

negative in the other. Since the interference arrives more or less equal to both wires this interference was canceled out by differential amplification techniques. The twisting of wire pairs ensured wires to be as close as possible to each other to make sure they suffered the same interference and maximum noise cancellation. In the meantime, hardware handshaking was becoming less and less used, in order to save money on the hardware (cables and connectors). Moreover, no male and female connectors are used for the cable. Additionally, in contrast to RS-232, where one wire (TX) is used for sending and the other (RX) for receiving, the same twisted-pair wires are used to both directions of communications. Thus, some kind of collision detection of sorts is needed. Finally, in many cases also some small power was needed for the connected equipment that did not have its own power supply.

Figure 5.13 shows a connector and the pin out. The speed of USB is up to 480 Mb/s for USB 2.0 and 5 Gb/s for USB 3.0. Rivaling technology is the Firewire standard. It differs from USB by having two twisted pairs instead of one. Technically speaking it is no longer serial communication since two bits are sent simultaneously. Other differences are the speed, nearly twice as fast (800 Mb/s) and a high power supply of up to 45 watts if the 6 pin connector is used; for a 4 pin connector, only the two signal twisted pairs are available. Moreover, where for USB connections there has to be a 'master', typically a computer, that controls the communication, the Firewire protocol (also known as (i-Link) is peer-to-peer. Yet, Firewire seems to be disappearing from the market, especially since the coming of USB 3.0.

Other communication protocols that are disappearing rapidly, but still find some use in electronic instrumentation interfacing are the parallel port (also known as Centronics for legacy printers) and GPIB. The latter is a twisted-pair type of parallel communications, with 8 bits sent simultaneously. Yet, this technology is prohibitively expensive and for that reason rapidly disappearing from the market. The reason why it is not discussed here.

### 5.3.5 OBD; On-board diagnostics

One important rapidly developing area of information communication is car. While not going into large depth here – also for lack of personal experience of

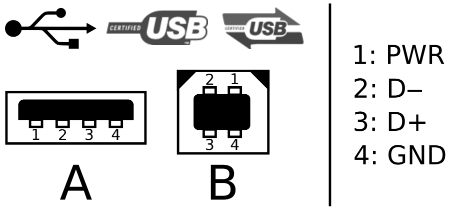


**Fig. 5.13**: USB. Three logos, two types of connectors (looking into the computer [A], or equipment [B]) and the pin numbering. It consists of a twisted pair (D+/D−) on which bidirectionally data is sent, a power line supplying 5 volt, and a ground

the author – it is worth to mention here. Moreover, all topics discussed in this book are relevant for automotive industry.

OBD, or on-board diagnostics, is a communication standard developed in the beginning of the century. All European cars have to have OBD-II or EOBD (European OBD) facilities. The idea is that the car is full of tiny processors, for the engine management, airbags, speed, etc, that all communicate over a bus (for example CAN, controller-area network bus) with a central processor (on-board computer) or even with each other. The local bus is similar to RS-232, in that it is a serial communication protocol, but has an essential difference, namely that the communication is many-to-many, there where RS-232 is one-to-one. In CAN, all devices can communicate on the same bus, at the same time. The one with highest priority gets access to the bus the first.

Since the automotive industry is very large, anything that is standard in that area, comes out very cheap. Devices exists that interface between CAN (and other standards of communication in cars) and RS-232. Normally, in modern packages, this device, ELM327, is accompanied by a FTDI (RS-232/USB) interface described before to result in an overall USB/CAN interface. Once connected to our computer on one side and the CAN bus (via a OBD-II connector) on the other side, we can communicate with all the sensors and processors of our car, see Figure 5.14.

## 5.4 Arduino

In this section we will use an external small processor to create small interfacing & stand-alone informatics projects. These are of the PIC type (peripheral interface controller), small user-programmable processors. More specifically, we will use hardware of the Arduino family. The Arduino is an Open Source hardware project that manages to pack a lot of computing and processing power in a small and cheap package. An engineers and hobbyist dream. It is a versatile platform that allows for interesting and powerful extensions and interfacing with hardware raging from LCD displays to stepper motors and can also communicate to a computer through a modern USB port, apart from being able

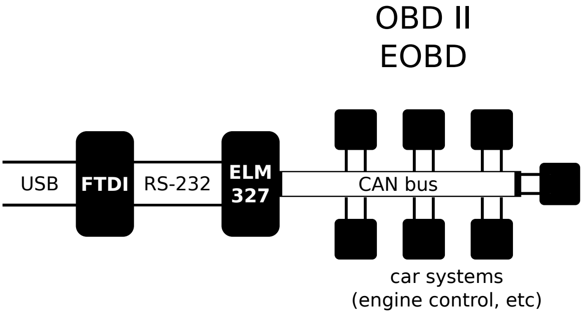


**Fig. 5.14**: OBD (on-board diagnostics) system including an interface (ELM327) connecting CAN to RS-232 and an FTDI interfacing RS-232 with USB

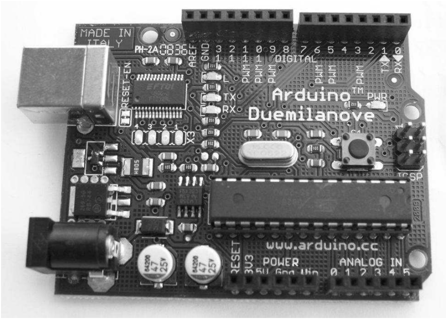


**Fig. 5.15**: A typical Arduino board

to work as a stand-alone device. As we will see, it can easily be expanded and configured to be full blown Internet server.

To start, we have to buy an Arduino board, which should cost around 20 euros/dollars, and the fun can begin. On the pages of Arduino (`http://arduino.cc`) we can download the IDE (Integrated Development Environment) for our operating system, the environment we will use to write, compile and upload or programs. On these pages we can see the Open Source character of the project; we can even find the EAGLE layout of the board, in case we would want to make the Arduino board ourselves (and we are then even allowed to sell it).

A basic Arduino board is show in Figure 5.15. It is contains the following components:

- Micro-controller. This is the CPU of the board where we can put our code that will be running. The latest version of the Arduino (Uno) has an ATmega328, which is an 8 bit processor (PIC, peripheral interface controller) running at 20 MHz and with a program memory of 32 kB and 1 kB of EEPROM memory space. Moreover, it has all the goodies that are useful for our projects: timers, ADCs, (pseudo)DACs, serial communication, etc; pins for this can be recognized on the Arduino board, see below. Code for the ATmega328 can be written in a C++-style of code on a computer in either Windows or Linux. The programs can then be compiled and uploaded to the processor through the USB bus.

- Reset Button.

- Interfacing and power pins. A row of 16 interfacing pins on one side and 12 on the other side

- USB Port & FTDI™ chip. The USB allows for the Arduino port to be connected to a computer (on Ubuntu Linux, the USB port is visible in the directory `/dev` under a name like `ttyUSB0`. In Windows it becomes a virtual COM port. In the Arduino IDE we can select these ports). The USB connection has a dual functionality. Primarily it is a way to put the written program on the processor. Secondarily, it is a way to have the program running on the processor communicate with the computer. The Arduino can be a stand-alone device, but if we want to export measurement data to an external computer, we can do it through this USB port. Note that, on the side of the Arduino, serial communication is done by the more-old-fashioned RS232 protocol. An FTDI™ (Future Technology Devices International) chip translates the RS232 signals to USB signals. Just behind the FTDI™ chip we can see two LEDs labelled 'TX' and 'RX' that show the serial transmission of information. Also note that the send and receive lines are available as digital pins (0 and 1), in case we want to communicate with other peripherals instead of the computer.

- Voltage regulator. As the name implies, this device regulates the voltages to levels useful for the hardware. Any voltage supply is transformed into two regulated voltages, 5 volt and 3.3 volt respectively. These are available on the power pins (see below). A power LED on the other side of the board shows the board is working. Older boards (for example Diecimila) had a jumper to chose where the power comes from, from a power supply or through the 5 volt USB power. Modern boards have an automatic selector. We can also decide to supply the power directly to the power pins (see below), 'Vin' and 'Gnd' instead of the power jack or via USB.

- Pins. The Arduino board has 28 power and signal pins, as shown in Figure 5.16. They are distributed over four connectors and divided into three functionalities: power, analog input (ADC) and digital I/O. Analog output (DAC) can be emulated with the digital pins.

  - Power pins. The regulated power voltages are made available to us on pins '3V3', '5V' and 'Gnd', respectively 3.3 V, 5 V and ground. The power supply voltage is also available at 'Vin'. If we want to supply the power ourselves, we can input it on this 'Vin' pin (and 'Gnd'). Ground is also available on the right side of the board.

**Table 5.V**: The ICSP connector pins

| Pin | Mnemonic | Description |
|---|---|---|
| 1 | MISO | Master In Slave Out. Data from the processor (slave) to the PC (master) |
| 2 | VCC | Power supply out (5 V) |
| 3 | SCK | The serial data clock |
| 4 | MOSI | Master Out Slave In. Data from the PC (master) to the processor (slave) |
| 5 | RST | Reset |
| 6 | GND | Ground |

- Analog input. The Arduino has 6 analog inputs that convert signals to 10 bit digital values. The voltage ranges can be programmed to be 5 V, 3.3 V, or the external voltage supplied to the 'AREF' pin.

- Digital I/O. The Arduino has 13 digital pins. Each pin can be programmed to be output or input using TTL (transistor transistor logic), high (5 V) = 1 and low (0) = 0. Two of the pins are shared with the serial communications, pin 0 is also 'RX' ('read'; output from Arduino) and pin 1 is also 'TX' ('send', input to Arduino) resulting in pseudo-RS232 communication (see Figure 5.10). It is better to avoid using these pins for digital I/O, least we need to disconnect the signal lines when we want to upload code to the Arduino.

  The other 11 pins can be programmatically configured as input or output.

- Analog output. Moreover, 6 of the digital pins (indicated by an asterisk, *) can be used to emulate an 8-bit-DAC analog output between 0 and 5 V. This is achieved by the processor through placing $n$ logical ones (and $2^8 - 1 - n$ logical zeros) in a so-called pulse-width modulation (PWN) technique, see Figure 5.17. The repetition frequency is high (about 500 Hz), so for many applications we can consider the output an average (DC) value that is equal to (5 V)$\times(n/255)$. This is a form of implementing an 8-bit DAC by oversampling a 1-bit-DAC as described in Section 5.2.4. The advantage is that the resulting DAC is very linear (limited only by the clock of the processor). The disadvantage is that we do not have a true DC signal, but it has a lot of high-frequency 'noise' instead.

- ICSP port. In Circuit Serial Programming connector. This is a way the processor communicates serially with other boards placed on top of the Arduino. We will not spend much time on this port, but it exists if we want to use it. We have access here to ground and 5 volt power too, if we need it, see Figure 5.18.

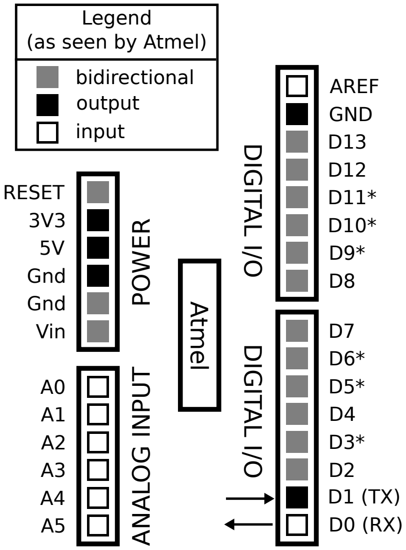


**Fig. 5.16**: Pin configuration of an Arduino board, as seen by the Atmel (an output, like TX transmit, is an output for the Atmel IC). Solid squares are output. Open squares are input. Gray squares are input and/or output, simultaneously for the power pins, or configurable for the digital pins. Digital pins (and on older boards, analog pins as well) are labeled only by numbers on the board; they are named D0, etc, here to be more unambiguous in the text. Digital pins marked with an asterisk (*) can emulate DAC output by placing a PWM code at the pin. Digital pins marked 'TX' and 'RX' are also used for serial communication

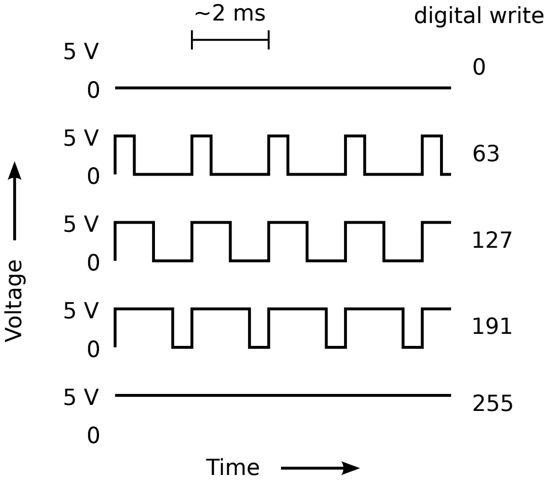


**Fig. 5.17**: Emulation of analog output by placing $n$ logical 1s (5 V) and $255-n$ logical 0s (0). The average voltage is a weighed average of ones and zeros. If needed, the high-frequency (approximately 500 Hz) can be filtered off

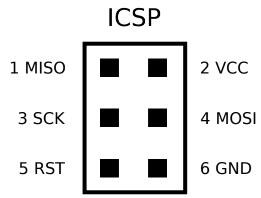
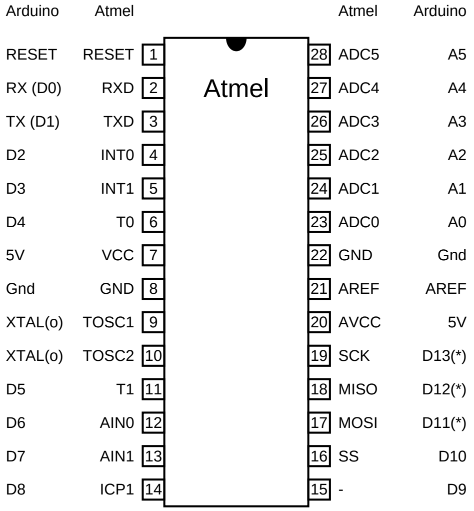


**Fig. 5.18**: The ICSP connector pins. MISO = Master In Slave Out, VCC = power supply (5 V), SCK = Serial clock, MOSI = Master out slave in, RST = reset, GND = ground



| Arduino | Atmel | | | Atmel | Arduino |
|---|---|---|---|---|---|
| RESET | RESET | 1 | 28 | ADC5 | A5 |
| RX (D0) | RXD | 2 | 27 | ADC4 | A4 |
| TX (D1) | TXD | 3 | 26 | ADC3 | A3 |
| D2 | INT0 | 4 | 25 | ADC2 | A2 |
| D3 | INT1 | 5 | 24 | ADC1 | A1 |
| D4 | T0 | 6 | 23 | ADC0 | A0 |
| 5V | VCC | 7 | 22 | GND | Gnd |
| Gnd | GND | 8 | 21 | AREF | AREF |
| XTAL(o) | TOSC1 | 9 | 20 | AVCC | 5V |
| XTAL(o) | TOSC2 | 10 | 19 | SCK | D13(*) |
| D5 | T1 | 11 | 18 | MISO | D12(*) |
| D6 | AIN0 | 12 | 17 | MOSI | D11(*) |
| D7 | AIN1 | 13 | 16 | SS | D10 |
| D8 | ICP1 | 14 | 15 | - | D9 |

**Fig. 5.19**: The Atmel IC and its pins and its corresponding connections on the Arduino. (*) is also available on the ICSP connector, (o) is on the crystal

Although the Arduino is a wonderful platform for studying processor-based applications (embedded systems), and are ideal for lecturing purposes (the main aim of this book), it is not needed to use the Arduino board. For some applications (the final, commercial version?) we can decide to leave out the board and directly connect everything to the Atmel processor. Figure 5.19 shows the IC and its connections

## 5.4.1 The Arduino board as a USB/RS232 adapter

The Arduino board can be used as a USB/RS232 adapter. That is, only the functionality of the FTDI chip can be used. This can be handy if we have a device that communicates through RS2323 and, before connecting it to our Arduino system, we want to test it by connecting it directly to a computer

**Table 5.VI**: The Atmel pins

| Pin | Mnemonic | Arduino | Pin | Mnemonic | Arduino |
|---|---|---|---|---|---|
| 1 | RESET | RESET | 28 | ADC5 | A5 |
| 2 | RXD | Rx (D0) | 27 | ADC4 | A4 |
| 3 | TXD | Tx (D1) | 26 | ADC3 | A3 |
| 4 | INT0 | D2 | 25 | ADC2 | A2 |
| 5 | INT1 | D3 | 24 | ADC1 | A1 |
| 6 | T0 | D4 | 23 | ADC0 | A0 |
| 7 | VCC | 5V | 22 | GND | Gnd |
| 8 | GND | Gnd | 21 | AREF | AREF |
| 9 | TOSC1 | Xtal$^o$ | 20 | AVCC | 5V |
| 10 | TOSC2 | Xtal$^o$ | 19 | SCK | D13/SCK$^*$ |
| 11 | T1 | D5 | 18 | MISO | D12/MISO$^*$ |
| 12 | AIN0 | D6 | 17 | MOSI | D11/MOSI$^*$ |
| 13 | AIN1 | D7 | 16 | SS | D10 |
| 14 | ICP1 | D8 | 15 | - | D9 |

o: On Crystal, *: On ICSP socket

and send commands. For this, we have to put the Atmel processor out of play, since we do not want it to receive the commands send by the computer destined for the device. One sure way to achieve this is by taking the Atmel out of its socket (but be careful not to destroy its pins). Another way is by constantly resetting it (connect the RESET pin to ground, GND). Note, however, that now the external RS232 device has to be connected differently, exchanging the TX and RX connections. Figure 5.20 clarifies this. In normal cases, with the RS232 device connected to the Arduino with Atmel, the Atmel sends (outputs) commands to the device on the TX pin and receives replies on the RX pin. The computer, however, connected to the USB/FTDI port sends commands on the RX pin and receives replies on the TX pin. We will later see an example of this with the XBee wireless communication shield where this is achieved by jumpers. Note that, in case both a computer is connected through the FTDI chip and a device is connected to the Arduino boards communications pins RX and TX, the external device 'wins' the communications conflict, i.e., the Atmel no longer receives information from the computer, but can still output to it.

