TFY 4190 Instrumentering

Institutt for fysikk, NTNU, 2010

January 25, 2011

Contents

Basic Vocabulary - Instrumentation

Accuracy: Smallest value measured with certainty.

Aliasing: Aliasing is an effect that causes different continuous signals to become indistinguishable (or aliases of one another) when under-sampled.

Aperture time The time during which the ADC is actively taking in signals for conversion.

Average mean: Mean value based on the average absolute value.

Baud Bits/second in communication. kB and MB are abbreviations used for kilobaud and megabaud.

BCD code A simple way to code a decimal number in binary number.

Bit A single binary digit, a "0" or a "1"

Block diagram An easy manner to represent a control/measurement system.

Buffer Memory or device which enables transferred data to be stored until it can be used by the receiving device, or sent by the sending device. Or a device that transforms a low output impedance signal to a high output impedance signal.

Bus A collection of conductors for transferring information within the computer.

Byte An eight-bit word (definition varies)

Clipping: Caused when the input signal exceeds the range of values that an A/D converter can represent at its output.

Closed control system Feedback is used for regulation.

CMRR Common mode rejection ratio, describes the ratio of amplification of the difference versus the common mode signal of a signal.

Control Use of an external influence to affect a physical parameter.

Daisychain Way to hook up instruments (along chain).

Differential linearity error: Error in amplitude step for one bit.

Digitization To represent quantities in digits

Dynamic range Resolution (amplitude/value of least significant value)

$e(t), E(s)$ Error signal.

Folding frequency: Half the sampling frequency, where folding will occur during sampling.

$g(t), G(s)$ Forward transfer function, often also called transfer function.

GPIB Old bus standard for instrument control. Max speed 1 Mbit/s.

Guarding To surround an object with a conductor held at the same potential to remove any stray capacitance to that object.

Gray code A code with little value change with changes in a single bit.

$h(t), H(s)$ Feedback transfer function.

Integration effect: When a sampler has a non-zero width in which the sample is measured.

Jitter Deviation from precisely-accurate sample timing intervals.

Linearity error: Error due to nonlinear conversion.

Lock-in amplifier Amplifier designed to send, pick up and analyze signals at a very narrow band-width.

LSB Least significant bit.

Measurement ground Ground connected to obtain an equipotential through a whole measurements system, often connected at one point to the safety ground.

MSB: Most significant bit.

Multi-drop When a device can contact many instruments in parallel at the same time.

Noise figure NF is defined as the logarithm of the ratio between the input signal-to-noise and the output signal to noise ratio.

Offset There is an bias added to the real signal.

Open control system No feedback is used for the control.

Parity For error correcting code an extra bit is added to always keep odd or even parity.

PID regulation The most simple and most used regulation system Proportional, integral and differential feedback is used to regulate.

Pull up resistance Needed to transform the open/close current logic of TTL to voltage signals.

Quantization error: round-off error introduced by representing each sample as an integer at the output of an A/D converter. For periodic signals $\sim 0.29Q$, where Q is the resolution.

$r(t), R(s)$ Reference input.

Regulation Use of an automatic system to control a physical value to certain set-point

Resolution: Smallest resolvable value.

Root Mean Square (RMS): Mean value based on the squared value, relates directly to the energy dissipation in electric circuits.

Safety ground Ground with the purpose to safeguard humans

Sampling theorem: States that to sample properly sampling frequency has to more than double the sampled frequency.

Shielding To surround an object with a conductor to screen away any effects of surrounding fields.

Skew Timing errors in a parallel signals.

Slew rate limit error: Error caused by an inability for an a/d converter output value to change sufficiently rapidly.

Slope error/amplification error: The conversion is made with an amplification error, this error linear with the amplitude of the signal.

TTL Transistor to transistor logic, very old standard for transferring information. Is mainly used for triggering nowadays.

Two complementary A common way to represent negative integers to avoid using subtractions in the processor.

USB Universal serial bus. Commonly used bus standard.

Word A series of bits grouped together to represent a number

$y(t), Y(s)$ Controlled variable.

Chapter 1

Introduction

This compendium is still in an evolving phase (as the field of instrumentation itself). It recollects important aspects of instrumentation from a large number of sources. Each chapter is followed by repetition questions which define what should be working knowledge after the course and during the evaluation.

Chapter 2

Measurements and Standards

Modern instrumentation has been undergoing a slow revolution due to the advancements of solid state electronics, circuit integration and the semiconductor industry. This has led to three major advancements:

- Precision. We can now measure with much higher precision than before (both in amplitude and time.
- Automated measurements. There is no longer a need for time-consuming procedures for storing data anymore.
- Data treatment. Unthinkable data flows are now possible through computers and buses, and computers are capable of treating these data and transforming it to meaningful information.

Modern instrumentation is the art of making use of these advantages for cheap and good measurements.

2.1 Standards

2.1.1 Measurements

Measuring is comparing to a well-known standard. The art of instrumentation is to do that with the best and the cheapest tools available to the best precision and accuracy possible. To assure that we are measuring the same quantities at any time, standards have been developed (the most well-known being the ones developed during the French revolution: the kilogram and the meter). Performing careful measurements involves minimizing the probability of error and realizing what errors actually will influence your measurements. As an example, if you are only interested in changes in magnitude of a certain measurement, it is unnecessary to have a traceable calibration to a national and international standard.

2.1.2 Standards

Through the SI system a chain of measurement standards is today maintained, and calibrated instruments from this chain of calibration can be bought. In addition to this measurement institutes and companies offer standard measurements and calibration of instruments. The standard system is adopted by international committees and if possible it is based on a physical principle or setup which (hopefully) will not change as time goes by. In Norway, Justervesenet (JV), the Norwegian Metrology Service, is responsible for legal metrology and the development and maintenance of national measurement standards in Norway. JV is an agency under the Department of Trade and Industry.

In the SI system There are four main types of standards in the international chain of standards:

- International standards: Defined by international agreements, kept by *International Bureau of Weights and Measures in France*. They are not used on a daily basis for calibration.
- Primary standards: Are the standards that are used for calibration within national laboratories. They are also the standards based on physical principles, which are compared between different countries to obtain good averaged values of world standards. These are used on a daily basis.
- Secondary standards: standards which are used as defacto standards when calibrating instruments that are to be used as working standards. They are kept strictly calibrated towards the primary standards.
- Working standards: standards that actually are used for calibration of ordinary instruments.

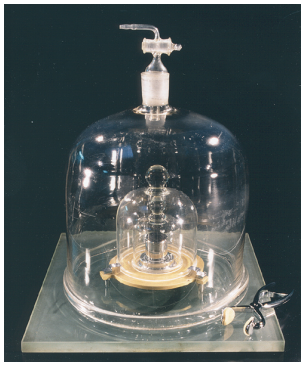
The following includes the current short version of the information brochure on the SI system:

A concise summary of the International System of Units, the SI

Metrology is the science of measurement, embracing all measurements, made at a known level of uncertainty, in any field of human activity.

The Bureau International des Poids et Mesures, the BIPM, was established by Article 1 of the Convention du Mètre, on 20 May 1875, and is charged with providing the basis for a single, coherent system of measurements to be used throughout the world. The decimal metric system, dating from the time of the French Revolution, was based on the metre and the kilogram. Under the terms of the 1875 Convention, new international prototypes of the metre and kilogram were made and formally adopted by the first Conférence Générale des Poids et Mesures (CGPM) in 1889. Over time this system developed, so that it now includes seven base units. In 1960 it was decided at the 11th CGPM that it should be called the Système International d'Unités, the SI (in English: the International System of Units). The SI is not static but evolves to match the world's increasingly demanding requirements for measurements at all levels of precision and in all areas of science, technology, and human endeavour. This document is a summary of the **SI Brochure**, a publication of the BIPM which is a statement of the current status of the SI.

The seven **base units** of the SI, listed in Table 1, provide the reference used to define all the measurement units of the International System. As science advances, and methods of measurement are refined, their definitions have to be revised. The more accurate the measurements, the greater the care required in the realization of the units of measurement.



The international prototype of the kilogram, $\mathcal{K}$, the only remaining artefact used to define a base unit of the SI.

Table 1 *The seven base units of the SI*

Quantity	Unit, symbol: definition of unit
length	**metre, m:** The metre is the length of the path travelled by light in vacuum during a time interval of 1/299 792 458 of a second. *It follows that the speed of light in vacuum, c_0, is* 299 792 458 m/s *exactly*.
mass	**kilogram, kg:** The kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram. *It follows that the mass of the international prototype of the kilogram, $m(\mathcal{K})$, is always* 1 kg *exactly.*
time	**second, s:** The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom. *It follows that the hyperfine splitting in the ground state of the caesium* 133 *atom, ν*(hfs Cs)*, is* 9 192 631 770 Hz *exactly.*
electric current	**ampere, A:** The ampere is that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross-section, and placed 1 metre apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newton per metre of length. *It follows that the magnetic constant, μ_0, also known as the permeability of free space is* $4\pi \times 10^{-7}$ H/m *exactly.*
thermodynamic temperature	**kelvin, K:** The kelvin, unit of thermodynamic temperature, is the fraction 1/273.16 of the thermodynamic temperature of the triple point of water. *It follows that the thermodynamic temperature of the triple point of water, T_{tpw}, is* 273.16 K *exactly.*
amount of substance	**mole, mol:** 1. The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon 12. 2. When the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, other particles, or specified groups of such particles. *It follows that the molar mass of carbon* 12*, $M(^{12}C)$, is* 12 g/mol *exactly.*
luminous intensity	**candela, cd:** The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} hertz and that has a radiant intensity in that direction of 1/683 watt per steradian. *It follows that the spectral luminous efficacy, K, for monochromatic radiation of frequency* 540×10^{12} Hz *is* 683 lm/W *exactly.*

The seven **base quantities** corresponding to the seven **base units** are length, mass, time, electric current, thermodynamic temperature, amount of substance, and luminous intensity. The **base quantities** and **base units** are listed, with their symbols, in Table 2.

Table 2 *Base quantities and base units used in the SI*

Base quantity	Symbol	Base unit	Symbol
length	l, h, r, x	metre	m
mass	m	kilogram	kg
time, duration	t	second	s
electric current	I, i	ampere	A
thermodynamic temperature	T	kelvin	K
amount of substance	n	mole	mol
luminous intensity	I_v	candela	cd

All other quantities are described as **derived quantities**, and are measured using **derived units**, which are defined as products of powers of the **base units**. Examples of **derived quantities** and **units** are listed in Table 3.

Table 3 *Examples of derived quantities and units*

Derived quantity	Symbol	Derived unit	Symbol
area	A	square metre	m^2
volume	V	cubic metre	m^3
speed, velocity	v	metre per second	m/s
acceleration	a	metre per second squared	m/s^2
wavenumber	$\sigma, \tilde{\nu}$	reciprocal metre	m^{-1}
mass density	ρ	kilogram per cubic metre	kg/m^3
surface density	ρ_A	kilogram per square metre	kg/m^2
specific volume	v	cubic metre per kilogram	m^3/kg
current density	j	ampere per square metre	A/m^2
magnetic field strength	H	ampere per metre	A/m
concentration	c	mole per cubic metre	mol/m^3
mass concentration	ρ, γ	kilogram per cubic metre	kg/m^3
luminance	L_v	candela per square metre	cd/m^2
refractive index	n	one	1
relative permeability	μ_r	one	1

Note that refractive index and relative permeability are examples of dimensionless quantities, for which the SI unit is the number one, 1, although this unit is not written.

Some **derived units** are given a **special name**, these being simply a compact form for the expression of combinations of **base units** that are used frequently. Thus, for example, the joule, symbol J, is by definition equal to $m^2\ kg\ s^{-2}$. There are 22 special names for units approved for use in the SI at present, and these are listed in Table 4.

Table 4 *Derived units with special names in the SI*

Derived quantity	Name of derived unit	Symbol for unit	Expression in terms of other units
plane angle	radian	rad	$m/m = 1$
solid angle	steradian	sr	$m^2/m^2 = 1$
frequency	hertz	Hz	s^{-1}
force	newton	N	$m\ kg\ s^{-2}$
pressure, stress	pascal	Pa	$N/m^2 = m^{-1}\ kg\ s^{-2}$
energy, work, amount of heat	joule	J	$N\ m = m^2\ kg\ s^{-2}$
power, radiant flux	watt	W	$J/s = m^2\ kg\ s^{-3}$
electric charge, amount of electricity	coulomb	C	s A
electric potential difference	volt	V	$W/A = m^2\ kg\ s^{-3}A^{-1}$
capacitance	farad	F	$C/V = m^{-2}\ kg^{-1}\ s^4\ A^2$
electric resistance	ohm	Ω	$V/A = m^2\ kg\ s^{-3}\ A^{-2}$
electric conductance	siemens	S	$A/V = m^{-2}\ kg^{-1}\ s^3\ A^2$
magnetic flux	weber	Wb	$V\ s = m^2\ kg\ s^{-2}A^{-1}$
magnetic flux density	tesla	T	$Wb/m^2 = kg\ s^{-2}\ A^{-1}$
inductance	henry	H	$Wb/A = m^2\ kg\ s^{-2}\ A^{-2}$
Celsius temperature	degree Celsius	°C	K
luminous flux	lumen	lm	cd sr = cd
illuminance	lux	lx	$lm/m^2 = m^{-2}$ cd
activity referred to a radionuclide	becquerel	Bq	s^{-1}
absorbed dose, specific energy (imparted), kerma	gray	Gy	$J/kg = m^2\ s^{-2}$
dose equivalent, ambient dose equivalent	sievert	Sv	$J/kg = m^2\ s^{-2}$
catalytic activity	katal	kat	s^{-1} mol

Although the hertz and the becquerel are both equal to the reciprocal second, the hertz is only used for cyclic phenomena, and the becquerel for stochastic processes in radioactive decay.

The unit of Celsius temperature is the degree Celsius, °C, which is equal in magnitude to the kelvin, K, the unit of thermodynamic temperature. The quantity Celsius temperature t is related to thermodynamic temperature T by the equation $t/°C = T/K - 273.15$.

The sievert is also used for the quantities directional dose equivalent and personal dose equivalent.

The last four special names for units in Table 4 were adopted specifically to safeguard measurements related to human health.

For each quantity, there is only one SI unit (although it may often be expressed in different ways by using the special names). However the same SI unit may be used to express the values of several different quantities (for example, the SI unit J/K may be used to express the value of both heat capacity and entropy). It is therefore important not to use the unit alone to specify the quantity. This applies both to scientific texts and also to measuring instruments (i.e. an instrument read-out should indicate both the quantity concerned and the unit).

Dimensionless quantities, also called quantities of dimension one, are usually defined as the ratio of two quantities of the same kind (for example, refractive index is the ratio of two speeds, and relative permittivity is the ratio of the permittivity of a dielectric medium to that of free space). Thus the unit of a dimensionless quantity is the ratio of two identical SI units, and is therefore always equal to one. However in expressing the values of dimensionless quantities the unit one, 1, is not written.

Decimal multiples and sub-multiples of SI units

A set of prefixes have been adopted for use with the SI units, in order to express the values of quantities that are either much larger than or much smaller than the SI unit used without any prefix. The SI prefixes are listed in Table 5. They may be used with any of the **base units** and with any of the **derived units** with special names.

Table 5 *The SI prefixes*

Factor	Name	Symbol	Factor	Name	Symbol
10^{1}	deca	da	10^{-1}	deci	d
10^{2}	hecto	h	10^{-2}	centi	c
10^{3}	kilo	k	10^{-3}	milli	m
10^{6}	mega	M	10^{-6}	micro	μ
10^{9}	giga	G	10^{-9}	nano	n
10^{12}	tera	T	10^{-12}	pico	p
10^{15}	peta	P	10^{-15}	femto	f
10^{18}	exa	E	10^{-18}	atto	a
10^{21}	zetta	Z	10^{-21}	zepto	z
10^{24}	yotta	Y	10^{-24}	yocto	y

When the prefixes are used, the prefix name and the unit name are combined to form a single word, and similarly the prefix symbol and the unit symbol are written without any space to form a single symbol, which may itself be raised to any power. For example, we may write: kilometre, km; microvolt, μV; femtosecond, fs; 50 V/cm = 50 V $(10^{-2}\ \mathrm{m})^{-1}$ = 5000 V/m.

When the **base units** and **derived units** are used without any prefixes, the resulting set of units is described as being **coherent**.

The use of a coherent set of units has technical advantages (see the **SI Brochure**). However the use of the prefixes is convenient because it avoids the need to use factors of 10^{n} to express the values of very large or very small quantities. For example, the length of a chemical bond is more conveniently given in nanometres, nm, than in metres, m, and the distance from London to Paris is more conveniently given in kilometres, km, than in metres, m.

The kilogram, kg, is an exception, because although it is a **base unit** the name already includes a prefix, for historical reasons. Multiples and sub-multiples of the kilogram are written by combining prefixes with the gram: thus we write milligram, mg, not microkilogram, μkg.

Units outside the SI

The SI is the only system of units that is universally recognized, so that it has a distinct advantage in establishing an international dialogue. Other units, i.e. non-SI units, are generally defined in terms of SI units. The use of the SI also simplifies the teaching of science. For all these reasons the use of SI units is recommended in all fields of science and technology.

Nonetheless some non-SI units are still widely used. A few, such as the minute, hour and day as units of time, will always be used because they are so deeply embedded in our culture. Others are used for historical reasons, to meet the needs of special interest groups, or because there is no convenient SI alternative. It will always remain the prerogative of a scientist to use the units that are considered to be best suited to the purpose. However when non-SI units are used, the conversion factor to the SI should always be quoted. A few non-SI units are listed in Table 6 below with their conversion factors to the SI. For a more complete list, see the **SI Brochure**, or the BIPM website.

Table 6 *A few non-SI units*

Quantity	Unit	Symbol	Relation to SI
time	minute	min	1 min = 60 s
	hour	h	1 h = 3600 s
	day	d	1 d = 86 400 s
volume	litre	L or l	1 L = 1 dm^3
mass	tonne	t	1 t = 1000 kg
energy	electronvolt	eV	1 eV ≈ 1.602×10^{-19} J
pressure	bar	bar	1 bar = 100 kPa
	millimetre of mercury	mmHg	1 mmHg ≈ 133.3 Pa
length	ångström	Å	1 Å = 10^{-10} m
	nautical mile	M	1 M = 1852 m
force	dyne	dyn	1 dyn = 10^{-5} N
energy	erg	erg	1 erg = 10^{-7} J

Symbols for units begin with a capital letter when they are named after an individual (for example, ampere, A; kelvin, K; hertz, Hz; coulomb, C). Otherwise they always begin with a lower case letter (for example, metre, m; second, s; mole, mol). The symbol for the litre is an exception: either a lower case

letter or a capital L may be used, the capital being allowed in this case to avoid confusion between the lower case letter l and the number one, 1.
The symbol for a nautical mile is given here as M; however there is no general agreement on any symbol for a nautical mile.

The language of science: using the SI to express the values of quantities

The value of a quantity is written as the product of a number and a unit, and the number multiplying the unit is the numerical value of the quantity in that unit. One space is always left between the number and the unit. For dimensionless quantities, for which the unit is the number one, the unit is omitted. The numerical value depends on the choice of unit, so that the same value of a quantity may have different numerical values when expressed in different units, as in the examples below.

The speed of a bicycle is approximately
$v = 5.0$ m/s = 18 km/h.
The wavelength of one of the yellow sodium lines is
$\lambda = 5.896 \times 10^{-7}$ m = 589.6 nm.

Quantity symbols are printed in an italic (slanting) type, and they are generally single letters of the Latin or Greek alphabet. Either capital or lower case letters may be used, and additional information on the quantity may be added as a subscript or as information in brackets.
There are recommended symbols for many quantities, given by authorities such as ISO (the International Organization for Standardization) and the various international scientific unions such as IUPAP and IUPAC. Examples are:

T	for temperature
C_p	for heat capacity at constant pressure
x_i	for the mole fraction (amount fraction) of species i
μ_r	for relative permeability
$m(\mathcal{K})$	for the mass of the international prototype of the kilogram $\mathcal{K}$.

Unit symbols are printed in a roman (upright) type, regardless of the type used in the surrounding text. They are mathematical entities and not abbreviations; they are never followed by a stop (except at the end of a sentence) nor by an s for the plural. The use of the correct form for unit symbols is mandatory, and is illustrated by the examples in the **SI Brochure**. Unit symbols may sometimes be more than a single letter. They are written in lower case letters, except that the first letter is a capital when the unit is named after an individual. However when the name of a unit is spelled out, it should begin with a lower case letter (except at the beginning of a sentence), to distinguish the unit from the man.

In writing the value of a quantity as the product of a numerical value and a unit, both the number and the unit may be treated by the ordinary rules of algebra. For example, the equation $T = 293$ K may equally be written $T/\mathrm{K} = 293$. This procedure is described as the use of quantity calculus, or the algebra of quantities. It is often useful to use the ratio of a quantity to its unit for heading the columns of tables, or labelling the axes of graphs, so that the entries in the table or the labels of the tick marks on the axes are all simply numbers. The example below shows a table of vapour pressure as a function of temperature, and the logarithm of vapour pressure as a function of reciprocal temperature, with the columns labelled in this way.

T/K	10^3 K/T	p/MPa	ln(p/MPa)
216.55	4.6179	0.5180	−0.6578
273.15	3.6610	3.4853	1.2486
304.19	3.2874	7.3815	1.9990

Algebraically equivalent forms may be used in place of 10^3 K/T, such as kK/T, or 10^3 (T/K)$^{-1}$.

In forming products or quotients of units the normal rules of algebra apply. In forming products of units, a space should be left between units (or alternatively a half high centred dot can be used as a multiplication symbol). Note the importance of the space, for example, m s denotes the product of a metre and a second, but ms denotes a millisecond. Also, when forming complicated products of units, use brackets or negative exponents to avoid ambiguities. For example, the molar gas constant R is given by:

$$pV_m/T = R = 8.314 \text{ Pa m}^3 \text{ mol}^{-1} \text{ K}^{-1}$$
$$= 8.314 \text{ Pa m}^3/(\text{mol K}).$$

When formatting numbers the decimal marker may be either a point (i.e. a stop) or a comma, as appropriate to the circumstances. For documents in the English language a point is usual, but for many continental European languages and in some other countries a comma is usual.
When a number has many digits, it is customary to group the digits into threes about the decimal point for easy reading. This is not essential, but it is often done, and is generally helpful. When this is done, the groups of three digits should be separated only by a (thin) space; neither a point nor a comma should be used. The uncertainty in the numerical value of a quantity may often be conveniently shown by giving the uncertainty in the least significant digits in brackets after the number.

Example: The value of the elementary charge is given in the 2002 CODATA listing of fundamental constants as

$e = 1.602\ 176\ 53\ (14) \times 10^{-19}$ C,

where 14 is the standard uncertainty in the final digits quoted for the numerical value.

For further information see the BIPM website, or the **SI Brochure** 8th edition, which is available at

http://www.bipm.org

This summary has been prepared by the Comité Consultatif des Unités (CCU) of the Comité International des Poids et Mesures (CIPM), and is published by the BIPM.

March 2006

Ernst Göbel, President of the CIPM
Ian Mills, President of the CCU
Andrew Wallard, Director of the BIPM

2.2 Signal description

There are two main ways in which to describe signals: in the time domain or in the frequency domain. The time domain is the natural way for many humans to think of signals, while the frequency domain can be very constructive when studying the path of a signal that is measured or generated, and especially of there are frequencies where the system is not stable. Both are complementary descriptions, as any time domain signal can be described by a complete basis set in the frequency domain.

2.2.1 Amplitude domain -Accuracy, precision and errors

The simplest description of a signal is given by its amplitude. We then usually consider either the *Maximal amplitude*, the *Maximal Error* or the *Signal to noise ratio - Maximal amplitude/ Error*.

In order to describe errors in amplitude, we need to define our vocabulary. Two important expressions are the *accuracy* and the *precision*. The Accuracy is how well you can measure the exact value of what you want to measure, while the precision is how well you can repeat your value. The *error* is how much your value can deviate from the exact value. It is usually expressed as percent value, even though the absolute error can also be interesting. The *limiting error* is the highest error that can be obtained in the system.

If Y_n is the true value of the signal, and M_n is the measured value, the accuracy can be described as:

$$A_n = 1 - |\frac{Y_n - M_n}{Y_n}|$$

The precision will be:

$$P_n = 1 - |\frac{M_n - \bar{M}}{\bar{M}}|$$

We can also express the error:

$$e_n = X_n - Y_n$$

which makes more sense to express in terms of percent:

$$\%e_n = 100|\frac{e_n}{Y_n}|$$

This can be related to the accuracy expressed in percent as:

$$\%A_n = 100 - \%e_n$$

Usually, we can not relate to exact measurements, as we do not know what the true value is. We therefore rely on statistics, estimates and calibration to establish these three basic quantities.

2.2.2 Amplitude Errors in measurements

2.2.3 Error sources

Noise in measurements are mainly due to three sources:

- Noise inherent to the measurements.
- Noise due to the electrical processing after sensing.
- Noise due to the digital conversion process.

These can all be minimized by sensible choices in building and setting up the measuring system. Generally, they limit the resolution of the measurement system. In addition comes errors induced though human errors and faulty handling/maintaining of equipment:

- Errors due to reading before steady state is reached.
- Errors due to misinterpreting data and/or errors in subsequent data treatment.
- Using uncalibrated equipment.
- Misuse of instrument e.g. not matching impedances.

An important distinction must be made between systematic and stochastic errors. Stochastic errors are generally easier to account for since they will contribute to a stochastic distribution of measurement values, while systematic errors will give a repeatable error of the same magnitude, and can only be reduced by pinpointing the error source.

2.3 Simple error estimations

It is very important to be able to detect possible errors in our measurements. Either, we calculate them using the known transfer characteristics of the system, or we estimated them using statistics. Often one has to resort to statistics.

Errors from reading off instruments are becoming more scarce, as most instruments are electronically controlled and the data is read through A/D conversion.

2.3.1 Statistics

From statistics, we have two interesting quantities that are related to the precision, namely the variance s^2 and the standard deviation s:

$$s^2 = \frac{1}{n-1}\Sigma_{i=1}^{n}(M_i - \bar{M})^2$$

$$s = \sqrt{s^2}$$

To relate this property to the probability of finding an error within a certain interval, the statistical distribution must be known. As an example, in a Gaussian distribution the probability of finding a value within an error limit of $\pm 0.6745s$ is 50%.

It is also evident from the above equations, that in order to increase the resolution we can increase the number of measurements. However, this will give a rather slow $1/\sqrt{N}$ reduction of the error limits.

2.3.2 Error estimation

When making simple error estimations, it is often enough to use simple linear models of the errors within the system. To do such an analysis, we need to find a proper expression for the transfer function and the limiting errors of the different parts of the instrument. Usually, they are given by the uncertainties in components such as amplifiers, resistors and sensors. Consider a signal M that depends on variables X_n. The actual measured signal $\hat{M}$ is then given by the variables X_n and their errors ΔX_n as:

$$\hat{M} = M(X_1 \pm \Delta X_1, X_2 \pm \Delta X_2 ...)$$

We can expand this into a Taylor series and ignore any higher order terms. We are interested in the error so we look at the difference between the signal without error and with error:

$$\Delta M = |M - \hat{M}| = \Sigma_{J=1}^{m} |\frac{\partial M}{\partial X_j} \Delta X_j|$$

The difficult part in performing simple error estimations is usually to find the correct transfer function and to estimate the errors of different components. A simple example is given by calculating the error in the power emitted by a heater:

$$P = I^2 R$$

Here there are two possible sources of errors, the current I and the resistance R. We can easily find the maximum error:

$$\Delta P_{Max} = 2IR\Delta I + I^2 \Delta R$$

To find the relative error, we divide by the power:

$$\frac{\Delta P_{Max}}{P} = 2|\frac{\Delta I}{I}| + |\frac{\Delta R}{R}|$$

If the maximum error in the resistance is 0.1 %, the maximum error of the ampere meter is 1% of the full scale (0-10A) and we are measuring the current 8A over 100Ω, the limiting error is:

$$\frac{\Delta P_{Max}}{P} = 2\frac{.1}{8} + \frac{.1}{100} = 0.025 + 0.001 = 0.026$$

It is evident from such a calculation that the error could be corrected by almost an order of magnitude by obtaining a better ampere meter.

2.3.3 Averaging

One common route to get a good estimate of most low frequency (typically below 10-50 MHz) components of your signal is to measure the mean value of the time-varying signal. For this there are two standards, *Root Mean Squared (RMS)* or *Average Rectified*. The root mean squared signal is defined as:

$$U_{RMS} = \sqrt{\frac{1}{\tau} \int_0^\tau [U(t)]^2 dt}$$

This expression directly relates to the average power dissipation in a DC circuit since

$$P = IU = \frac{U^2}{R} = I^2 R = \frac{1}{R\tau} \int [U(t)]^2 dt = \frac{R}{\tau} \int_0^\tau [I(t)]^2 dt$$

Accordingly the RMS value corresponds to the current or potential that would induce the same effect as the AC value. The dissipated power content of the signal can be very important in real systems, if you e.g. want to replace the signal with an equal signal dissipating the same energy, or simply want to estimate the energy consumption of a system.

The average rectified value:

$$U_{av} = \frac{1}{\tau} \int_0^\tau \sqrt{[U(t)]^2} dt$$

does not contain the same physical information but is much easier to actually perform electronically. Comparing the actual values of both we can evaluate the

pulse train with differently sized portions (given by duty cycle, the factor η) high and low. Assume the low value is 0 and the high A. The rectified average then simply equates to :

$$|f|_{av} = \eta A$$

However when you equate the RMS value you end up with:

$$f_{RMS} = \sqrt{\eta} A$$

To convert between the two types of average forms the *Form Factor (FF)* is used. This is different for different types of signals. For sinusoidal signals it is simply

$$\frac{\pi}{2\sqrt{2}} = 1.11$$

. The tricky thing is that FF changes with the signal, the pulse train above ha a form factor of $1\sqrt{\eta A}$ hence no conversion can be given for an arbitrary input signal.

The Average rectified value can easily be obtained through rectification through diodes, and a simple low-pass filter. True RMS is much harder to find, one important factor for correctly finding the RMS value is the *Crest Factor (CF)* which simply is defined as:

$$CF = \frac{|f|_{max}}{f_{RMS}}$$

The crest factor gives a number of how spiky the signal is, the higher crest factor the harder it is to obtain accurate measurements of any kind since there is a need for a higher dynamic reserve. Methods for conversion contains thermal techniques, analogue computation and digital computation.

2.3.4 Frequency domain description

The frequency domain description is very similar to the $j\omega$-method. However, in that case we only considered the harmonic solutions to an input: the sinusoidal stable solutions (basically the Fourier transform of the signals). Now we want to be able to describe any signal that is passing through the system.

A more suitable description in the frequency domain is then given by the Laplace transform. It is defined for any piece-wise continuous signal defined after a certain starting time, often set to $t = 0$. We observe that both the Fourier and Laplace transform has the same definition, except from two major details: the starting point ($-\infty$ compared to 0^-) and the integration domain (along $-\infty$ to ∞, compared to half the complex plane).

$$\mathcal{L}\{f(t)\} = \int_0^\infty e^{-st} f(t) dt$$

compared to

$$\mathcal{F}\{f(t)\} = \int_{-\infty}^\infty e^{-i\omega t} f(t) dt$$

The Laplace transform can describe all transfer functions, not only homogeneous. In many of the transfer functions obtained from ordinary circuit theory with the $j\omega$-method, we can simply replace the $j\omega$ in the transfer function with an s. This signifies the whole complex frequency plane of the transfer function. Typically we will encounter transfer functions of the type

$$\frac{V_{out}}{V_{in}} = \frac{K}{s + a}$$

which describes a simple low pass filter. This, or higher order low pass filters are the typical situation for most amplifier circuits. As an example of this can be any amplifier within a measurement system, it will impose an upper time limit in the response of the measurement system and the transfer function accordingly describes what limitations it will impose on the measurement system, something that is not as obvious in a time domain description of the same amplifier. Generally most systems can be characterised as low pass or band pass filters. More complex bandpass type signals can be found from sensors, which also have a low frequency response limit.

2.4 Laplace Transform

The Laplace transform is defined as:

$$F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty e^{-st} f(t) dt.$$

It is instructive to look at the expression for time convolution of the Laplace transform. This will yield a simple product in the transformed frequency domain. Remember the Laplace transform is defined from $t = 0$, ideal for problems with known starting values:

$$F(s)G(s) = \mathcal{L}\{f(t) * g(t)\} = \mathcal{L}\{\int_0^t f(t-\tau)g(\tau)d\tau\}.$$

Other important properties of the Laplace transform is the transform of the derivative and the integral:

$$\mathcal{L}[f'(t)] = \int_0^\infty \frac{df}{dt} e^{-st} dt = f e^{-st}|_0^\infty + s\int_0^\infty f(t)e^{-st}dt = sF(s) - f(0)$$

where partial integration was used at the second equality. In the same manner the expression for the integral can be found:

$$\mathcal{L}[\int_0^\infty f(t)dt] = \frac{F(s)}{s}.$$

We see that all integrations and derivations are reduced to algebraic expressions with the Laplace transform. These are relatively easy to handle, solving an differential equation then (simply?) reduces to a matter of finding and defining the system. Then the route is to perform Laplace transforming, reductions of the equations and finally transforming back the Laplace transform.

Limits are interesting to study how the system works, they can be evaluated in the frequency domain:

$$\lim_{t\to\infty} f(t) = \lim_{s\to 0} sF(s)$$

and

$$\lim_{t\to 0} f(t) = \lim_{s\to\infty} sF(s)$$

However, the first statement (final value theorem) might not always be true as it is only valid when the poles are in the left complex plane.

2.4.1 Linear response theory and why frequency space is natural for signal description

Above a typical system is depicted (fig. **??**). The system has a control variable $r(t)$ that renders an output $y(t)$ which is described by the transfer function $h(t)$. The

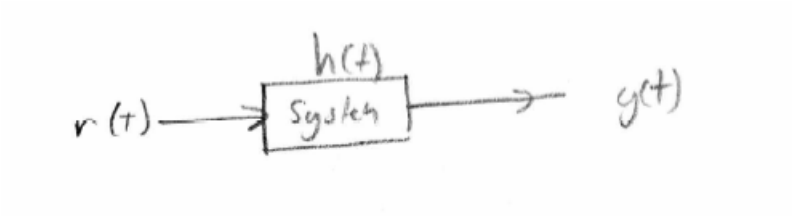


Figure 2.1: A simplistic interpretation of a general system characterised by the transfer function $h(t)$

transfer function describes the memory of the system at any point. However, to be able to analyse the system further we have to make two assumptions: linearity and time invariance.

- Linearity means that the relationship between the input and the output of the system satisfies the scaling and superposition properties. Formally, a linear system is a system which exhibits the following property: if the input of the system is
$$x(t) = Ax_1(t) + Bx_2(t),$$
then the output of the system will be
$$y(t) = Ay_1(t) + By_2(t),$$
for any constants A and B.
- Time invariance means that whether we apply an input to the system now or t seconds from now, the output will be identical, except a time delay of the t seconds. More specifically, an input affected by a time delay should affect a corresponding time delay in the output, hence time-invariant.

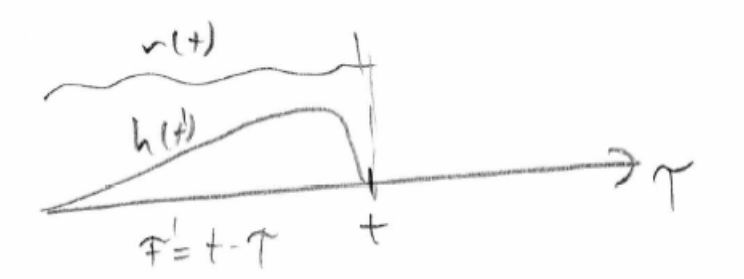


Figure 2.2: A schematic view of the two functions inside the convolution integral.

If a system behaves in this manner it is completely characterised by the response to a pulse (impulse response or $h(t)$). The output is simply the convolution of the input and the system's impulse response. For the system above we find the output as the convolution between the two

$$y(t) = \int_{-\infty}^{t} h(t-\tau)r(\tau)d\tau.$$

Accordingly we interpret $h(t)$ as a weight function defining what memory the system has of previous values (fig. 2.2). To find the value at a specific point, we have to

find the impulse response, and we can predict the behaviour of the system. This method of analysis is often called the time domain point-of-view. However it is rather inconvenient to always work with convolution integrals to find the response to an explicit excitation.

In the frequency domain the system is characterised by the system's transfer function, which is simply the Laplace transform of the system's impulse response. As a result of the properties of the Laplace transform, the output of the system in the frequency domain is the product of the transfer function and the Laplace transform of the input. In other words, convolution in the time domain is equivalent to multiplication in the frequency domain, and multiplication is much easier to perform. Thus the convolution written above for the corresponding system in the frequency domain is

$$Y(s) = H(s)X(s),$$

where $H(s), X(s)$ and $Y(s)$ are the Laplace transforms of $h(t), x(t)$ and $Y(t)$. As pointed out above this is only true for linear systems with constant coefficients.

An alternate way to consider this is as a system of differential equations. As long as these systems are linear differential equations with constant coefficients (linear and time invariant) the whole system can be solved by first Laplace transforming, finding the total response function of the system and then transforming back to time space. This offers a quick and safe route for solving these types of differential equations.

2.4.2 Description of systems: Block Diagrams

The most common way to describe a measurment system is through block diagrams, the natural mathematical formulation of the same system is then the Laplace transformed system since convolution can be described through simple multiplications that corresponds well to our perception on how a the signal path should be percieved, in the frequency. These provide a logical manner to analyse the system, and as you will see the time response can be evaluated straightforwardly utilising block diagrams and the Laplace transform. Block diagrams have three components:

- Block. A block contains an operation, typically multiplying by a constant, a filter or the physical response of a system.
- Connection. The connection between blocks is usually one-way, it forwards the output from one block to another one.
- Comparing device. Adds or subtracts signals.

Using these items a schematic description of almost any system can be made.

2.5 Repetition questions

1. Why is instrumentation important for you?
2. Explain the standard system?
3. What authorities (internationally and nationally) are responsible for the SI standards?
4. What are the base units of the SI system.
5. What are the general sources of errors, how are they detected and quantified?
6. What kind of information can be extracted from the statistics of an experiment?

7. How do you use linearization to make simple error estimates?
8. Calculate rectified-average and root-mean-squared values for simple signal forms.
9. Explain how the rms-value can be estimated from the rectified-average value of a signal.
10. Explain how the true rms- value can be obtained using electronic circuits.

Chapter 3

Sensors

This chapter will give an introduction to how to connect to sensors, and an overview of the type of sensors which are not based on micro-technology (which often are so specialised that there is no general type, but the performance is best read out from the data sheets.)

3.1 What is a sensor ?

Today much effort is put in to producing large amounts of miniaturised sensors for widespread applications. Through monitoring e.g. transport processes with a sensor it is possible to find out of any problems during the freight. This sensor explosion is often considered as an integrated part of the information technology. We will in the future have more and more detailed monitoring of more an more processes. To understand the possibilities and basic work mode of such sensors is vital to understand what information actually can be derived from measurements.

For understanding a general measurement situation it is also very important to understand the physical principle of the sensor, since it can affect the results. To minimise noise it is important to understand not only the sensor but also the whole signal conversion process.

A sensor as defined in this course is a device that can convert a quantity or magnitude into a electric signal, usually linearly dependent on the magnitude of the measured quantity. We are primarily interested in either mechanical or electrical quantities:

Table 3.1: Different quantities measured by sensors

Mechanical	Electrical
Motion and distance	Potential
Velocity	Current
Acceleration	Resistivity
Mass	Capacitance
Force	Inductance
Pressure	Magnetic fields
Flow	Charge
Temperature	

It is customary to divide sensors into active or passive sensors. The first need extra energy to work, while the latter convert energy from the quantity they are measuring:

- **Active** Active sensors convert energy from their measurement source, and are thus generally considered to disturb the measurement system more. Example: Piezoelectric sensors.
- **Passive** Passive sensors require an extra energy source, this energy is modulated by the sensor and the measuring system to obtain the signal. Example: Capacitor sensor.

3.1.1 Mathematical description of sensors

Sensors must as any part of a measure systems be described both in amplitude and frequency response. In general the manner which is mathematically most convenient to describe the system through is the transfer function of the system. However, this is rarely a practical entity to derive, and most commonly the transfer function is described by parameters and diagrams that together give a sufficient description of the performance. The amplitude conversion of sensors is usually described by the *sensitivity* which relates the measurement input to the measurement output. The response of a sensor will always be limited in bandwith, this is most commonly through a Amplitude/frequency diagram, or through simply stating the *bandwidth* of the sensor.

3.2 Connecting sensors

The important thing when connecting sensors is of course the type of output. A passive sensors usually changes a property like resistance, and must be connected in the correct manner to produce an output. Many sensors are also of a high impedance type, which means that correct connection is dependent on a high impedance input of the voltmeter used to measure the voltage. A typical example of such a sensor is the thermocouple which can not sustain any large currents, but need a high impedance input on the voltmeter to obtain a correct value. In general there are two main problems when connecting sensors:

- **Nulling offset:** Often there can be large offset levels in a sensor, a resistance based sensor will allways have a large base resistance that must be nulled by some sort of method. This can be done after digitization or amplification, but this will imply using a large portion of the range for a constant level, to obtain optimum resolution it is desirable to reduce the offset so that the full range can be used for registering signals of all levels. This can be done through the use of different electrical analog subtraction methods, as bridges described below.
- **Property conversion:** In many cases the sensing property is not a simple voltage, which easily can be connected to an electrical measurement system, but another entity like current, resistance, capacitance or inductance. Accordingly an circuit (passive or active) that converts this entity to a voltage must often be found. Some of them are described below, where especially different kind of bridges have been much used. Today intelligent use of modern electronic components can often replace these (like the Anderson loop).

3.2.1 Impedance matching

The first principal is to always match the impedance, that is to always make sure that your signal is not degraded by the impedance of the measuring device used to aquire signal from your sensor system. There are two ways to do this: to maximise transferred power - $Z_{meas} = Z^*_{sens}$, or to minimise disturbance - $Z_{meas} \gg Z_{sens}$.

3.2.2 Measurement bridges

There are a number of different kind of measurement bridges, the most commonly known is the Wheatstone bridge. Common for all of them is that they are used to measure signals from passive sensors that need excitation. In the Wheatstone case that includes resistive measurement devices. The main principle of a measurement bridge is to null out the offset imposed on the system by the drive current/potential through comparing it to an almost identical system.

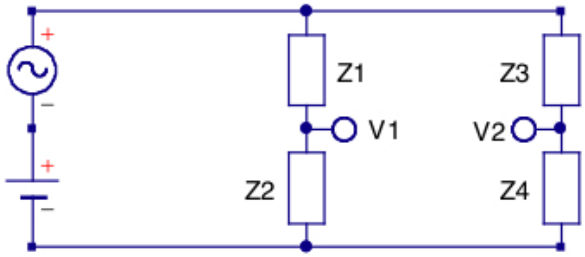


Figure 3.1: Schematic of the general layout of measurement bridges.

In general the equation for zero output ($V_{out} = V_1 - V2 = 0$) is:

$$\frac{Z_1}{Z_2} = \frac{Z_3}{Z_4}$$

To obtain the best sensitivity it can be shown that all impedances should be approximately the same. The most straightforward bridge to analyse is the wheatstone bridge which often is used in the configuration $R = R_1 = R_2 = R_4 \approx R_3$ where $R_3 = R + \delta R$ is the resistor that contains the sensor. This configuration is often called quarter bridge since only one quarter of the bridge is used for sensing. We then obtain an output voltage:

$$V_{out} = V_{bias}[\frac{R}{R+R} - \frac{R}{R+R+\delta R}]$$

To simplify interpretation we introduce the relative change in resistance:

$$\varepsilon = \frac{\delta R}{R}$$

We then find:

$$V_{out} = V_{bias}[\frac{1}{2} - \frac{1}{2+\varepsilon}]$$

This can be expanded in power series:

$$V_{out} = V_{bias}[\frac{\varepsilon}{4} - \frac{\varepsilon^2}{8}...]$$

Accordingly the change in bias with change in resistance is almost a linear relationship.

Famous types of bridges

There are many bridges that are useful, but remembering every bridge in detail is not really the point of this course, however, it should be noted that bridge measurements are one of the classical ways to measure R, L, or C with a high resolution and removing the background.

Wheatstone This bridge is used to interface resistive sensors, like strain gauges or gas sensitive meters.

Kelvin This bridge is used to measure very low resistances, like the ampere meter shunt resistors.

Schering This is a bridge particularly useful for measuring high loss capacitances.

De Sauty This is a simple bridge used to measure small changes in capacitance.

Wien This is not used in measurements, but as a frequency selective filter.

Maxwell This is a suitable bridge for measuring low Q value inductors, with Q in the range 0.02-10.

Hay This bridge is suitable for finding Q and L for high Q-value inductances.

3.2.3 Anderson loop

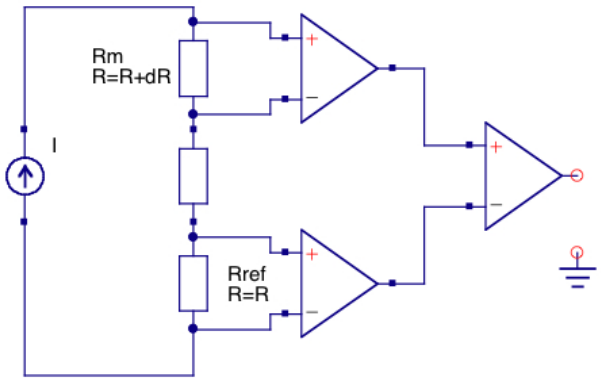


Figure 3.2: Schematic layout of the Anderson loop. The same current is sent through two measurement resistors, where one deviates slighly in value, the two voltages produced are compared and subtracted from each other. The ouput signal is only proportional to the difference in resistance between the two resistors.

In a modern measurements system, bridges can be replaced by Anderson loops. Instead of relying on identical base components, we can rely on using almost identical operational amplifiers. Instead of relying that the two nearly identical potential dividers will provide a nulling out effect, we here send the same current through two passive components and measure the potential using differential amplifiers with a fixed gain K. In a second stage this output is compared (the two resistances are nulled out against each other) through the third differential amplifier.The full output can then be written as:

$$V_{out} = KK(IR(1 + \delta R/R) - IR) = KKI\delta R$$

A linear relationship, with a very high common mode rejection ratio. This makes the circuit ideal for distributed sensor systems.

3.3 Temperature sensors

3.3.1 Resistive

Resistive temperature sensors utilise the change in resistance in materials due to the change in temperature. They are often called *thermistors*, and are either high purity metal sensors for precision measurements, or semiconducting or halfmetal devices for high sensitivity devices. Typically the resistance is expressed as an expansion around room temperature:

$$R(T) = 25R + \alpha(T - 25) + \beta(T - 25)^2$$

The temperature coefficient of a thermistor is defined as:

$$\alpha = \frac{dR(T)/DT}{R(T)}$$

at the desired temperature. Typical values are 0.00390 for platinum (used as measurement standard), and -0.05 for dedicated high sensitivity thermistors(which often are unlinear over larger temperature ranges).

3.3.2 Thermocouples/Thermopiles

Thermocuples are junctions between wires kept at different temperatures. The measuring principle is based on the Seebeck effect, that a current will flow in a loop made out of two materials if the two junctions are kept at different temperature (this in turn is dependent on the different electron density of the two materials, which will induce a asymmetry of the two materials, more of that in the solid state physics course). If the loop is broken we can instead measure a potential difference. Due to the temperature difference the electrons will migrate in the wire creating thermovoltages that are measurable at the boundaries. Thermopiles are stacks of thermocuples that multiply their efficiency.

When using thermocouples, the important thing is the temperature difference between the junctions. In fact, it is possible to work with a third metal as long as that connection is made at the same temperature. This simplifies connections since an ordinary copper wire can be used to connect to the measurement set-up and from that point the two different materials of the thermocouple can be used as long as the connections to the copper is kept at the same temperature. Most multimeters today come with a mode which reads some standard thermocouple directly in Kelvin.

Thermocouples are convenient, they operate over a large range but provide a rather small voltage ($\mu V/K$). Standard thermocouples are named by an letter, most common ones are: E, J, K, and T. It is usual to have use tables to look up what voltage difference corresponds to what temperature difference when using thermocouples (they are quite unlinear).

Type	Metal A - Metal B	Temperature Range	Sensitivity
E	Chromel-Constantan	-200 to +900 $°C$	$70\mu V/°C$
J	Iron-Constantan	0 to +750 °C	$55\mu V/°C$
K	Chromel-Alumel	-200 to +1250 $°C$	$40\mu V/°C$
T	Copper-Constantan	-200 to +350 $°C$	$50\mu V/°C$

Table 3.2: Four common type of thermocouples.

3.3.3 Diodes

It might be good to use diodes when high sensitivity is of essence, . The diode voltage/current dependence includes an exponential dependance on temperature which can be used for temperature monitoring. For the exact derivation of the epression of you will have to wait for the solid state physics course. However, when the diode is biased with a constant current (above the threshold current) it can be shown that the differential cange in voltage with temperature is:

$$\frac{dV}{dt} = \frac{V - V_g}{T} - \frac{3k_B}{q}$$

where V is the diode voltage, V_g the gap potential, and q the charge of the charge carrier. Typical values will render a sensitivity of 0.002 V/K. The drawback of diodes is that they dissipate heat.

3.4 Magnetic Fields

3.4.1 Magnetoresistive sensors

The Nobel prize 2007 was awarded the invention of modern magnetoresitive sensores which today are used in magnetic hard drives. The GMR (Giant MagntoResistive effect) sensors rely on that electrons of different spin will sense a higher resistance when sent through a two adjacent layers with anti parallel magnetisation compared parallel magnetisation. In essence the electrons will scatter much more if the magnetisation direction of the two adjacent layers are different, thereby increasing the resistance. Typical commercial senors wil induce a change in resistance in the order of 20% over a field variation of 100-200 Gauss. These sensors are good for reading out information but not particularly good for measuring exact fields. Typical measurement range is $10\mu T - 10^4 T$

Traditional magnetoresistive sensors rely on less responsive effects, the anistropic magnetoresistance (AMR), where the resistance depend on the relative orientation of the current with respect to the magentisation direction in the material. They are typically micromanufactured as balanced wheatstone bridges. Typical measurement range is $10pT - 1mT$

3.4.2 Hall effect sensors

When a charged particle moves along a solid it will interact with any transverse magnetic fields according to :

$$\vec{F} = q(\vec{v} \times \vec{B})$$

The induced force on the charge carriers is expressed as a voltage perpendicular to both the field and the current flow direction. The potential is directly proportional to the current and the applied magentic field:

$$V_H = \vec{B}\frac{I}{qtn}$$

where q is the charge of the charge carrier, t the thickness of the conductor and n the density of charge carriers. For metals this effect is very small, in the order of nV for Tesla size fields, but for semiconductors the voltage can be much higher, with a sensitivity of typically 1-10 V/T.

3.5 Light/radiation sensors

Light and radiation is usually detected through the release of free charge carriers, which induce reactions, current or charge. One of the most important factors when considering different kind of sensors is the dark current or the noise level at dark conditions which limits the use of many devices.

3.5.1 Photographic film

Photographic film is a sheet of plastic coated with an emulsion containing gelatin-bonded light-sensitive silver halide salts with variable crystal sizes. These determine the sensitivity, contrast and resolution of the film. When the emulsion is sufficiently exposed to light (or other forms of electromagnetic radiation such as X-rays), it forms a latent (invisible) image. Today photographic films are soon outdated but they can still be of good use when really good resolution is needed, or in certain harsh environments (although not radioactive).

In black-and-white photographic film there is usually one layer of silver salts. When the exposed grains are developed, the silver salts are converted to metallic silver, which block light and appear as the black part of the film negative.