### 5.4.2 Programming the Arduino; First examples: digital output and serial communication

The Arduino IDE (integrated development environment) uses a C++-like programming language that is then compiled into Atmel-understandable binary code which can be uploaded to the processor. We can get the Arduino IDE at `http://arduino.cc`, or in Ubuntu in the Ubuntu Software Center or any

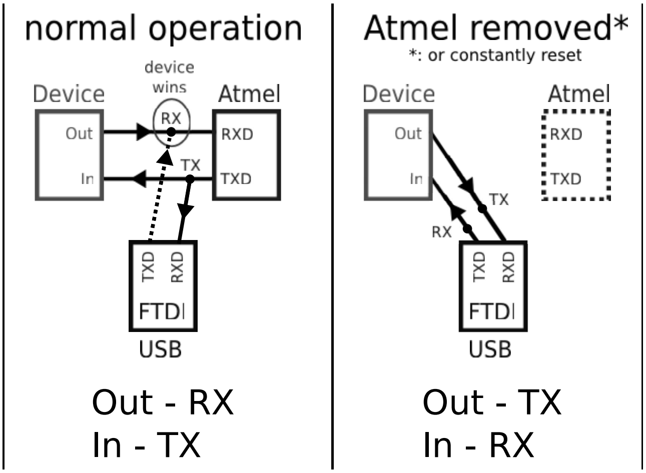
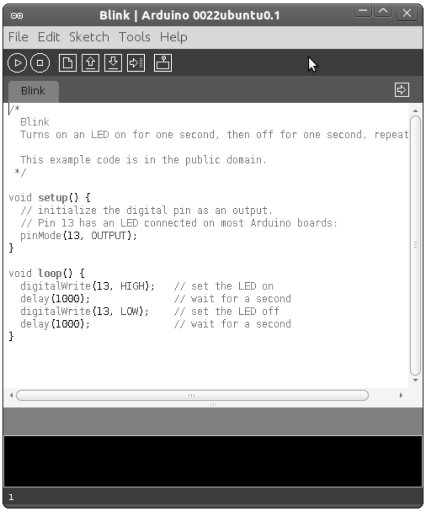


**Fig. 5.20**: The Arduino board as a USB/RS232 adapter. In normal operation the Atmel communicates with an external device through the serial com pins TX and RX. If we want to have the device directly communicate with a computer, we have to take the Atmel out of the game by physically removing it or resetting it constantly. Note the changing of the connections of the device when switching between Atmel and computer communication. Note also that if both computer and device communicate (send) to the Atmel, the device wins (while both can receive information from the Atmel simultaneously)

package manager. When we run it we see the editor in front of us, as can be seen in Figure 5.21, where a simple example program `Blink` is already loaded. The buttons on the top of the screen are for respectively: Compile (program in the editor), Stop (compiling), New (program), Load (program from file), Save (current program to file), Upload (compiled program to Arduino) and Serial Monitor. The black part at the bottom of the screen is for messages from the IDE, like compile errors, etc.

**Exercise**: Load the program `Blink`, compile it and upload it to the Arduino board. The program can be found under `File:Examples:1.Basics:Blink`. Make sure you select the correct Arduino board (including correct processor) and (virtual) COM port, both available under pull-down menu `Tools`. If you managed, the LED close to pin 13 should be blinking once every two seconds.

Any Arduino program consists of an initialization part called `setup()` and a routine `loop()` that is repeated over and over again. First the initialization routine is executed and then the loop is executed *ab-infinitum*, not ending until the power is removed from the device, see Figure 5.22. In either routine we place our code, including calls to new routines we placed outside these two routines.



```
/*
  Blink
  Turns on an LED on for one second, then off for one second, repeat

  This example code is in the public domain.
 */

void setup() {
  // initialize the digital pin as an output.
  // Pin 13 has an LED connected on most Arduino boards:
  pinMode(13, OUTPUT);
}

void loop() {
  digitalWrite(13, HIGH);   // set the LED on
  delay(1000);              // wait for a second
  digitalWrite(13, LOW);    // set the LED off
  delay(1000);              // wait for a second
}
```

**Fig. 5.21**: The Arduino IDE with an example program `Blink` loaded in the editor

The code for blink is

```
/*
  Blink
  Turns on an LED on for one second, then off for one second, repeatedly

  This example code is in the public domain.
*/

void setup() {
  // initialize the digital pin as an output.
```

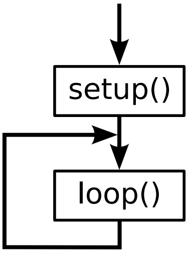


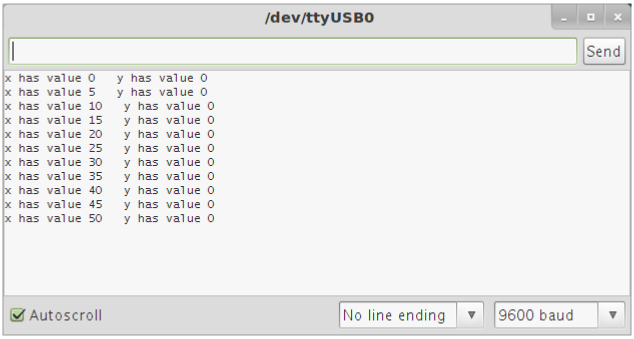
**Fig. 5.22**: The basic structure of an Arduino program consists of an initialization routine setup() that is executed once and a routine loop() that is repeated over and over again. In these routines we can place our code, including calls to routines outside these two

```
  // Pin 13 has an LED connected on most Arduino boards:
  pinMode(13, OUTPUT);
}

void loop() {
  digitalWrite(13, HIGH);   // set the LED on
  delay(1000);              // wait for a second
  digitalWrite(13, LOW);    // set the LED off
  delay(1000);              // wait for a second
}
```

A complication of programming an external processor of embedded systems is that we cannot easily debug the program. In most conventional IDEs (integrated development environment) we can run the program step-by-step,or place breakpoints at strategic points and view the state of the memory, i.e. the values of variables. Since th Arduino is running outside our computer, we do not have this facilities. We therefore have to fall back to good-old debugging techniques. In the developing versions, we let the program generate informative output at crucial places. This can be as simple as switching on a LED, but the most informative signals are text, for instance the values of variables, directly communicated to the programmer through the serial communications port. See the example below:

```
void setup() {
  // Set up the serial communication port
  Serial.begin(9600);
}

int x;

void loop() {
  int y;

  Serial.print("x has value ");
  Serial.print(x);
  Serial.print("   y has value ");
  Serial.println(y);
  delay(1000);
  x = x + 5;
  y = y + 5;
}
```

When the program is loaded into the Arduino and we open the serial monitor, we see the output as shown in Figure 5.23. We now see the maybe unexpected difference in behavior of global variables and local variables. Global variables in the Arduino are defined outside any procedure and they keep their



```
x has value 0   y has value 0
x has value 5   y has value 0
x has value 10   y has value 0
x has value 15   y has value 0
x has value 20   y has value 0
x has value 25   y has value 0
x has value 30   y has value 0
x has value 35   y has value 0
x has value 40   y has value 0
x has value 45   y has value 0
x has value 50   y has value 0
```

☑ Autoscroll    No line ending ▾    9600 baud ▾

**Fig. 5.23**: Output of program to the serial port to help us debugging the program. Note that this way we found out the difference in behavior between a global variable (x) that maintains its existence and value once the procedure loop is finished and a local variable (y) that is destroyed when ending the procedure and recreated and initialized when entering the procedure again

existence and value forever. Local variables are defined inside a procedure and they have their 'scope' limited to this procedure. Once the procedure ends (for example loop), the variable is destroyed. It is recreated and initialized when the procedure is entered again. We found this out by outputting values of variable values to the serial port.

One of the major advantages of the Arduino platform is that there already exist many examples available and the Open-Source character of the Arduino project means that there are many people out there that are trying to help us and contribute to the general advance of the knowledge. Whatever you are trying to do with your Arduinos, somebody probably already did it and published it on the internet.

An example is the conversion of two digital pins to become a serial port as well. This is handy when we have to communicate to two pieces of equipment, or if we want to maintain the possibility to communicate with the computer as well, apart from communicating to external equipment. In these situations we can make use of the SoftwareSerial program code. An example is given here below, where digital pins 2 and 3 are converted to serial ports RX and TX, respectively:

```
#include <SoftwareSerial.h>

//Creates a software serial port. (RX, TX)
SoftwareSerial mySerial(2, 3);

void setup()
{
  //Initialize serial port for communication.
  mySerial.begin(9600);
}

void loop() {
  mySerial.println("Hello World");
  delay(2000);
}
```

Note however that the software serial ports are not as good as the standard (pins-0,1) serial port. In some cases you may need to do some signal preparation (pull-down resistor, opamp buffering) to get clean communication.

### 5.4.3 Arduino extensions

A variety of hardware can be added to the Arduino. We can use the board to make robotics projects or signal processing tasks. We can even make it into an internet server if we want to. The most interesting Arduino extensions boards – 'shields' as they are more commonly known – are

- LCD screen. Typically a 2x16 (dual line, 16 characters per line) dot matrix screen. They come in two types. The simple type uses 4 data lines and 2 control lines, apart from power lines. That means that we will have to sacrifice 6 of our (precious) digital pins. An alternative is to use one that has an incorporated serial communications port which means we have to sacrifice our serial communications capabilities.

- Ethernet and SD-card-reader shield. These are useful if we want our Arduino to communicate through the internet or if we want to make a data logger. Or both. The SD card reader can address up to 4 GB of data, which is an immense increase from the kilobytes of data that can be stored inside the Atmel processor. To give an idea, if we take one floating-point measurement per second, we can make a datalogger that stores data for 40 years. And by that time, for sure, there will be shields with more capacity.

- XBee shield. An XBee chip of Digi implements the ZigBee protocol of communication. The XBee chip receives serial data (for instance from our serial pins of our Arduino) and communicates them wirelessly to another XBee chip that transmits them subsequently to a serial listener (for instance another Arduino). A shield is available to easily place the XBee on top of an Arduino, however, this shield is not essential and you can save the 20 euros or so. We could also opt for using ICs with the same functionality from Nordic

- Sensor shield. Most sensors and actuators come in the three-wire form: 'power', 'ground' and 'signal'. Commercial sensors and actuators have standardized connectors for this reason. An Arduino Sensor Shield has the 6 analog and 14 digital pins easily accessible by such three-pin connectors and the serial communication by a four-pin connector. This keeps our work organized. Once again, the shield is not essential, but can be handy anyway.

## 5.5 Examples of (cool) Arduino projects

Here we show some examples of projects. But remember that the possibilities with Arduino are sheer infinite. Just look on the internet; it is likely that you will get the answer to your project specifications or get inspiration for cool projects. The projects described here are just to get some inspiration and are as independent as possible.

- Communication, LCD display
- Temperature sensor & Controller
- Data logger
- Internet server
- Robotics (Stepper and servo motors)
- Sun dial
- Radio controlled clock
- Power-line controlled clock
- Programmable light dimmer
- GPS controlled clock
- Barcode reader
- Harddisk clock
- Remote control
- RFID

They will be described here separately. Of course, you can combine projects and make a radio-controlled internet time server, or a wirelessly-controlled robot. Or anything that you can come up with.

**Table 5.VII**: Serial communication functions for the Arduino. See the on-line pages at `http://arduino.cc` for more details and examples

| Function | Param. | Return | Description |
|---|---|---|---|
| `Serial.begin()` | baud rate | | Setup communication |
| `Serial.end()` | | | Stop communication |
| `Serial.available()` | | int | Returns the number of bytes available to read |
| `Serial.read()` | | int | read character (-1 if none available) |
| `Serial.peek()` | | int | Like read, but does not remove data from stream |
| `Serial.flush()` | | | Write buffer |
| `Serial.print()` | data | | Send string |
| `Serial.println()` | data | | Send string + CRLF |
| `Serial.write()` | data | | Like `print`. Send int values |
| `Serial.SerialEvent()` | | | Interrupt routine when data available |

## 5.5.1 Arduino serial communication & LCD display

We start with a small project that can be used for many of the projects described later. For many applications we can send the useful information via the RS232 channel to the computer by the instructions of the Serial object. This has the following functions as shown in Table 5.VII.

The nice thing about the Arduino is that we can use not only a tethered application, with the Arduino connected to the computer, but we can also make a stand-alone application. In this case, if we want somehow, we can use an LCD display. The standard LCD display for the Arduino is a 16-pin 16x2 (2 lines, 16 characters per line) dot matrix (7x5 pixels per character) display. Apart from the 5 volt power and ground connections (pins 1, 2, 15, 16) it uses 4 data pins for the data (pins 7, 8, 9 and 10), 2 lines for control (pins 4 and 6) and one line for contrast control (pin 3), something that can be done by a potentiometer selecting a voltage somewhere in between 0 and 5 V. See Figure 5.24. The Arduino software IDE already comes with the adequate libraries for such displays and we just have to include it and specify to what digital pins the display is connected.

A tradition in informatics is to write your first program that outputs 'Hello World'. The equivalent of this in Arduino is a program that writes it on an LCD display and sends it through the USB port to the computer:

```
/*******************************************\
 *   Generating output to USB              *
 *    and to an LCD display                *
\*******************************************/
```

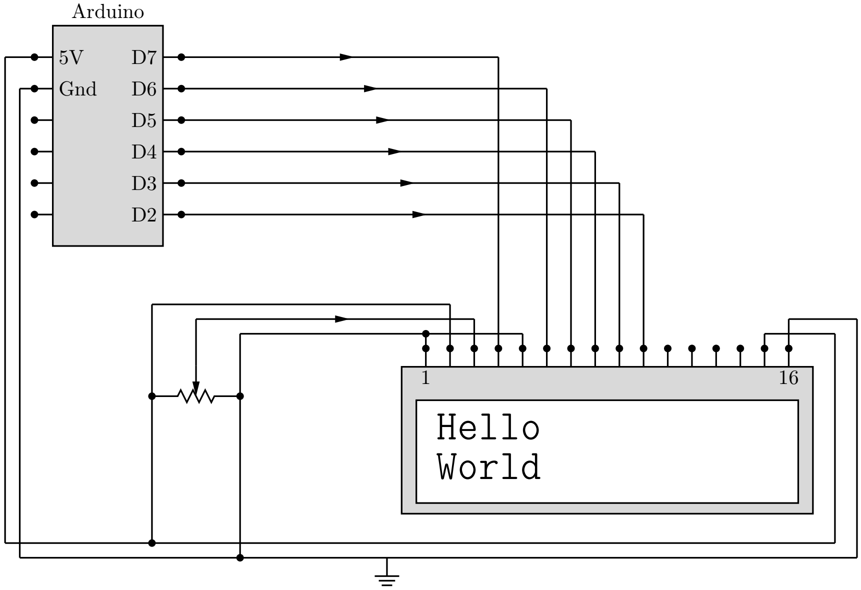


**Fig. 5.24**: Circuit for driving an LCD display

```
#include <LiquidCrystal.h>

// specify how the LCD is connected to Arduino:
LiquidCrystal lcd(7, 6, 5, 4, 3, 2);

void setup(){
  // setup serial communication:
  Serial.setup(9600);
  // set up the LCD's number of columns and rows:
  lcd.begin(16, 2);
}

void loop(){
  // output message to RS232/USB:
  Serial.println("Hello World");
  // output message to LCD display:
  lcd.setCursor(0, 0);
  lcd.print("Hello       ");
  lcd.setCursor(0, 1);
  lcd.print("World       ");
  // wait 2 seconds
  delay(2000);
}
```

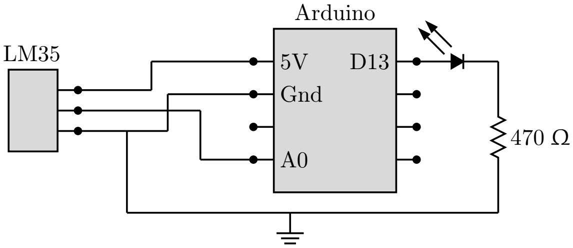


**Fig. 5.25**: Example of a thermometer with warning LED based on an LM35 sensor

## 5.5.2 Arduino temperature sensor

We can directly connect sensors that transduce physical parameters to voltage signals to the 6 ADC inputs. As an example we can use the LM35 temperature sensor. The circuit will light an LED when the temperature rises above 30 degrees and switches it off when the temperature drops below 20 degrees. The circuit for this is given in 5.25.

The program is given below. The LM35 temperature sensor has a sensitivity of $S$ = 10 mV/$^o$C. The ADCs of the Arduino are 10 bit ($2^{10} - 1 = 1023$ steps) and have a range of 0 to 5 volt, giving a digital resolution of $\Delta V$ = (5 V)/1023 = 4.89 mV. This translates into a temperature resolution of $\Delta T = \Delta V/S$ = (4.89 mV)/(10 mV/$^o$C) = 0.489 $^o$C, which is quite low. Fortunately the noise is quite large and we can make use of some kind of oversampling technique. This is accomplished by averaging. In the example a 1000-times averaging is used. This increases the resolution to half a millikelvin. The program also includes hysteresis effects. This is placed in a separate routine to also show how to write routines apart from `setup()` and `loop()`.

```
/*****************************************\
 *   Thermometer with a 20-30 degrees    *
 *    hysteresis warning LED             *
\*****************************************/

void hysteresis(float t){
  if (t>30.0)
    digitalWrite(13, HIGH);
  else if (t<20.0)
    digitalWrite(13, LOW);
}

void setup(){
  pinMode(13, OUTPUT);
}

void loop(){
  long int i;
  long int avg = 1000;
  long int adcsum = 0;

  // oversampling avg times:
  for (i=0; i<avg; i++){
    adcvalue = analogRead(0);
    adcsum += adcvalue;
  }
  //    0 = 0 volt = 0 degrees
  // 1023 = 5 volt = 500 degrees
  sensorvalue = 500.0*((float) adcsum) / 1023.0;
  sensorvalue /= (float) avg;
  hysteresis(sensorvalue);
}
```

### 5.5.3 Arduino Ethernet server for temperature sensor

The following project is using an Ethernet shield. The shield will function as a 'full-blown' HTML web server. That is, It waits for a request. Every character received is also echoed to the serial line, so that we can follow the internet activity.

```
/*******************************************************\
 *               Ethernet shield program               *
 * When called it shows the value of the temperature  *
 * measured with an LM35 temperature sensor connected *
 * to analog pin 0.                                    *
 *                                                     *
 * Based on program by Mellis and Igoe                 *
\*******************************************************/

// Ethernet shield libraries:
#include <SPI.h>
#include <Ethernet.h>

// Enter a MAC address and IP address for your controller below.
byte mac[] = { 0xDE, 0xAD, 0xBE, 0xEF, 0xFE, 0xED };
byte ip[] = { 10,10,21,97 };

// Initialize the Ethernet server library
// (port 80 is default for HTTP):
Server server(80);

void setup()
{
  // start the Ethernet connection and the server:
  Ethernet.begin(mac, ip);
  server.begin();
}

void loop()
{
  float t;

  // listen for incoming clients
  Client client = server.available();
  if (client) {
    // an http request ends with a blank line
    boolean currentLineIsBlank = true;
    while (client.connected()) {
      if (client.available()) {
        char c = client.read();
        // if you’ve gotten to the end of the line (received a newline
        // character) and the line is blank, the http request has ended,
        // so you can send a reply
        if (c == ’\n’ && currentLineIsBlank) {
          // send a standard http response header
          client.println("HTTP/1.1 200 OK");
          client.println("Content-Type: text/html");
          client.println();

          client.print("Temperature ");
          // LM35: 0.01 V/degree, ADC: 5 V = 1024
          t = (5.0*((float) analogRead(0))/1024.0)/0.01;
          client.print(t);
          client.println("<br />");
          break;
        }
        if (c == ’\n’) {
          // you’re starting a new line
          currentLineIsBlank = true;
        }
        else if (c != ’\r’) {
          // you’ve gotten a character on the current line
          currentLineIsBlank = false;
        }
      }
    }
    // give the web browser time to receive the data
    delay(1);
```

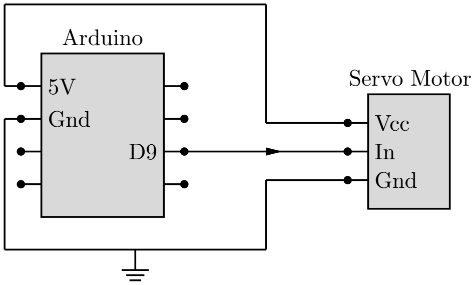


**Fig. 5.26**: Circuit for connecting a servo motor to an Arduino

```
    // close the connection:
    client.stop();
  }
}
```

### 5.5.4 Arduino robotics. Stepper motors & servo motors

In this project we will create mechanical movement by using motors. For this purpose we have two types of motors: stepper motors and servo motors. Which ones are better depends on the application we want. A servo motor is used when we want to program an absolute angle of the axis – 'rotor' – of the motor (and object connected to it). It normally has a high resolution but limited typically to one full turn. (See Chapter 3). A stepper motor, on the other hand, programs only a relative position, we can program a step of the rotor, without knowing the absolute position of axis. It can rotate in the same direction forever. If we want to have knowledge of the absolute position of an object with a stepper motor we have to let it pass over a calibration position and from there count the number of steps. See Table 4.V of Chapter 3 for a comparison of stepper and servo motors.

Using a typical servo motor is very simple. For a simple DC servo motor all we have to do is feed a voltage between the 0 and $V_{\mathrm{cc}}$ to the input pin of the servo motor. This can be done by a pseudo-DAC of the Arduino, namely one of PWM digital pins, by using the `analogWrite()` routine. The setup is shown in Figure 5.26. The resolution of the standard PWM of the Arduino is quite low, 8 bit gives only 20 mV voltage resolution.

For more advanced servo motors that actually require the signal by the PWM technique, we have to supply the information of desired angle through a digital pulse of defined width. A sample code is given below:

```
/****************************************\
 *   Servo motor (PWM)                   *
\****************************************/

int servopin = 9;
```

```
void setup(){
  pinMode(servopin, OUTPUT);
}

void loop(){
  float degrees;
  int pulsewidth;
  int i;

  for(degrees = 0.0; degrees <= 180.0; degrees += 0.5)
  {
    pulsewidth = 544 + (int) ((2400.0-544.0)*((float) degrees)/180.0);
    digitalWrite(servopin, HIGH);
    delayMicroseconds(pulsewidth);
    digitalWrite(servopin, LOW);
    delayMicroseconds(16000-pulsewidth);
  }
  for(degrees = 180.0; degrees >= 0.0; degrees -= 0.5)
  {
    pulsewidth = 544 + (int) ((2400.0-544.0)*((float) degrees)/180.0);
    digitalWrite(servopin, HIGH);
    delayMicroseconds(pulsewidth);
    digitalWrite(servopin, LOW);
    delayMicroseconds(16000-pulsewidth);
  }
}
```

For the Arduino specific code is written for servo motors that use this PWM technique, but written in interrupt-driven code, that thus liberates the processor for doing other tasks. It uses a repetition frequency of only 50 Hz instead of the 500 Hz of the standard PWM of the Atmel, and a pulse-time resolution of one microsecond. 50 Hz implies a PWM pulse width of anything between 0 (0) and 20,000 µs (5 V), this is thus effectively a $\log_2(20000) = 14$-bit DAC. For a standard servo motor, an angle of 0° corresponds to 544 µs and 180° to a pulse width of 2400 µs. To make use of this functionality, we have to include the servo library (`Servo.h`), after which we can attach a servo motor with the library object method `Servo.attach(pinnumber)`. Changing the position of the motor is then done by the `Servo.write(angle)` method.

### 5.5.5 Arduino sun dial

We are going to use it in a system where an arrow always points towards the sun. This means that we first have to calculate the position of the Sun in the sky. This is more difficult than it seems. Because of the irregular orbit of the Earth and the earth axis, the position of the sun in the sky is not a regular path.

The code below gives the Arduino routine for calculating the azimuth (angle from South) and altitude (angle from horizon) of the position of the sun in the sky for any time and for any position on Earth. Figure 5.27 gives an example for the calculation of the position of the Sun in the sky at our university and shows the complexity of the movement of the sun (in the sky). If we determine the position every day at the same hour, an '8' results, which is a so-called analemma.

```
void AziAlti(double y, double m, double d, double h, double mins,
             double north, double east, double *azi, double *alti){
/*****************************************************************
 *   Finds position of the sun in the sky: azimuth and altitude  *
 *   returned in variables azi and alti                          *
 *    azi: 0=South, -90=East, 90=West                            *
 *    alti: 0=horizon, 90=Zenith                                 *
 *   Input: y=year, m=month, d=day, h=hour, mins=minutes         *
 *          north=latitude, east=longitude (east=positive)       *
 *             unit: degrees                                     *
 *****************************************************************/
  double L, xx, yy, a, g, t, epsilon, alpha, beta, lambda, theta0,
        delta, theta, tau, alt, az;
  double pi = 3.14159265358979323846;

// PART I: Calculate Sun position in firmament
// (based on http://www.stargazing.net/kepler/sun.html
//      by Keith Burnett)
  h = h + mins/60.0;
//1. Find the day:
  d =  367 * y - floor(7 * (y +floor( (m + 9)/12))/ 4) +
   floor(275 * m / 9) + d - 730531.5 + h / 24;
//2. Find the Mean Longitude (L) of the Sun
  L = 280.461 + 0.9856474 * d;
  while (L<0)
    L = L+360.0;
//3. Find the Mean anomaly (g) of the Sun
  g = 357.528 + 0.9856003 * d;
//4. Find the ecliptic longitude (lambda) of the sun
  lambda = L + 1.915 * sin(g*pi/180.0) + 0.020 * sin(2*g*pi/180.0);
//5. Find the obliquity of the ecliptic plane (epsilon)
  epsilon = 23.439 - 0.0000004 * d;
//6. Find the Right Ascension (alpha) and Declination (delta) of  the Sun
  yy = cos(epsilon*pi/180.0) * sin(lambda*pi/180.0);
  xx = cos(lambda*pi/180.0);
  a = (180.0/pi)*atan(yy/xx);
  if (xx < 0)
```

```
    alpha = a + 180.0;
  else {
    if ((yy < 0.0) && (xx > 0.0))
      alpha = a + 360;
    else
      alpha = a;
  }
  delta = (180.0/pi)*asin(sin(epsilon*pi/180.0)*sin(lambda*pi/180.0));
// PART II: Calculate Sun position in sky
// (based on http://www.geoastro.de/elevaz/basics/
//    by Juergen Giesen
// compute Sidereal time (in degrees) at Greenwich:
  t = d/36525.0;
  theta0 = 280.46061837 + 0.98564736629*d + 360.0*(d-floor(d))
   + 0.000387933*t*t - t*t*t/38710000.0;
  while (theta0<0.0)
    theta0 = theta0+360.0;
  while (theta0>360.0)
    theta0 = theta0-360.0;
  theta = theta0 + east;
  tau = theta - alpha;
  beta = pi*north/180.0;
  delta = pi*delta/180.0;
  tau = pi*tau/180.0;
  alt = asin(sin(beta)*sin(delta) + cos(beta)*cos(delta)*cos(tau));
  az = atan2(-sin(tau), (cos(beta)*tan(delta) - sin(beta)*cos(tau)));
  *alti = 180.0*alt/pi;
  *azi = 180.0*az/pi-180.0;
  if (*azi<-180.0)
    *azi = *azi+360.0;
  if (*azi>180.0)
    *azi = *azi-360.0;
}
```

For a simple system we can use this to open and close the shutters of our house the moment the first rays of sunshine hit the windows. The sun has a finite diameter and thus the sun comes up some moments before the altitude is zero; the upper limb of the Sun's disk is just touching the horizon when its center has an altitude of $-0.833^{\circ}$.

```
void loop(){
  double alti, azi;
  int state;
  double y, m, d, h, mins, secs, north, east;
```

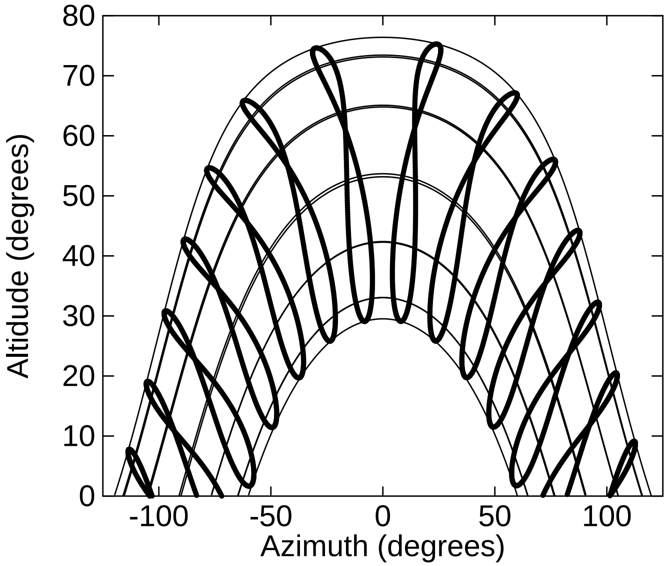


**Fig. 5.27**: Position of the Sun in the sky during 2011 at our university (37.028° North, 7.972° West). Every '8' (a so-called analemma) is the position of the Sun at a certain hour of the day along the year. The arcs are the position of the sun along the 21st day of every month. Calculated with an Octave version of the Arduino code given here

```
  DetermineTime(&y, &m, &d, &h, &mins, &secs);
  DeterminePos(&north, &east);
  AziAlti(y, m, d, h, mins, north, east, &azi, &alti);
  if ((state==0) && (alti>=-0.833)){
    MoveShutters(1);
    state=1;}
  if ((state==1) && (alti<-0.833)){
    MoveShutters(0);
    state=0;}
  delay(1000);
}
```

The determination of the time and position are treated somewhere else in this chapter. The determination of the position can be done in the setup part of the Arduino code (unless you live in an RV). The above code is for a winter situation (closing the shutters at night, open in the day time to let in light). For summer settings, to block out soaring heat and radiate heat at night, we can invert the state. The movement of the shutters itself is something more complicated since it involves power electronics. We can use a relay to operate the shutter engine (servo motors and stepper motors are less adequate since they are not power elements).

### 5.5.6 Arduino radio-controlled clock

In this section we will build a clock based on the radio synchronization signal emitted by the 77.5 kHz radio wave originating from Mainflingen in Germany (close to Frankfurt am Main), named DCF77. The 25 kW emission reach of this signal is some 2000 km, so in most part of Europe this clock should work. More or less from Moscow to the entire Iberian peninsula. From Greece to the edges of Sweden. In practice, hundreds of millions of clocks are synchronized by this signal. Equivalent signals exist in other parts of the world. Find your own closest antenna and necessary hardware through an on-line search. The text given here is specifically for the Mainflingen signal.

The information is transmitted once per minute in a 59-bit binary-coded-decimal string. (BCD: a decimal digit from 0 to 9 is coded by 4 bits 0000 to 1001, the six other patterns from 1010 to 1111 are not used). These bits have the meaning as shown in Table 5.VIII.

The bit pattern is sent over the 77.5 kHz carrier one bit per second by pulse-width encoding, see Figure 5.29. At the beginning of each second, the amplitude of the carrier is reduced by about 25%. For a logical 1 this pulse is longer (0.2 s; 15500 cycles) than a logical 0 (0.1 s; 7750 cycles). The rest of the second – 0.8 s and 0.9 s respectively – the carrier is back up to 100% of its amplitude. This way, a bit is transmitted for the seconds 0 unto 58 of each minute. During the 59th second no bit is encoded. This way, once per minute the actual time and date is transmitted.

Later, phase modulation was added while maintaining backward compatibility. During the full-power 'idle' part of a 1 second time frame, 512 rapid bits are placed by the phase-modulation technique. Every logical one has a different phase – about 26° – compared to a logical zero. With each rapid bit taking exactly 120 cycles of the carrier, the total bit sequence takes 61440 cycles, about 0.793 s. The bit sequence starts at 0.2 s and ends at 0.993 s, i.e. separated from the amplitude-pulse-width modulation. This secondary bit sequence itself is a (pseudo)random bit pattern. For a logical zero of the DCF77 signal, the bit pattern is maintained, for a logical 1 it is reversed. (We know the non-inverted bit sequence because the DCF77 string always starts with a logical zero, see Table 5.VIII).

This rapid bit pattern has as many zeros as ones, thus maintaining the overall central frequency at exactly 77.5 kHz, which allows for the continuation of the use of the radio signal itself as a frequency standard, it's original purpose; the deviation is less than $10^{-12}$ per day and less than $2 \times 10^{-13}$ per 100 days, a resolution that is achieved due to the fact that it is linked to an atomic clock in the Physical Institute in Braunschweig.

Quite complicated. Theoretically we should be able at this point to make a receiver ourselves, see the chapter on electronic oscillators (we make a resonator circuit at $\omega = 1/\sqrt{LC}$, rectifier, etc.). However, we can avoid all confusion. To acquire the bit pattern we can use a commercial transducer. They are cheap

**Table 5.VIII**: The 59 DCF77 signal bits sent at a rate of 1 b/s, with an example for Saturday 27 August 2011, 13:43 = '00000000000000000010111000011110010111100101100010110100000'

| Bits | Example | Meaning (when set to 1) |
|---|---|---|
| 00 | 0 | Start of minute, always 0 |
| 01-14 | 00000000000000 | Civil warning bits |
| 15 | 0 | Abnormal operation |
| 16 | 0 | Daylight-saving time imminent (one hour before actual change) |
| 17 | 0 | Daylight-saving time effective |
| 18 | 1 | Standard ('winter') time effective |
| 19 | 0 | Leap second imminent (one hour before actual leap second) |
| 20 | 1 | Start bit time code (always 1) |
| 21-27* | 1100001 (43) | Minutes 00-59 |
| 28° | 1 | (Even) parity check of minute bits |
| 29-34* | 110010 (13) | Hours 00-23 |
| 35° | 1 | (Even) parity check of hour bits |
| 36-41* | 111001 (27) | Day of month 01-31 |
| 42-44* | 011 (Saturday) | Day of Week (Monday = 1, Sunday = 7) |
| 45-49* | 00010 (8) | Month number 01-12 |
| 50-57* | 11010000 | Year 00-99 (century is not transmitted) |
| 58° | 0 | (Even) parity check of bits 36-57 |
| 59 | | This bit is not sent (carrier at full 100% amplitude) |

*: LSB first. Bits have weight 1, 2, 4, 8, 10, 20, 40, 80, resp.

o: Equal to value needed to make total number of 1s, including parity, even

enough to buy them ready-made instead, see for example the antenna/receiver in Figure 5.28. What interests us at this point is only the informatics description of Table 5.VIII, and what are the pins of the transducers. For that we have to read the datasheet.