Color film uses at least three layers. Dyes, which adsorb to the surface of the silver salts, make the crystals sensitive to different colors. Typically the blue-sensitive layer is on top, followed by the green and red layers. During development, the exposed silver salts are converted to metallic silver, just as with black and white film. But in a color film, the by-products of the development reaction simultaneously combine with chemicals known as color couplers that are included either in the film itself or in the developer solution to form colored dyes. Because the by-products are created in direct proportion to the amount of exposure and development, the dye clouds formed are also in proportion to the exposure and development. Following development, the silver is converted back to silver salts in the bleach step. It is removed from the film in the fix step. This leaves behind only the formed color dyes, which combine to make up the colored visible image

3.5.2 Photoconductive sensors

These are simply light-dependent resitors. Typically they are semiconductors where the incoming photons change the number of available charge carriers, which affect the resistivity of the sensor. As they are semiconductors they are as most sensitive to radioation which has energy above their band gap, whith a clear peak in the induced loss in resitivity due to radiation when the energy of the photons fit the band gap.

Matrial	Energy gap (eV)	Wavelength (Å)
PbSe	0.27	45982
PbTe	0.29	42811
PbS	0.37	33554
CdSe	1.73	7176
CdSS	2.42	5130

Table 3.3: Material constants for typical photoconductor materials.

The relationship between incident power and the resistance is typically nonlinear and approximatly follows this relationship:

$$log_{10}R = a - blog10P$$

where P is the incident power and R the resitivity.Typically they offer a dynamic range of four decades and they have quite slow response time (of 100 ms).

3.5.3 Photodiodes

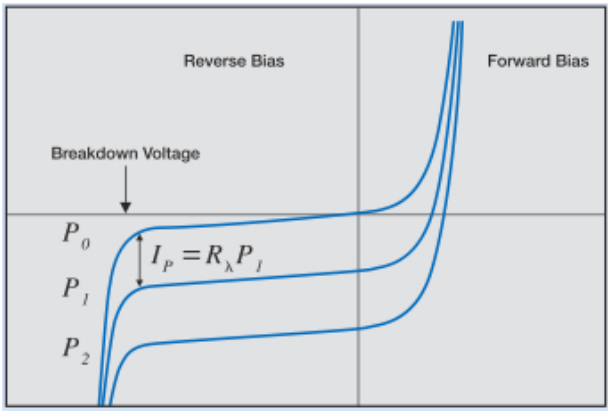


Figure 3.3: Photodiode operated in reverse-bias. Three different incedent power curves are plotted, and the offset in current due to the photoinduced current is cleerly shown

In a p-n junction the current is controlled by the charge balance in the depleted region between the differently doped region. Through excitations due to photons this charge balance is changed through the formation of electron-hole pairs that get separated at the junction. At zero bias a current will flow, at the same time as photvoltage will be created if the current is kept at zero. This is the basis of ordinary photvoltaic cells used as solar cells to harvest energy, but can also be used for fast light detection.

Photovoltaic mode

When used in zero bias or photovoltaic mode, the flow of photocurrent out of the device is restricted and a voltage builds up which is proportional to the impinging light. However, soon a charge balance is built up and this results in an unlinear realtionship between impinging power and voltage.

Photoconductive mode

In this mode the diode is often (but not always) reverse biased. This increases the width of the depletion layer, which decreases the junction's capacitance resulting in faster response times. The reverse bias induces only a small amount of current (known as saturation or back current) along its direction while the photocurrent remains virtually the same. Although this mode is faster (typical respondse time $1\mu s$), the photovoltaic mode tends to exhibit less electronic noise.

Avalanche mode

Avalanche photodiodes have a similar structure to regular photodiodes, but they are operated with much higher reverse bias. This allows each photo-generated carrier to be multiplied by avalanche breakdown, resulting in internal gain within the photodiode, which increases the effective responsivity of the device.

Phototransitors

Phototransistors also consist of a photodiode with internal gain. A phototransistor is in essence nothing more than a bipolar transistor that is encased in a transparent case so that light can reach the base-collector junction. The electrons that are generated by photons in the base-collector junction are injected into the base, and this current is amplified by the transistor operation. Note that although phototransistors have a higher responsivity for light they are unable to detect low levels of light any better than photodiodes. Phototransistors also have slower response times.

3.5.4 Geiger-Müller tubes

A Geiger-Müller tube consists of a tube filled with an inert gas such as helium, neon or argon between which there is a voltage of several hundred volts, but no current flowing. The walls of the tube are either metal or the inside is coated with metal or graphite to form the cathode while the anode is a wire passing up the center of the tube.

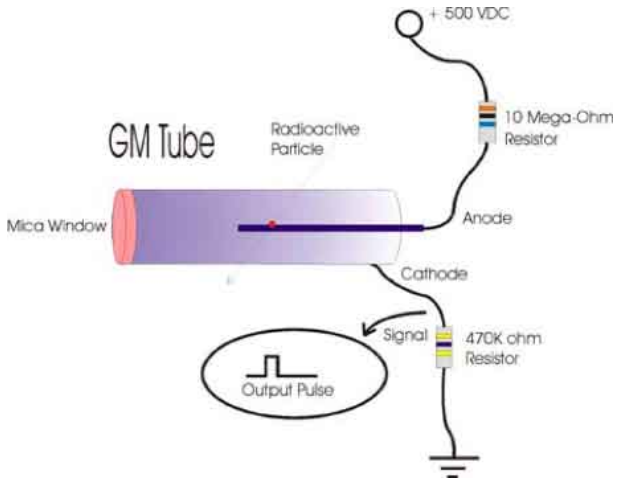


Figure 3.4: Schematic figure of the use of a GM-tube.

The usual form of tube is an end-window tube (fig. 3.4). This type is so-named because the tube has a window at one end through which ionizing radiation can easily penetrate. The other end normally has the electrical connectors. There are two types of end-window tubes: the glass-mantle type and the mica window type. The glass window type will not detect alpha radiation since it is unable to penetrate the glass, but is usually cheaper and will only detect beta radiation and X-rays. The mica window type will allow for detection of alpha radiation but is more fragile.

When ionizing radiation passes through in to the tube, some of the gas molecules are ionized, creating positively charged ions, and electrons. The strong electric field created by the tube's electrodes accelerates the ions towards the cathode and the electrons towards the anode. The ion pairs gain sufficient energy to ionize further gas molecules through collisions on the way, creating an avalanche of charged particles. This results in a short, intense pulse of current which passes (or cascades) from the negative electrode to the positive electrode and is measured or counted.

Most tubes will detect gamma radiation, and usually beta radiation above about 2.5 MeV. Geiger-Müller tubes will not normally detect neutrons since these do not ionise the gas. However, neutron-sensitive tubes can be produced which either have the inside of the tube coated with boron or contain boron trifluoride or helium-3 gas. The neutrons interact with the boron nuclei, producing alpha particles or

with the helium-3 nuclei producing hydrogen and tritium ions and electrons. These charged particles then trigger the normal avalanche process.

To prevent the current from flowing continuously there are several techniques to stop, or quench the discharge. Quenching is important because a single particle entering the tube is counted by a single discharge, and so it will be unable to detect another particle until the discharge has been stopped, and because the tube is damaged by prolonged discharges.

3.5.5 Photon multiplier tubes / channeltrons

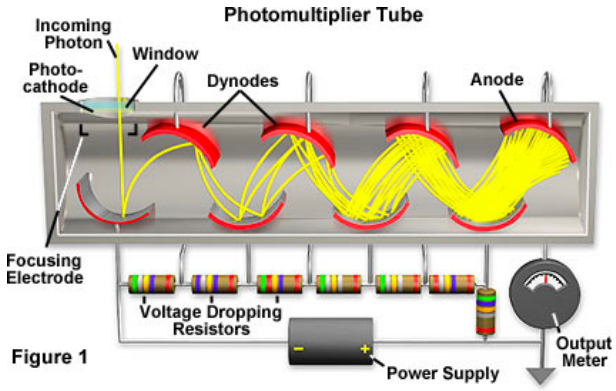


Figure 3.5: The use of a typical PM-tube.

Photomultiplier tubes are extremely sensitive detectors of light in the ultraviolet, visible and near infrared regime. PM tubes multiply the signal produced by incident light by as much as 10^8. The combination of high gain, low noise, high frequency response and large area of collection have meant that these devices find applications in nuclear and particle physics, astronomy and medical imaging.

Photomultipliers are constructed from a glass vacuum tube which houses a photocathode, several dynodes, and an anode. Incident photons strike the photocathode material which is present as a thin deposit on the entry window of the device, with electrons being produced as a consequence of the photoelectric effect. These electrons are directed by the focusing electrode towards the electron multiplier, where electrons are multiplied by the process of secondary emission.

The electron multiplier consists of a number of electrodes, called dynodes. Each dynode is held at a more positive voltage than the previous one. The electrons leave the photocathode, having the energy of the incoming photon (minus the work function of the photocathode). As they move towards the first dynode they are accelerated by the electric field and arrive with much greater energy. On striking the first dynode, more low energy electrons are emitted and these, in turn, are accelerated toward the second dynode. The geometry of the dynode chain is such that a cascade occurs with an ever-increasing number of electrons being produced at each stage. Finally the anode is reached where the accumulation of charge results in a sharp current pulse indicating the arrival of a photon at the photocathode.

For PM-tubes, channeltrons and microchannel plates there is an voltage of operation usually in the order of 1-4 kV, where dark current is very low but very high amplification is present. Usually there is a plateau in operating voltage where each burst has the same current, this is the safe operation voltage range. Using higher voltages will induce a dark current and may destroy the tube, using lower voltages is suboptimal. This range is sensitive to aging, and adjustment during the lifetime of the detector is needed.

Channeltron

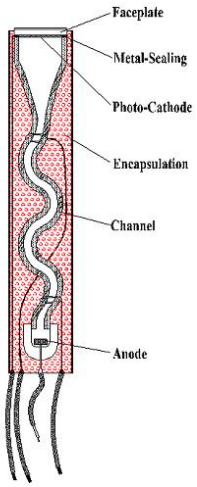


Figure 3.6: The appearance of a typical channeltron.

The channeltron is very similar but instead of dynodes there is a long tube with a large voltage applied over the ends. The tube is bent, and an impinging electron will create a cascade of new electrons which will hit the dynode surface again, multiplying the electrons until they are an easily detectable bunch. Channeltrons are typically used only to detect electrons in experiments where electrons are collected.

Microchannel plate

Sometimes it is desirable to detect single photons or charged particles over an image area. This can be done through micro-channel plate which are plates with small holes within them, where each hole serve as a channeltron. A voltage is applied to each of the sides of the plate and photons or charged particles entering the tube are enhanced in the same manner as in the channeltron.

3.5.6 Charge coupled device (CCD) and CMOS imaging chips

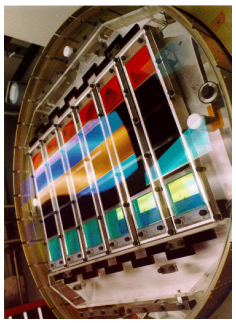


Figure 3.7: Array of 30 CCDs used on Sloan Digital Sky Survey telescope imaging camera.

Both CCD (charge-coupled device) and CMOS (complimentary metal-oxide semiconductor) image sensors work through a three step process. First they have

to convert light into electrons. This is done through excitation of an electron hole-pair in the semiconductor. The next step is to store the charge: this is done through defining p or n-doped regions on the chip, which provide a pn-junction that splits the electron-hole pair and store one part within a well defined region. This region is essentially the pixel (thermal excitations add to the dark current, which can be reduced by cooling the chip).

The last step is to read the value (accumulated charge) of each cell in the image. In a CCD device, the charge is actually transported across the chip and read at one corner of the array. This is done by electric gates on top of the surface, which deforms the potential at the surface to transport the electrons from pixel to pixel. Accordingly the readout is serial. An analog-to-digital converter turns each pixel's value into a digital value. In most CMOS devices, there are several transistors at each pixel that amplify and move the charge using more traditional wires. The CMOS approach is more flexible because each pixel can be read in parallel.

CCDs containing grids of pixels are used in digital cameras, optical scanners and video cameras as light-sensing devices. They commonly respond to 70% of the incident light (meaning a quantum efficiency of about 70%) making them far more efficient than photographic film, which captures only about 2% of the incident light. Most common type of CCDs are sensitive to infrared light, which allows infrared photography, night-vision devices, and zero lux (or near zero lux) video-recording/photography. Because of their sensitivity to infrared, CCDs used in astronomy are usually cooled to liquid nitrogen temperatures, because infrared black body radiation is emitted from room-temperature sources. For extra good efficiency in the UV region CCD:s are produced which have been polished from the backside. They are operated by projecting light onto the backside, where the read out electrodes are not disturbing the adsorption and excitation of charges.

CCDs use a special manufacturing process to create the ability to transport charge across the chip without distortion. This process leads to very high-quality sensors in terms of fidelity and light sensitivity. CMOS chips, on the other hand, use traditional manufacturing processes to create the chip the same processes used to make most microprocessors. Because of the manufacturing differences, there have been some noticeable differences between CCD and CMOS sensors.

- CCD sensors, as mentioned above, create high efficiency, high-quality, low-noise images. CMOS sensors, traditionally, are more susceptible to noise.
- Because each pixel on a CMOS sensor has several transistors located next to it, the light sensitivity of a CMOS chip tends to be lower. Many of the photons hitting the chip hit the transistors instead of the photodiode.
- CMOS traditionally consumes little power. Implementing a sensor in CMOS yields a low-power sensor.
- CCDs use a process that consumes lots of power. CCDs consume as much as 100 times more power than an equivalent CMOS sensor.
- CMOS chips can be fabricated on just about any standard silicon production line, so they tend to be extremely inexpensive compared to CCD sensors.
- CCD sensors have been mass produced for a longer period of time, so they are more mature. They tend to have higher quality and more pixels.

Based on these differences, you can see that CCDs tend to be used in applications that focus on high-quality images with lots of pixels and excellent light sensitivity. CMOS sensors traditionally have lower quality, lower resolution and lower sensitivity. CMOS sensors are just now improving to the point where they reach near parity with CCD devices in some applications.

3.6 Strain

Strain is measured using strain gauges. They often take the form of patches that are glued onto the surface of interest, once glued on they will yield a sensor signal proportional to the strain the patch is exposed to (which hopefully proportional to the strain of the surface where the patch is glued.

For resistive strain gauges the important sensitivity id the gauge factor:

$$G = \frac{dR/R}{dL/L}$$

where R is the resistivity and L the length in the strain sensitive direction. Suitable material for strain gauge are metal film gauges. Standard materials are 120 Ω and 350 Ω, and typically they can withstand strains of 5%, and the actual gauge value is approximately 2. It is possible with good set-ups to achieve resolutions in strain to approximately 1 parts per million.

There are also piezo-resistive gauges that have a much higher sensitivity (around 120), but they are more limited in the possible extension they can endure, and they are generally much more unlinear.

3.7 Pressure

Instruments used to measure pressure are called pressure gauges or vacuum gauges. A manometer is a pressure measuring instrument, usually limited to measuring pressures near to atmospheric. A vacuum gauge is used to measure the pressure in a vacuum — which is further divided into two subcategories: high and low vacuum (and sometimes ultra-high vacuum). By combining several different types of gauge, it is possible to measure system pressure continuously from 10 mbar down to 10^{-11} mbar.

Although pressure is an absolute quantity, everyday pressure measurements, such as for tire pressure, are usually made relative to ambient air pressure. In other cases measurements are made relative to a vacuum. When distinguishing between these zero references, the following terms are used:

Absolute pressure is zero referenced against a perfect vacuum, so it is equal to gauge pressure plus atmospheric pressure.

Gauge pressure is zero referenced against ambient air pressure, so it is equal to absolute pressure minus atmospheric pressure. Negative signs are usually omitted.

Differential pressure is the difference in pressure between two points.

There are a number of different pressure and vacuum gauges. Mechanical based pressure gauges, manometers are based on the mechanical displacement of a wall, which can be sensed either through changes in capacitance or resistance. Vacuum gauges rely on more sophisticated methods where changes in heat conductivity of the vacuum, or the ionization efficiency of gas atoms in the vacuum is measured.

3.8 Other sensors

There is an enormous amount of sensors for measuring different quantities, and many more coming in micro machined form. These are the topic of other courses.

3.9 Repetition questions

The important thing with this chapter is how to connect a sensor, and to know the basics of some of the more important sensors.

1. How are measurement bridges used?
2. Derive the general output of a quarter Wheatstone bridge.
3. Describe the Anderson loop.
4. Describe the different kinds of temperature sensors, how they are connected and what sensitivity they have.
5. Describe the different kinds of magnetic field sensors.
6. Describe the working manners of different kind of light and /radiation sensors.
7. Compare the different uses of different light/radiation sensors.
8. What types of pressure can you measure?

Chapter 4

Number representation

In this chapter we discuss some topics fundamental for understanding how numbers are represented in computers. It is meant to serve as a primer for further work. It is also important in case you should come across any low level programming, such as programming a micro-processor for communication.

4.1 Why digitization?

Digitization is to represent quantities in digits. This is a favorable way to represent quantities since it makes them easy to manipulate, compare and include in algorithms at the same time as it provides for noise immunity of the signal.

In experimental science we want to depict and represent the reality in order to extract new knowledge about our surroundings and the laws that govern them. For many sciences (like economics and social sciences) general descriptions can be enough. In natural sciences we often need to be much more accurate. Today, computers are used to push the limits of data collection. This is absolutely necessary when imaging materials in 3D, following fast transitions or when finding the few exciting events that can get you a Nobel Price at CERN. The only way we as humans can understand these large data sets is through a computer interface, which can reorder and filter the data into something with a physical meaning, like a curve or an image.

In order to utilize the number representation and their use in modern techniques fully, it useful to know how both humans and computers can represent reality through different kind of representations.

4.2 Number systems

The symbolic system used to represent a number greatly affects what kind of operations we can be preformed with these numbers. There are a multitude of ways to represent a number, the simplest being the one often used to tick of consumption of beverages: lines. For example 15 is written like:

One higher level of abstraction is to let groups of different numbers be denoted by different symbols, as used by the Romans and Greeks. They used different symbols to represent different numbers, and a given number was expressed by writing down the numbers in a sequence. Most of you are familiar with the roman system, where the position of the lower number indicates whether it will be subtracted or added to the higher number.

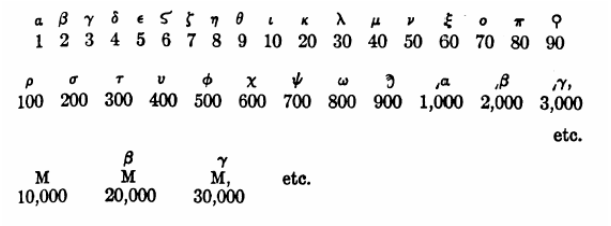


Figure 4.1: The main symbols used in the ancient Greek number system, the symbols could be put in any order, and were just added up to yield the full number.

However, representing increasingly larger numbers becomes a problem, as different symbols are required for each new level. In the Attic or Greek number system, indicated in Fig. 4.1 the symbols could be written down in any order and still represent the same number. They would write 2725 like:

Roman: `MMDCCXXV`
Greek: $\beta\psi\lambda\epsilon$

respectively. These two systems have two main disadvantages:

- Inability to cope with large numbers. For each large number a new symbol has to be found.
- No good algorithms for basic calculations, its just done by counting each symbol separately.

One of the more important inventions to the number systems was the 0. This represents both the pivoting point of negative and positive numbers, and no value. It is fundamental for representing numbers in a position number system. The zero and the number position system was introduced to the Europeans from Arabia, and is accordingly called the Arabic number system. The development of this number system (this time only ten symbols are used) is displayed in Figure 4.2.

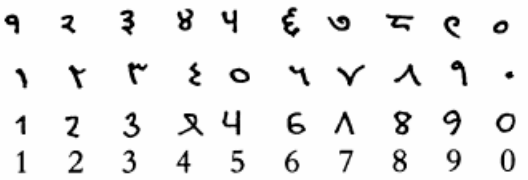


Figure 4.2: The evolution of the modern Arabic number system from Hindu to Arabic over medieval European to modern numbers.

4.2.1 The position number system

The modern number system utilizes one important detail: the position of the number is significant for the actual value it represents. This is simply given by the value times the *base* (r) to the power of the position of the value. Accordingly we interpret 2735:

$$2\cdot 10^3 + 7\cdot 10^2 + 3\cdot 10^1 + 5\cdot 10^0 \longrightarrow 2000 + 700 + 30 + 5$$

Decimal	Binary	Hexadecimal
0	0	0
1	1	1
2	10	2
3	11	3
4	100	4
5	101	5
6	110	6
7	111	7
8	1000	8
9	1001	9
10	1010	A
11	1011	B
12	1100	C
13	1101	D
14	1110	E
15	1111	F

Table 4.1: Conversion table for the most common number bases used in computers.

More generally, we write an n position number with digits d as a series:

$$N = \Sigma_{i=0}^{n-1} d_i r^i$$

It is customary to call the digit most to the left the most significant digit (MSD) and the one most to the right the least significant digit (LSD) of the number.

One of the most important features of the position based number system is that it easy to perform calculations with. They can be performed just by following simple rules, algorithms [1]

The *base* of a number system,(r), is often inherent to the system. For example, our modern number system is based on the number ten, since we have ten fingers to count with. Another common system is the hexadecimal system with the base 16. This is widely used because it is a good way for humans to represent 4 binary digits as an easily readable entity. A pair of two hexadecimal numbers represent the content of a byte. Two other systems of historic interest is the base 12 system and the base 60 system, which were used by the Babylonians, and still have a great influence on our time-base and within astronomy.

The inherit base for computers is 2. This is the simplest way to represent a number. It is also the only thing you need for doing Boolean algebra, which is how computers operate. A base two system can easily be represented by different physical states:

- High and low potentials (0 and 5V as in TTL signals).
- Absence or presence of a reflecting surface (CD-ROMS).
- Different magnetization directions on a magnetic surface (Hard discs).

[1] The name algorithm originates from astronomer Muhammad ibn Musa abu Abdallah al-Khwarizima al-Madjusi al-Qutrubelli, or Al-Korazmi. He wrote the book in the 9:th century which introduced the position system to European Scholars. The calculation were performed through a comparatively simple set of rules, and such sets has since then been called after him: algorithms.

- Presence of light or not (optical fibers).

4.2.2 Limit in representation, use of different bases

It is easy to predict the largest possible number and the number of different states that can be represented with n digits in a positional system. The largest number is

$$r^n - 1$$

and the number of states is:

$$r^n$$

The base used in number presentation is usually not expressed since it given by the situation. If expressed, it is often written after the number in subscript:

$$(1110)_2 = (16)_8 = (14)_{10} = (E)_{hex}$$

The easiest way to convert numbers between bases is to start with the least significant digit. The number that is to be converted is divided by this number, the remainder gives the value of the lowest significant digit while the quotient is used for the next operation. For the next digits the same procedure is repeated until only a remainder is left after the division. To convert $(211)_{10}$:

$$\begin{array}{rll} 211/2 = & 105 + 1/2 & d_0 = 1 \\ 105/2 = & 52 + 1/2 & d_1 = 1 \\ 52/2 = & 26 + 0 & d_2 = 0 \\ 26/2 = & 13 + 0 & d_3 = 0 \\ 13/2 = & 6 + 1/2 & d_4 = 1 \\ 6/2 = & 3 + 0 & d_5 = 0 \\ 3/2 = & 1 + 1/2 & d_6 = 1 \\ 1/2 = & 0 + 1/2 & d_7 = 1 \\ \Longrightarrow & & \\ & (211)_{10} = (11010011)_2 & . \end{array}$$

The representation of a number in any base can be found in a similar manner.

One can represent almost any number through a large enough number of *on* or off states (bits) and there is really no need to represent numbers in other ways. There have been attempts to build computers with other bases (all mechanical calculators used until the ∼1980 were 10-based) but today the binary system is dominating.

An upcoming field today is quantum computing, where data is stored in qubits. The information in qubits is not only a 1 or a 0, but any linear combination of both, and can thus represent more complex states at the same time. Consequently, the arithmetic becomes very complicated (often one talk of entanglement). Quantum computers have the potential to perform specific tasks with high very high speed compared to bit-based computers, but so far only a few qubits have been realized. It is usually estimated by optimistic researchers that it will take 15-25 years before this type of computers get operable. The basis of quantum computing is not a topic of this course.

4.2.3 Binary numbers: bits and bytes.

There are two often used terms: bits which stands for binary digit and bytes which is 8 bits[2]. The 8 bit byte can represent 256 symbols which is enough to obtain

[2]According to Wikipedia, the term byte was introduced in 1956 by Werner Buchholz when working in the IBM stretch computer, however at that time it only contained 6 bits,

functional coding of information and programs in most western languages. These two units are, as you probably are aware of, the standard unit for determining size for information transfer. As already mentioned, a good way for humans to represent the content of a byte is by two hexadecimal digits.

4.3 Binary arithmetic.

Now some basic algorithms for performing binary maths.

Addition

Adding in a position number system is very easily done. Start at the least significant number position (usually to the right) just add the numbers of the same position together, if they add up to more than the base, divide the number by the base and add the integer part to the next higher digit.
For the binary system we need the basic adding rules:

$$0 + 0 = 0$$

$$1 + 0 = 0 + 1 = 1$$

$$1 + 1 = 10$$

Then we are ready to work an example lets try $27 + 13 = 40$

$$\begin{array}{rl} & 11110 \quad \longleftarrow \textit{carry} \\ & 11011 \\ + & 01101 \\ \hline = & 101000 \end{array}$$

Observe that while doing this we have actually finished up with a number that is longer than the ones we originally had. If we had not allocated space in our storage (on the paper), we would have had overflow, and the answer would have been wrong.

4.3.1 Subtraction: Ordinary method

Again we start with at the least significant number position just subtract the numbers of the same position, if needed an extra value is added from the higher digit (borrowed). Subtraction rules:

$$0 - 0 = 1 - 1 = 0$$

$$1 - 0 = 10 - 01 = 1$$

Then we are ready to work an example, lets try $6 - 3 = 3$

$$\begin{array}{rl} & {\scriptstyle 10\ 10} \quad \longleftarrow \textit{borrow} \\ & 110 \\ - & 011 \\ \hline = & 011 \end{array}$$

the common size for in/out buses at that time

4.3.2 Subtraction: negative numbers.

There are many ways to represent negative numbers in a number system, One way to do this is to use one bit to signify the sign of the number. Often the most significant bit is used, and the method is called *signed magnitude*. A more complex method, but simplifying for low level calculations, is to represent negative numbers through the 2's complement. In this coding there is still a bit for the sign, but the rest of the number is represented in such a manner that simple addition of the two numbers will yield a subtraction. Thus only a one type of mathematical operation has to be performed. The general base complement is defined as

$$N^{comp} = r^n - N$$

If we take the base complement again we end up with:

$$(N^{comp})^{comp} = r^n - (r^n - N) = N.$$

Note that r^n is one digit longer than the n digit number. The complement has the fundamental property that addition of the number will effectively subtract the number:

$$N + N^{comp} = N + (r^n - N) = r^n$$

For 1011 in the two-base, represented with an extra bit (the most significant bit, MSB), we get

$$+1011 \longrightarrow 01011.$$

The two complement number will be:

$$100000 - 01011 \longrightarrow 10101.$$

Note that we have to find the 2-complement to the whole positive number 01011. Another way to obtain the same result is to take the one-complement (with an extra zero for the sign bit), that is change every 1 to a 0 and every 0 to a 1 and then add 1 at the end. Accordingly:

$$-1011 \longrightarrow 10100 + 1 \longrightarrow 10101$$

another example can be -011 (which will be used below)

$$-011 \longrightarrow -0011 \longrightarrow 1100 + 1 \longrightarrow 1101$$

4.3.3 Subtraction: Two complement method

Through base complements all possible negative operations can be performed through additions:

Subtraction	Addition
$A - B$	$A + (-B)$
$(-A) - B$	$(-A) + (-B)$
$A - (-B)$	$A + (-(-B))$
$(-A) - (-B)$	$(-A) + (-(-B))$

This makes implementing operations in a processor significantly easier. Lets try the last example again:

$$110 - 011 \longrightarrow 110 + (1100 + 1) \longrightarrow 110 + 1101 \longrightarrow (1)0011$$

Linux	MacOS	Windows
LF	CR	CRLF

Table 4.2: Symbols used for end of lines in text files for different operating systems

4.4 ASCII Coding

There is often a need to represent something more than just values with a binary number. Binary coding is the most efficient for storing measurement data, but often other kinds of information need to be stored. ASCII code is probably the most important coding. It is general for numbers and alphanumeric characters and can be summarized as:

- The name stands for American Standard Code for Information Interchange (ASCII)
- The extended version use 1 byte (8 bits) and can represent 256 symbols.
- Simple tables are found to encode and code. Here, each character corresponds to a number.
- The most straightforward way to encode the number is through a two digit hexadecimal code, each digit codes 4 bits, together making 8 bits.
- The first 32 characters in the ACSCII-code ($000\text{-}031_{10}$, $00\text{-}1F_{hex}$) forms a special set of non-printing characters called control characters. They perform various printer/display operations rather than displaying symbols. Unfortunately there is very little standardization for these kind of operations.

4.4.1 ASCII files of different operating systems

One example of poor standardization is how different operating systems store text files. It turns out that the different systems use different ways to symbolize a new line in a text file. This is usually not a problem when migrating files from multi-platform programs like Acrobat, but it can cause problems when transforming data files.

The carriage return is often referred to by the capital letters CR. On a Macintosh, every line has a CR at the end. Under Linux (a variant of Unix), the end of a line is indicated by a line feed (LF). Every line ends with a line feed. Calling the end of a line an LF versus a CR is not just semantics. A CR is a 13 in the ASCII table of characters and an LF is a 10. Contributing to the confusion is that fact that Microsoft Windows does things yet another way. Under Microsoft Windows, lines end with a combination of 2 characters – a CR followed by an LF.

Fortunately most file transfer software takes care of this conversion for you.

4.5 Unicode

Unicode is an industry standard whose goal is to provide the means by which text of all forms and languages can be encoded for use by computers.

The establishment of Unicode involved an ambitious project to replace existing character encoding schemes, many of which are very limited in size and incompatible with multilingual environments. Unicode has become the largest and most complete character encoding scheme, serving as the dominant such method in the internationalization and localization of computer software. The standard has been

implemented in many recent technologies, including XML, the Java programming language and modern operating systems. Unicode reserves 1114112 code points [3], and currently assigns characters to more than 96,000 of those code points. The first 256 codes precisely match those of ISO 8859-1 (ASCII versions). The majority of the 96,000 encodings are at this time, used for Chinese and Korean characters.

4.6 Real numbers in computers.

So far we have only considered integer numbers. The principles are the same for real numbers (rational and irrational numbers). To represent fractions of a number, positions are added after the least significant digit of the integer. This position is marked by the decimal point (the *radix*). Otherwise all principles of the positional system is followed: for each position the significance decreases and we express an n digit number with m decimals as:

$$N = \Sigma_{i=0}^{n-1} d_i r^i . \Sigma_{i=1}^{m} d_{-i} r^{-i}$$

There are two main strategies to represent real numbers: fixed point or floating point. The fixed point saves you the trouble of keeping track of where your radix is located. It is defined by the number. However, there is no versatility for expressing large numbers. Accordingly floating numbers are dominating in real number representation in computers. With the floating point representation very large numbers can be expressed. As an example, 123.456 becomes:

$$1.23456 \cdot 10^2$$

as floating point. In the hexadecimal system, 123.abc becomes:

$$1.23abc \cdot 16^2$$

The most common standard used for computers is the IEEE standard 754 for floating point numbers. It is a base two system, where each number is represented by three parts: the sign, the fraction and the exponent. Since binary number expressed as a floating point number will start with a 1, this is not expressed. Neither is the base of the exponent. Thus, only the following parts are stored (single precision):

- 1 bit for the sign (bit 31).
- 8 bits for the exponent (bits 30-23).
- 23 bits for the fraction (bits 22-00).

The exponent is expressed with a bias of 127_{10} $(01111111)_2$. To find the real exponent this bias is subtracted from the original number. The range for a single precision number is: $10^{-44.5}$-$10^{38.53}$ and the precision is 2^{23} around 7 digits in the decimal system. For double precision we have the following:

- 1 bit for the sign (bit 63).
- 11 bits for the exponent (bits 62-52).
- 52 bits for the fraction (bits 51-00).

Here the bias of the exponent is 1023, while the precision is around 19 decimal digits. There are also three special combinations: +infinity, -infinity and NaN (not a number) to be used when the representation can not be used to store the number.

[3] A code point represnts a single sign, accordingly Unicode can represent 1114112 signs.

4.7 Error detecting representation

It is very important for the receiver to know that the information received is the correct one. One brute method is to send every message twice. This is called oversampling in the CD and DVD player business. There are less demanding methods involving bit encoding:

- **Parity Check** By adding an extra bit to the transmission it can be made even odd or even. Doing this consequently, (either with odd or even parity) single errors can be found in binary transmissions.
- **Check sum** By adding all data words in a transmission and then transmitting the last bit(s)/Byte(s), a more thorough check can be performed. The result is sent to the receiver who perform the same check.
- **CRC** Cyclic redundancy check. Each data word is divided by a polynomial. The remainder is included in the next division and through this operation a running remainder is created. This is transmitted as the last word and compared with a remainder calculated in the same manner by the recipient. Division of these two should give a zero remainder, otherwise a transmission error has occurred. This method is easy to implement in hardware and is often used in common applications like data storage on disk.

4.8 Repetition questions

You should know what forms numbers can be represented in and how to convert between them. Essential points to be able to understand/perform are:

1. Convert (positive, negative and real) numbers from decimal to binary, and from binary to decimal.
2. Convert hexadecimal numbers to decimal and binary, and vice verse.
3. Do arithmetic with binary numbers (addition and subtraction).
4. Do subtraction using twos complement for negative numbers (using the correct amount of bits).
5. Explain why twos complement is a good way to represent negative numbers.
6. Calculate the highest (positive) number that can be represented using n bits.
7. Explain how the 8421-BCD code is used to code/represent decimal numbers.
8. Explain the advantage of the Gray code (over the 8421-BCD code).
9. Convert a 8-bit binary number to a ASCII-character.
10. Convert letters and (decimal) numbers to ASCII-code.
11. Explain why we need the ASCII-code.
12. Convert numbers represented using the IEEE Standard 754 (single and double precision) to decimal numbers.

Chapter 5

Digital conversion

This chapter will give an overview of different aspects of sampling and problems associated with sampling a signal. Digitisation of signal involves conversion from continuous to discrete signal as well as quantization of the amplitude. The main topics of the chapter is how to achieve this and how to avoid possible artifacts due to sampling: Aliasing, practical upper frequency limits and dithering are important examples.

5.1 Initial considerations

Analog to digital conversion is very dependent on the hardware for speed and resolution. Typical important characteristics to look for is the:

- Resolution (smallest measurable step)
- Dynamic Range (highest measurable signal/resolution)
- Accuracy (both amplitude and timing)
- Speed
- Costs

This has to be compared with the characteristics of the signal to be sampled and the actual information required. Sampling is performed through an instrument such as digital oscilloscopes, multimeters or general data acquisition cards (more about that in later chapters). The basic component in these instruments are the converters - Analog to Digital Converters (ADC) and Digital to Analog Converters (DAC).

5.2 Time-varying signals

In principle the faster the more expensive the measurement equipment and the faster the development of measurement techniques. Ordinary oscilloscopes can cope with time varying signals of bandwidth (or highest frequency) of typically 50 Mhz. Multimeters are usually limited to Hz measurements, while high speed digitisers go well up in radio frequency range (maximum sampling rate is today somewhere at 50-100 GHz. To probe phenomena at a smaller timescale optical and frequency mixing techniques must be used. Today atto-second pulses puts the absolute limit for time-resolved studies almost 8 orders of magnitude over the limit for electrical measurements. With no doubt you will experience how his limit will be expanded even further in the future.

As speed is very cost-driving it is necessary to limit your selection of instrument to as low frequency as you can. Typically you have to consider:

- The characteristics of the sampled signal (Amplitude, frequency).
- The information you want from the signal.

This decide what kind of amplitude, resolution (dynamic range) and the sampling rate you actually need.

5.3 Sampling

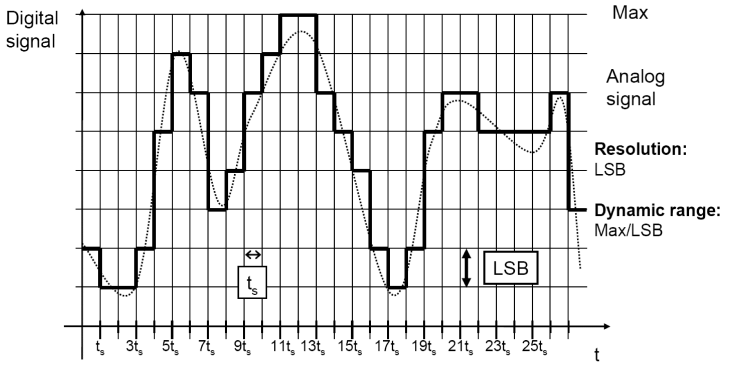


Figure 5.1: The way a analogue signal is discretised when performing sampling.

Sampling is performed in minimum two dimensions (as a minimum): the signal amplitude dimension and the time dimension. The signal amplitude is limited by the the maximum resolution as well as the maximum amplitude, while the time domain is limited by the maximum sampling rate.

5.3.1 Amplitude Signal

For the analogue to digital converter there are the some errors introduced by the digitization. Basically the signal is represented by a number, limited by the number of bits n used to represent the number. The number of levels are 2^n. For a n-bit converter, the maximum resolution (the value of the least significant bit) is given as:

$$Q = (U_{max} - U_{min})/(2^n - 1)$$

.

It can be instructive to study the number of levels and the resolution for a given number of bits:

Note that the resolution for the high-end by far exceeds the resolution of most analogue circuits (roughly 65 dB for a standard design). Converting data early in the signal chain has the advantage of keeping high dynamics for the signal, which also is true for storing of data.

5.3.2 Amplitude errors of sampling

Errors are either the inherit errors due to digitization and the errors due to the non-ideal characteristics of components. For all conversion the inherit maximum error

Bits	Decimal digits	levels	parts	parts	dB
6		0-63	1:64	1.6%	36dB
	2	0-99	1:100	1%	40dB
8		0-255	1:256	0.4%	48dB
	3	0-999	1:1000	.1%	60dB
12		0-4095	1:4096	0.02%	72dB
16		0-65535	1:65536	15ppm	96dB
	5	0-99999	1:100000	10ppm	100dB
18		0-262143	1:262144	1.6ppm	108dB
24		0-16777215	1:16777216	.00006ppm	144dB

Table 5.1: Resolution for a given number of bits.

is limited to the half value of the least significant bit. This is the error that is due to the deviation that we have from the real signal due to the quantised levels. The root mean square of this quantisation error, E_Q is given by a simple calculation:

$$E_q = \sqrt{1/T \int u^2 dt} \approx 0.29Q$$

where Q is the resolution with the peak distortion of $Q/2$.

If the signal is critically time dependent, another amplitude error is given by the time delay of the signal, where the time offset will yield errors proportional to the time derivative of the signal. Of course this is only a problem if you are actually comparing signals in time. This is the case for most modern I/O boards which multiplex all converter signals to the same converter. Ways to solve this problem is to use a high sampling rate, or AD converters with sample and hold circuits which freezes the signal from all signals at the same time for conversion later.

In addition to this there are generally four types of errors associated with the converters:

Offset There is an bias added to the real signal.

Slope error/amplification error The conversion is made with an amplification error, this error linear with the amplitude of the signal.

Linearity error The conversion is slightly nonlinear.

Differential linearity error Error in amplitude step for one bit.

5.3.3 Speed and accuracy

To sample a signal accurately, that is within 1/2 of the least significant bit, the time-varying signal must not change more than during the conversion time. To estimate the limit this sets, we can consider a sinusoidal input signal and how much the maximum change of that one will be, that sets a limit for how long time we can sample each bit (or we will find an average).

The maximum rate of change occurs at the zero crossover. For a $A\sin(2\pi ft)$ signal we get a maximum derivative at $t = 0$:

$$\frac{dV}{dt} = A2\pi f$$

Furthermore, if the conversions speed or aperture time is τ_{conv}, and we want the voltage to rise maximally $V_{LSB}/4$ and we get an upper limit of the derivative:

$$\frac{dV}{dt} = \frac{V_{LSB}}{4\tau_{conv}}.$$

We can accordingly obtain an expression for the maximum frequency sinusoidal that can be measured with acceptable accuracy f_{max}:

$$f_{max} = \frac{V_{LSB}}{A2\pi 4\tau_{conv}}$$

.

One example is a converter working at 10 volt peak to peak input range, 12bit ADC converting at $10\mu s$. For a maximum amplitude signal (5V) we obtain:

$$f_{max} = \frac{10}{((2^{12})8\pi 10^{-6}5} = \frac{10}{4096 \cdot 40\pi 10^{-6}} = 2Hz$$

We can now apply a sample and hold circuit, which will shorten the sampling time (or aperture time). The sampling time is shortened through an analogue circuit that briefly connects to the signal and stores the signal for the much slower sampling sequence. This can increase the accuracy dramatically.

It is possible to calculate the aperture needed to get the ADC above to work at a 10 kHz:

$$\tau_{conv} = \frac{V_{LSB}}{A2\pi 4 f_{max}} = \frac{10}{((2^{12})8\pi 10^4 5} = \frac{10}{4096 \cdot 40\pi 10^4} = 2 \cdot 10^{-9}.$$

5.3.4 Dithering

One would think that it would be possible to overcome quantisation problems through sampling many times and then perform averaging. For a stable signal that always yield exactly the same digital number this will not yield any additional information. To succeed the average error of the input signal must be significantly larger than resolution of the A/D converter. One way to overcome this problem is to perform *dithering*. Dithering is performed through adding Gaussian noise on the input and subsequently removing it through averaging. It is then possible to achieve quite high resolution with a low resolution A/D converter.

5.3.5 Offset nulling

One often encountered problem is that you want to detect small variations in a rather large signal. Typically this means that you will use a unwanted part of your digitisation range just to describe the constant part of the signal. A common way to circumvent this problem is to subtract the constant level from your signal before digitisation. It is then possible to further amplify the signal to optimise the resolution of the measurement.

5.4 Time domain

There are two ways to consider periodic time signals in the time domain or in the frequency domain. Understanding sampling is to understanding these two domains and their interconnection. This is mainly transform theory which I can not recapitulate here, but I can remind you of the main results as described in fig. 5.2.

5.5 Initial example - time response

DC signals are seemingly simple to measure. There are a number of questions regarding how you properly measure DC signals. However, when you connect to your circuit it can be considered to consist of a supply voltage connected to a switch through an equivalent resistor, R and the stray or load capacitance C. Together

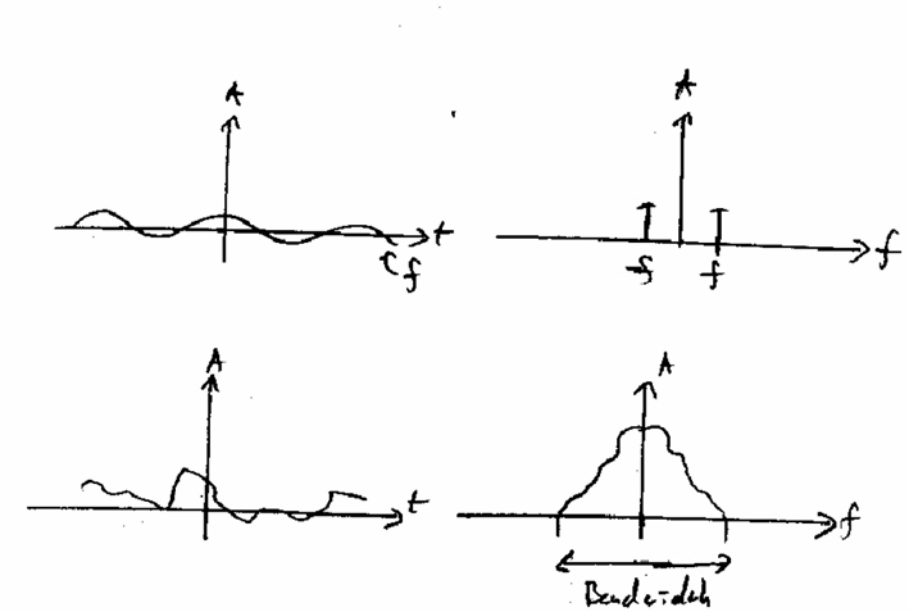


Figure 5.2: Upper: the time and frequency domain of a sinusoidal wave. This gives rise two two peaks at the frequency of the wave in the frequency domain. Lower: For a more complex wave it can be characterised by the highest frequency wave needed to describe it, the frequency where this occurs is the bandwidth of the signal.

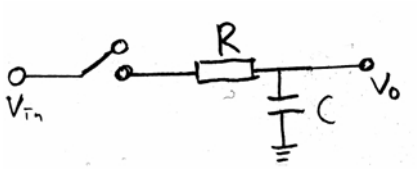


Figure 5.3: Schematic of the connection of a signal V_{in} to the conversion device.

these form a simple low pass filter which actually will make it impossible to measure the voltage instantaneously (fig. 5.3).

If the initial level is zero we can write the voltage after the switch closure (connection to DC-signal):

$$V_0 = V_{in}[1 - e^{\frac{-t}{RC}}]$$

For the output to reach 90% of the input we need:

$$\frac{V_0}{V_{in}} = 0.9 = 1 - e^{\frac{-t}{RC}}.$$

This yields the minimum waiting time:

$$t = -RCln(0.1) = 2.3RC.$$

To reach a decade better signal we have to wait another $2.3RC$ and so fourth. This is not surprising since currents must flow from the sampled signal. This is critical for all measurements: minimise the charges (currents) that has to flow to obtain good readings.

5.5.1 Conversion speed

Conversion speed and dynamic range are the two most important factors, in general the better the more expensive. The speed limit is usually characterized by an *equivalent resistance* as well as the *stray and load capacitance* of the circuit (as in the DC circuit). Again we get the minimal conversion time of $2.3RC$ per decade. However, as the signal now is sampled repetitively it is very important to balance this minimal acquisition time to the minimal sampling time.

5.5.2 Aliasing: under-sampling, folding

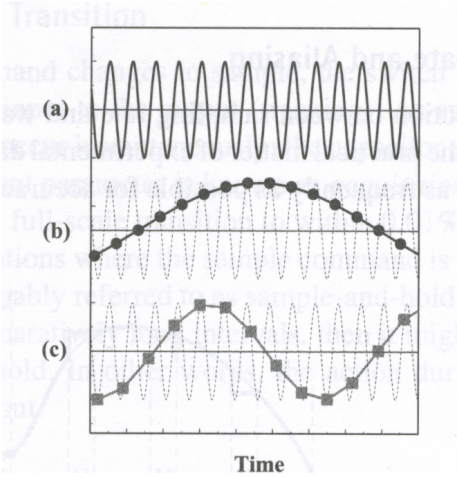


Figure 5.4: Why is it called Aliasing? If a true signal (a) is sampled at a certain sampling frequency, many different waveforms cane be fitted to the same curve, providing many aliases for the same sampling points. There is no longer a 1:1 relationship between the signal and the possible candidates for signals.

One of the profound results of the sampling is the possibility of aliasing. If the signal is sampled with to few samples, or to low frequency, we end up with false signals (fig. 5.5, that is low frequency components are mixed with higher frequency components through folding. By not sampling at fast enough rates we can not distinguish between signals at multiples of the sampling frequency. Thus higher frequency components do not only add noise to measurements, but signals at false frequencies.

The exact frequency when this occurs is the folding frequency (half the sampling frequency). When sampling above this frequency the observed frequency will be offset by f_s. Accordingly, the frequency will decrease until f_s has been reached when the frequency will be zero (fig. 5.6). If the frequency again increases a new cycle will start. The actual behaviour of the signal is as of the spectral axis was folded back and fourth between the origin and the folding frequency, giving rise to the term folding.

This is of course only true for periodic signals, non periodic signals (noise) will give rise to noise signals. However, the sampling theorem also indicate that any signal, which can be described quite well through a Fourier or Laplace series, will only be described by components up to half the sampling frequency thereby offering another way to compute maximum response times (which will not be dealt with here).

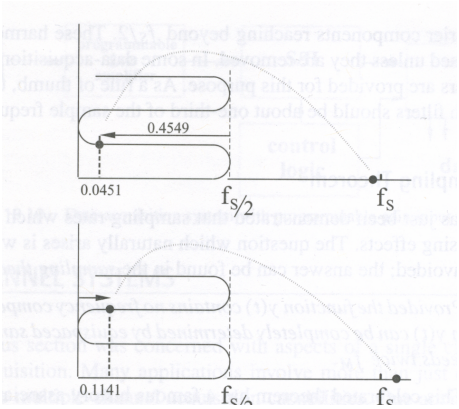


Figure 5.5: One way to consider the consequences of undersampling is through folding, the sampled frequencies can actually

5.6 Sampling considerations

In order to describe a periodic signal properly through sampling it is vital to sample with a high enough frequency. The question usually arises, what is a high enough frequency? For evaluating frequencies in a signal it can be just a factor two higher than the frequency of the signal, while for high resolution sampling it can be a factor ten or higher. The ultimate limit has been investigated and formulated through the sampling theorem.

Consider a sinusoidal signal $x(t)$ of frequency f_0:

$$x(t) = \sin(2\pi f_0 t)$$

this is sampled at even intervals with the sampling frequency f_s at the time interval τ_s. We can write down the value of the n:th sample:

$$x_n = \sin(2\pi f_0 n\tau_s).$$

The sinus or any periodic function repeats itself every 2π, accordingly we can insert an arbitrary integer constant $2\pi m$ into the signal *without changing it*:

$$x_n = \sin(2\pi f_0 n\tau_s) = \sin(2\pi f_0 n\tau_s + 2\pi m) = \sin(2\pi n(f_0 + \frac{m}{n\tau_s})\tau_s).$$

Now it is possible to replace the constant m with a new integer constant, $k = \frac{m}{n}$.

$$x_n = \sin(2\pi f_0 n\tau_s) = \sin(2\pi n(f_0 + \frac{m}{n\tau_s})\tau_s) = \sin(2\pi n(f_0 + \frac{k}{\tau_s})\tau_s).$$

Finally we can again identify that

$$\frac{1}{\tau_s} = f_s.$$

Thus we can write down the equivalent expression for the n:th sample:

$$x_n = \sin(2\pi n(f_0 + kf_s)\tau_s)$$

Thus the sampled signal can represent the base frequency, *but also all signals offset by multiples of the sampling frequency.*

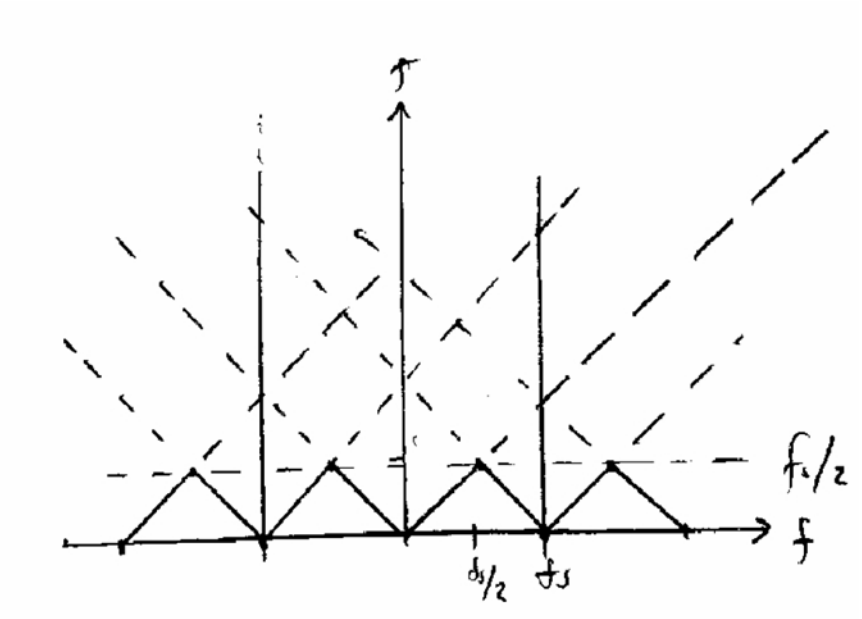


Figure 5.6: A plot indicating the observed frequency for a sinusoidal wave when sampled at f_s. The straight lines are replicated for each multiple of the sampling frequency. Thus in the observable window (up to $f_s/2$ on the y-axis) the frequency will go up and down according to the aliasing frequency of the under sampled frequencies.