The proposed circuit is given in Figure 5.30. The antenna module is powered by the Arduino through the 5V and Gnd pins. Gnd is also connected to the PON (power on) pin of the antenna module. From the module the demodulated signal (rectified, low-pass filtered and level shifted to 0-5 V) is send to the Arduino. The Arduino will interpret this 1 bit-per-second signal coming in through digital port 3 and send the decoded information to the serial port (USB) or to an LCD screen.

### 5.5.7 Arduino power-line controlled clock

Because of the interconnection of electricity power lines in Europe (and in other regions of the world as well) they need to be synchronized with each other.

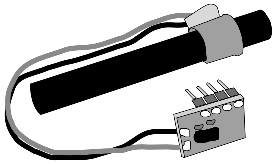
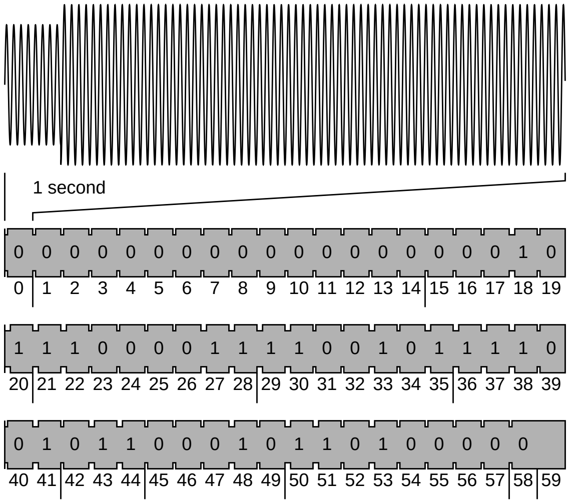


**Fig. 5.28**: A DCF77 antenna and transducer consisting of a ferrite nucleus with windings and a signal demodulator



0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
0 | 1 2 3 4 5 6 7 8 9 10 11 12 13 14 | 15 16 17 18 19

1 1 1 0 0 0 0 1 1 1 1 0 0 1 0 1 1 1 1 0
20 | 21 22 23 24 25 26 27 28 | 29 30 31 32 33 34 35 | 36 37 38 39

0 1 0 1 1 0 0 0 1 0 1 1 0 1 0 0 0 0 0
40 41 | 42 43 44 | 45 46 47 48 49 | 50 51 52 53 54 55 56 57 | 58 | 59

**Fig. 5.29**: Example of a 1-minute DFC77 signal as given in Table 5.VIII, Saturday 27 August 2011, 13:43. The frequency in the 1-second zoom in has been exaggerated a factor 1000 (It is here 77.5 Hz instead of 77.5 kHz). Markers are placed to indicate the various parts of the bit pattern

Imagine what would happen with a coupled grid if one of the centrals is 180 degrees out of phase; they'd be working against each other, each being the energy sink of the other. To avoid power inefficiency, the power cycles are synchronized with each other. As a side effect, the electricity centrals try to synchronize with an atomic clock, to keep the frequency as close to 50 Hz as possible. Ideally they are all working at 50 Hz in Europe, but heavy loads can make the frequency shift a little. In the long run they will make an average frequency of 50 Hz. We can make use of this by counting the number of power-line zero crossings to feed a clock.

The circuit is very simple: we just feed the mains power to a digital pin (D2). To avoid placing our entire Arduino at dangerous high voltages, we use an optocoupler to electrically separate it from the mains network. See Figure 5.31. An optocoupler is light-emitting diode and a photo-transistor in a sealed package. The signal is passed in the form of light and this makes the entrance

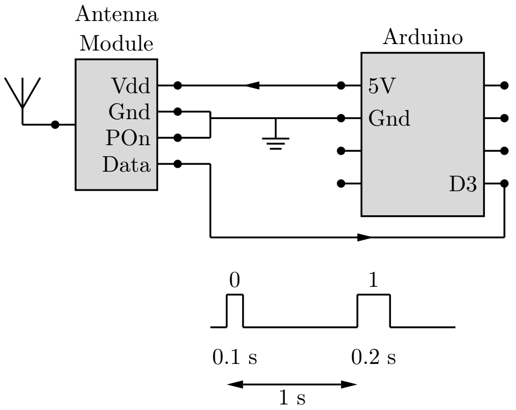


**Fig. 5.30**: Circuit used for radio-controlled clock. The antenna module (as shown in Figure 5.28) is power-supplied with 5 volt from the Arduino. The power-on PON pin is connected to ground. At the data line will appear the demodulated digital signal that can be analyzed by the Arduino

and the exit of the optocoupler electrically separated.

In the circuit the mains voltage is passed through the input of an optocoupler in series with a 1 megaohm resistance to limit the current. For the positive part of the mains cycle, the LED will emit light and the photo transistor opens; the transistor is effectively a short circuit and the output is connected to ground. In the other half of the cycle the transistor is closed and the pull-up resistor of 4.7 kiloohm makes sure the output is 5 volt. The output of the optocoupler is fed to a digital pin (2). In this way, the digital input will be low when the cycle of the mains is positive.

Another novelty we use is interrupts in the programming of the Arduino. The Arduino can be programmed to execute a routine when a certain type of event occurs, in this case when the pin (digital 2, which is interrupt pin 0) goes from high to low, see the code below. This allows for the program to execute other tasks instead of constantly monitoring – 'polling' – the pin.

```
/*******************************************************
 *  Part of program for power-line controlled clock     *
 *  Interrupts on digital pin 2 (interrupt pin 0)       *
 *  when pin goes from high to low and executes routine *
 *  myclock()                                           *
 *******************************************************/
int h=0, m=0, s=0, cs=0;

void setup(){
  pinMode(2, INPUT);
  // interrupt 0 is digital pin 2
  attachInterrupt(0, myclock, FALLING);
```

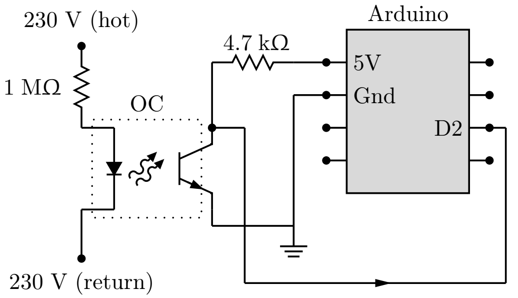


**Fig. 5.31**: A clock synchronized by the 240 volt network. To not have our Arduino on high voltage, we use an optocoupler (OC), to separate it electrically from the network. The signal, that is low when the the power cycle is positive, is connected to digital pin 2. This pin can be used for interrupts, see text

```
}

void myclock(){
  if (++cs==50){ // change to 60 for 60 Hz power line
    cs=0;
    if (++s==60){
      s=0;
      if (++m==60){
        m=0;
        if (++h==24){
          h=0;
} } } } }
```

The code can be used, for instance, in combination with an LCD display to make a full-blown autonomous clock. The instantaneous divergence from the real time can be relatively large (up to a minute, or so), but in the long run the electricity central will try to make as-close-as-possible a 50 Hz line cycle and our clock will therefore always go back to its correct time without any intervention.

### 5.5.8 Arduino light dimmer

The light dimmer is a project similar to the power-line controlled clock before. The Arduino will detect a zero crossing of a power-line cycle, wait a programmed amount of time and then open a TRIAC (triode for alternating current) by a pulse. A TRIAC is a member of the thyristor family. See Figure 5.32. It is bistable; it can either conduct large currents thus being effectively short circuit, or be effectively open circuit. The state can be programmed by the signal at the gate[1]. When the TRIAC is opened by a signal at the gate, large currents can flow through it. The TRIAC only goes back to the insulating state when

[1] An unfortunate name, since a TRIAC is not based on field-effector geometry but p-n junctions instead

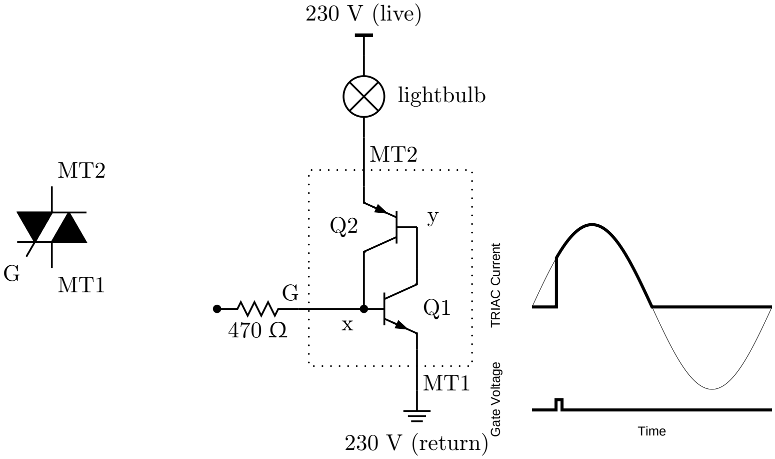


**Fig. 5.32**: Symbol of a TRIAC and an example of a simplified effective equivalent circuit in operation for positive voltages (so-called quadrant 1). The conduction is bistable. In the off state, with the gate voltage at x smaller than 0.7 V, transistor Q1 is closed. Since no current can flow through Q1 and thus also through Q2, the emitter-base drop of transistor Q2 must be less than 0.7 V. The voltage at y must therefore be something close to 240 V since, without current, also no voltage drop occurs in the lightbulb. Both transistors are closed. If the gate voltage is raised (temporarily) above 0.7 V, transistor Q1 opens and this pulls voltage at y down, increasing the emitter-base voltage drop of Q2 which thus also opens. Even if now the gate voltage is removed, the TRIAC remains conductive, until the end of a mains half-cycle, when current drops to zero. The 470-ohm resistance is used to limit the gate-signal source current just like for any transistor. The bottom image sketches this behavior

the gate signal is removed *and* the current drops below a certain value. The TRIAC then goes back to idle, waiting for a new pulse at the gate. We make use of this for our dimmer to open the TRIAC and switch on a lamp at an exact phase of the power line.

Once again, to separate the high voltage part from the low-voltage Arduino, we can use an optocoupler at the output part of our circuit. In fact, there exist special TRIAC driver optocouplers.

Figure 5.33 gives an example of a circuit for this light dimmer. The live voltage is sampled by digital port D2. The Arduino program is interrupted on a zero-crossing (going to negative) of the line voltage. It then waits a programmed amount of time using the `delayMicroseconds()` routine. It then opens the TRIAC by placing a pulse on port D3 and goes back to idle, waiting for an interrupt on port D2.

```
/*****************************************************
 *  Light dimmer program                             *
 *  Waits for zero crossing of line                  *
```

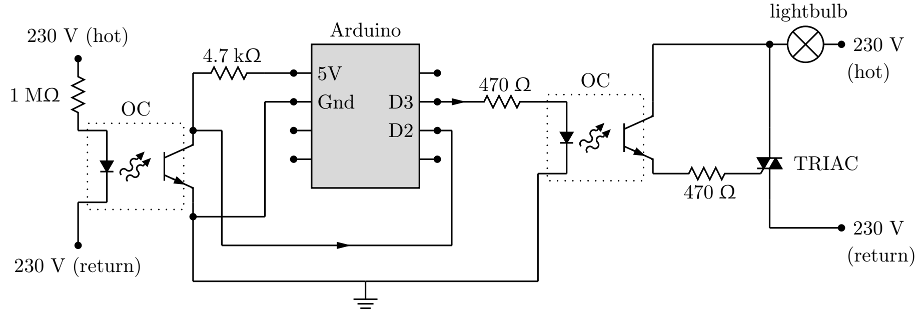


**Fig. 5.33**: Circuit for using an Arduino as a programmable light dimmer

```
 *  and switches TRIAC output by pulse                *
 ****************************************************/
int offpercent = 20;

void setup(){
  pinMode(2, INPUT);
  pinMode(3, OUTPUT);
  // interrupt 0 is digital pin 2
  attachInterrupt(0, myinterrupt, CHANGING);
}

void myinterrupt(){
  // for 50 Hz: 1 half cycle is 10000 us
  delayMicroseconds(offpercent*10);
  // Put pulse on output, switching TRIAC
  digitalWrite(3, HIGH);
  delayMicroseconds(10);
  digitalWrite(3, LOW);
}
```

You might want to add some filter elements around the TRIAC, especially if you are going to switch inductive elements such as motors. For simplicity these elements are not shown here. Look at the datasheets of the opto-coupler and TRIAC components for advice on how to use them in the best way.

The same thing can also be done by just putting a power FET inside a bridge rectifier and PWM-output a regular lamp without access to the zero crossing, but because the switching will not be synchronized with the power line, you might get some funny interference effects.

There also exist simpler light dimmers on the market, but don't forget that with the Arduino, we can link it with other parts. An example is making a light dimmer that is controllable through internet or controlled by a light detector.

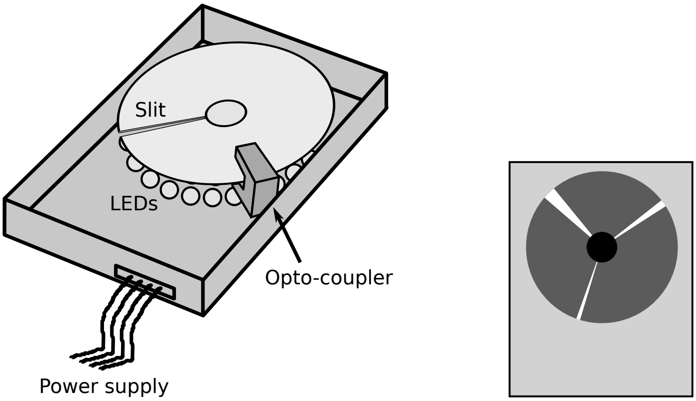


**Fig. 5.34**: Schematic drawing showing the setup of the harddisk clock (left) and how it looks when working (right), here showing 10:10:34

### 5.5.9 Arduino harddisk clock

This project is entertaining and informative. We are going to build a clock from mostly scrap material. The heart of the clock is a destroyed harddisk. That is, one that is still rotating, but no longer stores data reliably. It should not be so difficult to obtain one. Any one will do, as long as it still rotates when powered.

Open the harddisk and remove the read/write head completely. If the harddisk consists of several disks, remove all of them, except one (the top one). In this disk, cut a slit of some millimeters. (For stability reasons it is better not to cut it all the way to the center). Below the disk place (power) LEDs. See Figure 5.34. Flange the disk by an opto-coupler pair (emitter-detector). Such optocouplers can be found in computer mouses.

The opto-coupler will detect the passage of the slit and will signal it to the Arduino. The Arduino waits a certain amount of time – depending on the time, the seconds – and lights all the LEDs for a certain short amount of time. Because this procedure is repeated rapidly (in a conventional harddisk of 5400 rpm, it will be 90 times per second; 1/60th of a revolution will take 185 μs) the optical illusion is a stroboscopic effect of a stationary line at a certain place. This will represent the seconds hand. The same procedure is repeated for the minutes and hours hands. We will light these a little longer than the seconds hand, to make the line thicker.

The electronic circuit is shown in Figure 5.35. The resistances for the opto-coupler (O-C) used (here 470 Ω and 4.7 kΩ) depend on the type of optocoupler. It is best to check the signal with an oscilloscope to find adequate values. The shunt resistor of 100 Ω at the gate of the power FET is, in principle, not needed, but also does not do any harm and this way we can also use junction transistors. The power for the LEDs comes from the (computer) 12 V power supply that also powers the harddisk. The 5 volt from the same power supply is used to power the Arduino.

The program code is given below. Note that without knowledge of the real actual time, the program just assumes something. The program can be combined with another Arduino project described here to give a real clock.

```
/*********************************************************
 *  Harddisk clock                                       *
 *  Wait for trigger at pin 12                           *
 *  then wait correct amount of time and switch on LEDs *
 *  at pin 13. Repeat procedure for seconds, minutes     *
 *  and hours                                            *
 *********************************************************/

int ledPin =  13;
int triggerPin = 12;
long startseconds, offset, starth, startm, starts;

void setup(){
  pinMode(ledPin, OUTPUT);
  pinMode(triggerPin, INPUT);
  startseconds = (millis()/1000);
  // without knowledge of real time, we just set it here
  // at something random here.
  starth = 10;
  startm = 10;
  starts = 34;
  offset = starth*3600 + startm*60 +starts;
}

void loop(){
  int val;
  long seconds, minutes, hours, s, m, h;

  seconds = (millis()/1000)-startseconds;
  seconds += offset;
  s = (seconds % 60);
  minutes = seconds / 60;
  m = minutes % 60;
  hours = minutes / 60;
  h = 5*(hours % 12)+m/12;

  // correct for that sensor is not at 12 o'clock:
  s = (s+38) % 60;
  m = (m+38) % 60;
  h = (h+37) % 60; // hours-bar is wider

  val = 1;
```

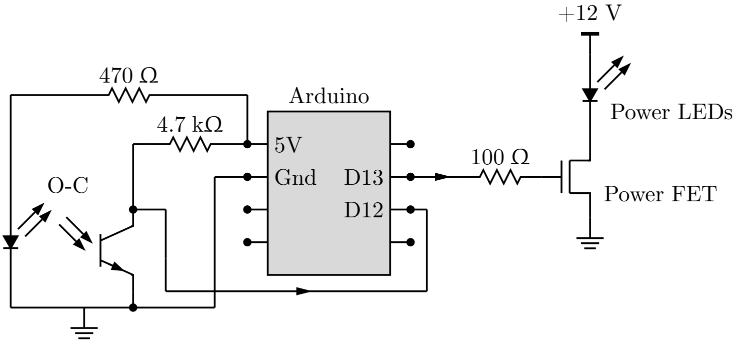


**Fig. 5.35**: Circuit for the harddisk clock. O-C is the opto-coupler, the choice of connected resistances depends on the opto-coupler specifications

```
  while(val)
    val = digitalRead(triggerPin);
  delayMicroseconds(185*s);
  digitalWrite(ledPin, HIGH);
  delayMicroseconds(80);
  digitalWrite(ledPin, LOW);
  val = 1;
  while(val)
    val = digitalRead(triggerPin);
  delayMicroseconds(185*m);
  digitalWrite(ledPin, HIGH);
  delayMicroseconds(200);
  digitalWrite(ledPin, LOW);
  val = 1;
  while(val)
    val = digitalRead(triggerPin);
  delayMicroseconds(185*h);
  digitalWrite(ledPin, HIGH);
  delayMicroseconds(400);
  digitalWrite(ledPin, LOW);
}
```

### 5.5.10 Arduino acoustic position detector

Piezo elements are sensitive pressure sensors that can be used as microphones. They are not of very high-fidelity, meaning that their frequency characteristics are not flat in the human-audible range of 10 Hz - 20 kHz, but they compensate by being extremely cheap and sensitive. As such, ideal for a motion detector. With some care, it should be easy to detect movement in a room. Here we will extend the idea a little and make a position detector. It works on the principle

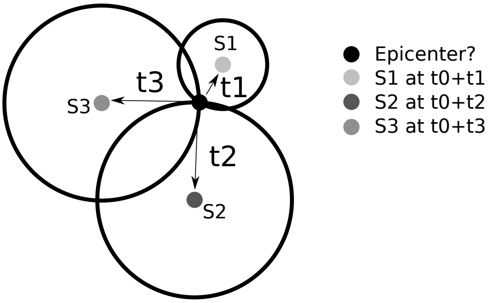


**Fig. 5.36**: Example of trilateration. Knowing the time-of-arrival of acoustic waves at three sensors, we can reconstruct the position of origin of the sound. Two sensors are enough to give two possibilities by intersection of circles. However, because the time of origin of the sound is not known, three sensors are needed. If the speed of sound is not known, for instance because the medium of the acoustic waves is not known, a fourth sensor is needed

that we determine the time of arrival of the acoustic waves at four different piezo sensors. Knowing the velocity of sound, we can calculate the distance of the origin to the four sensors and by trilateration we can determine the location of the source, see Figure 5.36. Two sensors are enough to determine the position on a plane. Or better to say to limit it to two possibilities, namely the two intersection points of the two circles. However, if we do not know the time of the origin of the sound, we need a third sensor. Thus, in principle, three sensors are enough to find the location on a plane, but if we also do not know the speed of sound, for instance because we do not know the material the sound travels through, a fourth sensor is needed. These ideas are very similar to the trilateration technique used for determining the position of an earthquake. It is also similar to GPS (global positioning system). In GPS, however, the role of emitter and receiver are reversed. Whereas our system has fixed sensors and a unknown emitter position, in GPS the emitters are fixed and we want to know the position of the single sensor.

An application of this piezo-sensor system is a projection screen for computer presentations. In the four places around the screen we glue our four piezo-sensors. When the user taps on the screen this can be interpreted as a mouse-click.

It is not worth it here to give a full Arduino program, because it depends a lot on the exact sensors used and the configuration. Yet, a basic idea for a two-sensor linear-position detection system is shown in the code below. For the hardware an ingredient is a high-gain amplifier. In the circuit of Fig. 5.37 we use operational amplifiers (for instance half of a 358 dual-opamp IC). Without feedback (as shown), the gain is theoretically infinite and the opamp functions as a comparator, with the output saturating at either $+V_{\mathrm{cc}}$ (+5 V) or $-V_{\mathrm{cc}}$ (0) depending on the sign of the sensor signal. In practice, the gain is some tens of thousands and we still have an analog signal at the output. For this reason,

**Table 5.IX**: Speed of sound in various materials at room temperature

| Material | Speed of sound |
|---|---|
| Air | 344 m/s |
| Wood (Oak) | 4470 m/s |
| Cork | 500 m/s |
| Iron | 5900 m/s |
| Water | 1480 m/s |

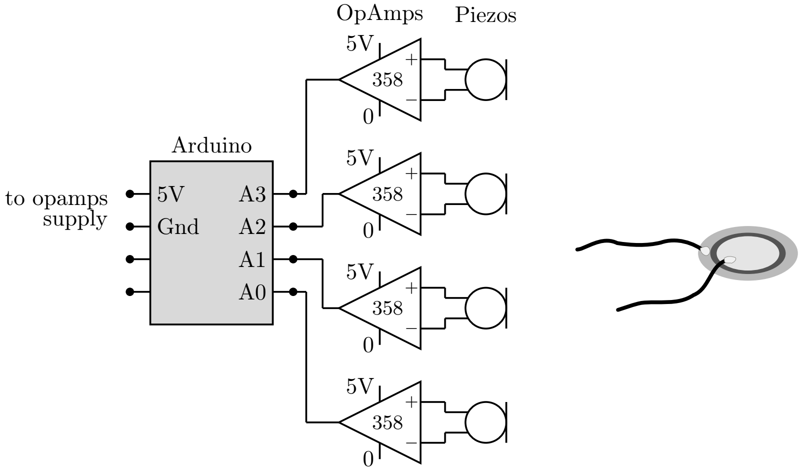


**Fig. 5.37**: Circuit for using the Arduino with four piezo sensors to detect the position of a sound source

the circuit here uses analog inputs (A0 to A3) of the Arduino instead of digital ones. The Arduino program times the arrival of the circuits. To convert the times found to position, we have to know the speed of sound. Table 5.IX gives the speed of sound in some common materials. Also note that the speed of sound is depending on the temperature, roughly

$$v = 331.3 + 0.606 \times T \text{ (m/s)}, \tag{5.4}$$

with $T$ the temperature in degrees Celsius (°C).

```
/********************************************************
 *  Linear position detector based on two Piezo sensors *
 *  connected to analog ports 4 and 5                   *
 *  If one of them receives a signal its time is        *
 *  remembered and the program waits for signal from    *
 *  the other sensor. If this doesn't come fast enough  *
 *  the system is reset (t4=t5=0)                       *
 *  Threshold levels for the sensor are 500 and this    *
 *  can be adjusted for each system                     *
 *  Output of timing is done to serial port, where a    *
 *  listening computer can process the signal           *
 *  Arduino idea by Marco Silva and Andre Cardoso       *
 ********************************************************/
unsigned long t4, t5;

void setup() {
  t4 = 0;
  t5 = 0;
}

void loop() {
  int v4, v5;
  long tdiff;

  if (t4==0){
    v4 = analogRead(4);
    if (v4>500)
      t4 = micros();
    tdiff = (t4-t5);
    if (tdiff>20000)
      t5 = 0;
  }
  if (t5==0){
    v5 = analogRead(5);
    if (v5>500)
      t5 = micros();
    tdiff = (t5-t4);
    if (tdiff>20000)
      t4 = 0;
  }
 if ((t4>0) && (t5>0)){
    if (t5>t4) tdiff = t5-t4;
    else tdiff = -(t4-t5);
    Serial.print("tdiff=");
    Serial.println(tdiff);
    delay(1500);
    t4=0;
    t5=0;
 }
}
```

### 5.5.11 Arduino remote control

Many home appliances nowadays use remote control. This ranges from television sets, where we can change the channels and volume settings etc., to other audio-visual equipment such as DVD players, tuners, media stations. My computer, actually, even has a remote control unit which comes in very handy when giving presentations. All these remote controls basically work the same way, apart from the fact that there exist different standards. The most common is the RC-5 protocol developed by Philips in the late 1980s. This will be explained here.

The RC-5 protocol uses triple modulation to send a bit stream of 14 bits, see Figure 5.38. The first step of the modulation is to turn a bit into two bits, '01' for a logical '1' and '10' for a logical '0', a technique that is called Manchester encoding. The ensures that in a stream there are as many high levels as low levels; the number of 1s and 0s in a bitstream are always equal. Moreover, since in this way there can never be more than two equal consecutive bits, the information is always passed in high frequency; even a bitstream of only 1s will have many state changes and the frequency spectrum of the signal is residing in high frequencies and thus avoid the earlier mentioned 1/f noise. A classic technique of telecommunications. We will also use this in the program, as a side effect.

Each bit of this doubled bitstream takes exactly 888.8 µs. The total pattern thus takes $14 \times 2 \times (888.8\ \mu\text{s}) = 24.88$ ms. Note that the first physical '0' of the first logical '1' cannot be detected, since it is a silence. The pattern can be repeated after 114 ms.

The second step in the modulation is filling each '1' with pulses 'high' with a repetition frequency of 36 kHz. The duty cycle (the percentage of time 'high') can be anything between 25% and 33%, but the total time between pulses is 27.78 µs. A '1' is thus a sequence of $(888.8\ \mu\text{s})/(27.78\ \mu\text{s}) = 32$ pulses. A '0', obviously, has zero pulses.

The third step is multiplying this pulse signal with the carrier, for which normally an infra-red LED is used, typically of 950 nm wavelength. Technically speaking, each pulse is filled with 315 THz electromagnetic radiation.

The 14-bit pattern itself always starts with a logical '1' (physical '01', low-high). Originally, the second bit was also a synchronization bit like the first one (and was always a logical '1'). Later it was used to extend the number of commands and could also be '0'. The third bit is a bit that is toggled with each button press. This way, there is a difference between two times pressing the same button (where the bit toggles) and keeping the button pressed down (where the bit doesn't toggle). The next 5 bits are reserved for identifying the equipment. This avoids that we change the volume level of the radio when we wanted to change it of the television. The last 6 bits are the actual command.

We could now design a receiver based on the above information. However, we can make a nice shortcut and buy a ready-made IR receiver module. For instance a TSOP312xx from Vishay or a SFH 506-xx of Siemens, where xx stands for the RF frequency used, 36 kHz in our case. These modules do the first two

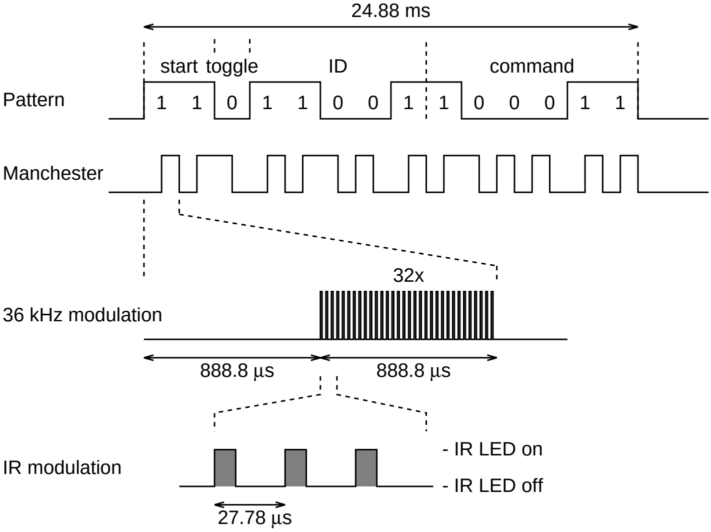


**Fig. 5.38**: Philips RC-5 modulation. Starting with a 14-bit pattern consisting of two start bits, one key-toggle bit, 5 device ID bits and 6 command bits, the first step is Manchester coding converting each '1' to a '01' combination and each '0' into a '10' combination. This ensures equal number of 1s and 0s in the pattern. The next step is filling each '1' with 36-kHz repetition-rate pulses. The final step is multiplying this pulse sequence with the infrared carrier (typically 950 nm), i.e., switching on and off an IR LED. Each bit takes 2×888.8 μs. The entire sequence 24.88 ms

steps in demodulation and result in a high-low Manchester coded sequence, as shown on the second line in Figure 5.38. Moreover, they use automatic gain control techniques, to be able to work in low-signal high-noise environments. Our Arduino only has to decode the Manchester code to determine the underlying bit pattern and find the code for the key pressed. The circuit is as simple as the one shown in Figure 5.39. The sensor uses inverted voltages, a logical '0' is 5 volt and a logical '1' is 0 volt.

We can even take it a step further. Most producers of equipment use their own protocol. While RC-5 was one of the most used, it is not the only one. Another protocol was developed by Sony. It consists of pulse-width modulation. A bit always starts with physical '0' of 600 μs, which is then followed by a physical '1' of 600 μs for a logical 0 and 1200 μs for a logical 1. Once again, each physical 1s is filled with a 36 kHz pulse train. Everything is preceded by a 3600 μs physical '1'. Figure 5.40 clarifies this.

Modern remote control units take it a step further. They just use the time between short pulses, see Figure 5.41, which shows the sequence of a generic RC unit. Instead of filling (half or two-thirds of) a bit by (24, 32 or 48) pulses, only a very short single pulse is generated at the beginning of a 'high'. The width of

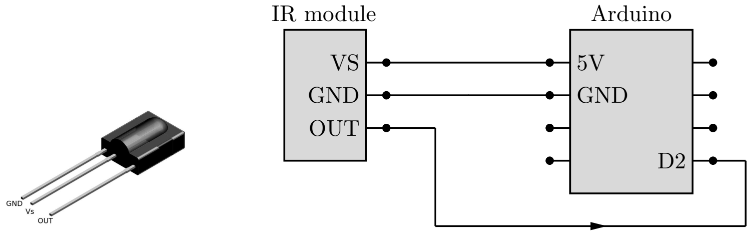


**Fig. 5.39**: Arduino with an IR receiver module for remote control

this pulse is actually irrelevant, but the electronics of the receiver module likes best a pulse of 48 µs. The information lies in the *distance* between the pulses. The distance between the starts of two consecutive pulses is always a multiple of the base time, for example 3360 µs. Actually, it is always 1, 2, 3 or 4 times this value. A typical code is thus something like the 13131312112211 as shown in Figure 5.41. The advantage of this technique is that it consumes much less power, because the LED is much less time on, because it uses one pulse instead of 32. This prolongs the lifetime of the battery of the remote control unit.

The Philips RC-5 and Sony RC protocols are actually compatible with this technique. For instance, the sequence in Figure 5.38 would have a base time of 888.8 µs and a code 2423323232. The Sony sequence of Figure 5.40 has a base time of 600 µs and a code 423323222322. If the receiver is smart and can understand this efficient signal-pulse protocol, a generic remote control can be used for it, both for Philips and Sony equipment. We can thus best use an empirical program to analyze the signals and deal with them.

A program to detect this is given below. It is based on an empirical determination of the timings and the codes of a generic remote control unit. The time found of the shortest distance is about 3360 µs, remarkable close to twice the time of a bit in th RC-5 code.