5.7 Considerations: transform approach

Here I indicate the more important points of the basic mathematics using transform theory(prerequisite: a large portion of transform theory).

Let us first consider a signal,$u(t)$, that is sampled at a frequency f_s. We can write the signal as a train of delta pulses with the internal delay $T_s = 1/f_s$ multiplied with the amplitude signal, $u(t)$:

$$u_s(t) = \sum_{n-\infty}^{\infty} \delta(t - nT_s)u(t).$$

Consider the properties of the pulse train or the Dirac comb function:

$$\sum_{n-\infty}^{\infty} \delta(t - nT_s),$$

when transformed into the frequency domain it forms another pulse train:

$$\frac{1}{T}\sum_{k-\infty}^{\infty} \delta(f - \frac{k}{T_s}) = \frac{1}{T}\sum_{k-\infty}^{\infty} \delta(f - kf_s).$$

The complete signal now transforms into a convolution which is easily solved since it contains the Dirac function.

$$U_s(f) = U_s(f) * \frac{1}{T}\sum_{k-\infty}^{\infty} \delta(f - kf_s) = \frac{1}{T}\sum_{k-\infty}^{\infty} U_s(f - kf_s)$$

Instead of our signal centered around zero frequency we obtain a train of signals separated by the sampling frequency through the whole of frequency space.

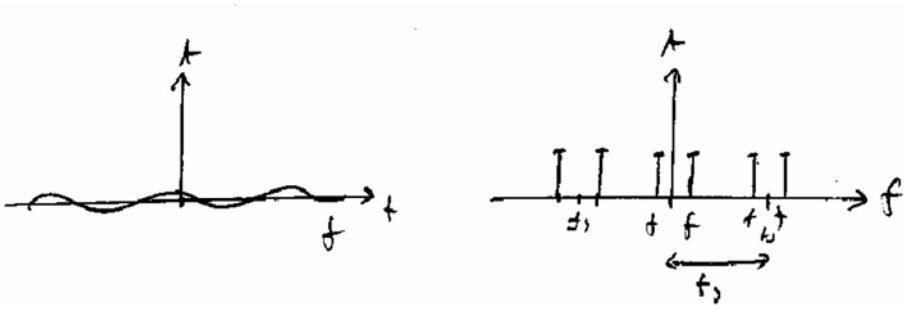


Figure 5.7: When a signal is sampled there arises an ambiguity whether the signal is the sampled signal or a signal of that frequency plus a multiple of the sampling frequency. This effect is real and is due to loss in information from the continuous signal.

5.7.1 The sampling theorem

We now have the background to tell when we sample the signal without mixing of other signals. As the sampled signal will reappear at the sampling frequency, we have to confine our signal to frequencies below half the sampling frequencies, otherwise components from the higher frequencies will perturb our signal.

- The sampling theorem states that for a limited bandwidth (band-limited) signal with maximum frequency f_{max}, the equally spaced sampling frequency f_s must be greater than twice of the maximum frequency f_{max}

$$f_s > 2 \cdot f_{max}$$

  in order to have the signal be uniquely reconstructed without aliasing.

- The frequency $2 \cdot f_{max}$ is called the Nyquist sampling rate. Half of this value, f_{max}, is sometimes called the Nyquist frequency or the folding frequency.

The effects of the sampling theorem, the repetition of the signals at higher frequencies is a very real effect of sampling. The high frequency structures will continue to infinity. Ambiguities regarding this can only be solved through limiting the known sampling region to below the folding frequency.

5.7.2 How to avoid aliasing

Aliasing is avoided through two different measures:

- High sampling frequencies. By sampling faster than the highest component of the signal you are not disturbed by the frequency ambiguity in sampling.
- Low pass filters. By filtering away all signals above $F_s/2$ you remove any folding frequencies, this is called anti-aliasing filters.

5.8 Digital Input and Output

In modern measurement technique the A/D and the D/A interface is one of the most crucial ones. Understanding the physics of the different conversion principles is to understand what choices that can be made.

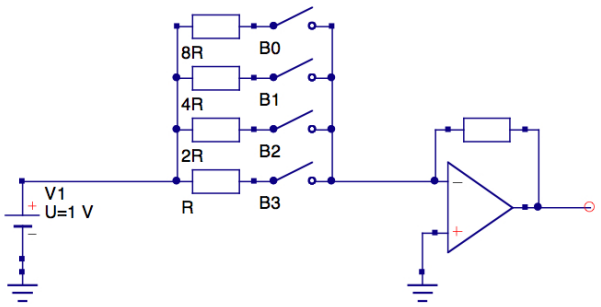


Figure 5.8: The schematic layout of a binary weighted DA converter. As each of the switches B0, B1, B2, B3 are closed currents proportional to their bit values will flow, these are converted through a current to voltage amplifier to a voltage.

5.8.1 Digital to Analogue Conversion (D/A, DAC)

There are three used techniques for digital to analogue conversion. All are dependent on being able to send a controlled current, by opening or closing switches to reference sources of different kinds.

- binary weighted resistor
- R-2R-ladder

Binary weighted resistor

The easiest to understand, but also one of the harder to realise is the binary weighted resistor DA-converter. In this all switches that are controlled by the input signal are connected to the same voltage, these are connected through resistances to a current-to-voltage converter (Fig. 5.8). By choosing different resistances the same potential will send different currents, and to comply with the weight of single digits of the number the most significant bit set a switch connected to a R resistance. The second most significant bit should send half the current, accordingly the resistance must be double that of the most significant bit. For the full conversion, n resistors are needed, and the values range from R to $2^n R$. The converter is made complete with a current to voltage converter for voltage signals, or using a current buffer.

$R-2R$ ladder

It is not very efficient to produce a ladder with very large differences in the resistance (as the binary weighted converter) is not very efficient when it comes to integration with semiconductors, therefore another scheme, the $R-2R$ ladder has been implemented. The basic principal of this converter is the same, for each step in the ladder, half the current should flow, and it should be done by closing switch. It is made possible through a very special configuration, where the total resistance of the remaining ladder always is $2R$. Accordingly the current divided at each branch, which goes through a $2R$ resistance to ground is half of the current flowing in to the branch. Independent on if the current is used for conversion or just dumped (fig. 5.9). Accordingly half the current that goes to the switch goes to the rest of the ladder, where in the next step the same thing occurs.

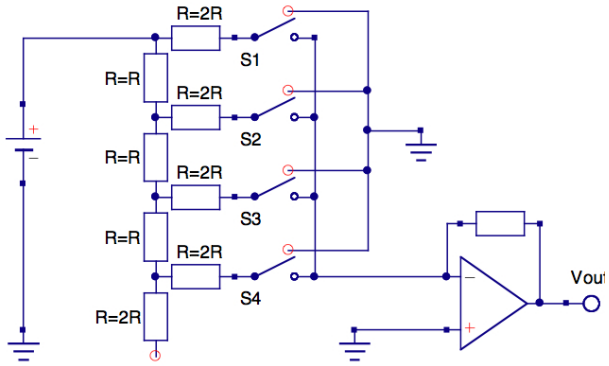


Figure 5.9: The schematic layout of a 2R-R DA converter, the switches S1-S4 are digitally controlled and will send a bit weighted current to the current to voltage converter.

5.8.2 Analogue to Digital Conversion (A/D, ADC) - Hardware

The input to an A/D converter can be coupled in three different manners:

Single-ended The input potential is measured with reference to an internal ground of the AD-device. Usually you can also access this ground through a separate connector/pin, by the way this is the most usual way an oscilloscope is coupled, the reference always grounded.

Pseudo-differential The input potential is measured with respect to a reference given by a input signal. However, the reference signal can not be input over the same range as the input signal but is intended to be used to offset slight variations (typically $< 100mV$) in the reference signal.

True differential The input signal is measured as the true difference between two signals, and the input range is the same for both. In DAQ devices this is often done through reserving one channel for each input, making the numbers of input channels half the number of single ended channels, but making it more easy to configure the right input. There is a problem with this approach and that is that it is done *after* digitisation which can cause larger noise than performing the same operation on the analogue signal (why?)[1].

The choice is dependent on the demand and the general noise level. In general single ended measurements calls for low noise surroundings.

Conversion strategies

A large number of strategies have been developed for A/D conversion. The most important factors is the sampling time, the conversion speed and the resolution. Generally the A/D converter is connected to an oscillator, generating clock pulses the conversion speed is therefore limited by clock speed. The conversion speed is expressed in number of clock cycles N_t. The resolution is expressed in number of bits n. While the conversion speed is generally not expressed.

- Flash Converter

[1]Because co-varying high frequency noise can be suppressed totally when subtracted in the analogue channel, while the digital signal has to be sampled, and accordingly the sampling has to be done at exactly the same time for this to work also on th digital side.

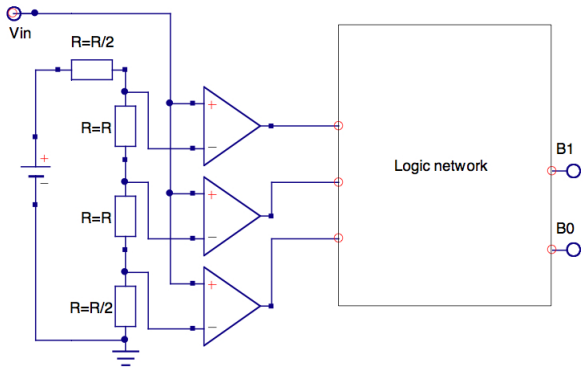


Figure 5.10: Schematic layout of the flash converter, the signal is compared for each level that is to resolved and the signals is sent to a binary network that outputs the corresponding binary number.

- Successive approximation converters
- Tracking converter
- Dual slope, integrating converter
- Delta Sigma ADC

Flash converter

The most straightforward strategy for conversion is the flash converter. In principle it is a resistance ladder connected to a large array of comparators (operational amplifiers designed only to compare the signal levels at the input) (fig. 5.10). The converter can do the comparison in one clock cycle, which renders a very fast sampling speed (limited by the speed of the comparators and the clock cycle). It is only limited by the input circuit characteristics regarding the conversion speed. The big disadvantage is the large numbers of complicated comparators, one is required for each level, so any high resolution is hard to achieve.

Successive approximation converters

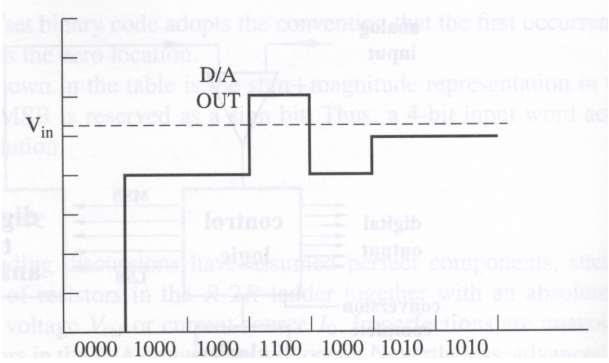


Figure 5.11: Example of how the result of a successive approximation converter convereges toward the input signal.

One of the more efficient procedures to find an arbitrary value is to use successive approximation. Here a comparison is made between a generated signal and th input signal. Only one comparator is needed. The conversion starts with comparing the value corresponding to the most significant bit with the input level. After comparison the value (higher or lower) is saved, and the comparison continues with comparing the input with the sum of the most and the next most significant bit, this yields the value of the most significant bit. This continues until the least significant bit has been reached. As each setting and comparison takes two clock cycles the total conversion time is bit dependent: $2n$ clock cycles. Today these are one of the dominating breeds of converters, implemented through charging capacitors and discharging them in a sophisticated order.

Tracking converter

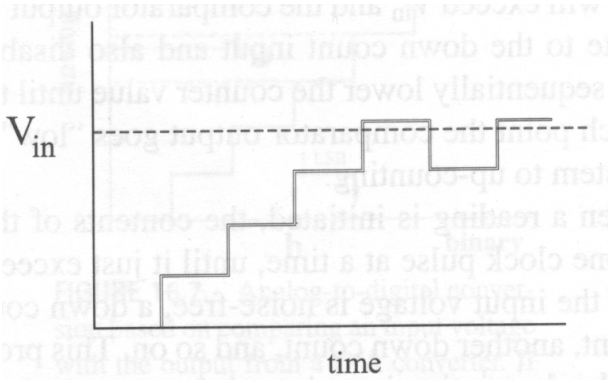


Figure 5.12: Example of how the result of a tracking converter converges toward the input signal.

In the tracking converter the signal is compared to the output of the AD converter and a up/down counter. Whenever the signal is higher than the output the counter will count upwards until it becomes larger than the signal level. It will then reverse the counting direction. It follows smooth signals very well, but has a maximum slope of the input signal of which is the height of the LSB divided by the clock period time.

Dual slope, integrating converter

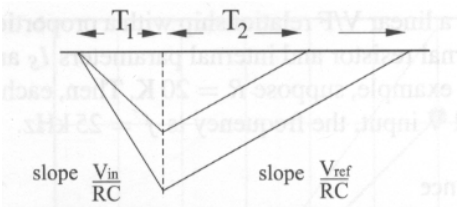


Figure 5.13: Runup and rundown of a dual slope converter. The runup slope is dependant on the input voltage while the rundown is fixed by the refernce voltage.

The work-horse of AD-conversion is the dual slope integrating converter. The conversion process takes place in two steps: first the run up, or integration, when a

capacitor is charged form a zero level by the supplied voltage, V_{in}, the second is the run-down where the capacitor is de-charged through a specified reference voltage, V_{ref}. The converter simply let the capacitor charge for a specific time. And then discharge it using a specified voltage, but keeping track of the time. It is thus possible to keep track of the charge collected in the capacitor, during run up

$$Q_1 = T_1 \frac{V_{in}}{RC}$$

and for the run down:

$$Q_2 = T_2 \frac{V_{out}}{RC}$$

It is then simple to find the input voltage:

$$V_{in} = V_{ref} \frac{T2}{T1}$$

The basic setback of the converter is the slow conversion (approximately 2^{n+1} clock cycles). The great advantage is the high resolution for an affordable price.

Delta Sigma converter

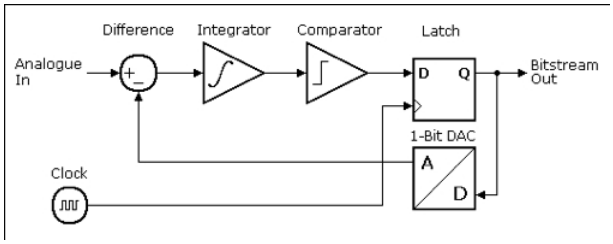


Figure 5.14: Schematic of a 1-bit delta-sigma converter. A feedback loop is utilised to keep the input close to zero, and the bistream generated to keep it at zero gives the converted signal (the bitstream is later converted to one number, reducing the time resolution but increasing the amplitude resolution).

The Delta-Sigma converter is built on the principle of using a poor resolution converter over a long time to obtain a high resolution average. To do this is not trivial, simply connecting a low resolution converter will yield the same result at each conversion and finally an extremely bad resolution. Instead a feedback system is implemented, where the integrated output signal over time is subtracted from the input. A feedback system is then used to keep the output as close to zero as possible.

The Delta Sigma converter has revolutionised the AD-converter market and marks a new paradigm data acquisition, offering superior sampling conditions to many of our everyday uses (audio sampling being one of the most prominent uses). It will be superior for many slower applications ¡1 MHz in the future.

5.9 Repetition questions

Basically you should know the limitations of sampling and the different conversion mechanism that can be used. The following repetition questions cover most of the topic:
Data conversion:

1. Explain why you need analog-to-digital and digital-to-analog conversion in instrumentation.
2. Explain the main difference between an analog and a digital signal
3. Describe the three main types of (time variation of) analog signals.
4. List the factors that determine the performance of a AD converter.
5. Explain what resolution and dynamic range for a n-bit AD or DA converter is.
6. Explain where the quantisation error comes from, and relate it to the converter resolution.
7. Calculate the quantisation error, resolution and dynamic range of a n-bit converter.
8. Explain how a single slope (ramp) AD converter works 9. Explain the advantage of the dual-slope converter over the single-slope converter.
9. Explain how a successive approximation AD converter works.
10. Explain how a flash converter works and list its main advantage and disadvantage.
11. Explain how a delta-sigma converter works and list its main advantage and disadvantage.
12. List the two main digital-to-analog (DA) conversion principles.
13. Explain why the current in the 2R-R resistance network is divided in half at each node.
14. Explain how a current can be converted into a voltage.

Sampling

1. Calculate the time needed for a AD converter with equivalent resistance R and capacitance C, to rise from zero at t=0 to within 1 LSB (the resolution) of the converter.
2. Understand how this rise time increases when the resolution decreases.
3. Draw the frequency spectrum of a simple sine wave, and a band-limited signal.
4. Draw the frequency spectrum of a sampled signal, and identify the difference between the original signals frequency spectrum and the sampled signals frequency spectrum.
5. Explain where the aliasing frequencies comes from.
6. Give an expression for the aliasing frequencies resulting from sampling a signal at frequency f0 at sampling frequency fs.
7. Explain how one can get rid of the aliasing frequencies higher than fs/2.
8. Explain why fs/2 is called the folding frequency.

9. Give an expression for the Nyquist frequency, and explain how sampling at frequencies higher than the Nyquist frequency combined with low-pass filtering (letting through only frequencies smaller than fs/2) will ensure identification of the correct frequency.
10. Explain how to use (analog) low-pass filters in connection to sampling to avoid aliasing high frequency noise into the frequency band of interest (-fs/2 to +fs/2).
11. Calculate the maximum frequency for an n-bit converter with a specified time needed for the conversion.
12. Explain when a sample-and-hold circuit must be used.
13. Calculate the sample time for a given signal frequency.

Chapter 6

Generic equipment

This chapter will deal with generic instruments, that is instruments for quite general use and their limitations.

6.1 Multimeter

A multimeter or a multitester is an electronic measuring instrument that combines several functions in one unit. The most basic instruments include a voltmeter, an ohmmeter, usually an ammeter, and often other testing equipment, such as capacitor and diode testing circuits. Analog multimeters are sometimes referred to as volt-ohm meters, abbreviated VOM. Digital multimeters are often abbreviated DMM.

Multimeters are the experimentalists best friend, the handheld variants are invaluable for faultfinding. Benchtop instruments are usually used in conjunction with an instrument bus (GPIB, LXI) and connected to a more complex measurement system.

6.1.1 Resolution

The resolution of a multimeter is often specified in "digits" of resolution. The producers simply specify the maximum resolution of the multimeter based on the digital display. By convention, a half digit can display either a zero or a one, while a three-quarters digit can display a numeral higher than a one but not nine. Commonly, a three-quarters digit refers to a maximum count of 3 or 5. The fractional digit is always the most significant digit in the displayed value. A 51/2 digit multimeter would have five full digits that display values from 0 to 9 and one half digit that could only display 0 or 1. Such a meter could show positive or negative values from 0 to 199,999. A 33/4 digit meter can display a quantity from 0 to 3,999 or 5,999, depending on the manufacturer.

6.1.2 Accuracy

Better circuitry and electronics have improved meter accuracy. Older analog meters might have basic accuracies of three to five percent. Modern portable DMMs may have accuracies as good as $\pm 0.01\%$, and high-end bench-top instruments can have accuracies in the hundredths of parts per million figures. At the other end of the spectrum, meters with $\pm 1\%$ basic accuracy are available for less than 200 NOK.

6.1.3 Manufacturers

- Fluke: http://www.fluke.com/
- Agilent (former HP): http://www.home.agilent.com/agilent/home.jspx
- Kiethley: http://www.keithley.com/
- Extech: http://www.extech.com/
- Meterman: http://www.metermantesttools.com
- Mastech: http://www.p-mastech.com/
- And many many more...

6.2 Oscilloscope

An oscilloscope (sometimes abbreviated CRO, for cathode-ray oscilloscope, or commonly just scope or O-scope) is a type of electronic test equipment that allows signal voltages to be viewed, usually as a two-dimensional graph of one or more electrical potential differences (vertical axis) plotted as a function of time or of some other voltage (horizontal axis)

6.2.1 Portable vs PC-based

Although most people think of an oscilloscope as a self-contained instrument in a box, a new type of "oscilloscope" is emerging that consists of a DAQ board (which can be an external USB or Parallel port device, or an internal add-on PCI or ISA card). The hardware itself usually consists of an electrical interface providing insulation and automatic gain controls, several hi-speed analogue-to-digital converters and some buffer memory, or even on-board DSPs. Depending on the exact hardware configuration, the hardware could be best described as a digitiser, a data logger or as a part of a specialised automatic control system. The PC provides the display, control interface, disc storage, networking and often the electrical power for the acquisition hardware. The viability of PC-based oscilloscopes depends on the current widespread use and low cost of standardised PCs. The acquisition hardware, in certain cases, may only consist of a standard sound card or even a game port, if only audio and low frequency signals are involved. The advantages of PC-based oscilloscopes include:

- Lower cost compared to a stand-alone oscilloscope.
- Easy exporting of data to standard PC software such as spreadsheets and word processors.
- Ability to control the instrument by running a custom program on the PC.
- Use of the PC's networking and disc storage functions, which cost extra when added to a self-contained oscilloscope.
- PC's typically have larger and higher resolution colour displays which can be easier to read.
- Easier portability when used with a laptop PC.

There are also some disadvantages, which include:

- Need for the owner to install oscilloscope software on the PC.
- Time taken for the PC to boot, compared with the almost instant start-up of a self-contained oscilloscope.

- Reduced portability when used with a desktop PC.
- Inconvenience of using part of the PC's screen for the oscilloscope display.
- If a sound card is used instead of dedicated signal acquisition hardware, frequency response is usually limited in the audio range, the number of inputs is limited by the number of recording channels (usually no more than the two usual stereo channels) and the inputs can handle only line-level voltages without the risk of damage.

One of the large manufacturers of oscilloscopes have produced a good general introduction/primer to oscilloscopes which is given on the following pages.:

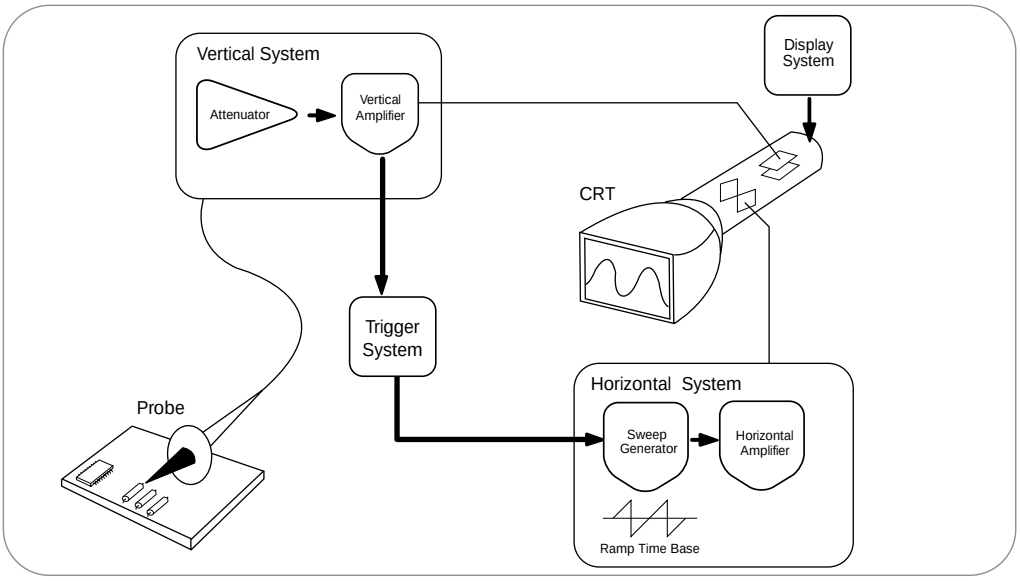


Figure 13. *The architecture of an analog oscilloscope.*

The Types of Oscilloscopes

Electronic equipment can be classified into two categories: analog and digital. Analog equipment works with continuously variable voltages, while digital equipment works with discrete binary numbers that represent voltage samples. A conventional phonograph is an analog device, while a compact disc player is a digital device.

Oscilloscopes can be classified similarly – as analog and digital types. For many applications, either an analog or digital oscilloscope will do. However, each type has unique characteristics that may make it more or less suitable for specific applications. Digital oscilloscopes can be further classified into digital storage oscilloscopes (DSOs), digital phosphor oscilloscopes (DPOs) and sampling oscilloscopes.

Analog Oscilloscopes

Fundamentally, an analog oscilloscope works by applying the measured signal voltage directly to the vertical axis of an electron beam that moves from left to right across the oscilloscope screen – usually a cathode-ray tube (CRT). The back side of the screen is treated with luminous phosphor that glows wherever the electron beam hits it. The signal voltage deflects the beam up and down proportionally as it moves horizontally across the display, tracing the waveform on the screen. The more frequently the beam hits a particular screen location, the more brightly it glows.

The CRT limits the range of frequencies that can be displayed by an analog oscilloscope. At very low frequencies, the signal appears as a bright, slow-moving dot that is difficult to distinguish as a waveform. At high frequencies, the CRT's writing speed defines the limit. When the signal frequency exceeds the CRT's writing speed, the display becomes too dim to see. The fastest analog oscilloscopes can display frequencies up to about 1 GHz.

When you connect an oscilloscope probe to a circuit, the voltage signal travels through the probe to the vertical system of the oscilloscope. Figure 13 illustrates how an analog oscilloscope displays a measured signal. Depending on how you set the vertical scale (volts/div control), an attenuator reduces the signal voltage and an amplifier increases the signal voltage.

Next, the signal travels directly to the vertical deflection plates of the CRT. Voltage applied to these deflection plates causes a glowing dot to move across the screen. The glowing dot is created by an electron beam that hits the luminous phosphor inside the CRT. A positive voltage causes the dot to move up while a negative voltage causes the dot to move down.

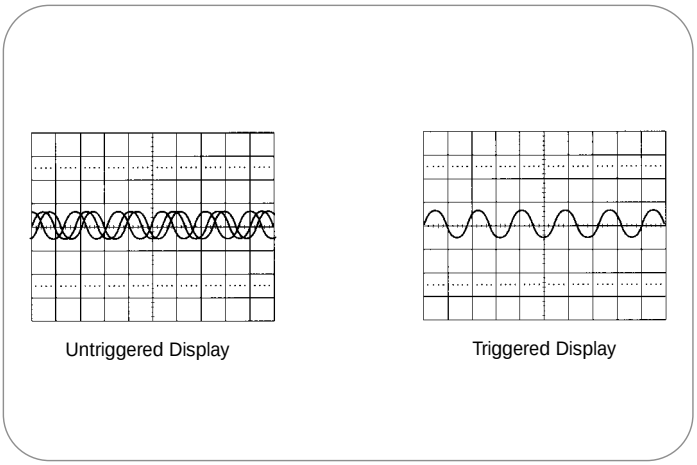


Figure 14. *The trigger stabilizes a repetitive waveform, creating a clear picture of the signal.*

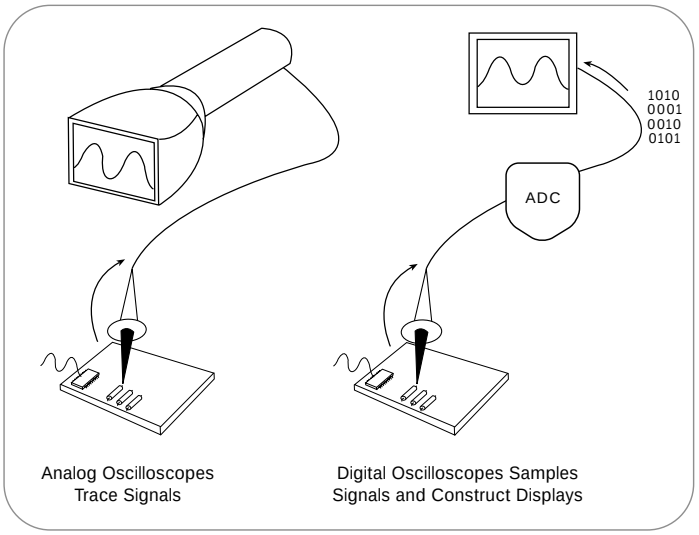


Figure 15. *Analog oscilloscopes trace signals, while digital oscilloscopes sample signals and construct displays.*

The signal also travels to the trigger system to start, or trigger, a horizontal sweep. Horizontal sweep refers to the action of the horizontal system that causes the glowing dot to move across the screen. Triggering the horizontal system causes the horizontal time base to move the glowing dot across the screen from left to right within a specific time interval. Many sweeps in rapid sequence cause the movement of the glowing dot to blend into a solid line. At higher speeds, the dot may sweep across the screen up to 500,000 times per second.

Together, the horizontal sweeping action and the vertical deflection action trace a graph of the signal on the screen. The trigger is necessary to stabilize a repeating signal – it ensures that the sweep begins at the same point of a repeating signal, resulting in a clear picture as shown in Figure 14.

In addition, analog oscilloscopes have focus and intensity controls that can be adjusted to create a sharp, legible display.

People often prefer analog oscilloscopes when it is important to display rapidly varying signals in "real time" – or, as they occur. The analog oscilloscope's chemical phosphor-based display has a characteristic known as intensity grading that makes the trace brighter wherever the signal features occur most often. This intensity grading makes it easy to distinguish signal details just by looking at the trace's intensity levels.

Digital Oscilloscopes

In contrast to an analog oscilloscope, a digital oscilloscope uses an analog-to-digital converter (ADC) to convert the measured voltage into digital information. It acquires the waveform as a series of samples, and stores these samples until it accumulates enough samples to describe a waveform. The digital oscilloscope then re-assembles the waveform for display on the screen. (see Figure 15)

Digital oscilloscopes can be classified into digital storage oscilloscopes (DSOs), digital phosphor oscilloscopes (DPOs), and sampling oscilloscopes.

The digital approach means that the oscilloscope can display any frequency within its range with stability, brightness, and clarity. For repetitive signals, the bandwidth of the digital oscilloscope is a function of the analog bandwidth of the front-end components of the oscilloscope, commonly referred to as the –3dB point. For single-shot and transient events, such as pulses and steps, the bandwidth can be limited by the oscilloscope's sample rate. Please refer to the Sample Rate section under Performance Terms and Considerations for a more detailed discussion.

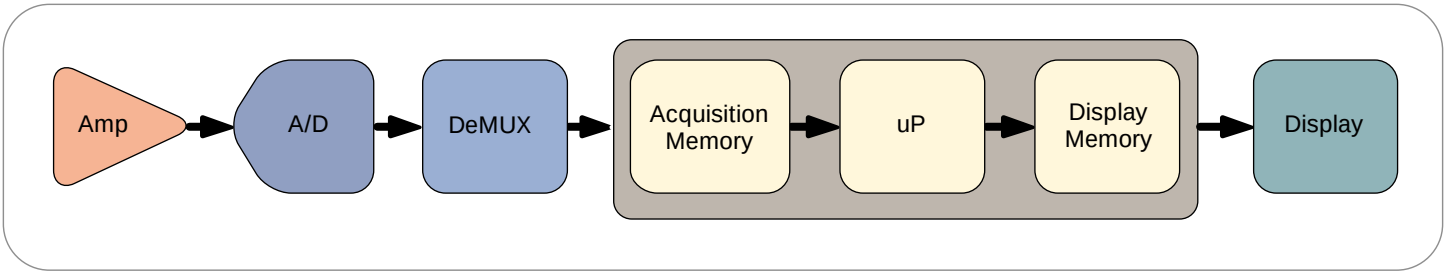


Figure 16. *The serial-processing architecture of a digital storage oscilloscope (DSO).*

Digital Storage Oscilloscopes

A conventional digital oscilloscope is known as a digital storage oscilloscope (DSO). Its display typically relies on a raster-type screen rather than luminous phosphor.

Digital storage oscilloscopes (DSOs) allow you to capture and view events that may happen only once – known as transients. Because the waveform information exists in digital form as a series of stored binary values, it can be analyzed, archived, printed, and otherwise processed, within the oscilloscope itself or by an external computer. The waveform need not be continuous; it can be displayed even when the signal disappears. Unlike analog oscilloscopes, digital storage oscilloscopes provide permanent signal storage and extensive waveform processing. However, DSOs typically have no real-time intensity grading; therefore, they cannot express varying levels of intensity in the live signal.

Some of the subsystems that comprise DSOs are similar to those in analog oscilloscopes. However, DSOs contain additional data-processing subsystems that are used to collect and display data for the entire waveform. A DSO employs a serial-processing architecture to capture and display a signal on its screen, as shown in Figure 16. A description of this serial-processing architecture follows.

Serial-processing Architecture

Like an analog oscilloscope, a DSO's first (input) stage is a vertical amplifier. Vertical controls allow you to adjust the amplitude and position range at this stage.

Next, the analog-to-digital converter (ADC) in the horizontal system samples the signal at discrete points in time and converts the signal's voltage at these points into digital values called **sample points**. This process is referred to as **digitizing** a signal. The horizontal system's sample clock determines how often the ADC takes a sample. This rate is referred to as the **sample rate** and is expressed in samples per second (S/s).

The sample points from the ADC are stored in acquisition memory as waveform points. Several sample points may comprise one waveform point. Together, the waveform points comprise one waveform record. The number of waveform points used to create a waveform record is called the record length. The trigger system determines the start and stop points of the record.

The DSO's signal path includes a microprocessor through which the measured signal passes on its way to the display. This microprocessor processes the signal, coordinates display activities, manages the front panel controls, and more. The signal then passes through the display memory and is displayed on the oscilloscope screen.

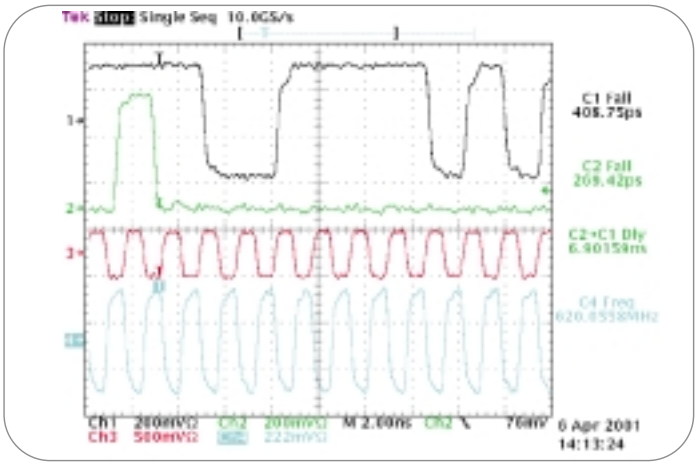


Figure 17. *The TDS694C delivers high-speed, single-shot acquisition across multiple channels, increasing the likelihood of capturing elusive glitches and transient events.*

Depending on the capabilities of your oscilloscope, additional processing of the sample points may take place, which enhances the display. Pre-trigger may also be available, enabling you to see events before the trigger point. Most of today's digital oscilloscopes also provide a selection of automatic parametric measurements, simplifying the measurement process.

A DSO provides high performance in a single-shot, multi-channel instrument (see Figure 17). DSOs are ideal for low-repetition-rate or single-shot, high-speed, multi-channel design applications. In the real world of digital design, an engineer usually examines four or more signals simultaneously, making the DSO a critical companion.

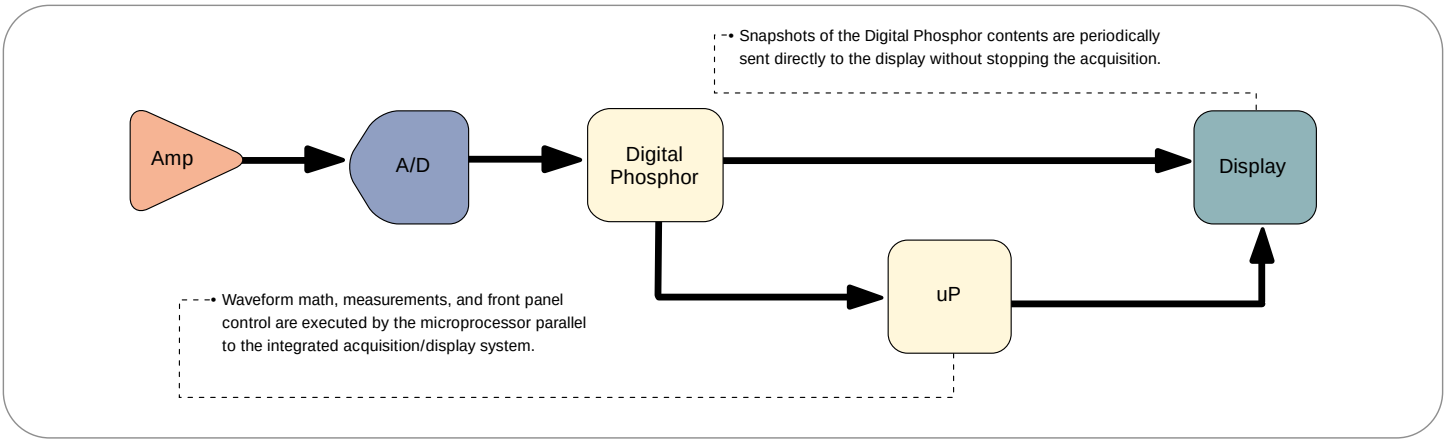


Figure 18. *The parallel-processing architecture of a digital phospor oscilloscope (DPO).*

Digital Phosphor Oscilloscopes

The digital phosphor oscilloscope (DPO) offers a new approach to oscilloscope architecture. This architecture enables a DPO to deliver unique acquisition and display capabilities to accurately reconstruct a signal.

While a DSO uses a serial-processing architecture to capture, display and analyze signals, a DPO employs a parallel-processing architecture to perform these functions, as shown in Figure 18. The DPO architecture dedicates unique ASIC hardware to acquire waveform images, delivering high waveform capture rates that result in a higher level of signal visualization. This performance increases the probability of witnessing transient events that occur in digital systems, such as runt pulses, glitches and transition errors. A description of this parallel-processing architecture follows.

Parallel-processing Architecture

A DPO's first (input) stage is similar to that of an analog oscilloscope – a vertical amplifier – and its second stage is similar to that of a DSO – an ADC. But, the DPO differs significantly from its predecessors following the analog-to-digital conversion.

For any oscilloscope – analog, DSO or DPO – there is always a holdoff time during which the instrument processes the most recently acquired data, resets the system, and waits for the next trigger event. During this time, the oscilloscope is blind to all signal activity. The probability of seeing an infrequent or low-repetition event decreases as the holdoff time increases.

It should be noted that it is impossible to determine the probability of capture by simply looking at the display update rate. If you rely solely on the update rate, it is easy to make the mistake of believing that the oscilloscope is capturing all pertinent information about the waveform when, in fact, it is not.

The digital storage oscilloscope processes captured waveforms serially. The speed of its microprocessor is a bottleneck in this process because it limits the waveform capture rate.

The DPO rasterizes the digitized waveform data into a digital phosphor database. Every 1/30th of a second – about as fast as the human eye can perceive it – a snapshot of the signal image that is stored in the database is pipelined directly to the display system. This direct rasterization of waveform data, and direct copy to display memory from the database, removes the data-processing bottleneck inherent in other architectures. The result is an enhanced "live-time" and lively display update. Signal details, intermittent events, and dynamic characteristics of the signal are captured in real-time. The DPO's microprocessor works in parallel with this integrated acquisition system for display management, measurement automation and instrument control, so that it does not affect the oscilloscope's acquisition speed.

A DPO faithfully emulates the best display attributes of an analog oscilloscope, displaying the signal in three dimensions: time, amplitude and the distribution of amplitude over time, all in real time.

Unlike an analog oscilloscope's reliance on chemical phosphor, a DPO uses a purely electronic digital phosphor that's actually a continuously updated database. This database has a separate "cell" of information for every single pixel in the oscilloscope's display. Each time a waveform is captured – in other words, every time the oscilloscope triggers – it is mapped into the digital phosphor database's cells. Each cell that represents a screen location and is touched by the waveform is reinforced with intensity information, while other cells are not. Thus, intensity information builds up in cells where the waveform passes most often.

When the digital phosphor database is fed to the oscilloscope's display, the display reveals intensified waveform areas, in proportion to the signal's frequency of occurrence at each point – much like the intensity grading characteristics of an analog oscilloscope. The DPO also allows the display of the varying frequency-of-occurence information on the display as contrasting colors, unlike an analog oscilloscope. With a DPO, it is easy to see the difference between a waveform that occurs on almost every trigger and one that occurs, say, every 100th trigger.

Digital phosphor oscilloscopes (DPOs) break down the barrier between analog and digital oscilloscope technologies. They are equally suitable for viewing high and low frequencies, repetitive waveforms, transients, and signal variations in real time. Only a DPO provides the Z (intensity) axis in real time that is missing from conventional DSOs.

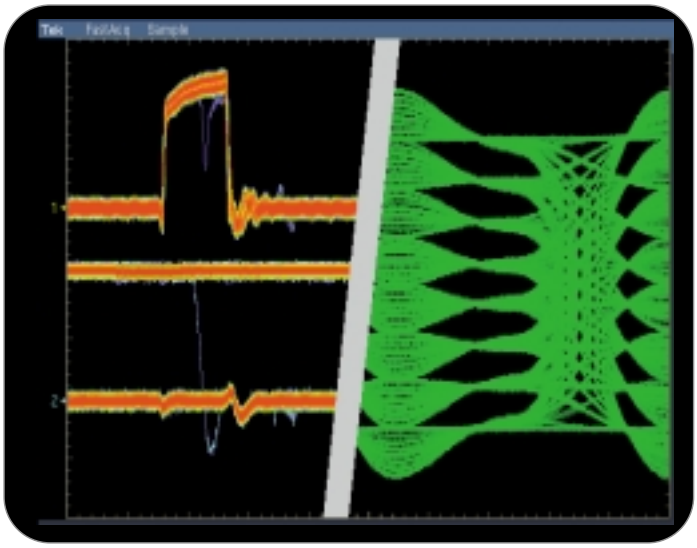


Figure 19. *Some DPOs can acquire millions of waveforms in just seconds, significantly increasing the probability of capturing intermittent and elusive events and revealing dynamic signal behavior.*

A DPO is ideal for those who need the best general-purpose design and troubleshooting tool for a wide range of applications (see Figure 19). A DPO is exemplary for communication mask testing, digital debug of intermittent signals, repetitive digital design and timing applications.

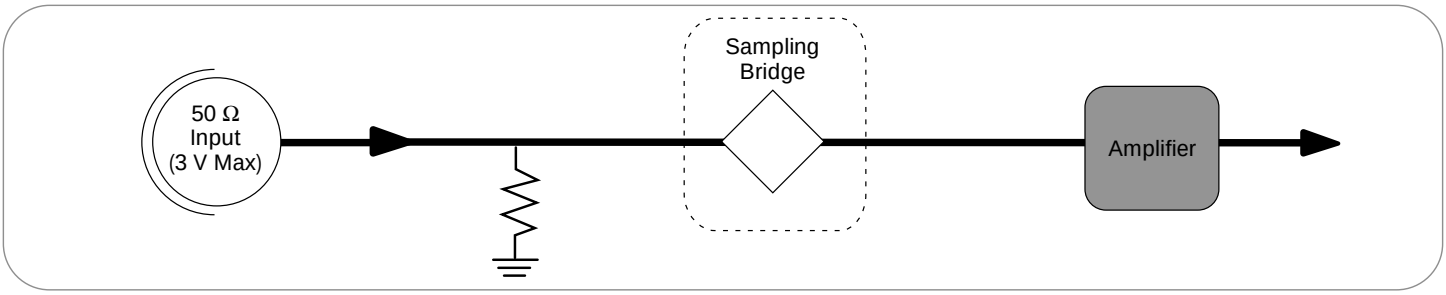


Figure 20. *The architecture of a digital sampling oscilloscope.*

Digital Sampling Oscilloscopes

When measuring high-frequency signals, the oscilloscope may not be able to collect enough samples in one sweep. A digital sampling oscilloscope is an ideal tool for accurately capturing signals whose frequency components are much higher than the oscilloscope's sample rate (see Figure 21). This oscilloscope is capable of measuring signals of up to an order of magnitude faster than any other oscilloscope. It can achieve bandwidth and high-speed timing ten times higher than other oscilloscopes for repetitive signals. Sequential equivalent-time sampling oscilloscopes are available with bandwidths to 50 GHz.

In contrast to the digital storage and digital phosphor oscilloscope architectures, the architecture of the digital sampling oscilloscope reverses the position of the attenuator/amplifier and the sampling bridge, as shown in Figure 20. The input signal is sampled before any attenuation or amplification is performed. A low bandwidth amplifier can then be utilized after the sampling bridge because the signal has already been converted to a lower frequency by the sampling gate, resulting in a much higher bandwidth instrument.

The tradeoff for this high bandwidth, however, is that the sampling oscilloscope's dynamic range is limited. Since there is no attenuator/amplifier in front of the sampling gate, there is no facility to scale the input. The sampling bridge must be able to handle the full dynamic range of the input at all times. Therefore, the dynamic range of most sampling oscilloscopes is limited to about 1 V peak-to-peak. Digital storage and digital phosphor oscilloscopes, on the other hand, can handle 50 to 100 volts.

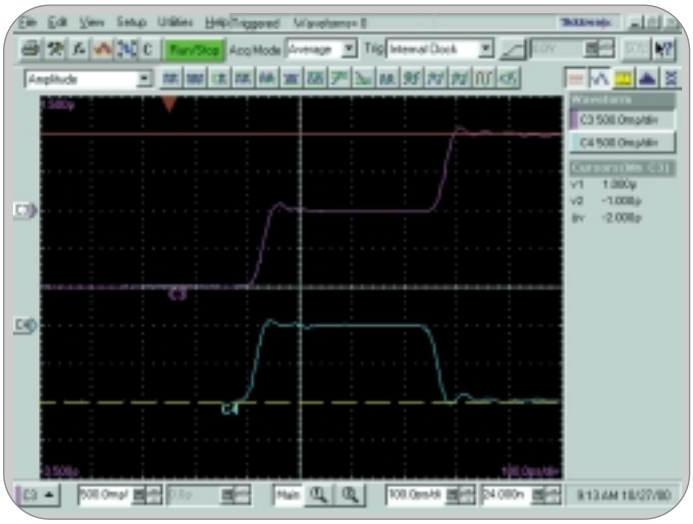


Figure 21. *Time domain reflectometry (TDR) display from a TDS8000 digital sampling oscilloscope and 80E04 20-GHz sampling module.*

In addition, protection diodes cannot be placed in front of the sampling bridge as this would limit the bandwidth. This reduces the safe input voltage for a sampling oscilloscope to about 3 V, as compared to 500 V available on other oscilloscopes.

We have described the basic oscilloscope controls that a beginner needs to know about. Your oscilloscope may have other controls for various functions. Some of these may include:

- Automatic parametric measurements
- Measurement cursors
- Keypads for mathematical operations or data entry
- Printing capabilities
- Interfaces for connecting your oscilloscope to a computer or directly to the Internet

Look over the other options available to you and read your oscilloscope's manual to find out more about these other controls.

The Complete Measurement System

Probes

Even the most advanced instrument can only be as precise as the data that goes into it. A probe functions in conjunction with an oscilloscope as part of the measurement system. Precision measurements start at the probe tip. The right probes matched to the oscilloscope and the device-under-test (DUT) not only allow the signal to be brought to the oscilloscope cleanly, they also amplify and preserve the signal for the greatest signal integrity and measurement accuracy.

- **To ensure accurate reconstruction of your signal, try to choose a probe that, when paired with your oscilloscope, exceeds the signal bandwidth by 5 times.**

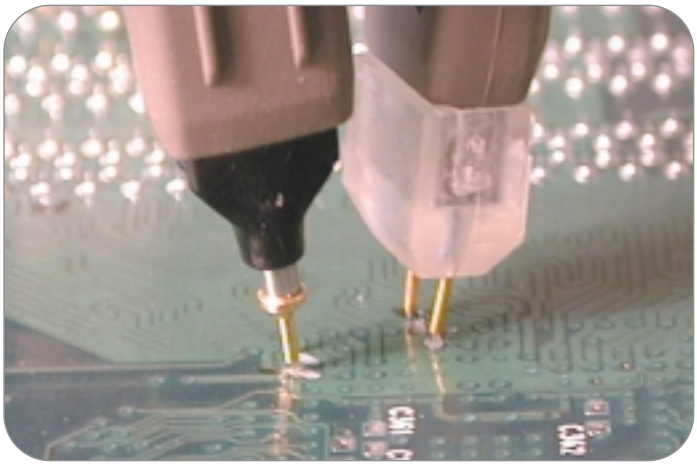


Figure 40. *Dense devices and systems require small form factor probes.*

Probes actually become part of the circuit, introducing resistive, capacitive and inductive loading that inevitably alters the measurement. For the most accurate results, the goal is to select a probe with minimal loading. An ideal pairing of the probe with the oscilloscope will minimize this loading, and enable you to access all of the power, features and capabilities of your oscilloscope.

Another consideration in the selection of the all-important connection to your DUT is the probe's form factor. Small form factor probes provide easier access to today's densely packed circuitry (see Figure 40).

A description of the types of probes follows. Please refer to Tektronix' *ABCs of Probes* primer for more information about this essential component of the overall measurement system.

Figure 41. *A typical passive probe with accessories.*

Passive Probes

For measuring typical signal and voltage levels, passive probes provide ease-of-use and a wide range of measurement capabilities at an affordable price. The pairing of a passive voltage probe with a current probe will provide you with an ideal solution for measuring power.

Most passive probes have some attenuation factor, such as 10X, 100X, and so on. By convention, attenuation factors, such as for the 10X attenuator probe, have the X after the factor. In contrast, magnification factors like X10 have the X first.

The 10X (read as "ten times") attenuator probe reduces circuit loading in comparison to a 1X probe and is an excellent general-purpose passive probe. Circuit loading becomes more pronounced for higher frequency and/or higher impedance signal sources, so be sure to analyze these signal/probe loading interactions before selecting a probe. The 10X attenuator probe improves the accuracy of your measurements, but also reduces the signal's amplitude at the oscilloscope input by a factor of 10.

Because it attenuates the signal, the 10X attenuator probe makes it difficult to look at signals less than 10 millivolts peak-to-peak. The 1X probe is similar to the 10X attenuator probe but lacks the attenuation circuitry. Without this circuitry, more interference is introduced to the circuit being tested. Use the 10X attenuator probe as your general-purpose probe, but keep the 1X probe accessible to measure slow-speed, low-amplitude signals. Some probes have a convenient feature for switching between 1X and 10X attenuation at the probe tip. If your probe has this feature, make sure you are using the correct setting before taking measurements.

Many oscilloscopes can detect whether you are using a 1X or 10X probe and adjust their screen readouts accordingly. However with some oscilloscopes, you must set the type of probe you are using or read from the proper 1X or 10X marking on the volts/div control.

The 10X attenuator probe works by balancing the probe's electrical properties against the oscilloscope's electrical properties. Before using a 10X attenuator probe you need to adjust this balance for your particular oscilloscope. This adjustment is known as compensating the probe and is described in more detail in the Operating the Oscilloscope section of this primer.