The Arduino program records the times when transitions high-low occur. (Note that the sensor uses inverse voltage; it is 5 volts in idle and 0 when IR light is received). It does this by detecting the state changes at digital pin 2 that receives the output from the IR receiver module. The time since last high-low transition in microseconds is stored in an array `times[]`. We start decoding when there has been silence for more than 20 ms. The procedure `decode()` is called that handles the decoding part. The procedure converts the times to integer numbers, multiples of the base time. It then compares the found code – the string of numbers – to a small database, in this case just the keystrokes '0' to '1', that were empirically found. If we want to find the base time and the sequences of a remote-control unit, we can uncomment the lines outputting the time array to the serial port. The output is then something like shown in Figure 5.42.

```
/****************************************
 *  Empirical IR remote control decoding *
 *****************************************/
```

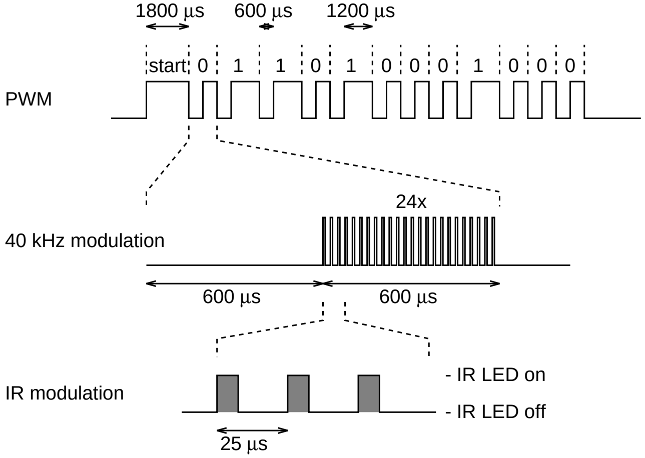


**Fig. 5.40**: Sony RC triple-modulation protocol. After a start signal of 1800 μs every bit of the code is first pulse-width modulated: a 600 μs 'low' is followed by 600 μs 'high' for a logical '0' and 1200 μs for a logical '1'. The next step is filling each 'high' pulse with 40-kHz repetition-rate pulses, namely 24 or 48 of them. The final step is multiplying this pulse sequence with the infrared carrier, i.e., switching on and off an IR LED. The total time of a sequence depends on the starting bit pattern

```
unsigned long times[100]; // the array with times
int ti;                   // the index in the above array
char code[80];            // the sent code
unsigned long prevt; // the previous time of interrupt
int prevstate;            // the previous state of the sensor

char *buttoncodes[] = {
    "13131312112211",   // 0
    "13131312211112",   // 1
    "131313122212",     // 2
```

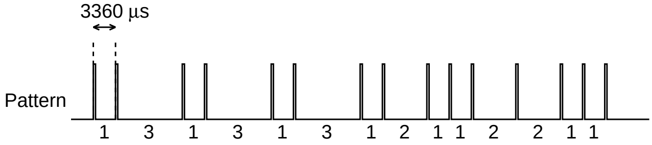


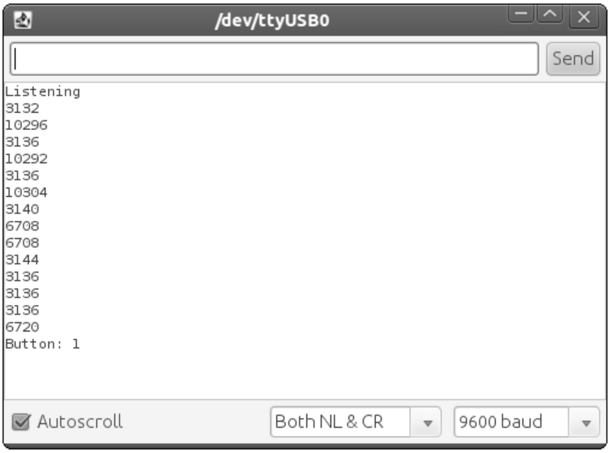
**Fig. 5.41**: Empirical remote control protocol. Pulses separated by a multiple of 3360 μs. In this case 13131312112211, corresponding to the key '1' of a generic remote control unit

```
    "1313131221122",    // 3
    "131313122222",     // 4
    "131313121111112",  // 5
    "13131312112112",   // 6
    "13131312111122",   // 7
    "1313131211222",    // 8
    "131313122111111"}; // 9

void reset(){
  ti = 0;
  prevstate = HIGH;
  prevt = 0;
}

void setup() {
  pinMode(2, INPUT);
  Serial.begin(9600);
  Serial.println("Listening");
  reset();
}

void decode(){
  // The procedure that converts the times into a code
  int i, j, n;
  int ci;

  if ((ti<6) || (ti>78)) { // Noise. Exit!
     reset();
     return;
  }
  ci = 0;
  for (i=0; i<ti; i++){
    n = (times[i]+1680) / 3360;
      code[ci++] = 48+n; // convert to ASCII and store
     // if you want to find the information of an RC unit
     // uncomment the following line:
     // Serial.println(times[i]);
  }
  code[ci] = 0; // terminate string
  // compare to database:
  for (i=0; i<=9; i++) {
    if (!strcmp(code, buttoncodes[i])) {
      Serial.print("Button: ");
      Serial.println(i);
      break; // key found; exit for loop
    }
  }
  reset();
}

void loop() {
  // The main loop that does the polling of the sensor
  // and checks if it is already time to decode the pattern
  unsigned long t;
  int state;

  t = micros();

  if (ti>90)  // too many transitions. something went wrong
    reset();

   // After 20 ms of silence, the pattern for sure is
   // finished and we can decode it:
  if ((prevt>0) && (t-prevt>20000))
    decode();

  state = digitalRead(2);
  if (state!=prevstate){
    if (state==LOW){
      if (prevt!=0)
        // if not first transition, save time since last one
        // and increment times-array index:
        times[ti++] = t-prevt;
      prevt = t;    // save last time
    }
    prevstate = state;
  }
}
```

To make our Arduino take on the role of the remote control unit, all we have to do is connect an infra-red LED to one of the ports and generate the correct pulse sequence.

The Arduino code is quite simple, especially when compared to the decoding program given before. Here is an example that cycles sending the codes for buttons '0' to '9' one keystroke per second:

```
/****************************************
 *  Empirical IR remote control encoding *
 ****************************************/

char *buttoncodes[] = {
```



```
Listening
3132
10296
3136
10292
3136
10304
3140
6708
6708
3144
3136
3136
3136
6720
Button: 1
```

**Fig. 5.42**: Typical output of empirical RC program

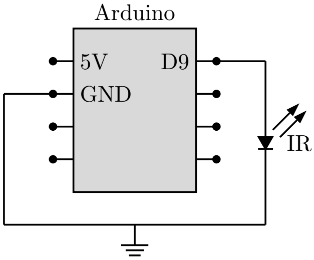


**Fig. 5.43**: Arduino with an IR LED to be a remote control unit and control a television set or any other home equipment

```
   "13131312112211",   // 0
   "1313131221112",    // 1
   "131313122211",     // 2
   "131313122112 2",   // 3
```

```
   "13131312112211",   // 0
   "1313131221111 2",   // 1
```

```
   "13131312112211",   // 0
   "1313131221111 2",   // 1
   "131313122211 2",    // 2
   "131313122112 2",    // 3
   "13131312222 2",     // 4
   "13131312111111 2",  // 5
   "1313131211211 2",   // 6
   "1313131211112 2",   // 7
   "131313121122 2",    // 8
   "13131312211111"}; // 9

void setup() {
  pinMode(9, OUTPUT);
  Serial.begin(9600);
  Serial.println("Talking");
}
```

```
void encode(int n){
  // The procedure that converts the code into times
  int i;

   // start pulse:
  digitalWrite(9, HIGH);
  delayMicroseconds(48);
  digitalWrite(9, LOW);
  for (i=0; buttoncodes[n][i]!=0; i++){
    delayMicroseconds((buttoncodes[n][i]-48)*3360-48);
    digitalWrite(9, HIGH);
    delayMicroseconds(48);
    digitalWrite(9, LOW);
  }
}

void loop() {
  int j;

  for (j=0; j<=9; j++) {
    Serial.println(j);
    encode(j);
    delay(1000);
  }
}
```

### 5.5.12 Arduino barcode reader

A barcode is a set of lines coding a product number. The most famous barcode is UPC (Universal Product Code), used for most products in shops (in supermarkets, for food in general sometimes shorter versions are found). A version of this is EAN-13 (European Article Number with 13 digits) which is encountered in European shops and which we will discuss here.

The product number has 12 codified digits, divided in two groups of 6. This is preceded by one (uncoded) digit. (This one plus the first or first two coded digits represent the country of origin, for example, '560' is Portugal, '00' is United States).

Each digit is coded by two black and two white bars of width 1 to 4 units, with a total width of 7 units. Before the code, at the end of the code, and in the middle a synchronization pattern is placed consisting of two single-width black bars spaced by single-width white bars. They can be recognized by them being a little bit longer than the others. The codes for these are given in Table 5.X ('0' is white, '1' is black). A barcode therefore always has exactly 30 black lines on a white background. The total width from first black line to last black line inclusive is 95 times the width of the thinnest line. An example is given in

**Table 5.X**: Start, stop and middle bit pattern of barcode ('0' is white, '1' is black)

| Synchronization code | Pattern |
|---|---|
| Start | 0101 |
| Middle | 01010 |
| Stop | 1010 |

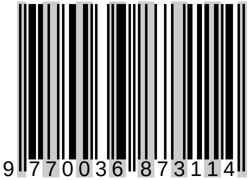


**Fig. 5.44**: Example of a barcode (of the Scientific American magazine). The individual digits of the code are highlighted by gray background for every odd digit (this is not part of the barcode, of course)

Figure 5.44 where the individual patterns are highlighted.

A digit can be coded in three different ways, depending on its position in the number. Digits on the right side (last six digits) are always coded normally and always start with a black line and end with a white line. Digits on the left side (first 6 digits) can be either coded the inverse of the normal pattern (1 becomes 0 and vice versa), or its reverse order (first bit becomes last bit, etc.). Table 5.XI clarifies what is meant with this.

The six left digits are either coded inverse (I) or reverse (R) depending on their position in the number and the value of the zeroth non-coded digit. The first (coded) digit is always coded inverse. The others are coded according to

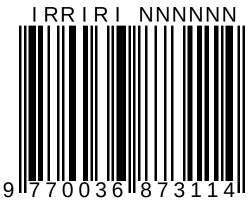


**Fig. 5.45**: Example of a barcode with the type of coding for each digit indicated above. Because the non-coded digit before the barcode is a 9, the first 6 coded digits are coded by 'IRRIRI' (I is inverse, R is reverse), see Tables 5.XI and 5.XII. The rest of the digits, after the synchronization pattern in the middle, are coded normally (N)

**Table 5.XI**: Normal, inverse and reverse bit patterns of decimal digits of barcode. The second half of a barcode is always coded normally, while for the first six digits – left of the middle-synchronization pattern – the coding inverse or reverse depends on the position of the digit and the first non-coded digit, see Table 5.XII. ('0' is white, '1' is black)

| Digit | Normal code (N) | Inverse code (I) | Reverse code (R) |
|---|---|---|---|
| 0 | 1110010 | 0001101 | 0100111 |
| 1 | 1100110 | 0011001 | 0110011 |
| 2 | 1101100 | 0010011 | 0011011 |
| 3 | 1000010 | 0111101 | 0100001 |
| 4 | 1011100 | 0100011 | 0011101 |
| 5 | 1001110 | 0110001 | 0111001 |
| 6 | 1010000 | 0101111 | 0000101 |
| 7 | 1000100 | 0111011 | 0010001 |
| 8 | 1001000 | 0110111 | 0001001 |
| 9 | 1110100 | 0001011 | 0010111 |

**Table 5.XII**: The type of coding for the first half of a barcode – left of the middle-synchronization pattern – is depending on the position of the digit and the first non-coded digit. 'N' is normal, 'I' is inverse, 'R' is reverse, see Table 5.XI

| Non-coded digit | Coding of coded digits |
|---|---|
| 0 | IIIIII-NNNNNN |
| 1 | IIRIRR-NNNNNN |
| 2 | IIRRIR-NNNNNN |
| 3 | IIRRRI-NNNNNN |
| 4 | IRIIRR-NNNNNN |
| 5 | IRRIIR-NNNNNN |
| 6 | IRRRII-NNNNNN |
| 7 | IRIRIR-NNNNNN |
| 8 | IRIRRI-NNNNNN |
| 9 | IRRIRI-NNNNNN |

**Table 5.XIII**: Country codes for EAN-13 barcodes. These indicate the country where the code was issued, not necessarily where the product was made

| | | | | | |
|---|---|---|---|---|---|
| 00-13 | USA & Canada | 20-29 | In-Store Functions | 30-37 | France |
| 40-44 | Germany | 45,49 | Japan | 46 | Russian Federation |
| 471 | Taiwan | 474 | Estonia | 475 | Latvia |
| 477 | Lithuania | 479 | Sri Lanka | 480 | Philippines |
| 482 | Ukraine | 484 | Moldova | 485 | Armenia |
| 486 | Georgia | 487 | Kazakhstan | 489 | Hong Kong |
| 50 | United Kingdom | 520 | Greece | 528 | Lebanon |
| 529 | Cyprus | 531 | Macedonia | 535 | Malta |
| 539 | Ireland | 54 | Belgium & Luxembourg | 560 | Portugal |
| 569 | Iceland | 57 | Denmark | 590 | Poland |
| 594 | Romania | 599 | Hungary | 600,601 | South Africa |
| 609 | Mauritius | 611 | Morocco | 613 | Algeria |
| 619 | Tunisia | 622 | Egypt | 625 | Jordan |
| 626 | Iran | 64 | Finland | 690-692 | China |
| 70 | Norway | 729 | Israel | 73 | Sweden |
| 740 | Guatemala | 741 | El Salvador | 742 | Honduras |
| 743 | Nicaragua | 744 | Costa Rica | 746 | Dominican Republic |
| 750 | Mexico | 759 | Venezuela | 76 | Switzerland |
| 770 | Colombia | 773 | Uruguay | 775 | Peru |
| 777 | Bolivia | 779 | Argentina | 780 | Chile |
| 784 | Paraguay | 785 | Peru | 786 | Ecuador |
| 789 | Brazil | 80-83 | Italy | 84 | Spain |
| 850 | Cuba | 858 | Slovakia | 859 | Czech Republic |
| 860 | Yugoslavia | 869 | Turkey | 87 | Netherlands |
| 880 | South Korea | 885 | Thailand | 888 | Singapore |
| 890 | India | 893 | Vietnam | 899 | Indonesia |
| 90,91 | Austria | 93 | Australia | 94 | New Zealand |
| 955 | Malaysia | 977 | ISSN | 978 | ISBN |
| 979 | ISMN | 980 | Refund receipts | 981,982 | Common currency |
| 99 | Coupons | | | | |

Table 5.XII. An example is given in Figure 5.45.

Note that, since the pattern of coding types of the first 6 digits is unique, in this way it is possible to recover the zeroth non-coded digit! (This shows the history of the EAN-13 code, as a derivative of the UPC-12 code – which is equal to the the EAN-13 with the zeroth digit equal to 0 – compatibility prohibited the zeroth digit to be coded directly).

The meaning of the number is the following: The first digits codify the country of origin. This includes the zeroth non-coded digit plus one or two more digits. The rest of the first six digits are reserved for the producer. The first five digits are the choice of the producer. The last digit is a checksum code. Table 5.XIII shows the country codes (where the institute that issued a code resides, not necessarily the country where the product was made!). An example is given in Figure 5.46 with the individual patterns highlighted.

We have a way to verify if everything is correct by looking at the checksum:

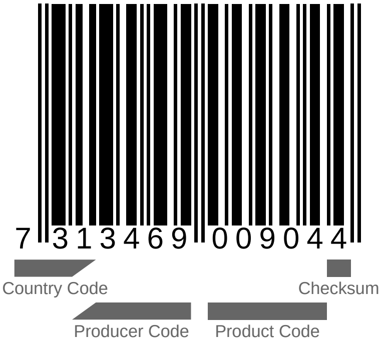
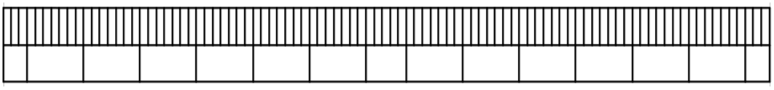


**Fig. 5.46**: Items of a barcode. The first two or three digits (including the non-coded digit) are the country code. The next digits, up to the half-way synchronization pattern are the producers code. The digits after the synchronization pattern are the product code

The sum of all digits, including the checksum, needs to be a multiple of 10. For this calculation, every even digit (counting from the right) has to be multiplied by 3 before adding to the sum. As an example, the above barcode of Figure 5.46 is 7 313469 009044, and $7 + 3 \times 3 + 1 + 3 \times 3 + 4 + 3 \times 6 + 9 + 3 \times 0 + 0 + 3 \times 9 + 0 + 3 \times 4 + 4 = 100$, and the checksum is correct because 100 is a multiple of 10.

Exercises
1) What is the code for the barcodes shown in Figure 5.47 (including the zeroth digit!). Give also the country and the checksum.
2) Draw the barcode for 5602007192198 in the figure below



3) What is the checksum of the barcode in Figure 5.48?
4) The price is not coded into the EAN. Why not?
5) The Da Vince code? In every code there is the number 666. Find out where.
Answers: 1) 4 009993 902090 (40: Germany, CS: 3), 3 501365 988035 (35: France, CS: 0), 8 411359 037586 (84: Spain, CS: 6). 3) 5. 4) Price is not constant. Varies from shop to shop, from day to day, from currency to currency. 5) The code for a '6' is two single-width black lines separated by a single-width white line. These can be found in the synchronization patterns at the beginning, end and in the middle; 666.

The hardware for this project is once again very simple. We will use a photo-transistor or photo-LED with a resistance to convert the optical signal to a voltage. The user drags this unit across the barcode and the signal is fed

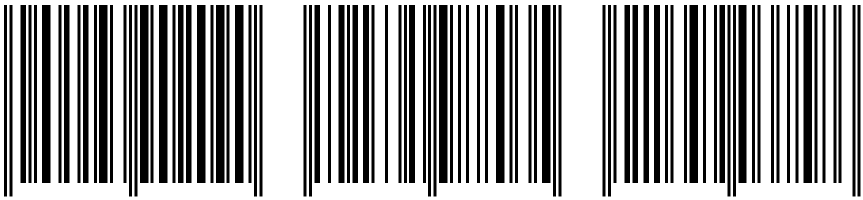


**Fig. 5.47**: Some barcodes. See exercise on p. 360

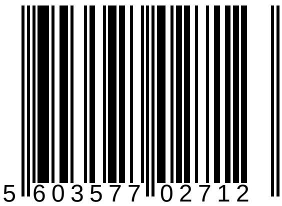


**Fig. 5.48**: Barcode with the checksum information removed. What is it? See exercise on p. 360

to one of the analog ports or digital ports, if we want to do preprocessing of the signal by software or hardware, respectively. This signal will then be converted into dark and light times and then into width of the bars.