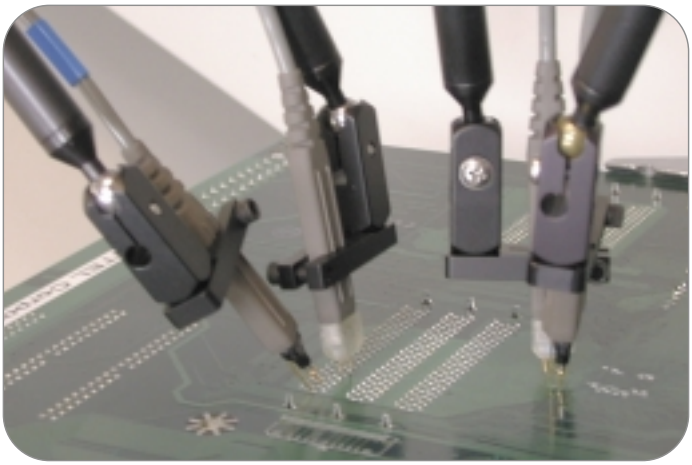


Figure 42. *High-performance probes are critical when measuring the fast clocks and edges found in today's computer buses and data transmission lines.*

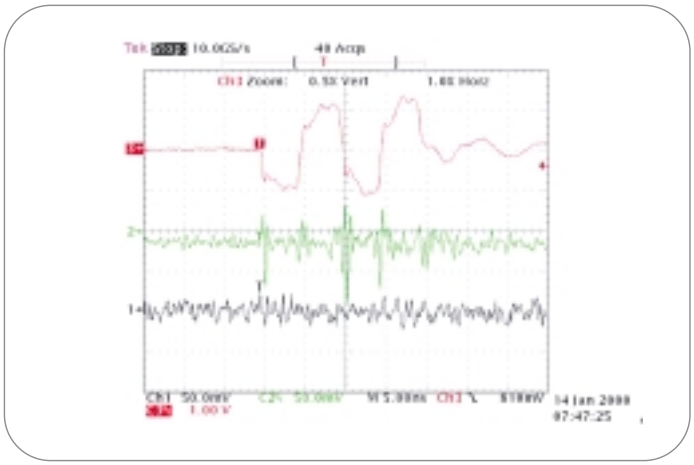


Figure 43. *Differential probes can separate common-mode noise from the signal content of interest in today's fast, low-voltage applications – especially important as digital signals continue to fall below typical noise thresholds found in integrated circuits.*

Passive probes provide excellent general-purpose probing solutions. However, general-purpose passive probes cannot accurately measure signals with extremely fast rise times, and may excessively load sensitive circuits. The steady increase in signal clock rates and edge speeds demands higher speed probes with less loading effects. High-speed active and differential probes provide ideal solutions when measuring high-speed and/or differential signals.

Active and Differential Probes

Increasing signal speeds and lower-voltage logic families make accurate measurement results difficult to achieve. Signal fidelity and device loading are critical issues. A complete measurement solution at these high speeds includes high-speed, high-fidelity probing solutions to match the performance of the oscilloscope (see Figure 42).

Active and differential probes use specially developed integrated circuits to preserve the signal during access and transmission to the oscilloscope, ensuring signal integrity. For measuring signals with fast rise times, a high-speed active or differential probe will provide more accurate results.

Figure 44. *The Tektronix TekConnect™ interface preserves signal integrity to 10 GHz and beyond to meet present and future bandwidth needs.*

Figure 45. *The Tektronix SF200A and SF500 Series SureFoot™ adapters provide reliable short-lead length probe tip connection to a specific pin on an integrated circuit.*

Probe Accessories

Many modern oscilloscopes provide special automated features built into the input and mating probe connectors. In the case of intelligent probe interfaces, the act of connecting the probe to the instrument notifies the oscilloscope about the probe's attenuation factor, which in turn scales the display so that the probe's attenuation is figured into the readout on the screen. Some probe interfaces also recognize the type of probe – that is, passive, active or current. The interface may act as a DC power source for probes. Active probes have their own amplifier and buffer circuitry that requires DC power.

Ground lead and probe tip accessories are also available to improve signal integrity when measuring high-speed signals. Ground lead adapters provide spacing flexibility between probe tip and ground lead connections to the DUT, while maintaining very short lead lengths from probe tip to DUT.

Please refer to Tektronix' *ABCs of Probes* primer for more information about probe accessories.

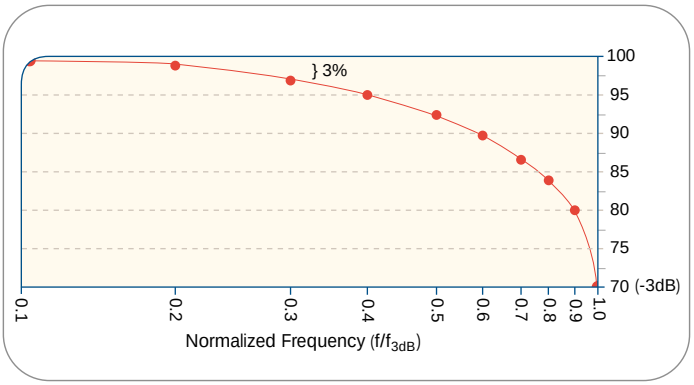


Figure 46. *Oscilloscope bandwidth is the frequency at which a sinusoidal input signal is attenuated to 70.7% of the signal's true amplitude, known as the –3 dB point.*

Performance Terms and Considerations

As previously mentioned, an oscilloscope is analogous to a camera that captures signal images that we can observe and interpret. Shutter speed, lighting conditions, aperture and the ASA rating of the film all affect the camera's ability to capture an image clearly and accurately. Like the basic systems of an oscilloscope, the performance considerations of an oscilloscope significantly affect its ability to achieve the required signal integrity.

Learning a new skill often involves learning a new vocabulary. This idea holds true for learning how to use an oscilloscope. This section describes some useful measurement and oscilloscope performance terms. These terms are used to describe the criteria essential to choosing the right oscilloscope for your application. Understanding these terms will help you to evaluate and compare your oscilloscope with other models.

Bandwidth

Bandwidth determines an oscilloscope's fundamental ability to measure a signal. As signal frequency increases, the capability of the oscilloscope to accurately display the signal decreases. This specification indicates the frequency range that the oscilloscope can accurately measure.

Oscilloscope bandwidth is specified as the frequency at which a sinusoidal input signal is attenuated to 70.7% of the signal's true amplitude, known as the –3 dB point, a term based on a logarithmic scale (see Figure 46).

Figure 47. *The higher the bandwidth, the more accurate the reproduction of your signal, as illustrated with a signal captured at 250 MHz, 1 GHz and 4 GHz bandwidth levels.*

Without adequate bandwidth, your oscilloscope will not be able to resolve high-frequency changes. Amplitude will be distorted. Edges will vanish. Details will be lost. Without adequate bandwidth, all the features, bells and whistles in your oscilloscope will mean nothing.

The 5 Times Rule
Oscilloscope Bandwidth Required = Highest Frequency Component of Measured Signal x 5

To determine the oscilloscope bandwidth needed to accurately characterize signal amplitude in your specific application, apply the "5 Times Rule."

An oscilloscope selected using the 5 Times Rule will give you less than +/-2% error in your measurements – typically sufficient for today's applications. However, as signal speeds increase, it may not be possible to achieve this rule of thumb. Always keep in mind that higher bandwidth will likely provide more accurate reproduction of your signal (see Figure 47).

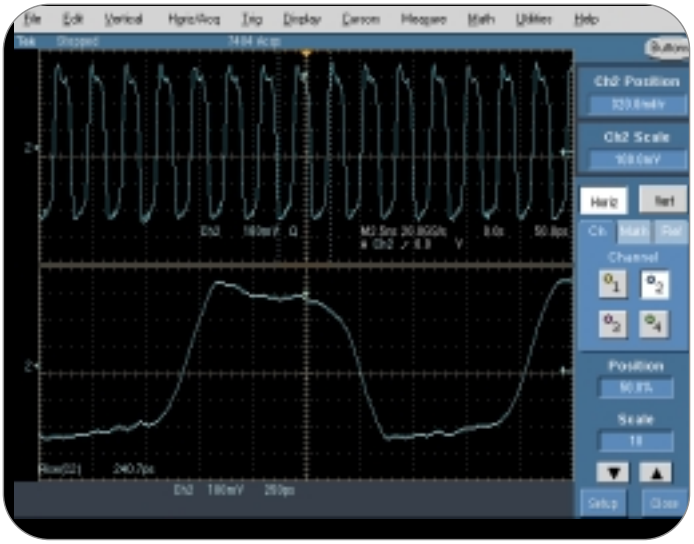


Figure 48. Rise time characterization of a high-speed digital signal.

Logic Family	Typical Signal Rise Time	Calculated Signal Bandwidth
TTL	2 ns	175 MHz
CMOS	1.5 ns	230 MHz
GTL	1 ns	350 MHz
LVDS	400 ps	875 MHz
ECL	100 ps	3.5 GHz
GaAs	40 ps	8.75 GHz

Figure 49. Some logic families produce inherently faster rise times than others.

Rise Time

In the digital world, rise time measurements are critical. Rise time may be a more appropriate performance consideration when you expect to measure digital signals, such as pulses and steps. Your oscilloscope must have sufficient rise time to accurately capture the details of rapid transitions.

Rise time describes the useful frequency range of an oscilloscope. To calculate the oscilloscope rise time required for your signal type, use the following equation:

Oscilloscope Rise Time Required = Fastest Rise Time of Measured Signal ÷ 5

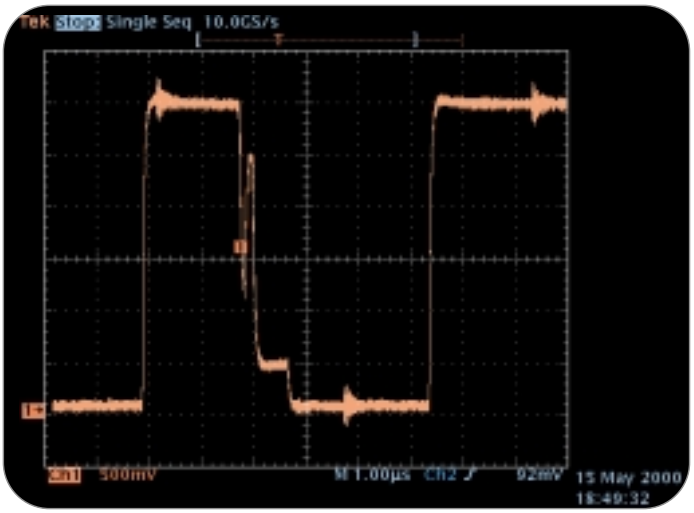
Note that this basis for oscilloscope rise time selection is similar to that for bandwidth. As in the case of bandwidth, achieving this rule of thumb may not always be possible given the extreme speeds of today's signals. Always remember that an oscilloscope with faster rise time will more accurately capture the critical details of fast transitions.

In some applications, you may know only the rise time of a signal. A constant allows you to relate the bandwidth and rise time of the oscilloscope, using the equation:

$$\text{Bandwidth} = \frac{k}{\text{Rise Time}}$$

where k is a value between 0.35 and 0.45, depending on the shape of the oscilloscope's frequency response curve and pulse rise time response. Oscilloscopes with a bandwidth of <1 GHz typically have a 0.35 value, while oscilloscopes with a bandwidth >1 GHz usually have a value between 0.40 and 0.45.

Some logic families produce inherently faster rise times than others, as illustrated in Figure 49.



▶ ***Figure 50.*** *A higher sample rate provides greater signal resolution, ensuring that you'll see intermittent events.*

Sample Rate

Sample rate – specified in samples per second (S/s) – refers to how frequently a digital oscilloscope takes a snapshot or sample of the signal, analogous to the frames on a movie camera. The faster an oscilloscope samples (i.e., the higher the sample rate), the greater the resolution and detail of the displayed waveform and the less likely that critical information or events will be lost, as shown in Figure 50. The minimum sample rate may also be important if you need to look at slowly changing signals over longer periods of time. Typically, the displayed sample rate changes with changes made to the horizontal scale control to maintain a constant number of waveform points in the displayed waveform record.

How do you calculate your sample rate requirements? The method differs based on the type of waveform you are measuring, and the method of signal reconstruction used by the oscilloscope.

In order to accurately reconstruct a signal and avoid aliasing, Nyquist theorem says that the signal must be sampled at least twice as fast as its highest frequency component. This theorem, however, assumes an infinite record length and a continuous signal. Since no oscilloscope offers infinite record length and, by definition, glitches are not continuous, sampling at only twice the rate of highest frequency component is usually insufficient.

In reality, accurate reconstruction of a signal depends on both the sample rate and the interpolation method used to fill in the spaces between the samples. Some oscilloscopes let you select either sin (x)/x interpolation for measuring sinusoidal signals, or linear interpolation for square waves, pulses and other signal types.

▶ **For accurate reconstruction using sin(x)/x interpolation, your oscilloscope should have a sample rate at least 2.5 times the highest frequency component of your signal. Using linear interpolation, sample rate should be at least 10 times the highest frequency signal component.**

Some measurement systems with sample rates to 20 GS/s and bandwidths to 4 GHz have been optimized for capturing very fast, single-shot and transient events by oversampling up to 5 times the bandwidth.

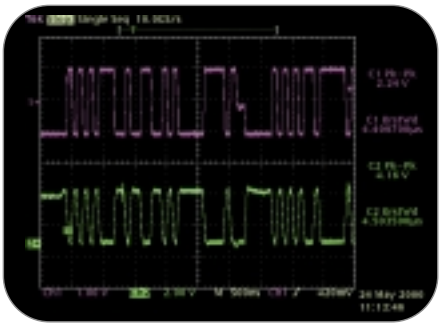


Figure 51. *A DSO provides an ideal solution for non-repetitive, high-speed, multi-channel digital design applications.*

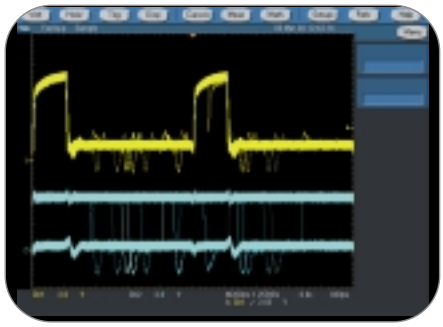


Figure 52. *A DPO enables a superior level of insight into signal behavior by delivering vastly greater waveform capture rates and three-dimensional display, making it the best general-purpose design and troubleshooting tool for a wide range of applications.*

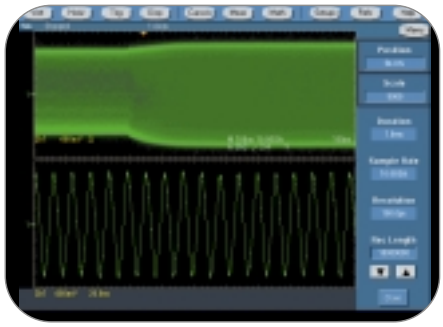


Figure 53. *Capturing the high frequency detail of this modulated 85 MHz carrier requires high resolution sampling (100 ps). Seeing the signal's complete modulation envelope requires a long time duration (1 ms). Using long record length (10 MB), the oscilloscope can display both.*

Waveform Capture Rate

All oscilloscopes blink. That is, they open their eyes a given number of times per second to capture the signal, and close their eyes in between. This is the waveform capture rate, expressed as waveforms per second (wfms/s). While the sample rate indicates how frequently the oscilloscope samples the input signal within one waveform, or cycle, the waveform capture rate refers to how quickly an oscilloscope acquires waveforms.

Waveform capture rates vary greatly, depending on the type and performance level of the oscilloscope. Oscilloscopes with high waveform capture rates provide significantly more visual insight into signal behavior, and dramatically increase the probability that the oscilloscope will quickly capture transient anomalies such as jitter, runt pulses, glitches and transition errors. (Refer to Figures 51 and 52.)

Digital storage oscilloscopes (DSOs) employ a serial-processing architecture to capture from 10 to 5,000 wfms/s. Some DSOs provide a special mode that bursts multiple captures into long memory, temporarily delivering higher waveform capture rates followed by long processing dead times that reduce the probability of capturing rare, intermittent events.

Most digital phosphor oscilloscopes (DPOs) employ a parallel-processing architecture to deliver vastly greater waveform capture rates. Some DPOs can acquire millions of waveforms in just seconds, significantly increasing the probability of capturing intermittent and elusive events and allowing you to see the problems in your signal more quickly. Moreover, the DPO's ability to acquire and display three dimensions of signal behavior in real time – amplitude, time and distribution of amplitude over time – results in a superior level of insight into signal behavior.

Record Length

Record length, expressed as the number of points that comprise a complete waveform record, determines the amount of data that can be captured with each channel. Since an oscilloscope can store only a limited number of samples, the waveform duration (time) will be inversely proportional to the oscilloscope's sample rate.

$$\text{Time Interval} = \frac{\text{Record Length}}{\text{Sample Rate}}$$

Modern oscilloscopes allow you to select record length to optimize the level of detail needed for your application. If you are analyzing an extremely stable sinusoidal signal, you may need only a 500-point record length, but if you are isolating the causes of timing anomalies in a complex digital data stream, you may need a million points or more for a given record length.

Triggering Capabilities

An oscilloscope's trigger function synchronizes the horizontal sweep at the correct point of the signal, essential for clear signal characterization. Trigger controls allow you to stabilize repetitive waveforms and capture single-shot waveforms.

Please refer to the Trigger section under Performance Terms and Considerations for more information regarding triggering capabilities.

Effective Bits

Effective bits represent a measure of a digital oscilloscope's ability to accurately reconstruct a sinewave signal's shape. This measurement compares the oscilloscope's actual error to that of a theoretical "ideal" digitizer. Because the actual errors include noise and distortion, the frequency and amplitude of the signal must be specified.

Frequency Response

Bandwidth alone is not enough to ensure that an oscilloscope can accurately capture a high frequency signal. The goal of oscilloscope design is a specific type of frequency response: Maximally Flat Envelope Delay (MFED). A frequency response of this type delivers excellent pulse fidelity with minimum overshoot and ringing. Since a digital oscilloscope is composed of real amplifiers, attenuators, ADCs, interconnects, and relays, MFED response is a goal that can only be approached. Pulse fidelity varies considerably with model and manufacturer. (Figure 46 illustrates this concept.)

Vertical Sensitivity

Vertical sensitivity indicates how much the vertical amplifier can amplify a weak signal – usually measured in millivolts (mV) per division. The smallest voltage detected by a general-purpose oscilloscope is typically about 1 mV per vertical screen division.

Sweep Speed

Sweep speed indicates how fast the trace can sweep across the oscilloscope screen, enabling you to see fine details. The sweep speed of an oscilloscope is represented by time (seconds) per division.

Gain Accuracy

Gain accuracy indicates how accurately the vertical system attenuates or amplifies a signal, usually represented as a percentage error.

Horizontal Accuracy (Time Base)

Horizontal, or time base, accuracy indicates how accurately the horizontal system displays the timing of a signal, usually represented as a percentage error.

Vertical Resolution (Analog-to-Digital Converter)

Vertical resolution of the ADC, and therefore, the digital oscilloscope, indicates how precisely it can convert input voltages into digital values. Vertical resolution is measured in bits. Calculation techniques can improve the effective resolution, as exemplified with hi-res acquisition mode. Please refer to the Horizontal System and Controls section under The Systems and Controls of an Oscilloscope section.

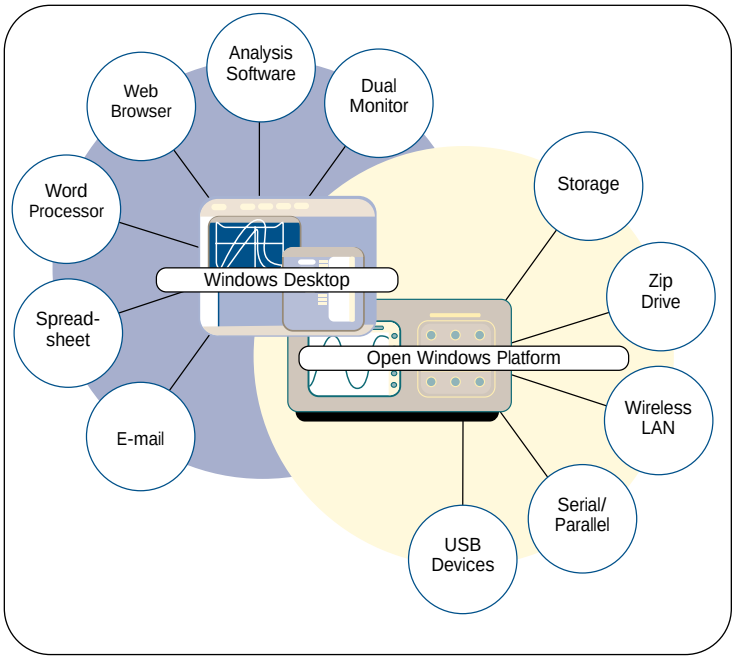


Figure 54. *A TDS7000 Series oscilloscope connects people and equipment to save time and increase total work group productivity.*

Figure 55. *A TDS3000 Series oscilloscope provides a wide array of communications interfaces, such as a standard Centronics port and optional Ethernet/RS-232, GPIB/RS-232, and VGA/RS-232 modules.*

Connectivity

The need to analyze measurement results remains of utmost importance. The need to document and share information and measurement results easily and frequently over high-speed communication networks has also grown in importance.

The connectivity of an oscilloscope delivers advanced analysis capabilities and simplifies the documentation and sharing of results. Standard interfaces (GPIB, RS-232, USB, Ethernet) and network communication modules enable some oscilloscopes to deliver a vast array of functionality and control.

Some advanced oscilloscopes also let you:

- Create, edit and share documents on the oscilloscope – all while working with the instrument in your particular environment
- Access network printing and file sharing resources
- Access the Windows® desktop
- Run third-party analysis and documentation software
- Link to networks
- Access the Internet
- Send and receive e-mail

Figure 56. *The TDSJIT2 optional software package for the TDS7000 Series oscilloscope is specifically designed to meet jitter measurement needs of today's high-speed digital designers.*

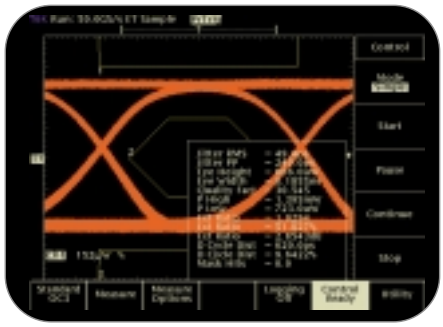


Figure 57. *Equip the TDS700 Series oscilloscope with the TDSCEM1 application module for communications mask compliance testing.*

Figure 58. *The TDS3SDI video module makes the TDS3000 Series oscilloscope a fast, tell-all tool for video troubleshooting.*

Expandability

An oscilloscope should be able to accommodate your needs as they change. Some oscilloscopes allow you to:

- Add memory to channels to analyze longer record lengths
- Add application-specific measurement capabilities
- Complement the power of the oscilloscope with a full range of probes and modules
- Work with popular third-party analysis and productivity Windows-compatible software
- Add accessories, such as battery packs and rackmounts

Application modules and software may enable you to transform your oscilloscope into a highly specialized analysis tool capable of performing functions such as jitter and timing analysis, microprocessor memory system verification, communications standards testing, disk drive measurements, video measurements, power measurements and much more.

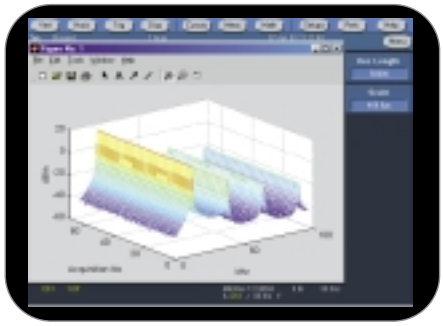


Figure 59. *Advanced analysis and productivity software, such as MATLAB®, can be installed in the TDS7000 Series oscilloscope to accomplish local signal analysis.*

Figure 60. *Traditional, analog-style knobs control position, scale, intensity, etc. – precisely as you would expect.*

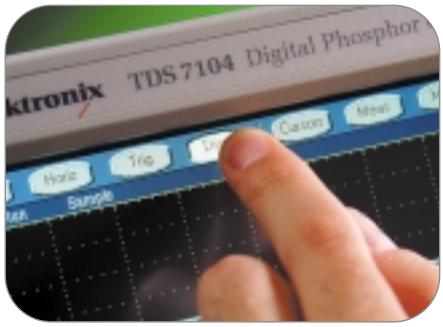


Figure 61. *Touch-sensitive display naturally solves issues with cluttered benches and carts, while providing access to clear, on-screen buttons.*

Figure 62. *Use graphical control windows to access even the most sophisticated functions with confidence and ease.*

Ease-of-Use

Oscilloscopes should be easy to learn and easy to use, helping you work at peak efficiency and productivity. Just as there is no one typical car driver, there is no one typical oscilloscope user. There are both traditional instrument users and those who have grown up in the Windows®/Internet era. The key to satisfying such a broad group of users is flexibility in operating style.

Many oscilloscopes offer a balance between performance and simplicity by providing the user with many ways to operate the instrument. A front-panel layout provides dedicated vertical, horizontal and trigger controls. An icon-rich graphical user interface helps you understand and intuitively use advanced capabilities. Touch-sensitive display solves issues with cluttered benches and carts, while providing access to clear, on-screen buttons. On-line help provides a convenient, built-in reference manual. Intuitive controls allow even occasional oscilloscope users to feel as comfortable driving the oscilloscope as they do driving a car, while giving full-time users easy access to the oscilloscope's most advanced features. In addition, many oscilloscopes are portable, making the oscilloscope efficient in many different operating environments – in the lab or in the field.

Figure 63. *The portability of many oscilloscopes makes the instrument efficient in many operating environments.*

Probes

A probe functions as a critical component of the measurement system, ensuring signal integrity and enabling you to access all of the power and performance in your oscilloscope. Please refer to The Complete Measurement System under the Systems and Controls of the Oscilloscope section, or the Tektronix' *ABCs of Probes* primer, for additional information.

6.2.2 Manufacturers

- Agilent (former HP): http://www.home.agilent.com/agilent/home.jspx
- Tektronix: http://www.tek.com/
- LeCroy: http://www.lecroy.com
- Pico technology: http://www.picotech.com/
- And many many more...

6.3 DAQ-devices

The most expanding generic instrumentation is the DAQ-devices, that enables computers to tap data directly onto files, control hardware and digital signals. These are the modern swiss army knife of instrumentation and are built to be versatile and usable. They are becoming more and more easy to work with and makes the basic set up of measuring systems very easy.

6.4 General build-up

DAQ-cards often contain multiple components (multiplexer, ADC, DAC, TTL-IO, high speed timers, RAM). These are accessible via a bus which connect the computer to a small micro controller that controls the input and output of the DAQ device. The programs of the microcontroller are controlled through a more advanced software interface (C, Labview Matlab etc). The DAQ device can also contain digital signal processors that spend a lot of silicon on arithmetic and allow tight control loops or filters. The connection and control with a PC allows for comfortable data import into a native data treatment interface as well as easy compilation, control and debugging. Using an external housing a modular design with slots in a bus the device can can grow with the needs of the user.

6.4.1 Different types

DAQ hardware is what usually interfaces between the signal and a PC. It could be in the form of modules that can be connected to the computer's ports (parallel, serial, USB, etc...) or cards connected to slots (PCI, ISA) in the mother board.

The disadvantage of using a internal slot is that the space on the back of a PCI card is too small for all the connections needed, so an external breakout box is required. The cable between this Box and the PC is expensive due to the many wires and the required shielding and because it is exotic. The advantage is that it is connected to a high speed bus in the computer, so the bus itself will not impair the speed or timing of the DAQ device. Today the usual interface for slower signals and less demanding situations is USB or ethernet. Oscilloscope-like DAQ-cards are named digitisers, and have generally much higher resolution than old style oscilloscopes. DAQ components with only input features are usually called data loggers.

6.4.2 Common features

Generally the following features can be found

- Input: Generally several analog signals with 10-24 bits resulution and a sampling rate from a few kHz to several MHz. In addition to this often a digital input. The input range can usually be controlled through a programmable gain amplifier (PGA) from 10 V down to 0.01 V.

- Output: Typically two output signals 10-16 bits from hundreds of Hz u to hundreds of KHz. Digital output.
- Synchronisation: For more expensive devices there is usually several sync signals for hardware synchronisation of devices so that they can be run in parallel without being limited by the timing resolution of the bus they are run by.
- Power: Usually there are some contacts for supplying power, like +5V.

6.4.3 Manufacturers

- National instruments: http://www.ni.com
- IOtech: http://www.iotech.com/
- Labjack: http://www.labjack.com/
- Pico Technology: http://www.picotech.com/
- Data translation: http://www.datx.com/
- Eagle technology: http://www.eagledaq.com/
- Data Translation: http://www.datx.com/
- And many many more...

6.5 Lock-in amplifier

Lock-in amplifiers are used to measure the amplitude and phase of signals buried in noise. They achieve this by acting as a narrow bandpass filter which allows only signals at the wanted frequency in phase with the reference frequency to pass. Thus the vast majority of the noise can be rejected during measurements.

The frequency of the signal to be measured and hence the passband region of the filter is set by a reference signal, which has to be supplied to the lock-in amplifier along with the unknown signal. The reference signal must be at the same frequency or a multiple of the frequency of the modulation of the signal to be measured. Tuning the lock-in -amplifier to higher-order frequencies can allow for directly analysing a beyond first order physical phenomena.

A basic lock-in amplifier can be split into 4 stages: an input gain stage, the reference circuit, a demodulator and a low pass filter.

- Input Gain Stage: The variable gain input stage pre-processes the signal by amplifying it to a level suitable for the demodulator. Nothing complicated here, but high performance amplifiers are required.
- Reference Circuit: The reference circuit allows the reference signal to be generated and phase shifted.
- Demodulator: The demodulator is a multiplier. It takes the input signal and the reference and multiplies them together. When you multiply two waveforms together you get the sum and difference frequencies as the result. As the input signal to be measured and the reference signal are of the same frequency, the difference frequency is zero and you get a DC output which is proportional to the amplitude of the input signal and the cosine of the phase difference between the signals. By adjusting the phase of the reference signal using the reference circuit, the phase difference between the input signal and the reference can be brought to zero and hence the DC output level from the multiplier is proportional to the input signal. The noise signals will still be present at the output of the demodulator and may have amplitudes 1000x larger than the DC offset.
- Low Pass Filter: As the various noise components on the input signal are at different frequencies to the reference signal, the sum and difference frequencies will be non zero and will not contribute to the DC level of the output signal. This DC level (which is proportional to the input signal) can now be recovered by passing the output from the demodulator through a low pass filter.

The above gives an idea of how a basic lock-in amplifier works. Actual lock-in amplifiers are more complicated, as there are instrument offsets that need to be removed, but the basic principle of operation is the same.

For further information see the instructive note from one of the manufacturers of lock-in amplifiers.

About Lock-In Amplifiers

Application Note #3

Lock-in amplifiers are used to detect and measure very small AC signals—all the way down to a few nanovolts. Accurate measurements may be made even when the small signal is obscured by noise sources many thousands of times larger. Lock-in amplifiers use a technique known as phase-sensitive detection to single out the component of the signal at a specific reference frequency and phase. Noise signals, at frequencies other than the reference frequency, are rejected and do not affect the measurement.

Why Use a Lock-In?

Let's consider an example. Suppose the signal is a 10 nV sine wave at 10 kHz. Clearly some amplification is required to bring the signal above the noise. A good low-noise amplifier will have about 5 nV/√Hz of input noise. If the amplifier bandwidth is 100 kHz and the gain is 1000, we can expect our output to be 10 μV of signal (10 nV × 1000) and 1.6 mV of broadband noise (5 nV/√Hz × √100 kHz × 1000). We won't have much luck measuring the output signal unless we single out the frequency of interest.

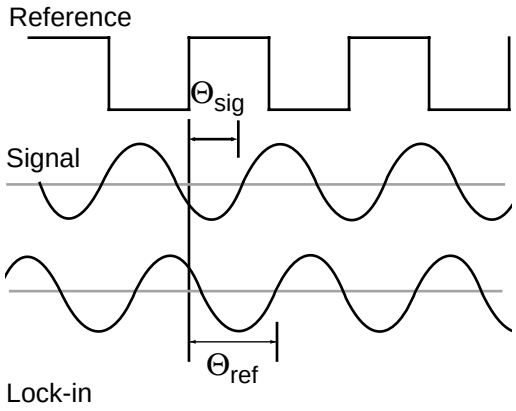
If we follow the amplifier with a band pass filter with a Q=100 (a VERY good filter) centered at 10 kHz, any signal in a 100 Hz bandwidth will be detected (10 kHz/Q). The noise in the filter pass band will be 50 μV (5 nV/√Hz × √100 Hz × 1000), and the signal will still be 10 μV. The output noise is much greater than the signal, and an accurate measurement can not be made. Further gain will not help the signal-to-noise problem.

Now try following the amplifier with a phase-sensitive detector (PSD). The PSD can detect the signal at 10 kHz with a bandwidth as narrow as 0.01 Hz! In this case, the noise in the detection bandwidth will be 0.5 μV (5 nV/√Hz × √.01 Hz × 1000), while the signal is still 10 μV. The signal-to-noise ratio is now 20, and an accurate measurement of the signal is possible.

What is Phase-Sensitive Detection?

Lock-in measurements require a frequency reference. Typically, an experiment is excited at a fixed frequency (from an oscillator or function generator), and the lock-in detects the response from the experiment at the reference frequency. In the following diagram, the reference signal is a square wave at frequency ω_r. This might be the sync output from a function generator. If the sine output from the function generator is used to excite the experiment, the response might be the signal waveform shown below. The signal is $V_{sig}\sin(\omega_r t + \theta_{sig})$ where V_{sig} is the signal amplitude, ω_r is the signal frequency, and θ_{sig} is the signal's phase.

Lock-in amplifiers generate their own internal reference signal usually by a phase-locked-loop locked to the external reference. In the diagram, the external reference, the lock-in's reference, and the signal are all shown. The internal reference is $V_L\sin(\omega_L t + \theta_{ref})$.



The lock-in amplifies the signal and then multiplies it by the lock-in reference using a phase-sensitive detector or multiplier. The output of the PSD is simply the product of two sine waves.

$$V_{psd} = V_{sig}V_L\sin(\omega_r t + \theta_{sig})\sin(\omega_L t + \theta_{ref})$$

$$= \tfrac{1}{2}V_{sig}V_L\cos([\omega_r - \omega_L]t + \theta_{sig} - \theta_{ref}) - \tfrac{1}{2}V_{sig}V_L\cos([\omega_r + \omega_L]t + \theta_{sig} + \theta_{ref})$$

The PSD output is two AC signals, one at the difference frequency ($\omega_r - \omega_L$) and the other at the sum frequency ($\omega_r + \omega_L$).

If the PSD output is passed through a low pass filter, the AC signals are removed. What will be left? In the general case, nothing. However, if ω_r equals ω_L, the difference frequency component will be a DC signal. In this case, the filtered PSD output will be:

$$V_{psd} = \tfrac{1}{2}V_{sig}V_L\cos(\theta_{sig} - \theta_{ref})$$

This is a very nice signal—it is a DC signal proportional to the signal amplitude.

It's important to consider the physical nature of this multiplication and filtering process in different types of lock-ins. In traditional analog lock-ins, the signal and reference are analog voltage signals. The signal and reference are multiplied in an analog multiplier, and the result is filtered with one or more stages of RC filters. In a digital lock-in, such as the SR830 or SR850, the signal and reference are represented by sequences of numbers. Multiplication and filtering are performed mathematically by a digital signal processing (DSP) chip. We'll discuss this in more detail later.

Narrow Band Detection

Let's return to our generic lock-in example. Suppose that instead of being a pure sine wave, the input is made up of signal plus noise. The PSD and low pass filter only detect

signals whose frequencies are very close to the lock-in reference frequency. Noise signals, at frequencies far from the reference, are attenuated at the PSD output by the low pass filter (neither $\omega_{noise} - \omega_{ref}$ nor $\omega_{noise} + \omega_{ref}$ are close to DC). Noise at frequencies very close to the reference frequency will result in very low frequency AC outputs from the PSD ($|\omega_{noise} - \omega_{ref}|$ is small). Their attenuation depends upon the low pass filter bandwidth and rolloff. A narrower bandwidth will remove noise sources very close to the reference frequency; a wider bandwidth allows these signals to pass. The low pass filter bandwidth determines the bandwidth of detection. Only the signal at the reference frequency will result in a true DC output and be unaffected by the low pass filter. This is the signal we want to measure.

Where Does the Lock-In Reference Come From?

We need to make the lock-in reference the same as the signal frequency, i.e. $\omega_r = \omega_L$. Not only do the frequencies have to be the same, the phase between the signals can not change with time. Otherwise, $\cos(\theta_{sig} - \theta_{ref})$ will change and V_{psd} will not be a DC signal. In other words, the lock-in reference needs to be phase-locked to the signal reference.

Lock-in amplifiers use a phase-locked loop (PLL) to generate the reference signal. An external reference signal (in this case, the reference square wave) is provided to the lock-in. The PLL in the lock-in amplifier locks the internal reference oscillator to this external reference, resulting in a reference sine wave at ω_r with a fixed phase shift of θ_{ref}. Since the PLL actively tracks the external reference, changes in the external reference frequency do not affect the measurement.

Internal Reference Sources

In the case just discussed, the reference is provided by the excitation source (the function generator). This is called an external reference source. In many situations the lock-in's internal oscillator may be used instead. The internal oscillator is just like a function generator (with variable sine output and a TTL sync) which is always phase-locked to the reference oscillator.

Magnitude and Phase

Remember that the PSD output is proportional to $V_{sig}\cos\theta$, where $\theta = (\theta_{sig} - \theta_{ref})$. θ is the phase difference between the signal and the lock-in reference oscillator. By adjusting θ_{ref} we can make θ equal to zero. In which case we can measure V_{sig} ($\cos\theta = 1$). Conversely, if θ is 90°, there will be no output at all. A lock-in with a single PSD is called a single-phase lock-in and its output is $V_{sig}\cos\theta$.

This phase dependency can be eliminated by adding a second PSD. If the second PSD multiplies the signal with the reference oscillator shifted by 90°, i.e. $V_L\sin(\omega_L t + \theta_{ref} + 90°)$, its low pass filtered output will be:

$$V_{psd2} = \tfrac{1}{2} V_{sig} V_L \sin(\theta_{sig} - \theta_{ref})$$

$$V_{psd2} \sim V_{sig}\sin\theta$$

Now we have two outputs: one proportional to $\cos\theta$ and the other proportional to $\sin\theta$. If we call the first output X and the second Y,

$$X = V_{sig}\cos\theta \qquad Y = V_{sig}\sin\theta$$

these two quantities represent the signal as a vector relative to the lock-in reference oscillator. X is called the 'in-phase' component and Y the 'quadrature' component. This is because when $\theta = 0$, X measures the signal while Y is zero.

By computing the magnitude (R) of the signal vector, the phase dependency is removed.

$$R = (X^2 + Y^2)^{1/2} = V_{sig}$$

R measures the signal amplitude and does not depend upon the phase between the signal and lock-in reference.

A dual-phase lock-in has two PSDs with reference oscillators 90° apart, and can measure X, Y and R directly. In addition, the phase (θ) between the signal and lock-in is defined as:

$$\theta = \tan^{-1}(Y/X)$$

Digital PSD vs. Analog PSD

We mentioned earlier that the implementation of a PSD is different for analog and digital lock-ins. A digital lock-in, such as the SR830, multiplies the signal with the reference sine waves digitally. The amplified signal is converted to digital form using a 16-bit A/D converter sampling at 256 kHz. The A/D converter is preceeded by a 102 kHz anti-aliasing filter to prevent higher frequency inputs from aliasing below 102 kHz.

This input data stream is multiplied, a point at a time, with the computed reference sine waves described previously. Every 4 µs the input signal is sampled, and the result is multiplied by both reference sine waves (90° apart).

The phase sensitive detectors (PSDs) in the digital lock-in act as linear multipliers; that is, they multiply the signal with a reference sine wave. Analog PSDs (both square wave and linear) have many problems associated with them. The main problems are harmonic rejection, output offsets, limited dynamic reserve, and gain error.

The digital PSD multiplies the digitized signal with a digitally computed reference sine wave. Because the reference sine waves are computed to 20 bits of accuracy, they have very low harmonic content. In fact, the harmonics are at the –120 dB level! This means that the signal is multiplied by a single reference sine wave (instead of a reference and its many harmonics), and only the signal at this single reference frequency is detected. The SR810, SR830 and SR850 digital lock-ins are completely insensitive to signals at harmonics of the reference. In contrast, a square wave multiplying lock-in will detect at all of the odd harmonics of the reference (a square wave contains many large odd harmonics).

Output offset is a problem because the signal of interest is a DC output from the PSD, and an output offset contributes to error and zero drift. The offset problems of analog PSDs are eliminated using the digital multiplier. There are no erroneous DC output offsets from the digital multiplication of the signal and reference. In fact, the actual multiplication is virtually error free.

The dynamic reserve of an analog PSD is limited to about 60 dB. When there is a large noise signal present, 1000 times (or 60 dB) greater than the full-scale signal, the analog PSD measures the signal with an error. The error is caused by non-linearity in the multiplication (the error at the output depends upon the amplitude of the input). This error can be quite large (10 % of full scale) and depends upon the noise amplitude, frequency and waveform. Since noise generally varies quite a bit in these parameters, the PSD error causes a lot of output uncertainty.

In the digital lock-in, dynamic reserve is limited by the quality of the A/D conversion. Once the input signal is digitized, no further errors are introduced. Certainly, the accuracy of the multiplication does not depend on the size of the numbers. The A/D converter used in the SR810, SR830 and SR850 is extremely linear, meaning that the presence of large noise signals does not impair its ability to correctly digitize a small signal. In fact, the dynamic reserve of these lock-ins can exceed 100 dB without any problems. We'll talk more about dynamic reserve a little later.

A linear, analog PSD multiplies the signal by an analog reference sine wave. Any amplitude variation in the reference amplitude shows up directly as a variation in the overall gain. Analog sine-wave generators are susceptible to amplitude drift: especially as a function of temperature. The digital reference sine wave has a precise amplitude and never changes. This avoids a major source of gain error common to analog lock-ins.

The overall performance of a lock-in amplifier is largely determined by the performance of its phase sensitive detectors. In virtually all respects, the digital PSD outperforms its analog counterparts.

What Does a Lock-In Measure?

So what exactly does the lock-in measure? Fourier's theorem basically states that any input signal can be represented as the sum of many sine waves of differing amplitudes, frequencies and phases. This is generally considered as representing the signal in the "frequency domain". Normal oscilloscopes display the signal in the "time domain". Except in the case of clean sine waves, the time domain representation does not convey very much information about the various frequencies which make up the signal.

A lock-in multiplies the signal by a pure sine wave at the reference frequency. All components of the input signal are multiplied by the reference simultaneously. Mathematically speaking, sine waves of differing frequencies are orthogonal, i.e. the average of the product of two sine waves is zero unless the frequencies are EXACTLY the same. The product of this multiplication yields a DC output signal proportional to the component of the signal whose frequency is exactly locked to the reference frequency. The low pass filter (which follows the multiplier) provides the averaging which removes the products of the reference with components at all other frequencies.

A lock-in amplifier, because it multiplies the signal with a pure sine wave, measures the single Fourier (sine) component of the signal at the reference frequency. Let's take a look at an example. Suppose the input signal is a simple square wave at frequency f. The square wave is actually composed of many sine waves at multiples of f with carefully related amplitudes and phases. A 2 Vpp square wave can be expressed as:

$$S(t) = 1.273\sin(\omega t) + 0.4244\sin(3\omega t) + 0.2546\sin(5\omega t) + ...$$

where $\omega = 2\pi f$. The lock-in, locked to f, will single out the first component. The measured signal will be $1.273\sin(\omega t)$, not the 2 Vpp that you'd measure on a scope.

In the general case, the input consists of signal plus noise. Noise is represented as varying signals at all frequencies. The ideal lock-in only responds to noise at the reference frequency. Noise at other frequencies is removed by the low pass filter following the multiplier. This "bandwidth narrowing" is the primary advantage that a lock-in amplifier provides. Only inputs with frequencies at the reference frequency result in an output.

RMS or Peak?

Lock-in amplifiers, as a general rule, display the input signal in volts rms. When a lock-in displays a magnitude of 1 V (rms), the component of the input signal (at the reference frequency) is a sine wave with an amplitude of 1 Vrms, or 2.8 Vpp.

Thus, in the previous example with a 2 Vpp square wave input, the lock-in would detect the first sine component, $1.273\sin(\omega t)$. The measured and displayed magnitude would be 0.90 Vrms (or $1.273/\sqrt{2}$).

Degrees or Radians?

In this discussion, frequencies have been referred to as f (Hz) and ω ($2\pi f$ radians/s). This is because people measure frequencies in cycles per second, and math works best in radians. For purposes of measurement, frequencies as measured in a lock-in amplifier are in Hz. The equations used to explain the actual calculations are sometimes written using ω to simplify the expressions.

Phase is always reported in degrees. Once again, this is more by custom than by choice. Equations written as $\sin(\omega t + \theta)$ are written as if θ is in radians, mostly for simplicity. Lock-in amplifiers always manipulate and measure phase in degrees.

Dynamic Reserve

The term "dynamic reserve" comes up frequently in discussions about lock-in amplifiers. It’s time to discuss this

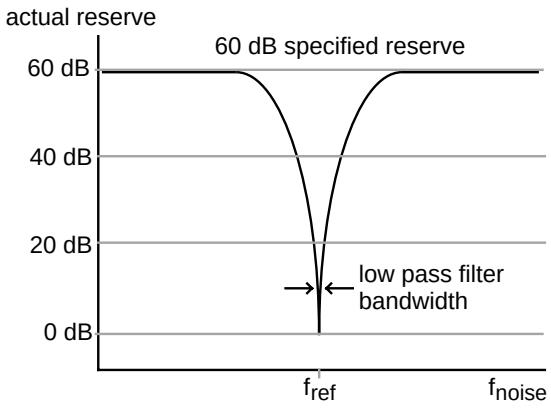
term in a little more detail. Assume the lock-in input consists of a full-scale signal at f_{ref} plus noise at some other frequency. The traditional definition of dynamic reserve is the ratio of the largest tolerable noise signal to the full-scale signal, expressed in dB. For example, if full scale is 1 μV, then a dynamic reserve of 60 dB means noise as large as 1 mV (60 dB greater than full scale) can be tolerated at the input without overload.

The problem with this definition is the word "tolerable". Clearly, the noise at the dynamic reserve limit should not cause an overload anywhere in the instrument—not in the input signal amplifier, PSD, low pass filter or DC amplifier. This is accomplished by adjusting the distribution of the gain. To achieve high reserve, the input signal gain is set very low so the noise is not likely to overload. This means that the signal at the PSD is also very small. The low pass filter removes the large noise components from the PSD output which allows the remaining DC component to be amplified (a lot) to reach 10 V full scale. There is no problem running the input amplifier at low gain. However, as we have discussed previously, analog lock-ins have a problem with high reserve because of the linearity of the PSD and the DC offsets of the PSD and DC amplifier. In an analog lock-in, large noise signals almost always disturb the measurement in some way.

The most common problem is a DC output error caused by the noise signal. This can appear as an offset or as a gain error. Since both effects are dependent upon the noise amplitude and frequency, they can not be offset to zero in all cases and will limit the measurement accuracy. Because the errors are DC in nature, increasing the time constant does not help. Most lock-ins define tolerable noise as levels which do not affect the output more than a few percent of full scale. This is more severe than simply not overloading.

Another effect of high dynamic reserve is to generate noise and drift at the output. This comes about because the DC output amplifier is running at very high gain, and low-frequency noise and offset drift at the PSD output or the DC amplifier input will be amplified and appear large at the output. The noise is more tolerable than the DC drift errors since increasing the time constant will attenuate the noise. The DC drift in an analog lock-in is usually on the order of 1000 ppm/°C when using 60 dB of dynamic reserve. This means that the zero point moves 1 % of full scale over 10 °C temperature change. This is generally considered the limit of tolerable.

Lastly, dynamic reserve depends on the noise frequency. Clearly noise at the reference frequency will make its way to the output without attenuation. So the dynamic reserve at f_{ref} is 0 dB. As the noise frequency moves away from the reference frequency, the dynamic reserve increases. Why? Because the low pass filter after the PSD attenuates the noise components. Remember, the PSD outputs are at a frequency of $|f_{noise} - f_{ref}|$. The rate at which the reserve increases depends upon the low pass filter time constant and rolloff. The reserve increases at the rate at which the filter rolls off. This is why 24 dB/oct filters are better than 6 or 12 dB/oct filters. When the noise frequency is far away, the reserve is limited by the gain distribution and overload level of each gain element. This reserve level is the dynamic reserve referred to in the specifications.



The above graph shows the actual reserve vs. the frequency of the noise. In some instruments, the signal input attenuates frequencies far outside the lock-in's operating range (f_{noise}>>100 kHz). In these cases, the reserve can be higher at these frequencies than within the operating range. While this creates a nice specification, removing noise at frequencies very far from the reference does not require a lock-in amplifier. Lock-ins are used when there is noise at frequencies near the signal. Thus, the dynamic reserve for noise within the operating range is more important.

Dynamic Reserve in Digital Lock-Ins

The SR810, SR830 and SR850, with their digital phase sensitive detectors, do not suffer from DC output errors caused by large noise signals. The dynamic reserve can be increased to above 100 dB without measurement error. Large noise signals do not cause output errors from the PSD. The large DC gain does not result in increased output drift.

In fact, the only drawback to using ultra-high dynamic reserves (>60 dB) is the increased output noise due to the noise of the A/D converter. This increase in output noise is only present when the dynamic reserve is increased above 60 dB *and* above the minimum reserve. (If the minimum reserve is 80 dB, then increasing to 90 dB may increase the noise. As we'll discuss next, the minimum reserve does not have increased output noise: no matter how large it is.)

To set a scale, the digital lock-in's output noise at 100 dB dynamic reserve is only measurable when the signal input is grounded. Let's do a simple experiment. If the lock-in reference is at 1 kHz, and a large signal is applied at 9.5 kHz, what will the lock-in output be? If the signal is increased to the dynamic reserve limit (100 dB greater than full scale), the output will reflect the noise of the signal at 1 kHz. The spectrum of any pure sine generator always has a noise floor, i.e. there is some noise at all frequencies. So even though the

applied signal is at 9.5 kHz, there will be noise at all other frequencies, including the 1 kHz lock-in reference. This noise will be detected by the lock-in and appear as noise at the output. This output noise will typically be greater than the lock-in's own output noise. In fact, virtually all signal sources will have a noise floor which will dominate the lock-in output noise. Of course, noise signals are generally much noisier than pure sine generators and will have much higher broadband noise floors.

If the noise does not reach the reserve limit, the digital lock-in's own output noise may become detectable at ultra-high reserves. In this case, simply lower the dynamic reserve and the DC gain will decrease, and the output noise will decrease also. In general, do not run with more reserve than necessary. Certainly don't use ultra-high reserve when there is virtually no noise at all.

The frequency dependence of dynamic reserve is inherent in the lock-in detection technique. The SR810, SR830 and SR850, by providing more low-pass filter stages, can increase the dynamic reserve close to the reference frequency. The specified reserve applies to noise signals within the operating range of the lock-in, i.e. frequencies below 100 kHz. The reserve at higher frequencies is actually greater but is generally not that useful.

Minimum Dynamic Reserve

The SR810, SR830 and SR850 always have a minimum amount of dynamic reserve. This minimum reserve changes with the sensitivity (gain) of the instrument. At high gains (full-scale sensitivity of 50 µV and below), the minimum dynamic reserve increases from 37 dB at the same rate as the sensitivity increases. For example, the minimum reserve at 5 µV sensitivity is 57 dB. In many analog lock-ins, the reserve can be lower. Why can't the digital lock-ins run with lower reserve at this sensitivity?

The answer to this question is: “Why would you want lower reserve?” In an analog lock-in, lower reserve means less output error and drift. In the SR800 series lock-ins, more reserve does not increase the output error or drift. But, more reserve can increase the output noise. However, if the analog signal gain before the A/D converter is high enough, the 5 nV/√Hz noise of the signal input will be amplified to a level greater than the input noise of the A/D converter. At this point, the detected noise will reflect the actual noise at the signal input and not the A/D converter's noise. Increasing the analog gain (decreasing the reserve) will not decrease the output noise. Thus, there is no reason to decrease the reserve. At a sensitivity of 5 µV, the analog gain is sufficiently high so that A/D converter noise is not a problem. Sensitivities below 5 µV do not require any more gain since the signal-to-noise ratio will not be improved (the front-end noise dominates). The SR800 series lock-ins do not increase their gain below the 5 µV sensitivity. Instead, the minimum reserve increases. Of course, the input gain can be decreased and the reserve increased; in which case, the A/D converter noise might be detected in the absence of any signal input.

Dynamic Reserve in Analog Lock-Ins

Because of the limitations of their PSDs, analog lock-in amplifiers must use different techniques to improve their dynamic reserve. The most common of these is the use of analog prefilters. The SR510 and SR530 have tunable, band-pass filters at their inputs. The filters are designed to automatically track the reference frequency. If an interfering signal is attenuated by a filter before it reaches the lock-in input, the dynamic reserve of the lock-in will be increased by that amount. For the SR510 and SR530, a dynamic reserve increase of up to 20 dB can be realized using the input band pass filter. Of course, such filters add their own noise and contribute to phase error: so they should only be used if necessary.

A lock-in can measure signals as small as a few nanovolts. A low-noise signal amplifier is required to boost the signal to a level where the A/D converter can digitize the signal without degrading the signal-to-noise. The analog gain in the SR850 ranges from roughly 7 to 1000. As discussed previously, higher gains do not improve signal-to-noise and are not necessary.

The overall gain (AC and DC) is determined by the sensitivity. The distribution of the gain (AC versus DC) is set by the dynamic reserve.

Input Noise

The input noise of the SR810, SR830 or SR850 signal amplifier is about 5 nVrms/√Hz. The SR530 and SR510 lock-ins have 7 nVrms/√Hz of input noise. What does this noise figure mean? Let's set up an experiment. If an amplifier has 5 nVrms/√Hz of input noise and a gain of 1000, then the output will have 5 µVrms/√Hz of noise. Suppose the amplifier output is low-pass filtered with a single RC filter (6 dB/oct rolloff) with a time constant of 100 ms. What will be the noise at the filter output?

Amplifier input noise and Johnson noise of resistors are Gaussian in nature. That is, the amount of noise is proportional to the square root of the bandwidth in which the noise is measured. A single stage RC filter has an equivalent noise bandwidth (ENBW) of 1/4T, where T is the time constant (R × C). This means that Gaussian noise at the filter input is filtered with an effective bandwidth equal to the ENBW. In this example, the filter sees 5 µVrms/√Hz of noise at its input. It has an ENBW of 1/(4 × 100 ms) or 2.5 Hz. The voltage noise at the filter output will be 5 µVrms/√Hz × √2.5 Hz, or 7.9 µVrms. For Gaussian noise, the peak-to-peak noise is about 5 times the rms noise. Thus, the output will have about 40 µVpp of noise.

Input noise for a lock-in works the same way. For sensitivities below about 5 µV full scale, the input noise will determine the output noise (at minimum reserve). The amount of noise at the output is determined by the ENBW of the low pass filter. The ENBW depends upon the time constant and filter rolloff. For example, suppose the lock-in is set to 5 µV full scale, with a 100 ms time constant, and 6 dB/oct of filter rolloff. The lock-in

6.5.1 Manufacturers

- Stanford Research Systems: http://www.thinksrs.com/
- Signal recovery: http://www.signalrecovery.com/
- Boston electronics: http://www.boselec.com/
- Scitech: http://www.scitec.uk.com/

6.6 Repetition questions

1. What us the typical measurements that a multimeter can perform?
2. What is the typical resolution of a cheap multimeter?
3. What is an oscilloscope?
4. What are the differences between computer based and portable oscilloscopes?
5. What are the four different types of oscilloscopes?
6. What ere the great advantage of a DPO?
7. What oscilloscopes do you expect FFT functionality from?
8. What is the basic criteria you should set on the bandwidth of an oscilloscope?
9. How does the rise time connect with the bandwidth?
10. What sample rate would you select for your oscilloscope?
11. When is a high capture rate needed in an oscilloscope?
12. What would you expect from a general DAQ device?
13. Why would you use a lock-in amplidier?
14. Explain the working principle for a lock in amplifier.

Chapter 7

Basics of measurement systems

Measurement systems are usually built from many components, the most common ones will be described in chapter 8 on generic instruments. This chapter will briefly introduce you to the current standards for control of instruments.

7.1 Measurement systems - basic components

7.1.1 Real-time versus data logging

One important consideration when building a measurements system is whether you need a data logger or a real-time system. Usually real-time systems are only needed in control situations. Data loggers are less expensive, and in industrial systems, there are a number of systems for automatic data logging in distributed measurement systems. Of these more and more are working through wire-less connections.

7.1.2 Isolated instruments

Today there are fewer and fewer instruments apart from handheld that does not contain any buses. However, isolated instruments have the advantage of being battery-operated and often not connected to any external potential. This is important when working with high potentials or sensitive systems.

7.1.3 Computer controlled acquisition

Most serious data acquisition is made utilising computers. This can be done using a standard computer with PC data acquisition and/or instruments connected through buses, or through dedicated computer systems.

The most common way to do this is through a windows based PC, which is both quite cheap and offers means for data treatment compatible with the rest of the analysis chain. However, a windows computer is not deterministic, there is no way to know when the computer will perform a task. This makes it awkward to work with in demanding real-time measurement situations. The PC i not a very good measurement system by itself but must be interfaced either by connecting a DAQ device, or by connecting instruments to it using ordinary communication buses or dedicated instrument buses.

PC Data acquisition (DAQ) cards

DAQ cards are installed in internal I/O slots in the PC or connected directly to the PC through the USB port. A DAQ card is designed to acquire analog signals on one or many channels of input. It can usually also generate analog signals on one or more output lines, and read and write digital data.

Most modern high speed cards use the PCI bus. The Peripheral Component Interconnect standard (in practise almost always shortened to PCI) specifies a computer bus for attaching peripheral devices to a computer motherboard. These devices can take the form of either integrated circuits fitted on the motherboard itself or expansion cards that fit in sockets.

The PCI bus is common in modern PCs, where it has displaced ISA and VESA Local Bus as the standard expansion bus, but it also appears in many other computer types. As a development of the PCI bus a faster version has been developed. PCI-express. It was introduced on 2004 by Intel, and has now reached the market in full. It is primarily used when high data-rates are needed. In PCI- express, the data transfer is divided up in lanes, where each slot can carry up to 32 lanes (thus this is point to pint connections with full duplex). The maximum transfer rate per lane is determined by the standard number, currently (PCIe 1.1) it is 250 MByte/s. But it doubles with each standard number (so PCIe 3 will carry 1GByte/s.

VXI, PXI and LXI buses

To extend the computer interface beyond the computer cabinet, and the PCI interface. To allow for better noise reduction, more measurement modules or instruments to be connected through instrument buses extends buses has been defined. Apart from offering a fast route of data, they often offer internal buses for data and fast triggering in extension of the PCI bus. There are four main standards for instrument communication: GPIB, VXI, PXI and LXI all which are described later in this chapter.

7.1.4 Deterministic (Real-time) computer

Computers are gaining in speed and ability by the second and standard PC:s or their robust industrial cousins are getting more and more useful for real-time control and measurement. Accordingly most slower control that used to be done by micro-controllers can be performed by computers today. The cost is higher, but the user-friendliness is staggering compared to low-level programming of a micro controllers. Thus for unconventional single assignments, there will in the future be a twist towards more complex controllers that can be programmed through high-level programming languages which allow for full modularity for all possible measuring situations.

For an ordinary PC to work in a deterministic fashion it can not be run under windows. Thus for these systems real-time operating systems are necessary.

Micro-controllers

A micro-controller (MCU) is a computer-on-a-chip used to control electronic devices. It is a type of microprocessor emphasising self-sufficiency and cost-effectiveness, in contrast to a general-purpose processor, the kind used in a PC. A typical micro-controller contains all the memory, peripherals and input/output interfaces needed, whereas a general purpose microprocessor requires additional chips to provide these functions.

A micro-controller is quite easy to make into a working computer, with a minimum of external support chips. The idea is that the micro-controller is placed in

the device to be controlled, hooked up to power and any information it needs. This makes them ideal for some measurement applications.

A traditional microprocessor does not allow you to do this. It requires all of these additional tasks to be handled by other chips. For example, a number of RAM or Flash memory chips must be added. The amount of memory provided is more flexible in the traditional approach, but at least a few external memory chips must be provided, which requires numerous connections to pass the data back and forth to them.

Micro controllers trade away speed and flexibility to gain ease of equipment design and low cost. There is only so much room on the chip to include functionality, so for every I/O device or memory increase the micro-controller includes, some other circuitry has to be removed.

Digital signal processing (DSP) boards

A digital signal processor (DSP) is a specialised microprocessor designed specifically for digital signal processing, generally in real-time. The commands for and the architecture of the processor itself is optimised for fast digital signal processing to solve problems in real-time. Often boards can be bought with DAC and ADC:s to enable development of prototype systems.

Field programmable gate array (FPGA)

A field-programmable gate array or FPGA is a semiconductor device containing programmable logic components and programmable interconnects. The programmable logic components can be programmed to duplicate the functionality of basic logic gates (such as AND, OR, XOR, NOT) or more complex combinatorial functions such as decoders or simple math functions. In most FPGAs, these programmable logic components also include memory elements, which may be simple flip-flops or more complete blocks of memories.

As logic generally can perform operations much faster than computers. FPGA:s are useful for fast real-time systems, where signals can be acquired and processed in a single step. However they are currently not very convenient to program and they are still quite expensive.

7.2 Analogue communication

To use analogue communication in-between instruments is rather common. The manner of communication depends on the resolution and speed needed. The following types of lines are available:

- Single wire lines. Satisfactory for low frequency communication <400Hz.
- Open pairs.
- Twisted pairs (shielded or unshielded) Not used for high frequencies, but for reducing magnetic cross talk. However many pairs nearby will suffer from crosstalk unless the sum of the current flowing through a pair is zero (balanced).
- Multipair. A bunch of twisted pairs working in a balanced mode, where each current sent through one part of a pair is sent in the other direction in the other part of the pair. Separately shielded to avoid stray capacitive coupling.
- Coaxial cables (single shield or multi-shield). Minimises radiation losses, and can accordingly carry higher frequency signals. The generated fields are confined in space, especially with balanced signals.

To reduce noise it is important to work with a balanced signal. Balanced signals are sent with reverse phase on each conductor (there the return path is not ground). Less stray fields are generated improving decreasing noise and cross-talk, however this is generally more expensive and not necessary for low bandwidth systems.

Wires that are to be connected have to be in a good configuration, they have to be terminated in the correct manner. For most signal communication that implies a high impedance. It is worthwhile to remember that the 50Ω marking on a cable has little to do with the resistance but everything to do with the impedance, that is how a wave is transported along a cable, the actual resistance is not 50Ω per unit length but much much lower.

7.2.1 Standard voltages/currents

Most free standing instruments today can accept input voltages to approx. 220 V if used for ordinary household measurements, while more delicate instruments have selectable input ranges from ±10 V down to ±.1 V.

Much of todays electronics is designed for low voltage to decrease power consumption. Thus the old standard of ±15 volt as a supply and ±10 V as signal levels is not always compatible with modern components. Today, 0 to 5 volts, 0 to 3.3 and 0 to 2.2 are common drive ranges and also input ranges of many circuits.

7.2.2 Current loops

One industry standard that is used within signal communication is the 4-20 mA current standard, where the signal is sent in form of a current. Analog current loops are used for any purpose where a device needs to be either monitored or controlled remotely over a pair of conductors. The "live zero" at 4 mA allows the receiving instrumentation to distinguish between a zero signal and a broken wire or a dead instrument. This standard was developed in the 1950s and is still widely used in industry.

Benefits of the 4-20 mA convention are that it is widely used by many manufacturers, relatively low-cost to implement, and it can reject many forms of electrical noise. The live zero also allows low-power instruments to be directly powered from the loop, saving the cost of extra wires. Current loop is also much easier to understand and debug than more complicated digital fieldbuses. Using fieldbuses and solving related problems usually requires much more education and understanding than required by simple current loop solutions.

MIDI (Musical Instrument Digital Interface) and fire wire are a digital current loop interfaces.

7.3 Digital communication

7.3.1 Computer bus

In computer architecture, a bus is a subsystem that transfers data or power between computer components inside a computer or between computers. Unlike a point-to-point connection, a bus can logically connect several peripherals over the same set of wires. Early computer buses were literally parallel electrical buses with multiple connections (i.e. GPIB), but the term is now used for any physical arrangement that provides the same logical functionality as a parallel electrical bus. Modern computer buses can use both parallel and bit-serial connections, and can be wired in either a multidrop (electrical parallel) or daisy chain topology, or connected by switched hubs, as in the case of USB. Today much communication is also performed through Ethernet, which is not a direct bus but offers connectivity and speeds that

	Ethernet	USB	IEEE-1394	RS485
Number of devices	Many	127	63	32
Max distance	100m	100 feet	72m	4000 feet
Max rate	1000Mbps	480Mbps	400Mbps	10Mbps

Table 7.1: Overview of the fastest versions of computer communications right now.

compares to the fastest buses (however timing is an issue at Ethernet). The most common buses are collected in table 7.1.

7.3.2 Serial and Parallel

Serial data is any data that is sent one bit at a time using a single electrical signal. In contrast, parallel data is sent 8, 16, 32, or even 64 bits at a time using a signal line for each bit. Data that is sent without the use of a master clock is said to be asynchronous serial data.

Buses can be parallel buses, which carry data words striped across multiple wires, or serial buses, which carry data in bit-serial form. The addition of extra power and control connections, differential drivers, and data connections in each direction usually means that most serial buses have more conductors than the minimum of two. As data rates increase, the problems of timing skew and crosstalk across parallel buses become more and more difficult to circumvent.

Often, a serial bus can actually be operated at higher overall data rates than a parallel bus, despite having fewer electrical connections, because a serial bus inherently has no timing skew or crosstalk. USB, firewire, and Serial ATA are examples of this. Multidrop connections do not work well for fast serial buses, so most modern serial buses use daisy-chain or hub designs. In a multidrop connection, one primary computer communicates simultaneously with multiple secondary computers/devices.

7.3.3 Buffer

In telecommunication, a buffer is a routine or storage medium used to compensate for a difference in rate of flow of data, or time of occurrence of events, when transferring data from one device to another. Buffers are used for many purposes, such as (a) interconnecting two digital circuits operating at different rates, (b) holding data for use at a later time, (c) allowing timing corrections to be made on a data stream, (d) collecting binary data bits into groups that can then be operated on as a unit, (e) delaying the transit time of a signal in order to allow other operations to occur.

In computing, a buffer is a region of memory used to temporarily hold output or input data. The data can be output to or input from devices outside the computer or processes within a computer. Buffers can be implemented in either hardware or software, but the vast majority of buffers are implemented in software. Buffers are used when there is a difference between the rate at which data is received and the rate at which it can be processed, or in the case that these rates are variable, for example in a printer spooler.

7.3.4 Transistor-Transistor Logic (TTL)

TTL became popular with electronic systems designers in 1962 after Texas Instruments introduced the 7400 series of ICs, which had a wide range of digital logic

block functions. TTL became important because it was the first time that low-cost integrated circuits made digital techniques economically practical for tasks previously done by analog methods.

Today TTL is simply a standard for easily transmitting digital trigger signals. The high level is at 5 Volts while the low is at 0 V. It is still used quite a lot, and sometimes you will need to interface such a port. Then you will often need a pull-up resistance to translate an open/closed current state to a high/low voltage state.

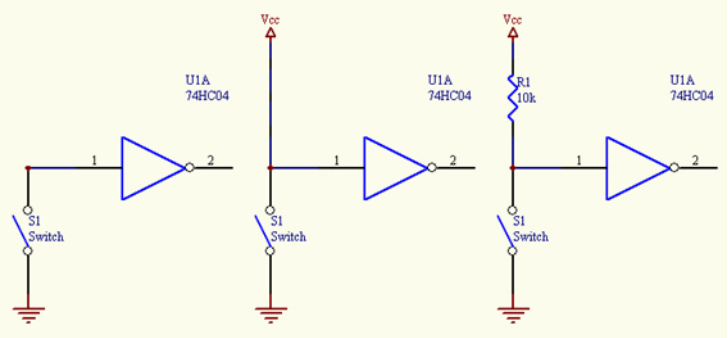


Figure 7.1: The reason why pull-up resistors are used, without any pull-up resistance the potential at the input is undefined.

Consider this left most schematic in fig. 7.1. The gate U1A has an input (pin 1) and an output (pin 2). The input state of most logic gates is called a high impedance. This means it provides no real power of its own. Therefore, if nothing is connected to pin 1, the value of the input is considered to be floating. Most gates will float towards a high state. This is a very weak condition, and any electrical noise could cause the input to go low.

When switch S1 is closed (on), the input state at pin 1 is connected to ground - the state of the pin is stable. When switch S1 is open (off), then input pin 1 is susceptible to a wide array of electrical problems. What is needed here is a way to connect pin 1 to an electrical potential that can be removed when the switch is closed. This electrical potential will allow the pin to keep a steady state.

One thought is to tie the pin to Vcc (+5 volts) to insure that pin 1 doesn't float (middle schematic fig. 7.1). The circuit to the right certainly does that. With pin 1 tied directly to Vcc, the line does not float, and has an ON state. The problem with this circuit is what happens when switch S1 is closed. This creates a direct electrical connection between Vcc and GND. In other words, it will short out the circuit. If you are lucky, it will just stop your entire system from working. If you are unlucky,it will burn up the wires!

Now consider the rightmost schematic in fig. 7.1, which is similar to the first but has added a pull-up resistor. This resistors function is to limit the amount of current that can flow through the circuit. When switch S1 is open (off), pin 1 is tied to Vcc through the resistor. Since pin1 is a high impedance input, a voltage meter or logic probe placed on pin 1 will show Vcc (+5v) if connected to pin 1. When switch S1 is closed (on), pin 1 has a direct connection to GND, which takes it to the low state. The pin1 side of R1 also has direct connection to ground. Current will flow from Vcc, through R1, and to ground. It isn't considered a short, however, because R1 will limit the amount of current that can flow to a very small amount. In fact, you can compute this using Ohms law:

$$I = \frac{U}{R} = \frac{5}{10000} = 0.5mA.$$

Which is a very modest current that will not destroy your equipment.

7.4 Parallel communication

7.4.1 Timing Skew

Timing skew is a problem that can occur on many kinds of computer buses. When signals are transmitted down parallel paths, they will not arrive at exactly the same time due to unavoidable variations in wire transmission properties and transistor sizing, but the signals will arrive close to each other in time. As the frequencies of these circuits increases, this variation will become more and more erratic. If the timing skew is large enough, the clock signal may arrive while the data signal is still transitioning between the previous and current values. If this happens, it will be impossible to determine what value was transmitted from the detected value, resulting in a functional error.

7.4.2 Daisychain

Within computer engineering a daisy chain is a bus wiring scheme in which, for example, device A is wired to device B, device B is wired to device C, device C to device D etc. The first and last devices are normally wired to a resistor called a terminator. All devices may receive identical signals or, in contrast to a simple bus, each device in the chain may modify one or more signals before passing them on.

Daisy chain topologies have the advantage of simplicity in protocols, as each node only needs to know whether the information is relevant to it, or which way to forward it. The disadvantage is that each node introduces a delay in the signal, so long chains have a relatively high latency compared to broadcast topologies. A damaged or crashed node is also likely to partition the network, unlike wireless broadcast or star topologies.

7.4.3 GPIB (IEEE-488)

The GPIB bus has been along for many years and today many labs have a number of GPIB instruments. The interface is today not dominating, but it was 10 years ago, which means that you will most likely encounter it. However we will not cover the details of the standard in this course. With modern programming most instrument builders have ready drivers for the instruments, and hopefully you will not have to deal with the low-level driver writing (and if you do, it is just a matter of writing the correct commands).

The original GPIB was developed in the late 1960s by Hewlett-Packard (where it is called the HP-IB) to connect and control programmable instruments that Hewlett-Packard manufactured. With the introduction of digital controllers and programmable test equipment, the need arose for a standard, high-speed interface for communication between instruments and controllers from various vendors. In 1975, the Institute of Electrical and Electronic Engineers (IEEE) published ANSI/IEEE Standard 488-1975, IEEE Standard Digital Interface for Programmable Instrumentation. This bus is now used worldwide and is known by three names:

- General Purpose Interface Bus (GPIB)
- Hewlett-Packard Interface Bus (HP-IB)
- IEEE 488 Bus,

In 1990, the IEEE 488.2 specification included the Standard Commands for Programmable Instrumentation (SCPI) document. SCPI defines specific commands that each instrument class (which usually includes instruments from various vendors) must obey. Thus, SCPI guarantees complete system compatibility and configurability among these instruments. It is no longer necessary to learn a different command set for each instrument in an SCPI-compliant system, and it is easy to replace an instrument from one vendor with an instrument from another.

IEEE-488 allows up to 15 intelligent devices to share a single bus by daisy-chaining, with the slowest device participating in the control and data transfer handshakes to determine the speed of the transaction.

- You can link devices in either a linear, star or combination configuration using a shielded 24-conductor cable.
- One byte (8 bit) digital information is sent in parallel each time.
- The maximum data rate is about one megabyte per second.
- The IEEE-488 bus (cable) specifies a maximum total cable length of 20 meters.
- A maximum separation of 4 meters between devices and an average separation of 2 meters over the full bus should be followed. Bus extenders and expanders are available to overcome these system limits.

7.5 Serial communication

In theory a serial link would only need two wires, a signal line and a ground, to move the serial signal from one location to another. But in practise this doesn't really work well at higher frequencies, some bits might get lost in the signal and thus altering the ending result. If one bit is missing at the receiving end, all succeeding bits are shifted resulting in incorrect data when converted back to a parallel signal. So to establish reliable serial communications you must overcome these bit errors that can emerge in many different forms.

There are two basic types of serial communications, synchronous and asynchronous. With synchronous communications, the two devices initially synchronise themselves to each other through a timing device, and then continually send characters to stay in sync. Even when data is not really being sent, a constant flow of bits allows each device to know where the other is at any given time. That is, each character that is sent is either actual data or an idle character. Synchronous communications allows faster data transfer rates than asynchronous methods, because additional bits to mark the beginning and end of each data byte are not required.

Asynchronous means "no synchronisation", and thus does not require sending and receiving idle characters for timing. By introducing a start bit which indicates the start of a short data stream, the position of each bit can be determined by timing the bits at regular intervals, by sending start bits in front of each 8 bit streams, the two systems don't have to be synchronised by a clock signal, the only important issue is that both systems must be set at the same port speed.When the receiving end of the communication receives the start bit it starts a short term timer. By keeping streams short, there's not enough time for for the timer to get out of sync. This method is known as asynchronous communication because the sending and receiving end of the communication are not precisely synchronised by the means of a signal line. The serial ports on IBM-style PCs are asynchronous devices and therefore only support asynchronous serial communications.

7.5.1 Parity, bits and stop bits

Parity is a simple form of error checking used in serial communication. There are four types of parity – even, odd, marked (1), and spaced (0). You also can use no parity. For even and odd parity, the serial port sets the parity bit (the last bit after the data bits) to a value to ensure that the transmission has an even or odd number of logic-high bits. For example, if the data is 011, for even parity, the parity bit is 0 to keep the number of logic-high bits even. If the parity is odd, the parity bit is 1, resulting in 3 logic-high bits. Marked and spaced parity does not actually check the data bits but simply sets the parity bit high for marked parity or low for spaced parity. This allows the receiving device to know the state of a bit so the device can determine if noise is corrupting the data or if the transmitting and receiving device clocks are out of sync.

Data bits are a measurement of the actual data bits in a transmission. When the computer sends a packet of information, the amount of actual data may not be a full 8 bits. If the data you are transferring is simple text (standard ASCII), sending 7 bits of data per packet is sufficient for communication. By convention, the least significant bit of the word is sent first and the most significant bit is sent last. A data frame refers to a single byte transfer, including start/stop bits, data bits, and parity.

Stop bits are used to signal the end of communication for a single packet. Typical values are 1, 1.5, and 2 bits. Because the data is clocked across the lines and each device has its own clock, it is possible for the two devices to become slightly out of sync. Therefore, the stop bits not only indicate the end of transmission but also give the computers some room for error in the clock speeds. The more bits used for stop bits, the greater the lenience in synchronising the different clocks, but the slower the data transmission rate.

7.5.2 Low-voltage differential signalling (LVDS)

LVDS is a differential signalling system, which means that it transmits two different voltages which are compared at the receiver. LVDS uses this difference in voltage between the two wires to encode the information. The transmitter injects a small current, nominally 3.5 mA, into one wire or the other, depending on the logic level to be sent. The current passes through a resistor of about 100 to 120 (matched to the characteristic impedance of the cable) at the receiving end, then returns in the opposite direction along the other wire. From Ohm's law, the voltage difference across the resistor is therefore about 350 mV. The receiver senses the polarity of this voltage to determine the logic level.

The small amplitude of the signal and the tight electric- and magnetic-field coupling between the two wires reduces the amount of radiated electromagnetic noise.

The low common-mode voltage (the average of the voltages on the two wires) of about 1.25 V allows LVDS to be used with a wide range of integrated circuits with power supply voltages down to 2.5 V or lower. The low differential voltage, about 350 mV as stated above, causes LVDS to consume very little power compared to other systems. For example, the static power dissipation in the LVDS load resistor is 1.2 mW, compared to the 90 mW dissipated by the load resistor for an RS-422 signal. Without a load resistor the whole wire has to be loaded and unloaded for every bit of data. Using high frequencies and a load resistor so that a single bit only covers a part of the wire (while travelling near light speed) is more power efficient.

LVDS only became popular in the latter half of the 1990s. Before that, computers were too slow to make use of such fast data rates, and the need to run twice as many wires for the same amount of data outweighed the speed benefits. It is currently the basis for Fire wire.

7.5.3 Recommended or Radio Standard (RS-232, RS-422, RS-485)

Serial is a device communication protocol that is standard on almost every PC. (Do not confuse it with universal serial bus, USB). Most computers include two RS-232-based serial ports. Serial is also a common communication protocol for instrumentation in many devices, and numerous GPIB compatible devices come with an RS-232 port. Furthermore, you can use serial communication for data acquisition in conjunction with a remote sampling device. For a short comparison of the recommended standard see table 7.2.

RS-232 is the serial connection found on IBM-compatible PCs. Engineers use it for many purposes, such as connecting a mouse, printer, or modem, as well as for industrial instrumentation. Because of line driver and cable improvements, applications often increase the performance of RS-232 beyond the distance and speed listed in the standard. RS-232 is limited to point-to-point connections between PC serial ports and devices (single ended, susceptible to noise). You can use RS-232 hardware for serial communication for distances up to 50 ft, but multidrop is not possible (one RS232 port per device). RS use a large number of extra pins which are used for different control signals described more closely in figure 7.2

Signal	DB-25	DE-9	EIA/TIA 561	Yost	RJ-50
Common Ground	7	5	4	4,5	6
Transmitted Data (TD)	2	3	6	3	8
Received Data (RD)	3	2	5	6	9
Data Terminal Ready (DTR)	20	4	3	2	7
Data Set Ready (DSR)	6	6	1	7	5
Request To Send (RTS)	4	7	8	1	4
Clear To Send (CTS)	5	8	7	8	3
Carrier Detect (DCD)	8	1	2	7	10
Ring Indicator (RI)	22	9	1	-	2

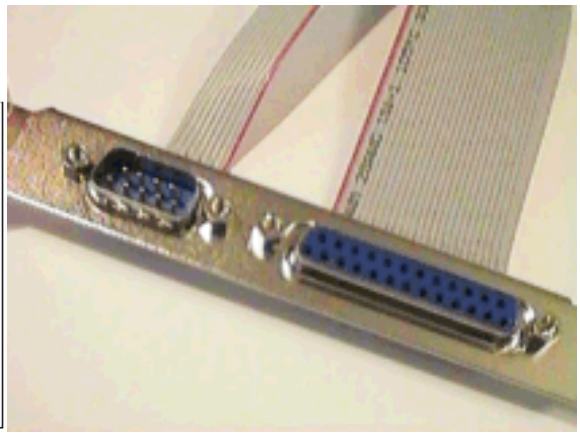


Figure 7.2: Left: Signals and their corresponding pins for RS-232. Right two D-sub connectors (9 and 25).

RS-422 (EIA RS-422-A Standard) is the serial connection used on Apple Macintosh computers. RS-422 uses a differential electrical signal, as opposed to the unbalanced signals referenced to ground with RS-232 (four wires instead of two). Differential transmission, which uses two lines each for transmit and receive signals, results in greater noise immunity and longer distances as compared to RS-232. The greater noise immunity and distance are big advantages in industrial applications. Multidrop, up to 10 devices and 500ft cable length.

RS-485 (EIA-485 Standard) is an improvement over RS-422 because it increases the number of devices from 10 to 32 and defines the electrical characteristics necessary to ensure adequate signal voltages under maximum load. With this enhanced multidrop capability, you can create networks of devices connected to a single RS-485 serial port. The noise immunity and multidrop capability make RS-485 the serial connection of choice in industrial applications requiring many distributed devices networked to a PC. RS-485 is a superset of RS-422; thus, all RS-422 devices may be controlled by RS-485. You can use RS-485 hardware for serial communication for up to 4000 ft of cable.

Most equipment using RS-232 serial ports use a DB-25 type connector even

	RS232	RS422	RS485
Cabling	Single ended	Single ended/multi-drop	multi-drop
Transmitters	1	5	32
Receivers	1	10	32
Max distance	50 feet 19.2 Kbps	4000 feet at 100 Kbps	4000 feet at 100Kbps
Max date rate	19.2 KBPS at 50 feet	10 Mbps at 50 feet	10 Mbps at 50 feet
Signalling	unbalanced	balanced	balanced
Input level	± 3 V	0.2 V difference	0.2 V difference
Output current	0.5 A	150mA	250mA

Table 7.2: Comparison between different RS interfaces

if the original documents didn't specify a specific connector, many PCs today use DB-9 connectors since all you need in asynchronous mode is 9 signals. Normally the male connector is on the PC side and the female connector is on the device side even if this is not always the case.

7.5.4 IEEE-1394 Fire wire

FireWire was developed by Apple Computer in the 1990s and is the most well-known of the LVDS standards. The system is commonly used for connection of data storage devices and digital video cameras, but is also popular in industrial systems for machine vision and professional audio systems. It is used instead of the more common USB due to its faster effective speed, higher power distribution capabilities, and because it does not need a computer host. Perhaps more importantly, FireWire makes full use of all SCSI capabilities and, compared to USB 2.0 High Speed, has higher sustained data transfer rates, a feature especially important for audio and video editors.

FireWire can connect together up to 63 peripherals in an acyclic network structure (hubs). It allows for hot-swapping (unhooking without turning off. It allows peer-to-peer device communication, such as communication between a scanner and a printer, to take place without using system memory or the CPU. FireWire also supports multiple hosts per bus. USB requires a special chipset to perform the same function, effectively resulting in the need for a unique and expensive cable, whereas FireWire requires only a cable with the correct number of pins on either end - (normally 6). Its six-wire cable can supply up to 45 watts of power per port, allowing moderate-consumption devices to operate without a separate power cord. The Sony-branded i.Link usually omits the power part of the cable/connector system and uses a 4-pin connector. Power is provided by a separate power adaptor.

FireWire 400 can transfer data between devices at 100, 200, or 400 Mbit/s data rates. Although USB2 claims to be capable of higher speeds (480Mbit/s), FireWire is, in practice, faster. Cable length is limited to 4.5 metres but up to 16 cables can be daisy chained yielding a total length of 72 meters under the specification.

FireWire 800 was introduced commercially by Apple in 2003. This newer 1394 specification and corresponding products allow a transfer rate of 786.432 Mbit/s with backwards compatibility to the slower rates and 6-pin connectors of FireWire 400.

The full IEEE 1394b specification supports optical connections up to 100 metres in length and data rates all the way to 3.2 Gbit/s. With this new technology, FireWire, which was arguably already slightly faster, is now substantially faster than Hi-Speed USB.

Firewire S800, 1394c was published as standard in 2007, and allows for the use of the RJ45 connector (ordinary LAN connector) with a transfer rate of up to 800

MB/s. It will also allow for using the same port both for firewire and Ethernet.

7.5.5 USB

USB was originally seen as a complement to FireWire, which was designed as a high-speed serial bus which could efficiently interconnect peripherals such as hard disks, audio interfaces, and video equipment. USB originally operated at a far lower data rate and used much simpler hardware, and was suitable for small peripherals such as keyboards and mice. However, because FireWire ports were more costly to implement than USB ports, primarily due to their per-port licence fee, they were rarely provided as standard equipment on computers, and peripheral manufacturers offered many more USB devices. The introduction of USB 2.0 Hi-Speed, with its widely advertised 480 Mbit/s signalling rate, convinced many consumers that FireWire was outdated (although this was not necessarily the case.

Today, USB Hi-Speed is rapidly replacing FireWire in consumer products. FireWire retains its popularity in many professional settings, where it is used for audio and video transfer, and data storage.

A USB system has an asymmetric design, consisting of a host controller and multiple daisy-chained devices. Additional USB hubs may be included in the chain, allowing branching into a tree structure, subject to a limit of 5 levels of branching per controller. Not more than 127 devices, including the bus devices, may be connected to a single host controller. Modern computers often have several host controllers, allowing a very large number of USB devices to be connected. USB cables do not need to be terminated.

USB was designed to allow peripherals to be connected without the need to plug expansion cards into the computer's ISA, EISA, or PCI bus, and to improve plug-and-play capabilities by allowing devices to be hot-swapped (connected or disconnected without powering down or rebooting the computer). When a device is first connected, the host enumerates and recognises it, and loads the device driver it needs.

USB can connect peripherals such as mice, keyboards, gamepads and joysticks, scanners, digital cameras, printers, external storage, networking components, etc. For many devices such as scanners and digital cameras, USB has become the standard connection method. USB is also used extensively to connect non-networked printers, replacing the parallel ports which were widely used; USB simplifies connecting several printers to one computer. As of 2004 there were about 1 billion USB devices in the world. As of 2005, the only large classes of peripherals that cannot use USB are displays and monitors, and high-quality digital video components, because they need a higher data rate than USB can provide.

The USB connector provides a single nominally 5 volt wire from which connected USB devices may power themselves. In practice, delivered voltage can drop well below 5 V, to only slightly above 4 V. The compliance spec requires no more than 5.25 V anywhere and no less than 4.375 V at the worst case; a low-power function after a bus-powered hub. In typical situations the voltage is close to 5 V. A given segment of the bus is specified to deliver up to 500 mA. This is often enough to power several devices, although this budget must be shared among all devices downstream of an unpowered hub. A bus-powered device may use as much of that power as allowed by the port it is plugged into.

There are four speeds for USB: low speed 187kB/s (USB 1.1), full speed 1.5 MB/s (USB 1.1, USB 2.0), Hi-speed 60 MB/s. Most hubs fall back to full speed, while Hi-speed is maintained by pure USB 2.0 connections. There is a development going on towards the USB 3.0 and super-speed of 600 MB/s. It is to include an optical link, and is likely to arrive in 2009 or 2010.

7.5.6 Local area networks (LAN)

Ethernet is based on the idea of peers on the network sending messages in what was essentially a radio system, captive inside a common wire or channel, sometimes referred to as the ether. Each peer has a unique 48-bit key known as the MAC address to ensure that all systems in an Ethernet network have distinct addresses. By default network cards come programmed with a globally unique address but this can generally be changed and there are a number of reasons for doing so. Since Ethernet is the standard computer communication nowadays it is getting very popular for transmitting data between computers, also at a lower level transferring files directly without any programmatic overlayers. LabVIEW contains functionality for doing this automatically.

Despite the huge changes in Ethernet from a thick coaxial cable bus running at 10 Mbps to point-to-point links running at 1 Gbps and beyond, the different variants remain essentially the same from the programmer's point of view and are easily interconnected using readily available inexpensive hardware.

The highest standard distribution net for ethernet right now is 1000BASE-T which is a standard for Gigabit Ethernet over copper wiring. In a departure from both 10BASE-T and 100BASE-TX, 1000BASE-T uses all four cable pairs for simultaneous transmission in both directions through the use of echo cancellation and a 5-level pulse amplitude modulation (PAM-5) technique. Each network segment can have a maximum distance of 100 metres. This usually consists of 90 m horizontal (inside the building), 9 m at the patch panel, and 1 m from the port to the computer or node. The future is likely to see much more of WLAN communication.

7.5.7 Radio communication

WLAN is an extension of local area networks, using a number of standards for data communication over radio. Today with 54 Mbit/s 802.11a (5 GHz) and 802.11g (2.4 GHz) they are becoming very common and a swift way to transfer data (802.11a, is much faster than 802.11b, with a 54Mbps maximum data rate. 802.11g maintains compatibility with 802.11b and offers data rates comparable with 802.11a). However they are not very safe and in measurements the WLAN networks can induce large noise problems .

7.5.8 Optical data transfer

Today most of our internet communication is made over optical fibers. In optical fibers the digital data is transferred as light signals within thin glass fibers, where the light is confined through total internal reflection. Theoretically fibers can communicate with bandwidths of up to THz regime, while the current limit is GHz. They are excellent for point to point contact where large data streams are to be passed between instruments. They also have a superior advantage when it comes to noise immunity since light is totally insensitive to electrical disturbances.

Typically a fibre can carry GHz data streams over 100:s of kilometre. Today the electro-optical interfaces are quite cheap (you find them in audio-visual equipment, and they are finding use in instrumentation, and will probably be large increase in use over the years to come.

7.5.9 Serial Instrument buses: VXI, PXI and LXI

There are three main instrument buses developed for serial interfaces (although USB should be included as a one of the most used ones not offering the same triggering capabilities).

The VXI bus architecture is an open standard platform for automated test based upon VMEbus. VXI stands for VME eXtensions for Instrumentation, defining additional bus lines for timing and triggering as well as mechanical requirements and standard protocols for configuration, message-based communication, multi-chassis extension, and other features.

In 1997 a more recent standard, PXI, was launched, it is of a similar architecture to VXI, but has a smaller form factor and is based on the PCI bus. Pci eXtensions for Instrumentation (PXI) is a modular instrumentation platform originally introduced in 1997 by National Instruments. PXI is designed for measurement and automation applications that require high-performance and a rugged industrial form-factor. With PXI, you can select the modules from a large number of vendors and easily integrate them into a single PXI system, over 1150 module types at 2005.

PXI is based on CompactPCI the standard for insertion cards into personal computers, offering an very good interfacing with modern computers. It offers all of the benefits of the PCI architecture, but adds integrated timing and synchronisation that is used to route synchronisation clocks, and triggers internally. The open architecture allows for more cards than a PC and for the hardware to be reconfigured to provide new facilities and features that are difficult to emulate in comparable bench instruments. PXI System performance now approaches and often exceeds the performance of the older VXI test standard.

PXI modules providing the instrument functions are plugged into a PXI chassis which may include its own controller running an industry standard Operating System such as Windows XP or Windows 2000, or a PCI to PXI bridge that provides a high speed link to a desktop PC controller. The PXI Standard was updated in 2005 with an additional specification termed PXI Express, this is based on the PCI Express technology, initial products are expected to focused on modules with very high bandwidth requirements.

A related standard, LXI (Lan Extensions for Instrumentation) was introduced in Sept 2005, based around Ethernet communication, LXI is expected to challenge the much older and formally dominant IEEE-488 Instrumentation Bus and to a lesser extent the VXIbus bus. It is thought of as an replacement of the GPIB bus, and as of today most dominating measurement companies are involved in developing both PXI and LXI buses. This will probably lead to LXI as the preferred choice for instrument to instrument communication, simple control, and PXI for complex real-time demanding combination of modular instruments. The LXI Standard has three key functional attributes:

- A standardised LAN interface that provides a framework for web based interfacing and programmatic control. The LAN interface can include wireless connectivity as well as physically connected interfaces. The interface supports peer to peer operation as well as master slave operation.
- A trigger facility based on the IEEE 1588 Precision Timing Protocol that enables modules to have a sense of time that allows modules to time stamp actions and initiate triggered events over the LAN interface.
- A physical wired trigger system based on a LVDS electrical interface that tightly synchronises the operation of multiple LXI boxes.

7.6 Repetition questions

Measurement systems:

1. Describe the advantages and disadvantages of isolated instruments and connected instrument set-ups.
2. Describe different types of control systems (PC, Microprocessor, dedicated instrument, FPGA), what are the choices and when do you use what kind of system? Order them in appropriate sequence when considering acquisition time, determinism, user friendliness and data speed.
3. Describe the difference between a real time system and a non real time system.

Computer communication:

1. Describe different ways a signal can be degraded as it is sent along a wire/data cable.
2. Explain what a computer bus is.
3. Why is a serial bus more effective in transferring data than a parallel bus?
4. What are the advantages of using a current loop for communication?
5. Explain what serial and parallel communication is.
6. List the most important parameters (max cable length, speed, number of devices, communication mode etc) for various computer bus standards (RS232, RS422, RS485, USB, Firewire).
7. List the instrument buses available. What communications speeds do they allow for, what are their advantages compared to ordinary computer communication buses?

Chapter 8

Planning and Performing Experimental work

Experimental work is 95% planning and being well prepared for the unexpected, a tricky art to master. This chapter is intended to give you a guideline on how to be better prepared.

8.1 General work-flow of experimental work

The traditional image of a good experimentalist is that the person loves to spend all time in the lab and never exits the lab. However, nothing could be more wrong. It might seem that the person always is doing the same thing all the time, but a successful experimentalist never gets detached from the theoretical, scientific and economic surroundings, and always works with a goal in mind (although the goal might change from time to time).

- **Setting a goal:** An experiment starts with setting a preliminary goal for what to achieve. This goal has to be founded in a good sense for the theory and the experiments that has been performed before. This goal should always be kept in mind and updated when the experiments proceeds.
- **Reading up:** It is time to collect information on how the experiment can be performed, if it has been done before and how it has been done before. This is also the time to get a full understanding of the theoretical description of the problem.
- **Planning:** The goal must now be extended to a master-plan. What should actually be measured, decisions must be made on acquisitions, and a preliminary plan for what data to obtain should be ready by now. An extra check must be made that all equipment actually meets the requirements. Funding for equipment and time to perform the experiment has to be found, maybe external sources of funding has to be contacted. Often a project plan must be submitted at this stage. The experiment might also need collaborations for experimental/theoretical work, then find nice people to collaborate with.
- **Prepare set-up:** A decision has to be made what set-up to actually perform the experiment on.
- **Perform experiments:** Ordinarily at least two sets of experiments has to be performed: one fast to scan the relevant parameter space, then dedicated experiments aimed at yielding the best performance of your set-up giving the most clear-cut data to reach your goal.

- **Finalise the results:** Now results have to be presented in a nice manner. This can only be done if this was thought of during the experiments and in the planning process!

8.2 Preparations

8.2.1 Setting a goal

When performing experiments it is important to be oriented towards a specific problem which you want to solve. It might start out as a very general wish, but the nature of experimental work will force you to narrow down your scope. However, when you do this you have to have the knowledge of the system and the options to do it. Well planned experiments are always marked by the correct, well informed choices.

Experimental work is goal oriented, but with a dynamic goal as the situation will always change. It is always a need for a guideline when making choices, that can come with short notice when you are performing experiments. This guideline is your goal, your knowledge and your intuition based on that same knowledge.

8.2.2 Information sources

During preparation you need to take all information sources into account:

- Text books
- Journal Papers
- Databases - web of science is highly recommended - www.isiknowledge.com
- Friends. Use your network, this is the easiest way to get good information, and often you will also get help on the way.
- Internet. Not a validated source but if you read with care and apply a sound judgement it can give a lot of helpful information.
- Companies. For some things it is actually better to just pay the money and get the expertise.

8.2.3 Theory

Physics does not exist without theory, physical theory does not exist without experiment. As such all physics experiments can only be performed with a sound knowledge of theory. Often it is necessary to both get deep into the describing theory of the physics and all prior experiments to understand the system under study. This involves both the intuitive feeling, and the theoretical understanding for a system necessary to guide you along the experiment. Therefore no good physics experiments are performed without a sound knowledge of the theory behind the experiment. This has to come before the planning, since it is crucial for the correct design of your experimental set-up

8.2.4 Experimental background

When studying the experimental background, there are several sources that has to be followed up. First of all any similar experiments will give you hints on actual set-ups and parameters used. Then the experimental facilities has to be investigated so that the limitations set by it is clearly understood. This then has to be coupled to theory, is there another manner to perform the experiment? Can the set-up be altered slightly to gain more in knowledge?

8.2.5 Planning

When you know the prerequisites it is time to plan the experiment, this is an iterative process, you will have to go back to theory and experiments to check. There is no short-cut here. In the real world this is often the phase when you have to obtain the money, time and equipment needed for the experiment, which can take up to 15 years in extreme cases. In more complex cases this will mean that you before you make the experiment have to write a detailed project plane with budget, expected results etc. At the University your counterpart will be funding agencies like the research council, in a company it will be the directors which you have to convince.

To plan properly is to be able to manage resources well over time. Many tools where developed for this with start at the second world war. An open software that provide good tolls for this is found at http://openproj.org/. Basically this gives you the means to graphically plan events and how resources should be used through Gantt charts.

8.2.6 Equipment

The last part of the preparations is usually to collect the equipment and see to that everything that is needed is bought, that extra help that is needed is there. Always try to identify potential problematic parts of the experiment. Always have an extra part of consumables ready in case you run out, or anything breaks, especially if you are working outside your home laboratory.

8.3 Performing experiments

8.3.1 Health and safety executive

When performing experiments HSE issues must be taken seriously. However important the experiment might seem it is not worth to risk your health. Nowadays every professionally run laboratory has HSE guidelines that might be different from place to place, always check that your experiment is compatible with those before you go to another lab. At NTNU these can be found in the HMS handbook (http://www.ntnu.no/adm/hms/handbok and in the Lab Handbook (referenced on that page).

8.3.2 The art of planned experiments

Often the possibilities you have when it comes to performing experiments are very wide. Therefore experiments are often made on a need to know basis, where the goal set up in the beginning is your guideline when it comes to what to do when. In a situation where you have a very simple measurement to make this is not a problem. But with a less straightforward experiment it often pays off to cover a large set of the parameter space first. The procedure will then be:

- **Initial experiments:** With your goal in mind make an experimental investigation of the parameter space you deem as important. Make it quite fast, and use the time to fine-tune settings to get the best out of your instrumentation.
- **Analyse and update plans** Analyse the results as you measure, do they show what you expect? if so then proceed according to plan, if they do not (as usually is the case) revise you experimental plan to focus on the most interesting part. Remake the plans for the full experimental period to make sure that you will get all the data you need for the purpose of the experiment.

- **Perform dedicated experiments:** Now you have to make the experiments you need to provide you with the data you want. It is now necessary to have in mind in what manner you want to present your data, and how much data you need to prove your point. The instrumentation should be fine-tuned to fit your experiment, and make sure to get all the data you need. Repeating experiments is time-consuming and often they can not be repeated without performing the full procedure again. In complicated experiments this can be years of work!
- **Double check:** that you have the data you need before leaving the set-up.

8.3.3 Documenting experiments

Documenting experiments has a three-fold purpose:

- **Traceability:** For health, company code and research ethics all experiments should be documented and be able to trace so that any event during experiment can be investigated.
- **Memory:** Without documentation you will never remember what you did after three months, clear and definite records are essential. It is important to log all possible details, you never know what influences your experiments.
- **Scientific overview:** It is very important to document in a fashion that allows you to have an overview of the experiment(s). Today many different techniques are often combined and a good structure in your documentation is needed to draw any conclusions from such complex data sets.

Today there is no problem to keep electronic records, do that and do it from the start of the experiment.

8.4 Summing up experiments

As mentioned this should already start when performing the experiments. Condensing data to easily understandable figures is one of the first steps, this should have been prepared for already in the planning and experimental phase through collecting data for educative and conclusive figures. Post processing can take as long as the experiment, so it is important to not spend too much time on this before you know exactly what data you want to use. The best thing is again to obtain an overview of the data and after that prepare the text and figures on the portions you select to present.

8.4.1 Dividing the work

Often there is a need to divide work. This is not problematic as long as everyone involved in the experiment meet and discuss what to do with the different parts. It is important to do this often, since differences in opinions often strengthen an argument.

8.5 Codes of honor

8.5.1 Oath of the European physical society

The European physical society suggest the following oath as a guideline for research:

- In all my scientific work I will be honest and I will not do anything which in my view is to the obvious detriment of the human race.

- If, later, I find that my work is being used - in my view - to the detriment of the human race, I will endeavour to stop these developments.

8.5.2 Guidelines for professional conduct

The American physical society has the following guidelines for professional conduct:

Each physicist is a citizen of the community of science. Each shares responsibility for the welfare of this community. Science is best advanced when there is mutual trust, based upon honest behavior, throughout the community. Acts of deception, or any other acts that deliberately compromise the advancement of science, are unacceptable. Honesty must be regarded as the cornerstone of ethics in science. Professional integrity in the formulation, conduct, and reporting of physics activities reflects not only on the reputations of individual physicists and their organizations, but also on the image and credibility of the physics profession as perceived by scientific colleagues, government and the public. It is important that the tradition of ethical behavior be carefully maintained and transmitted with enthusiasm to future generations.

The following are the minimal standards of ethical behavior relating to several critical aspects of the physics profession. Physicists have an individual and a collective responsibility to ensure that there is no compromise with these guidelines.

Research Results

The results of research should be recorded and maintained in a form that allows analysis and review. Research data should be immediately available to scientific collaborators. Following publication, the data should be retained for a reasonable period in order to be available promptly and completely to responsible scientists. Exceptions may be appropriate in certain circumstances in order to preserve privacy, to assure patent protection, or for similar reasons.

Fabrication of data or selective reporting of data with the intent to mislead or deceive is an egregious departure from the expected norms of scientific conduct, as is the theft of data or research results from others.

Publication and Authorship Practices

Authorship should be limited to those who have made a significant contribution to the concept, design, execution or interpretation of the research study. All those who have made significant contributions should be offered the opportunity to be listed as authors. Other individuals who have contributed to the study should be acknowledged, but not identified as authors. The sources of financial support for the project should be disclosed. Plagiarism constitutes unethical scientific behavior and is never acceptable. Proper acknowledgement of the work of others used in a research project must always be given. Further, it is the obligation of each author to provide prompt retractions or corrections of errors in published works.

Peer Review

Peer review provides advice concerning research proposals, the publication of research results and career advancement of colleagues. It is an essential component of the scientific process.

Peer review can serve its intended function only if the members of the scientific community are prepared to provide thorough, fair and objective evaluations based on requisite expertise. Although peer review can be difficult and time-consuming, scientists have an obligation to participate in the process. Privileged information or

ideas that are obtained through peer review must be kept confidential and not used for competitive gain. Reviewers should disclose conflicts of interest resulting from direct competitive, collaborative, or other relationships with any of the authors, and avoid cases in which such conflicts preclude an objective evaluation.

Conflict of Interest

There are many professional activities of physicists that have the potential for a conflict of interest. Any professional relationship or action that may result in a conflict of interest must be fully disclosed. When objectivity and effectiveness cannot be maintained, the activity should be avoided or discontinued. It should be recognized that honest error is an integral part of the scientific enterprise. It is not unethical to be wrong, provided that errors are promptly acknowledged and corrected when they are detected.

8.6 Repetition questions

1. Describe the main points in planning and performing an experiment.

Chapter 9

Communicating knowledge

This is not a course on communicating knowledge, but after your education that is one of the primary aspects of your education. Without communication knowledge is dead. This chapter hopefully only repeats what you already know, but is meant to serve as checklist when writing reports.

9.1 Communication

The most important point to address before communicating is to think through what you want to have said and to whom. The second most important thing is to remember this when you are writing/giving an presentation.