One of the problems of this simple bar-code reader is that it is not straightforward to translate times into bar widths. For two basic reasons. First, there is a lot of noise; exactly, when something is black and when white? A good question and difficult to answer, especially in varying illumination conditions; more light makes all white bars wider and black bars narrower. Second, we do not know how fast the user moved the pen. To avoid the second problem we can use the synchronization patterns in the beginning, in the middle and at the end to determine the speed of the movement of the reader unit. However, this speed is not constant and we should also make use of interpolation. All-in-all the project is quite challenging and the Arduino code can be lengthy. The reason why no example code is given here. Trial-and-error is the best approach here.

The electronic circuit on the other hand is quite simple, see Fig. 5.49. The light intensity is detected by a photo-transistor that we pass over the barcode. To increase the signal strength, a normal LED is placed close to the phototransistor, pointing in the direction of the barcode, though this component is not essential. The speed of information transmitted is quite limited and we can thus filter off a large amount of noise. Typically, a person passes the barcode 'wand' in about a quarter of a second over barcode. For 96 bits, this makes a total of some 400 b/s. We can safely filter off anything above, say, 2 kHz. In the figure a low-pass filter was added by bridging the signal pin to ground, with $1/2\pi RC = 2$ kHz, and $R = 4.7$ k$\Omega$, we get a capacitor of 80 μF. We use 100

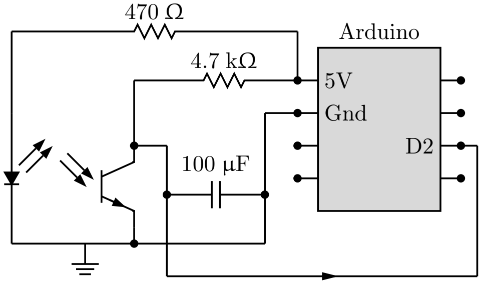


Fig. 5.49: Circuit for using the Arduino to read barcodes. It is based on a photo-transistor. Additionally, a normal LED is used to add locally light

µF to be on the safe side. In any case, we have to determine what values of the components work best for our system.

The Arduino program to acquire the barcode is actually very similar to the one used for acquiring the remote-control code. I leave it up to the reader to develop a program.

After acquiring the barcode, it is best to send it to a main computer via RS232/USB. On the computer a database can be maintained with actualized products and prices and to make statistics on products sold. The Arduino is not adequate for that task.

### 5.5.13 Arduino RFID

RFID (radio-frequency identification) is a way to label things by using the low-frequency part of the electromagnetic spectrum. In an earlier section we have already seen how optical EAN labels, more commonly known as barcodes, can be attached to products to identify them optically. With RFID, the code is stored in microelectronics and communicated to the receiver by ways of a short radio-frequency signal. In most cases the RFID tag is passive. That is, the electronics are powered by the RF pulse emitted by the final reader. The RFID tag itself does not have any power source. The distance at which communication can be done depends heavily on the wavelength (See Table 5.XIV). That is because the efficiency of the antenna system is given by Friis' formula, giving the ratio of power received ($P_\mathrm{R}$) and power transmitted ($P_\mathrm{T}$) as

$$\frac{P_\mathrm{R}}{P_\mathrm{T}} = \frac{A_\mathrm{R}A_\mathrm{T}}{\lambda^2 L^2}, \tag{5.5}$$

with $A_\mathrm{R}$ and $A_\mathrm{T}$ the area of the receiving and transmitting antennas, resp., $\lambda$ the wavelength, and $L$ the distance. For longer wavelengths, the efficiency drops off quadratically and this is compensated by necessarily being at a shorter distance $L$. This limits the range at which the ID works. The antenna size is often increased by a long loop added to the circuit, see for instance the low frequency RFID tag shown in Figure 5.50. An advantage of using low frequencies is that they permit using ultra-low-cost electronics, such as printable (organic)

**Table 5.XIV**: RFID protocols

| Frequency | Range |
|---|---|
| 125 kHz | 10 cm |
| 13.56 MHz | 1 m |
| 433 MHz | 100 m |
| 866 MHz | 10 m |
| 2.45-5.8 GHz | 2 m |
| 3.1-10 GHz | 200 m |

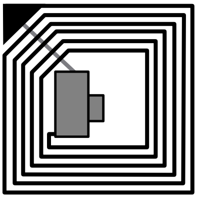


**Fig. 5.50**: Typical (low frequency) RFID tag. Most space is taken by the antenna, the spiral loop. The electronics are placed in the center. The power for driving the circuit comes from the reader that emits a strong RF pulse that is captured by the antenna and converted into electrical energy

electronics. I.e., an electronic RFID can be printed on top of a product as easily as printing an optical barcode.

With the Arduino we can make an RFID (radio-frequency identity) reader station. Here we use the low-frequency (125 kHz) technique. There exist ready-made modules that read the card and send its ID through a serial line. In this example we use an ID-12 module from Innovations because it is one that does not need any external components such as antennas, or anything. All we have to do is power it (from our Arduino), and read the serial signal coming from it. One tiny complication is that the serial line is also used to upload our program to the Arduino, so we have to disconnect the module (or at least the serial line) when uploading our sketch. Figure 5.51 shows the basic circuit, and below is given the code. Everytime the module sends a character via the serial port (RX), this character is added to a string `idnr`. A code is normally finished with ASCII character 3. When this happens (actually when any ASCII character below 32 is received), the existing string is compared (`strcmp`) to a 'database' a 10-element array of 13-character strings `legalids` (12 characters for the code and one '0' for terminating the string). If the code checks, the green LED connected to digital port 13 is switched on for one second. If it fails, the red LED (digital port 12) is switched on. The information is also sent via the serial/USB port to a computer.

The program has a novelty in that it stores the legal card numbers non-volatile memory (EEPROM), so that when power-cycles occur, the database is

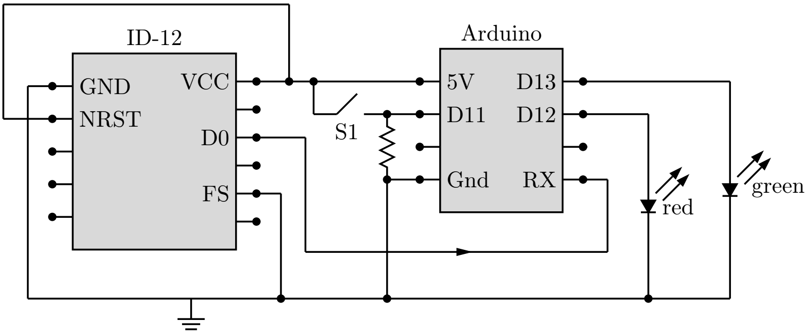


**Fig. 5.51**: Circuit for RFID reader based on a ID-12 module from Innovations

not lost. To add a card to the database, a button (S1) connected to digital pin 11 has to be pressed (on a non-user accessible place, duh!) when the card is passed in front of the sensor. The green LED will flash twice rapidly. The same button, when pressed at power start up, resets the entire database (see the part in the procedure `setup()`). The unlabeled resistor is a pull-down resistor (some 10 kΩ) to ensure that the pin is at ground when the button is open, to avoid having a 'floating', unpredictable input.

```
/********************************************
 *        RFID with EEPROM database        *
 ********************************************/

#include <EEPROM.h>

char legalids[10][13]; // databse of 10 cards
char idnr[13];   // current ID
int numids;      // number of stored legal IDs
int idlen = 0;
int eeptr;       // pointer to first free space in EEPROM

void setup() {
  int i, j;

  Serial.begin(9600);
  pinMode(11, INPUT);
  pinMode(12, OUTPUT);
  pinMode(13, OUTPUT);
  eeptr = 0;
  numids = EEPROM.read(eeptr++);
  if (digitalRead(11)) {   // HIGH on pin 11 causes reset
    numids=0;
    EEPROM.write(0, numids);
    Serial.print("Databse reset");
  }
  else {
    Serial.print("Number of cards in database: ");
    Serial.println(numids);
  }
  for (i=0; i<numids; i++) {
    for (j=0; j<12; j++)
      legalids[i][j] = EEPROM.read(eeptr++);
    legalids[i][12] = 0;  // terminate string
  }
}

void loop() {
  char inchar;
  int i, j, nomatch;

  if (Serial.available() > 0) {
    inchar = Serial.read();
    if ((inchar>32) && (idlen<13))
      idnr[idlen++] = inchar;
    else if (idlen>0) {
        idnr[idlen]=0;
        Serial.println("actual code:");
        Serial.println(idnr);
        i=0;
        nomatch=1;
        while ((i<numids) && (nomatch)) {
          nomatch = (strcmp(idnr, legalids[i]));
          i++;
        }
        if (nomatch) {
          if (digitalRead(11)) { // HIGH on pin 11: add ID
            strcpy(legalids[numids], idnr);
            numids++;
            Serial.println(">> ADDED");
            Serial.print("card number: ");
            Serial.println(numids);
            digitalWrite(13, HIGH);
            delay(200);
            digitalWrite(13, LOW);
            delay(300);
            digitalWrite(13, HIGH);
            delay(200);
            digitalWrite(13, LOW);
            EEPROM.write(0, numids); // save # cards
            for (j=0; j<12; j++) // save card in EEPROM
               EEPROM.write(eeptr++, legalids[i][j]);
          }
          else {
            Serial.println(">> REJECT");
            digitalWrite(12, HIGH);
            delay(1000);
            digitalWrite(12, LOW); }
        }
        else {
          Serial.println(">> PASS");
          digitalWrite(13, HIGH);
          delay(1000);
          digitalWrite(13, LOW);
        }
        idlen=0;
    }
    else
      idlen=0;
  }
}
```

### 5.5.14 Arduino wireless

For large sensor networks, a huge amount of cabling can become prohibitively difficult to manage. Imagine an industry project of monitoring of hundreds of fishtanks, including important vital parameters such as temperature, pH, salinity, etc. Each tank will have a tiny microprocessor-oriented measurement environment (for instance an Arduino described before). The data will have to be communicated to a central processing unit, possibly making the data available on-line. The communication to the central processor can be done by cables via RS232, however, the amount of cables makes the system unreliable, inflexible and cumbersome to say the least.

For these situations, we can revert to wireless communications. A good option is to use ZigBee protocol that translates standard RS232 (serial) communication into wireless packages and back to serial. And, in contrast to the one-to-one RS2323 protocol, this can be done in a network of devices, all talking to all, see Figure 5.52. An implementation of the ZigBee protocol is the XBee IC. As an exception to the rule of making this book in the spirit of the Open Source community, this part is about a commercial product. When using the XBee, you will immediately notice the difference. The price of the product will shoot up, the ease of operation has gone and the community is trying to protect the knowledge seemingly at all cost, resulting in a remarkable absence of good blogs and forums. Nevertheless, it is worth mentioning here.

For the Arduino a dedicated shield is made for the XBee IC, see Figure

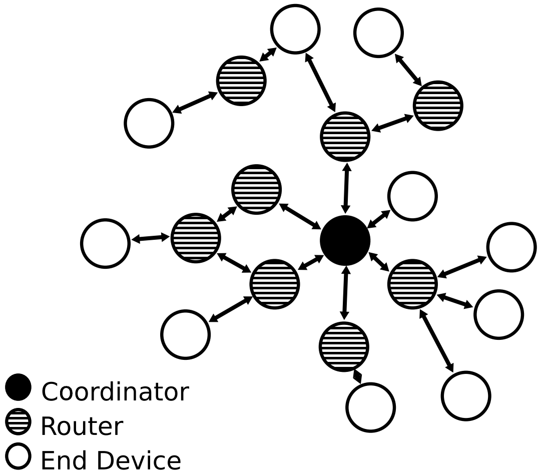


**Fig. 5.52**: A multi-point ZigBee wireless network consisting of a coordinator and several routers and end devices. Communication can be from any device to any other device and can involve hops to neighboring device

5.53. It allows for the Atmel processor of the Arduino board to communicate through the XBee with another XBee unit, as if it were connected by a serial cable. In other words, from that moment on, everything we do with the serial port is done through the XBee. This is done in a fully transparent way. That is, the Arduino does not even know it is communicating through an XBee.

First we have to set up the XBee unit. There exist dedicated XBee programming kits for this purpose, but we can also use the Arduino board plus XBee shield to program the XBee unit. To get access to the XBee, we have set the jumpers on the XBee shield correct. Figure 5.54 shows the three configurations of the jumpers. The jumpers themselves have two positions, 'XBee' and 'USB'. When in 'XBee' position, the Atmel of the Arduino board communicates with the XBee model. This is the normal mode of operation, when we have our program up-and-running on an autonomous Arduino system, for instance when not tethered to a computer. On the other hand, when we want to upload a program to the Arduino, while the XBee is present, we have to set the jumpers to the 'USB' position. The Atmel will now talk to the USB port via the FTDI (RS232/USB) chip. Finally, if we want to upload firmware or settings to the XBee module we have to silence the Atmel processor. This can be done by physically removing it, or by constantly resetting it. The latter method is preferred and can be achieved by keeping the reset pin on the Arduino board connected to ground. (See Figure 5.16). We now have access to the XBee module and can configure it to our wishes. In principle, everything can be done through the USB interface using any terminal communication program on the computer –

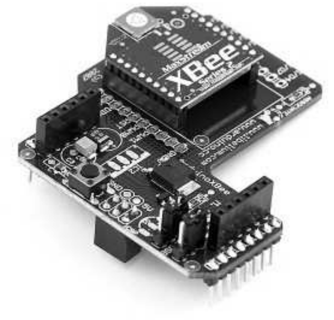


**Fig. 5.53**: XBee module on an Arduino XBee shield

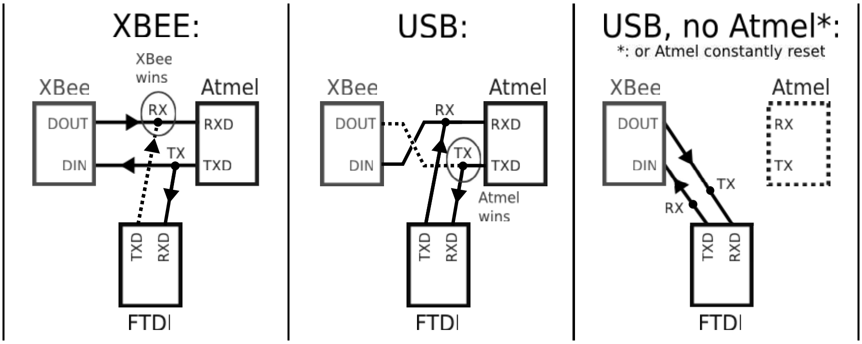


**Fig. 5.54**: Jumper settings of XBee shield

especially the settings are easy to change this way – but to upload firmware, we can make use of the handy utility X-CTU supplied by the manufacturers of the XBee, namely Digi.com.

We can now setup the XBees. The first thing to notice is that we have to make one of the units a 'controller'. Out of the box they all come as 'routers'. The difference between a router and a controller is that a controller is the one that manages the network. There can be only one controller in a network, just the same as there can be only one captain on a ship. Therefore, per network, there has to be one – and only one – controller. For this the firmware has to be adjusted, for instance with the X-CTU program. Apart from the controller and router types, there also exists an 'end-point' device, which is basically a router that works on low-power consumption mode and sleeps most of the time to conserve (battery) energy.

Setting up communication consists of making the XBees belong to the same network (PAN). Different networks can exist side by side in the same space. Only units with the same network ID can communicate with each other. Apart from this, the units also have to be on the same 'channel', which is the physical

**Table 5.XV**: Basic XBee AT commands

| Command | Description |
|---|---|
| `RE` | Reset device to factory settings |
| `WR` | Write parameters to non-volatile memory |
| `VR` | Echo firmware version |
| `CN` | Exit modem-setup mode |

carrier frequency (in the 2.4 GHz range) the modules will use for communication, very similar to WiFi computer-communication equipment. These channels are numbered from 11 to 26. (Note that not all channels are available to all types of XBees). We can directly specify the channel or specify the choice of channels that are allowed to be used and leave it up to the coordinator to decide which one is best. This is especially useful if we have a multi-network environment; the coordinator will select the channel with highest performance. Apart from that, each unit has an address which will be attributed by the controller. In summary, the following items are relevant:

- The network ID (also known as PAN)
- The channel
- The address

Commands can be specified to the XBee by putting it into modem-setup mode. Just like for legacy Hayes modems, this is achieved by typing "+++" (without the quotes) and waiting 1 second. The XBee will reply with "OK" to signal it is now in modem-setup mode. We can now specify commands. This is done by writing "AT" followed by the command, for instance "ATID" to query the network ID of the XBee module, to which it replies with something like "234". The basic commands are given in Table 5.XV. It shows, for instance, how the device parameters can be reset to factory values (RE). Don't forget to write the settings to non-volatile memory (WR), otherwise all settings will be lost forever when removing the power. To exit modem-set-up mode, we can use the 'CN' command, or wait 10 seconds idle to exit automatically.

A list of some useful AT-commands for setting up an XBee is given in Table 5.XVI. When you have set up the network correctly, the coordinator will blink slowly (once per second) and the router rapidly (twice per second). Congratulations, a network has been formed. If we had mounted both XBees on Arduinos with them both talking to the FTDI/USB port (rightmost situation in Fig. 5.54), we could now type on one computer and the text will appear on the other.

The XBee also has ADC inputs for sampling of sensors, so it is a little like an Arduino itself. It is thus a little overkill for our projects (Nordic Semiconductor has cheaper ZigBee devices with less functionality), but nevertheless interesting for our wireless communication survey. Also since there exists an Arduino community out there to help us when we have any problem with the XBee.