9.2 On the content

The content of any kind of report has to be selfcontained, that is, the report should be able to be read without access to sources. Main points and interpretations should be possible to follow without any extra material. To do this clearly there has to be a tight connection in what it presented and how it is presented. The tradditional style of writing is often thought of as a time-glass. The scope of the text should be shaped as a time glass, with a wide field of interest in mind when writing the introduction and at the end to put the work into a context. On the other hand the scientific depth of the text should have the inverse relationship: it should widen at the centre where all details are discussed. The transition between the two should be seemless and with a good flow.

9.3 Presenting data

When presenting data, it is rarely that original untreated data is presented. This is only natural since we today can collect data at a unprecedented speed, and accordingly the essence of the data has to be found and presented. Today many research papers present data that might be a summary of a summary of a summary. That is might a graph that represents the slopes of a certain parameter, which has been extruded through a tedious analysis of time-sequenced images.

It is important that the presentation of data always should be made in conjunction with the text, strengthening the point that you want to make. For each and every data representation (specially images/graphs) there must be specific message that you want to communicate, and that should be contained within that image. Comparisons should be made in one image or within one table and not in-between

different tables or images. It is your responsibility as a presenter/author to make a condensed and easy to read presentation of your data.

9.3.1 Graphs

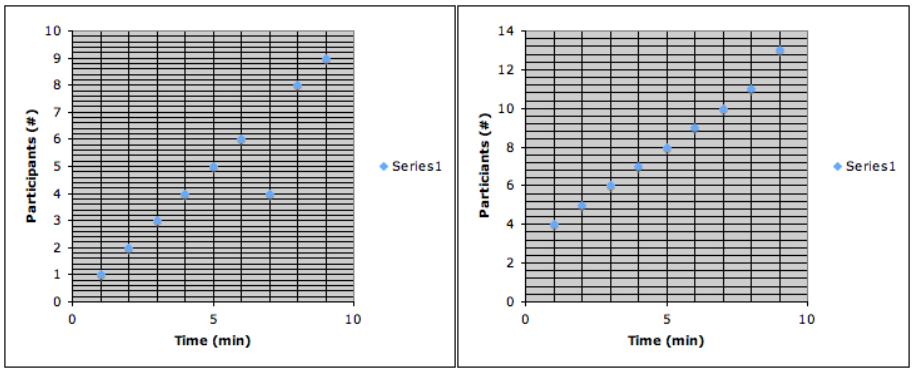


Figure 9.1: Two bad examples of graphs trying to illustrate two data sets that should be compared (typical output of micrsoft excel)

One of the most important means to representing data is through graphs, there are some golden rules about graphs that always has to be checked:

- Always maximise the use of the area, only present the interesting region.
- Always use good labelling.
- Always condense the data as much as possible

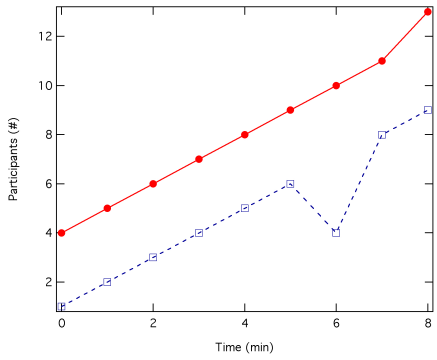


Figure 9.2: The same data as in last graphs, but plotted in a good manner, readible in black in white and optimising the paper use.

9.4 Written reports

All experiments, both in ministry and research must documented. This is to communicate the results achieved, but also to give an understanding for the fundamental problem that was considered, and the way it was solved in. A report has its structure because it is a good way to transfer knowledge, not because it should be a very interesting piece of literature, that is very important to understand when

writing a report. The different parts are placed where they should be to make it very easy for the reader to find the information and to easily grasp your findings. The schematic outline of a report is described in the following sections

Front page

The title should generally be descriptive, enclosing what is contained in the report, that is contain keywords of the experiment and the type of study but still be readable. For a general university report it should contain the title, the name of the students performing the work, course title/number.

Summary

A summary gives a short summary of the intent of the exercise described in the report, the method that has been used and the main results, finally the main conclusion should be stated. A summary does not contain equations, references or discussions. It should be self-contained, that is be possible to read without knowledge to content of the rest of the text.

Table of contents

If the report is longer than four-five pages a table of content should be included. There is often no page numbering on the front page, roman numbers for abstract and table of contents, and the first ordinary number shooed start with the introduction.

Introduction / Background

An introduction should make clear what is the intention of the investigation/exercise. It should not be a simple word-by-word restatement of the assignment. Worthwhile to mention here is a short background as a motivation, fundamental theory and equations. It should be evident from reading the introduction what was interesting with the assignment. By reading the background there should be no problem to follow the rest of the text. However, if you know the field you should be able to skip the background and still understand the full text.

Method / Experiment

This section must be adopted to the assignment that the report is describing. It should give all necessary information for recreating the assignment/experiment. For experiments that includes to describe the type of equipment used, how data was collected and treated.

Results

Here, the main results should be presented. It is always much easier to grasp a result presented in a figure or graph than in the form of text or tables. Always put a major effort in presenting you results in a condensed form through figures. That is not the same as not describing them in the text. *Both* the text and the figures should be readable without each other. Do not forget to report uncertainties. If remotly necessary for the investigation, always perform an error estimation of the original data and how that is transformed through the data interpretation.

Discussion

The purpose of the discussion is to evaluate and compare your results, how do they compare with known values, what could have been done in a better manner? This is often you only chance to display you wits!

Conclusions

This basically contains the main results again (the third time), but with reflections on what could be improved, and guidlines for the future.

References/bibliography

Whenever something is stated that is not drawn purely from you own wisdom, it has to be traced back to the source. This is the foundation of academic writing and allows for other people to follow in you trace. Always give references in text, for physics that is usually in order of appearance, with reference numbers. Typically they are written like:

Book Author: Title, Edition, Publisher, Location (of printer), Year

Reasearch paper : Author: title, journal, Year, Volume, Issue, Pages

WWW Author: Title, ¡web address¿, (Date when information was collected)

Appendix

The appendix is solely meant to spare your reader from long boring presentations of derivations, and extra data. However, the appendix should never contain new data or data not discussed in he other parts, only details of that data that parts of the readers might find interesting.

9.5 Check list - language

- Have you adopted your text to the reader, is it always at the same level or one level lower than the anticipated reader?
- Explain for your reader what has happened, not what you think have happened.
- Do not write a dairy, present the parts that were successful, and express what you leaned from the other parts of your work.
- Explain one thing at a time.
- Try to use active tense, do not describe the images, let the images illustrate important points you want to make.
- Spell check, let someone else read through the report, never blindly trust a spell/grammar check programme.
- Avoid spoken language.
- Try to avoid long complicated words, do not use any advanced language without explanations.
- Write readable, no sentences longer than 20 words.

9.6 Check list - Graphics/Data

Has or are all:

- images, graphs and tales relevant and utilising all available area?
- images, graphs and tables numbered and referred to/linked to the text?
- images and graphs given with self explanatory texts that underneath them which indicate what is presented?
- tables given with self explanatory texts that above them which indicate what is presented?
- all table values given with units?
- all graphs have axes with labels and units, with properly expressed data points with error bars?
- curve fits explained?
- all images adopted from sources given with references?
- does all data have the accurate number of digits and correct units?

Chapter 10

Control and regulation

Measurements are in many cases only half-way to the goal, most often there is also a need for control. In such a case the measurement system will form an integral part of the system. In modern processes this is vital, changes in process parameters can totally change the outcome of an industrial process. This chapter will address the basic concepts of control, the PID controller and the use of mathematical analysis for understanding control and regulation.

In many situations it is necessary to regulate and control a system. A daily experience is the shower. In the old type showers temperature control was made by setting the warm water and cold water valves separately. This often lead to surprises as the wanted value (*reference input, r(t)*) often deviated from the actual value (*controlled variable, y(t)*). The cause of this was twofold: 1) either that there was large stepwise changes in the flow of water (*disturbance input, V(s)*) due to uneven water pressure, or 2) due to that the sensitivity of the tap was to high, so that when a difference between the reference input and the controlled variable (*error signal, e(t),*) occurred the error was overcompensated. This is a typical example of a badly designed control system since no automatic mechanism compensated for incorrect water temperature.

In a modern warm water installation all these problems are compensated through the thermostat which sets the correct water temperature. This is a closed loop control system.

10.0.1 Open and closed loop control system

One often differ between open and closed control systems. In open control systems a parameter is set and is not changed regardless of the outcome (fig. 10.1). This often works very well if there are few external disturbances (ex. fire alarm, cold water flow on). When external disturbances do matter a closed loop control system is to be preferred, since then there is at least a chance to achieve a good control of the system (examples: thermostats).

10.0.2 Negative feedback control system

A simple example of a closed loop control system consists of a sensor, coparator and controller (fig. 10.2). The sensor, givs the current value of the controlled parameter. The output from this sensor is then compared with the set-point, or the reference input (a voltage value). The difference is then connected to the control unit which controls an input which can affect the controlled parameter (a heater element).

To achieve any control the feedback signal must be subtracted from the reference signal, otherwise the feedback loop will work in a self amplifying manner, and the system will not be stable.

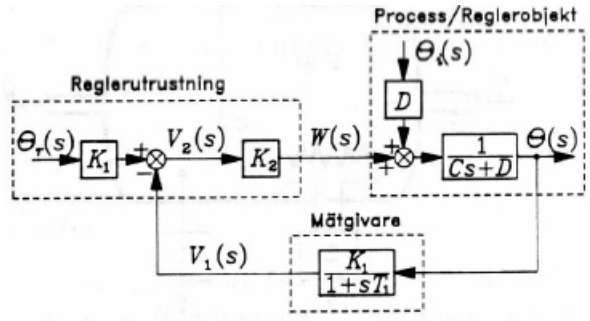


Figure 10.1: Open and closed loop control systems.

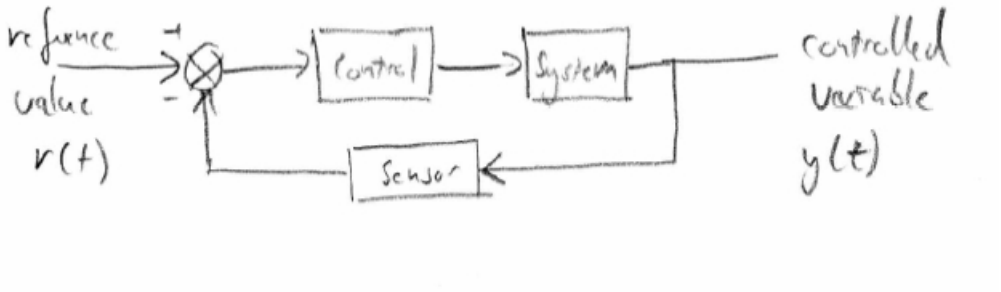


Figure 10.2: A block diagram of a feedback system with negative feedback.

Now let us consider the stability of a dynamic system: as long as there is no time lag in the system, there is no problem of stability and every error will automatically be compensated for. Problems starts to appear when a time lag is introduced, for a time-varying sinusoidal signal this means that a phase-shift will be introduced into the system. As long as the phase difference is below 180° the error signal will be amplified in the correct manner. However, when larger the system will compensate out of phase and the system will be become unstable.

10.0.3 Negative feedback revisited

Due to the importance of this system there are number of signals labels that have become standard for control situations:

$r(t), R(s)$ Reference input.

$y(t), Y(s)$ Controlled variable.

$e(t), E(s)$ Error signal.

$g(t), G(s)$ Forward transfer function.

$h(t), H(s)$ Feedback transfer function.

The product (or convolution in the time domain) $G(s)H(s)$ becomes a very important property, labelled the loop transfer function. From the block diagram we find the following:

$$E(s) = R(s) - H(s)Y(s),$$
$$Y(s) = G(s)E(s)$$

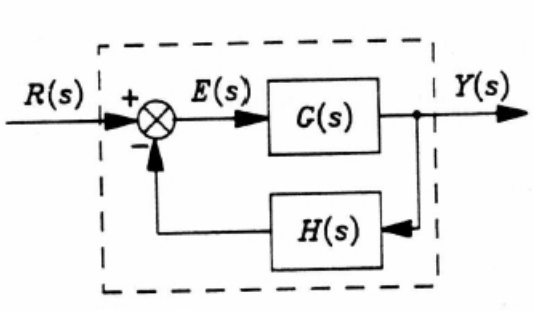


Figure 10.3: The specific case of negative feedback.

The full transfer function $Y(s)/R(s)$, that is the response to a control variation, can now be found through elimination of the error signal:

$$Y(s) = G(s)R(s) - G(s)H(s)Y(s)$$

$$\frac{Y(s)}{R(s)} = \frac{G(s)}{1 + G(s)H(s)}$$

.

Now, it can be seen that the stability that we can change the position and the numbers of poles by choice of the feedback transfer system (and function).

10.0.4 When will a typical system be stable?

Now the interesting part is to see what stability this system has. One vital question is whether the system really works as a regulator, that is, if the error goes to zero. To investigate this we can study steady state error function, e_s generally written as:

$$e_s = \lim_{t \to \infty} e(t)$$

However, in frequency space the same limit for the transform transforms into a limit of low frequencies through the final value theorem (very small frequencies are significant for actions that take very long time):

$$e_s = \lim_{s \to 0} s[R(s) - Y(S)].$$

To make things simple we can limit ourselves to investigate a very simpe example were the feedback that simply contains a simple time-independent constant

$$\frac{V_{out}}{V_{in}} = H(s) = K$$

and we have a disturbance coming adjusting the output

$$Y(s) = G(s)R(s) + I(s)$$

We can investigate the response at very long times for our system for three different cases:

1. A delta distributed disturbance α in $I(s)$.
2. A step disturbance in the input temperature of height β in $I(s)$.
3. A ramp in the input temperature with the growth in amplitude of γ $I(s)$.

To simplify the calculations we put the regulated parameter $R(t)$, to 0. Accordingly we get the following error function to be evaluated for the different types of response:

$$e_s = \lim_{t\to\infty} s[0 - I(S)H(S)] = \lim_{s\to 0} -sK(s).$$

We proceed to analyze the three cases (we already now that the last case should give an infinite value as answer).

1. $I(s) = \alpha$
$$e_s = -\lim_{s\to 0} sK\alpha = 0$$
2. $I(s) = \frac{\beta}{s}$
$$e_s = -\lim_{s\to 0} sK\frac{\beta}{s} = -\beta\frac{K}{a}$$
3. $I(s) = \frac{\gamma}{s^2}$
$$e_s = -\lim_{s\to 0} sK\frac{\gamma}{s^2} \longrightarrow -\infty$$

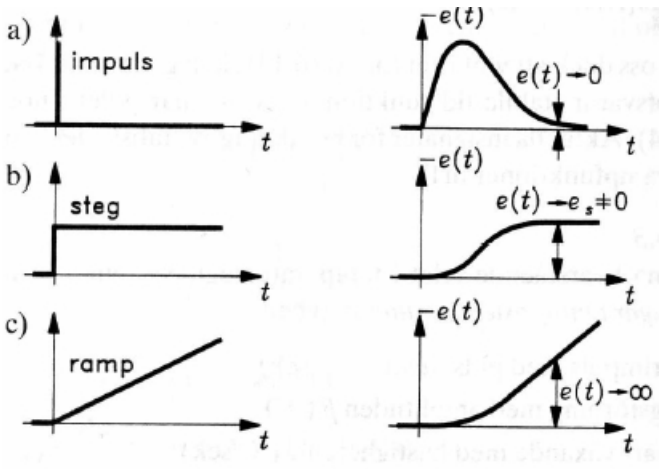


Figure 10.4: The process disturbance and the corresponding regulation error.

Thus we see (fig 10.4)that it is only impulse deviations that does not give any remaining error. stepwise changes give an error which can be minimised by increasing the amplification, but it will always remain. This is the main reason for not only using proportional feedback with a constant feedback. It this will lead small errors that can not be completely removed.

10.1 PID control

One way to correct for the problems with a remaining error is to work with feedback circuits that also contains other components. The most usual is the integrate the error, or differentiate the error (multiply or divide by s in transform space). This has the benefit of removing any sustained errors (when integrating) or speeding up the control process (when differentiating). For the differentiating part it is instructive to study the case of an oscillating error signal, it is evident that the as the error signal is the same at t_1 and t_2 the P regulator will do the same thing at both points. However, the derivative is different and actually an compensation that accounts for this might be helpful. A differential input can be added to faster counteract the action of the difference (fig. 10.5).

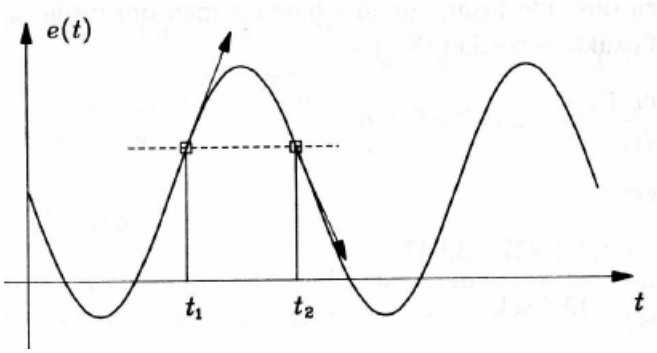


Figure 10.5: Motivation for the use of D control.

PID controllers have been industry standard for more than fifty years, and are still very common in most applications. It is instructive to look at the step response of the PID controller/regulator. To find this we have to look at what the controller does:

$$y(t) = y_0(t) + y_P + y_I(t) + y_D(t)$$

$$= y_0(t) + K_P e(t) + \frac{K_P}{T_I}\int_0^t e(\tau)d\tau + K_P T_D \frac{de(t)}{dt}$$

$$y(t) = y_0(t) + K_P[e(t) + \frac{1}{T_I}\int_0^t e(\tau)d\tau + T_D\frac{de(t)}{dt}].$$

This can easily be Laplace transformed:

$$Y(s) = y_0 + K_p[E(s) + \frac{1}{T_I}\frac{E(s)}{s} + T_D s E(s) + T_D e(0)]$$

$$Y(s) = y_0 + K_p[E(s) + \frac{1}{T_I}\frac{E(s)}{s} + T_D s E(s)]$$

10.1.1 Step response

It is instructive study the step response of the regulator (remember: this is a step response of the regulator itself, not a complete system). However, if there is a constant level even with an error signal there is a risk that error will remain when the regulator is put into a loop. The step response is a standard way to analyse a control loop. The response has a number of properties that has been defined: overshoot M ($y_{max} - 1$), Rise time T_s (the time it takes for the signal to rise from 10% to 90 % of the signal) and settling time T_δ (the time before the signal is within 2% or 5% of the final value).

To analyse the PID controller assume that the step occurs at $t = 0$ and that it has the height A, this has A/s as Laplace transform. We have three different contributions:

1. Proportional part:
$$Y(s) = K_P\frac{A}{s}$$
This yields the not surprising result of a step function with the amplitude amplified by K_P.

2. Integral part: $y(t) = \frac{K_P}{T_I}\int_0^t e(\tau)d\tau$ which transforms to:
$$Y(s) = K_P\frac{A}{T_i s^2}$$

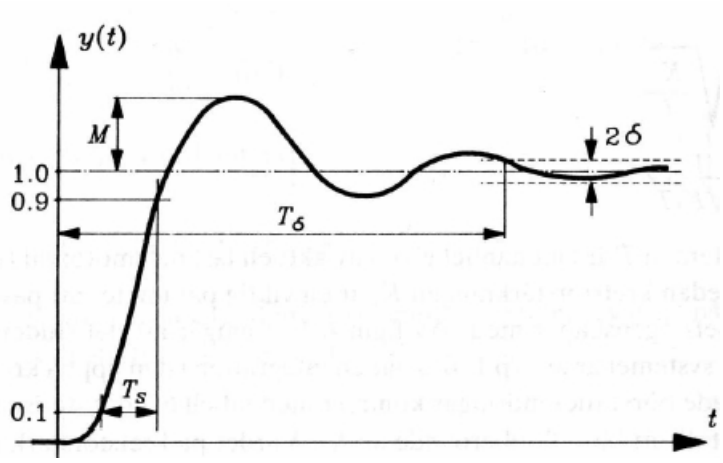


Figure 10.6: Graphical representation of the three characteristics of the step response.

and we get the final result:

$$K_p \frac{A}{T_i} t$$

. This is a linear ramp, accordingly errors that remain will be amplified and hopefully compensated away.

3. Differential part:

$$y(t) = K_P T_D \frac{de(t)}{dt}$$

which transforms to into:

$$Y(s) = T_D s E(s)$$

The final result is a delta peak of area $K_D T_D$. Often this filter is actually modelled by another function with a slight low pass filtering function since it is very hard to actually construct a controller with a infinite pulse height.

Accordingly we see that the total result yields two extra control dimensions, the integral which diminishes the remaining error while the derivative gives a faster reaction.

10.2 Stability

To understand stability we need to understand what destabilises a system, in a very simplified picture (which often is enough to understand these things) this has to do with balancing the feedback so that it always is negative, yielding a net negative contribution. By amplifying extra or adding time delays the circuit will become unstable.

The feedback circuit depicted in figure 10.7 serves as a good example to understand this. Compare the situation with the switch closed and the switch opened. If the switch is opened, the response $E_2(s)$ will be given by the transfer function $-G(s)H(s)E_1(s)$. Now consider the situation when the transfer function has a phase change of a full 360°. That is 180° from the negative feedback and a further 180° from the transfer function. The stability can then estimated from the gain in the transfer function at the frequency when the phase change is 180°.

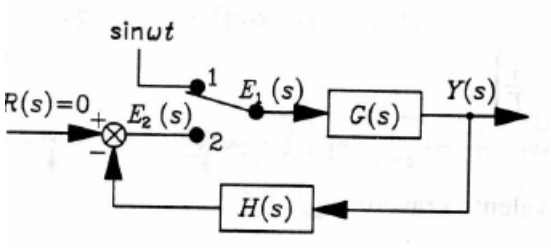


Figure 10.7: Simplified block diagram for understanding choice of parameters for stability.

a $|G(j\omega)H(j\omega)|<1$ The system will not amplify the signal when the switch is closed and will be stable.

b $|G(j\omega)H(j\omega)| = 1$ Then the system will exhibit self sustained oscillations if the switch is closed.

c $|G(j\omega)H(j\omega)|>1$ Any disturbance at that frequency will be amplified and the system will be unstable.

This forms the basis of the simplified Nyquist criterion.

10.2.1 Bode diagram

The basis of the Bode diagram and frequency analysis of the transfer function is the interpretation of the transfer function as a response to a sinusoidal excitation. To do this the input is considered only as occurring at a single frequency. We the consider only input at the imaginary axis and accordingly we exchange s with $j\omega$. It is then possible to prove that the if we excite the system with a signal $r(t) = Asin(\omega t)$ the corresponding stationary transfer function (say $g(t)$) can be written like (given no poles in the left half plane):

$$y_s(t) = A|G(j\omega)|\sin(\omega t + \angle G(j\omega)$$

Where $|G(j\omega)|$ is the absolute value and $\angle G(j\omega)$ the argument of the transfer function. The physical interpretation is accordingly that the $|G(j\omega)|$ gives the amplification and $\angle G(j\omega)$ the phase shift of a sinusoidal input.

In the Bode diagram both these parts are represented separately as function of frequency. The amplitude as the dB

$$|G(j\omega)| = 20\log^{10}|G(j\omega)|$$

. When integrating this means that we will have a phase shift of $-90°$ and a slope of -20dB/decade in frequency. In a opposite manner the phase shift will be $-90°$ and the slope 20dB/decade in frequency.

We can now consider the Nyquist criterion in this context. The Bode diagram of a system of different gain is marked in fig. 10.8 to the left. The important factor is the gain of the transfer function when we have a phase lag of 180°, and the different cases of stability can easily be found. To get a measure of the stability we consider two measures, one is the *phase margin* φ_m, and the *amplitude margin*, A_m. They are defined as:

$$A_m = \frac{1}{|G(j\omega)H(j\omega)|}$$

$$\varphi_m = 180° + \angle G(j\omega)H(j\omega)$$

Ordinarily you would expect a phase margin of at least 40°- 60° and a amplitude margin of 2-5 (6-12 dB).

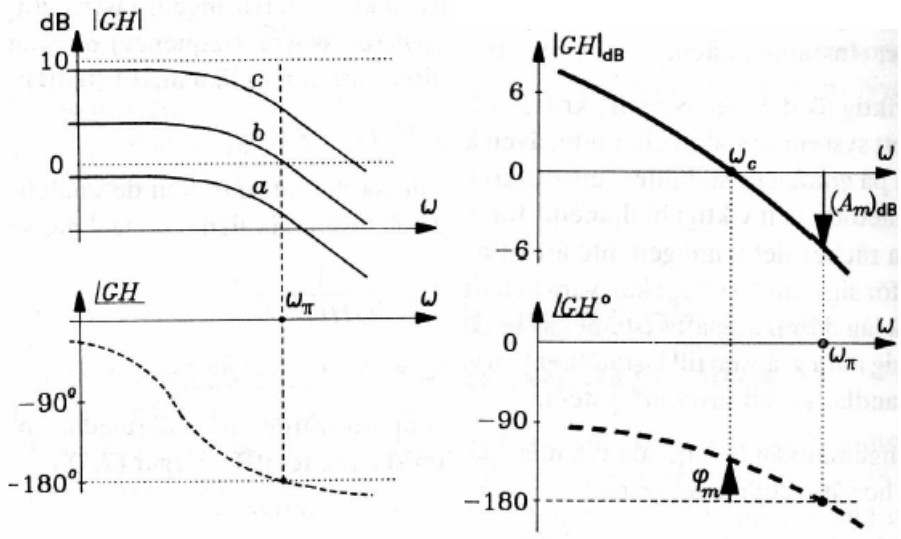


Figure 10.8: The three cases of stability displayed in a Bode diagram, and the definition of the phase margin and the amplitude margin.

10.3 Tuning

Tuning a control loop is the adjustment of its control parameters (gain/proportional band, integral/reset, derivative/rate) to the optimum values for the desired control response. The optimum behaviour on a process change or setpoint change varies depending on the application. Some processes must not allow an overshoot of the process variable from the setpoint. Other processes must minimise the energy expended in reaching a new setpoint. Generally stability of response is required and the process must not oscillate for any combination of process conditions and setpoints. Tuning of loops is made more complicated by the response time of the process; it may take minutes or hours for a setpoint change to produce a stable effect. Some processes have a degree of non-linearity and so parameters that work well at full-load conditions don't work when the process is starting up from no-load. This section describes some traditional manual methods for loop tuning.

There are several methods for tuning a PID loop. The choice of method will depend largely on whether or not the loop can be taken "offline" for tuning, and the response speed of the system. If the system can be taken offline, the best tuning method often involves subjecting the system to a step change in input, measuring the output as a function of time, and using this response to determine the control parameters.

Most modern industrial facilities no longer tune loops using the manual calculation methods shown above. Instead, PID tuning and loop optimisation software are used to ensure consistent results. These software packages will gather the data, develop process models, and suggest optimal tuning. Some software packages can even develop tuning by gathering data from reference changes. However, you will often encounter systems which does not have automatic tuning.

10.3.1 General on tuning:

If the system must remain online, one tuning method is to first set the I and D values to zero. Increase the P until the output of the loop oscillates. Then increase I until oscillation stops. Finally, increase D until the loop is acceptably quick to reach its reference. A fast PID loop tuning usually overshoots slightly to reach the

setpoint more quickly; however, some systems cannot accept overshoot. An general overview of the effect of tuning parameters is given in table 10.1.

Regulator	Rise time	Overshoot	settling time	S-S error
K_p	Decrease	Increase	Small change	Decrease
$K_I = K_p/T_i$	Decrease	Increase	Increase	Eliminate
$K_D = K_p T_D$	Small change	Decrease	Decrease	Small Change

Table 10.1: Effect on increasing parameters.

10.3.2 Practical parameters: Ziegler Nichols method

To practically obtain a fast tuning of a control it is advantageous to have some general rules to follow. One of the most common is Ziegler Nichols method. For a actual working system it actually forces you to bring the system into oscillation, which might not be a good idea. However it helps to get a feeling for how and what to do to increase the speed. The following points should be followed:

- Use the regulator as a pure P regulator (this is to only probe the transfer function of the system). Put $T_i = \infty$ and $T_d = 0$.
- Increase the proportional gain, K_P until the system starts to oscillate with constant amplitude. Note the gain, K_0 and period of the, T_0 of the oscillation.
- Change the parameters according to the table below.

Regulator type	K_P	T_I	T_D
P-regulator	$0.5K_0$	-	-
PI-regulator	$0.45K_0$	$T_0/1.2$	-
PID-regulator	$0.6K_0$	$T_0/2$	$T_0/8$

Table 10.2: Regulation parameters of the Ziegler Nichols method

10.3.3 Theoretical parameters: Ziegler Nichols method

The same method can be used to find parameters for a theoretical system. The basis for the evaluation is the a Bode diagram. In the Bode diagram you then have to find the situation which corresponds to case above, a constant oscillation. This can only occur when the phase shift of the transfer function i 180° and the amplification is 1. The first point is easy to find and the corresponding frequency for the π phase shift is denoted ω_π. From this we can find the period time of the oscillation of the theoretical system:

$$T_0 = \frac{2\pi}{\omega_\pi}.$$

The amplifications when oscillation can occur is when the amplification of the regulator gives a total response of 1:

$$K_0|GH(j\omega_\pi)| = 1$$

$$K_0 = \frac{1}{|GH(j\omega_\pi)|}$$

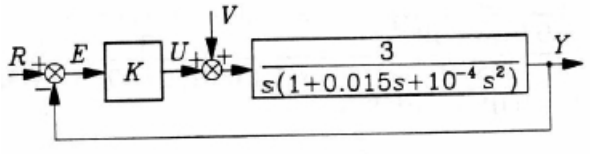


Figure 10.9: Block diagram of the regulation system.

10.3.4 Example

Use Ziegler Nichols method to find the regulation parameters for a system characterised by the following forward transfer function (see fig 10.9):

$$G(s) = \frac{3}{s(1 + 0.015s + 10^{-4}s^2}$$

.

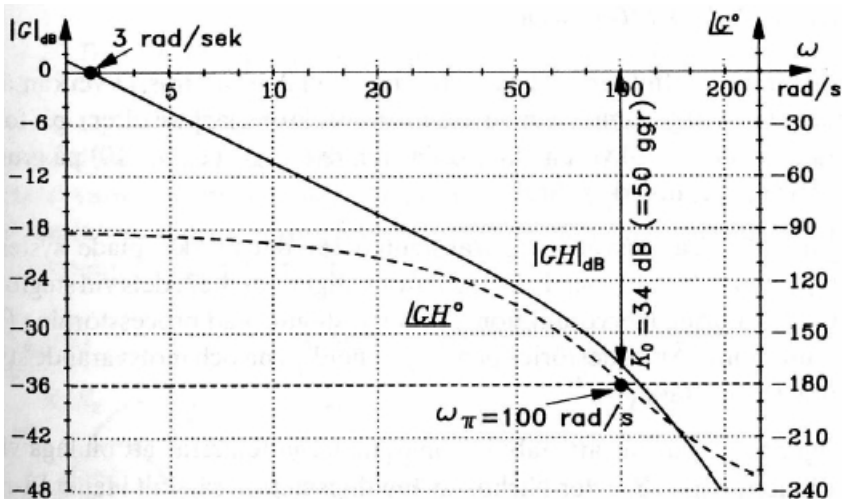


Figure 10.10: Bode diagram of the forward transfer function

We first have to find K_0. This is done through inspecting the Bode diagram. At the π phase shift we have gain -34 dB or 50 times. Accordingly we have to have the gain $K_0 = 50$ for the system to oscillate at a constant amplitude. The period length is given by the frequency at which this occurs and from the Bode diagram we get:

$$T_0 = \frac{2\pi}{\omega_\pi} = \frac{2\pi}{100} = 0.063$$

Now we obtain the control parameters for any PID configuration we might want to use as controller/regulator for the system.

Regultor type	K_P	T_I	T_D
P-regulator	$25K_0$	-	-
PI-regulator	$22.5K_0$	0.053	-
PID-regulator	$30K_0$	0.031	0.008

Table 10.3: Regulation parameters of the Ziegler Nichols method

The response to step function either in control value or ($R(s)$) or process value can now be evaluated and plotted (fig. 10.11). This gives an idea of the drawbacks

of the different types of regulation. Clearly it can be seen that the PID regulator settles faster. It can also be noted that the Ziegler Nichols method is prone to give slightly unstable systems.

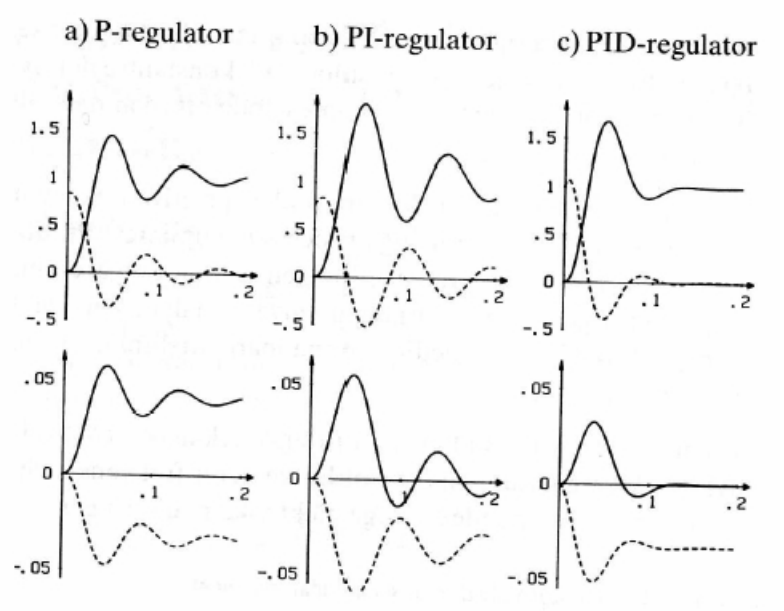


Figure 10.11: The response of a Ziegler-Nichols optimised regulators.

10.4 Repetition questions

The essential parts of this chapter is to understand what PID control can achieve and how the parameters of a PID controller should be tuned.

1. Draw a block diagram of a feedback control system, and identify the various signals (reference, error, control input, process output, measured process output/process variable) and parts (regulator, process, sensor) in the system.
2. Write an expression for the PID control function.
3. Know which term in the PID control function that leads to zero deviation (error) between the output and the reference.
4. Explain why Laplace transforms are useful in connection with control systems.
5. Explain what the transfer function is.
6. Explain how you use the Laplace transform to find the time response to an input signal, if the transfer function is known.
7. Explain what the impulse response is, and what the link to the transfer function is.
8. Describe the step response for an (ideal) PID regulator.
9. Explain how decreasing Kp will affect the step response of a PID-regulator.
10. Explain at least three different methods to tune a PID-regulator.

Chapter 11

Noise suppression

The resolution of all measurements are limited by noise. It is often it is problematic to find the source of the noise and a fundamental knowledge of noise sources and their coupling to your measurement system is important. It is necessary to be systematic both when building measurement setups, and when analysing noise. Otherwise it is very easy to make things worse. This chapter is meant as an introduction to noise and the art of systematic noise suppression. However, it should be noted, detailed knowledge about the system and the parts it contains is a necessity to properly address these kind of problems.

11.1 Types of noise

The main destinction in noise is made between inherent (to the measurement setup) and external noise. The external noise is often harmonic and due to man-made sources, while inherent noise often is broad-band. Accordingly external noise can often exhibit the sam akind of phenomena as any harmonic signal different kind of interference effects like beating.

11.2 Grounding

11.2.1 Measurement Ground, Safety ground and Earth

When discussing grounding it is essential to know the difference between measurement ground and safety ground. Safety ground has the single purpose of keeping anything connected to it at a safe potential for humans. This is definitely not the same thing as the purpose of measurement ground, which is to obtain a common reference potential, *equipotential*, for the measurement system.

- A good measurement ground will ensure the same potential all over the measurements system. Is signified by low resistance and low pick-up of external disturbances through avoiding extra impedances and any ground loops that can pick up electromagnetic noise. It does not have to be connected to safety ground, preferably it is not, since safety ground usually is an especially noisy reference.
- A good safety ground will keep the earthed component at potential safe for humans, whatever happens to the electronics installed close to the component. For measurements purposes it is often very bad, fluctuating 0.1-1 V at different parts of the system. However, for safety reasons *one* point should be connected to safety ground. But connections should be limited to this point (star-coupling).

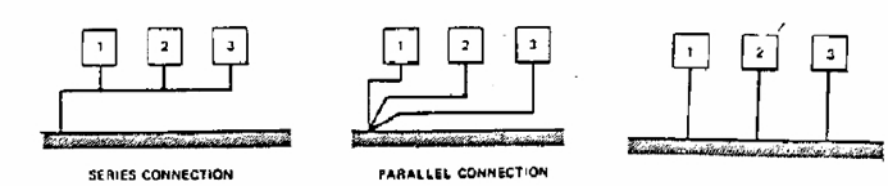


Figure 11.1: Common types of grounding.

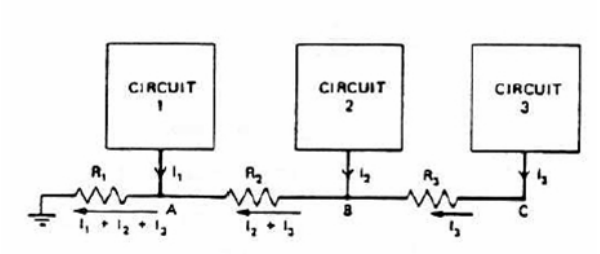


Figure 11.2: Serial grounding.

Often safety ground is physically connected to the earth (through long iron rods driven into earth, or dug down copper plates). This is why safety ground often is named earth. However the potential of the earth is very different at different spots. This is a good way for safety earthing, but not for obtaining a good steady potential for a measurement system.

11.2.2 Proper grounding

When grounding for measurement purposes, the important issue is to obtain a good constant potential for all parts of the system. This can be obtained through a number of configuration, the most common are (fig. 11.1):

- Single point earth, serial connection, one wire is used to obtain a single reference potential for all components.
- Parallel single point earth. Several wires are used to connect all components to the same ground point.
- Several points earth. Components are wired to earth through several wires that does not meet.

When evaluating the different types of configurations it is important to understand the implications they can have for different kind of measurement set-ups. Two facts are dominant when analysing this:

- All conductors are characterised by a certain impedance (R, C and L). At high frequencies the reactive part will dominate and long stretches of wire should be avoided.
- Two separate ground points rarely are rarely at the same potential.

It is now possible to analyse the different grounding systems:

The first method with one point and serial connection will accumulate errors due to currents to ground and between the different points (fig. 11.2). We can write the potential at the first point as:

$$V_A = (I_1 + I_2 + I_3)R_1.$$

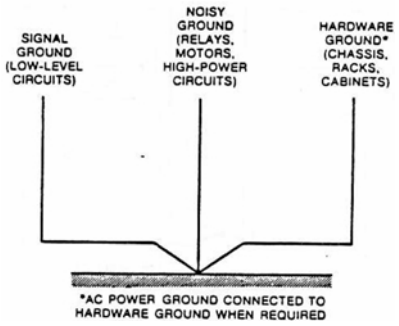


Figure 11.3: The minimum ground levels needed for most equipments

While for the end point we have:

$$V_C = I_1 + I_2 + I_3)R_1 + (I_2 + I_3)R_2 + I_3R_3.$$

This way of grounding should not be used for circuits with high currents floating in the ground cables. However, this method can be the optimal grounding since it offers a simple coupling scheme, without much double wiring.

The star coupling is preferred since then only the current and resistance to ground are separated.

At low frequencies (< 1 MHz) the preferable method for grounding is the star coupling, with separate leads for every component. This reduces all drifting potentials to a minimum and offers a maximum of control of the potentials. The main problem is the complicated wiring, often with long wires with high impedance. Accordingly at higher frequencies (> 10 MHz) it is better to operate with very short leads to a common ground plane. This will reduce the effects due to impedances in the circuits. Wiring at high frequencies is complicated, since the exact routing of wires and ground planes will affect the characteristics of the circuit, and will not be topic of this course

11.2.3 Practical grounding

When performing grounding a mixed type of grounding is often used, where systems are separated by their different sensitivity, and their potential to produce a noisy ground. Thus sensitive amplifiers are grounded separately as well as hardware (racks and boxes) that need a safety ground. Noisy equipment like motors and relays are connected in their own ground. Although the reverse is very common these grounds should be coupled in a star configuration (fig. 11.3).

11.3 Noise limits: sources of random noise

Noise is measured an analysed in terms of power, RMS deviation in potential or RMS deviation in current in a frequency interval. Typically nV/Hz. As expected noise is dependent on the bandwidth of the system. Measuring with an unnecessary bandwidth will therefore hamper your measurements by imposing too large noise levels.

The fundamental limit of measurements depends on basic noise limits of the measurement system due to inherent uncertainties in the physical processes of the measurements. For electrical measurements these are typically linked to the uncertainty of electron energy of the electrons conducting the current, or to the fundamental conduction processes found in different devices.

Noise in the from of random signals are characterised with a certain power density depending on type. Basic noise is categorised after how it decays when approaching higher frequencies. Constant noise level is called white, while noise with high low frequency content is called pink or red depending on decay in noise intensity with frequency, ($1/f$ and $1/f^2$ respectively).

11.3.1 White noise

White noise is a random signal (or process) with a flat power spectral density. In other words, the signal's power spectral density has equal power in any band, at any centre frequency, having a given bandwidth.

An infinite-bandwidth white noise signal is purely a theoretical construct. By having power at all frequencies, the total power of such a signal is infinite. In practice, a signal can be "white" with a flat spectrum over a defined frequency band.

11.3.2 Pink Noise

Pink noise is a noise that decays like $1/f$, instead of producing all frequencies equally. Sometimes pronounced as one over f noise, it is also called flicker noise. For this noise there is equal energy in all octaves. In terms of power at a constant bandwidth, 1/f noise falls off at 3 dB per octave.

11.3.3 Brown (or red) noise

Brown noise is similar to pink noise, but with a power density decrease of 6 dB per octave with increasing frequency (density proportional to $1/f^2$ over a frequency range which does not include DC. It can be generated by an algorithm which simulates Brownian motion or by integrating white noise. Brown noise is not named for a power spectrum that suggests the colour brown; rather, the name is a corruption of Brownian motion. Also known as "random walk" or "drunkard's walk" noise.

11.4 Sources of random noise

11.4.1 Thermal noise (Johnson noise)

This is the most fundamental noise form and is due to the randomised energy of electrons in a conductor. This leads to noise in the system which is linked to the absolute temperature (T). The power can be expressed like:

$$P_{thermal} = 4k_B T \Delta f,$$

where k_B is Boltzmann's constant ($1.38 \cdot 10^{-23}$ J/K) and Δf is the bandwidth of the system.

It is more interesting to now the noise expressed in potential differences or currents. We can assign the power to either a Thevenin or Norton equivalent, equating the power to the one just found:

$$P = \frac{v_{RMS}^2}{R}$$

$$P = i_{RMS}^2 R.$$

This yields the following RMS values for the thermal noise of a resistor:

$$v_{RMS} = \sqrt{PR} = \sqrt{4k_B T \Delta f R}$$

and

$$i_{RMS} = \sqrt{\frac{P}{R}} = \sqrt{\frac{4k_B T \Delta f}{R}}$$

11.4.2 Shot noise

In many electronic components charges are flowing across potential barriers independent of each others. The quantum behaviour of these electrons produce a current which is stochastic function on time. This noise is also white to nature but the amplitude is dependent on current:

$$I_{shot} = \sqrt{2eI\Delta f} \approx 5.66 \cdot 10^{-10} \sqrt{I \Delta f}.$$

Where I is the average current and Δf the frequency interval. Except for very low currents (pA) the shot noise is negligible.

11.5 Noise calculations

Most noise calculations are done through replacing an ideal component with the noise equivalent, that is either a Norton or Thevenin equivalent of the the noise generating component. From this the calculations are straightforward using ordinary circuit theory

11.5.1 Example

Consider a nanoampere meter, consisting of only a single feedback resistor of 100MΩ and a bandwidth of 3kHz. We want to find the noise floor of that circuit. The peak to peak value is roughly 8 times the RMS value, and accordingly we get the noise current in the feedback resistor to be:

$$I_{Johnson} \approx 8\sqrt{\frac{4 \cdot 300 \cdot 1.38 \cdot 10^{-23} \cdot 3 \cdot 10^3}{10^8}} \approx 5.6pA$$

This is small value often in the order of the noise of most fast low low current amplifiers. This value can be compared to the shot noise:

$$I_{shot} = \sqrt{2eI\Delta f} \approx 3.1 \cdot 10^{-8} \sqrt{I}.$$

So when measuring roughly 0.1 nA, we have a shot noise of $3 \cdot 10^{-13}$ A or 0.3 pA. This is also roughly the contemporary limit for relatively fast measurements of current.

11.6 Sources of external noise

All unwanted signals in a measurement system can be considered as a noise. For noise to exist three requirements have to be fulfilled:

- Noise source
- Coupling channel
- Receiver in the measurement system.

Accordingly there are three possibilities to remove noise, to eliminate the noise source, remove the coupling channel or to render the measurement circuit insensitive to noise. It is always preferable to remove the noise channel. However, it might often also be part of the measurement system so a detailed knowledge is needed about measurement systems to be able to do all three.

11.6.1 Coupling channels: direct, capacitive and magnetic coupling

There are generally three different ways noise can couple into a system, either through capacitive coupling, galvanic contact or magnetic coupling. The first generally couples to the system either through parasitic capacitance or through direct coupling which often means improper grounding. The second is purely due to bad wiring and unconsidered wiring. The third couples through parts of the circuit that works as antennas, picking up electromagnetic radiation at different frequencies.

11.7 Galvanic coupling

galvanic coupling most often occurs either through ground, or through the power supply. In figure 11.4 two systems are depicted that are connected through a common ground cable with a finite resistance R_g. It is evident from this that the finite resistance will yield a coupling in terms of connecting ground currents from the two different circuits to a common varying ground potential.

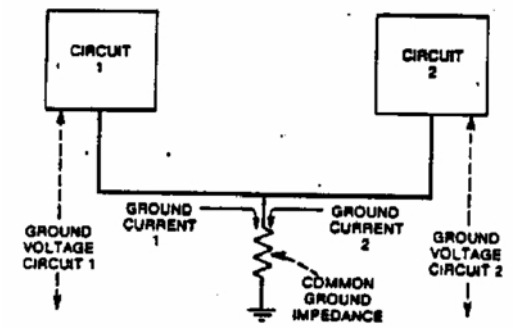


Figure 11.4: Two circuits connected through a common non-ideal ground.

A more common situation is the existence of *ground loops*. Here the problem (as depicted in fig. 11.5) is that the two different ground points very seldom is at exactly the same potential (apart from probably forming a nice antenna for electromagnetic radiation). Thus currents will be induced in the ground lead due to the potential difference and the V_{tr} will contain not only errors due to the currents passing the resistors

$$V_{err} = V_{tr} \frac{R_1 + R_2}{Z_i}$$

but also from the potential differences of the two ground points V_{cm} with the corresponding error current:

$$I_{cm} = \frac{V_{cm}}{R_2}$$

and:

$$I_{cm} = \frac{V_{cm}}{R_1}.$$

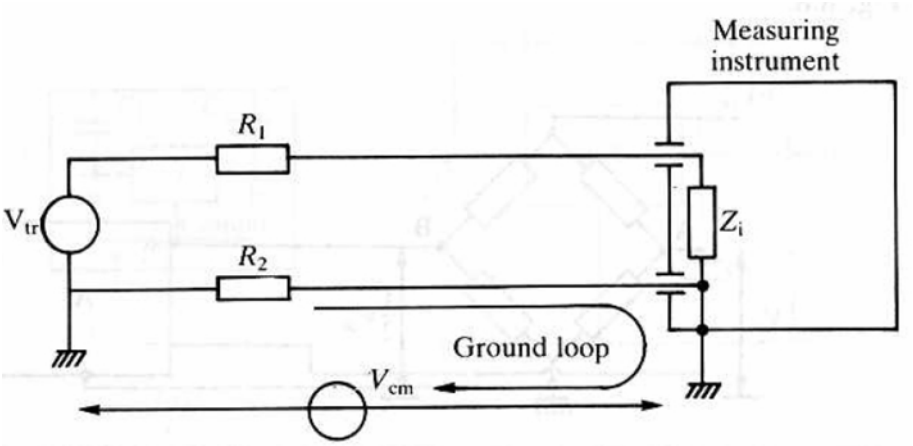


Figure 11.5: Basics of a ground loop, the potential difference at different ground points will drive an extra current through the leads.

Apart from ground it is very common to also have some kind of power supply to drive your circuit. These are often not ideal and will not be constant, but contain small, often periodic (50 Hz) deviations, *ripple*. For example, in the measurement situation given in figure 11.6 where a bridge configuration is displayed, here any deviation will be transported to the two outputs of the bridge. Often these can be suppressed through using instrumentation amplifiers with good common mode rejection. But if the imbalance of one of the current branches is large then an asymmetric signal will be fed to the amplifier which can not be corrected for through a good CMRR.

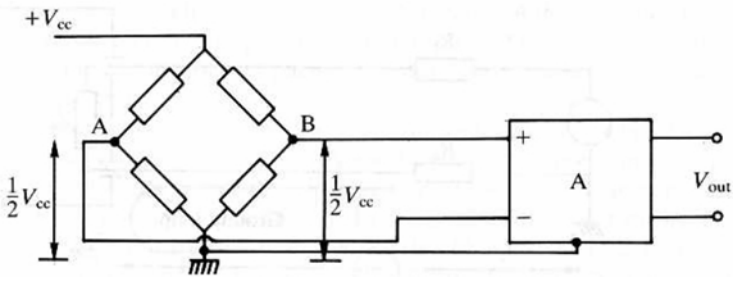


Figure 11.6: Common mode noise in a bridge measurement set-up

Another example is given in figure 11.7 where the two components are connected to the same power supply. The actual potential at the components will depend on the resistances in the leads and the currents flowing to the components, something that can be very time dependent.

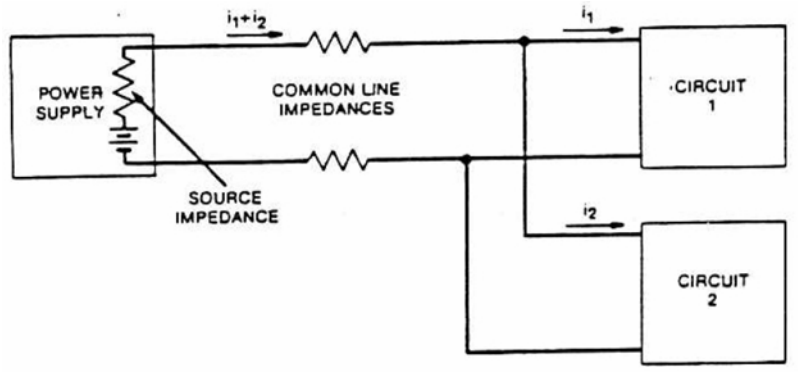


Figure 11.7: Coupling through a common current source.

11.8 Decoupling electronic disturbances

The only effective manner to reduce ground loops is to break them, and then no current can flow that will disturb the measurements. For stabilising power lines, adding capacitance's at the input will reduce the potential fluctuations. It is very common to add when building your own circuits to add large capacitance's in the immediate vicinity of operational amplifiers and other power consuming components. These will work as low pass filters, effectively filtering out any higher frequency components that might otherwise transport through the power supply network.

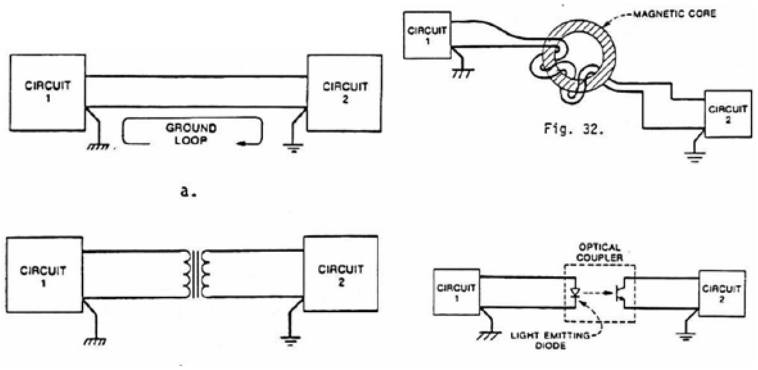


Figure 11.8: Different ways to decouple signals electrically

Other ways to break contacts is to use transformers (usually quite noisy), opto-couplers, or magnetic cores which increases resistances at higher frequencies (Fig. 11.8).

11.9 Capacitive coupling

Consider any components in vicinity of each other held at different potential. The existence of a possible potential difference will lead to the two components to work as a capacitance. One often talks about *stray capacitance* of the components. At low frequencies this is usually not a problem but with increasing frequency this will lead to a serious coupling between the two components. To start with we can consider two conductors in vicinity of each other (fig.**??**. If the first is kept at a varying potential characterised by a potential U_1 and frequency ω. The induced

noise in the second conductor will depend on the second conductors resistance to ground R and the mutual capacitance C_{12} :

$$V_N = j\omega R C_{12} U1_1.$$

This effect can of be reduced by reducing the mutual capacitance or by reducing the resistance. Most efficiently it can be prevented by surrounding one of the conductors with another conductor held at a constant potential or the same potential as the enclosed conductor (either *shielding* or *guarding*).

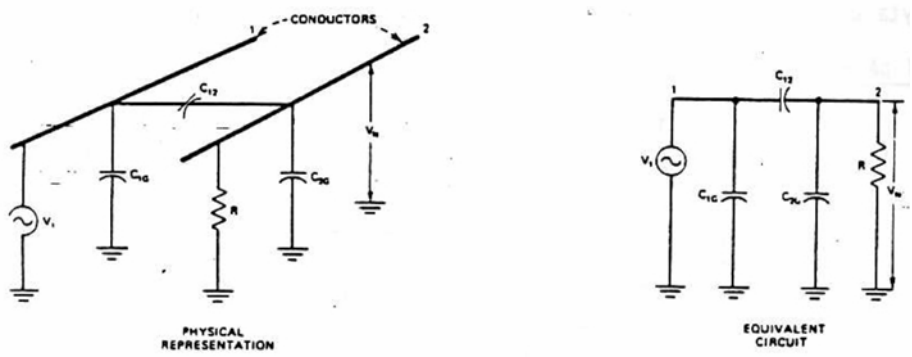


Figure 11.9: Real situation and model of a the stray capacitance between two wires

11.9.1 Shielding

By enclosing the lead totally inside another good conductor, we can accordingly shield away the stray capacitance. The ideal model for this is considered in figure 11.10

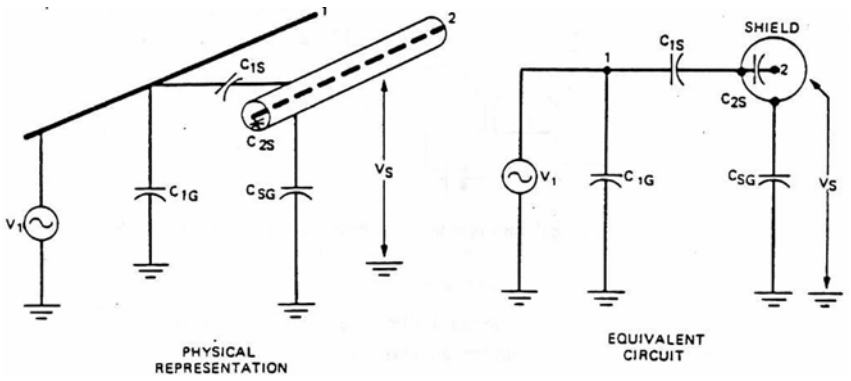


Figure 11.10: Ideal shielding

Most often the shield cannot be extended along the whole length of the cable, although modern coaxial cables can allow you to follow the ideal scheme quite well. Instead there will be a stray capacitance at the ends. However the surface of these leads are very much reduced since the stray capacitance is reduced, hence the capacitively coupled noise can be reduced with orders of magnitude.

The shield of a cable does not have to be connected to ground, but can be connected to any other potential. This is advantageous when there is need to not have any connection to ground, for floating measurements. It is also better to

connect the shield to the measuring point instead of to the ground since this will reduce any noise induced by the potential difference between the two.

11.9.2 Enclosures

When using amplifiers in enclosures the stray capacitance between the in put and output leads can be short-cut through the stray capacitance of the common lead. This is unfortunate since this is often the path of the feedback loop. To prevent this the whole enclosure must be kept at the same level as common (fig. 11.11).

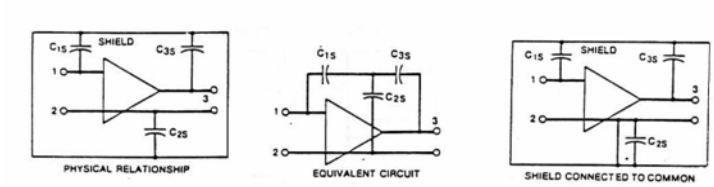


Figure 11.11: Sketches of the situation inside a shielded enclosure.

When connecting shielding enclosures it is necessary to keep the enclosure at the same potential as the shield. This is not a problem for when the signal is grounded at the enclosure, but is more tricky when grounding is made from the sensor. Then often an extra enclosure has to be added to protect the shielding enclosure of the instrument to come in contact with ground.

11.9.3 Current amplifier again - Active guarding

If we look back to the picoampere meter again we can consider the coaxial cable used to hook up the the amplifier to the signal. This typically has a capacitance 100 pF per meter. Just bending the cable or acoustically vibrating it with sound will make the capacitance change. The current will also change according to:

$$I = \frac{dQ}{dt} = C\frac{dV}{dt} + V\frac{dC}{dt}.$$

For example if the voltage is 10 mV and we have a noise level of 1pA at 1kHz due to acoustic coupling the resulting noise current is 60pA, not a very nice noise level when we want to measure pA currents.

If we want to avoid this we can instead of connecting the shield to a measurement ground connect it to a copy of the applied voltage on the inner conductor, supplied through a buffer. Through this guard variations in material and position of the two conductors will not be able to induce any noise currents into the system.

11.10 Magnetic coupling

Currents through a circuit will induce magnetic fields proportional to the current. If the current (or area) varies with time, the flux will vary, and any closed loops nearby will be able to pick up the variations. The coupling can be diminished by reducing the physical area of the receiving loop, by twisting pairs, or by moving leads closer to the ground plane. For two magnetically coupled loops with a coupling of M_{12} we find that the induced current in the second circuit is:

$$V_N = M_{12}\frac{di_1}{dt} = j\omega M_{12} I_1$$

A nonmagnetic not earthed screen placed around one of the loops will *not* affect the noise level. However a noise signal will be induced within the screen,

connecting the screen at one point (not closing any loop) will not affect the noise level. However if the current is fed back through the shield a counteracting filed is built which reduces the coupling (fig 11.12).

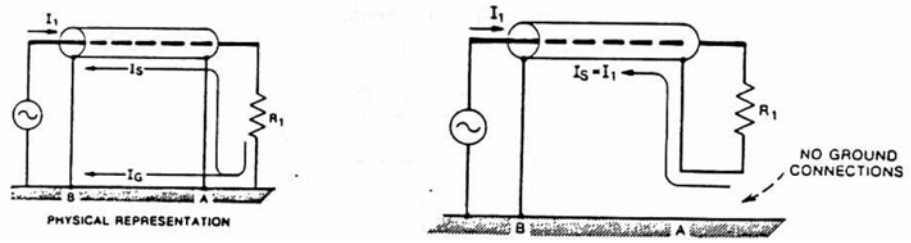


Figure 11.12: The problem with large loops provided by grounding can be reduced by feeding the current back into the shield.

11.10.1 Reducing magnetic coupling

The only way to counteract induced electromagnetic fields is to induce an opposite magnetic field that can counteract the induced field. One way to do this is to use the screen around in a coaxial arrangement to send the induced current back the same way.

- Use twisted pair or coaxial cables for carrying large currents (in this case the cables are the source of the noise).
- Shield the circuit with the proper material for the frequency at hand.
- Place magnetic sensitive parts as far as possible from magnetic field sources.
- Avoid ground loops.
- Minimise cable length.

The effectiveness of different setups (with and without shielding) is depicted in figure 11.13.

Why the coaxial cable is so effective can be realised by considering the field induced by two counter-directed currents. In a coaxial configuration they will exactly cancel each other.

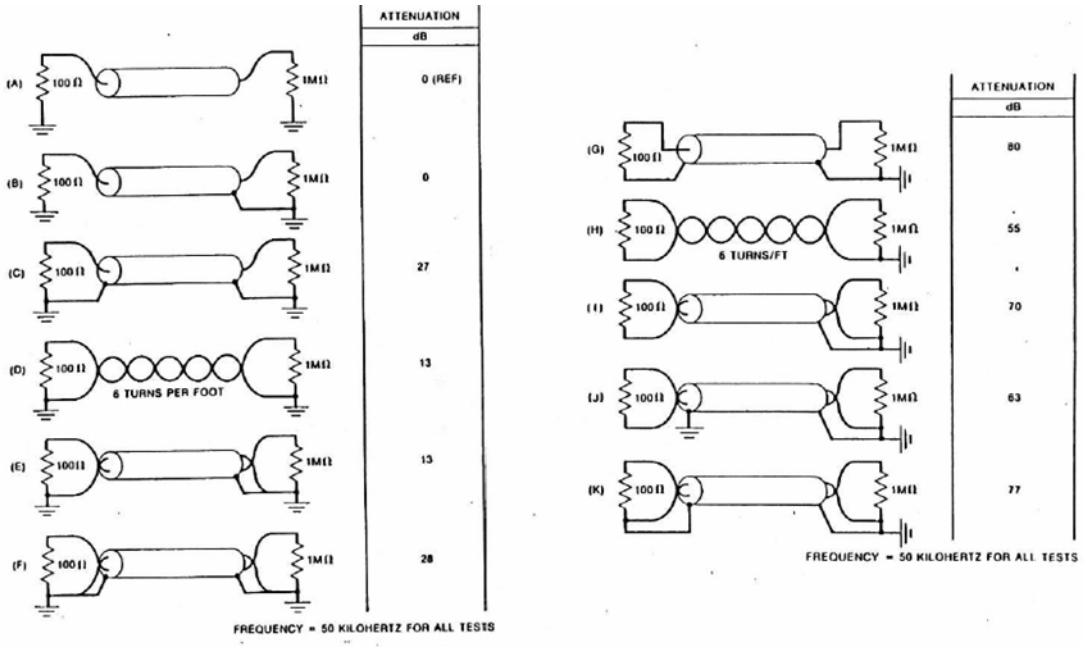


Figure 11.13: The reduction in noise level induced by a signal in the depicted circuits using a coil next to a coil of the depicted set-up. The combination of coaxial cable in combination with shield yields the best configuration.

11.11 Guide to reduction of noise in total system

11.11.1 Signal conditioning

Place your amplifier as early as you can in the measurement chain. In most cases it is also favourable to amplify as much as possible as early as possible in order to reduce noise. If possible filter the signal and use only the interesting frequency components of the system.

11.11.2 How to measure: CMRR, and balanced setup

A very good tool to render the system inert to noise is to use instrumentation amplifiers with a high common mode rejection. High common mode rejection indicates that only the differential input is amplified, while co-varying noise on both inputs will be suppressed.

To reduce noise it is important to work with balanced signal. Balanced signals are with reverse phase on each conductor (there the return path is not ground). Then less stray fields will be generated. This is generally more expensive and not necessary for low count systems. The types of cables used are summarised below.

- Single wire lines. Satisfactory for low frequency communication <400Hz.
- Open pairs.
- Twisted pairs (shielded or unshielded) Not used for high frequencies, but for reducing magnetic cross talk. However any pair nearby will suffer from crosstalk unless the sum of the current flowing through a pair is zero (unbalanced).

- Multipair. A bunch of twisted pairs working in a balanced mode, where each current sent through one part of a pair is sent in the other direction in the other part of the pair. Separately shielded to avoid stray capacitive coupling.
- Coaxial cables (single shield or multi-shield). Minimises radiation losses, and can accordingly carry higher frequency signals. The generated fields are confined in space, especially with balanced signals.

When connecting it is very important that no ground loops are formed and that the area picking up electronic noise is low recommended configurations are displayed in figure 11.14.

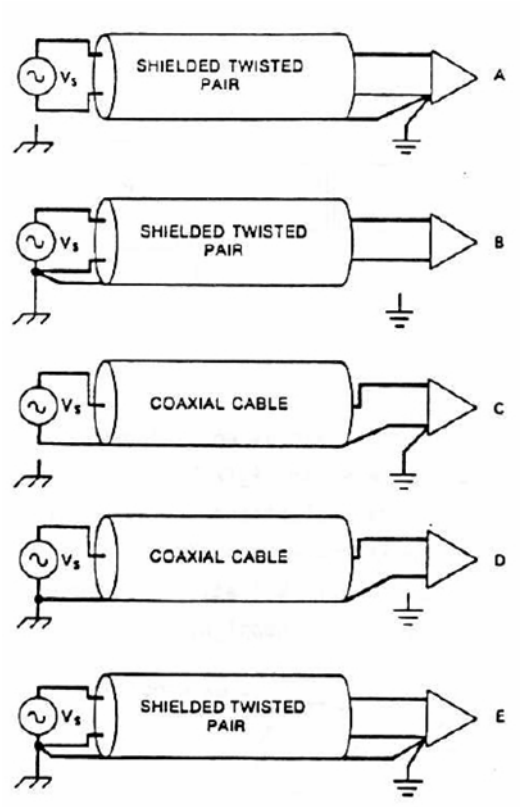


Figure 11.14: Recommended ways to connect cables.

11.11.3 Identification of noise

To identify noise it is often necessary to characterise the noise signal in the time domain. That is to investigate both the frequency of the noise signal, and any correlation of the noise signal with the status of possible signal sources.

For the first case the best method today i to analyse the signal through a FFT oscilloscope. This yields direct information on the frequencies involved in the noise signal. Motors, radio senders are characterised by their own specific frequencies.

To identify the noise source systematically turn off suspected noise sources that can be turned off. If that is not possible try to correlate times of good measurement conditions with deployment of other equipment that might interfere with your measurement equipment.

11.11.4 Reduction of noise

Perform the following:

- Remove any identified noise sources.
- Ensure that you have proper grounding for your system, break up any ground loops using separate power supplies, optical couplers or isolation amplifiers. Be sure to use ground straps that also can carry high frequency signals (wide bands with large surface area).
- For radio frequency noise ferrites around one of the conductors can be used to increase the resistance of high frequency signals. Otherwise ordinary filters might work well.
- If this does not work, rebuild you system in more sophisticated manner.

11.11.5 Numerical averaging

A more and more common method to perform noise reduction is to actually sample many more samples than you need, and later make numerical averaging. What you effectively do is low pass filter your signal. However the method has an advantage with modern I/O boards which often have a locked acquisition time. This means that by sampling more you are actually probing your signal for a longer time, which gives you the possibility of lowering the noise level. You can also numerically filter away unwanted signals through Fourier filtering. In general the noise level will follow the ordinary low for standard deviation when increasing the number of samples, that is the noise level will decrease

$$\overline{V}_N \frac{V_N}{\sqrt{N}}$$

where N is the factor you oversample the signal with.

11.12 Repetition questions

The important topic of this chapter is to understand the basics of noise generation in circuits, and that there are three means to reduce noise: reduce the source, reduce the coupling strength and reduce the receptibility in the circuit.

1. Explain ground, safety ground and earth?
2. Explain good things with a serial and a star grounding configuration.
3. Explain what noise is.
4. Describe the origin, strength and frequency dependence of shot noise and thermal noise and how it is modelled.
5. List types of noise and how the effect of them can be reduced
6. Explain what white noise is
7. Describe cross-talk between signal leads and how it can be avoided
8. Describe how you would go about to find a noise source.
9. Describe the measures you can take to reduce noise.

Chapter 12

High Frequency Signal Transmission

This chapter deals with the high frequency properties of signal cables. At high frequencies (when the wavelength of the signal is shorter than the length of the transmission line) we have to take into account how waves propagate in wires. The main point here is that any differences on the transmission line may reflect the wave and cause interference. Accordingly the properties of a signal line at every point is important to transmit signals or power in a secure manner.

12.1 Background

If you are only familiar with low frequency circuits, you may be used to treat all lines connecting the various circuit elements as perfect wires, with no voltage drop and no impedance associated to them (lumped impedance circuits). This is a reasonable procedure as long as the length of the wires is much smaller than the wavelength of the signal. At any given time, the measured voltage and current are the same for each location on the same wire.

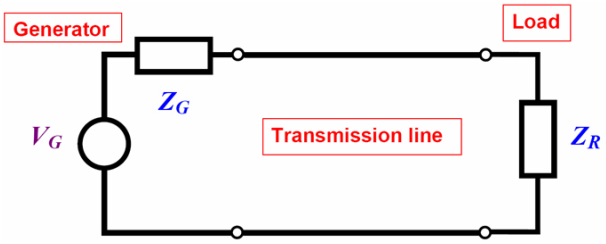


Figure 12.1: The basic setup discussed in this chapter.

At sufficiently high frequencies the wavelength is comparable with the length of conductors in a transmission line. The signal propagates as a wave of voltage and current along the line, because it cannot change instantaneously at all locations. Therefore, we cannot neglect the impedance properties of the wires (distributed impedance circuits). Note that the equivalent circuit of a generator consists of an ideal alternating voltage generator in series with its actual internal impedance. When the generator is open ($Z_R \to \infty$) we have:

$$I_{in} = 0$$

and

$$V_{in} = V_g$$

If the generator is connected to a load Z_R we instead have:

$$I_{in} = \frac{V_g}{Z_R + Z_g}$$

and

$$V_{in} = \frac{V_g Z_R}{Z_R + Z_g}$$

The most simple circuit problem that we can study consists of a voltage generator connected to a load through a uniform transmission line. In general, the impedance seen by the generator is not the same as the impedance of the load. The only exception is some very particular cases when the wavelength of the generated signal fits with the length of the transmission line:

$$Z_{in} = Z_R$$

only if

$$L = n\frac{\lambda}{2}$$

where n is an integer. At all other places the wavelength will usually prohibit a straightforward analysis of the problem. Our goal is to determine the equivalent impedance seen by the generator, that is, the input impedance of the line terminated by the load. Once that is known, standard circuit theory can be used.

12.2 General transmission line

A uniform transmission line is a *distributed circuit* that can be described as a cascade of identical cells with infinitesimal length. The conductors used to realise the line possess a certain series *inductance*, L, and *resistance*, R. In addition, there is a *shunt capacitance*, C between the conductors, and even a *shunt conductance*, R, if the medium insulating the wires is not perfect. We use the concept of shunt conductance, rather than resistance, because it is more convenient for adding the parallel elements of the shunt. We can represent the uniform transmission line with the distributed circuit below (general lossy line)

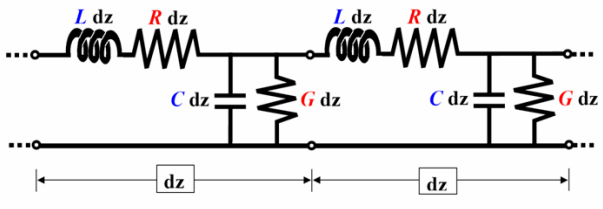


Figure 12.2: A general model of a transmission line.

The impedance parameters L, R, C, and G represent:

L series inductance per unit length (Hm^{-1}).

R series resistance per unit length (Ωm^{-1}).

C shunt capacitance per unit length (Fm^{-1}).

G shunt conductance per unit length (Sm^{-1}).

Each cell of the distributed circuit will have impedance elements with values: Ldz, Rdz, Cdz, and Gdz, where dz is the infinitesimal length of the cells. If we can determine the differential behaviour of an elementary cell of the distributed circuit, in terms of voltage and current, we can find a global differential equation that describes the entire transmission line. We can do so, because we assume the line to be uniform along its length. So, all we need to do is to study how voltage and current vary in a single elementary cell of the distributed circuit.

The solution for a uniform lossy transmission line can be obtained using the equivalent circuit for the elementary cell shown in the figure below. The series impedance determines the variation of the voltage from input to output of the cell, according to the sub-circuit The corresponding circuit equation is

$$(V + dV) - V = -(J\omega Ldz + Rdz)I$$

from which we obtain a first order differential equation for the voltage

$$\frac{dV}{dz} = -(J\omega L + R)I = -ZI$$

where Z is the series impedance per unit length. The current flowing through the shunt admittance determines the input-output variation of the current, according to the sub-circuit, The corresponding circuit equation is:

$$dI = -(j\omega Cdz + Gdz)(V + dV) = -(j\omega Cdz + Gdz)V + -(j\omega Cdz + Gdz)dV$$

The second term (including dV dz) can be ignored as it only contains second order terms. This yieds a first order differential equation for the current:

$$\frac{dI}{dz} = -(j\omega C + G)V = -YV$$

where Y is the shunt admittance per unit length. It is customary to make a difference between lossy lines which contain R and G and loss-less transmission lines in which no energy is lost through resistances.

We now have a system of coupled first order differential equations that describe the behaviour of voltage and current on the lossy transmission line:

$$\begin{aligned} \frac{dV}{dz} &= -(J\omega L + R)I \\ \frac{dI}{dz} &= -(j\omega C + G)V \end{aligned}$$

These are the telegraphers equations for the lossy transmission line case. One can easily obtain a set of uncoupled equations by differentiating with respect to the coordinate z:

$$\begin{aligned} \frac{d^2V}{dz^2} &= -(J\omega L + R)\frac{dI}{dz} = (J\omega L + R)(j\omega C + G)V = ZYV \\ \frac{d^2I}{dz^2} &= -(j\omega C + G)\frac{dV}{dZ} = (j\omega C + G)(J\omega L + R)I = YZI \end{aligned}$$

These are the telephonists equations for the lossy line.

The telephonists equations for the lossy transmission line are uncoupled second order differential equations and are again wave equations. The general solution for the voltage equation is:

$$V(z) = V^+e^{-\gamma z} + V^-e^{\gamma z} = V(z) = V^+e^{-\alpha z}e^{-i\beta z} + V^-e^{\alpha z}e^{-i\beta z}$$

Where we have split up the propagation constant in two parts:

$$\gamma = \sqrt{(j\omega L + R)(j\omega C + G)} = \alpha + i\beta$$

The real part α of the propagation constant γ describes the attenuation of the signal due to resistive losses. β describe the phase of the wave in space. For a loss-less line we have:

$$\beta = \omega\sqrt{LC}.$$

The current distribution on a lossy transmission line can be readily obtained by differentiation of the result for the voltage:

$$\frac{dV}{dz} = -(J\omega L + R)I = -\gamma V^{+}e^{-\gamma z} + \gamma V^{-}e^{\gamma z}$$

which gives:

$$I = \sqrt{\frac{(J\omega C + G)}{(J\omega L + R)}}(V^{+}e^{-\gamma z} - V^{-}e^{\gamma z}) = \frac{1}{Z_0}(V^{+}e^{-\gamma z} - V^{-}e^{\gamma z})$$

with the characteristic impedance, Z_0, of the lossy transmission line:

$$Z_0 = \sqrt{\frac{(J\omega L + R)}{(J\omega C + G)}}$$

12.2.1 Characteristic impedance Z_0

For both loss-less and lossy transmission lines the characteristic impedance does not depend on the line length but only on the metal of the conductors, the dielectric material surrounding the conductors and the geometry of the line cross-section. One must be careful not to interpret the characteristic impedance as some lumped impedance that can replace the transmission line in an equivalent circuit. The characteristic impedance only characterises the local transport properties of the transmission line.

Now let us consider the general characteristic impedance:

$$Z_0 = \sqrt{\frac{(J\omega L + R)}{(J\omega C + G)}}$$

Generally R/G is much higher than L/C so that a transmission will occur at a few kHz when the transmission line is dominated by the inductive and capacitive behaviour of the line. Typically $L = 250nH/m$ and $C = 100pF/M$ which yields a high frequency impedance of 50Ω.

12.2.2 Propagation velocity

The most interesting case is when L and C is dominating (low loss case). If the incoming signal is characterised by an angular frequency ω and a wavelength λ we have:

$$\upsilon = \frac{\omega}{2\pi}\lambda = \frac{\omega}{\beta} = \frac{1}{\sqrt{LC}}$$

For typical values of transmission lines (50Ω or 75Ω) this gives a $\upsilon = 2 \cdot 10^8$ m/s about two thirds of the velocity of light.

12.3 Characterisation of transmission lines:

Since $V(z)$ and $I(z)$ are the solutions of second order differential (wave) equations, we must determine two unknowns, V^{+} and V^{-}, which represent the amplitudes

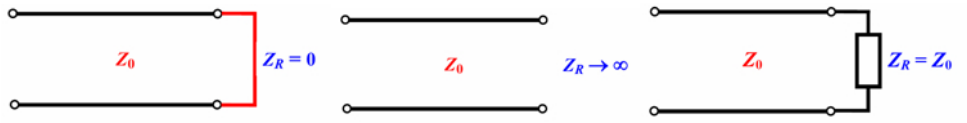


Figure 12.3: The three cases considered as examples.

of steady-state voltage waves, travelling in the positive and in the negative direction, respectively. Therefore, we need two boundary conditions to determine these unknowns. This is usually done by considering the effect of the load and of the generator connected to the transmission line. Before we consider the boundary conditions, it is very convenient to shift the reference of the space coordinate so that the zero reference is at the location of the load instead of the generator. Since the analysis of the transmission line normally starts from the load itself, this will considerably simplify the problem later. We will also let z increase when moving from load to generator along the transmission line. At the load ($z = 0$) we have, for both cases:

$$I = \frac{1}{Z_0}(V^+ - V^-)$$
$$V = V^+ + V^-$$

For a given load impedance Z_R , the load boundary condition is:

$$V^+ + V^- = Z_R I(0)$$
$$V^+ + V^- = \frac{Z_R}{Z_0}(V^+ - V^-)$$

From this we can obtain the general voltage load reflection coefficient:

$$\Gamma_R = \frac{V^-}{V^+} = \frac{Z_R - Z_0}{Z_R + Z_0}$$

To find the a more general expression along the loss-less transmission line we can define a generalised reflection coefficient by considering the phase behaviour of the distance from the point of reflection:

$$\Gamma(d) = \Gamma_R e^{-2\gamma d}$$

where we have the line equations:

$$V(d) = V^+ e^{\gamma d}(1 + \Gamma(d)) \tag{12.1}$$
$$I(d) = \frac{V^+ e^{\gamma d}}{Z_0}(1 - \Gamma(d)) \tag{12.2}$$

Finally we get the line impedance which is the ratio between the voltage and current at any given point d from the load resistance:

$$Z(d) = \frac{V(d)}{I(d)} = Z_0 \frac{1 + \Gamma(d)}{1 - \Gamma(d)} = Z_0 \frac{Z_R + Z_0 \tanh \gamma d}{Z_R \tanh \gamma d + Z_0}$$

This describes the behaviour of potential and current in a satisfactory manner.

12.3.1 Specific cases: short circuit, open circuit, impedance match

Now we can consider three specific cases: when the load is a short circuit ($Z_R \to 0$) and open circuit ($Z_R \to \infty$) or matched ($Z_R = Z_0$). We consider the potential at the load:

$$V(d = 0) = V^+ e^{\gamma 0}(1 + \Gamma_R e^{-2\gamma 0}) = V^+(1 + \Gamma_R) = 0$$

$$\Rightarrow \qquad \Gamma_R = -1$$

Since

$$\Gamma_R = \frac{V^-}{V^+}$$

$$\Rightarrow \qquad V^- = -V^+$$

As we could have expected the reflected wave is phase-shifted π at the load, which corresponds to the reflection of a wave at the fixed point of a string.

We can also look at the distributed impedance:

$$Z(d) = jZ_0 \tan \beta d$$

Thus we have a changing impedance with the distance from the short, (as we have standing waves formed). This can be utilised for tuning circuits since a short circuit can be moved and accordingly the element changed just by setting the length.

Another important case concerns the open circuit at which the current at the load must be zero:

$$I(0) = \frac{V^+ e^{\gamma 0}}{Z_0}(1 - \Gamma_R e^{-2\gamma 0}) = \frac{V^+ e^{\gamma 0}}{Z_0}(1 - \Gamma_R e^{-2\gamma 0}) = \frac{V^+}{Z_0}(1 - \Gamma_R 1) = 0$$

$$\Rightarrow \qquad \Gamma_R = 1$$

$$\Rightarrow \qquad V^- = V^+$$

Again we end up with the expected result with the same wave reflected back along the line. For the matched load we expect no reflection since the signal does not meet any obstruction. Glancing at the reflection coefficient that is exactly what we obtain:

$$\Gamma_R = \frac{V^-}{V^+} = \frac{Z_R - Z_0}{Z_R + Z_0} = 0.$$

Furthermore we find that

$$Z(d) = Z_0$$

that we do not obtain any standing waves which also is as expected.

12.3.2 Voltage standing wave ratio (VSWR)

The standing wave patterns provide the envelopes that bound the time-oscillations of voltage and current along the line. In other words, the standing wave patterns provide us with information on where the maximum values that voltage and current will be established at transmission line for given load and generator. The standing wave pattern gives a clear representation of wave interference in a transmission line. The patterns present a succession of maxima and minima which repeat in space with a period of $\lambda/2$, due to constructive or destructive interference between forward and reflected waves. Generally we can state the following:

- If the load is matched to the transmission line ($Z_R = Z_0$) the voltage standing wave pattern is flat, with value $|V^+|$.
- If the load is real and $Z_R > Z_0$, the voltage standing wave pattern starts with a maximum at the load.
- If the load is real and $Z_R < Z_0$, the voltage standing wave pattern starts with a minimum at the load.
- If the load is complex and $Im(Z_R) > 0$ (inductive reactance), the voltage standing wave pattern initially increases when moving from load to generator and reaches a maximum first.

- If the load is complex and $Im(Z_R) < 0$ (capacitive reactance), the voltage standing wave pattern initially decreases when moving from load to generator and reaches a minimum first.

The voltage standing wave ratio (VSWR) is an indicator of load matching which is widely used in engineering applications. The appearance of standing waves is crucially dependent on the matching of the load impedance, Z_R, with the line impedance Z_0. A perfect match means no standing wave formation (as there is no reflection). The ratio between maxima and minima (VSWR) becomes 1 as it is defined as:

$$VSWR = \frac{|V_{max}|}{|V_{min}|} = \frac{1 + |\Gamma_R|}{1 - |\Gamma_R|}$$

And we can deduce that:

$$\Gamma_R = 0 \qquad \Rightarrow \qquad VSWR = 0$$

If the lead is a short circuit or open conductor we find that:

$$|\Gamma_R| = 1 \qquad \Rightarrow \qquad VSWR \to \infty$$

A measurement of the voltage standing wave pattern provides the locations of the first voltage maximum and of the first voltage minimum with respect to the load. The ratio of the voltage magnitude at these points gives the VSWR directly. This information is sufficient to determine the load impedance Z_R , if the characteristic impedance of the transmission line Z_0 is known.

- STEP 1: The VSWR provides the magnitude of the load reflection coefficient Γ_R.
$$|\Gamma_R| = \frac{VSRW - 1}{VSRW + 1}$$
- STEP 2: The distance from the load of the first maximum gives the phase ϕ of the load reflection coefficient.
$$\phi = 2\beta d_{max}$$
- STEP 3: The load impedance is obtained by inverting the expression for the reflection coefficient.
$$\Gamma_R = |\Gamma_R| e^{j\phi} = \frac{Z_R - Z_0}{Z_R + Z_0}$$
$$\Rightarrow \qquad Z_R = \frac{1 + |\Gamma_R| e^{j\phi}}{1 - |\Gamma_R| e^{j\phi}}$$

12.4 Reflections

The concept of reflection coefficients is instructive to study to understand transients in signal systems. Consider a very simple system characterised by three impedance of the generator, Z_G, load, Z_L and the transmission line ,Z_0. A switch is closed at the generator side at a time. The question is how the pulse will look? It is handy to know the transit time along the lead:

$$\tau = \sqrt{LC} l$$

where l is the length of the transmission lime. At $t = \tau$ the step is reflected at the load end, and we will have a reflection given by:

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

After another time τ the pulse will come back and be reflected at the generator side:

$$\Gamma_{gen} = \frac{Z_G - Z_0}{Z_G + Z_0}$$

.

For each reflection the potential will be altered by the reflection and the pulse will gradually reach a saturated level.

12.4.1 Example

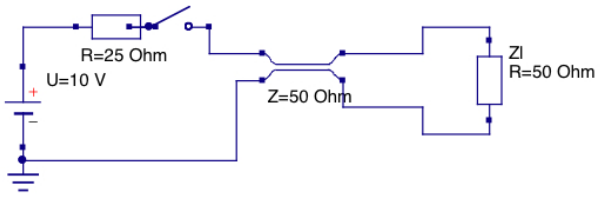


Figure 12.4: Example circuit for wave reflections at impedance boundaries.

Consider the case with a load resistance $Z_L = 75\Omega$ on a transmission line $Z_0 = 50\Omega$, a battery with internal resistance 25Ω is connected to a this lead (fig. 12.4). The important reflection coefficients are:

$$\Gamma_{load} = \frac{75 - 50}{75 + 50} = 0.2$$

$$\Gamma_{gen} = \frac{25 - 50}{25 + 50} = -0.333$$

We also need the initial potential as the battery is connected:

$$V_1 = \frac{10(50)}{25 + 50}$$

We can write the potential at all reflections:

$$V_1 = 6.67V$$

$$V_2 = 0.2V_1 = 1.33V$$

$$V_3 = -0.333V_2 = -0.455V$$

$$V_4 = .2V_3 = -0.089V$$

$$V_5 = -0.333V_4 = -0.0297V$$

$$V_6 = 0.2V_5 = 0.0059V$$

If we want to know the voltage at the generator end we have to keep track of the waves propagating forward and back in the transmission line. Accordingly after two passes of the signal we will have a potential of $6.67V + 1.33V = 8V$.

Another way to consider this system is that at the moment when a potential is applied it will only be affected by the impedance of the lead, the effect of the load comes later when the information of the load (the impedance of the load) has been reflected back, and it will continue in this manner until a steady state as been reached.

12.4.2 Investigating transmission line errors

As the delay time of a pulse depends on the actual distance and the impedance of a lead, sending a pulse or transient is actually a good way to check the integrity of a line. If the transmission line has been broken there will be a change in impedance which will reflect portions of the wave. Accordingly any error in the line can be detected and exactly located by probing the line properly.

12.5 Repetition questions

The are two important things to remember from this chapter: the true meaning of impedance in a wave propagation sense and the concept of wave propagation as mean to understand high speed electric circuits.

1. Give an expression for the characteristic impedance in a transmission line as a function of its inductance and capacitance .
2. Give an expression for the reflection coefficient at the position of the load in the transmission line, as a function of the load impedance Z_L and the characteristic impedance of the transmission line Z_0.
3. Calculate the reflection coefficient for open circuit and short circuit.
4. Explain why the reflection coefficient is zero if the load has the same impedance as the transmission line.
5. Sketch how a signal at a given point on the transmission line will vary with time, depending on the values of the impedance of the source, transmission line and load.

Chapter 13

Laboratory infrastructures

Today experimental work is often split up between home laboratories and shared laboratories. This is due to that instrumentation in extreme cases can be very expensive, and must be shared locally/nationally and internationally. In this chapter a background will be given to the only shared facility you will enter during the course: a clean room, and also the overview over future european facilities that are supported through the 2008 esfrii project.

13.1 Generally on shared facilities

Shared facilities need another degree of control to maintain a safe and good, often you apply for user time, and depending on the facility it might be that you are not even allowed to touch the instrumentation.

13.2 Clean rooms

A cleanroom is an environment, typically used in manufacturing or scientific research, that has a low level of environmental pollutants such as dust, airborne microbes, aerosol particles and chemical vapours. In addition to this a clean room often incorporates infrastructure like gas lines, cooling water and vacuum, vibration damping and gas supplies. The cleanrooms was initially adapted to lower the failure rate for production, and as usual it has its offspring from the second world war, when weapons construction was reaching levels which called for particle control. Cleanrooms can be very large. Entire manufacturing facilities can be contained within a cleanroom with factory floors covering thousands of square meters. They are used extensively in semiconductor manufacturing, biotechnology, the life sciences and other fields that are very sensitive to environmental contamination.They are also built as research facilities mainly for research on and production of micro- and nano-scale model systems utilising the tools of semiconductor industry.

A cleanroom has a controlled level of contamination that is specified by the number of particles per cubic meter at a specified particle size. To give perspective, the ambient air outside in a typical urban contain as many as 35,000,000 particles per cubic meter, 0.5 μm and larger in diameter, corresponding to an ISO 9 cleanroom.

The air entering a cleanroom from outside is filtered to exclude dust, and the air inside is constantly recirculated through high efficiency particulate air (HEPA) and ultra low penetration air (ULPA) filters to remove internally generated contaminants. Staff enter and leave through airlocks (sometimes including an air shower

Table 13.1: ISO 14644-1 cleanroom standards. The numbers a maximum number of particles of that size per m^2

Class	$>0.1\mu m$	$>0.2\mu m$	$>0.3\mu m$	$>0.5\mu m$	$>1\mu m$	$>5\mu m$	US standard
ISO 1	10	2					
ISO 2	100	24	10	4			
ISO 3	1000	237	102	35	8		Class 1
ISO 4	10000	2370	1020	352	83		Class 10
ISO 5	100000	23700	10200	3520	832	29	Class 100
ISO 6	1000000	237000	102000	35200	8320	293	Class 1000
ISO 7				352000	83200	2930	Class 10000
ISO 8				3520000	832000	29300	Class 100 000
ISO 9				35200000	8320000	293000	Room air

stage), and wear protective clothing such as hats, face masks, gloves, boots and cover-alls.

Equipment inside the cleanroom is designed to generate minimal air contamination. There are even specialised mops and buckets. Cleanroom furniture is also designed to produce a minimum of particles and to be easy to clean. Common materials such as paper, pencils, and fabrics made from natural fibers are often excluded; however, alternatives are available. Cleanrooms are not sterile (i.e., free of uncontrolled microbes) and more attention is given to airborne particles. Particle levels are usually tested using a particle counter.

Behaviour in the clean room is adapted to the environment. Since the room is kept clean through laminar air flow from the filters in the ceiling all work and equipment has to be adapted to this. Every movement should be done accordingly with the airflow in mind to protect samples.

Most cleanrooms are kept at a positive pressure so that if there are any leaks, air leaks out of the chamber instead of unfiltered air coming in. Entering a cleanroom usually requires wearing a cleanroom suit. In cheaper cleanrooms, in which the standards of air contamination are less rigorous, the entrance to the cleanroom may not have an air shower. There is an anteroom, in which the special suits must be put on, but then a person can walk in directly to the room.

13.2.1 cleanroom classes

Cleanrooms are classified according to the number and size of particles permitted per volume of air. Large numbers like "class 100" or "class 1000" refer to US FED STD 209E, and denote the number of particles of size 0.5 μm or larger permitted per cubic foot of air. The standard also allows interpolation, so it is possible to describe e.g. "class 2000". Small numbers refer to ISO 14644-1 standards, which specify the decimal logarithm of the number of particles 0.1 μm or larger permitted per cubic metre of air. So, for example, an ISO class 5 cleanroom has at most 105 = 100,000 particles per m3. Ordinary room air is approximately class 1,000,000 or ISO 9.

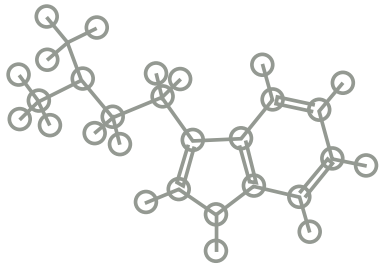
13.3 Large facilities

Today, a large number of large laboratories are being built. The reasons for this two-fold, either the detector and experimental set-up needed to perform an experiment is so large that it can not be maintained or contained in a single home laboratory examples of such infrastructures are CERN and Telescopes. The other reason is

that environment or excitition source for studying something is very expensive examples of such instalations are synchetron light sources and free-electron lasers which can supply electromagnetic radiation with very high brilliance at a tunable waevlength. Acces to such facilities is often restricted and given on project basis. In most cases an application has to be submitted which has to be judged and ranked by a scientific committee.

>Materials and Analytical Facilities

The development of new materials contributes to all areas of human activity from energy generation and storage through to medical implants and computer components. Advances in materials from steel blades to biological materials, including fluids and plasmas, has been fuelled by the capability to observe, design and assemble or manipulate them at an ever increasing definition of scale. This has supported a century-long durable and effective industrial and economic growth based on increasingly sophisticated new products, from catalysers or cell phones to new pharmaceutical drugs, and the continuous improvement in traditional products, from car engines to glass cover for housing or fabrics for clothing.



It is now increasingly possible to operate at the nanometre scale, observing and manipulating, as well as designing, atom by atom, increasingly complex materials.

Most techniques can be available in relatively small laboratory environments (like the atomic force microscope, or the atomic-layer deposition chambers), but, when it comes to operating with increasing definition on larger pieces of materials, the need is to be able to "illuminate and reach" all atoms of the materials under investigation. This requires "large facilities" capable of providing the adequate "brilliances", much like the need for a strong light to explore a dark environment.

The overall situation of the main research infrastructures in this field in Europe can be summarised as follows.

Photon Sources.

Light photons are only one, but the most flexible, of the many complementary "probes" which can be used. Large related instruments are Synchrotrons, Integrated Laser Laboratories or High Power Lasers. A recent technological breakthrough is adding the Free Electron Lasers (FELs) capable not only of much higher brilliances but also of short time "flashes" opening the dynamic "filming" of atom-related properties. Photon sources are needed over a large range of "colours", from the low Infrared range up to the Hard X-Rays. High power lasers and Synchrotron light are also used to produce and study plasmas, e.g. the conditions for energy production by fusion, or to produce devices through photolithography.

Europe has a long tradition of excellence in the development and use of these sources and related instruments. Several new 3rd and 4th generation light sources are now in operation, under construction or planned in order to satisfy the rapidly increasing demand for beam time and new capabilities. The **European Synchrotron Radiation Facility (ESRF)** and its planned 10-year upgrade programme is the prime example of the evolution of the landscape in this field due to the continuous effort made by several EU countries.

The new capabilities offered by FELs will allow the exploration of a new terra incognita. State of the art 3rd generation synchrotron sources will be surpassed in peak intensity by 8-9 orders of magnitude and by 3-4 orders of magnitude in average intensity and short-time capabilities, pushing research to new frontiers and opening novel areas of research.

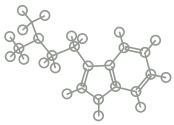
The first exciting results in atomic and molecular physics are already coming out from the operation of the *FLASH* facility in Hamburg (UV to soft X-ray range). In this range of photon energies, several other national projects of European relevance and with different characteristics are currently either under construction or in the design stage. Some of these are *FERMI@ELETTRA* (Italy), *MAX IV* (Sweden), and *PSI FEL* (Switzerland). Some other facilities are currently under consideration. **EuroFEL** is coordinating these activities aiming at integrating them into a unique distributed infrastructure.

The most notable success of the roadmap so far is certainly the progress of the **European XFEL**. The new facility is hosted in Hamburg, with the participation of 11 EU Member and Associated States. This has also been a pioneer project for developing new finance, governance and management models for European research infrastructures.

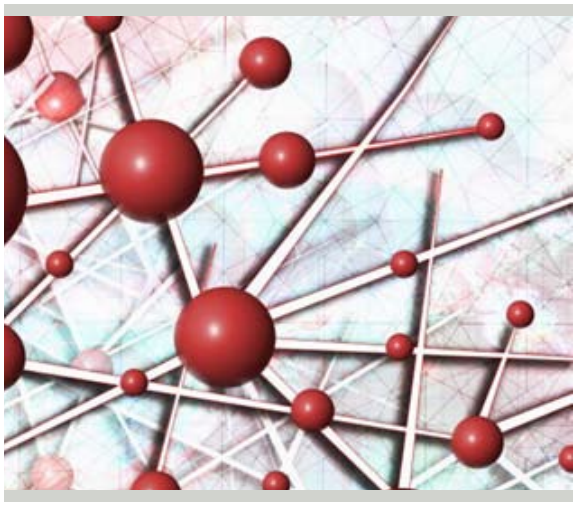
Neutron Sources.

The use of neutrons as a probe of matter is complementary to photons and they are unique for measuring and detecting very important aspects of materials and biological matter. Infrastructures for neutron spectroscopy use low energy neutrons. Their magnetic moment allows the detection of magnetic structures, while their small momentum allows the measurement of thermal and mechanical properties, and the differential measurement of hydrogen and deuterium enables important sites in biological matter to be studied. As a result of the great transparency of materials to neutrons, neutron radiography of large and thick machinery is possible, to study directly some important engineering aspects.

Beams of neutrons can be generated either by fission in nuclear reactors or by spallation by the impact of protons on targets. Europe has been leading in both types of neutron sources in the past 20-25 years, but new and increasing competition from the USA and Japan is rapidly eroding this advantage. Currently there are about 5000 users of neutron spectroscopy in Europe. This number is continuously increasing and attracting researchers in biology, biochemistry, biophysics and nanoscience.



To ensure that Europe has access to world leading facilities, the European Spallation Source (ESS) is a high priority together with the upgrade of the Institut Laue-Langevin (ILL). The European Spallation Source will become the world's most powerful neutron source. It will be particularly suited to elucidate complex biological structures and processes. Studies of biomaterials, foods, pharmaceuticals, and systems relevant to environmental health will also be possible in addition to the existing applications. Candidate countries to host the European Spallation Source are Hungary, Spain and Sweden. If a decision on the site of this 5-MW long-pulse source is taken in 2008 and construction starts in 2009, the facility may become operational in 2019/20. This will ensure that Europe's leading role in neutron scattering is kept and strengthened.



High magnetic fields.

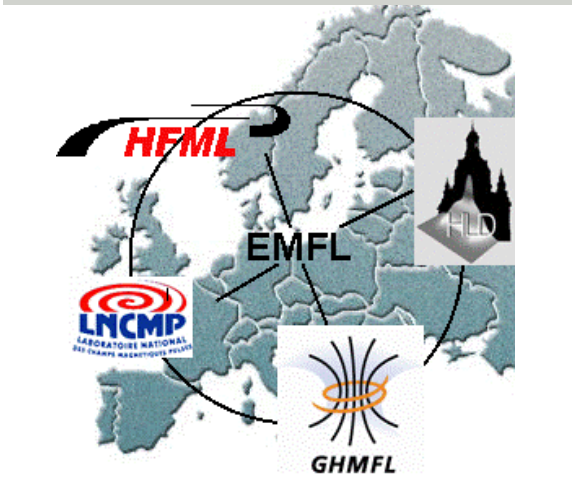
A powerful tool to modify and study materials is also provided by high magnetic field laboratories, operating both in continuous and pulsed modes. High magnetic fields are used in a variety of applications, ranging from materials research (including superconducting materials) to physics, chemistry and life science. The current plan to upgrade and integrate the four existing European high magnetic field laboratories into one distributed pan-European research infrastructure, the European Magnetic Field Laboratory (EMFL) will allow Europe to strengthen its competitiveness in this field of research.

The connection of this new infrastructure with the neutron and photon sources, in particular ILL and ESRF in Grenoble, is also currently under discussion and, if approved, it will provide a truly multi-source multi-disciplinary user instrument for the investigation of matter.

EMFL - European Magnetic Field Laboratory

The facility:

EMFL will be a dedicated magnet field laboratory providing the highest possible fields (both continuous and pulsed) to European researchers. It will be operated as a single distributed research infrastructure which integrates and upgrades the four already existing major European high magnetic field laboratories: the Grenoble High Magnetic Field Laboratory (GHMFL), the Laboratoire National des Champs Magnétiques Pulsés (LNCMP) in Toulouse, the Hochfeld-Magnetlabor Dresden (HLD), and the High Field Magnet Laboratory (HFML) in Nijmegen. EMFL will allow Europe to take the lead in the production and use of very high magnetic fields for scientific goals.



Background.

High magnetic fields, both static and pulsed, provide the most powerful tools available to scientists for the study, the modification and the control of the state of matter. They are extensively used in a variety of scientific domains, from physics and material science to chemistry and life sciences. Technological applications include the characterisation of superconductive materials. Europe has been leading the development of high magnetic fields and their use for science and technology. At present, however, the USA's National High Magnetic Field Laboratory (NHMFL), distributed over three sites is the leading facility in the field. For Europe to regain its competitiveness, it is urgently necessary to coordinate and upgrade Europe's high field activities to an effective size and efficiency comparable to that of the NHMFL.

What's new? Impact foreseen?

The multi-site EMFL will develop common magnet technology and magnets with the goals to take the lead in the production of very high fields and to develop unique experimental facilities worldwide. EMFL will coordinate the science and technology programs by developing complementary scientific specialisations at each of the four sites. HFML will concentrate on advanced spectroscopy through the unique combination with a FEL and the dedicated continuous 40 T low vibration hybrid magnet for local nano-spectroscopy, while GHMFL will house a record field hybrid magnet (50 T) with a new 40 MW installation. ESRF and ILL (Grenoble) are also planning a major upgrade of their installations by combining very high magnetic fields with neutrons and synchrotron radiation. To do so, they will cooperate with the EMFL to design, build and operate the necessary magnets for beam scattering and will share the new high power installation. HLD will fully exploit the coupling to the ELBE FEL for infrared spectroscopy and will develop magnets for the production of the highest available pulsed fields, while LNCMP will expand its activities in X-ray and visible spectroscopy, and strengthen its magnet materials development program. EMFL will also manage the scientific access of its users to all its installations, the selection of the proposals being made by an independent external Selection Committee.

>Timeline.

The construction of the new EMFL facility is expected to start in 2011 after a 2 years preparatory phase, and to last for 5 years. The facility should remain in operation for at least 15 years.

>Estimated costs.

Preparation costs:	10 M€.
Total construction costs:	~120 M€.
Operation costs:	8 M€/year additional, or 22 M€/year including existing budgets.
Decommissioning costs:	not applicable.

>Website: http://www.emfl.eu/

ESRF Upgrade

STARTED

The facility:

The European Synchrotron Radiation Facility (ESRF), located in Grenoble, France, is a joint facility set up by international agreement, supported and shared by 18 European countries and Israel. It operates the most powerful high energy synchrotron light source in Europe and brings together a wide range of disciplines including physics, chemistry and materials science as well as biology, medicine, geophysics and archaeology. There are many industrial applications, including pharmaceuticals, cosmetics, petrochemicals and microelectronics.

Background.

The ESRF's 6 GeV storage-ring light source built in the early nineties was the first insertion-device-based ("third generation") synchrotron radiation (SR) source. The ESRF has been extremely successful, both in terms of technical innovation and also where the very large volume of new and exciting science is concerned. With some 6200 scientific user visits each year, resulting in more than 1500 refereed publications, the ESRF is recognised as one of the world's most innovative and productive synchrotron light sources. This success is also measured by requests for beamtime from the community of users of the ESRF, consistently greatly exceeding the available beamtime.

What's new? Impact foreseen?

In order to maintain its leading role and to respond to the emerging scientific challenges, the ESRF is planning an ambitious Upgrade Programme, comprising (i) the extension of the experimental hall to enable the construction of new and upgraded beamlines with largely improved performance and new scientific opportunities, as well as improved infrastructures for the preparation of experiments, (ii) a programme of improvements of the accelerator complex, (iii) a strong supporting programme of engineering and technology developments, and (iv) the development of productive science and technology driven partnerships. The option for a joint high magnetic field laboratory with ILL and other European high field laboratories is also foreseen. The planned Upgrade will enable significant progress in fields such as nanoscience and nanotechnology, structural and functional biology, health, environment, energy and transport, information technology, and materials engineering. The Science case and the related technological challenges are laid out in the so-called Purple Book, available on the ESRF website.

>Timeline.

Pending a positive decision of the ESRF Council, ESRF is expected to start its 10-year upgrade programme in 2009.

>Estimated costs.