**Table 5.XVI**: XBee Command Reference Table (most relevant set-up commands)

| AT command | Description | Parameter range | Default |
|---|---|---|---|
| `ID` | Network choice ID. 64 bit. If set to 0 the coordinator will select a random ID and the router will join any available network | 0 - F..F | 0 |
| `OP` | Echo actual operating network ID (64 bit; `OI` is 16 bit version). If ID > 0, ID = OP | | read only |
| `SC` | Range of allowed channels. Bitpattern ('0' = disabled, '1' = allowed), bit 0 = channel 11, bit 1 = channel 12 ... bit 15 = channel 26. Not all XBee models can use all channels | 0 - FFFF | 1FFE |
| `CH` | Echo actual operating channel | 0B - 1A | read only |
| `DH` | Destination address high (32 bit) | 0 - F..F | 0 |
| `DL` | Destination address low (32 bit) | 0 - F..F | 0 (router)<br>FFFF (coordinator) |
| `SH` | Serial number high (32 bit) | 0 - F..F | factory |
| `SL` | Serial number low (32 bit) | 0 - F..F | factory |
| `MY` | Echo actual address of module, 16 bit (FFFE means no address obtained yet) | 0 - FFFE | read only |

**Excercise**: Set up communication between two computers via Arduino/XBee pairs.

**Excercise**: Make an Arduino send alternating "0" and "1" once per second to a computer connected to an Arduino/XBee.

**Hint**: In some cases, the XBees communicate but it is intermittent, with pauses of up to ten seconds or so. In this case, you must set the destination address of the sender (see `DH` and `DL`) equal to the serial number of the receiver (see `SH`, `SL`).

**Hint**: It is very likely that you will brick the XBee when configuring it. It will no longer communicate with your computer running X-CTU in any way. Don't panic. It is possible to recover your XBee. If you have an official XBee programmers kit, the procedure is quite simple. If you have only your Arduino with XBee shield, it can involve some soldering, but it is possible. The exact procedure, however, depends on the hardware (Arduino and XBee). You will rapidly find the correct way of un-bricking it on the internet.

### 5.5.15 Arduino Internet

The next step is to include some internet functionality and interactivity to our Arduino-based system. We can think here about things like monitoring the state

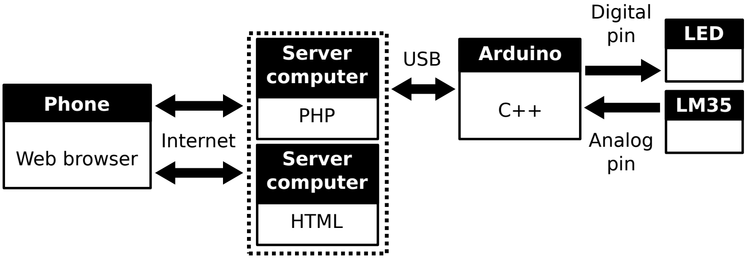


**Fig. 5.55**: Internet application where a telephone switches on and off a LED and reads the temperature. The server computer hosting the HTML file and the one with the PHP program can be the same or can be in other parts of the world. The computer with the PHP script is physically connected to the Arduino and communicates with it through USB. The Arduino switches the LED connected to a digital pin and reads the LM35 temperature sensor connected to an analog input pin

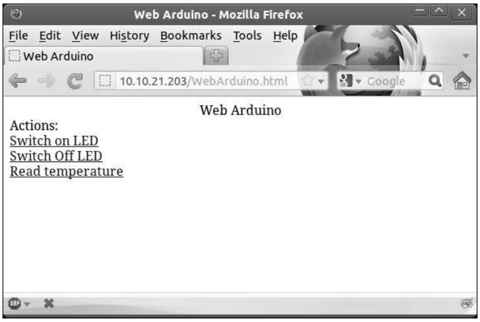
of our house, including the temperature, the state of the doors and windows, etc. And this combined with interactivity, i.e., the possibility to switch on the heater via an application on a mobile telephone. In a subsection above, the Arduino itself served as web host and this was adequate for small web applications. However, for larger systems, we want a more powerful webserver, basically a standard PC. This will be treated in this subsection. Figure 5.55 shows a schematic of a simple system that switches on and off an LED and reads the temperature through the internet.

On the mobile telephone a standard web browser is opened, for instance Firefox. It loads a page that is stored on the server computer at home, see Figure 5.56. (Note that the computer has to be accessible and visible to the rest of the world and has a known IP address. And be careful with security; the rest of the world will be able to operate your Arduino). The page is written in standard HTML. It has three hyperlinks `<a...>...</a>`, the syntax of a hyperlink in HTML is the following

```
<a href="linked-page">Clickable text</a>
```

In this case, the first two hyperlinks point to the same ’page’, a program (PHP) file called `switchLED.php` that also resides on the server in the same directory (it can also reside somewhere else, but that host computer must be physically connected to the Arduino). To pass information to that PHP program we can use the link itself by adding a question mark and a list of informations separated by commas. In this case we only have one piece of information, the value of a parameter `x` that takes a value 1 if we want our Arduino to switch on the led and 0 when it should be switched off. So, for instance, the link to switch on the LED is

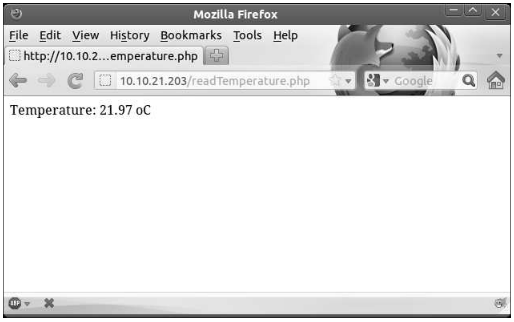
```
switchLED.php?x=1
```


Web Arduino - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Web Arduino
10.10.21.203/WebArduino.html
Google

Web Arduino

Actions:
Switch on LED
Switch Off LED
Read temperature

**Fig. 5.56**: Screen shot of the HTML page in a Firefox browser showing the three links


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://10.10.2...emperature.php
10.10.21.203/readTemperature.php
Google

Temperature: 21.97 oC

**Fig. 5.57**: Browser screen shot after a click on the 'read temperature' link

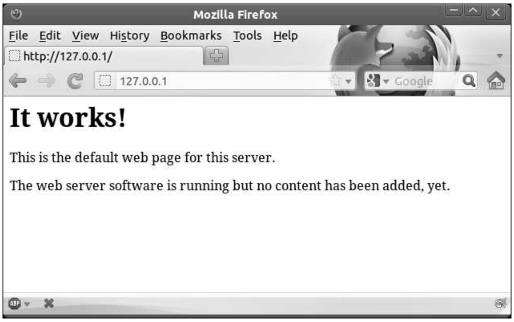
The PHP program consists of checking if the variable x is passed (by the function call isset). If not, the program exits. If it exists, the value of x is written to the serial port. For this, the object-oriented PHP has to create an instance of the phpSerial class, initialize it and execute the method sendMessage with the value of x as a parameter.

The third hyperlink of the HTML page points to a PHP program, that queries the Arduino for a temperature reading of a connected sensor and returns it to the calling page, see Figure 5.57.

The Arduino program constantly monitors the serial line, receives the 'command' and reacts accordingly. Changing the state of the LED of digital pin 13 when received a '0' or a '1', and returning the value of the temperature measured with an LM35 sensor connected to analog input 0 when receiving a 't'.

Before we begin we have to do the following steps. These steps are specifically for Linux (Ubuntu and Mint). For Windows they might be slightly different:

- Make sure that we have the correct software on the computers. For the client computer (the one that the final user uses to control the Arduino,



# It works!

This is the default web page for this server.

The web server software is running but no content has been added, yet.

**Fig. 5.58**: Opening a browser on the server computer and point it to the home page of the computer itself ('local host' = `http://127.0.0.1/`)

for instance a smart phone) not much is needed. We need just a simple HTML browser. Any one will do. On the server computer we must have HTML-server (for instance Apache) and PHP-server software. This can be done in the Software Center of the host computer; like all programs found there, they are free of charge!

- Test the server software by opening a browser on that computer and point it to itself ('local host') by typing `http://127.0.0.1/` in the URL field. It should load a page saying "It works!", see Figure 5.58.
- Test the same page from the client computer/smart phone. Open a terminal on the host computer and find its IP address by typing `ifconfig`. Now on the client computer open a bowser with that IP in the URL field, for instance `http://10.10.21.203/` and it should show the same page with "It works!".
- Make sure the permissions of the USB port to which the Arduino is connected are set correctly.

  ```
  ls -l /dev/ttyUSB*
  ```

  will show the USB ports and their permissions. If there is not three times written a 'wr-' pair, not everybody has permission to write there. We can change that by typing

  ```
  sudo chmod a+rw /dev/ttyUSB*
  ```
- Put the HTML and PHP files in the directory `/var/www/`

In total the following files exists:

file `WebArduino.html` on the server:

```
<html><head>
  <meta name="Author" content="Peter Stallinga">
  <title>Web Arduino</title>
</head>
```

```
<body>
<center>Web Arduino</center>
Actions:<br>
<a href="switchLED.php?x=1">Switch on LED</a><br>
<a href="switchLED.php?x=0">Switch Off LED</a><br>
<a href="readTemperature.php">Read temperature</a><br>
</body></html>
```

file `switchLED.php` on the server:

```
<?php
if(isset($_GET['x'])){
  $x = $_GET['x'];
  // open serial port for write:
  $fp = fopen("/dev/ttyUSB0", "w");
  fwrite($fp, $x);
  fclose($fp);
  if ($x==1)
    echo "<html>LED switched on</html>";
  else if ($x==0)
    echo "<html>LED switched off</html>";
  else
    echo "<html>Illegal value for x</html>";
}
else
  echo "<html>Illegal parameter</html>";
?>
```

file `readTemperature.php` on the server:

```
<?php
  // open serial port for read and write:
  $fp = fopen("/dev/ttyUSB0", "r+");
  fwrite($fp, 't');
  $c = '\0';
  $txt = "";
  while ($c!=chr(10)){
    $c = fgetc($fp);
    if ( ($c!=chr(10)) and ($c!=chr(13)) )
      $txt = $txt . $c;
  }
  fclose($fp);
  echo "<html>Temperature: $txt oC</html>";
?>
```

On the Arduino:

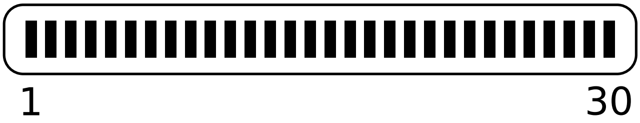
**Table 5.XVII**: Some mobile telephone modem commands

| Command | Meaning |
|---|---|
| `AT+CGMI` | Output manufacturer |
| `AT+CGMM` | Output telephone model |
| `AT+CGMR` | Output telephone revision |
| `AT+CGSN` | Output serial number (IMEI) |
| `AT+CPIN=`<pin> | Enter PIN code |
| `AT+CMGF=`<n> | Set message format ($n = 0$: PDU, $n = 1$: TXT) |
| `AT+CMSS=`<index> | Send SMS stored in memory (element <index>) |
| `AT+CMGS=`<number>[CR]<br>SMS Text[CTRL-Z]/[ESC] | Send "SMS Text" to <number> |
| `AT+CMGR=`<index> | Read SMS from memory <index> |

```
void setup(){
  Serial.begin(9600);
  pinMode(13, OUTPUT);
}

void loop(){
  unsigned char state;
  float t;

  if (Serial.available()>0){
    state = Serial.read();
    if (state=='0')
      digitalWrite(13, LOW);
    else if (state=='1')
      digitalWrite(13, HIGH);
    else if (state=='t'){
      // LM35: 0.01 V/degree, ADC: 5 V = 1024
      t = (5.0*((float) analogRead(0))/1024.0)/0.01;
      Serial.println(t);
    }
  }
}
```

### 5.5.16 Arduino Mobile telephone SMS

It is very easy to connect a telephone to the Arduino to make it send SMSs. Most mobile telephones have modem compatible with standard Hayes modem protocol, also sometime called 'AT command' because every command send to the modem starts with 'AT'. See the section on modems. To enter mobile-telephone specific commands, the AT prefix is extended by '+C'. A typical command is thus



1 GND
2 Video GND
3 R audio out
4 L audio out
5 R audio in
6 L audio in
7 -
8 S-video (Pb) out
9 S-video (Y) out
10 S-video (Pr) out
11 Audio switch
12 Tx
13 Rx
14 -
15 GND
16 USB GND
17 -
18 3.3V
19 12V (firewire)
20 12V (firewire)
21 Accesory
22 Firewire TPA-
23 5V USB
24 Firewire TPA+
25 USB TP-
26 Firewire TPB-
27 USB TP+
28 Firewire TPB+
29 Firewire GND
30 Firewire GND

**Fig. 5.59**: iPod/iPhone/iPad pinout

AT+CGMI

to which the telephone replies with the manufacturer name. Table 5.XVII gives a summary of some extended-Hayes mobile telephone modem commands. Probably the most interesting for us is the possibility to send an SMS from our Arduino:

AT+CMGS=<number>[CR]SMS Text[CTRL-Z]/[ESC]

Which sends the text "SMS Text" to telephone number <number> (substitute the desired number). The [CR] and [CRTL-Z] are control codes, 'carriage return' (ASCII 13) and 'escape' (ASCII 27), respectively. In older phones, the things are a little more complicated because they only understand PDU format messages. The command to change it to TXT format will not work:

AT+CMGF=1

The idea is to find out in the datasheet of the telephone what is the pinout of the connector, and what kind of handshaking is used. Very likely, no handshaking whatsoever is used. Once this is known, rests just to insert these Arduino signals to these pins. See for instance the pinout on the iPod/iPhone in Figure 5.59.

```
/*******************************************
 *     Sending SMS with a mobile phone     *
 *******************************************/
void setup() {
  Serial.begin(9600);
}

void sendSMS(char *number, char *message)
{
```

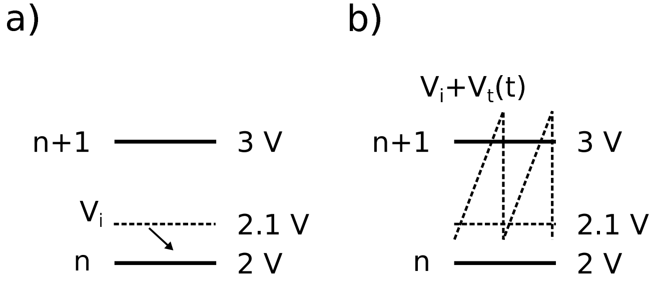


**Fig. 5.60**: a) Digitalization error. The input signal $V_i$ will always result in a rounded-down value. b) By adding a triangular wavefunction and using oversampling, the systematic error can be removed

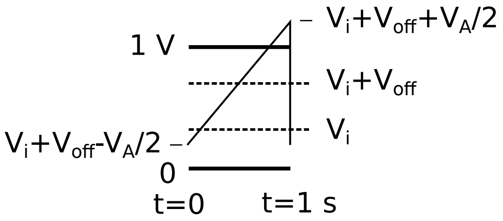
```
  Serial.println("AT+CMGF=1");
  delay(1);
  Serial.print("AT+CMGS=\"");
  Serial.print(number);
  Serial.println("\"");
  delay(1);
  Serial.print(message);
  Serial.print(modem,"\x1A");
  delay(1);
}
```

## 5.6 Exercises

1. Adding noise and using oversampling can increase the resolution of an ADC. Without the added signal, the ADC will always result in the same digital value, thus introducing digitalization noise, see for example the figure 5.60 where the input voltage is always round down causing a systematic error. Show that by adding a sawtooth or triangular wavefunction to the input signal and using oversampling this error can be removed. What is the amplitude and offset of the signal? (Assume the digitalization distance, the digital resolution equal $\Delta V = 1$ V and the ADC is rounding down to the nearest level). Can the same be achieved by other wave functions, or adding (white) noise?

## 5.7 Answers

1 For the sake of simplicity of the calculation it is assumed that the difference between two digital levels is 1 volt and that the lower level is 0. The rest is just a matter of scale. Also, the time scale is assumed to be 0-1 s, i.e., one sample per second. We get the following picture:



With offset and amplitude, the signal is

$$V(t) = V_\mathrm{i} + V_\mathrm{off} - V_\mathrm{A}/2 + V_\mathrm{A}t. \tag{5.6}$$

This line crosses the $V = 1$ level at $t = t_\mathrm{x}$ for which we find an expression

$$t_\mathrm{x} = \frac{1 - (V_\mathrm{i} + V_\mathrm{off}) + V_\mathrm{A}/2}{V_\mathrm{A}}. \tag{5.7}$$

For $t < t_\mathrm{x}$ the ADC value returned is 0, while for $t > t_\mathrm{x}$ the value is 1 volt. The average for an infinite number of t's actual is

$$V_\mathrm{ADC} = t_\mathrm{x} \times 0 + (1 - t_\mathrm{x}) \times (1\ \mathrm{V}). \tag{5.8}$$

We want this to be equal to $V_\mathrm{i}$, thus, after rearranging of terms:

$$V_\mathrm{i}\left(1 - \frac{(1\ \mathrm{V})}{V_\mathrm{A}}\right) = \frac{1}{2} + \frac{V_\mathrm{off} - (1\ \mathrm{V})}{V_\mathrm{A}}. \tag{5.9}$$

This should be true for *all* values of $V_\mathrm{i}$, and thus: $V_\mathrm{A} = 1$ V and $V_\mathrm{off} = 0.5$ V.