Preparation costs:	6.8 M€
Total construction costs:	capital costs 238 M€ (in 2008 prices), of which 77 M€ from the regular budget, recurrent costs 28 M€, personnel costs 21 M€.
Operation costs:	83 M€/year.
Decommissioning costs:	not applicable.

>Website:

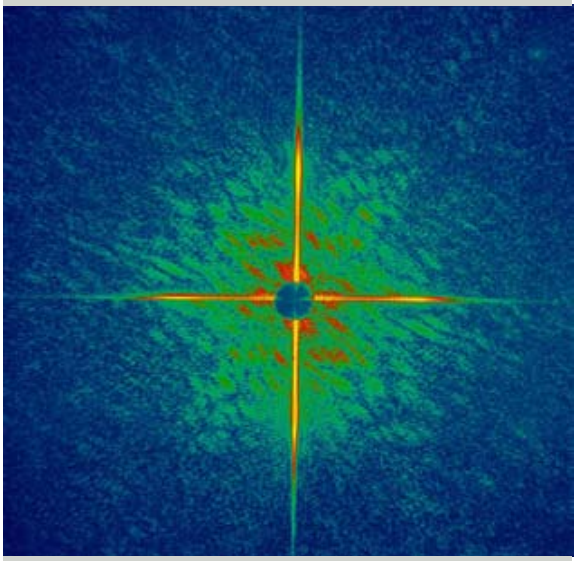
http://www.esrf.fr/

http://www.esrf.fr/AboutUs/Upgrade/purple-book/

EuroFEL (ex-IRUVX-FEL)

The facility:

Intense light beams with infrared to soft X-ray wavelengths are the major probe to study the electronic properties of matter, involving a very large user community. Free Electron Lasers (FELs) can now produce such beams of coherent, femtosecond light pulses with unprecedented intensities. The EuroFEL Consortium (previously called IRUVX-FEL) aims at integrating the national FEL based facilities currently in operation or emerging in Europe in a single, distributed and internationally open research infrastructure. The integration will exploit in the best way the complementary specifications and instruments of each facility for wide-ranging studies of matter by a large science community.



Background.

Recent technological advances have allowed developing new light sources based on free electron lasers for a broad spectral range from infrared to X-ray wavelengths. These sources produce collimated beams of femto-second pulses with high coherence and unprecedented brilliance to study the structure and dynamical behaviour of materials at the atomic level. The development of a set of complementary FEL sources as an integrated pan-European research infrastructure will give Europe a first-class infrastructure, unprecedented worldwide, complementing present synchrotron radiation sources and "table top" lasers.

What's new? Impact foreseen?

The interaction between matter and intense electromagnetic radiation in this spectral range is virtually unexplored. The photon beams of soft X-ray FELs have completely new qualities compared to those of synchrotrons and also exceed other sources based on conventional lasers. Europe has the unique chance of consolidating its world leadership in a field of highest relevance. Scientific challenges and opportunities will open for a wide range of scientific disciplines, ranging from nanosciences, materials and biomaterials sciences, plasma physics, molecular and cluster, femto- and nano-physics and chemistry, with various connections to life, environmental, astrophysical and earth sciences and the development of technologies ranging from micro electronics to energy. Some novel emerging synchrotron techniques, like holographic coherent imaging or ultra fast pump-probe studies will greatly benefit from the enhanced beam properties.

>Timeline.

The overall EuroFEL facility could be realised in the next 8–12 years. One facility, FLASH in Hamburg, is already in operation, another, FERMI in Trieste, is under construction, with an overall financing provided by the two hosting country and regional governments, the European Investment Bank and EU projects, for a total estimated cost, including design and preparatory phases, of about 300 M€. Assuming that all currently proposed FEL projects will be financed, the EuroFEL consortium may finally include up to eight facilities.

>Estimated costs.

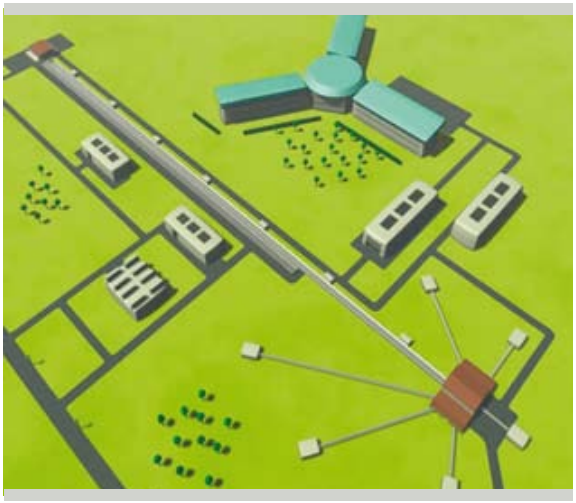
Preparation costs:	150-200 M€ (approximately 15% of the investment costs. EC contribution 5.7M€).
Total construction costs:	1200 - 1600 M€, including FLASH and FERMI, costing between about 150 and 200 M€.
Operation costs:	approximately 10% of the overall investment costs.
Decommissioning costs:	not applicable.

>Website: http://www.eurofel.eu

ESS - European Spallation Source

The facility:

The European Spallation Source will be the world's most powerful source of neutrons. Its built-in upgradeability (more than the initial 20 instruments, higher power) will make it the most cost-effective top tier source for 40 years or more. A genuine pan-European facility, it will serve a community of 5,000 researchers across many areas of science and technology.



Background.

Fine analysis of matter requires the complementary use of diverse "probes" and techniques: light scattering, neutron scattering, NMR, computer modelling and simulations and so on. Among them, neutrons are particularly important for soft and hard condensed matter, magnetism and biology as well as for nuclear physics. The intense beams of low energy neutrons which will available at the ESS will create entirely new opportunities for real time, real size, in situ, in vivo measurements, including movies of nano-scale events. They will allow understanding of the structure, dynamics and functions of increasingly complex systems covering the broad field of inorganic and organic materials as well as biomaterials.

What's new? Impact foreseen?

Neutron beams produced by reactors are inherently intensity-limited. The ESS R&D and design phase involving all major European Labs, has demonstrated the feasibility of MW-power spallation sources. In line with the global neutron strategy endorsed by OECD ministers in 1999, the US is now commissioning its facility SNS at Oak Ridge, and Japan is preparing for the first neutrons at J-PARC. In comparison with them, the long pulse configuration of ESS provides substantially higher power, maximum complementarities and the largest instrument innovation potential. Its performances guarantee a long-term world leading position. ESS will also offer new modes of operation and user support to maximally facilitate the industrial access, next to university and research lab users. The higher neutron flux will allow advanced and more effective investigations of ultra thin and laterally confined structures for ICT reading devices, active site structures in enzymes, technologies for storing hydrogen, multi-component complex fluids in porous media for tertiary oil production, the templating of nanostructures for catalysts, medical implants, pharmaceuticals, photonic materials etc. Requirements for novel detectors, instrument and software technologies will be additional drivers of innovation. ESS, a multifunctional facility with applications in many industries, will also have a marked regional impact.

>Timeline.

The science case and the preliminary baseline range (design and costing) are ready and allow formal negotiations to start in parallel to ongoing work to complete the detailed engineering design including detailed costing and optimisation. Sweden (leading a Scandinavian consortium), Spain and Hungary are presently candidates to host the ESS. Further characterisation of the three sites is currently underway, assisted by ESFRI. Assuming that construction will start in 2009, first neutrons will be delivered in 2017 and the facility will be opened for access in 2019/2020.

>Estimated costs.

Preparation costs:	30 M€.
Total construction costs:	1300 M€.
Operation costs:	110 M€/year.
Decommissioning costs:	300 M€.

>Website: http://neutron.neutron-eu.net/n_ess

European XFEL

STARTED

The facility:

The European X-ray Free Electron Laser, which is being built in Hamburg, Germany, will be a world leading facility for the production of intense, short pulses of X-rays for scientific research in a wide range of disciplines.

Background.

The European X-ray Free Electron Laser (European XFEL) project foresees the construction in Hamburg, Germany of a new international user facility for the production and scientific use of ultra-bright and ultra-short pulses of spatially coherent hard X-rays. The facility comprises a 1.7 km long superconducting linear accelerator accelerating electrons up to energy of 17.5 GeV, which will distribute up to ~30,000 electron bunches per second, with an energy of 17.5 GeV, into three undulators. These will generate spatially coherent X-ray radiation pulses shorter than 100 fs in duration and with peak power exceeding 10 GW, in a wavelength range from 0.1 nm to 1.6 nm (or even longer at lower electron energy). An additional set of two undulators can later on be added to generate hard X-rays down to 0.01 nm wavelength by spontaneous emission. The facility includes an initial set of six (expandable to ten) experimental stations with state of the art equipment for the scientific use of the radiation.

What's new? Impact foreseen?

It is anticipated that the availability of X-ray pulses with peak brilliance of up to nine orders of magnitude greater than existing 3rd generation light sources shall allow the performance of presently impossible and potentially revolutionary experiments in a variety of disciplines from condensed matter and materials physics to nanoscience, from plasma physics to chemistry and to structural biology. The European XFEL will provide a new tool for many different research fields. The detailed understanding of chemical reactions and the way how molecular machines work will be essential for future drug and material design. The big leaps in brilliance and pulse duration have already triggered the development of novel detector technologies and high power optical lasers, which could be expected to be drivers of further innovation. The European XFEL will use a new superconducting technology to accelerate electrons at high repetition rate, enabling a combination of high peak and high average brilliance. This technology is expected to be the basis of many future accelerators. With the experience gained with the realisation of the XFEL, industry in Europe could achieve a world leadership in these technologies.

>Timeline.

Starting in 2004, the Memorandum of Understanding (MoU) on the preparatory phase of the European XFEL has been signed by representatives of 11 EU Member States, China, Russia and Switzerland. The start of the construction was formally announced in a communiqué signed by the representatives of ten MoU partner states, the Free and Hanseatic City of Hamburg and the Federal State Schleswig-Holstein on 5 June 2007. The contracts for the main civil construction will be signed in November 2008, while the signature of the intergovernmental convention is scheduled in early 2009. The accelerator and the first six experimental stations will be commissioned starting in 2014.

>Estimated costs.

At 2005 prices:

Preparation costs:	39 M€.
Total construction costs:	1043 M€ (including commissioning).
Operation costs:	84 M€/year
Decommissioning costs:	100 M€.

>Website: http://www.xfel.eu

ILL 20/20 Upgrade

STARTED

The facility:

The reactor-based laboratory at the Institut Laue Langevin (ILL) is recognised as the world's most productive and reliable source of slow neutrons for the study of condensed matter, and its overall upgrade is the most cost-effective response in the short to medium term to users' requirements.

Copyright: ILL/Artechnique

Background.

The ILL has been and remains a European success story, having been set up with a full complement of beam-lines and experimental instruments from its beginning in the mid 70's, and continuously improved. The near-siting of the European Synchrotron Research Facility has added value for European users offering them access to complementary techniques and joint support laboratories at the Grenoble site. Recently the second phase of a wide-ranging upgrade programme – the Millennium Programme – began; this will ensure world level competitiveness and scientific value for the international user community. The programme includes the optimisation of the neutron source and its moderators, the neutron delivery (guides and beam tubes) and the neutron instrumentation. Access to new scientific areas will be strengthened through enhanced support facilities for users. The first such facility, the Partnership for Structural Biology, a joint project together with ESRF and three other laboratories, opened in November 2005. It provides special services, such as the growing of deuterated single crystals, for visiting research teams using the neutron and synchrotron instruments on site.

What's new? Impact foreseen?

The renewal of the neutron production and instrumentation of the Institut Laue-Langevin in Grenoble, the so-called 20/20 plan, will give a longer perspective of competitiveness. Added value is given by the partnerships for materials science and engineering, and soft condensed matter. A High Magnetic Field laboratory in collaboration with ESRF will support instruments on both the neutron and the synchrotron radiation sources. Specific measures will be implemented to facilitate Technology Transfer.

>Timeline.

The upgrade will be implemented in two consecutive 5 years phases from 2007 to 2012 and from 2013 to 2017.

>Estimated costs.

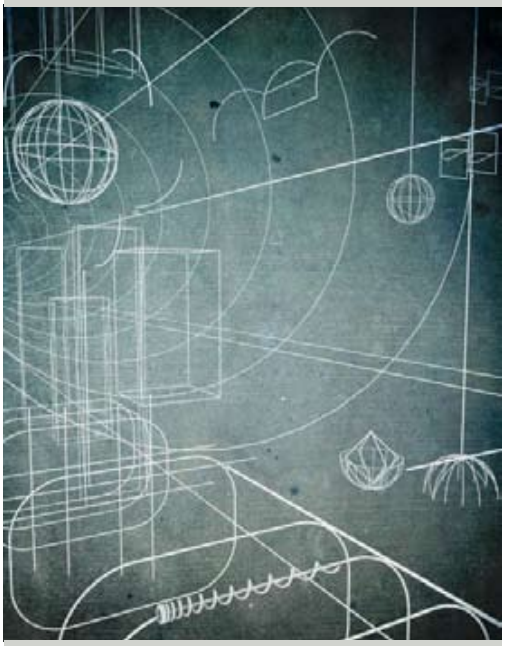
Preparation costs:	6.2 M€.
Total construction costs:	171 M€, including 15M€ of regional and local government funding towards additional infrastructural aspects for the proposed joint site together with ESRF.
Operation costs:	5 M€/year additional to the current operation costs.
Decommissioning costs:	under definition.

>Website: http://www.ill.eu/about/future-planning/perspectives-opportunities/

>Physical Sciences and Engineering

Physical sciences deal with phenomena at all scales and complexities, from the extremes of the universe to the smallest elementary particles. Through the years, scientific progress has generated many new fundamental questions which are today on the agenda of astrophysics, astroparticle physics, particle and nuclear physics, and are in many ways interconnected.

In the course of these developments, the facilities for fundamental physics and astronomy have become much larger, technically more complicated and expensive. More than ever before it has become a necessity to join intellectual and financial resources to realise these projects. Such facilities drive the development of new technologies and new ways of working for instance in ICT and in the applications of nanotechnology and superconductivity.



Astronomy and Astroparticle Physics

Europe has a long tradition in astronomy with a strong community working across a diverse range of fields from studies of the interaction of the solar wind with the Earth's upper atmosphere to cosmology.

Discoveries in recent years have raised new fundamental problems. These include the nature of dark energy and dark matter, the emergence of the first stars and galaxies in the universe and their evolution, the description of gravity, and planet formation around other stars. To tackle these and other questions new instruments are required to provide data across the electromagnetic spectrum:

- *Ground-based optical astronomy* can now use as largest instruments a set of 8-10 m telescopes, but it has become clear that the challenge posed by the new fundamental questions requires still a larger collecting area and stronger angular resolution. The European Extremely Large Telescope (E-ELT) is the follow-up project of the current generation of optical telescopes. With segmented mirrors and built-in adaptive optics, it is feasible to build a 40-m class telescope. Locations that are currently being characterised as possible sites of the E-ELT include Cerro Macón (Argentina), North Paranal and Vizcachas (Chile), Aklim (Morocco) and La Palma (Spain).

- *Neutrino astronomy:* The Cubic Kilometre Neutrino Telescope (KM3NeT) will consist of thousands of optical sensors distributed in a volume of about one cubic kilometre in the depth of the Mediterranean Sea. The sensors detect the light which is produced in the water by charged particles originated from neutrinos. It aims to monitor the universe continuously together with the IceCube Neutrino Detector, currently under construction at the South Pole.

- In *radio astronomy* the next generation telescope should be the Square Kilometre Array (SKA). The SKA will have a collecting area of one million square metres distributed over a distance of at least 3000 km. This area, necessary to collect the faint signals from the early universe, will result in a 100 times higher sensitivity compared to existing facilities. The radically new concept of an "electronic" telescope will allow very fast surveys. Candidate sites for SKA are Australia/New Zealand and Southern Africa.

- In *gamma-ray astronomy* a similar role will be played by the Cherenkov Telescope Array (CTA). The pioneering Cherenkov telescopes HESS and MAGIC have observed a multitude of gamma ray sources both in our galactic centre and outside our galaxy. The CTA will greatly extend the reach of these two projects and allow for further exciting scientific discoveries. The CTA will be deployed in two locations, one in the southern hemisphere and one in the northern hemisphere (likely sites are in Namibia and in the Canary Islands).

Particle Physics and Space Physics

Particle physics is entering in a new and exciting era of discovery, exploring new domains and probing the deep structure of space-time. European particle physics is based on strong national institutes, universities and laboratories and on the CERN Organisation.

The CERN Council has adopted a strategy for the field and follows up its implementation in regular European Strategy Sessions. An update is foreseen for 2011. The current strategy is:

- The Large Hadron Collider (LHC) at CERN, now starting, will be the energy frontier machine for the foreseeable future and it has the highest priority to fully exploit its physics potential. Depending on the nature of the discoveries made at the LHC, higher-statistics studies of these phenomena would naturally call for an increase in luminosity. This upgrade – referred to as Super-LHC – should increase the luminosity by a factor ten.
- It is vital to strengthen the advanced accelerator R&D programme in Europe, providing a strong technological basis for future projects in particle physics.
- It is fundamental to complement the results of the LHC with measurements at an electron-positron linear collider. Such a linear collider will provide a unique scientific opportunity at the precision and energy frontiers. This programme can be carried out by the International Linear Collider (ILC) or, if multi-TeV energies are needed, by a novel design called the Compact Linear Collider (CLIC) which has the potential to deliver such energies. For essentially every new physics scenario involving particles in the linear collider energy range, detailed and very promising research programmes have been formulated. The linear collider studies are in the R&D and these studies will, together with results from the LHC, guide the way towards realisation.
- Neutrino physics opens another exciting window to study physics beyond the standard model. Recent measurements of neutrino oscillations and masses, and the possibility of observing CP violation in this sector, point forward to the need of constructing more advanced neutrino facilities, and design studies are ongoing. Which route to take, depends on the result of accelerator R&D, and on results from experiments now starting.
- Several important experiments take place and are planned in the overlap region between Particle and Astroparticle Physics, or between Particle and Nuclear Physics. Examples of such experiments can be found in Europe's four world-class deep underground laboratories: Boulby (UK), Canfranc (Spain), Gran Sasso (Italy) and Modane (France). These facilities study neutrinos – including in some cases long baseline experiments with accelerator neutrinos - and search for dark matter and proton decays.
- New initiatives and plans are being developed in the field of flavour physics where the Super-B facility at the INFN National Laboratory of Frascati is a possibility being pursued.

In the field of **Space Science**, the European Space Agency has outlined the future plans for future space missions in its Cosmic Vision paper.

Nuclear Physics

Modern nuclear physics has two main goals. At the larger scale one wants to understand the limits of nuclear stability by producing exotic nuclei with vastly different numbers of neutrons and protons. At the smaller scale one wants to explore the substructure of the constituent neutrons and protons, for it is in the interaction of their constituent parts that the ultimate description of nuclei must lie. Such a description is much needed in order to reduce uncertainties on the nuclear data upon which design and operation of fission reactors are based.

There are two approaches to producing radioactive beams – the "In-Flight Fragmentation" and the "Isotope-Separation On-Line" (ISOL) techniques. The In-Flight production technique is fast and can produce the shortest-lived radioactive nuclei, whereas the ISOL technique can provide more intense and better controlled beams for detailed studies. So both techniques are complementary.

The leading in-flight facility is the **Facility for Antiproton and Ion Research (FAIR)**, which will soon become an international research centre in Darmstadt (Germany). The technical plan for the first stage and the legal documents will allow starting construction in 2009.

SPIRAL2, a major expansion of the SPIRAL facility at GANIL in Caen (France), will help to maintain the European leadership in the ISOL development. It is an essential step on the road to EURISOL, the ultimate ISOL facility. The objective is to realise the project with international participation.

Engineering

Within engineering, nanotechnology is the manipulation or self-assembly of individual atoms, molecules, or molecular clusters into structures to create material and devices with new properties. Nano-science and technology is and will be one of the major research and development areas for the coming decade. As a result of its multidisciplinary nature, the question of research infrastructure needs is quite different from that of other fields. A broad range of often smaller but dedicated and complementary equipment for nano-scale synthesis is needed in order to be able to perform all processing and characterisation steps needed.

The **Pan-European Research Infrastructure for Nano-Structures (PRINS)** will be a distributed facility of leading centres, smaller centres and research groups from all countries to form a large pan European infrastructure dealing with the ultimate silicon and heterogeneous integration. This distributed facility will collaborate with the European nano-electronics community through the strategic Joint Technology Initiative "ENIAC". A significant amount of the total cost is expected to be raised by industry in a public-private partnership.

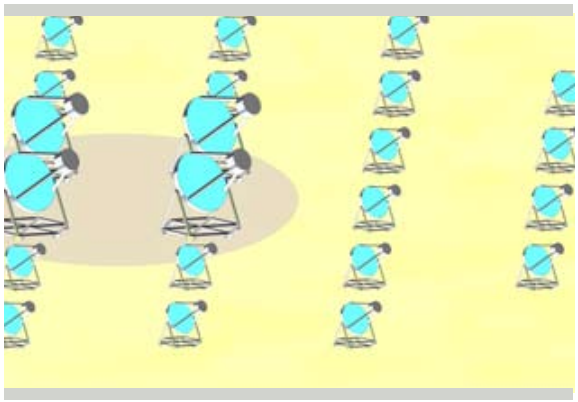
High Intensity Lasers

The frontier of laser science is progressing at an extremely steep gradient in many different directions, opening new perspectives in basic research (ultra-relativistic intensity regime) as well as in applied areas (particle acceleration, development of efficient, compact secondary sources of electrons, ions and photons). The high power, short pulse laser installation **ELI** appears thus appropriate to maintain or even increase the European leadership in this very rapidly evolving domain. Important societal applications will greatly benefit from it (compact accelerators, hadron and radiation therapies, medical imaging, etc.).

CTA – Cherenkov Telescope Array

The facility:

The Cherenkov Telescope Array will be an advanced facility for ground-based high-energy gamma-ray astronomy. With two sites, in both the southern and northern hemispheres, it will extend the study of astrophysical origin of gamma-rays at energies of a few tens of GeV and above. It will provide the first complete and detailed view of the universe in this part of the radiation spectrum and will contribute towards a better understanding of astrophysical and cosmological processes.



Background:

In the last years the present generation of imaging atmospheric Cherenkov telescopes have allowed the first detailed observations of the sky using gamma-rays of energies of a TeV and above. They have revealed a sky unexpectedly rich in gamma-rays features such as extended sources with complex and resolved structure lining the central band of the Milky Way, and extragalactic sources at very large distances with some showing very fast variability on a time scale of minutes. Extending these observations is an important future avenue of inquiry for astronomy.

What's new? Impact foreseen?

The proposed facility will consist of arrays of Cherenkov telescopes which will increase the sensitivity for observations of distant or faint objects by another order of magnitude, provide better angular resolution and lead to improved images of the structure of gamma-ray sources, allow a wider field of view enhancing all-sky survey capability and the study of transient phenomena, enhance all sky survey capability, and have wide and uniform coverage for gamma-ray energies from tens of GeV to hundreds of TeV. The array will be built at two separate sites, one in the southern hemisphere with wide gamma-ray energy range and high resolution to cover the plane of the Milky Way, and the second in the northern hemisphere specialised for lower energies, which will focus on extragalactic and cosmological objects.

The CTA will investigate cosmic non-thermal processes, in cooperation with observatories in other wavelength ranges of the electromagnetic radiation spectrum, as well as with those using other messenger types (i.e. neutrino telescopes, cosmic ray arrays). This multi-messenger approach to astronomy will lead to deeper understanding of major astrophysical processes and of the evolution of the universe.

The CTA facility will be operated as a proposal-driven observatory, with a Science Data Centre providing transparent access to data, analysis tools, and user training.

>Timeline.

Technical design and prototype construction 2006-2011; construction 2012-2017; operation 2018 (partial operation starting after 2013); expected lifetime is 20-30 years.

>Estimated costs.

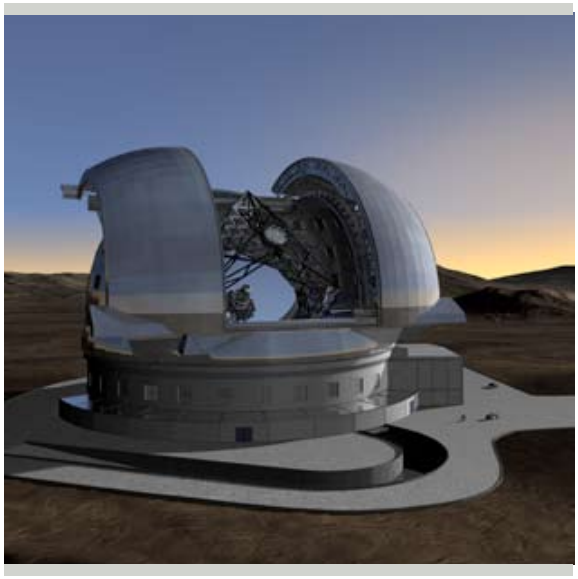
Preparation costs:	~8 M€.
Total construction costs:	~150 M€ (100 M€ for the southern site, 50 M€ for the northern site).
Operation costs:	~10 M€/year.
Decommissioning costs:	~10 M€.

>Website: http://www.mpi-hd.mpg.de/CTA

E-ELT – European Extremely Large Telescope

The facility:

ELTs are seen world-wide as one of the highest priorities in ground-based astronomy. They will vastly advance astrophysical knowledge allowing detailed studies of inter alia planets around other stars, the first objects in the Universe, super-massive Black Holes, and the nature and distribution of the Dark Matter and Dark Energy which dominate the Universe. The 42m European Extremely Large Telescope (E-ELT) project will maintain and reinforce Europe's position at the forefront of astrophysical research.



Background:

Extremely Large Telescopes allow the next major step in addressing the most fundamental properties of the universe. All areas of known astronomy, from studies of our own solar system to the farthest observable objects at the edge of the universe, will be advanced by the enormous improvements attainable in collecting area and angular resolution. Following a resolution by the ESO Council in 2004 instructing ESO to ensure European leadership in ground-based optical/near infrared astronomy in the ELT era, ESO completed at the end of 2006 the Reference Design of the 42 meter European Extremely Large Telescope (E-ELT). In parallel, the E-ELT's scientific case has been developed and is being refined by the astrophysical community through the EC-funded OPTICON program as well as by various ESO Committees. Major enabling technologies are being pursued by European research institutes and high-tech companies within the ELT Design Study FP6 project, with ESO and the Commission as the main funders. These efforts are conducted in close contact with the other similar projects all around the world.

What's new? Impact foreseen?

Astronomy is a technology-enabled science. Recent technology developments, especially in real-time control of complex systems, now allow the next generation of telescopes to be built. Improvements in light collection and spatial resolution, needed to go from the present 8-10 metres to over 30 metres in diameter, will improve on current limits by tens to hundreds of times, providing the critical increase in sensitivity and resolution to solve outstanding scientific questions and almost certainly open new ones. Astronomy is known to be the most effective topic attracting young people to science and technology careers. Astronomical telescopes, being large precision opto-mechanical systems in hostile environments, develop advanced technologies in many state-of-the-art areas with spin-offs ranging from medicine to image data processing.

>Timeline.

ESO is currently undertaking a 3-year 57 M€ detailed design study. The preparatory and design phase of the E-ELT will last until 2009, with final site selection in 2010. Construction is expected in the period 2010-2017.

>Estimated costs.

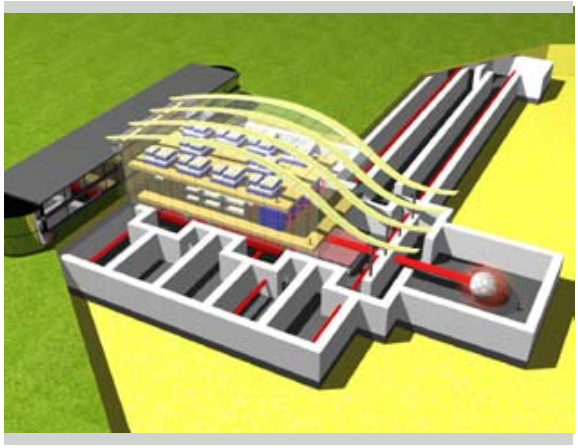
Preparation costs:	100 M€ is covered mainly by ESO, with additional funding by the EU.
Total construction costs:	950 M€, including 1st instrument complement.
Operation costs:	30 M€/year.
Decommissioning costs:	to be evaluated following negotiations with the host country.

>Website: http://www.eso.org/projects/e-elt/

ELI – Extreme Light Infrastructure

The facility:

ELI will be an international research infrastructure open to scientists dedicated to the investigation and applications of laser matter interaction at the highest intensity level, i.e. more than 6 orders of magnitude higher than today's state of the art. ELI will comprise three branches: ultra high field science that will explore laser matter interaction up to the nonlinear QED limit including the investigation of pair creation and vacuum structure; attosecond laser science designed to conduct temporal investigation at the attosecond scale of electron dynamics in atoms, molecules, plasmas and solids; lastly, the high energy beam facility devoted to the development of dedicated beam lines of ultra short pulses of high energy radiation and particles up to 100GeV for users.



Background:

Laser intensities have increased by 6 orders of magnitude in the last few years. These are now so large that the laws of optics change in a fundamental way. This new optics field is called relativistic optics. Among the important by-products of this field are the generation of particle, x-ray and gamma-ray beams. The wealth of discoveries made in the relativistic regime justifies going further to the ultra-relativistic regime. One important aspect of ELI is the possibility to produce ultra-short pulses of high energy photons, electrons, protons, neutrons, muons, and neutrinos in the attosecond and possibly zeptosecond regimes on demand. Time-domain studies will allow unravelling the attosecond dynamics in atomic, molecular physics and plasma physics.

What's new? Impact foreseen?

ELI will be the first facility in the world dedicated to laser-matter interaction in the ultra-relativistic regime, providing unprecedented intensity levels. It will be the gateway to new regimes in physics. At the same time, it will also promote new technologies such as relativistic microelectronics with the development of compact laser-accelerators delivering >100GeV particles and photon sources. ELI will have a large societal benefit in medicine with new radiography and hadron therapy methods, in material sciences with the possibility to unravel and slow down the ageing process in a nuclear reactor and in environment by offering new ways to treat nuclear waste. The completed machine will provide laser pulses with a peak power above 200PW, a power level 200000 higher than the power of the entire European electric grid, but only for a millionth of a billionth of a second.

>Timeline.

ELI is in its preparatory phase for the next two years, followed by two years of design study and a five year construction period. The conceptual design calls for construction in three stages. The first pulses at the 100 TW level will be available for users at the end of the second construction year, while the next stage, providing pulses with a few PW level, will operate from the fourth construction year.

>Estimated costs.

Preparation costs:	85 M€.
Total construction costs:	400 M€.
Operation costs:	50 M€/year.
Decommissioning costs:	30 M€.

>Website: www.eli-laser.eu

FAIR – Facility for Antiproton and Ion Research

STARTED

The facility:

FAIR will provide high energy primary and secondary beams of ions of highest intensity and quality, including an "antimatter beam" of antiprotons allowing forefront research in five different disciplines of physics. The accelerator facility foresees the broad implementation of ion storage/cooler rings and of in-ring experimentation with internal targets. Two superconducting synchrotrons will deliver high intensity ion beams up to 35 GeV per nucleon for experiments with primary beams of ion masses up to Uranium and the production of a broad range of radioactive ion beams.

Background:

The concept for the Facility for Antiproton and Ion Research (FAIR), planned for construction at the GSI Laboratory in Darmstadt, Germany, has been developed by international working groups. In 2001, GSI, together with a large international science community, presented a Conceptual Design Report (CDR). Following an in-depth evaluation by the German Wissenschaftsrat (the science advisory committee of the German government) and its recommendation to realise the facility, the German Federal Government announced in February 2003 approval of the project, with Germany providing up to 75% of the needed funding. Since 2004 the FAIR project is governed by the International Steering Committee with 14 countries participating as members, which declared to participate in the construction and operation of the facility. The FAIR project developed significantly since the CDR. About 2500 scientists from 44 countries submitted Letters of Intent in 2004 and Technical Proposals in 2005 for the experiment programs at FAIR. Significant R&D has been carried out and detailed design has been developed. A large fraction of this effort is funded by the European Commission. The 3500-page long FAIR Baseline Technical Report was published in April 2006.

What's new? Which impacts?

FAIR has a broad scientific scope allowing forefront research in nuclear structure physics and nuclear astrophysics with radioactive ion beams, QCD studies with cooled beams of anti-protons, physics of hadronic matter at the highest baryon density, plasma physics at very high pressure, density and temperature, atomic physics and applied sciences. It will provide the European science communities with a world-wide competitive facility. The central program, nuclear physics, is in its totality of first class internationally. FAIR is also unique in areas such as highly-compressed intense heavy-ion beams for plasma physics, and in its unparalleled research program with cooled antiproton beam and internal-target storage-ring capabilities for QCD studies.

>Timeline.

The start of the construction is projected for 2009, following the political declaration launching the project in November 2007 and the expected signature of the legal documents and establishment of the FAIR Company with international participation towards the end of 2008. The initial agreement is expected to cover a start-up version of the facility, while additional contributions (also from additional partner countries) will be sought to complete the full facility by 2015. The full performance with the parallel operation of all experimental programs will be reached in 2016.

>Estimated costs.

Preparation costs:	~120 M€.
Total construction costs:	~1187 M€ (total investment, plus costs for manpower - price index 2005).
Operation costs:	120 M€/year (price index 2005).
Decommissioning costs:	to be estimated.

>Website: http://www.gsi.de/fair/index_e.html

KM3NeT – Kilometre Cube Neutrino Telescope

The facility:

KM3NeT will be a deep-sea research infrastructure in the Mediterranean Sea hosting a cubic-kilometre sized deep-sea neutrino telescope for the astronomy based on the detection of high-energy cosmic neutrinos and giving access to long-term deep-sea measurements.

Background:

Since they are not deflected and can travel cosmological distances without absorption, neutrinos are ideal messengers for studying the highest-energy, most violent processes in the universe. However, due to their weak interaction with ordinary matter, huge detectors are required to measure them. Several first generations of such neutrino telescopes in the Mediterranean Sea are currently in operation or under construction. However, only future installations of cubic-kilometre size will exploit the full scientific potential of neutrino astronomy. These installations can be built in synergy with environmental observation underwater stations such as EMSO.

What's new? Impact foreseen?

The KM3NeT neutrino telescope will be the leading European facility for neutrino astronomy. It will be the only deep-sea installation of this size in the world and only be complemented by the US-led IceCube project currently under construction in the Antarctic ice at the South Pole. Compared to IceCube, KM3NeT will determine direction and energy of the neutrinos with higher precision, it will have a significantly higher sensitivity for source detection and it will have the major advantage of being able to observe neutrinos originating from the central region of the Milky Way. The design of KM3NeT poses substantial challenges concerning e.g. photo-detection, data acquisition and processing, deep-sea technology, installation and maintenance procedures, cost effectiveness and stability of operation. These issues are being addressed in a FP6 design study (2006-9), building on technology at the forefront of science. KM3NeT will be a truly interdisciplinary research infrastructure: It will provide access to neutrino observations for the astronomy, astrophysics, astroparticle and particle physics communities and, in addition, allow for long-term measurements in the deep-sea environment that are of utmost interest for biologists, geophysicists and oceanographers.

>Timeline.

By 2009, the design study will culminate in a Technical Design Report laying the technical foundations for the construction of the KM3NeT infrastructure. A 3-year preparatory phase project started in March 2008. Thereafter, 4 to 5 years time will be required to establish funding, for industrialisation and deployment.

>Estimated costs.

Preparation costs:	Design Study 20 M€; preparatory phase 12 M€.
Total construction costs:	a solid estimate of the construction cost will result from the design study; the objective is to achieve a price tag of 200 M€ or below for a cubic-kilometre installation (salaries not included).
Operation costs:	~5 M€/year.
Decommissioning costs:	~5 M€.

>Website: http://www.km3net.org/

PRINS – Pan-European Research Infrastructure for Nanostructures

The facility:

The Pan-European Research Infrastructure for Nano-Structures (PRINS) is the research infrastructure arm of a broader initiative, the ENIAC European Technology Platform. PRINS will bridge the area between research and market-driven applications and provide Europe with the ability to master the revolutionary transition from Microelectronics to Nano-electronics, i.e. down to the level of individual atoms.

Background:

PRINS has been conceived as a distributed infrastructure based essentially in 3 European countries (Belgium, France and Germany) that will jointly address the new challenges in a coordinated and complementary way. Academic access to these pre-existing centres of excellence (IMEC, CEA-LETI and Fraunhofer Group for Microelectronics, respectively) will be put under a common umbrella. These three scientific and technical integration centres will be supported by a complementary network of flexible rapid-prototyping laboratories. Their role will be the validation of innovative device and materials steps in the nano-scale CMOS and beyond-CMOS areas as well as More-than-Moore.

What's new? Impact foreseen?

PRINS will enable European research to move into the ultimate scaling of electronic components, the combination of digital signal processing with other types of functionality, the exploration of novel device concepts and the integration of components and materials into Systems in a Package (SiP). PRINS will contribute to realising the goals of the ENIAC Strategic Research Agenda. It will bring together an unprecedented array of equipment and know-how in topics like high-resolution lithography, advanced process steps and modules, electronic systems integration, imaging devices, silicon-based micro-systems, and miniaturized devices addressing the nano-bio convergence. It will give a boost to European RTD performance in the area of Nano-electronics and combined Nano-Structures. The applications that PRINS will generate will serve the future demands of European society, will increase high-skilled employment, will reinforce the competitiveness of European industry and will secure global leadership in high-tech multidisciplinary research.

>Timeline.

The PRINS project is now executing the preparatory phase, to define the operational modes and required infrastructure at the three integration centres. The latter will be built in a modular way in the period 2009-2015. The infrastructure will be partly operational in 2009, while major additional research equipment will be brought in until 2015, in response to quickly changing needs and new technologies becoming available.

>Estimated costs.

Preparation costs:	3.5 M€.
Total construction costs:	1400 M€.
Operation costs:	300 M€.
Decommissioning costs:	not applicable.

>Website: www.prins-online.eu

SKA – Square Kilometre Array

GLOBAL

The facility:

The Square Kilometre Array will be the next generation radio telescope. With an operating frequency range of 70 MHz - 25 GHz and a collecting area of about 1.000.000 m², it will be 50 times more sensitive than current facilities. With its huge field-of-view it will be able to survey the sky more than 10,000 times faster than any existing radio telescope. The SKA will be a machine that transforms our view of the universe.

Background:

The development of radio telescopes and radio interferometers over recent decades has helped drive a continuous advance in our knowledge of the universe, its origins and evolution, and the enormously powerful phenomena that give rise to star and galaxy formation. Radio astronomy also provides one of the most promising search techniques in humanity's quest to determine if life exists elsewhere in the universe.

What's new? Impact foreseen?

The huge collecting area of the SKA will result in sensitivity 50 times greater than any existing interferometer, a requirement to see the faint radio signals from the early universe. The radically new concept of an "electronic" telescope with a huge field-of-view and multiple beams will allow very fast surveys. The SKA will be the most sensitive radio telescope ever built and will attack many of the most important problems in cosmology and fundamental physics. Observations of pulsars will detect cosmic gravitational waves and test Einstein's General Theory of Relativity in the vicinity of black holes. The SKA will study the distribution of neutral hydrogen (the most common element in the universe) in a billion galaxies across cosmic history, thus making it possible to map the formation and evolution of galaxies, study the nature of dark energy and probe the epoch when the first stars were born. The SKA will be the only instrument that will map magnetic fields across the universe, allowing us for the first time to study the nature of magnetism. Last but not least, the SKA will study the formation of planetary systems and address the question "does life exist elsewhere in the Universe?"

>Timeline.

Preliminary design and technology development: 2000-2007; costed system design from 2008-2012. Phase 1 construction and first data: 2012-2016, completion of the full SKA at low and mid frequencies (up to 10 GHz): 2016-2020; construction of high frequency segment: post-2020.

>Estimated costs.

Preparation costs:	~150 M€.
Total construction costs:	~1500 M€ (low and mid frequencies). Cost of the high frequency segment to be defined.
Operation costs:	100-150 M€/year.
Decommissioning costs:	to be defined during the preparatory phase.

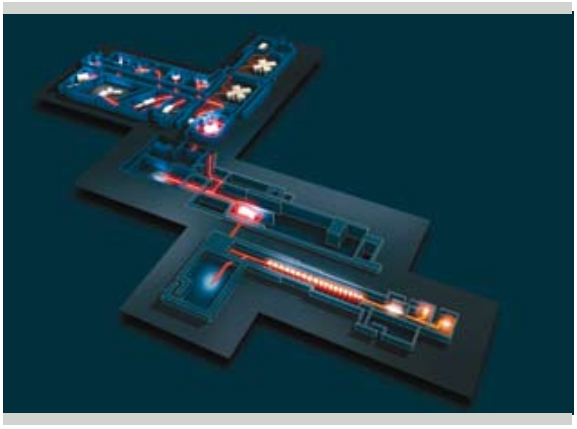
>Website: http://www.skatelescope.org/

SPIRAL2

STARTED

The facility:

SPIRAL2 is a new European facility to be built at GANIL laboratory in Caen, France. The project aims at delivering stable and rare isotope beams with intensities not yet available with present machines. SPIRAL2 will reinforce the European leadership in the field of nuclear physics based on exotic nuclei.



Background:

The frontier of nuclear physics is advancing through the study of nuclear reactions between diverse types of stable and rare isotope ions, covering the widest possible range of different nuclei and energies. The approach of SPIRAL2, complementary to that of FAIR, is based on the ISOL (Isotope Separation On Line) technique and aims at two orders of magnitude increase of the rare isotope beams available for nuclear physics studies. The SPIRAL2 facility will produce beams of excellent optical quality for moderately short-living radioactive nuclei in the energy range from 30keV to 20 MeV/nucleon.

What's new? Impact foreseen?

The scientific programme, prepared by a team of six hundred world class specialists, proposes the investigation of the most challenging nuclear and astrophysics questions aiming at the deeper understanding of the nature of matter. SPIRAL2 will contribute to the physics of nuclei far from stability, nuclear fission and fusion based on the collection of unprecedented detailed basic nuclear data, to the production of rare radio-isotopes for medicine, to radio-biology and to material science. The SPIRAL2 project is an intermediate step toward EURISOL, the most advanced nuclear physics research facility presently imaginable and based on the ISOL principle. It is expected that the realisation of SPIRAL2 will substantially increase the know-how of technical solutions to be applied not only for EURISOL but also in a number of other European and world projects. The current negotiations with international partner countries should allow them to join the ongoing construction and future operation phase, turning the present GANIL facility into a fully international legal entity.

>Timeline.

The construction will last about seven years (2006-2013).

>Estimated costs.

Preparation costs:	8.8 M€.
Total construction costs:	196 M€.
Operation costs:	~6.6 M€/year.
Decommissioning costs:	10 M€.

>Website:

http://www.ganil.fr/research/developments/spiral2/

Appendix A

Laplace Transforms

Table of Laplace Transforms	
$f(t)$ for $t \geq 0$	$\mathcal{L}(f)$
1	$\dfrac{1}{s}$
e^{at}	$\dfrac{1}{s-a}$
t^n	$\dfrac{n!}{s^{n+1}}$ $(n = 0, 1, \ldots)$
$\sin at$	$\dfrac{a}{s^2+a^2}$
$\cos at$	$\dfrac{s}{s^2+a^2}$
$\sinh at$	$\dfrac{a}{s^2-a^2}$
$\cosh at$	$\dfrac{s}{s^2-a^2}$
$H_a(t)$	$\dfrac{e^{-as}}{s}$
$\delta(t-a)$	e^{-as}
$f'(t)$	$s\mathcal{L}(f) - f(0)$
$f''(t)$	$s^2\mathcal{L}(f) - sf(0) - f'(0)$

where

$$H_a(t) = \begin{cases} 0 & t \leq a \\ 1 & t > a \end{cases}$$

TABLE OF LAPLACE TRANSFORMS
Revision G

By Tom Irvine
Email: tomirvine@aol.com

November 24, 2010

Operation Transforms		
N	F(s)	f (t) , t > 0
1.1	$Y(s) = \int_0^\infty \exp(-st)y(t)dt$	y(t) , definition of Laplace transform
1.2	Y(s)	$y(t) = \frac{1}{j2\pi}\int_{c-j\infty}^{c+j\infty} \exp(st)Y(s)ds$ inversion formula
1.3	$sY(s) - y(0)$	$y'(t)$, first derivative
1.4	$s^2Y(s) - sy(0) - y'(0)$	$y''(t)$, second derivative
1.5	$s^nY(s) - s^{n-1}[y(0)] - s^{n-2}[y'(0)] - \ldots - s[y^{(n-2)}(0)] - [y^{(n-1)}(0)]$	$y^{(n)}(t)$, nth derivative
1.6	$\frac{1}{s}F(s)$	$\int_0^t Y(\tau)d\tau$, integration
1.7	F(s)G(s)	$\int_0^t f(t-\tau)g(\tau)d\tau$, convolution integral
1.8	$\frac{1}{\alpha}F\left(\frac{s}{\alpha}\right)$	f(αt), scaling
1.9	F(s - α)	exp(αt) f(t), shifting in the s plane
1.10	$\frac{1}{\alpha}F\left(\frac{s}{\alpha} - \beta\right)$	$\exp(\alpha\beta t)f(\alpha t)$, combined scaling and shifting

Function Transforms		
N	F(s)	f (t) , t > 0
2.1	1	$\delta(t)$, unit impulse at t = 0
2.2	s	$\frac{d}{dt}\delta(t)$, doublet impulse at t = 0
2.3	$\exp(-\alpha s),\ \alpha \geq 0$	$\delta(t-\alpha)$
2.4	$\frac{1}{s}$	u(t) , unit step
2.5	$\frac{1}{s}\exp(-\alpha s)$	$u(t-\alpha)$
2.6	$\frac{1}{s^2}$	t
2.7a	$\frac{1}{s^n}$, $n = 1,2,3,....$	$\frac{t^{n-1}}{(n-1)!}$
2.7b	$\frac{n!}{s^{n+1}}$, $n = 1,2,3,....$	t^n
2.8	$\frac{1}{s^k}$, k is any real number > 0	$\frac{t^{k-1}}{\Gamma(k)}$, the Gamma function is given in Appendix A
2.9	$\frac{1}{s+\alpha}$	$\exp(-\alpha t)$
2.10	$\frac{1}{(s+\alpha)^2}$	$t\exp(-\alpha t)$

Function Transforms		
N	F(s)	f (t) , t > 0
2.11	$\frac{1}{(s+\alpha)^n}, \quad n = 1, 2, 3, ...$	$\left[\frac{t^{n-1}}{(n-1)!}\right]\exp(-\alpha t)$
2.12	$\frac{\alpha}{s(s+\alpha)}$	$1-\exp(-\alpha t)$
2.13	$\frac{1}{(s+\alpha)(s+\beta)}, \quad \alpha \neq \beta$	$\frac{1}{(\beta-\alpha)}\left[\exp(-\alpha t)-\exp(-\beta t)\right]$
2.14	$\frac{1}{s(s+\alpha)(s+\beta)}, \quad \alpha \neq \beta$	$\frac{1}{\alpha\beta}+\frac{\exp(-\alpha t)}{\alpha(\alpha-\beta)}+\frac{\exp(-\beta t)}{\beta(\beta-\alpha)}$
2.15	$\frac{s}{(s+\alpha)(s+\beta)}, \quad \alpha \neq \beta$	$\frac{1}{(\alpha-\beta)}\left[\alpha\exp(-\alpha t)-\beta\exp(-\beta t)\right]$
2.16a	$\frac{\alpha}{s^2+\alpha^2}$	$\sin(\alpha t)$
2.16b	$\frac{[\sin(\phi)]s+[\cos(\phi)]\alpha}{s^2+\alpha^2}$	$\sin(\alpha t+\phi)$
2.17	$\frac{s}{s^2+\alpha^2}$	$\cos(\alpha t)$
2.18	$\frac{s^2-\alpha^2}{\left(s^2+\alpha^2\right)^2}$	$t\cos(\alpha t)$
2.19	$\frac{1}{s\left(s^2+\alpha^2\right)}$	$\frac{1}{\alpha^2}\left[1-\cos(\alpha t)\right]$
2.20	$\frac{1}{\left(s^2+\alpha^2\right)^2}$	$\frac{1}{2\alpha^3}\left[\sin(\alpha t)-\alpha t\cos(\alpha t)\right]$
2.21	$\frac{s}{\left(s^2+\alpha^2\right)^2}$	$\frac{1}{2\alpha}\left[t\sin(\alpha t)\right]$

Function Transforms		
N	F(s)	f (t) , t > 0
2.22	$\frac{s^2}{\left(s^2+\alpha^2\right)^2}$	$\frac{1}{2\alpha}\left[\sin(\alpha t)+\alpha t\cos(\alpha t)\right]$
2.23	$\frac{1}{\left(s^2+\omega^2\right)\left(s^2+\alpha^2\right)}$, $\alpha \neq \omega$	$\left\{\frac{1}{\omega^2-\alpha^2}\right\}\left\{\frac{1}{\alpha}\sin(\alpha t)-\frac{1}{\omega}\sin(\omega t)\right\}$
2.24	$\frac{\alpha}{s^2(s+\alpha)}$	$t-\frac{1}{\alpha}\left[1-\exp(-\alpha t)\right]$
2.25	$\frac{\beta}{(s+\alpha)^2+\beta^2}$	$\exp(-\alpha t)\sin(\beta t)$
2.26	$\frac{s+\alpha}{(s+\alpha)^2+\beta^2}$	$\exp(-\alpha t)\cos(\beta t)$
2.27	$\frac{s+\lambda}{(s+\alpha)^2+\beta^2}$	$\exp(-\alpha t)\left\{\cos(\beta t)+\left[\frac{\lambda-\alpha}{\beta}\right]\sin(\beta t)\right\}$
2.28	$\frac{s+\alpha}{s^2+\beta^2}$	$\frac{\sqrt{\alpha^2+\beta^2}}{\beta}\sin(\beta t+\phi),\quad \phi=\arctan\left(\frac{\beta}{\alpha}\right)$
2.29	$\frac{1}{s^2-\alpha^2}$	$\frac{1}{\alpha}\sinh(\alpha t)$
2.30	$\frac{s}{s^2-\alpha^2}$	$\cosh(\alpha t)$
2.31	$\arctan\left(\frac{\alpha}{s}\right)$	$\frac{1}{t}\sin(\alpha t)$
2.32	$\frac{1}{\sqrt{s}}$	$\frac{1}{\sqrt{\pi t}}$

Function Transforms		
N	F(s)	f (t) , t > 0
2.33	$\frac{1}{\sqrt{s+\alpha}}$	$\frac{1}{\sqrt{\pi t}}\exp(-\alpha t)$
2.34	$\frac{1}{\sqrt{s^3}}$	$2\sqrt{\frac{t}{\pi}}$
2.35	$\frac{1}{\sqrt{s^2+\alpha^2}}$	$J_o(\alpha t)$, Bessel function given in Appendix A
2.36	$\frac{1}{\left(s^2+\alpha^2\right)^{3/2}}$	$\left(\frac{t}{\alpha}\right)J_1(\alpha t)$
2.37	$\frac{1}{\sqrt{s^2-\alpha^2}}$	$I_o(\alpha t)$, Modified Bessel function given in Appendix A
2.38	$\frac{1}{\left(s^2-\alpha^2\right)^{3/2}}$	$\left(\frac{t}{\alpha}\right)I_1(\alpha t)$
2.39	$\sqrt{s-\alpha}-\sqrt{s-\beta}$	$\frac{1}{2t\sqrt{\pi t}}\left[\exp(\beta t)-\exp(\alpha t)\right]$

References

1. Jan Tuma, Engineering Mathematics Handbook, McGraw-Hill, New York, 1979.
2. F. Oberhettinger and L. Badii, Table of Laplace Transforms, Springer-Verlag, N.Y.,1972.
3. M. Abramowitz and I. Stegun, editors, Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables, National Bureau of Standards, Washington, D.C., 1964.